5 Common Functional Implied Volatility Analysis

Michal Benko and Wolfgang Härdle

5.1 Introduction

Trading, hedging and risk analysis of complex option portfolios depend on accurate pricing models. The modelling of implied volatilities (IV) plays an important role, since volatility is the crucial parameter in the Black-Scholes (BS) pricing formula. It is well known from empirical studies that the volatilities implied by observed market prices exhibit patterns known as volatility smiles or smirks that contradict the assumption of constant volatility in the BS pricing model. On the other hand, the IV is a function of two parameters: the strike price and the time to maturity and it is desirable in practice to reduce the dimension of this object and characterize the IV surface through a small number of factors. Clearly, a dimension reduced pricing-model that should reflect the dynamics of the IV surface needs to contain factors and factor loadings that characterize the IV surface itself and their movements across time.

A popular dimension reduction technique is the principal components analysis (PCA), employed for example by Fengler, Härdle, and Schmidt (2002) in the IV surface analysis. The discretization of the strike dimension and application of PCA yield suitable factors (weight vectors) in the multivariate framework. Noting that the IVs of fixed maturity could also be viewed as random functions, we propose to use the functional analogue of PCA. We utilize the truncated functional basis expansion described in Ramsay and Silverman (1997) to the IVs of the European options on the German stock index (DAX). The standard functional PCA, however, yields weight functions that are too rough, hence a smoothed version of functional PCA is proposed here.

Like Fengler, Härdle, and Villa (2003) we discover similarities of the resulting weight functions across maturity groups. Thus we propose an estimation procedure based on the Flury-Gautschi algorithm, Flury (1988), for the simultaneous estimation of the weight functions for two different maturities. This procedure yields common weight functions with the level, slope, and curvature interpretation known from the financial literature. The resulting common factors of the IV surface are the basic elements to be used in applications, such as simulation based pricing, and deliver a substantial dimension reduction.

The chapter is organized as follows. In Section 5.2 the basic financial framework is presented, while in Section 5.3 we introduce the notation of the functional data analysis. In the following three sections we analyze the IV functions using functional principal components, smoothed functional principal components and common estimation of principal components, respectively.

5.2 Implied volatility surface

Implied volatilities are derived from the BS pricing formula for European options. Recall that European call and put options are derivatives written on an underlying asset S driven by the price process S_t, which yield the pay-off $\max(S_T - K, 0)$ and $\max(K - S_T, 0)$ respectively, at a given expiry time T and for a prespecified strike price K. The difference $\tau = T - t$ between the day of trade and day of expiration (maturity) is called time to maturity. The pricing formula for call options, Black and Scholes (1973), is:

$$C_t(S_t, K, \tau, r, \sigma) = S_t \Phi(d_1) - K e^{-r\tau} \Phi(d_2)$$

$$d_1 = \frac{\ln(S_t/K) + (r + 1/2\sigma^2)\tau}{\sigma\sqrt{\tau}}, \ d_2 = d_1 - \sigma\sqrt{\tau},$$

(5.1)

where $\Phi(\cdot)$ is the cumulative distribution function of the standard normal distribution, r is the riskless interest rate, and σ is the (unknown and constant) volatility parameter. The put option price P_t can be obtained from the put-call parity $P_t = C_t - S_t + e^{-\tau r} K$.

For a European option the implied volatility $\hat{\sigma}$ is defined as the volatility $-\sigma$, which yields the BS price C_t equal to the price $\tilde{C}_t$ observed on the market. For a single asset, we obtain at each time point t a two-dimensional function – the IV surface $\hat{\sigma}_t(K, \tau)$. In order to standardize the volatility functions in time, one
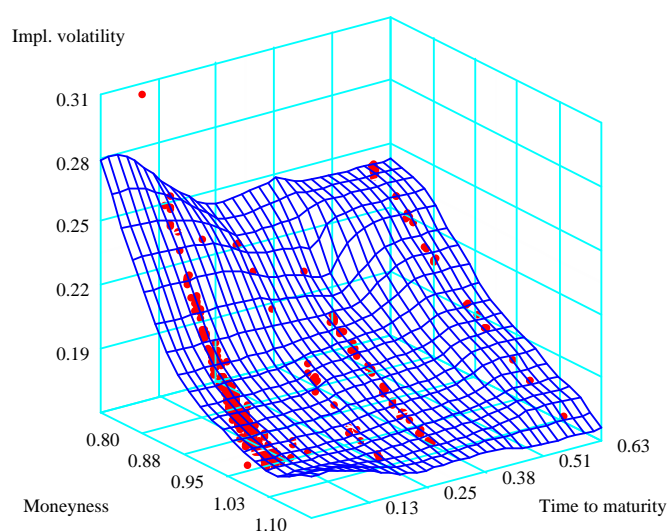
Volatility Surface

Figure 5.1: Implied volatility surface of ODAX on May 24, 2001.

STFfda01.xpl

may rescale the strike dimension by dividing K by the future price $F_t(\tau)$ of the underlying asset with the same maturity. This yields the so-called moneyness $\kappa = K/F_t(\tau)$. Note that some authors define moneyness simply as $\kappa = K/S_t$. In contrast to the BS assumptions, empirical studies show that IV surfaces are significantly curved, especially across the strikes. This phenomenon is called a volatility smirk or smile. Smiles stand for U-shaped volatility functions and smirks for decreasing volatility functions.

We focus on the European options on the German stock index (ODAX). Figure 5.1 displays the ODAX implied volatilities computed from the BS formula (red points) and the IV surface on May 24, 2001 estimated using a local polynomial

estimator for $\tau \in [0, 0.6]$ and $\kappa \in [0.8, 1.2]$. We can clearly observe the "strings" of the original data on maturity grid $\tau \in \{0.06111, 0.23611, 0.33333, 0.58611\}$, which corresponds to 22, 85, 120, and 211 days to maturity. This maturity grid is structured by market conventions and changes over time. The fact that the number of transactions with short maturity is much higher than those with longer maturity is also typical for the IVs observed on the market.

The IV surface is a high-dimensional object – for every time point t we have to analyze a two-dimensional function. Our goal is to reduce the dimension of this problem and to characterize the IV surface through a small number of factors. These factors can be used in practice for risk management, e.g. with vega-strategies.

The analyzed data, taken from MD*Base, contains EUREX intra-day transaction data for DAX options and DAX futures (FDAX) from January 2 to June 29, 2001. The IVs are calculated by the Newton-Raphson iterative method. The correction of Hafner and Wallmeier (2001) is applied to avoid influence of the tax-scheme in the DAX. For robustness, we exclude the contracts with time to maturity of less than 7 days and maturity strings with less than 100 observations. The approximation of the "riskless" interest rate with a given maturity is obtained on a daily basis from the linear interpolation of the 1, 3, 6, and 12 month EURIBOR interest rates (obtained from Datastream).

The resulting data set is analyzed using the functional data analysis framework. One advantage of this approach, as we will see later in this chapter, is the possibility of introducing smoothness in the functional sense and using it for regularization. The notation of the functional data analysis is rather complex, therefor the theoretical motivation and the basic notation will be introduced in the next section.

5.3   Functional data analysis

In the functional data framework, the objects are usually modelled as realizations of a stochastic process $X(t), t \in J$, where J is a bounded interval in $\mathbb{R}$. Thus, the set of functions

$$x_i(t), \ i = 1, 2, \ldots n, \ t \in J,$$

represents the data set. We assume the existence of the mean, variance, and covariance functions of the process $X(t)$ and denote these by $\mathrm{E}X(t)$, $\mathrm{Var}(t)$ and $\mathrm{Cov}(s, t)$ respectively.

For the functional sample we can define the sample-counterparts of $\mathrm{E}X(t)$, $\mathrm{Var}(t)$ and $\mathrm{Cov}(s,t)$ in a straightforward way:

$$
\begin{aligned}
\bar{X}(t) &= \tfrac{1}{n} \sum_{i=1}^{n} x_i(t), \\
\widehat{\mathrm{Var}}(t) &= \tfrac{1}{n-1} \sum_{i=1}^{n} \left\{ x_i(t) - \bar{X}(t) \right\}^2, \\
\widehat{\mathrm{Cov}}(s,t) &= \tfrac{1}{n-1} \sum_{i=1}^{n} \left\{ x_i(s) - \bar{X}(s) \right\} \left\{ x_i(t) - \bar{X}(t) \right\}.
\end{aligned}
$$

In practice, we observe the function values $\mathcal{X} \stackrel{\text{def}}{=} \{x_i(t_{i1}), x_i(t_{i2}), \ldots, x_i(t_{ip_i}); i = 1, \ldots, n\}$ only on a discrete grid $\{t_{i1}, t_{i2}, \ldots, t_{ip_i}\} \in J$, where p_i is the number of grid points for the ith observation. One may estimate the functions $x_1, \ldots, x_n$ via standard nonparametric regression methods, Härdle (1990). Another popular way is to use a truncated functional basis expansion. More precisely, let us denote a functional basis on the interval J by $\{\Theta_1, \Theta_2, \ldots, \}$ and assume that the functions x_i are approximated by the first L basis functions $\Theta_l, l = 1, 2, \ldots, L$:

$$
x_i(t) = \sum_{l=1}^{L} c_{il} \Theta_l(t) = \mathbf{c}_i^\top \boldsymbol{\Theta}(t), \tag{5.2}
$$

where $\boldsymbol{\Theta} = (\Theta_1, \ldots, \Theta_L)^\top$ and $\mathbf{c}_i = (c_{i1}, \ldots, c_{iL})^\top$. The number of basis functions L determines the tradeoff between data fidelity and smoothness. The analysis of the functional objects will be implemented through the coefficient matrix

$$
\mathbf{C} = \{c_{il}, \ i = 1, \ldots, n, \ l = 1, \ldots, L\}.
$$

The mean, variance, and covariance functions are calculated by:

$$
\begin{aligned}
\bar{X}(t) &= \bar{\mathbf{c}}^\top \boldsymbol{\Theta}(t), \\
\widehat{\mathrm{Var}}(t) &= \boldsymbol{\Theta}(t)^\top \mathrm{Cov}(\mathbf{C}) \boldsymbol{\Theta}(t), \\
\widehat{\mathrm{Cov}}(s,t) &= \boldsymbol{\Theta}(s)^\top \mathrm{Cov}(\mathbf{C}) \boldsymbol{\Theta}(t),
\end{aligned}
$$

where $\bar{\mathbf{c}}_l \stackrel{\text{def}}{=} \tfrac{1}{n} \sum_{i=1}^{n} c_{il}, \ l = 1, \ldots, L$ and $\mathrm{Cov}(\mathbf{C}) \stackrel{\text{def}}{=} \tfrac{1}{n-1} \sum_{i=1}^{n} (\mathbf{c}_i - \bar{\mathbf{c}})(\mathbf{c}_i - \bar{\mathbf{c}})^\top$.
The scalar product in the functional space is defined by:

$$
\langle x_i, x_j \rangle \stackrel{\text{def}}{=} \int_J x_i(t) x_j(t) dt = \mathbf{c}_i^\top \mathbf{W} \mathbf{c}_j,
$$

where

$$\mathbf{W} \stackrel{\text{def}}{=} \int_J \mathbf{\Theta}(t)\mathbf{\Theta}(t)^\top \, dt. \tag{5.3}$$

In practice, the coefficient matrix $\mathbf{C}$ needs to be estimated from the data set $\mathcal{X}$.

An example for a functional basis is the Fourier basis defined on J by:

$$\Theta_l(t) = \left\{ \begin{array}{ll} 1, & l = 0, \\ \sin(r\omega t), & l = 2r - 1, \\ \cos(r\omega t), & l = 2r, \end{array} \right.$$

where the frequency ω determines the period and the length of the interval $|J| = 2\pi/\omega$. The Fourier basis defined above can be easily transformed to the orthonormal basis, hence the scalar-product matrix in (5.3) is simply the identity matrix.

Our aim is to estimate the IV-functions for fixed $\tau = 1$ month (1M) and 2 months (2M) from the daily-specific grid of the maturities. We estimate the Fourier coefficients on the moneyness-range $\kappa \in [0.9, 1.1]$ for maturities observed on particular day i. For $\tau^* = $ 1M, 2M we calculate $\hat{\sigma}_i(\kappa, \tau^*)$ by linear interpolation of the closest observable IV string with $\tau \le \tau^*$, $\widehat{\sigma}_i(\kappa, \tau^*_{i-})$ and $\tau \ge \tau^*$, $\widehat{\sigma}_i(\kappa, \tau^*_{i+})$:

$$\hat{\sigma}_i(\kappa, \tau^*) = \hat{\sigma}_i(\kappa, \tau^*_{i-}) \left(1 - \frac{\tau^* - \tau^*_{i-}}{\tau^*_{i+} - \tau^*_{i-}} \right) + \hat{\sigma}_i(\kappa, \tau^*_{i+}) \left(\frac{\tau^* - \tau^*_{i-}}{\tau^*_{i+} - \tau^*_{i-}} \right),$$

for i where τ^*_{i-} and τ^*_{i-} exist. In Figure 5.2 we show the situation for $\tau^* = $1M on May 30, 2001. The blue points and the blue finely dashed curve correspond to the transactions with $\tau^*_- = $16 days and the green points and the green dashed curve to the transactions with $\tau^*_+ = 51$ days. The solid black line is the linear interpolation at $\tau^* = 30$ days.

The choice of $L = 9$ delivers a good tradeoff between flexibility and smoothness of the strings. At this moment we exclude from our analysis those days, where this procedure cannot be performed due to the complete absence of the needed maturities, and strings with poor performance of estimated coefficients, due to the small number of contracts in a particular string or presence of strong outliers. Using this procedure we obtain 77 "functional" observations $x^{1M}_{i_1}(\kappa) \stackrel{\text{def}}{=} \hat{\sigma}_{i_1}(\kappa, 1M)$, $i_1 = 1, \ldots, 77$, for the 1M-maturity and 66 observations $x^{2M}_{i_2}(\kappa) \stackrel{\text{def}}{=} \hat{\sigma}_{i_2}(\kappa, 2M)$, $i_2 = 1, \ldots, 66$, for the 2M-maturity, as displayed in Figure 5.3.
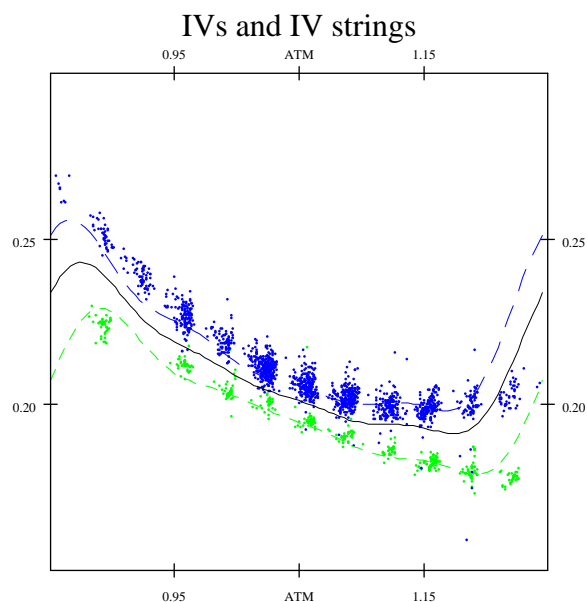
Figure 5.2: Linear interpolation of IV strings on May 30, 2001 with $L = 9$.

STFfda02.xpl

5.4 Functional principal components

Principal Components Analysis yields dimension reduction in the multivariate framework. The idea is to find normalized weight vectors $\gamma_m \in \mathbb{R}^p$, for which the linear transformations of a p-dimensional random vector $\mathbf{x}$, with $\mathrm{E}[\mathbf{x}] = 0$:

$$f_m = \gamma_m^\top \mathbf{x} = \langle \gamma_m, \mathbf{x} \rangle, \ m = 1, \ldots, p, \tag{5.4}$$

have maximal variance subject to:

$$\gamma_l^\top \gamma_m = \langle \gamma_l, \gamma_m \rangle = I(l = m) \text{ for } l \leq m.$$

Where I denotes the identificator function. The solution is the Jordan spectral decomposition of the covariance matrix, Härdle and Simar (2003).

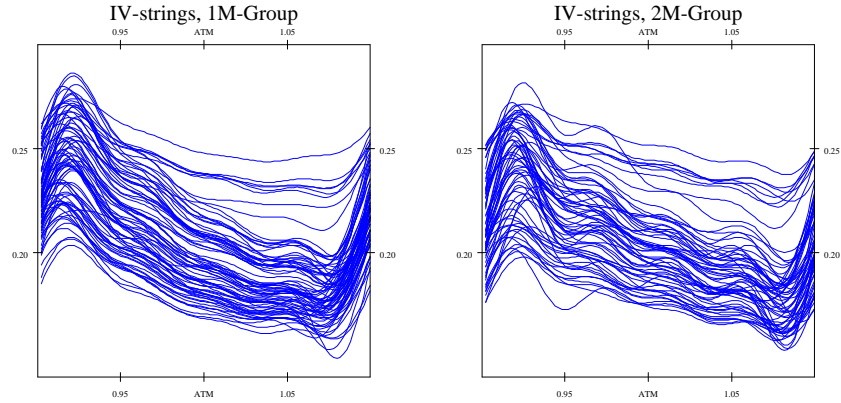IV-strings, 1M-Group   IV-strings, 2M-Group

Figure 5.3: Functional observations estimated using Fourier basis with $L = 9$, $\hat{\sigma}_{i_1}(\kappa, 1M)$, $i_1 = 1, \ldots, 77$, in the left panel, $\hat{\sigma}_{i_2}(\kappa, 2M)$ $i_2 = 1, \ldots, 66$ in the right panel.

STFfda03.xpl

In the Functional Principal Components Analysis (FPCA) the dimension reduction can be achieved via the same route, i.e. by finding orthonormal weight functions $\gamma_1, \gamma_2, \ldots$, such that the variance of the linear transformation is maximal. In order to keep notation simple we assume $EX(t) = 0$. The weight functions satisfy:

$$
\begin{aligned}
||\gamma_m||^2 &= \int \gamma_m(t)^2 dt = 1, \\
\langle \gamma_l, \gamma_m \rangle &= \int \gamma_l(t)\gamma_m(t)dt = 0, \; l \neq m.
\end{aligned}
$$

The linear transformation is:

$$
f_m = \langle \gamma_m, X \rangle = \int \gamma_m(t)X(t)dt,
$$

and the desired weight functions solve:

$$
\underset{\langle \gamma_l, \gamma_m \rangle = I(l=m), l \leq m}{\arg \max} \operatorname{Var} \langle \gamma_m, X \rangle, \tag{5.5}
$$

or equivalently:

$$\underset{\langle \gamma_l, \gamma_m \rangle = I(l=m), l \leq m}{\arg\max} \int \int \gamma_m(s) \mathrm{Cov}(s,t) \gamma_m(t) ds dt.$$

The solution is obtained by solving the Fredholm functional eigenequation

$$\int \mathrm{Cov}(s,t) \gamma(t) dt = \lambda \gamma(s). \tag{5.6}$$

The eigenfunctions $\gamma_1, \gamma_2, \ldots$ sorted with respect to the corresponding eigenvalues $\lambda_1 \geq \lambda_2 \geq \ldots$ solve the FPCA problem (5.5). The following link between eigenvalues and eigenfunctions holds:

$$\lambda_m = \mathrm{Var}(f_m) = \mathrm{Var}\left[\int \gamma_m(t) X(t) dt\right] = \int \int \gamma_m(s) \mathrm{Cov}(s,t) \gamma_m(t) ds dt.$$

In the sampling problem, the unknown covariance function $\mathrm{Cov}(s,t)$ needs to be replaced by the sample covariance function $\widehat{\mathrm{Cov}}(s,t)$. Dauxois, Pousse, and Romain (1982) show that the eigenfunctions and eigenvalues are consistent estimators for λ_m and γ_m and derive some asymptotic results for these estimators.

5.4.1   Basis expansion

Suppose that the weight function γ has expansion

$$\gamma = \sum_{l=1}^{L} \mathbf{b}_l \Theta_l(t) = \mathbf{\Theta}^\top \mathbf{b}.$$

Using this notation we can rewrite the left hand side of eigenequation (5.6):

$$\int \mathrm{Cov}(s,t) \gamma(t) dt = \int \mathbf{\Theta}(s)^\top \mathrm{Cov}(\mathbf{C}) \mathbf{\Theta}(t) \mathbf{\Theta}(t)^\top \mathbf{b} dt$$
$$= \mathbf{\Theta}^\top \mathrm{Cov}(\mathbf{C}) \mathbf{W} \mathbf{b},$$

so that:

$$\mathrm{Cov}(\mathbf{C}) \mathbf{W} \mathbf{b} = \lambda \mathbf{b}.$$

The functional scalar product $\langle \gamma_l, \gamma_k \rangle$ corresponds to $\mathbf{b}_l^\top \mathbf{W} \mathbf{b}_k$ in the truncated basis framework, in the sense that if two functions γ_l and γ_k are orthogonal, the corresponding coefficient vectors $\mathbf{b}_l, \mathbf{b}_k$ satisfy $\mathbf{b}_l^\top \mathbf{W} \mathbf{b}_k = 0$. Matrix $\mathbf{W}$ is
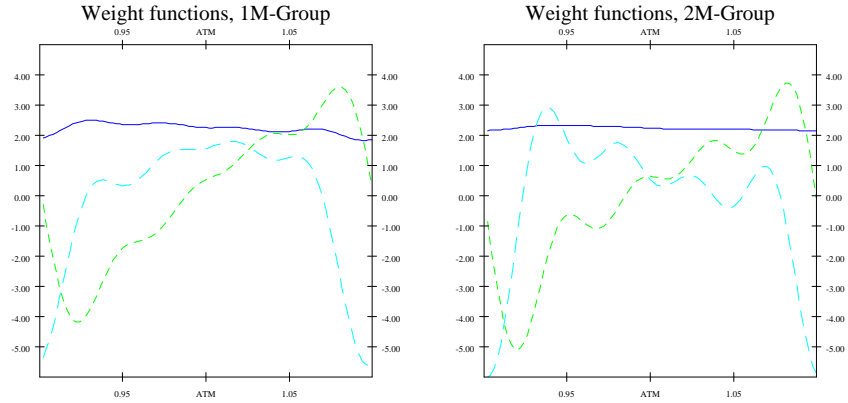
Figure 5.4: Weight functions for $1M$ and $2M$ maturity groups. Blue solid lines, $\hat{\gamma}_1^{1M}$ and $\hat{\gamma}_1^{2M}$, are the first eigenfunctions, green finely dashed lines, $\hat{\gamma}_2^{1M}$ and $\hat{\gamma}_2^{2M}$, are the second eigenfunctions, and cyan dashed lines, $\hat{\gamma}_3^{1M}$ and $\hat{\gamma}_3^{2M}$, are the third eigenfunctions.

STFfda04.xpl

symmetric by definition. Thus, defining $\mathbf{u} = \mathbf{W}^{1/2}\mathbf{b}$, one needs to solve finally a symmetric eigenvalue problem:

$$\mathbf{W}^{1/2}\text{Cov}(\mathbf{C})\mathbf{W}^{1/2}\mathbf{u} = \lambda\mathbf{u},$$

and to compute the inverse transformation $\mathbf{b} = \mathbf{W}^{-1/2}\mathbf{u}$. For the orthonormal functional basis (i.e. also for the Fourier basis) $\mathbf{W} = \mathbf{I}$, i.e. the problem of FPCA is reduced to the multivariate PCA performed on the matrix $\mathbf{C}$.

Using the FPCA method on the IV-strings for 1M and 2M maturities we obtain the eigenfunctions plotted in Figure 5.4. It can be seen, that the eigenfunctions are too rough. Intuitively, this roughness is caused by the flexibility of the functional basis. In the next section we present a way of incorporating the smoothing directly into the PCA problem.

5.5   Smoothed principal components analysis

As we can see in Figure 5.4, the resulting eigenfunctions are often very rough. Smoothing them could result in a more natural interpretation of the obtained weight functions. Here we apply a popular approach known as roughness penalty. The downside of this technique is that we loose orthogonality in the L^2 sense.

Assume that the underlying eigenfunctions of the covariance operator have a continuous and square-integrable second derivative. Let $\mathcal{D}\gamma = \gamma'(t)$ be the first derivative operator and define the roughness penalty by $\Psi(\gamma) = ||\mathcal{D}^2\gamma||^2$. Moreover, suppose that γ_m has square-integrable derivatives up to degree four and that the second and the third derivatives satisfy one of the following conditions:

1. $\mathcal{D}^2\gamma$, $\mathcal{D}^3\gamma$ are zero at the ends of the interval J,

2. the periodicity boundary conditions of γ,$\mathcal{D}\gamma$, $\mathcal{D}^2\gamma$, and $\mathcal{D}^3\gamma$ on J.

Then we can rewrite the roughness penalty in the following way:

$$
\begin{aligned}
||\mathcal{D}^2\gamma||^2 &= \int \mathcal{D}^2\gamma(s)\mathcal{D}^2\gamma(s)ds \\
&= \mathcal{D}\gamma(u)\mathcal{D}^2\gamma(u) - \mathcal{D}\gamma(d)\mathcal{D}^2\gamma(d) - \int \mathcal{D}\gamma(s)\mathcal{D}^3\gamma(s)ds \quad (5.7) \\
&= \gamma(u)\mathcal{D}^3\gamma(u) - \gamma(d)\mathcal{D}^3\gamma(d) - \int \gamma(s)\mathcal{D}^4\gamma(s)ds \quad (5.8) \\
&= \langle \gamma, \mathcal{D}^4\gamma \rangle, \quad (5.9)
\end{aligned}
$$

where d and u are the boundaries of the interval J and the first two elements in (5.7) and (5.8) are both zero under any of the conditions mentioned above.

Given a eigenfunction γ with norm $||\gamma||^2 = 1$, we can penalize the sample variance of the principal component by dividing it by $1 + \alpha\langle \gamma, \mathcal{D}^4\gamma \rangle$:

$$
PCAPV \stackrel{\text{def}}{=} \frac{\int \int \gamma(s)\widehat{\text{Cov}}(s,t)\gamma(t)dsdt}{\int \gamma(t)(\mathcal{I} + \alpha\mathcal{D}^4)\gamma(t)dt}, \quad (5.10)
$$

where $\mathcal{I}$ denotes the identity operator. The maximum of the penalized sample variance (PCAPV) is an eigenfunction γ corresponding to the largest eigenvalue of the generalized eigenequation:

$$
\int \widehat{\text{Cov}}(s,t)\gamma(t)dt = \lambda(\mathcal{I} + \alpha\mathcal{D}^4)\gamma(s). \quad (5.11)
$$

As already mentioned above, the resulting weight functions (eigenfunctions) are no longer orthonormal in the L^2 sense. Since the weight functions are used as smoothed estimators of principal components functions, we need to rescale them to satisfy $||\gamma_l||^2 = 1$. The weight functions γ_l can be also interpreted as orthogonal in the modified scalar product of the Sobolev type

$$(f, g) \stackrel{\text{def}}{=} \langle f, g \rangle + \alpha \langle \mathcal{D}^2 f, \mathcal{D}^2 g \rangle.$$

A more extended theoretical discussion can be found in Silverman (1991).

5.5.1   Basis expansion

Define $\mathbf{K}$ to be a matrix whose elements are $\langle \mathcal{D}^2 \Theta_j, \mathcal{D}^2 \Theta_k \rangle$. Then the generalized eigenequation (5.11) can be transformed to:

$$\mathbf{W} \operatorname{Cov}(\mathbf{C}) \mathbf{W} \mathbf{u} = \lambda (\mathbf{W} + \alpha \mathbf{K}) \mathbf{u}. \tag{5.12}$$

Using Cholesky factorization $\mathbf{L} \mathbf{L}^\top = \mathbf{W} + \alpha \mathbf{K}$ and defining $\mathbf{S} = \mathbf{L}^{-1}$ we can rewrite (5.12) as:

$$\{\mathbf{S} \mathbf{W} \operatorname{Cov}(\mathbf{C}) \mathbf{W} \mathbf{S}^\top\} (\mathbf{L}^\top \mathbf{u}) = \lambda \mathbf{L}^\top \mathbf{u}.$$

Applying Smoothed Functional PCA (SPCA) to the IV-strings, we get the smooth-eigenfunctions plotted in Figure 5.5. We use $\alpha = 10^{-7}$, the aim is to use a rather small degree of smoothing, in order to replace the high frequency fluctuations only. Some popular methods, like cross-validation, could be employed as well, Ramsay and Silverman (1997).

The interpretation of the weight functions displayed in Figure 5.5 is as follows: The first weight function (solid blue) represents clearly the level of the volatility – weights are almost constant and positive. The second weight function (finely dashed green) changes sign near the at-the-money point, i.e. can be interpreted as the in-the-money/out-of-the-money identification factor or slope. The third (dashed cyan) weight function may play the part of the measure for a deep in-the-money or out-of-the-money factor or curvature. It can be seen that the weight functions for the 1M $(\widetilde{\gamma}_1^{1M}, \widetilde{\gamma}_2^{1M}, \widetilde{\gamma}_3^{1M})$ and 2M maturities $(\widetilde{\gamma}_1^{2M}, \widetilde{\gamma}_2^{2M}, \widetilde{\gamma}_3^{2M})$ have a similar structure. From a practical point of view it can be interesting to try to get common estimated eigenfunctions (factors in the further analysis) for both groups. In the next section, we introduce the estimation motivated by the Common Principal Component Model.
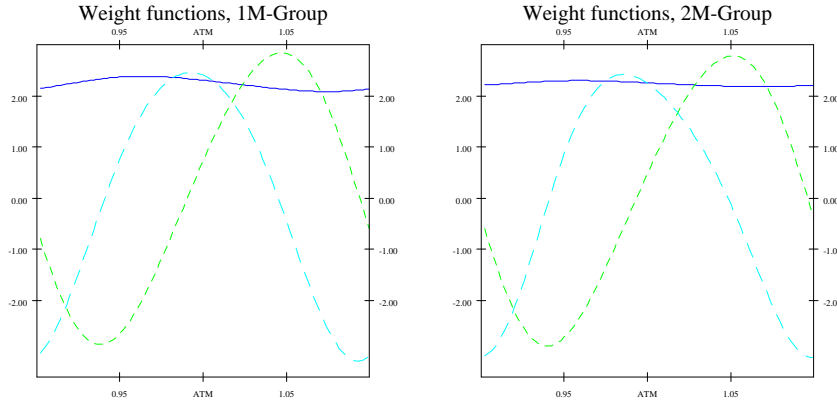
Figure 5.5: Smoothed weight functions with $\alpha = 10^{-7}$. Blue solid lines, $\hat{\gamma}_1^{1M}$ and $\hat{\gamma}_1^{2M}$, are the first eigenfunctions, green finely dashed lines, $\hat{\gamma}_2^{1M}$ and $\hat{\gamma}_2^{2M}$, are the second eigenfunctions, and cyan dashed lines, $\hat{\gamma}_3^{1M}$ and $\hat{\gamma}_3^{2M}$, are the third eigenfunctions.

STFfda05.xpl

5.6 Common principal components model

The Common Principal Components model (CPC) in the multivariate setting can be motivated as the model for similarity of the covariance matrices in the k-sample problem, Flury (1988). Having k random vectors, $\mathbf{x}_{(1)}, \mathbf{x}_{(2)}, \dots, \mathbf{x}_{(k)} \in \mathbb{R}^p$ the CPC-Model can be written as:

$$\boldsymbol{\Psi}_j \stackrel{\text{def}}{=} \mathrm{Cov}(\mathbf{x}_{(j)}) = \boldsymbol{\Gamma}\boldsymbol{\Lambda_j}\boldsymbol{\Gamma}^\top,$$

where $\boldsymbol{\Gamma}$ is an orthogonal matrix and $\boldsymbol{\Lambda}_j = \mathrm{diag}(\lambda_{i1}, \dots, \lambda_{ip})$. This means that eigenvectors are the same across samples and just the eigenvalues – variances of the principal component scores (5.4) differ.

Using the normality assumption, the sample covariance matrices $\mathbf{S}_j, j = 1, \dots, k$, are Wishart-distributed:

$$\mathbf{S}_j \sim W_p(n_j, \boldsymbol{\Psi}_j/n_j),$$

and the CPC model can be estimated using maximum likelihood estimation with likelihood-function:

$$\text{L}(\boldsymbol{\Psi}_1, \boldsymbol{\Psi}_2, \ldots, \boldsymbol{\Psi}_k) = C \prod_{j=1}^{k} \exp\left\{ \text{tr}\left(-\frac{n_j}{2} \boldsymbol{\Psi}_j^{-1} \mathbf{S}_j \right) \right\} (\det \boldsymbol{\Psi_j})^{-n_j/2}.$$

Here C is a factor that does not depend on the parameters and n_j is the number of observations in group j. The maximization of this likelihood function is equivalent to:

$$\prod_{j=1}^{k} \left\{ \frac{\det \text{diag}(\boldsymbol{\Gamma}^\top \mathbf{S_j} \boldsymbol{\Gamma})}{\det(\boldsymbol{\Gamma}^\top \mathbf{S_j} \boldsymbol{\Gamma})} \right\}^{n_j}, \tag{5.13}$$

and the maximization of this criterion is performed by the so-called Flury-Gautschi(FG)-algorithm, Flury (1988).

As shown in Section 5.4, using the functional basis expansion, the FPCA and SPCA are basically implemented via the spectral decomposition of the "weighted" covariance matrix of the coefficients. In view of the minimization property of the FG algorithm, the diagonalization procedure optimizing the criterion (5.13) can be employed. However, the obtained estimates may not be maximum likelihood estimates.

Using this procedure for the IV-strings of 1M and 2M maturity we get "common" smoothed eigenfunctions. The first three common eigenfunctions ($\widetilde{\gamma}_1^c$, $\widetilde{\gamma}_2^c$, $\widetilde{\gamma}_3^c$) are displayed in Figures 5.6–5.8. The solid blue curve represents the estimated eigenfunction for the 1M maturity, the finely dashed green curve for the 2M maturity and the dashed black curve is the common eigenfunction estimated by the FG-algorithm.

Assuming that $\hat{\sigma}_i(\kappa, \tau)$ are centered for $\tau = 1M$ and $2M$ (we subtract the sample mean of corresponding group from the estimated functions), we may use the obtained weight functions in the factor model of the IV dynamics of the form:

$$\widetilde{\sigma}_i(\kappa, \tau) = \sum_{j=1}^{R} \widetilde{\gamma}_j^c(\kappa) \langle \widetilde{\gamma}_j^c(\kappa), \hat{\sigma}_i(\kappa, \tau) \rangle, \tag{5.14}$$

where $\tau \in \{1M, 2M\}$ and R is the number of factors. Thus $\widetilde{\sigma}_i$ is an alternative estimation of σ_i. This factor model can be used for simulation applications like Monte Carlo VaR. Especially the usage of Common Principal Components $\widetilde{\gamma}_j^c(\kappa)$ reduces the high-dimensional IV-surface problem to a small number of functional factors.
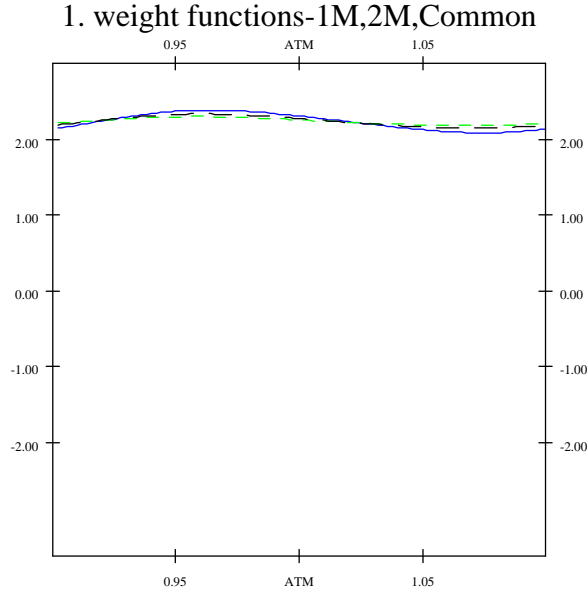
1. weight functions-1M,2M,Common

Figure 5.6: First weight functions, $\alpha = 10^{-7}$, solid blue line is the weight function of the 1M maturity group $(\hat{\gamma}_1^{1M})$, finely dashed green line of the 2M maturity group $(\hat{\gamma}_1^{2M})$, and dashed black line is the common eigenfunction $(\widetilde{\gamma}_1^c)$, estimated from both groups.

In addition, an econometric approach, successfully employed by Fengler, Härdle, and Mammen (2004) can be employed. It consists of fitting an appropriate model to the time series of the estimated principal component scores, $\widetilde{f}_{ij}^c(\tau) = \langle \widetilde{\gamma}_j^c(\kappa), \ \hat{\sigma}_i(\kappa, \tau) \rangle$, as displayed in Figure 5.9. Note that $\hat{\sigma}_i(\kappa, \tau)$ are centered again (sample means are zero). The fitted time series model can be used for forecasting future IV functions.

There are still some open questions related to this topic. First of all, the practitioner would be interested in a good automated choice of the parameters of our method (dimension of the truncated functional basis L and smoothing parameter α). The application of the Fourier coefficients in this framework seems to be reasonable for the volatility smiles (U-shaped strings), however for the volatility smirks (typically monotonically decreasing strings) the performance
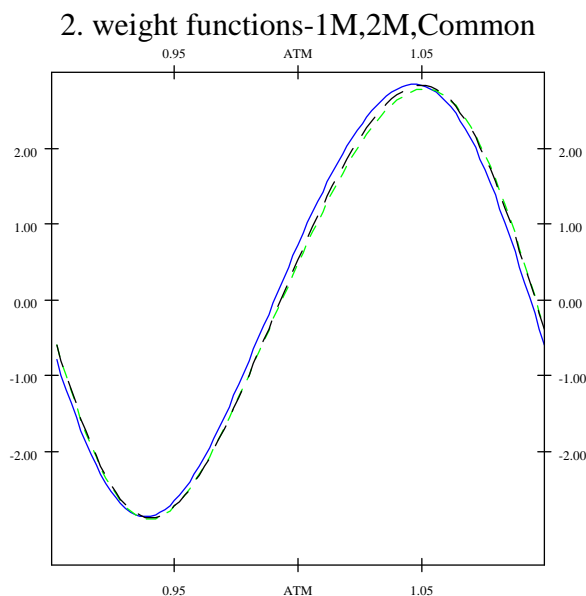
2. weight functions-1M,2M,Common

Figure 5.7: Second eigenfunctions, $\alpha = 10^{-7}$, solid blue line is the weight function of the 1M maturity group $(\hat{\gamma}_2^{1M})$, finely dashed green line of the 2M maturity group $(\hat{\gamma}_2^{2M})$, and dashed black line is the common eigenfunction $(\widetilde{\gamma}_2^c)$, estimated from both groups.

is rather bad. In particular, the variance of our functional objects and the shape of our weight functions at the boundaries is affected. The application of regression splines in this setting seems to be promising, but it increases the number of smoothing parameters by the number and the choice of the knots – problems which are not generally easy to deal with. The next natural question, which is still open concerns the statistical properties of the technique and the testing procedure for the Functional Common PCA model. Finally, using the data for a longer time period one may also analyze the longer maturities like 3 months or 6 months.

3. weight functions-1M,2M,Common
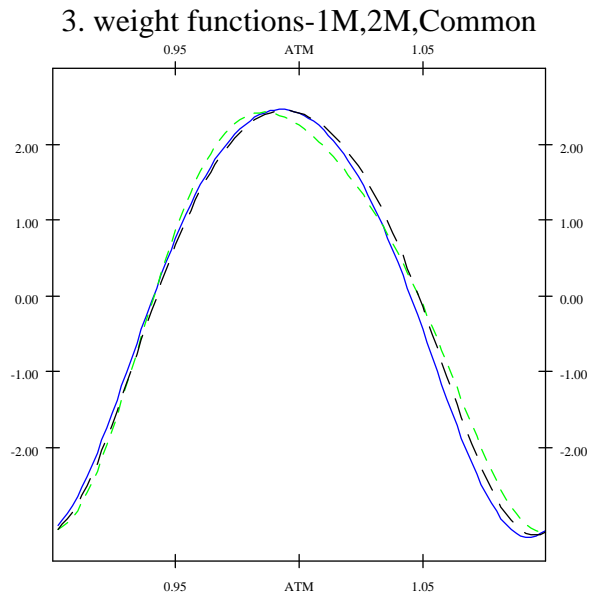
Figure 5.8: Third eigenfunctions, $\alpha = 10^{-7}$, solid blue line is the weight function of the 1M maturity group $(\hat{\gamma}_3^{1M})$, finely dashed green line of the 2M maturity group $(\hat{\gamma}_3^{2M})$, and dashed black line is the common eigenfunction $(\widetilde{\gamma}_3^c)$, estimated from both groups.

Figure 5.9: Estimated principal component scores, $\widetilde{f}^c_{i1}(1M)$, $\widetilde{f}^c_{i2}(1M)$, and $\widetilde{f}^c_{i3}(1M)$ for 1M maturity – first row, and $\widetilde{f}^c_{i1}(2M)$, $\widetilde{f}^c_{i2}(2M)$, and $\widetilde{f}^c_{i3}(2M)$ for 2M maturity – second row; $\alpha = 10^{-7}$.

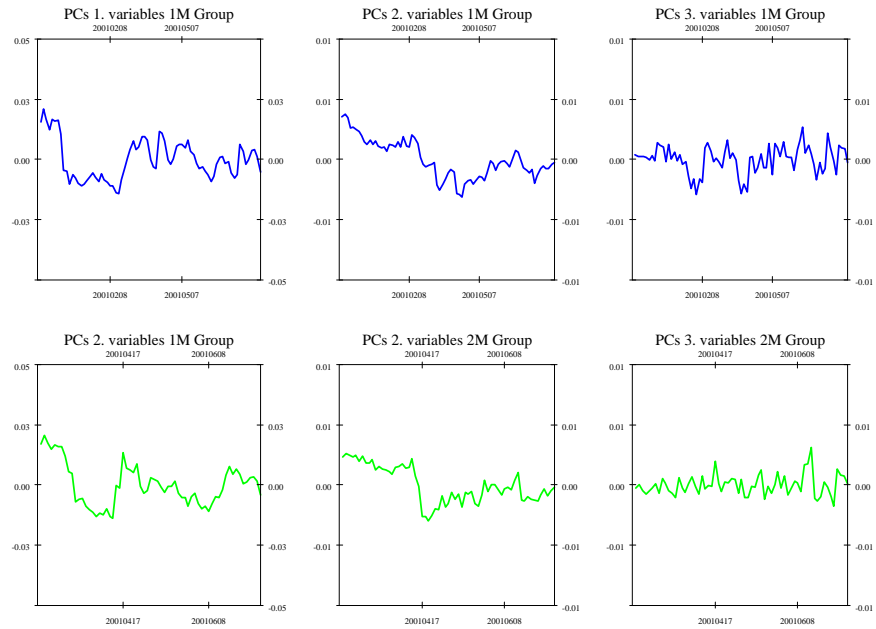# Bibliography

Black, F. and Scholes, M. (1973). The pricing of options and corporate liabilities, *Journal of Political Economy*, **81**: 637:654.

Dauxois, J., Pousse, A., and Romain, Y. (1982). Asymptotic Theory for the Principal Component Analysis of a Vector Random Function: Some Applications to Statistical Inference, *Journal of Multivariate Analysis* **12**: 136-154.

Flury, B. (1988). *Common Principal Components and Related Models*, Wiley, New York.

Fengler, M., Härdle, W., and Schmidt, P. (2002). Common Factors Governing VDAX Movements and the Maximum Loss, *Journal of Financial Markets and Portfolio Management* **16(1)**: 16-29.

Fengler, M., Härdle, W., and Villa, P. (2003). The Dynamics of Implied Volatilities: A common principle components approach, *Review of Derivative Research* **6**: 179-202.

Fengler, M., Härdle, W., and Mammen, E. (2004). Implied Volatility String Dynamics, CASE Discussion Paper, http://www.case.hu-berlin.de.

Föllmer, H. and Schied A. (2002). *Stochastic Finance*, Walter de Gruyter.

Härdle, W. (1990). *Applied Nonparametric Regression*, Cambridge University Press.

Hafner, R. and Wallmeier, M. (2001). The Dynamics of DAX Implied Volatilities, *International Quarterly Journal of Finance* **1(1)**: 1-27.

Härdle, W. and Simar, L. (2003). *Applied Multivariate Statistical Analysis*, Springer-Verlag Berlin Heidelberg.

Kneip, A. and Utikal, K. (2001). Inference for Density Families Using Functional Principal Components Analysis, *Journal of the American Statistical Association* **96**: 519-531.

Ramsay, J. and Silverman, B. (1997). *Functional Data Analysis*, Springer, New York.

Rice, J. and Silverman, B. (1991). Estimating the Mean and Covariance Structure Nonparametrically when the Data are Curves, *Journal of Royal Statistical Society, Series B* **53**: 233-243.

Silverman, B. (1996). Smoothed Functional Principal Components Analysis by Choice of Norm, *Annals of Statistics* **24**: 1-24.

8 FFT-based Option Pricing

Szymon Borak, Kai Detlefsen, and Wolfgang Härdle

8.1 Introduction

The Black-Scholes formula, one of the major breakthroughs of modern finance, allows for an easy and fast computation of option prices. But some of its assumptions, like constant volatility or log-normal distribution of asset prices, do not find justification in the markets. More complex models, which take into account the empirical facts, often lead to more computations and this time burden can become a severe problem when computation of many option prices is required, e.g. in calibration of the implied volatility surface. To overcome this problem Carr and Madan (1999) developed a fast method to compute option prices for a whole range of strikes. This method and its application are the theme of this chapter.

In Section 8.2, we briefly discuss the Merton, Heston, and Bates models concentrating on aspects relevant for the option pricing method. In the following section, we present the method of Carr and Madan which is based on the fast Fourier transform (FFT) and can be applied to a variety of models. We also consider briefly some further developments and give a short introduction to the FFT algorithm. In the last section, we apply the method to the three analyzed models, check the results by Monte Carlo simulations and comment on some numerical issues.

8.2 Modern pricing models

The geometric Brownian motion (GBM) is the building block of modern finance. In particular, in the Black-Scholes model the underlying stock price is

assumed to follow the GBM dynamics:

$$dS_t = rS_t dt + \sigma S_t dW_t, \tag{8.1}$$

which, applying Itô's lemma, can be written as:

$$S_t = S_0 \exp\left\{ \left(r - \frac{\sigma^2}{2} \right) t + \sigma W_t \right\}. \tag{8.2}$$

The empirical facts, however, do not confirm model assumptions. Financial returns exhibit much fatter tails than the Black-Scholes model postulates, see Chapter 1. The common big returns that are larger than six-standard deviations should appear less than once in a million years if the Black-Scholes framework were accurate. Squared returns, as a measure of volatility, display positive autocorrelation over several days, which contradicts the constant volatility assumption. Non-constant volatility can be observed as well in the option markets where "smiles" and "skews" in implied volatility occur. These properties of financial time series lead to more refined models. We introduce three such models in the following paragraphs.

8.2.1   Merton model

If an important piece of information about the company becomes public it may cause a sudden change in the company's stock price. The information usually comes at a random time and the size of its impact on the stock price may be treated as a random variable. To cope with these observations Merton (1976) proposed a model that allows discontinuous trajectories of asset prices. The model extends (8.1) by adding jumps to the stock price dynamics:

$$\frac{dS_t}{S_t} = rdt + \sigma dW_t + dZ_t, \tag{8.3}$$

where Z_t is a compound Poisson process with a log-normal distribution of jump sizes. The jumps follow a (homogeneous) Poisson process N_t with intensity λ (see Chapter 14), which is independent of W_t. The log-jump sizes $Y_i \sim N(\mu, \delta^2)$ are i.i.d random variables with mean μ and variance δ^2, which are independent of both N_t and W_t.

The model becomes incomplete which means that there are many possible ways to choose a risk-neutral measure such that the discounted price process is a martingale. Merton proposed to change the drift of the Wiener process and to leave the other ingredients unchanged. The asset price dynamics is then given by:

$$S_t = S_0 \exp \left(\mu^M t + \sigma W_t + \sum_{i=1}^{N_t} Y_i \right),$$

where $\mu^M = r - \sigma^2 - \lambda \{ \exp(\mu + \frac{\delta^2}{2}) - 1 \}$. Jump components add mass to the tails of the returns distribution. Increasing δ adds mass to both tails, while a negative/positive μ implies relatively more mass in the left/right tail.

For the purpose of Section 8.4 it is necessary to introduce the characteristic function (cf) of $X_t = \ln \frac{S_t}{S_0}$:

$$\phi_{X_t}(z) = \exp \left[t \left\{ -\frac{\sigma^2 z^2}{2} + i \mu^M z + \lambda \left(e^{-\delta^2 z^2 / 2 + i \mu z - 1} \right) \right\} \right], \qquad (8.4)$$

where $X_t = \mu^M t + \sigma W_t + \sum_{i=1}^{N_t} Y_i$.

8.2.2   Heston model

Another possible modification of (8.1) is to substitute the constant volatility parameter σ with a stochastic process. This leads to the so-called "stochastic volatility" models, where the price dynamics is driven by:

$$\frac{dS_t}{S_t} = r dt + \sqrt{v_t} dW_t,$$

where v_t is another unobservable stochastic process. There are many possible ways of choosing the variance process v_t. Hull and White (1987) proposed to use geometric Brownian motion:

$$\frac{dv_t}{v_t} = c_1 dt + c_2 dW_t. \qquad (8.5)$$

However, geometric Brownian motion tends to increase exponentially which is an undesirable property for volatility. Volatility exhibits rather a mean

reverting behavior. Therefore a model based on an Ornstein-Uhlenbeck-type process:

$$dv_t = \kappa(\theta - v_t)dt + \beta dW_t, \tag{8.6}$$

was suggested by Stein and Stein (1991). This process, however, admits negative values of the variance v_t.

These deficiencies were eliminated in a stochastic volatility model introduced by Heston (1993):

$$\begin{aligned}
\frac{dS_t}{S_t} &= rdt + \sqrt{v_t}dW_t^{(1)}, \\
dv_t &= \kappa(\theta - v_t)dt + \sigma\sqrt{v_t}dW_t^{(2)},
\end{aligned} \tag{8.7}$$

where the two Brownian components $W_t^{(1)}$ and $W_t^{(2)}$ are correlated with rate ρ:

$$\mathrm{Cov}\left(dW_t^{(1)}, dW_t^{(2)}\right) = \rho dt, \tag{8.8}$$

for details see Chapter 7. The term $\sqrt{v_t}$ in equation (8.7) simply ensures positive volatility. When the process touches the zero bound the stochastic part becomes zero and the non-stochastic part will push it up.

Parameter κ measures the speed of mean reversion, θ is the average level of volatility, and σ is the volatility of volatility. In (8.8) the correlation ρ is typically negative, which is consistent with empirical observations (Cont, 2001). This negative dependence between returns and volatility is known in the market as the "leverage effect."

The risk neutral dynamics is given in a similar way as in the Black-Scholes model. For the logarithm of the asset price process $X_t = \ln\frac{S_t}{S_0}$ one obtains the equation:

$$dX_t = \left(r - \frac{1}{2}v_t\right)dt + \sqrt{v_t}dW_t^{(1)}.$$

The cf is given by:

$$
\begin{aligned}
\phi_{X_t}(z) &= \frac{\exp\{\frac{\kappa\theta t(\kappa - i\rho\sigma z)}{\sigma^2} + iztr + izx_0\}}{(\cosh\frac{\gamma t}{2} + \frac{\kappa - i\rho\sigma z}{\gamma}\sinh\frac{\gamma t}{2})^{\frac{2\kappa\theta}{\sigma^2}}} \\
&\quad \cdot \ \exp\left\{-\frac{(z^2 + iz)v_0}{\gamma\coth\frac{\gamma t}{2} + \kappa - i\rho\sigma z}\right\},
\end{aligned}
\tag{8.9}
$$

where $\gamma = \sqrt{\sigma^2(z^2 + iz) + (\kappa - i\rho\sigma z)^2}$, and x_0 and v_0 are the initial values for the log-price process and the volatility process, respectively.

8.2.3   Bates model

The Merton and Heston approaches were combined by Bates (1996), who proposed a model with stochastic volatility and jumps:

$$
\begin{aligned}
\frac{dS_t}{S_t} &= rdt + \sqrt{v_t}dW_t^{(1)} + dZ_t, \tag{8.10} \\
dv_t &= \kappa(\theta - v_t)dt + \sigma\sqrt{v_t}dW_t^{(2)}, \\
\mathrm{Cov}(dW_t^{(1)}, dW_t^{(2)}) &= \rho\,dt.
\end{aligned}
$$

As in (8.3) Z_t is a compound Poisson process with intensity λ and log-normal distribution of jump sizes independent of $W_t^{(1)}$ (and $W_t^{(2)}$). If J denotes the jump size then $\ln(1 + J) \sim N(\ln(1 + \bar{k}) - \frac{1}{2}\delta^2, \delta^2)$ for some $\bar{k}$. Under the risk neutral probability one obtains the equation for the logarithm of the asset price:

$$
dX_t = (r - \lambda\bar{k} - \frac{1}{2}v_t)dt + \sqrt{v_t}dW_t^{(1)} + \tilde{Z}_t,
$$

where $\tilde{Z}_t$ is a compound Poisson process with normal distribution of jump magnitudes.

Since the jumps are independent of the diffusion part in (8.10), the characteristic function for the log-price process can be obtained as:

$$
\phi_{X_t}(z) = \phi_{X_t}^D(z)\phi_{X_t}^J(z),
$$

where:

$$
\phi_{X_t}^D(z) \;=\; \frac{\exp\left\{\frac{\kappa\theta t(\kappa - i\rho\sigma z)}{\sigma^2} + izt(r - \lambda\overline{k}) + izx_0\right\}}{\left(\cosh\frac{\gamma t}{2} + \frac{\kappa - i\rho\sigma z}{\gamma}\sinh\frac{\gamma t}{2}\right)^{\frac{2\kappa\theta}{\sigma^2}}}
$$

$$
\cdot\;\; \exp\left\{-\frac{(z^2 + iz)v_0}{\gamma\coth\frac{\gamma t}{2} + \kappa - i\rho\sigma z}\right\} \tag{8.11}
$$

is the diffusion part cf and

$$
\phi_{X_t}^J(z) = \exp\{t\lambda(e^{-\delta^2 z^2/2 + i(\ln(1+\overline{k}) - \frac{1}{2}\delta^2)z} - 1)\}, \tag{8.12}
$$

is the jump part cf. Note that (8.9) and (8.11) are very similar. The difference lies in the shift $\lambda\overline{k}$ (risk neutral correction). Formula (8.12) has a similar structure as the jump part in (8.4), however, μ is substituted with $\ln(1+\overline{k}) - \frac{1}{2}\delta^2$.

8.3   Option Pricing with FFT

In the last section, three asset price models and their characteristic functions were presented. In this section, we describe a numerical approach for pricing options which utilizes the characteristic function of the underlying instrument's price process. The approach has been introduced by Carr and Madan (1999) and is based on the FFT. The use of the FFT is motivated by two reasons. On the one hand, the algorithm offers a speed advantage. This effect is even boosted by the possibility of the pricing algorithm to calculate prices for a whole range of strikes. On the other hand, the cf of the log price is known and has a simple form for many models considered in literature, while the density is often not known in closed form.

The approach assumes that the cf of the log-price is given analytically. The basic idea of the method is to develop an analytic expression for the Fourier transform of the option price and to get the price by Fourier inversion. As the Fourier transform and its inversion work for square-integrable functions (see Plancherel's theorem, e.g. in Rudin, 1991) we do not consider directly the option price but a modification of it.

Let $C_T(k)$ denote the price of a European call option with maturity T and strike $K = \exp(k)$:

$$C_T(k) = \int_k^\infty e^{-rT}(e^s - e^k)q_T(s)ds,$$

where q_T is the risk-neutral density of $s_T = \log S_T$. The function C_T is not square-integrable because $C_T(k)$ converges to S_0 for $k \to -\infty$. Hence, we consider a modified function:

$$c_T(k) = \exp(\alpha k)C_T(k), \tag{8.13}$$

which is square-integrable for a suitable $\alpha > 0$. The choice of α may depend on the model for S_t. The Fourier transform of c_T is defined by:

$$\psi_T(v) = \int_{-\infty}^\infty e^{ivk}c_T(k)dk.$$

The expression for ψ_T can be computed directly after an interchange of integrals:

$$
\begin{aligned}
\psi_T(v) &= \int_{-\infty}^\infty e^{ivk}\int_k^\infty e^{\alpha k}e^{-rT}(e^s - e^k)q_T(s)ds\,dk \\
&= \int_{-\infty}^\infty e^{-rT}q_T(s)\int_{-\infty}^s (e^{\alpha k+s} - e^{(\alpha+1)k})e^{ivk}dk\,ds \\
&= \int_{-\infty}^\infty e^{-rT}q_T(s)\left(\frac{e^{(\alpha+1+iv)s}}{\alpha+iv} - \frac{e^{(\alpha+1+iv)s}}{\alpha+1+iv}\right)ds \\
&= \frac{e^{-rT}\phi_T(v-(\alpha+1)i)}{\alpha^2+\alpha-v^2+i(2\alpha+1)v},
\end{aligned}
$$

where ϕ_T is the Fourier transform of q_T. A sufficient condition for c_T to be square-integrable is given by $\psi_T(0)$ being finite. This is equivalent to

$$E(S_T^{\alpha+1}) < \infty.$$

A value $\alpha = 0.75$ fulfills this condition for the models of Section 8.2. With this choice, we follow Schoutens et al. (2003) who found in an empirical study that this value leads to stable algorithms, i.e. the prices are well replicated for many model parameters.

Now, we get the desired option price in terms of ψ_T using Fourier inversion

$$C_T(k) = \frac{\exp(-\alpha k)}{\pi}\int_0^\infty e^{-ivk}\psi(v)dv.$$

This integral can be computed numerically as:

$$C_T(k) \approx \frac{\exp(-\alpha k)}{\pi} \sum_{j=0}^{N-1} e^{-iv_j k} \psi(v_j) \eta, \tag{8.14}$$

where $v_j = \eta j$, $j = 0, \ldots, N-1$, and $\eta > 0$ is the distance between the points of the integration grid.

Lee (2004) has developed bounds for the sampling and truncation errors of this approximation. Formula (8.14) suggests to calculate the prices using the FFT, which is an efficient algorithm for computing the sums

$$w_u = \sum_{j=0}^{N-1} e^{-i\frac{2\pi}{N} j u} x_j, \text{ for } u = 0, \ldots, N-1. \tag{8.15}$$

To see why this is the case see Example 1 below, which illustrates the basic idea of the FFT. In general, the strikes near the spot price are of interest because such options are traded most frequently. We consider thus an equidistant spacing of the log-strikes around the log spot price s_0:

$$k_u = -\frac{1}{2} N\zeta + \zeta u + s_0, \text{ for } u = 0, \ldots, N-1, \tag{8.16}$$

where $\zeta > 0$ denotes the distance between the log strikes. Substituting these log-strikes yields for $u = 0, \ldots, N-1$:

$$C_T(k_u) \approx \frac{\exp(-\alpha k)}{\pi} \sum_{j=0}^{N-1} e^{-i\zeta \eta j u} e^{i\{(\frac{1}{2} N\zeta - s_0) v_j\}} \psi(v_j) \eta.$$

Now, the FFT can be applied to

$$x_j = e^{i\{(\frac{1}{2} N\zeta - s_0) v_j\}} \psi(v_j), \text{ for } j = 0, \ldots, N-1,$$

provided that

$$\zeta \eta = \frac{2\pi}{N}. \tag{8.17}$$

This constraint leads, however, to the following trade-off: the parameter N controls the computation time and thus is often determined by the computational setup. Hence the right hand side may be regarded as given or fixed.

One would like to choose a small ζ in order to get many prices for strikes near the spot price. But the constraint implies then a big η giving a coarse grid for integration. So we face a trade-off between accuracy and the number of interesting strikes.

Example 1

The FFT is an algorithm for computing (8.15). Its popularity stems from its remarkable speed: while a naive computation needs N^2 operations the FFT requires only $N \log(N)$ steps. The algorithm was first published by Cooley and Tukey (1965) and since then has been continuously refined. We illustrate the original FFT algorithm for $N = 4$. Writing u and j as binary numbers:

$$u = 2u_1 + u_0, \; j = 2j_1 + j_0,$$

with $u_1, u_0, j_1, j_0 \in \{0, 1\}$ $u = (u_1, u_0)$, $j = (j_1, j_0)$ the formula (8.15) is given as:

$$w_{(u_1, u_0)} = \sum_{j_0=0}^{1} \sum_{j_1=0}^{1} x_{(j_1, j_0)} W^{(2u_1 + u_0)(2j_1 + j_0)},$$

where $W = e^{-2\pi i/N}$. Because of

$$W^{(2u_1 + u_0)(2j_1 + j_0)} = W^{2u_0 j_1} W^{(2u_1 + u_0) j_0},$$

we get

$$w_{(u_1, u_0)} = \sum_{j_0=0}^{1} \left(\sum_{j_1=0}^{1} x_{(j_1, j_0)}, W^{2u_0 j_1} \right) W^{(2u_1 + u_0) j_0}.$$

Now, the FFT can be described by the following three steps

$$w^1_{(u_0, j_0)} = \sum_{j_1=0}^{1} x_{(j_1, j_0)} W^{2u_0 j_1},$$

$$w^2_{(u_0, u_1)} = \sum_{j_0=0}^{1} w^1_{(u_0, j_0)} W^{(2u_1 + u_0) j_0},$$

$$w_{(u_1, u_0)} = w^2_{(u_0, u_1)}.$$

While a naive computation of (8.15) requires $4^2 = 16$ complex multiplications the FFT needs only $4 \log(4) = 8$ complex multiplications. This explains the speed of the FFT because complex multiplications are the most time consuming operations in this context.
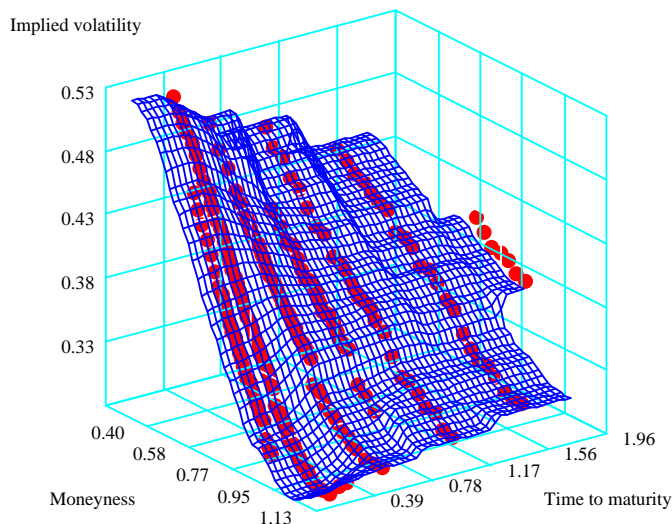
Figure 8.1: Implied volatility surface of DAX options on January 4, 1999.

STFfft01.xpl

8.4   Applications

In this section, we apply the FFT option pricing algorithm of Section 8.3 to the models described in Section 8.2. Our aim is to demonstrate the remarkable speed of the FFT algorithm by comparing it to Monte Carlo simulations. Moreover, we present an application of the fast option pricing algorithm to the calibration of implied volatility (IV) surfaces. In Figure 8.1 we present the IV surface of DAX options on January 4, 1999 where the red points are the observed implied volatilities and the surface is fitted with the Nadaraya-Watson kernel estimator. For analysis of IV surfaces consult Fengler et al. (2002) and Chapter 5.

In order to apply the FFT-based algorithm we need to know the characteristic function of the risk neutral density which has been described in Section 8.2 for the Merton, Heston, and Bates models. Moreover, we have to decide on

the parameters α, N, and η of the algorithm. Schoutens et al. (2003) used $\alpha = 0.75$ in a calibration procedure for the Eurostoxx 50 index data. We follow their approach and set α to this value. The computation time depends on the parameter N which we set to 512. As the number of grid points of the numerical integration is also given by N, this parameter in addition determines the accuracy of the prices. For parameter η, which determines the distance of the points of the integration grid, we use 0.25. A limited simulation study showed that the FFT algorithm is not sensitive to the choice of η, i.e. small changes in η gave similar results. In Section 8.3, we have already discussed the relation between these parameters.

For comparison, we computed the option prices also by Monte Carlo simulations with 500 time steps and 5000 repetitions. Such simulations are a convenient way to check the results of the FFT-based algorithm. The calculations are based on the following parameters: the price of the underlying asset is $S_0 = 100$, time to maturity $T = 1$, and the interest rate $r = 0.02$. For demonstration we choose the Heston model with parameters: $\kappa = 10$, $\theta = 0.2$, $\sigma = 0.7$, $\rho = -0.5$, and $v_0 = 0.2$. To make our comparison more sound we also calculate prices with the analytic formula given in Chapter 7. In the left panel of Figure 8.2 we show the prices of European call options as a function of the strike price K. As the prices obtained with the analytical formula are close to the prices obtained with the FFT-based method and the Monte Carlo prices oscillate around them, this figure confirms that the pricing algorithm works correctly. The different values of the Monte Carlo prices are mainly due to the random nature of this technique. One needs to use even more time steps and repetitions to get better results. The minor differences between the analytical and FFT-based prices come form the fact that the latter method gives the exact values only on the grid (8.16) and between the grid points one has to use some interpolation method to approximate the price of the option. This problem can be more clearly observed in the right panel of Figure 8.2, where percentage differences between the analytical and FFT prices are presented. In order to preserve the great speed of the algorithm we simply use linear interpolation between the grid points. This approach, however, slightly overestimates the true prices since the call option price is a convex function of the strike. It can be clearly seen that near the grid points the prices obtained by both methods coincide, while between the grid points the FFT-based algorithm generates higher prices than the analytical solution.

Although these methods yield similar results they need different computation time. In Table 8.1 we compare the speed of C++ implementations of the Monte Carlo and FFT methods. We calculate Monte Carlo prices for 20 different
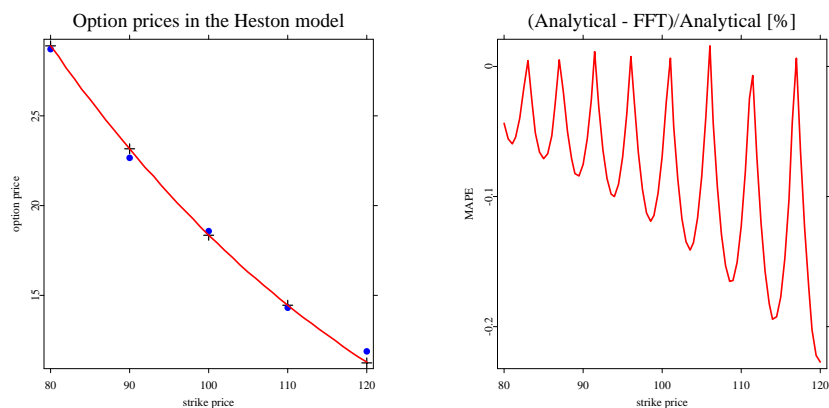
Figure 8.2: *Left panel:* European call option prices obtained by Monte Carlo simulations (filled circles), analytical formula (crosses) and the FFT method (solid line) for the Heston model. *Right panel:* Percentage differences between analytical and FFT prices.

STFfft02.xpl

Table 8.1: The computation times in seconds for the FFT method and the Monte Carlo method for three different models. Monte Carlo prices were calculated for 20 different strikes, with 500 time steps and 5000 repetitions.

Model	FFT	MC
Merton	0.01	31.25
Heston	0.01	34.41
Bates	0.01	37.53

strikes for each of the three models. The speed superiority of the FFT-based method is clearly visible. It is more than 3000 times faster than the Monte Carlo approach.

As an application of the fast pricing algorithm we consider the problem of model calibration. Given option prices observed in the market we look for model parameters that can reproduce the data well. Normally, the market prices are given by an implied volatility surface which represents the implied volatility of option prices for different strikes and maturities. The calibration can then be done for the implied volatilities or for the option prices. This decision depends on the problem considered. As a measure of fit one can use the Mean Squared Error (MSE):

$$MSE = \frac{1}{\text{number of options}} \sum_{\text{options}} \frac{(\text{market price - model price})^2}{\text{market price}^2}, \quad (8.18)$$

but other choices like the Mean Absolute Percentage Error (MAPE) or Mean Absolute Error (MAE) are also possible:

$$MAPE = \frac{1}{\text{number of options}} \sum_{\text{options}} \frac{|\text{ market price - model price }|}{\text{market price}},$$

$$MAE = \frac{1}{\text{number of options}} \sum_{\text{options}} |\text{ market price - model price }|.$$

Moreover, the error function can be modified by weights if some regions of the implied volatility surface are more important or some observations should be ignored completely.

The calibration results in a minimization problem of the error function MSE. This optimization can be carried out by different algorithms like simulated annealing, the Broyden-Fletcher-Goldfarb-Shanno-algorithm, the Nelder-Mead simplex algorithm or Monte Carlo Markov Chain methods. An overview of optimization methods can be found in Čížková (2003). As minimization algorithms normally have to compute the function to be minimized many times an efficient algorithm for the option prices is essential. The FFT-based algorithm is fairly efficient as is shown in Table 8.1. Moreover, it returns prices for a whole range of strikes at one maturity. This is an additional advantage because for the calibration of an implied volatility surface one needs to calculate prices for many different strikes and maturities.

As an example we present the results for the Bates model calibrated to the IV surface of DAX options on January 4, 1999. The data set, which can be found in MD*Base, contains 236 option prices for 7 maturities (for each maturity there is a different number of strikes). We minimize (8.18) with respect to 8 parameters of the Bates model: $\lambda, \delta, \overline{k}, \kappa, \theta, \sigma, \rho, v_0$. Since the function (8.18)
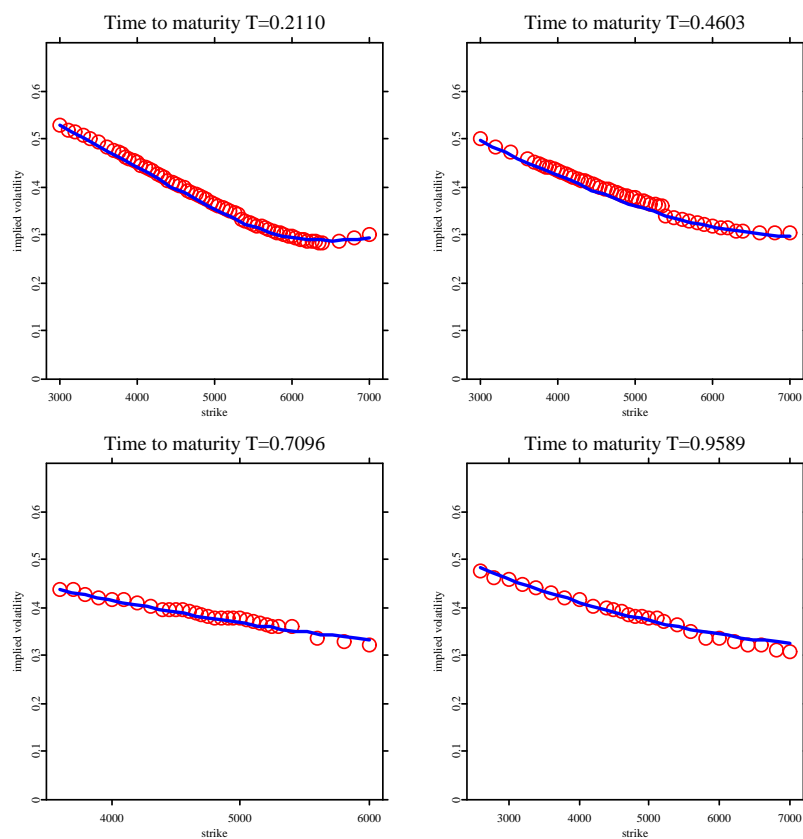
Figure 8.3:   The observed implied volatilities of DAX options on January 4, 1999 (circles) and fitted Bates model (line) for 4 different maturity strings.

STFfft03.xpl

has many local minima, we use the simulated annealing minimization method, which offers the advantage to search for a global minimum, combined with the Nelder-Mead simplex algorithm. As a result we obtaine the following estimates for the model parameters: $\widehat{\lambda} = 0.13$, $\widehat{\delta} = 0.0004$, $\widehat{\overline{k}} = -0.03$, $\widehat{\kappa} = 4.23$, $\widehat{\theta} =$

0.17, $\widehat{\sigma} = 1.39$, $\widehat{\rho} = -0.55$, $\widehat{v_0} = 0.10$, and the value of MSE is 0.00381. In Figure 8.3 we show the resulting fits of the Bates model to the data for 4 different maturities. The red circles are implied volatilities observed in the market on the time to maturities $T = 0.21, 0.46, 0.71, 0.96$ and the blue lines are implied volatilities calculated from the Bates model with the calibrated parameters. In the calibration we used all data points. As the FFT-based algorithm computes prices for the whole range of strikes the biggest impact on the speed of calibration has the number of used maturities, while the total number of observations has only minor influence on the speed.

On the one hand, the Carr-Madan algorithm offers a great speed advantage but on the other hand its applications are restricted to European options. The Monte Carlo approach instead works for a wider class of derivatives including path dependent options.

Thus, this approach has been modified in different ways. The accuracy can be improved by using better integration rules. Carr and Madan (1999) considered also the Simpson rule which leads – taking (8.17) into account – to the following formula for the option prices:

$$C_T(k_u) \approx \frac{\exp(-\alpha k)}{\pi} \sum_{j=0}^{N-1} e^{-i\zeta\eta ju} e^{i\{(\frac{1}{2}N\zeta - s_0)v_j\}} \psi(v_j)\frac{\eta}{3}\{3 + (-1)^j - I(-j = 0)\}.$$

This representation again allows a direct application of the FFT to compute the sum.

An alternative to the original Carr-Madan approach is to consider instead of (8.13) other modifications of the call prices. For example, Cont and Tankov (2004) used the (modified) time value of the options:

$$\tilde{c}_T(k) = C_T(k) - \max(1 - e^{k-rT}, 0).$$

Although this method also requires the existence of α satisfying $E(S_T^{\alpha+1}) < \infty$ the parameter does not enter into the final pricing formula. Thus, it is not necessary to choose any value for α. This freedom of choice of α makes the approach easier to implement. On the other hand, option price surfaces that are obtained with this method often have a peak for small maturities and strikes near the spot. This special form differs from the surfaces typically observed in the market. The peak results from the non-differentiability of the intrinsic value at the spot. Hence, other modifications of the option prices have been considered that make the modified option prices differentiable (Cont and Tankov, 2004).

The calculation of option prices by the FFT-based algorithm leads to different errors. The truncation error results from substituting the infinite upper integration limit by a finite number. The sampling error comes from evaluating the integrand only at grid points. Lee (2004) gives bounds for these errors and discusses error minimization strategies. Moreover, he presents and unifies extensions of the original Carr-Madan approach to other payoff classes. Besides the truncation and the sampling error, the implementation of the algorithm often leads to severe roundoff errors because of the complex form of the characteristic function for some models. To avoid this problem, which often occurs for long maturities, it is necessary to transform the characteristic function.

Concluding, we can say that the FFT-based option pricing method is a technique that can be used whenever time constraints are important. However, in order to avoid severe pricing errors its application requires careful decisions regarding the choice of the parameters and the particular algorithm steps used.

Bibliography

Bates, D. (1996). Jump and Stochastic Volatility: Exchange Rate Processes Implicit in Deutsche Mark Options, *Review of Financial Studies* **9**: 69–107.

Carr, P. and Madan, D. (1999). Option valuation using the fast Fourier transform, *Journal of Computational Finance* **2**: 61–73.

Čížková, L. (2003). Numerical Optimization Methods in Econometrics, *in* J.M. Rodriguez Poo (ed.) *Computer-Aided Introduction to Econometrics*, Springer-Verlag, Berlin.

Cooley, J. and Tukey, J. (1965). An algorithm for the machine calculation of complex Fourier series, *Math. Comput.* **19**: 297–301.

Cont, R. (2001). Empirical properties of assets returns: Stylized facts and statistical issues, *Quant. Finance* **1**: 1-14.

Cont, R. and Tankov, P. (2004). *Financial Modelling With Jump Processes*, Chapman & Hall/CRC.

Fengler, M., Härdle, W. and Schmidt, P. (2002). The Analysis of Implied Volatilities, *in* W. Härdle, T. Kleinow, G. Stahl (eds.) *Applied Quantitative Finance*, Springer-Verlag, Berlin.

Heston, S. (1993). A closed-form solution for options with stochastic volatility with applications to bond and currency options, *Review of Financial Studies* **6**: 327-343.

Hull, J. and White, A. (1987). The pricing of Options on Assets with Stochastic Volatilities, *Journal of Finance* **42**: 281–300.

Lee, R. (2004). Option pricing by transform methods: extensions, unification and error control, *Journal of Computational Finance* **7**.

Merton, R. (1976). Option pricing when underlying stock returns are discontinuous, *J. Financial Economics* **3**: 125-144.

Rudin, W. (1991). *Functional Analysis*, McGrawHill.

Schoutens, W., Simons, E., and Tistaert, J. (2003). A Perfect Calibration! Now What? *UCS Technical Report*, Catholic University Leuven.

Stein, E. and Stein, J. (1991). Stock price distribution with stochastic volatility: An analytic approach, *Review of Financial Studies* **4**: 727–752.

DSFM fitting of Implied Volatility Surfaces

Szymon Borak, Matthias Fengler and Wolfgang Härdle
CASE - Center for Applied Statistics and Economics
Humboldt-Universität zu Berlin
Spandauer Straße 1, 10178 Berlin, Germany
borak@wiwi.hu-berlin.de

Abstract

Implied volatility is one of the key issues in modern quantitative finance, since plain vanilla option prices contain vital information for pricing and hedging of exotic and illiquid options. European plain vanilla options are nowadays widely traded, which results in a great amount of high-dimensional data especially on an intra day level. The data reveal a degenerated string structure. Dynamic Semiparametric Factor Models (DSFM) are tailored to handle complex, degenerated data and yield low dimensional representations of the implied volatility surface (IVS). We discuss estimation issues of the model and apply it to DAX option prices.

1. Introduction

The Black-Scholes formula (BS) for calculating the price of a European plain vanilla option is one of the most recognized results in modern quantitative finance. The price is given as a function of the price of the underlying, a strike price, the interest rate, time to maturity and an unobserved volatility. By plugging the observed option price into the BS formula it is straightforward to calculate the implied volatility (IV). The surface (on day t) given by the mapping from moneyness κ (a measure of strikes) and from time to maturity τ $(\kappa, \tau) \rightarrow \hat{\sigma}_t(\kappa, \tau)$ is called implied volatility surface (IVS). The observed IVS, Figure 1, reveals a non-flat profile across moneyness (called "smile" or "smirk") and across time to maturity.

Despite the deficiencies of the BS model it is popular among practitioners to quote option prices in terms of IV due to its intuitive simplicity. However there were many efforts to model a non constant IVS. One possible approach is to change the dynamics of the process of the underlying asset by increasing its degree of freedom. This leads to parametric models with jumps like in [14], stochastic volatility like in [13] and [12] or models based on Lévy processes

like the generalized hyperbolic in [4] among many others. The models reproduce the smile phenomenon and their parameters are calibrated from the option prices by minimizing cost functional. One may also consider local volatility (LV) models like in [3] where the volatility is assumed to be a function of time and the price of the underlying. There exist an analytical formula which allows to calibrate this surface (LVS) straightforward from the IVS.

A drawback even of the most sophisticated models is the failure to correctly describe the dynamics of the IVS. This can be inferred from frequent recalibration of the model and has been best understood in the context of LV-models [10]. Consequently, studying the IVS as an additional market factor has become a vital stand of research. The main focus is on a low dimensional approximation of the IVS based on principal components analysis (PCA). The PCA is applied both to the term structure of the IVS ([16] or [7]) and strike dimension (eg. [15]). The common PCA for several maturity groups is studied in [8] and the functional PCA was discussed in [1] and [2].

Our approach is to represent the IVS as the sum of factors treated as two dimensional functions depending on moneyness and maturity. In [2] the factors are obtained using the functional PCA for the IVS fitted on a grid for each particular day. However this fit may be biased due to the degenerated data design. In [9] the IVS is obtained as a projection on parametric factors which has to be initially specified. In DSFM the factors are estimated from the data, which allows flexible modelling. Contrary to [2] the IVS is obtained as a fit to the factors smoothed in time, which reflects the dynamics of the whole system.

The paper is organized as follows: in the next section we describe the DSFM and present the estimation procedure. In Section 3 we discuss the estimation issues and proposed improvements of the algorithm. Section 4 presents the empirical results on DAX options and discusses briefly possible applications.
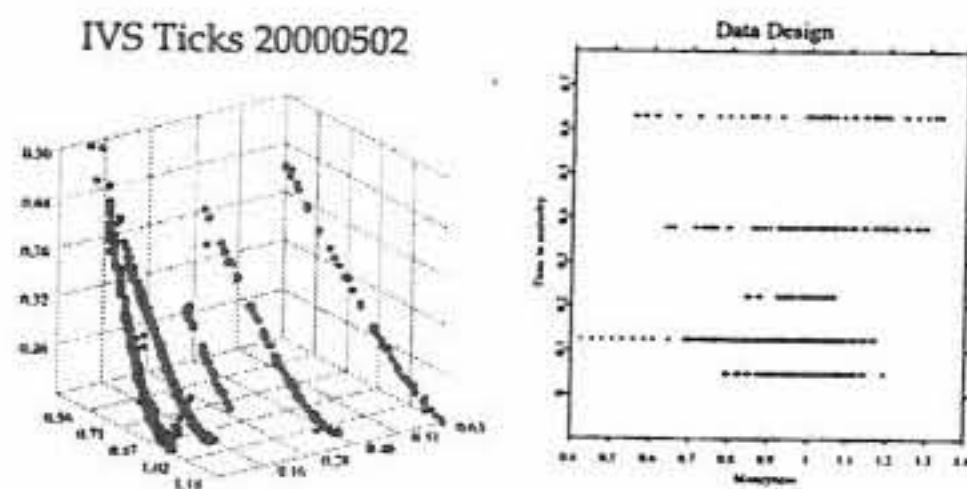
2. DSFM

Figure 1. Left panel: implied volatilities observed on 2nd May, 2000. Right panel: data design on 2nd May, 2000.

Institutional conventions of the option market entail a specific degenerated IV data design. Each day one observes only a small number of maturities which form typical 'strings'. The usual data pattern is visible in the right panel of Figure 1. Options belonging to the same string have a common time to maturity but different moneyness. In the left panel of the Figure 1 IV smiles are presented. One can easily see different curvatures for a each time to maturity. As time passes not only do the strings move through the space towards expiry but they also change their shape and level randomly.

In order to capture this complex dynamic structure of the IVS a DSFM was proposed in [6]. It offers a low-dimensional representation of the IVS, which is approximated by basis functions in a finite dimensional function space. The basis functions are unknown and have to be estimated from the data. The IVS dynamics are explained by loading coefficients, which form a multidimensional time series.

Let $Y_{i,j}$ be the log-implied volatility observed on a particular day. The index i is the number of the day, while the total number of days is denoted by I ($i = 1, ..., I$). The index j represents an intra-day trade on day i and the number of trades on that day is J_i ($j = 1, ...J_i$). Let $X_{i,j}$ be a two-dimensional variable containing moneyness $\kappa_{i,j}$ and maturity $\tau_{i,j}$. Among many moneyness settings we define it as $\kappa_{i,j} = \frac{K_{i,j}}{F_{t_i}}$, where $K_{i,j}$ is a strike and F_{t_i} the underlying futures price at time t_i. The DSFM regresses $Y_{i,j}$ on $X_{i,j}$ by:

$$Y_{i,j} = m_0(X_{i,j}) + \sum_{l=1}^{L} \beta_{i,l} m_l(X_{i,j}), \qquad (1)$$

where m_0 is an invariant basis function, m_l ($l = 1, ...L$) are the 'dynamic' basis functions and $\beta_{i,l}$ are the factor weights depending on time i.

2.1. Estimation

The estimates $\widehat{\beta}_{i,l}$ and $\widehat{m}_l$ are obtained by minimizing the following least squares criterion ($\beta_{i,0} = 1$):

$$\sum_{i=1}^{I} \sum_{j=1}^{J_i} \int \left\{ Y_{i,j} - \sum_{l=0}^{L} \widehat{\beta}_{i,l} \widehat{m}_l(u) \right\}^2 K_h(u - X_{i,j}) \, du, \qquad (2)$$

where K_h denotes a two-dimension kernel function. The possible choice for two-dimensional kernels is a product of one dimensional kernels $K_h(u) = k_{h_1}(u_1) \times k_{h_2}(u_2)$, where $h = (h_1, h_2)^\top$ are bandwidths and $k_h(v) = k(h^{-1}v)/h$ is a one dimensional kernel function.

The minimization procedure searches through all functions $\widehat{m}_l : \mathbb{R}^2 \longrightarrow \mathbb{R}$ ($l = 0, ..., L$) and time series $\widehat{\beta}_{i,l} \in \mathbb{R}$ ($i = 1, ..., I; l = 1, ..., L$).

To calculate the estimates an iterative procedure is applied. First we introduce the following notation for $1 \leq i \leq I$:

$$\widehat{p}_i(u) = \frac{1}{J_i} \sum_{j=1}^{J_i} K_h(u - X_{i,j}), \qquad (3)$$

$$\widehat{q}_i(u) = \frac{1}{J_i} \sum_{j=1}^{J_i} K_h(u - X_{i,j}) Y_{i,j}. \qquad (4)$$

We denote by $\widehat{m}^{(r)} = (\widehat{m}_0^{(r)}, ..., \widehat{m}_L^{(r)})^\top$ the estimate of the basis functions and $\widehat{\beta}_i^{(r)} = (\widehat{\beta}_{i,1}^{(r)}, ..., \widehat{\beta}_{i,L}^{(r)})^\top$ the factor loadings on the day i after r iterations. By replacing each function $\widehat{m}_l$ in (2) by $\widehat{m}_l + \delta g$ with arbitrary function g and taking derivatives with respect to δ one obtains:

$$\sum_{i=1}^{I} \sum_{j=1}^{J_i} \left\{ Y_{i,j} - \sum_{l=0}^{L} \widehat{\beta}_{i,l} \widehat{m}_l(X_{i,j}) \right\} \widehat{\beta}_{i,l'} K_h(u - X_{i,j}) = 0. \qquad (5)$$

Rearranging terms in (5) and plugging in (3)-(4) yields:

$$\sum_{i=1}^{I} J_i \widehat{\beta}_{i,l'} \widehat{q}_i(u) = \sum_{i=1}^{I} J_i \sum_{l=0}^{L} \widehat{\beta}_{i,l'} \widehat{\beta}_{i,l} \widehat{p}_i(u) \widehat{m}_l(u), \qquad (6)$$

for $0 \leq l' \leq L$. In fact (6) is a set of $L+1$ equations. Define the matrix $B^{(r)}(u)$ and vector $Q^{(r)}(u)$ by their elements:

$$(B^{(r)}(u))_{l,l'} = \sum_{i=1}^{I} J_i \widehat{\beta}_{i,l'}^{(r-1)} \widehat{\beta}_{i,l}^{(r-1)} \widehat{p}_i(u), \qquad (7)$$

$$(Q^{(r)}(u))_l = \sum_{i=1}^{I} J_i \widehat{\beta}_{i,l}^{(r-1)} \widehat{q}_i(u). \quad (8)$$

Thus (6) is equivalent to:

$$B^{(r)}(u)\widehat{m}^{(r)}(u) = Q^{(r)}(u) \quad (9)$$

which yields the estimate of $\widehat{m}^{(r)}(u)$ in the r-th iteration.

A similar idea has to be applied to update $\widehat{\beta}_i^{(r)}$. Replacing $\widehat{\beta}_{i,l}$ by $\widehat{\beta}_{i,l} + \delta$ in (2) and taking once more the derivative with respect to δ yields:

$$\sum_{j=1}^{J_i} \int \left\{ Y_{i,j} - \sum_{l=0}^{L} \widehat{\beta}_{i,l} \widehat{m}_l(X_{i,j}) \right\} \widehat{m}_{l'}(u) K_h(u - X_{i,j}) du = 0, \quad (10)$$

which leads to:

$$\int \widehat{q}_i(u)\widehat{m}_{l'}(u)\,du = \sum_{l=0}^{L} \widehat{\beta}_{i,l} \int \widehat{p}_i(u)\widehat{m}_{l'}(u)\widehat{m}_l(u)\,du, \quad (11)$$

for $1 \le l' \le L$. The formula (11) is now a system of L equations. Define the matrix $M^{(r)}(i)$ and the vector $S^{(r)}(i)$ by their elements:

$$(M^{(r)}(i))_{l,l'} = \int \widehat{p}_i(u)\widehat{m}_{l'}(u)\widehat{m}_l(u)\,du, \quad (12)$$

$$(S^{(r)}(i))_l = \int \widehat{q}_i(u)\widehat{m}_l(u)\,du - \int \widehat{p}_i(u)\widehat{m}_0(u)\widehat{m}_l(u)\,du. \quad (13)$$

An estimate of $\widehat{\beta}_i^{(r)}$ is thus given by solving:

$$M^{(r)}(i)\widehat{\beta}_i^{(r)} = S^{(r)}(i). \quad (14)$$

The algorithm stops when only minor changes occur:

$$\sum_{i=1}^{I} \int \left(\sum_{l=0}^{L} \widehat{\beta}_{i,l}^{(r)} \widehat{m}_l^{(r)}(u) - \widehat{\beta}_{i,l}^{(r-1)} \widehat{m}_l^{(r-1)}(u) \right)^2 du \le \epsilon \quad (15)$$

for some small ϵ. Obviously one needs to set initial values of $\widehat{\beta}_i^{(0)}$ in order to start the algorithm.

2.2. Orthogonalization and normalization

The estimates $\widehat{m} = (\widehat{m}_1, ..., \widehat{m}_L)^\top$ of the basis functions are not uniquely defined. They can be replaced by functions that span the same affine space. Define $\widehat{p}(u) = \frac{1}{I} \sum_{i=1}^{I} \widehat{p}_i(u)$ and the $L \times L$ matrix Γ by its elements

$$\Gamma_{l,l'} = \int \widehat{m}_l(u)\widehat{m}_{l'}(u)\widehat{p}(u)du.$$

The estimates $\widehat{m}$ are replaced by new functions $\widehat{m}^{new} = (\widehat{m}_1^{new}, ..., \widehat{m}_L^{new})^\top$:

$$\widehat{m}_0^{new} = \widehat{m}_0 - \gamma^\top \Gamma^{-1} \widehat{m}$$
$$\widehat{m}^{new} = \Gamma^{-1/2} \widehat{m}$$

such that they are now orthogonal in the $L^2(\widehat{p})$ space. The loading time series estimates $\widehat{\beta}_i = (\widehat{\beta}_{i,1}, ..., \widehat{\beta}_{i,L})^\top$ need to be substituted by:

$$\widehat{\beta}_i^{new} = \Gamma^{-1/2}(\widehat{\beta}_i + \Gamma^{-1}\gamma) \quad (16)$$

where γ is ($L \times 1$) vector with $\gamma_l = \int \widehat{m}_0(u)\widehat{m}_l(u)\widehat{p}(u)du$.

The next step is to choose an orthogonal basis such that for each $w = 1, ..., L$ the explanation achieved by the partial sum:

$$m_0(u) + \sum_{l=1}^{w} \beta_{i,l} m_l(u)$$

is maximal. One proceeds as in PCA. First define a matrix B with $B_{l,l'} = \sum_{i=1}^{I} \widehat{\beta}_{i,l}\widehat{\beta}_{i,l'}$ and $Z = (z_1, ..., z_L)$ where $z_1, ..., z_L$ are the eigenvectors of B. Then replace $\widehat{m}$ by $\widehat{m}^{new} = Z^\top \widehat{m}$ and $\widehat{\beta}_i$ by $\widehat{\beta}_i^{new} = Z^\top \widehat{\beta}_i$.

The orthonormal basis $\widehat{m}_1, ..., \widehat{m}_L$ is chosen such that $\sum_{i=1}^{I} \widehat{\beta}_{i,1}^2$ is maximal and given $\widehat{\beta}_{i,1}, \widehat{m}_0, \widehat{m}_1$ the quantity $\sum_{i=1}^{I} \widehat{\beta}_{i,2}^2$ is maximal and so forth.

3. Estimation issues

The estimation procedure encounter several computational challenges. The basis factor functions can be represented on the finite grid only, which obviously may not cover the whole desired estimation space. One also needs to choose some kernel function, the bandwidths and the initial loading time series $\widehat{\beta}_i^{(0)}$. Due to the degenerated design of the IV data proper decision of these points is a key issue in successful model estimation.

3.1. Implementation

As a numerical result of the estimation L time series $(\widehat{\beta}_{i,l})$ and $L + 1$ functions ($\widehat{m}_l$) given on the finite grid are obtained. The choice of the grid needs to be arbitrary and depends on the density of the data points. The data ($X_{i,j}$ and $Y_{i,j}$) comes into the computation only in $\widehat{p}_i(u)$ and $\widehat{q}_i(u)$, which has to be calculated before the main iteration procedure. The calculation of $\widehat{p}_i(u)$ and $\widehat{q}_i(u)$, however, is the main computation effort in the estimation procedure. Therefore we believe that the DSFM can be used efficiently in a 'sliding window' type of analysis. Updating of $\widehat{p}_i(u)$ and $\widehat{q}_i(u)$ requires calculations only for one additional day, which is not a big computational issue.

3.2. Bandwidths dependence

In derivative market one can observe fairly many different types of option contracts. Each day one may trade options with several different time to maturities and many different strikes. However the number of possible strikes is much higher than the number of maturities, which results in the string structure. Moreover the contracts with smaller maturities are traded more intensively and there tend to exist more contracts for the smaller time to maturities for which the difference between two successive expiry days is one month (1M, 2M, 3M), but for the next maturity range it increases to three months (6M, 9M, 12M).

Since the strings are moving in the maturity vs. moneyness plane towards expiry one needs to pool many days in order to fill the plane with observations. However due to an unequal distribution of data points one needs even more days to fill the range with bigger maturities than with smaller ones. Otherwise one faces gaps for some particular maturity range.

These gaps may obstruct the estimation procedure. If in any point u' the function $\hat{p}(u') = 0$ in (3) then obviously matrix $B^{(r)}(u')$ in (7) contains only 0 and is singular. This means that one may not estimate successfully any value of the IVS in this point.

This problem may be solved by increasing the bandwidths but it may lead also to a larger bias. One may also use a kernel function with infinite support like the Gaussian kernel but instead of analytical zeros numerical zeros creep in. Another possibility is to use the k-nearest neighbor estimator. In the range with many data, however, one takes into consideration only very few observations closest to the grid points. On the other hand in the range with few points the estimator is based on the observations far from the grid points. In order to cope with the degenerated data design local bandwidths can be applied. In (3) and (4) the fixed bandwidths are replaced by bandwidths dependent on time to maturity and moneyness:

$$\hat{p}_i(u) = \frac{1}{J_i} \sum_{j=1}^{J_i} K_{h(u)}(u - X_{i,j}), \qquad (17)$$

$$\hat{q}_i(u) = \frac{1}{J_i} \sum_{j=1}^{J_i} K_{h(u)}(u - X_{i,j}) Y_{i,j}. \qquad (18)$$

Due to the described data design we propose to keep the bandwidths in the moneyness direction constant and linearly increasing in the maturity dimension. For the optimal choice of the bandwidths we refer to [11].

3.3. Initial parameter dependence

The problem of gaps in the data cannot only be handled with the size of the bandwidths. Of course it is obligatory that $\hat{p}_i(u)$ needs to be non-zero for at least one i. However this is not a sufficient condition to ensure non singularity of the matrix $B^{(r)}(u')$. The initial estimates of $\hat{\beta}_i^{(0)}$ play also an important role.

In [6] a piecewise constant initial time series was proposed. The subintervals $I_1, ..., I_L$ are pairwise disjoint subsets of $\{1, ..., I\}$ and $\bigcup_{l=1}^{L} I_l$ is a strict subset of $\{1, ..., I\}$. The initial estimates are now defined by $\hat{\beta}_{i,l}^{(0)} = 1$ if $i \in I_l$ and $\hat{\beta}_{i,l}^{(0)} = 0$ if $i \notin I_l$. To complete the setting $\hat{\beta}_{i,0}^{(0)} = 1$ for each i.

However this kind of setting requires even more data to obtain the final estimates. For each subset I_l there needs to exist at least one day i such that $\hat{p}_i(u') \neq 0$, otherwise the row of zeros in (7) appears. The smaller is the length of I_l intervals the bigger bandwidths need to be taken. This deficiency can be removed by taking a random initial time series.

4. Results

For our analysis we employ tick statistics on DAX index options from January 1999 to February 2003. By inverting the BS formula one easily obtains IV. We regard as outliers observations with IV bigger that 0.8 and smaller than 0.04. We also remove observations with maturity less than 10 day since their behavior in this range is irregular due to expiry effect.

We apply the algorithm on an equidistant grid covering moneyness $\kappa \in [0.8, 1.2]$ and time to maturity measured in years $\tau \in [0.05, 1.00]$. In each direction our grid consists of 25 points. We set the number of dynamic basis functions to $L = 3$ like in [6]. In the moneyness direction we apply constant bandwidths $h_1 = 0.03$ and in order to get smoother estimates of the basis functions in the maturity direction we use linearly increasing bandwidths. On the smallest maturity grid points we set bandwidths on 0.02 and increase them linearly to 0.2 for the greatest maturity points. As the starting values of $\hat{\beta}^{(0)}$ we take a piecewise constant series on disjoint time intervals. The initial weights selection is discussed below.

Figure 2 presents the estimated factors loading $\hat{\beta}_1, \hat{\beta}_2$ and $\hat{\beta}_3$ respectively. The magnitude and variance of the $\hat{\beta}_1$ are much higher than for the other two time series, which suggests that the first basis function has the biggest explanatory power of the IVS variation. This is actually not surprising since the basis functions were ordered with respect to the biggest variance of loading factors.
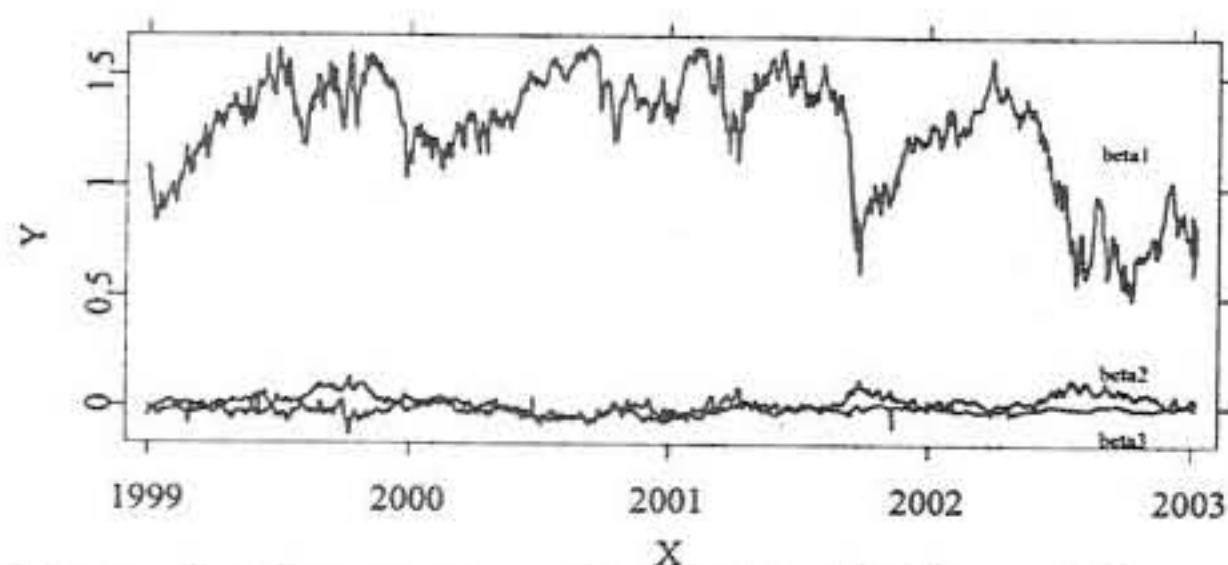
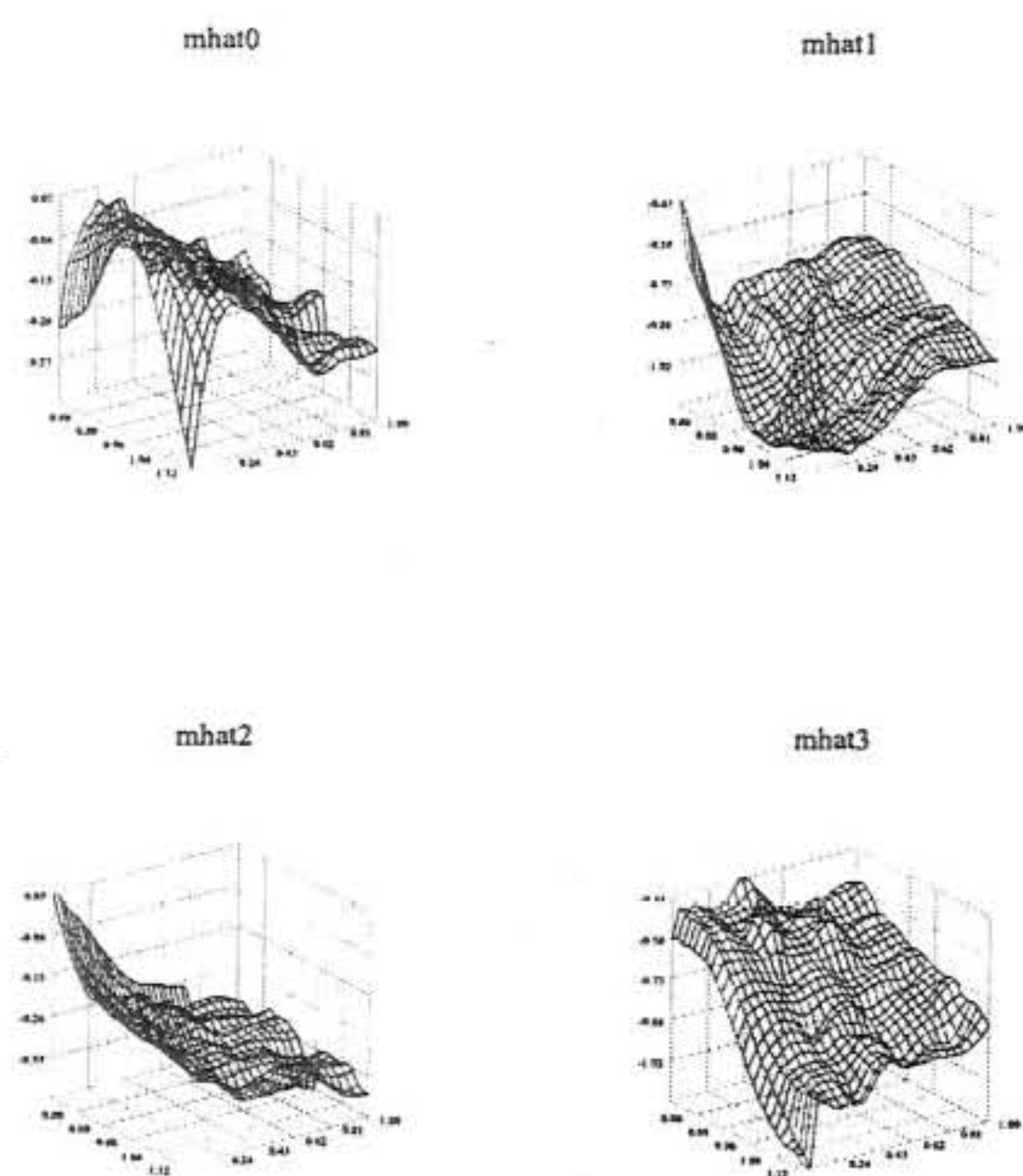**Figure 2. Time series of weights $\widehat{\beta}_1, \widehat{\beta}_2$ and $\widehat{\beta}_3$.**

Figure 3. Invariant basis function $\widehat{m}_0$ and dynamic basis functions $\widehat{m}_1, \widehat{m}_2$ and $\widehat{m}_3$.

Figure 3 displays the estimated basis functions $\widehat{m}_0$ - $\widehat{m}_3$. We find similar interpretations of the factors as in [6] or [2]. The first dynamic factor $\widehat{m}_1$ is relatively flat on almost the whole range and negative on all grid points. It reflects the up and down shifts of the entire log-IVS. For the small maturities a strong curvature can be seen. It corresponds to the empirical fact that near the expiry the 'smile' effect becomes stronger. The second function is positive for the small maturities and negative for the bigger maturities. The positive $\widehat{\beta}_2$ increases short term maturities IVs and simultaneously de-

creases the long term ones. The negative $\widehat{\beta}_2$ causes the opposite effect. This function provides term structure changes of the IVS. The last function $\widehat{m}_3$ reveals a strong slope in the moneyness direction changing from positive to negative near at-the-money. It reflects changes of the moneyness slope and smile curvature.
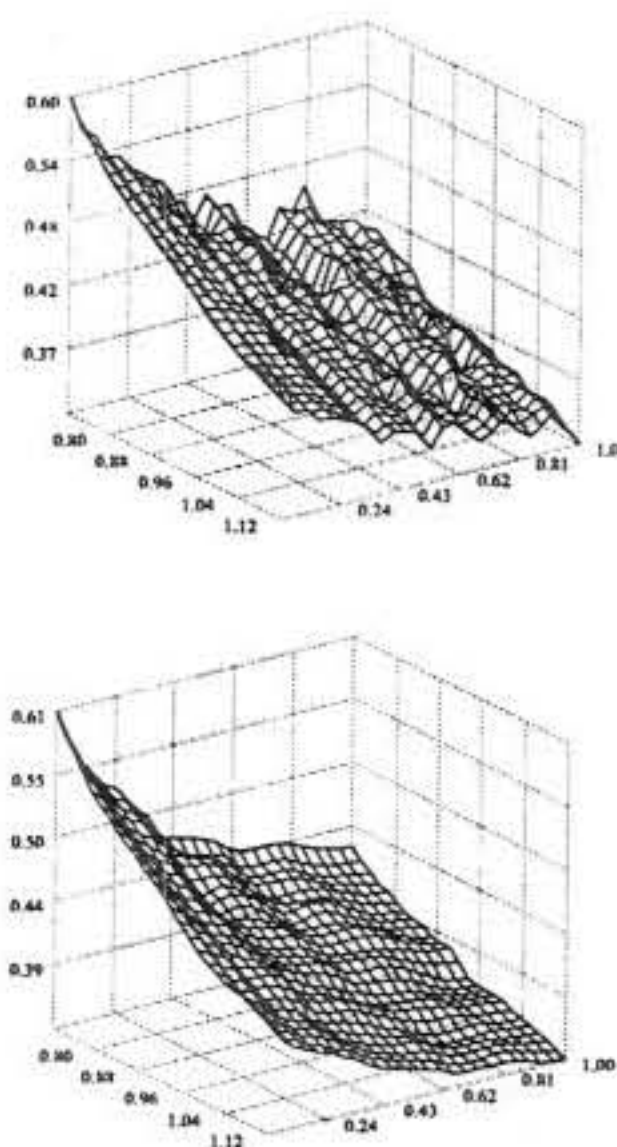

Figure 4. IVS estimates on February 25, 2003, fixed bandwidths $h_1 = 0.03, h_2 = 0.02$ (top), linearly increasing bandwidths (bottom).

In our estimation we used the local bandwidths linearly increasing in maturity. Figure 4 presents the comparison of the two different IVS estimates on February 25, 2003 obtained with fixed bandwidths $h_1 = 0.03, h_2 = 0.02$ and with local bandwidths. While in the fixed bandwidths approach in bigger maturities the estimated IVS is rough, in the local bandwidths approach it becomes smoother.

Another estimation issue is the choice of the initial times series $\widehat{\beta}^{(0)}$. We have recalculated the estimates for different starting values. Denote by PC_1, PC_2, PC_3 as the different settings of piecewise constant starting values as described in Section 3.3. Denote also by WN_1, WN_2, WN_3 settings where the algorithm starts from a white noise and BM_1, BM_2, BM_3 from a Brownian Motion. For each of the 9 settings we have obtained different estimates of weights. The correlation between different estimates of $\widehat{\beta}_1, \widehat{\beta}_2$ and $\widehat{\beta}_3$ are respectively:

$$\begin{pmatrix} 1.0 & -1.0 & -1.0 & -1.0 & -1.0 & 1.0 & -0.9 & 0.9 & -0.9 \\ & 1.0 & 1.0 & 1.0 & 1.0 & -1.0 & 0.9 & -0.9 & 0.9 \\ & & 1.0 & 1.0 & 1.0 & -1.0 & 0.9 & -0.9 & 0.9 \\ & & & 1.0 & 1.0 & -1.0 & 0.9 & -0.9 & 0.9 \\ & & & & 1.0 & -1.0 & 0.9 & -0.9 & 0.9 \\ & & & & & 1.0 & -0.9 & 0.9 & 0.9 \\ & & & & & & 1.0 & -1.0 & 1.0 \\ & & & & & & & 1.0 & -1.0 \\ & & & & & & & & 1.0 \end{pmatrix}$$

$$\begin{pmatrix} 1.0 & 1.0 & 1.0 & -1.0 & -1.0 & -1.0 & 0.3 & -0.3 & -0.3 \\ & 1.0 & 1.0 & -1.0 & -1.0 & -1.0 & 0.3 & -0.3 & -0.3 \\ & & 1.0 & -1.0 & -1.0 & -1.0 & 0.3 & -0.3 & -0.3 \\ & & & 1.0 & 1.0 & 1.0 & -0.3 & 0.3 & 0.3 \\ & & & & 1.0 & 1.0 & 0.3 & 0.3 & 0.3 \\ & & & & & 1.0 & -0.3 & 0.3 & 0.3 \\ & & & & & & 1.0 & -1.0 & -1.0 \\ & & & & & & & 1.0 & 1.0 \\ & & & & & & & & 1.0 \end{pmatrix}$$

$$\begin{pmatrix} 1.0 & -1.0 & -1.0 & -1.0 & -1.0 & 1.0 & -0.8 & 0.8 & -0.8 \\ & 1.0 & 1.0 & 1.0 & 1.0 & -1.0 & 0.8 & -0.8 & 0.8 \\ & & 1.0 & 1.0 & 1.0 & -1.0 & 0.8 & -0.8 & 0.8 \\ & & & 1.0 & 1.0 & -1.0 & 0.8 & -0.8 & 0.8 \\ & & & & 1.0 & -1.0 & 0.8 & -0.8 & 0.8 \\ & & & & & 1.0 & -0.8 & 0.8 & -0.8 \\ & & & & & & 1.0 & -0.8 & 0.8 \\ & & & & & & 1.0 & -1.0 & 1.0 \\ & & & & & & & 1.0 & -1.0 \\ & & & & & & & & 1.0 \end{pmatrix}$$

where the sequence of the settings is following: $PC_1, PC_2, PC_3, WN_1, BM_1, BM_2, WN_2, WN_3, BM_3$. The algorithm converges to two different solutions depending on the starting values since the settings form clearly two clusters: $(PC_1, PC_2, PC_3, WN_1, BM_1, BM_2)$ and (WN_2, WN_3, BM_3). Inside the clusters the weights are almost perfectly correlated - top left and bottom right corners of the matrices contain 1 or -1. Above we present only one digit precision but the correlation is similar with four digit precision. Of course if the correlation of the time series estimates is -1 the same factors are considered because they are identifiable only up to sign. Between the clusters the correlation is not so strong. In order to choose one solution other criteria like explained variance or smoothness of IVS need to be taken into account.

The DSFM can easily be applied in hedging or risk management. Computing sensitivity with respect to factor loadings changes simplify the vega hedge since the whole dynamics of the IVS is reduced to L factors. After estimating stochastic model for $\widehat{\beta}$, like in [5] where VAR(2) was detected, it can be used for scenario generation in Monte Carlo framework. Therefore it allows to compute the VaR for portfolios containing options.

5. Conclusion

We discuss estimation issues of the DSFM, which gives a flexible way of handling IV data and is a convenient modelling tool. We study the dependence on the starting $\widehat{\beta}$ and the bandwidths settings. These are the key issues in efficient application of the model, which is left for future research.

6. Acknowledgement

References

[1] M. Benko and W. Härdle. *Common Functional Implied Volatility Analysis* in P. Čížek, W. Härdle and R. Weron (eds) *Statistical Tools for Fianance and Insurance*, chapter 5, pages 115–134. Springer Verlag, 2005.

[2] R. Cont and J. da Fosenca. Dynamics of implied volatility surfaces. *Quantitative Finance*, 2:45–60, February 2002.

[3] B. Dupire. Pricing with a smile. *RISK*, 1(7):18–20, 1994.

[4] E. Eberlein and K. Prause. *The generalized hyperbolic model: Financial derivatives and risk measures* in H. German, D. Madan, S. Pliska, T. Vorst (eds) *Mathematical Finance - Bachelier Congress 2000* pages 245–267. Springer Verlag, 2002.

[5] M. Fengler. *Semiparametric Modelling of Implied Volatility*. PhD thesis, Humboldt-Universität zu Berlin, 2004.

[6] M. Fengler, W. Härdle, and E. Mammen. A Dynamic Semiparametric Factor Model for Implied Volatility String Dynamics. CASE Discussion Paper, Humboldt-Universität zu Berlin, 2004.

[7] M. Fengler, W. Härdle, and P. Schmidt. Common factors govering VDAX movements and the maximum loss. *Journal of Financial Markets and Portfolio Managment*, 1(16):16–19, 2002.

[8] M. Fengler, W. Härdle, and C. Villa. The Dynamics of Implied Volatilities: A Common Principal Components Approach. *Review of Derivatives Research*, 6:179–202, 2003.

[9] R. Hafner. *Stochastic Implied Volatility*. Springer-Verlag, Berlin Heidelberg, 2004.

[10] P. Hagan, D. Kumar, D. Lesniewski and D. Woodward. Manging smile risk. *Wilmott magazine*, 1:84–108, 2002.

[11] W. Härdle. *Applied Nonparametric Regression*. Cambridge University Press, 1990.

[12] S. Heston. A closed-form solution for options with stochastic volatility with applications to bond and currency options. *Review of Financial Studies*, 6:327–343, 1993.

[13] J. Hull, A. White. The pricing on option on assets with stochastic volatilities. *Journal of Finance*, 42:281–300, 1987.

[14] R. Merton. Option pricing when underlying stock returns are discontinuous. *Journal of Financial Economics*, 3:125–144, 1976.

[15] G. Skiadopoulos, S. Hodges, and L. Clewlow. The dynamics of the S&P 500 implied volatility surface. *Review of Derivatives Research*, 3:263–282, 1999.

[16] Y. Zhu and M. Avellaneda. An E-ARCH model for the term-structure of implied volatility of FX options. *Applied Mathematical Finance*, 4:81–100, 1997.

Value-at-Risk Calculations with Time Varying Copulae

Enzo Giacomini, Wolfgang Härdle
Center for Applied Statistics and Economics
Institut für Statistik und Ökonometrie
Humboldt-Universität zu Berlin
Spandauer Str. 1, D-10178 Berlin
Germany
E-mail: giacomini@wiwi.hu-berlin.de

Value-at-Risk (VaR) of a portfolio is determined by the multivariate distribution of the risk factors increments. This distribution can be modelled through copulae, where the copulae parameters are not necessarily constant over time. For an exchange rate portfolio, copulae with time varying parameters are estimated and the VaR simulated accordingly. Backtesting underlines the improved performance of time varying copulae.

Value-at-Risk and Copulae

At time t a linear portfolio composed of d positions $w = (w_1, \ldots, w_d)^\top$ on assets with prices $S_t = (S_{1,t}, \ldots, S_{d,t})^\top$ and log prices $Z_t = \ln S_t$, has value

$$(1) \qquad V_t = \sum_{j=1}^d w_j e^{Z_{j,t}}$$

The *profit and loss* (P&L) function is defined as $L_{t+1} = (V_{t+1} - V_t)$. Defining $X_{t+1} = (Z_{t+1} - Z_t)$ as the time increment in the risk factors from period t to $t+1$, the P&L can be expressed as:

$$(2) \qquad L_{t+1} = \sum_{j=1}^d w_j S_{j,t}(e^{X_{j,t+1}} - 1)$$

The *Value-at-Risk* (VaR) is calculated as the α-quantile from F_L, the distribution of L:

$$(3) \qquad VaR = F_L^{-1}(\alpha)$$

The 1-dimensional distribution F_L depends on the d-dimensional distribution F_X. Using *copulae*, the marginal distributions F_{X_j} from each univariate increment can be *separately* modelled from their dependence structure and then *coupled* together to form the multivariate distribution F_X.

In the following, the dependence parameter $\hat{\theta}$ and the joint distribution $\hat{F}_X$ from a sample $\{X_t\}_{t=1}^T$ of log returns from exchange rate positions are estimated with copulae. A Monte Carlo simulation based on $\hat{F}_X$ generates different P&L samples. The quantiles at different levels from these simulation samples are then used as estimators for the Value-at-Risk of differem portfolio.

Computing Value-at-Risk with Copulae

A copula is a d-dimensional distribution function $C : [0,1]^d \to [0,1]$ with uniform marginals on the interval $[0,1]$. As in Nelsen (1998), multivariate distributions can be modelled via:

Theorem 1 (Sklar's theorem) *Let F be a d-dimensional distribution function with marginals $F_{X_1} \ldots, F_{X_d}$. Then there exists a copula C with*

$$(4) \qquad F(x_1, \ldots, x_d) = C\{F_{X_1}(x_1), \ldots, F_{X_d}(x_d)\}$$

for every $x_1, \ldots, x_d \in \overline{\mathbb{R}}$. If $F_{X_1}, \ldots, F_{X_d}$ are continuous, then C is unique. On the other hand, if C is a copula and $F_{X_1}, \ldots, F_{X_d}$ are distribution functions, then the function F defined in (4) is a joint distribution function with marginals $F_{X_1}, \ldots, F_{X_d}$.

The estimation of the Value-at-Risk, based on an *i.i.d.* sample $\{X_t\}_{t=1}^T$ is implemented in the following procedure:

1. specification of marginal distributions $F_{X_j}(x_j)$

2. specification of copula $C(u_1, \ldots, u_d; \theta)$

3. fitting the copula C to obtain $\hat{\theta}$

4. generation of Monte Carlo data $X_{T+1} \sim C(u_1, \ldots, u_d; \hat{\theta})$

5. generation of a sample of portfolio losses $L_{T+1}(X_{T+1})$

6. estimation of $\widehat{VaR}_{T+1}$, the empirical quantile at level α from $L_{T+1}(X_{T+1})$.

For copulae belonging to a parametric family $C = \{C_\theta, \theta \in \Theta\}$ and univariate marginals $F_{X_j}(x_j; \delta_j)$, the density of X is given by:

$$f(x_1, \ldots, x_d; \delta_1, \ldots, \delta_d, \theta) = c\{F_{X_1}(x_1; \delta_1), \ldots, F_{X_d}(x_d; \delta_d); \theta\} \prod_{j=1}^d f_j(x_j; \delta_j)$$

where
$$c(u_1, \ldots, u_d) = \frac{\partial^d C(u_1, \ldots, u_d)}{\partial u_1 \ldots \partial u_d}$$

In the IFM (*inference for margins*) method, the log-likelihood function for each of the marginal distributions

$$(5) \qquad \ell_j(\delta_j) = \sum_{t=1}^T \ln f_i(x_{j,t}; \delta_j), j = 1, \ldots, d$$

is maximized to obtain estimates $(\hat{\delta}_1, \ldots, \hat{\delta}_d)^\top$. The function

$$(6) \qquad \ell(\theta, \hat{\delta}_1, \ldots, \hat{\delta}_d) = \sum_{t=1}^T [\ln c\{F_{X_1}(x_{1,t}; \hat{\delta}_1), \ldots, F_{X_d}(x_{d,t}; \hat{\delta}_d); \theta\}]$$

is then maximized over θ to get the dependence parameter estimate $\hat{\theta}$. The estimates $\hat{\theta}_{IFM} = (\hat{\delta}_1, \ldots, \hat{\delta}_d, \hat{\theta})^\top$ solve

$$(\partial \ell_1 / \partial \delta_1, \ldots, \partial \ell_d / \partial \delta_d, \partial \ell / \partial \theta) = 0$$

Backtesting

This procedure is applied to a daily exchange rate portfolio (DEM/USD and GBP/USD from 01.12.1979 to 01.04.1994) with $T = 250$. The univariate risk factor increments (log returns) are assumed to be Gaussian distributed with parameters estimated from the data. The selected copulae belong to the bivariate one-parametric Gumbel family:

$$(7) \qquad C(u, v) = \exp(-[(\ln u)^\theta + (\ln v)^\theta]^{\theta^{-1}}), 1 \leq \theta \leq \infty$$

To evaluate the performance of the copula in the VaR calculations, different portfolio compositions are used to generate P&L samples. The quantiles from the samples at four levels $\alpha_1 = 0.05$, $\alpha_2 = 0.01$, $\alpha_3 = 0.005$ and $\alpha_4 = 0.001$ are used as estimators for VaR.

The estimated VaR is compared with the realization of the P&L function, an *exceedance* occurring for each P&L value smaller than the estimated VaR. The ratio of the number of exceedances to the number of observations gives the empirical level $\hat{\alpha}$.
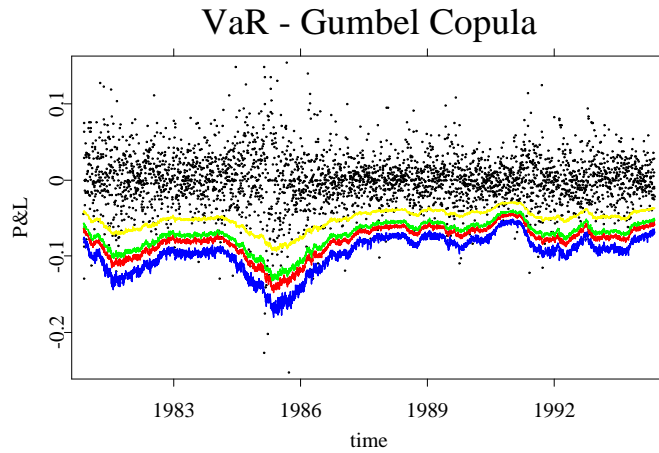

Figure 1: Value-at-Risk at levels $\alpha_1 = 0.05$ (yellow), $\alpha_2 = 0.01$ (green), $\alpha_3 = 0.005$ (red), and $\alpha_4 = 0.001$ (blue), P&L (black), estimated at each time from a Monte Carlo sample of 10.000 P&L values generated with Gumbel copula, $w = (2, 1)^\top$.
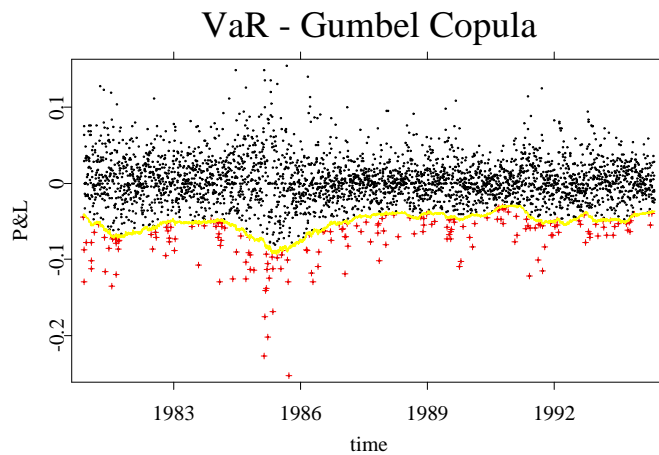

Figure 2: Value-at-Risk at level $\alpha = 0.05$ (yellow), P&L (black) and exceedances (red), $\hat{\alpha} = 0.0573$, $w = (2, 1)^\top$. P&L samples generated with Gumbel copula.

Table 1: *Gumbel copula, empirical levels* $\hat{\alpha}$ *for different portfolios.*

Portfolio $w^\top$	level $\alpha(\times 10^2)$			
	5	1	0.5	0.1
	empirical level $\hat{\alpha}(\times 10^2)$			
$(1,1)$	6.05	2.45	1.75	0.83
$(1,2)$	6.34	2.74	1.75	1.00
$(2,1)$	5.73	2.24	1.58	0.69
$(2,3)$	6.22	2.56	1.75	0.92
$(3,2)$	5.99	2.30	1.55	0.74
$(-1,2)$	1.64	0.37	0.20	0.11
$(1,-2)$	2.01	0.51	0.43	0.11
$(-2,1)$	4.44	1.49	0.95	0.40
$(2,-1)$	4.09	1.35	1.09	0.49

REFERENCES

Durrelman, V., Nikeghbali, A. and Roncalli, T. (2000). Which Copula is the Right One ?, *Groupe de Recherche Opérationnelle Crédit Lyonnais.*

Embrechts, P., McNeil, A. and Straumann, D. (1999). *Correlation and Dependence in Risk Management: Properties and Pitfalls Correlation, Risk Management: Value at Risk and Beyond,* Cambridge University Press, Cambridge.

Franke, J., Härdle, W. and Hafner, C. (2004). *Statistics of Financial Market*s, Springer-Verlag, Heidelberg.

Giacomini, E. and Härdle, W. (2005). Nonparametric Risk Management with Adaptive Copulae, *Discussion Paper, SFB 649.*

Härdle, W., Herwartz, H. and Spokoiny, V. (2003). Time Inhomogeneous Multiple Volatility Modeling, *Journal of Financial Econometrics*, 2003 1 (1): 55-95.

Härdle, W., Kleinow, T. and Stahl, G. (2002). *Applied Quantitative Finance*, Springer-Verlag, Heidelberg.

Joe, H. (1997). *Multivariate Models and Dependence Concepts*, Chapman & Hall, London.

Nelsen, R. (1998). *An Introduction to Copulas*, Springer-Verlag, New York.

RÉSUMÉ

La Value at Risk (VaR) d'un portefeuille est déterminée par la distribution multivariée des incréments des facteurs de risques. Cette distribution peut être modélisée par des copules dont les paramètres ne sont pas nécessairement constants par rapport au temps. Pour un portefeuille de taux de change, des copules dépendant du temps sont estimés et la VaR est ainsi simulée. Le backtesting confirme l'amélioration apportée par les copules dépendant du temps.

1 Stable Distributions

Szymon Borak, Wolfgang Härdle, and Rafał Weron

1.1 Introduction

Many of the concepts in theoretical and empirical finance developed over the past decades – including the classical portfolio theory, the Black-Scholes-Merton option pricing model and the RiskMetrics variance-covariance approach to Value at Risk (VaR) – rest upon the assumption that asset returns follow a normal distribution. However, it has been long known that asset returns are not normally distributed. Rather, the empirical observations exhibit fat tails. This heavy tailed or leptokurtic character of the distribution of price changes has been repeatedly observed in various markets and may be quantitatively measured by the kurtosis in excess of 3, a value obtained for the normal distribution (Bouchaud and Potters, 2000; Carr et al., 2002; Guillaume et al., 1997; Mantegna and Stanley, 1995; Rachev, 2003; Weron, 2004).

It is often argued that financial asset returns are the cumulative outcome of a vast number of pieces of information and individual decisions arriving almost continuously in time (McCulloch, 1996; Rachev and Mittnik, 2000). As such, since the pioneering work of Louis Bachelier in 1900, they have been modeled by the Gaussian distribution. The strongest statistical argument for it is based on the Central Limit Theorem, which states that the sum of a large number of independent, identically distributed variables from a finite-variance distribution will tend to be normally distributed. However, as we have already mentioned, financial asset returns usually have heavier tails.

In response to the empirical evidence Mandelbrot (1963) and Fama (1965) proposed the stable distribution as an alternative model. Although there are other heavy-tailed alternatives to the Gaussian law – like Student's t, hyperbolic, normal inverse Gaussian, or truncated stable – there is at least one good reason

for modeling financial variables using stable distributions. Namely, they are supported by the generalized Central Limit Theorem, which states that stable laws are the only possible limit distributions for properly normalized and centered sums of independent, identically distributed random variables.

Since stable distributions can accommodate the fat tails and asymmetry, they often give a very good fit to empirical data. In particular, they are valuable models for data sets covering extreme events, like market crashes or natural catastrophes. Even though they are not universal, they are a useful tool in the hands of an analyst working in finance or insurance. Hence, we devote this chapter to a thorough presentation of the computational aspects related to stable laws. In Section 1.2 we review the analytical concepts and basic characteristics. In the following two sections we discuss practical simulation and estimation approaches. Finally, in Section 1.5 we present financial applications of stable laws.

1.2   Definitions and Basic Characteristics

Stable laws – also called α-stable, stable Paretian or Lévy stable – were introduced by Levy (1925) during his investigations of the behavior of sums of independent random variables. A sum of two independent random variables having an α-stable distribution with index α is again α-stable with the same index α. This invariance property, however, does not hold for different α's.

The α-stable distribution requires four parameters for complete description: an index of stability $\alpha \in (0, 2]$ also called the tail index, tail exponent or characteristic exponent, a skewness parameter $\beta \in [-1, 1]$, a scale parameter $\sigma > 0$ and a location parameter $\mu \in \mathbb{R}$. The tail exponent α determines the rate at which the tails of the distribution taper off, see the left panel in Figure 1.1. When $\alpha = 2$, the Gaussian distribution results. When $\alpha < 2$, the variance is infinite and the tails are asymptotically equivalent to a Pareto law, i.e. they exhibit a power-law behavior. More precisely, using a central limit theorem type argument it can be shown that (Janicki and Weron, 1994; Samorodnitsky and Taqqu, 1994):

$$\begin{cases} \lim_{x \to \infty} x^{\alpha} \mathrm{P}(X > x) = C_{\alpha}(1 + \beta)\sigma^{\alpha}, \\ \lim_{x \to \infty} x^{\alpha} \mathrm{P}(X < -x) = C_{\alpha}(1 + \beta)\sigma^{\alpha}, \end{cases} \qquad (1.1)$$
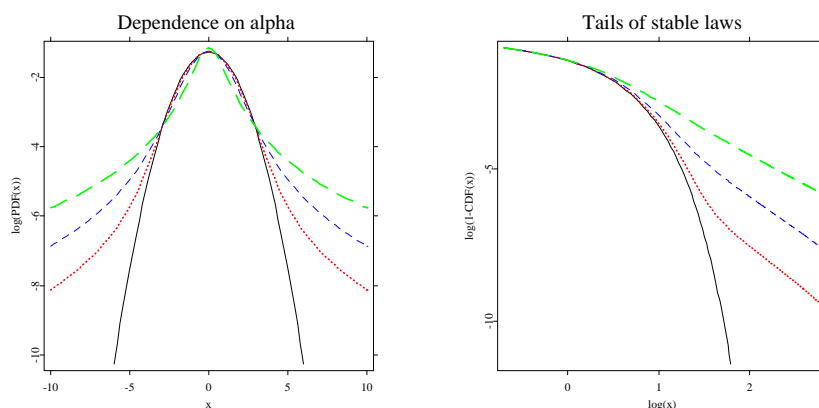
Figure 1.1: *Left panel*: A semilog plot of symmetric ($\beta = \mu = 0$) α-stable
probability density functions (pdfs) for $\alpha = 2$ (black solid line), 1.8
(red dotted line), 1.5 (blue dashed line) and 1 (green long-dashed
line). The Gaussian ($\alpha = 2$) density forms a parabola and is the
only α-stable density with exponential tails. *Right panel*: Right
tails of symmetric α-stable cumulative distribution functions (cdfs)
for $\alpha = 2$ (black solid line), 1.95 (red dotted line), 1.8 (blue dashed
line) and 1.5 (green long-dashed line) on a double logarithmic paper.
For $\alpha < 2$ the tails form straight lines with slope $-\alpha$.

<div align="right">STFstab01.xpl</div>

where:

$$C_\alpha = \left(2 \int_0^\infty x^{-\alpha} \sin(x) dx \right)^{-1} = \frac{1}{\pi} \Gamma(\alpha) \sin \frac{\pi\alpha}{2}.$$

The convergence to a power-law tail varies for different α's and, as can be seen
in the right panel of Figure 1.1, is slower for larger values of the tail index.
Moreover, the tails of α-stable distribution functions exhibit a crossover from
an approximate power decay with exponent $\alpha > 2$ to the true tail with exponent
α. This phenomenon is more visible for large α's (Weron, 2001).

When $\alpha > 1$, the mean of the distribution exists and is equal to μ. In general,
the pth moment of a stable random variable is finite if and only if $p < \alpha$. When
the skewness parameter β is positive, the distribution is skewed to the right,
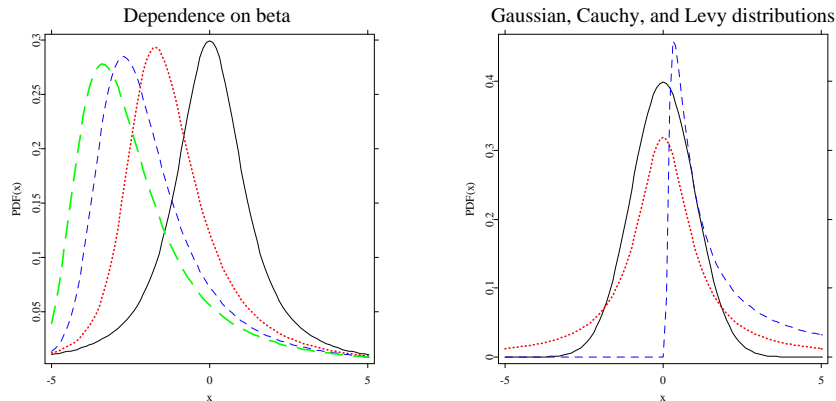
Figure 1.2: *Left panel*: Stable pdfs for $\alpha = 1.2$ and $\beta = 0$ (black solid line), 0.5 (red dotted line), 0.8 (blue dashed line) and 1 (green long-dashed line). *Right panel*: Closed form formulas for densities are known only for three distributions – Gaussian ($\alpha = 2$; black solid line), Cauchy ($\alpha = 1$; red dotted line) and Levy ($\alpha = 0.5, \beta = 1$; blue dashed line). The latter is a totally skewed distribution, i.e. its support is $\mathbb{R}_+$. In general, for $\alpha < 1$ and $\beta = 1 (-1)$ the distribution is totally skewed to the right (left).

STFstab02.xpl

i.e. the right tail is thicker, see the left panel of Figure 1.2. When it is negative, it is skewed to the left. When $\beta = 0$, the distribution is symmetric about μ. As α approaches 2, β loses its effect and the distribution approaches the Gaussian distribution regardless of β. The last two parameters, σ and μ, are the usual scale and location parameters, i.e. σ determines the width and μ the shift of the mode (the peak) of the density. For $\sigma = 1$ and $\mu = 0$ the distribution is called standard stable.

1.2.1 Characteristic Function Representation

Due to the lack of closed form formulas for densities for all but three distributions (see the right panel in Figure 1.2), the α-stable law can be most
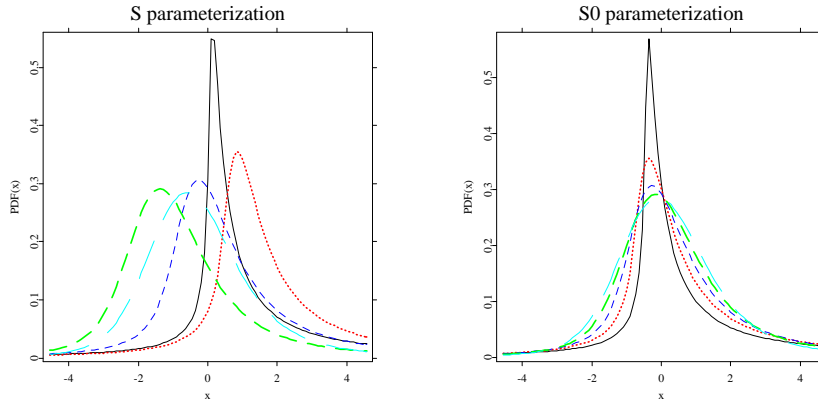
Figure 1.3: Comparison of S and S^0 parameterizations: α-stable pdfs for $\beta = 0.5$ and $\alpha = 0.5$ (black solid line), 0.75 (red dotted line), 1 (blue short-dashed line), 1.25 (green dashed line) and 1.5 (cyan long-dashed line).

STFstab03.xpl

conveniently described by its characteristic function $\phi(t)$ – the inverse Fourier transform of the probability density function. However, there are multiple parameterizations for α-stable laws and much confusion has been caused by these different representations, see Figure 1.3. The variety of formulas is caused by a combination of historical evolution and the numerous problems that have been analyzed using specialized forms of the stable distributions. The most popular parameterization of the characteristic function of $X \sim S_\alpha(\sigma, \beta, \mu)$, i.e. an α-stable random variable with parameters α, σ, β, and μ, is given by (Samorodnitsky and Taqqu, 1994; Weron, 2004):

$$
\ln \phi(t) = \begin{cases} -\sigma^\alpha |t|^\alpha \{1 - i\beta \mathrm{sign}(t) \tan \frac{\pi\alpha}{2}\} + i\mu t, & \alpha \neq 1, \\[2mm] -\sigma |t| \{1 + i\beta \mathrm{sign}(t) \frac{2}{\pi} \ln |t|\} + i\mu t, & \alpha = 1. \end{cases} \tag{1.2}
$$

For numerical purposes, it is often advisable to use Nolan's (1997) parameterization:

$$
\ln \phi_0(t) =
\begin{cases}
-\sigma^\alpha |t|^\alpha \{1 + i\beta\,\text{sign}(t)\tan\frac{\pi\alpha}{2}[(\sigma|t|)^{1-\alpha} - 1]\} + i\mu_0 t, & \alpha \neq 1, \\[2ex]
-\sigma|t|\{1 + i\beta\,\text{sign}(t)\frac{2}{\pi}\ln(\sigma|t|)\} + i\mu_0 t, & \alpha = 1.
\end{cases}
$$

$$(1.3)$$

The $S_\alpha^0(\sigma, \beta, \mu_0)$ parameterization is a variant of Zolotariev's (M)-parameterization (Zolotarev, 1986), with the characteristic function and hence the density and the distribution function jointly continuous in all four parameters, see the right panel in Figure 1.3. In particular, percentiles and convergence to the power-law tail vary in a continuous way as α and β vary. The location parameters of the two representations are related by $\mu = \mu_0 - \beta\sigma\tan\frac{\pi\alpha}{2}$ for $\alpha \neq 1$ and $\mu = \mu_0 - \beta\sigma\frac{2}{\pi}\ln\sigma$ for $\alpha = 1$. Note also, that the traditional scale parameter σ_G of the Gaussian distribution defined by:

$$
f_G(x) = \frac{1}{\sqrt{2\pi}\sigma_G}\exp\left\{-\frac{(x-\mu)^2}{2\sigma_G^2}\right\},
$$

$$(1.4)$$

is not the same as σ in formulas (1.2) or (1.3). Namely, $\sigma_G = \sqrt{2}\sigma$.

1.2.2   Stable Density and Distribution Functions

The lack of closed form formulas for most stable densities and distribution functions has negative consequences. For example, during maximum likelihood estimation computationally burdensome numerical approximations have to be used. There generally are two approaches to this problem. Either the fast Fourier transform (FFT) has to be applied to the characteristic function (Mittnik, Doganoglu, and Chenyao, 1999) or direct numerical integration has to be utilized (Nolan, 1997, 1999).

For data points falling between the equally spaced FFT grid nodes an interpolation technique has to be used. Taking a larger number of grid points increases accuracy, however, at the expense of higher computational burden. The FFT based approach is faster for large samples, whereas the direct integration method favors small data sets since it can be computed at any arbitrarily chosen point. Mittnik, Doganoglu, and Chenyao (1999) report that for $N = 2^{13}$ the FFT based method is faster for samples exceeding 100 observations and slower for smaller data sets. Moreover, the FFT based approach is less universal – it is efficient only for large α's and only for pdf calculations. When

computing the cdf the density must be numerically integrated. In contrast, in the direct integration method Zolotarev's (1986) formulas either for the density or the distribution function are numerically integrated.

Set $\zeta = -\beta \tan \frac{\pi \alpha}{2}$. Then the density $f(x; \alpha, \beta)$ of a standard α-stable random variable in representation S^0, i.e. $X \sim S_\alpha^0(1, \beta, 0)$, can be expressed as (note, that Zolotarev (1986, Section 2.2) used yet another parametrization):

- when $\alpha \neq 1$ and $x > \zeta$:

$$f(x; \alpha, \beta) = \frac{\alpha (x - \zeta)^{\frac{1}{\alpha-1}}}{\pi \mid \alpha - 1 \mid} \int_{-\xi}^{\frac{\pi}{2}} V(\theta; \alpha, \beta) \exp \left\{ -(x - \zeta)^{\frac{\alpha}{\alpha-1}} V(\theta; \alpha, \beta) \right\} d\theta,$$

$$(1.5)$$

- when $\alpha \neq 1$ and $x = \zeta$:

$$f(x; \alpha, \beta) = \frac{\Gamma(1 + \frac{1}{\alpha}) \cos(\xi)}{\pi (1 + \zeta^2)^{\frac{1}{2\alpha}}},$$

- when $\alpha \neq 1$ and $x < \zeta$:

$$f(x; \alpha, \beta) = f(-x; \alpha, -\beta),$$

- when $\alpha = 1$:

$$f(x; 1, \beta) = \begin{cases} \frac{1}{2|\beta|} e^{-\frac{\pi x}{2\beta}} \int_{-\frac{\pi}{2}}^{\frac{\pi}{2}} V(\theta; 1, \beta) \exp \left\{ -e^{-\frac{\pi x}{2\beta}} V(\theta; 1, \beta) \right\} d\theta, & \beta \neq 0, \\ \\ \frac{1}{\pi(1+x^2)}, & \beta = 0, \end{cases}$$

where

$$\xi = \begin{cases} \frac{1}{\alpha} \arctan(-\zeta), & \alpha \neq 1, \\ \frac{\pi}{2}, & \alpha = 1, \end{cases}$$

and

$$V(\theta; \alpha, \beta) = \begin{cases} (\cos \alpha \xi)^{\frac{1}{\alpha-1}} \left(\frac{\cos \theta}{\sin \alpha(\xi+\theta)} \right)^{\frac{\alpha}{\alpha-1}} \frac{\cos\{\alpha\xi + (\alpha-1)\theta\}}{\cos \theta}, & \alpha \neq 1, \\ \\ \frac{2}{\pi} \left(\frac{\frac{\pi}{2} + \beta\theta}{\cos \theta} \right) \exp \left\{ \frac{1}{\beta}(\frac{\pi}{2} + \beta\theta) \tan \theta \right\}, & \alpha = 1, \beta \neq 0. \end{cases}$$

The distribution $F(x; \alpha, \beta)$ of a standard α-stable random variable in representation S^0 can be expressed as:

- when $\alpha \neq 1$ and $x > \zeta$:

$$F(x; \alpha, \beta) = c_1(\alpha, \beta) + \frac{\text{sign}(1-\alpha)}{\pi} \int_{-\xi}^{\frac{\pi}{2}} \exp\left\{ -(x-\zeta)^{\frac{\alpha}{\alpha-1}} V(\theta; \alpha, \beta) \right\} d\theta,$$

where

$$c_1(\alpha, \beta) = \begin{cases} \frac{1}{\pi}\left(\frac{\pi}{2} - \xi\right), & \alpha < 1, \\ 1, & \alpha > 1, \end{cases}$$

- when $\alpha \neq 1$ and $x = \zeta$:

$$F(x; \alpha, \beta) = \frac{1}{\pi}\left(\frac{\pi}{2} - \xi\right),$$

- when $\alpha \neq 1$ and $x < \zeta$:

$$F(x; \alpha, \beta) = 1 - F(-x; \alpha, -\beta),$$

- when $\alpha = 1$:

$$F(x; 1, \beta) = \begin{cases} \frac{1}{\pi} \int_{-\frac{\pi}{2}}^{\frac{\pi}{2}} \exp\left\{ -e^{-\frac{\pi x}{2\beta}} V(\theta; 1, \beta) \right\} d\theta, & \beta > 0, \\[2mm] \frac{1}{2} + \frac{1}{\pi} \arctan x, & \beta = 0, \\[2mm] 1 - F(x, 1, -\beta), & \beta < 0. \end{cases}$$

Formula (1.5) requires numerical integration of the function $g(\cdot)\exp\{-g(\cdot)\}$, where $g(\theta; x, \alpha, \beta) = (x - \zeta)^{\frac{\alpha}{\alpha-1}} V(\theta; \alpha, \beta)$. The integrand is 0 at $-\xi$, increases monotonically to a maximum of $\frac{1}{e}$ at point θ^* for which $g(\theta^*; x, \alpha, \beta) = 1$, and then decreases monotonically to 0 at $\frac{\pi}{2}$ (Nolan, 1997). However, in some cases the integrand becomes very peaked and numerical algorithms can miss the spike and underestimate the integral. To avoid this problem we need to find the argument θ^* of the peak numerically and compute the integral as a sum of two integrals: one from $-\xi$ to θ^* and the other from θ^* to $\frac{\pi}{2}$.

1.3   Simulation of α-stable Variables

The complexity of the problem of simulating sequences of α-stable random variables results from the fact that there are no analytic expressions for the

inverse F^{-1} of the cumulative distribution function. The first breakthrough was made by Kanter (1975), who gave a direct method for simulating $S_\alpha(1, 1, 0)$ random variables, for $\alpha < 1$. It turned out that this method could be easily adapted to the general case. Chambers, Mallows, and Stuck (1976) were the first to give the formulas.

The algorithm for constructing a standard stable random variable $X \sim S_\alpha(1, \beta, 0)$, in representation (1.2), is the following (Weron, 1996):

- generate a random variable V uniformly distributed on $(-\frac{\pi}{2}, \frac{\pi}{2})$ and an independent exponential random variable W with mean 1;

- for $\alpha \neq 1$ compute:

$$X = S_{\alpha,\beta} \cdot \frac{\sin\{\alpha(V + B_{\alpha,\beta})\}}{\{\cos(V)\}^{1/\alpha}} \cdot \left[\frac{\cos\{V - \alpha(V + B_{\alpha,\beta})\}}{W}\right]^{(1-\alpha)/\alpha}, \quad (1.6)$$

where

$$\begin{aligned} B_{\alpha,\beta} &= \frac{\arctan(\beta \tan \frac{\pi\alpha}{2})}{\alpha}, \\ S_{\alpha,\beta} &= \left\{1 + \beta^2 \tan^2\left(\frac{\pi\alpha}{2}\right)\right\}^{1/(2\alpha)}; \end{aligned}$$

- for $\alpha = 1$ compute:

$$X = \frac{2}{\pi} \left\{ \left(\frac{\pi}{2} + \beta V\right) \tan V - \beta \ln\left(\frac{\frac{\pi}{2} W \cos V}{\frac{\pi}{2} + \beta V}\right) \right\}. \quad (1.7)$$

Given the formulas for simulation of a standard α-stable random variable, we can easily simulate a stable random variable for all admissible values of the parameters α, σ, β and μ using the following property: if $X \sim S_\alpha(1, \beta, 0)$ then

$$Y = \begin{cases} \sigma X + \mu, & \alpha \neq 1, \\ \sigma X + \frac{2}{\pi} \beta \sigma \ln \sigma + \mu, & \alpha = 1, \end{cases} \quad (1.8)$$

is $S_\alpha(\sigma, \beta, \mu)$. It is interesting to note that for $\alpha = 2$ (and $\beta = 0$) the Chambers-Mallows-Stuck method reduces to the well known Box-Muller algorithm for generating Gaussian random variables (Janicki and Weron, 1994). Although many other approaches have been proposed in the literature, this method is regarded as the fastest and the most accurate (Weron, 2004).
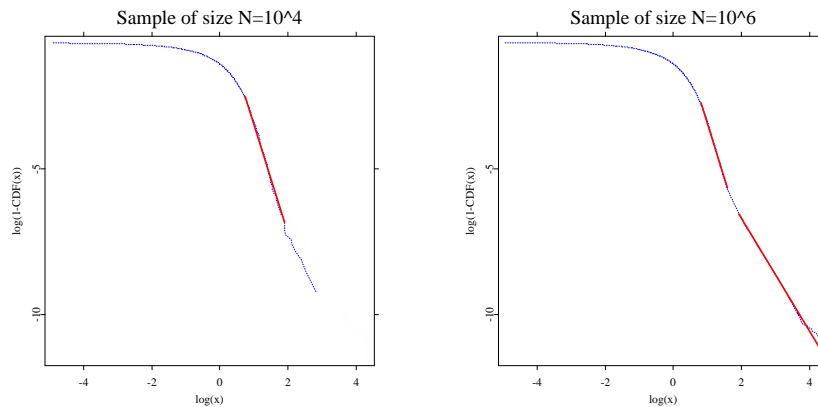
Figure 1.4: A double logarithmic plot of the right tail of an empirical symmetric 1.9-stable distribution function for a sample of size $N = 10^4$ (*left panel*) and $N = 10^6$ (*right panel*). Thick red lines represent the linear regression fit. The tail index estimate ($\hat{\alpha} = 3.7320$) obtained for the smaller sample is close to the initial power-law like decay of the larger sample ($\hat{\alpha} = 3.7881$). The far tail estimate $\hat{\alpha} = 1.9309$ is close to the true value of α.

STFstab04.xpl

1.4   Estimation of Parameters

Like simulation, the estimation of stable law parameters is in general severely hampered by the lack of known closed-form density functions for all but a few members of the stable family. Either the pdf has to be numerically integrated (see the previous section) or the estimation technique has to be based on a different characteristic of stable laws.

All presented methods work quite well assuming that the sample under consideration is indeed α-stable. However, if the data comes from a different distribution, these procedures may mislead more than the Hill and direct tail estimation methods. Since the formal tests for assessing α-stability of a sample are very time consuming we suggest to first apply the "visual inspection" tests to see whether the empirical densities resemble those of α-stable laws.

1.4.1 Tail Exponent Estimation

The simplest and most straightforward method of estimating the tail index is to plot the right tail of the empirical cdf on a double logarithmic paper. The slope of the linear regression for large values of x yields the estimate of the tail index α, through the relation $\alpha = -\text{slope}$.

This method is very sensitive to the size of the sample and the choice of the number of observations used in the regression. For example, the slope of about -3.7 may indicate a non-α-stable power-law decay in the tails or the contrary – an α-stable distribution with $\alpha \approx 1.9$. This is illustrated in Figure 1.4. In the left panel a power-law fit to the tail of a sample of $N = 10^4$ standard symmetric ($\beta = \mu = 0$, $\sigma = 1$) α-stable distributed variables with $\alpha = 1.9$ yields an estimate of $\hat{\alpha} = 3.732$. However, when the sample size is increased to $N = 10^6$ the power-law fit to the extreme tail observations yields $\hat{\alpha} = 1.9309$, which is fairly close to the original value of α.

The true tail behavior (1.1) is observed only for very large (also for very small, i.e. the negative tail) observations, after a crossover from a temporary power-like decay (which surprisingly indicates $\alpha \approx 3.7$). Moreover, the obtained estimates still have a slight positive bias, which suggests that perhaps even larger samples than 10^6 observations should be used. In Figure 1.4 we used only the upper 0.15% of the records to estimate the true tail exponent. In general, the choice of the observations used in the regression is subjective and can yield large estimation errors, a fact which is often neglected in the literature.

A well known method for estimating the tail index that does not assume a parametric form for the entire distribution function, but focuses only on the tail behavior was proposed by Hill (1975). The Hill estimator is used to estimate the tail index α, when the upper (or lower) tail of the distribution is of the form: $1 - F(x) = Cx^{-\alpha}$, see Figure 1.5. Like the log-log regression method, the Hill estimator tends to overestimate the tail exponent of the stable distribution if α is close to two and the sample size is not very large. For a review of the extreme value theory and the Hill estimator see Härdle, Klinke, and Müller (2000, Chapter 13) or Embrechts, Klüppelberg, and Mikosch (1997).

These examples clearly illustrate that the true tail behavior of α-stable laws is visible only for extremely large data sets. In practice, this means that in order to estimate α we must use high-frequency data and restrict ourselves to the most "outlying" observations. Otherwise, inference of the tail index may be strongly misleading and rejection of the α-stable regime unfounded.
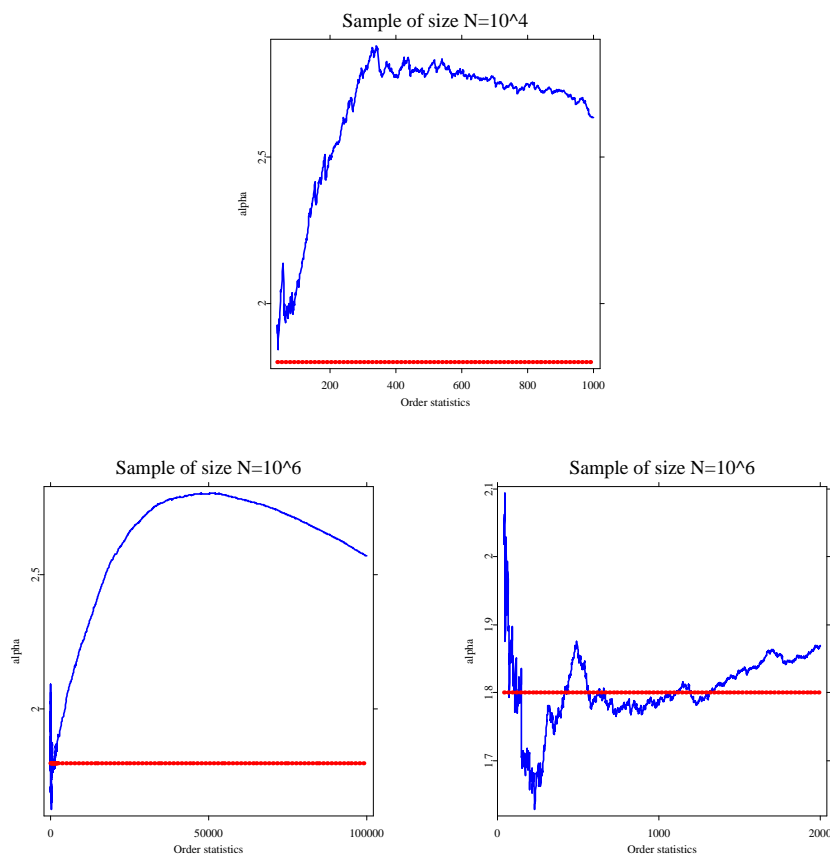
Figure 1.5: Plots of the Hill statistics $\hat{\alpha}_{n,k}$ vs. the maximum order statistic k for 1.8-stable samples of size $N = 10^4$ (*top panel*) and $N = 10^6$ (*left and right panels*). Red horizontal lines represent the true value of α. For better exposition, the right panel is a magnification of the left panel for small k. A close estimate is obtained only for $k = 500, ..., 1300$ (i.e. for $k < 0.13\%$ of sample size).

STFstab05.xpl

We now turn to the problem of parameter estimation. We start the discussion with the simplest, fastest and ... least accurate quantile methods, then develop the slower, yet much more accurate sample characteristic function methods and, finally, conclude with the slowest but most accurate maximum likelihood approach. Given a sample $x_1, ..., x_n$ of independent and identically distributed $S_\alpha(\sigma, \beta, \mu)$ observations, in what follows, we provide estimates $\hat{\alpha}$, $\hat{\sigma}$, $\hat{\beta}$, and $\hat{\mu}$ of all four stable law parameters.

1.4.2   Quantile Estimation

Already in 1971 Fama and Roll provided very simple estimates for parameters of symmetric ($\beta = 0, \mu = 0$) stable laws when $\alpha > 1$. McCulloch (1986) generalized and improved their method. He analyzed stable law quantiles and provided consistent estimators of all four stable parameters, with the restriction $\alpha \geq 0.6$, while retaining the computational simplicity of Fama and Roll's method. After McCulloch define:

$$v_\alpha = \frac{x_{0.95} - x_{0.05}}{x_{0.75} - x_{0.25}}, \qquad (1.9)$$

which is independent of both σ and μ. In the above formula x_f denotes the f-th population quantile, so that $S_\alpha(\sigma, \beta, \mu)(x_f) = f$. Let $\hat{v}_\alpha$ be the corresponding sample value. It is a consistent estimator of v_α. Now, define:

$$v_\beta = \frac{x_{0.95} + x_{0.05} - 2x_{0.50}}{x_{0.95} - x_{0.05}}, \qquad (1.10)$$

and let $\hat{v}_\beta$ be the corresponding sample value. v_β is also independent of both σ and μ. As a function of α and β it is strictly increasing in β for each α. The statistic $\hat{v}_\beta$ is a consistent estimator of v_β.

Statistics v_α and v_β are functions of α and β. This relationship may be inverted and the parameters α and β may be viewed as functions of v_α and v_β:

$$\alpha = \psi_1(v_\alpha, v_\beta), \ \ \beta = \psi_2(v_\alpha, v_\beta). \qquad (1.11)$$

Substituting v_α and v_β by their sample values and applying linear interpolation between values found in tables provided by McCulloch (1986) yields estimators $\hat{\alpha}$ and $\hat{\beta}$.

Scale and location parameters, σ and μ, can be estimated in a similar way. However, due to the discontinuity of the characteristic function for $\alpha = 1$ and $\beta \neq 0$ in representation (1.2), this procedure is much more complicated. We refer the interested reader to the original work of McCulloch (1986).

1.4.3 Characteristic Function Approaches

Given a sample $x_1, ..., x_n$ of independent and identically distributed (i.i.d.) random variables, define the sample characteristic function by

$$\hat{\phi}(t) = \frac{1}{n}\sum_{j=1}^{n} e^{itx_j}. \tag{1.12}$$

Since $|\hat{\phi}(t)|$ is bounded by unity all moments of $\hat{\phi}(t)$ are finite and, for any fixed t, it is the sample average of i.i.d. random variables $\exp(itx_j)$. Hence, by the law of large numbers, $\hat{\phi}(t)$ is a consistent estimator of the characteristic function $\phi(t)$.

Press (1972) proposed a simple estimation method, called the method of moments, based on transformations of the characteristic function. The obtained estimators are consistent since they are based upon estimators of $\phi(t)$, $\mathrm{Im}\{\phi(t)\}$ and $\mathrm{Re}\{\phi(t)\}$, which are known to be consistent. However, convergence to the population values depends on a choice of four points at which the above functions are evaluated. The optimal selection of these values is problematic and still is an open question. The obtained estimates are of poor quality and the method is not recommended for more than preliminary estimation.

Koutrouvelis (1980) presented a regression-type method which starts with an initial estimate of the parameters and proceeds iteratively until some prespecified convergence criterion is satisfied. Each iteration consists of two weighted regression runs. The number of points to be used in these regressions depends on the sample size and starting values of α. Typically no more than two or three iterations are needed. The speed of the convergence, however, depends on the initial estimates and the convergence criterion.

The regression method is based on the following observations concerning the characteristic function $\phi(t)$. First, from (1.2) we can easily derive:

$$\ln(-\ln|\phi(t)|^2) = \ln(2\sigma^\alpha) + \alpha \ln|t|. \tag{1.13}$$

The real and imaginary parts of $\phi(t)$ are for $\alpha \neq 1$ given by

$$\Re\{\phi(t)\} = \exp(-|\sigma t|^\alpha)\cos\left[\mu t + |\sigma t|^\alpha \beta \mathrm{sign}(t)\tan\frac{\pi\alpha}{2}\right],$$

and

$$\Im\{\phi(t)\} = \exp(-|\sigma t|^\alpha)\sin\left[\mu t + |\sigma t|^\alpha \beta \mathrm{sign}(t)\tan\frac{\pi\alpha}{2}\right].$$

The last two equations lead, apart from considerations of principal values, to

$$\arctan\left(\frac{\Im\{\phi(t)\}}{\Re\{\phi(t)\}}\right) = \mu t + \beta\sigma^\alpha \tan\frac{\pi\alpha}{2}\text{sign}(t)|t|^\alpha. \tag{1.14}$$

Equation (1.13) depends only on α and σ and suggests that we estimate these parameters by regressing $y = \ln(-\ln|\phi_n(t)|^2)$ on $w = \ln|t|$ in the model

$$y_k = m + \alpha w_k + \epsilon_k, \quad k = 1, 2, ..., K, \tag{1.15}$$

where t_k is an appropriate set of real numbers, $m = \ln(2\sigma^\alpha)$, and ϵ_k denotes an error term. Koutrouvelis (1980) proposed to use $t_k = \frac{\pi k}{25}, k = 1, 2, ..., K$; with K ranging between 9 and 134 for different estimates of α and sample sizes.

Once $\hat{\alpha}$ and $\hat{\sigma}$ have been obtained and α and σ have been fixed at these values, estimates of β and μ can be obtained using (1.14). Next, the regressions are repeated with $\hat{\alpha}$, $\hat{\sigma}$, $\hat{\beta}$ and $\hat{\mu}$ as the initial parameters. The iterations continue until a prespecified convergence criterion is satisfied.

Kogon and Williams (1998) eliminated this iteration procedure and simplified the regression method. For initial estimation they applied McCulloch's (1986) method, worked with the continuous representation (1.3) of the characteristic function instead of the classical one (1.2) and used a fixed set of only 10 equally spaced frequency points t_k. In terms of computational speed their method compares favorably to the original method of Koutrouvelis (1980). It has a significantly better performance near $\alpha = 1$ and $\beta \neq 0$ due to the elimination of discontinuity of the characteristic function. However, it returns slightly worse results for very small α.

1.4.4   Maximum Likelihood Method

The maximum likelihood (ML) estimation scheme for α-stable distributions does not differ from that for other laws, at least as far as the theory is concerned. For a vector of observations $x = (x_1, ..., x_n)$, the ML estimate of the parameter vector $\theta = (\alpha, \sigma, \beta, \mu)$ is obtained by maximizing the log-likelihood function:

$$L_\theta(x) = \sum_{i=1}^{n} \ln \tilde{f}(x_i; \theta), \tag{1.16}$$

where $\tilde{f}(\cdot; \theta)$ is the stable pdf. The tilde denotes the fact that, in general, we do not know the explicit form of the density and have to approximate it

numerically. The ML methods proposed in the literature differ in the choice of the approximating algorithm. However, all of them have an appealing common feature – under certain regularity conditions the maximum likelihood estimator is asymptotically normal.

Modern ML estimation techniques either utilize the FFT-based approach for approximating the stable pdf (Mittnik et al., 1999) or use the direct integration method (Nolan, 2001). Both approaches are comparable in terms of efficiency. The differences in performance result from different approximation algorithms, see Section 1.2.2.

Simulation studies suggest that out of the five described techniques the method of moments yields the worst estimates, well outside any admissible error range (Stoyanov and Racheva-Iotova, 2004; Weron, 2004). McCulloch's method comes in next with acceptable results and computational time significantly lower than the regression approaches. On the other hand, both the Koutrouvelis and the Kogon-Williams implementations yield good estimators with the latter performing considerably faster, but slightly less accurate. Finally, the ML estimates are almost always the most accurate, in particular, with respect to the skewness parameter. However, as we have already said, maximum likelihood estimation techniques are certainly the slowest of all the discussed methods. For example, ML estimation for a sample of a few thousand observations using a gradient search routine which utilizes the direct integration method is slower by 4 orders of magnitude than the Kogon-Williams algorithm, i.e. a few minutes compared to a few hundredths of a second on a fast PC! Clearly, the higher accuracy does not justify the application of ML estimation in many real life problems, especially when calculations are to be performed on-line.

1.5   Financial Applications of Stable Laws

Many techniques in modern finance rely heavily on the assumption that the random variables under investigation follow a Gaussian distribution. However, time series observed in finance – but also in other applications – often deviate from the Gaussian model, in that their marginal distributions are heavy-tailed and, possibly, asymmetric. In such situations, the appropriateness of the commonly adopted normal assumption is highly questionable.

It is often argued that financial asset returns are the cumulative outcome of a vast number of pieces of information and individual decisions arriving almost continuously in time. Hence, in the presence of heavy-tails it is natural

Table 1.1: Fits to 2000 Dow Jones Industrial Average (DJIA) index returns from the period February 2, 1987 – December 29, 1994. Test statistics and the corresponding p-values based on 1000 simulated samples (in parentheses) are also given.

Parameters:	α	σ	β	μ
α-stable fit	1.6411	0.0050	-0.0126	0.0005
Gaussian fit		0.0111		0.0003
Tests:	Anderson-Darling		Kolmogorov	
α-stable fit	0.6441		0.5583	
	(0.020)		(0.500)	
Gaussian fit	$+\infty$		4.6353	
	($<$0.005)		($<$0.005)	

Q STFstab06.xpl

to assume that they are approximately governed by a stable non-Gaussian distribution. Other leptokurtic distributions, including Student's t, Weibull, and hyperbolic, lack the attractive central limit property.

Stable distributions have been successfully fit to stock returns, excess bond returns, foreign exchange rates, commodity price returns and real estate returns (McCulloch, 1996; Rachev and Mittnik, 2000). In recent years, however, several studies have found, what appears to be strong evidence against the stable model (Gopikrishnan et al., 1999; McCulloch, 1997). These studies have estimated the tail exponent directly from the tail observations and commonly have found α that appears to be significantly greater than 2, well outside the stable domain. Recall, however, that in Section 1.4.1 we have shown that estimating α only from the tail observations may be strongly misleading and for samples of typical size the rejection of the α-stable regime unfounded. Let us see ourselves how well the stable law describes financial asset returns.

In this section we want to apply the discussed techniques to financial data. Due to limited space we chose only one estimation method – the regression approach of Koutrouvelis (1980), as it offers high accuracy at moderate computational time. We start the empirical analysis with the most prominent example – the Dow Jones Industrial Average (DJIA) index, see Table 1.1. The data set covers the period February 2, 1987 – December 29, 1994 and comprises 2000
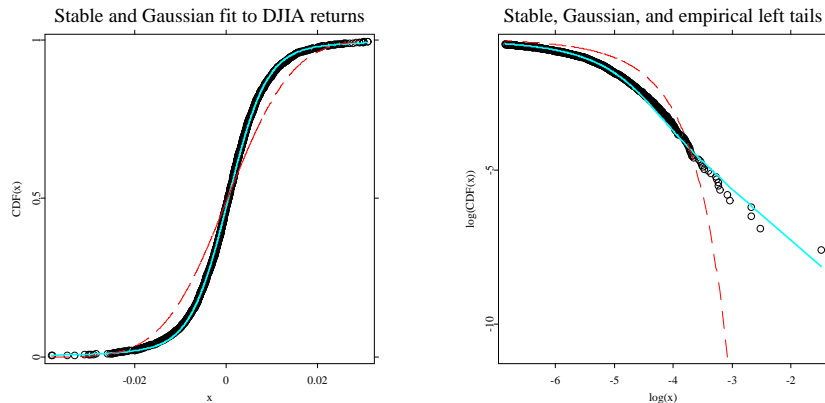
Figure 1.6: Stable (cyan) and Gaussian (dashed red) fits to the DJIA returns (black circles) empirical cdf from the period February 2, 1987 – December 29, 1994. Right panel is a magnification of the left tail fit on a double logarithmic scale clearly showing the superiority of the 1.64-stable law.

STFstab06.xpl

daily returns. Recall, that it includes the largest crash in Wall Street history – the Black Monday of October 19, 1987. Clearly the 1.64-stable law offers a much better fit to the DJIA returns than the Gaussian distribution. Its superiority, especially in the tails of the distribution, is even better visible in Figure 1.6.

To make our statistical analysis more sound, we also compare both fits through Anderson-Darling and Kolmogorov test statistics (D'Agostino and Stephens, 1986). The former may be treated as a weighted Kolmogorov statistics which puts more weight to the differences in the tails of the distributions. Although no asymptotic results are known for the stable laws, approximate p-values for these goodness-of-fit tests can be obtained via the Monte Carlo technique, for details see Chapter 13.   First the parameter vector is estimated for a given sample of size n, yielding $\hat{\theta}$, and the test statistics is calculated assuming that the sample is distributed according to $F(x; \hat{\theta})$, returning a value of d. Next, a sample of size n of $F(x; \hat{\theta})$-distributed variates is generated. The parameter
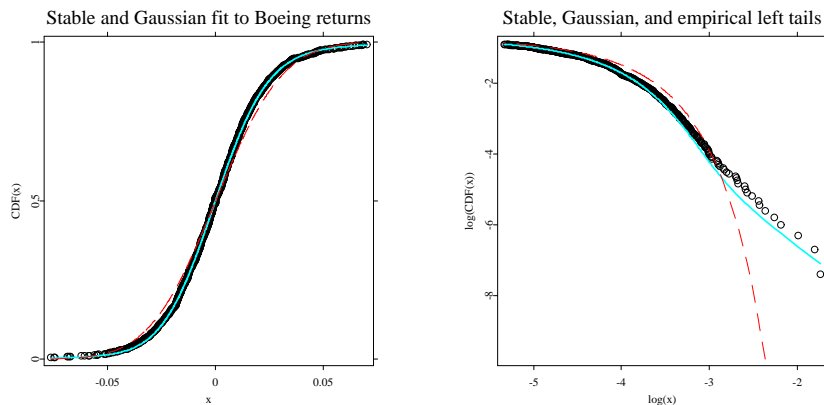
Figure 1.7: Stable (cyan) and Gaussian (dashed red) fits to the Boeing stock
returns (black circles) empirical cdf from the period July 1, 1997 –
December 31, 2003. Right panel is a magnification of the left tail
fit on a double logarithmic scale clearly showing the superiority of
the 1.78-stable law.

STFstab07.xpl

vector is estimated for this simulated sample, yielding $\hat{\theta}_1$, and the test statistics
is calculated assuming that the sample is distributed according to $F(x; \hat{\theta}_1)$. The
simulation is repeated as many times as required to achieve a certain level of
accuracy. The estimate of the p-value is obtained as the proportion of times
that the test quantity is at least as large as d.

For the α-stable fit of the DJIA returns the values of the Anderson-Darling and
Kolmogorov statistics are 0.6441 and 0.5583, respectively. The corresponding
approximate p-values based on 1000 simulated samples are 0.02 and 0.5 allowing
us to accept the α-stable law as a model of DJIA returns. The values of the
test statistics for the Gaussian fit yield p-values of less than 0.005 forcing us to
reject the Gaussian law, see Table 1.1.

Next, we apply the same technique to 1635 daily returns of Boeing stock prices
from the period July 1, 1997 – December 31, 2003. The 1.78-stable distribution
fits the data very well, yielding small values of the Anderson-Darling (0.3756)
and Kolmogorov (0.4522) test statistics, see Figure 1.7 and Table 1.2. The

Table 1.2: Fits to 1635 Boeing stock price returns from the period July 1, 1997
– December 31, 2003. Test statistics and the corresponding p-values
based on 1000 simulated samples (in parentheses) are also given.

Parameters:	α	σ	β	μ
α-stable fit	1.7811	0.0141	0.2834	0.0009
Gaussian fit		0.0244		0.0001

Tests:	Anderson-Darling		Kolmogorov	
α-stable fit	0.3756		0.4522	
	(0.18)		(0.80)	
Gaussian fit	9.6606		2.1361	
	($<$0.005)		($<$0.005)	

<div align="right">

Q STFstab07.xpl

</div>

approximate p-values based, as in the previous example, on 1000 simulated
samples are 0.18 and 0.8, respectively, allowing us to accept the α-stable law as
a model of Boeing returns. On the other hand, the values of the test statistics
for the Gaussian fit yield p-values of less than 0.005 forcing us to reject the
Gaussian distribution.

The stable law seems to be tailor-cut for the DJIA index and Boeing stock
price returns. But does it fit other asset returns as well? Unfortunately, not.
Although, for most asset returns it does provide a better fit than the Gaussian
law, in many cases the test statistics and p-values suggest that the fit is not
as good as for these two data sets. This can be seen in Figure 1.8 and Table
1.3, where the calibration results for 4444 daily returns of the Japanese yen
against the US dollar (JPY/USD) exchange rate from December 1, 1978 to
January 31, 1991 are presented. The empirical distribution does not exhibit
power-law tails and the extreme tails are largely overestimated by the stable
distribution. For a risk manager who likes to play safe this may not be a bad
idea, as the stable law overestimates the risks and thus provides an upper limit
of losses. However, from a calibration perspective other distributions, like the
hyperbolic or truncated stable, may be more appropriate for many data sets
(Weron, 2004).

Table 1.3: Fits to 4444 JPY/USD exchange rate returns from the period December 1, 1978 – January 31, 1991. Test statistics and the corresponding p-values (in parentheses) are also given.

Parameters:	α	σ	β	μ
α-stable fit	1.3274	0.0020	-0.1393	-0.0003
Gaussian fit		0.0049		-0.0001

Tests:	Anderson-Darling	Kolmogorov
α-stable fit	4.7833	1.4520
	(<0.005)	(<0.005)
Gaussian fit	91.7226	6.7574
	(<0.005)	(<0.005)

<div align="right">

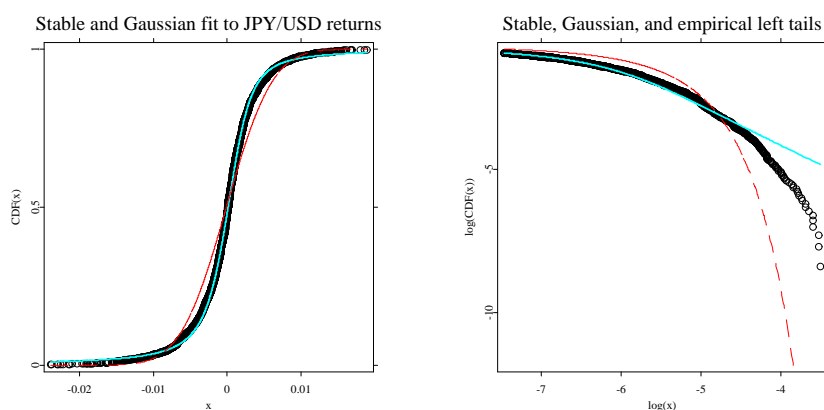![Q] STFstab08.xpl

</div>

Figure 1.8: Stable (cyan) and Gaussian (dashed red) fits to the JPY/USD exchange rate returns (black circles) empirical cdf from the period December 1, 1978 – January 31, 1991. Right panel is a magnification of the left tail fit on a double logarithmic scale. The extreme returns are largely overestimated by the stable law.

<div align="right">

![Q] STFstab08.xpl

</div>

Bibliography

Bouchaud, J.-P. and Potters, M. (2000). *Theory of Financial Risk*, Cambridge University Press, Cambridge.

Carr, P., Geman, H., Madan, D. B., and Yor, M. (2002). The fine structure of asset returns: an empirical investigation, *Journal of Business* **75**: 305–332.

Chambers, J. M., Mallows, C. L., and Stuck, B. W. (1976). A method for simulating stable random variables, *Journal of the American Statistical Association* **71**: 340–344.

D'Agostino, R. B. and Stephens, M. A. (1986). *Goodness-of-Fit Techniques*, Marcel Dekker, New York.

Embrechts, P., Kluppelberg, C., and Mikosch, T. (1997). *Modelling Extremal Events for Insurance and Finance*, Springer.

Fama, E. F. (1965). The behavior of stock market prices, *Journal of Business* **38**: 34–105.

Fama, E. F. and Roll, R. (1971). Parameter estimates for symmetric stable distributions, *Journal of the American Statistical Association* **66**: 331–338.

Gopikrishnan, P., Plerou, V., Amaral, L. A. N., Meyer, M. and Stanley, H. E. (1999). Scaling of the distribution of fluctuations of financial market indices, *Physical Review E* **60(5)**: 5305–5316.

Guillaume, D. M., Dacorogna, M. M., Dave, R. R., Müller, U. A., Olsen, R. B., and Pictet, O. V. (1997). From the birds eye to the microscope: A survey of new stylized facts of the intra-daily foreign exchange markets, *Finance & Stochastics* **1**: 95–129.

Härdle, W., Klinke, S., and Müller, M. (2000). *XploRe Learning Guide*, Springer.

Hill, B. M. (1975). A simple general approach to inference about the tail of a distribution, *Annals of Statistics* **3**: 1163–1174.

Janicki, A. and Weron, A. (1994). *Simulation and Chaotic Behavior of α-Stable Stochastic Processes*, Marcel Dekker.

Kanter, M. (1975). Stable densities under change of scale and total variation inequalities, *Annals of Probability* **3**: 697–707.

Koutrouvelis, I. A. (1980). Regression-type estimation of the parameters of stable laws, *Journal of the American Statistical Association* **75**: 918–928.

Kogon, S. M. and Williams, D. B. (1998). Characteristic function based estimation of stable parameters, *in* R. Adler, R. Feldman, M. Taqqu (eds.), *A Practical Guide to Heavy Tails*, Birkhauser, pp. 311–335.

Levy, P. (1925). *Calcul des Probabilites*, Gauthier Villars.

Mandelbrot, B. B. (1963). The variation of certain speculative prices, *Journal of Business* **36**: 394–419.

Mantegna, R. N. and Stanley, H. E. (1995). Scaling behavior in the dynamics of an economic index, *Nature* **376**: 46–49.

McCulloch, J. H. (1986). Simple consistent estimators of stable distribution parameters, *Communications in Statistics – Simulations* **15**: 1109–1136.

McCulloch, J. H. (1996). Financial applications of stable distributions, *in* G. S. Maddala, C. R. Rao (eds.), *Handbook of Statistics, Vol. 14*, Elsevier, pp. 393–425.

McCulloch, J. H. (1997). Measuring tail thickness to estimate the stable index α: A critique, *Journal of Business & Economic Statistics* **15**: 74–81.

Mittnik, S., Doganoglu, T., and Chenyao, D. (1999). Computing the probability density function of the stable Paretian distribution, *Mathematical and Computer Modelling* **29**: 235–240.

Mittnik, S., Rachev, S. T., Doganoglu, T. and Chenyao, D. (1999). Maximum likelihood estimation of stable Paretian models, *Mathematical and Computer Modelling* **29**: 275–293.

Nolan, J. P. (1997). Numerical calculation of stable densities and distribution functions, *Communications in Statistics – Stochastic Models* **13**: 759–774.

Nolan, J. P. (1999). An algorithm for evaluating stable densities in Zolotarev's (M) parametrization, *Mathematical and Computer Modelling* **29**: 229–233.

Nolan, J. P. (2001). Maximum likelihood estimation and diagnostics for stable distributions, *in* O. E. Barndorff-Nielsen, T. Mikosch, S. Resnick (eds.), *Lévy Processes*, Brikhäuser, Boston.

Press, S. J. (1972). Estimation in univariate and multivariate stable distribution, *Journal of the American Statistical Association* **67**: 842–846.

Rachev, S., ed. (2003). *Handbook of Heavy-tailed Distributions in Finance*, North Holland.

Rachev, S. and Mittnik, S. (2000). *Stable Paretian Models in Finance*, Wiley.

Samorodnitsky, G. and Taqqu, M. S. (1994). *Stable Non–Gaussian Random Processes*, Chapman & Hall.

Stoyanov, S. and Racheva-Iotova, B. (2004). Univariate stable laws in the field of finance – parameter estimation, *Journal of Concrete and Applicable Mathematics* **2(4)**, in print.

Weron, R. (1996). On the Chambers-Mallows-Stuck method for simulating skewed stable random variables, *Statistics and Probability Letters* **28**: 165–171. *See also* R. Weron, *Correction to: On the Chambers-Mallows-Stuck method for simulating skewed stable random variables*, Research Report HSC/96/1, Wrocław University of Technology, 1996, http://www.im.pwr.wroc.pl/~hugo/Publications.html.

Weron, R. (2001). Levy-stable distributions revisited: Tail index > 2 does not exclude the Levy-stable regime, *International Journal of Modern Physics C* **12**: 209–223.

Weron, R. (2004). Computationally intensive Value at Risk calculations, *in* J. E. Gentle, W. Härdle, Y. Mori (eds.) *Handbook of Computational Statistics*, Springer, Berlin, 911–950.

Zolotarev, V. M. (1986). *One-Dimensional Stable Distributions*, American Mathematical Society.

21 Working with the XQC

Szymon Borak, Wolfgang Härdle, and Heiko Lehmann

21.1 Introduction

An enormous number of statistical methods have been developed in quantitive finance during the last decades. Nonparametric methods, bootstrapping time series, wavelets, Markov Chain Monte Carlo are now almost standard in statistical applications. To implement these new methods the method developer usually uses a programming environment he is familiar with. Thus, automatically such methods are only available for preselected software packages, but not for widely used standard software packages like MS Excel. To apply these new methods to empirical data a potential user faces a number of problems or it may even be impossible for him to use the methods without rewriting them in a different programming language. Even if one wants to apply a newly developed method to simulated data in order to understand the methodology one is confronted with the drawbacks described above. A very similar problem occurs in teaching statistics at undergraduate level. Since students (by definition!) have their preferred software and often do not have access to the same statistical software packages as their teacher, illustrating examples have to be executable with standard tools. The delayed proliferation of new statistical technology over heterogeneous platforms and the evident student/teacher software gap are examples of inefficient distribution of quantitative methodology. This chapter describes the use of a platform independent client that is the basis for e-books, transparencies and other knowledge based systems.

In general, two statisticians are on either side of the distribution process of newly implemented methods, the provider (inventor) of a new technique (algorithm) and the user who wants to apply (understand) the new technique. The aim of the XploRe Quantlet client/server architecture is to bring these statisticians closer to each other. The XploRe Quantlet Client (XQC) represents the

front end – the user interface (UI) of this architecture allowing to access the XploRe server and its methods and data. The XQC is fully programmed in Java not depending on a specific computer platform. It runs on Windows and Mac platforms as well as on Unix and Linux machines.

The following sections contain a description of components and functionalities the XQC offers. Section 21.2.1 gives a short overview about possible configuration settings of the XQC, which allow influencing the behaviour of the client. Section 21.2.2 explains how to connect the XQC to an XploRe Quantlet Server. A detailed description of the XQC's components *desktop*, *Quantlet editor*, *data editor* and *method tree* is part of Sections 21.3 to 21.3.3. Section 21.3.4 finally explains graphical features offered by the XploRe Quantlet Client.

21.2   The XploRe Quantlet Client

The XploRe Quantlet Client can be initiated in two different ways. The way depends on whether the XQC is supposed to run as a standalone application or as an applet embedded within an HTML page. The XQC comes packed in a single Java Archive (JAR) file, which allows an easy usage. This JAR file allows for running the XQC as an application, as well as running it as an applet.

Running the XQC as an **application** does not require any programming skills. Provided that a Java Runtime Environment is installed on the computer the XQC is supposed to be executed on, the *xqc.jar* will automatically be recognized as an executable jar file that opens with the program *javaw*. If the XQC is embedded in a HTML page it runs as an **applet** and can be started right after showing the page.

21.2.1   Configuration

Property files allow configuring the XQC to meet special needs of the user. These files can be used to manage the appearance and behavior of the XQC. Any text editor can be used in editing the configuration files. Generally, the use of all information is optional. In its current version, the XQC works with three different configuration files. The *xqc.ini* file contains important information about the basic setup of the XploRe Quantlet Client, such as server and port information the client is supposed to connect to. It also contains information

Figure 21.1: Manual input for server and port number.

about the size of the client. This information can be maintained either relative to the actual size of the screen by using a factor or by stating its exact width and height. If this information is missing, the XQC begins by using its default values.

The *xqc_language.ini* allows for setting up the XQC's language. This file contains all texts used within the XQC. To localize the client, the texts have to be translated. If no language file can be found, the client starts with its default setup, showing all menus and messages in English.

The *xqc_methodtree.ini* file finally contains information about the method tree that can be shown as part of the METHOD/DATA window, see Section 21.3.2. A detailed description of the set up of the method tree will be part of Section 21.3.3.

21.2.2   Getting connected

After starting the XQC the client attempts to access and read information from the configuration files. If no configuration file is used error messages will pop up. If server and port information cannot be found, a pop up appears and enables a manual input of server and port number, as displayed in Figure 21.1.

The screenshot in Figure 21.2 shows the XQC after it has been started and connected to an XploRe server. A traffic light in the lower right corner of the screen indicates the actual status of the server. A green light means the client
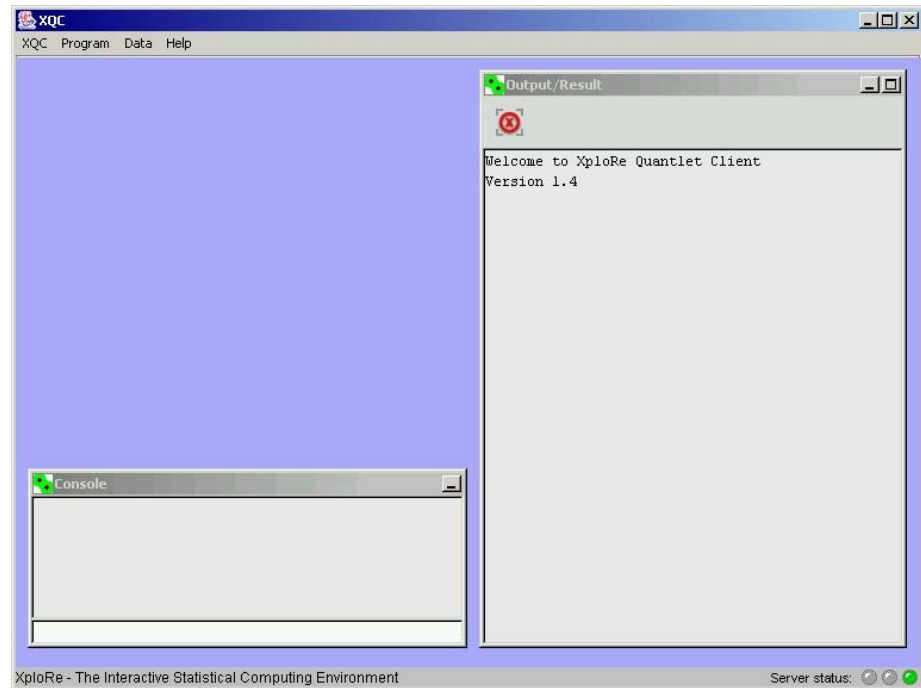
Figure 21.2: XQC connected and ready to work.

has successfully connected to the server and the server is ready to work. If the server is busy, computing previously received XploRe code, the traffic light will be set to yellow. A red light indicates that the XQC is not connected to the server.

21.3   Desktop

If no further restrictions or features are set in the configuration file (e.g. not showing any window or starting with executing a certain XploRe Quantlet) the XQC should look like shown in the screen shot. It opens with the two screen components CONSOLE and OUTPUT/RESULT window. The CONSOLE allows for the sending of single-line XploRe commands to the server to be

executed immediately. It also offers a history of the last 20 commands sent to the server. To repeat a command from the history, all that is required is a mouse click on the command, and it will be copied to the command line. Pressing the 'Return' key on the keyboard executes the XploRe command.

Text output coming from the XploRe server will be shown in the OUTPUT / RESULT window. Any text that is displayed can be selected and copied for use in other applications – e.g. for presentation of results within a scientific article. At the top of the screen the XQC offers additional functions via a menu bar. These functions are grouped into four categories. The **XQC** menu contains the features *Connect*, *Disconnect*, *Reconnect* and *Quit*.

Depending on the actual server status not every feature is enabled – e.g. if the client is not connected (the server status is indicated by a red traffic light) it does not make sense to disconnect or reconnect, if the client is already connected (server status equals a green light) the connect feature is disabled.

21.3.1   XploRe Quantlet Editor

The **Program** menu contains the features *New Program*, *Open Program (local). . .* and *Open Program (net). . . .* *New Program* opens a new and empty text editor window. This window enables the user to construct own XploRe Quantlets.

The feature *Open Program (local)* offers the possibility of accessing XploRe Quantlets stored on the local hard disk drive. It is only available if the XQC is running as an application or a certified applet. Due the Java sandbox restrictions running the XQC as an unsigned applet, it is not possible to access local programs.

If the user has access to the internet the menu item *Open Program (net)* can be useful. This feature allows the opening of Quantlets stored on a remote Web server. All it needs is the filename and the URL address at which the file is located.

Figure 21.3 shows a screen shot of the editor window containing a simple XploRe Quantlet. Two icons offer actions on the XploRe code:

- – Represents the probably most important feature – it sends the XploRe Quantlet to the server for execution.
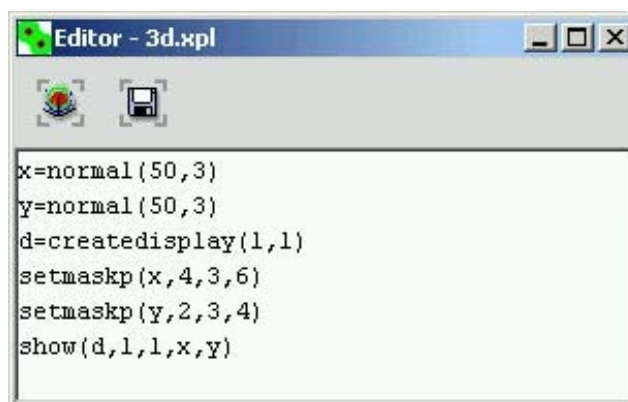
Figure 21.3: XploRe Editor window.

- – Saves the XploRe Quantlet to your local computer (not possible if running the XQC as an unsigned applet).

The Quantlet shown in Figure 21.3 assigns two three-dimensional standard normal distributions to the variables x and y. The generated data are formatted to a certain color, shape and size using the command `setmaskp`. The result is finally shown in a single display.

21.3.2   Data editor

The **Data** menu contains the features *New Data...*, *Open Data (local)...*, *Open Data (net)...*, *Download DataSet from Server...* and *DataSets uploaded to Server.*

New Data can be used to generate a new and empty data window. Before the data frame opens a pop-up window as shown in Figure 21.4 appears, asking for the desired dimension – the number of rows and cols – of the new data set. The XQC needs this information to create the spreadsheet. This definition does not have to be the exact and final decision, it is possible to add and delete rows and columns later on.

Figure 21.4: Dimension of the Data Set.

The menu item *Open Data (local)* enables the user to open data sets stored on the local hard disk. Again, access to local resources of the user's computer is only possible if the XQC is running as an application or a certified applet. The file will be interpreted as a common text format file. Line breaks within the file are considered as n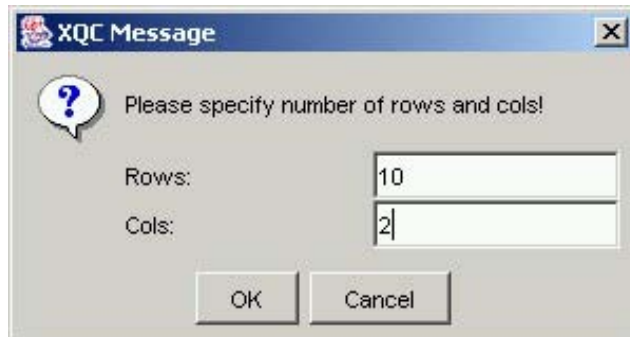ew rows for the data set. To recognize data belonging to a certain column the single data in one line must be separated by either using a ";" or a "tab" (separating the data by just a "space" will force the XQC to open the complete line in just on cell).

Open Data (net) lets the user open a data set that is stored on a web server by specifying the URL.

The menu item *Download DataSet from Server* offers the possibility to download data from the server. The data will automatically be opened in a new method and data window, offering all features of the method and data window (e.g. applying methods, saving, . . .) to them.

A helpful feature especially for research purposes is presented with the menu item *DataSets uploaded to Server*. This item opens a window that contains a list of objects uploaded to the server using the data window or the console. Changes of these objects are documented as an object history. Due to performance reasons only uploaded data and actions on data from the CONSOLE and the TABLE MODEL are recorded.

The appearance of the data window depends on the settings in the configuration file. If a method tree is defined and supposed to be shown, the window shows
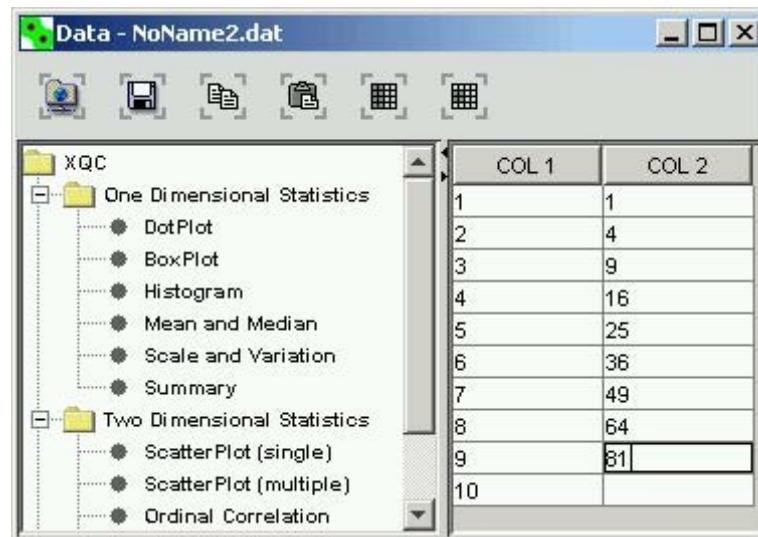
Figure 21.5: Combined Data and Method Window.

the method tree on the left part and data spreadsheet on the right part of the frame. If no method tree has been defined, only the spreadsheet will be shown. The method tree will be discussed in more detail in Section 21.3.3. Figure 21.5 shows a screen shot of the combined data and method frame.

Icons on the upper part of the data and method window offer additional functionalities:

- – If columns or cells are selected – this specific selection, otherwise the entire data set can be uploaded to the server with specifying a variable name.

- – Saves the data to your local computer (not possible if running the XQC as an unsigned applet).

- / – Copy and paste.

- ⊞ / ⊞ – Switches the column or cell selection mode on and off. Selected columns/cells can be uploaded to the server or methods can be executed on them.

The spreadsheet of the data and method window also offers a context menu containing the following items:

- Copy

- Paste

- No Selection Mode – Switches OFF the column or cell selection mode.

- Column Selection Mode – Switches ON the column selection mode.

- Cell Selection Mode – Switches ON the cell selection mode.

- Set Row as Header Line

- Set column Header

- Delete single Row

- Insert single Row

- Add single Row

- Delete single Column

- Add single Column

Most of the context menu items are self-explaining. However, there are two items that are worth taking a closer look at – *'Set Row as Header Line'* and *'Set column Header'*. The spreadsheet has the capability to specify a header for each column. This information can be used within XploRe Quantlets to name the axis within a plot, making it easier for the user to interpret graphics. A more detailed description is included in Section 21.3.3. Default values for the headers are COL1, COL2, . . . as shown in Figure 21.6. Naming a single column can be performed using the menu item *'Set column Header'*. The name has to be maintained within the pop up window that appears right after choosing this menu item. It can also be used to change existing column headers. The spreadsheet also offers the possibility to set column headers all at once. If the
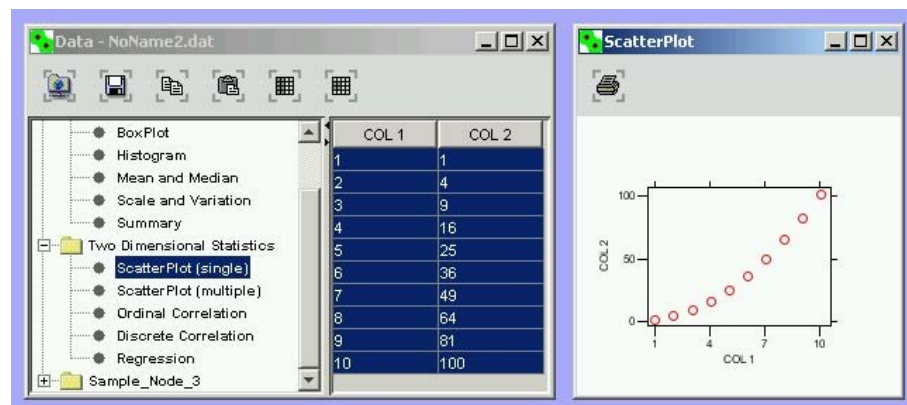
Figure 21.6: Working with the Data and Method Window.

data set already contains a row with header information – either coming from manual input or as part of an opened data set – these row can be set as header using the menu item *'Set Row as Header Line'*. The row with the cell that is active at that time will be cut out of the data set and pasted into the header line.

Setting the header is also possible while opening a data set. After choosing the data, a pop up asks whether or not the first row of the data set to be opened should be used as the header. Nevertheless, the context menu features just described above are of course still available, enabling the user to set or change headers afterwards.

Working with the XQC's method and data window does not require any XploRe programming knowledge. All it requires is a pointing device like the mouse. Applying, for example, the scatter-plot-method on the two columns would only mean to

- switch on the column selection mode

- mark both columns

- mouse click on the method "Scatter Plot"

Result will be a plot as shown in Figure 21.6. As stated above, the selected area can also be uploaded to the server using the icon in order to be used for further investigation. This new variable can be used within XploRe Quantlets written using the EDITOR window or manipulated via the CONSOLE.

21.3.3   Method tree

The METHOD TREE represents a tool for accessing statistical methods in an easy way. Its setup does not require any Java programming skills. All it needs is the maintenance of two configuration files.

Settings maintained within the *xqc.ini* file tell to the XQC whether there will be a method tree to be shown or not and where to get the tree information from. The client also needs to know where the methods are stored at. The `MethodPath` contains this information. Path statements can either be absolute statements or relative to the directory the XQC has been started in. For relative path information the path must start with `XQCROOT`. The settings in the example below tell the client to generate a method tree by using the file *xqc_methodtree.ini* with the XploRe Quantlets stored in the relative subdirectory `xqc_quantlets/`.

```
ShowMethodTree    = yes
MethodTreeIniFile = xqc_methodtree.ini
MethodPath        = XQCROOT/xqc_quantlets/
```

The actual method tree is set up in a separate configuration file that is given by the property of `MethodTreeIniFile`. This file contains a systematic structure of the tree – nodes and children, the method to be executed and its description to be shown within the tree frame.

```
Node_1 = path name
  Child_1.1 = method|description
  Child_1.2 = method|description
  Child_1.3 = method|description
  Node_2 = path name
    Node_2.1 = path name
      Child_2.1.1 = method|description
```

The name of the method has to be identical to the name of the XploRe program (Quantlet). The Quantlet itself has to have a procedure with the same name as the method. This procedure is called by the XQC on execution within the method tree.

Example

The following example shows how to set up a simple method tree. First of all, we choose XploRe Quantlets used within this e-book that we want to be part of the method tree. The aim of the Quantlet should be to generate graphics from selected data of the data spreadsheet or to just generate text output. Before being able to use the Quantlets within the method tree, they have to be 'wrapped' in a procedure. The name of the procedure – in our case for example 'STFstab08MT' – has to equal the name of the saved XploRe file. Our example Quantlet *STFstab08MT.xpl* is based on the original Quantlet *STFstab08.xpl* used in Chapter 1. The procedure must further have two parameters:

- `data` – Used for passing the selected data to the XploRe Quantlet.

- `names` – Contains the names of the selected columns taken from the header of the spreadsheet.

It might also be necessary to make some minor adjustments within the Quantlet in order to refer to the parameter handed over by the XQC. Those changes depend on the Quantlet itself.

```
library ("graphic")
proc() = STFstab08MT(data, names)
    ...
endp
```

Figure 21.7: STFstab08MT.xpl.

The XploRe coding within the procedure statement is not subject to any further needs or restrictions.

Once we have programmed the Quantlet it needs to be integrated into a method tree. For this purpose we define our own configuration file - *xqc_methodtree_STF* – with the following content shown in Figure 21.8.

```
Node_1 = Stable Distribution

    Node_1.1 = Estimation
        Child_1.1.1 = stabreg.xpl|Stabreg
        Child_1.1.2 = stabcull.xpl|Stabcull
        Child_1.1.3 = stabmom.xpl|Stabmom

    Node_1.2 = Examples
        Child_1.2.1 = STFstab08.xpl|STFstab08
        Child_1.2.2 = STFstab09.xpl|STFstab09
        Child_1.2.3 = STFstab10.xpl|STFstab10
```

Figure 21.8: sample_tree.ini

We create a node calling it 'Estimation'. Below this first node we set up the Quantlets *stabreg.xpl*, *stabcull.xpl* and *stabmom.xpl*. A second node – 'Examples' contains the Quantlets *STFstab08.xpl*, *STFstab09.xpl* and *STFstab10.xpl*. The text stated right beside each Quantlet (separated by the '|') represents the text we would like to be shown in the method tree.

Now that we have programmed the XploRe Quantlet(s) and set up the method tree we still need to tell the XQC to show our method tree upon opening data sets.

```
...

ShowMethodTree     = yes
MethodTreeIniFile  = xqc_methodtree_STF.ini
MethodPath         = XQCROOT/xqc_quantlets/

...
```

Figure 21.9: Extract of the xqc.ini.

The settings as shown in Figure 21.9 tell the XQC to show the method tree that is set up in our *xqc_methodtree_STF.ini* file and to use our XploRe Quantlet stored in a subdirectory of the XQC itself.

Our method tree is now ready for finally being tested. Figure 21.10 shows a screenshot of the final result – the method tree, set up above.

21.3.4   Graphical output

The previous sections contain some examples of graphical output shown within a display. The XQC's displays do not show only the graphical results received

Figure 21.10: Final result of our tree example.

from the XploRe server. Besides the possibility to print out the graphic it offers additional features that can be helpful for investigating data - especially for three-dimensional plots. Those features can be accessed via the display's context menu. Figure 21.11 shows three-dimensional plot of the 236 implied volatilities and fitted implied volatility surface of DAX from January 4th 1999.

The red points in the plot represent observed implied volatilities on 7 different maturities $T = 0.13, 0.21, 0.46, 0.71, 0.96, 1.47, 1.97$. The plot shows that implied volatilities are observed in strings and there are more observations on the strings with small maturities than on the strings with larger maturities. The surface is obtained with Nadaraya-Watson kernel estimator.

For a more detailed inspection three-dimensional plots can be rotated by using a pointing device such as a mouse (with the left mouse-button pressed) or by using the keyboards arrow-keys. Figure 21.12 shows the same plot as before – it has just been rotated by some degrees. Now, one can see implied volatilities "smiles" and "smirks" and recognize different curvature for different maturities. For further research it would be helpful to know which data point belongs to which string. Here the XQC's display offers a feature to show
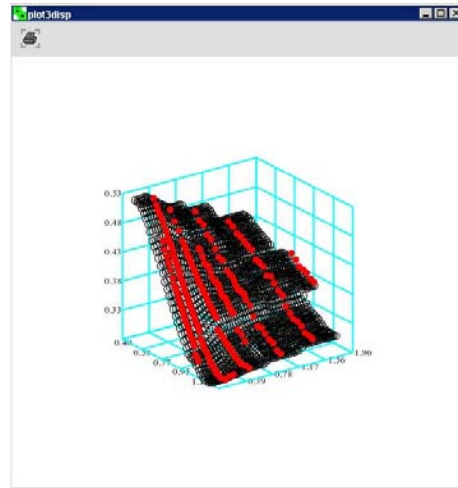
Figure 21.11: Plot of the implied volatility surface from January 4, 1999

the point's coordinates. This feature can be accessed via the display's context menu. 'Showing coordinates' is not the only option. The user could also switch between the three dimensions - 'Show X$\tilde{Y}$', 'Show X$\tilde{Z}$' and 'Show Y$\tilde{Z}$'.

After the 'Showing coordinates' has been chosen all it needs is to point the mouse arrow on a certain data point in order to get the information.

The possibility to configure the XploRe Quantlet Client for special purposes as well as its platform independence are features that recommends itself for the integration into HTML and PDF contents for visualizing statistical and mathematical coherences as already shown in this e-book.
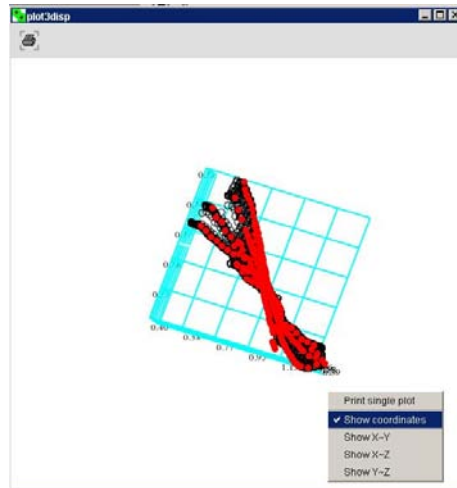
Figure 21.12: Rotating scatter plot showing the context menu.
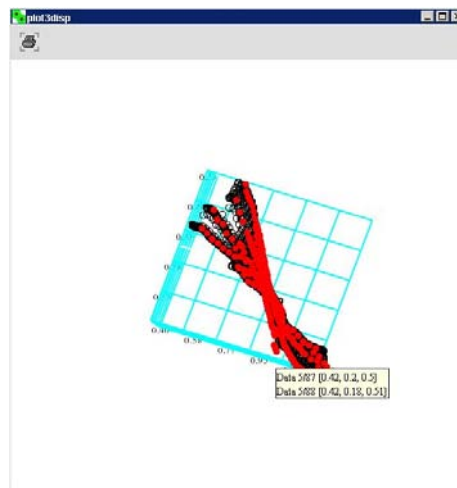
Figure 21.13: Showing the coordinates of a data point.

Value-at-Risk Calculations with Time Varying Copulae

Enzo Giacomini, Wolfgang Härdle
Center for Applied Statistics and Economics
Institut für Statistik und Ökonometrie
Humboldt-Universität zu Berlin
Spandauer Str. 1, D-10178 Berlin
Germany
E-mail: giacomini@wiwi.hu-berlin.de

Value-at-Risk (VaR) of a portfolio is determined by the multivariate distribution of the risk factors increments. This distribution can be modelled through copulae, where the copulae parameters are not necessarily constant over time. For an exchange rate portfolio, copulae with time varying parameters are estimated and the VaR simulated accordingly. Backtesting underlines the improved performance of time varying copulae.

Value-at-Risk and Copulae

At time t a linear portfolio composed of d positions $w = (w_1, \ldots, w_d)^\top$ on assets with prices $S_t = (S_{1,t}, \ldots, S_{d,t})^\top$ and log prices $Z_t = \ln S_t$, has value

$$(1) \qquad V_t = \sum_{j=1}^{d} w_j e^{Z_{j,t}}$$

The *profit and loss* (P&L) function is defined as $L_{t+1} = (V_{t+1} - V_t)$. Defining $X_{t+1} = (Z_{t+1} - Z_t)$ as the time increment in the risk factors from period t to $t+1$, the P&L can be expressed as:

$$(2) \qquad L_{t+1} = \sum_{j=1}^{d} w_j S_{j,t} (e^{X_{j,t+1}} - 1)$$

The *Value-at-Risk* (VaR) is calculated as the α-quantile from F_L, the distribution of L:

$$(3) \qquad VaR = F_L^{-1}(\alpha)$$

The 1-dimensional distribution F_L depends on the d-dimensional distribution F_X. Using *copulae*, the marginal distributions F_{X_j} from each univariate increment can be *separately* modelled from their dependence structure and then *coupled* together to form the multivariate distribution F_X.

In the following, the dependence parameter $\hat{\theta}$ and the joint distribution $\hat{F}_X$ from a sample $\{X_t\}_{t=1}^{T}$ of log returns from exchange rate positions are estimated with copulae. A Monte Carlo simulation based on $\hat{F}_X$ generates different P&L samples. The quantiles at different levels from these simulation samples are then used as estimators for the Value-at-Risk of differem portfolio.

Computing Value-at-Risk with Copulae

A copula is a d-dimensional distribution function $C : [0,1]^d \to [0,1]$ with uniform marginals on the interval $[0,1]$. As in Nelsen (1998), multivariate distributions can be modelled via:

Theorem 1 (Sklar's theorem) *Let F be a d-dimensional distribution function with marginals $F_{X_1} \ldots, F_{X_d}$. Then there exists a copula C with*

$$(4) \qquad F(x_1, \ldots, x_d) = C\{F_{X_1}(x_1), \ldots, F_{X_d}(x_d)\}$$

for every $x_1, \ldots, x_d \in \overline{\mathbb{R}}$. If $F_{X_1}, \ldots, F_{X_d}$ are continuous, then C is unique. On the other hand, if C is a copula and $F_{X_1}, \ldots, F_{X_d}$ are distribution functions, then the function F defined in (4) is a joint distribution function with marginals $F_{X_1}, \ldots, F_{X_d}$.

The estimation of the Value-at-Risk, based on an *i.i.d.* sample $\{X_t\}_{t=1}^T$ is implemented in the following procedure:

1. specification of marginal distributions $F_{X_j}(x_j)$

2. specification of copula $C(u_1, \ldots, u_d; \theta)$

3. fitting the copula C to obtain $\hat{\theta}$

4. generation of Monte Carlo data $X_{T+1} \sim C(u_1, \ldots, u_d; \hat{\theta})$

5. generation of a sample of portfolio losses $L_{T+1}(X_{T+1})$

6. estimation of $\widehat{VaR}_{T+1}$, the empirical quantile at level α from $L_{T+1}(X_{T+1})$.

For copulae belonging to a parametric family $C = \{C_\theta, \theta \in \Theta\}$ and univariate marginals $F_{X_j}(x_j; \delta_j)$, the density of X is given by:

$$f(x_1, \ldots, x_d; \delta_1, \ldots, \delta_d, \theta) = c\{F_{X_1}(x_1; \delta_1), \ldots, F_{X_d}(x_d; \delta_d); \theta\} \prod_{j=1}^d f_j(x_j; \delta_j)$$

where
$$c(u_1, \ldots, u_d) = \frac{\partial^d C(u_1, \ldots, u_d)}{\partial u_1 \ldots \partial u_d}$$

In the IFM (*inference for margins*) method, the log-likelihood function for each of the marginal distributions

$$(5) \qquad \ell_j(\delta_j) = \sum_{t=1}^T \ln f_i(x_{j,t}; \delta_j), j = 1, \ldots, d$$

is maximized to obtain estimates $(\hat{\delta}_1, \ldots, \hat{\delta}_d)^\top$. The function

$$(6) \qquad \ell(\theta, \hat{\delta}_1, \ldots, \hat{\delta}_d) = \sum_{t=1}^T [\ln c\{F_{X_1}(x_{1,t}; \hat{\delta}_1), \ldots, F_{X_d}(x_{d,t}; \hat{\delta}_d); \theta\}]$$

is then maximized over θ to get the dependence parameter estimate $\hat{\theta}$. The estimates $\hat{\theta}_{IFM} = (\hat{\delta}_1, \ldots, \hat{\delta}_d, \hat{\theta})^\top$ solve

$$(\partial \ell_1 / \partial \delta_1, \ldots, \partial \ell_d / \partial \delta_d, \partial \ell / \partial \theta) = 0$$

Backtesting

This procedure is applied to a daily exchange rate portfolio (DEM/USD and GBP/USD from 01.12.1979 to 01.04.1994) with $T = 250$. The univariate risk factor increments (log returns) are assumed to be Gaussian distributed with parameters estimated from the data. The selected copulae belong to the bivariate one-parametric Gumbel family:

$$(7) \qquad C(u, v) = \exp(-[(\ln u)^\theta + (\ln v)^\theta]^{\theta^{-1}}), 1 \le \theta \le \infty$$

To evaluate the performance of the copula in the VaR calculations, different portfolio compositions are used to generate P&L samples. The quantiles from the samples at four levels $\alpha_1 = 0.05$, $\alpha_2 = 0.01$, $\alpha_3 = 0.005$ and $\alpha_4 = 0.001$ are used as estimators for VaR.

The estimated VaR is compared with the realization of the P&L function, an *exceedance* occurring for each P&L value smaller than the estimated VaR. The ratio of the number of exceedances to the number of observations gives the empirical level $\hat{\alpha}$.

Figure 1: Value-at-Risk at levels $\alpha_1 = 0.05$ (yellow), $\alpha_2 = 0.01$ (green), $\alpha_3 = 0.005$ (red), and $\alpha_4 = 0.001$ (blue), P&L (black), estimated at each time from a Monte Carlo sample of 10.000 P&L values generated with Gumbel copula, $w = (2, 1)^\top$.

Figure 2: Value-at-Risk at level $\alpha = 0.05$ (yellow), P&L (black) and exceedances (red), $\hat{\alpha} = 0.0573$, $w = (2, 1)^\top$. P&L samples generated with Gumbel copula.

Table 1: *Gumbel copula, empirical levels $\hat{\alpha}$ for different portfolios.*

Portfolio $w^{\top}$	level $\alpha(\times 10^2)$			
	5	1	0.5	0.1
	empirical level $\hat{\alpha}(\times 10^2)$			
$(1, 1)$	6.05	2.45	1.75	0.83
$(1, 2)$	6.34	2.74	1.75	1.00
$(2, 1)$	5.73	2.24	1.58	0.69
$(2, 3)$	6.22	2.56	1.75	0.92
$(3, 2)$	5.99	2.30	1.55	0.74
$(-1, 2)$	1.64	0.37	0.20	0.11
$(1, -2)$	2.01	0.51	0.43	0.11
$(-2, 1)$	4.44	1.49	0.95	0.40
$(2, -1)$	4.09	1.35	1.09	0.49

REFERENCES

Durrelman, V., Nikeghbali, A. and Roncalli, T. (2000). Which Copula is the Right One ?, *Groupe de Recherche Opérationnelle Crédit Lyonnais.*

Embrechts, P., McNeil, A. and Straumann, D. (1999). *Correlation and Dependence in Risk Management: Properties and Pitfalls Correlation, Risk Management: Value at Risk and Beyond*, Cambridge University Press, Cambridge.

Franke, J., Härdle, W. and Hafner, C. (2004). *Statistics of Financial Market*s, Springer-Verlag, Heidelberg.

Giacomini, E. and Härdle, W. (2005). Nonparametric Risk Management with Adaptive Copulae, *Discussion Paper, SFB 649.*

Härdle, W., Herwartz, H. and Spokoiny, V. (2003). Time Inhomogeneous Multiple Volatility Modeling, *Journal of Financial Econometrics*, 2003 1 (1): 55-95.

Härdle, W., Kleinow, T. and Stahl, G. (2002). *Applied Quantitative Finance*, Springer-Verlag, Heidelberg.

Joe, H. (1997). *Multivariate Models and Dependence Concepts*, Chapman & Hall, London.

Nelsen, R. (1998). *An Introduction to Copulas*, Springer-Verlag, New York.

RÉSUMÉ

La Value at Risk (VaR) d'un portefeuille est déterminée par la distribution multivariée des incréments des facteurs de risques. Cette distribution peut être modélisée par des copules dont les paramètres ne sont pas nécessairement constants par rapport au temps. Pour un portefeuille de taux de change, des copules dépendant du temps sont estimés et la VaR est ainsi simulée. Le backtesting confirme l'amélioration apportée par les copules dépendant du temps.

Statistical Tools in Finance and Insurance (2005)

5 Nonparametric Productivity Analysis

Wolfgang Härdle, Seok-Oh Jeong

5.1 Introduction

Productivity analysis is an important part of classical microeconomic theory. Farrel (1957) studied the problem of measuring efficiency in an empirical way. Shephard (1970) formed the foundation of the modern theory of cost and frontier function. Färe, Grosskopf, and Lovell (1985) and Färe, Grosskopf, and Lovell (1994) provide a comprehensive survey of the problem of measuring efficiency and the theoretical formulations of frontier function.

The activity of production units such as banks, universities, governments and hospitals may be described and formalized by *the production set* of feasible input-output points (x, y):

$$\Psi = \{(x, y) \in \mathbb{R}_+^{p+q} \mid x \text{ can produce } y\}.$$

In many cases y is one-dimensional, so Ψ can be characterized by a function g called the *frontier function* or the *production function*:

$$\Psi = \{(x, y) \in \mathbb{R}_+^{p+1} \mid y \leq g(x)\}.$$

The *input (requirement) set* $X(y)$ is defined by:

$$X(y) = \{x \in \mathbb{R}_+^p \mid (x, y) \in \Psi\},$$

which is the set of all input vectors $x \in \mathbb{R}_+^p$ that yield at least the output vector y. The *output (correspondence) set* $Y(x)$ defined by:

$$Y(x) = \{y \in \mathbb{R}_+^q \mid (x, y) \in \Psi\},$$

Statistical Tools in Finance and Insurance (2005)

which is the set of all output vectors $y \in \mathbb{R}_+^q$ that is obtainable from the input vector x.

In productivity analysis one is interested in the input and output *isoquants* or *efficient boundaries*, denoted by $\partial X(y)$ and $\partial Y(x)$ respectively. They consist of the minimal attainable boundary in a radial sense: for $y \neq 0$,

$$\partial X(y) = \{x \mid x \in X(y),\ \theta x \notin X(y), 0 < \theta < 1\}$$

and for $Y(x) \neq \{0\}$,

$$\partial Y(x) = \{y \mid y \in Y(x),\ \lambda y \notin X(y), \lambda > 1\}.$$

For $y = 0$ we have $\partial X(y) = \{0\}$, and for $Y(x) = \{0\}$ $\partial Y(x) = \{0\}$.

Returns to scale is a characteristic of the surface of the production set. The production set exhibits *constant returns to scale* (CRS) if, for $\alpha \geq 0$ and $P \in \Psi$, $\alpha P \in \Psi$; it exhibits *non-increasing returns to scale* (NIRS) if, for $0 \leq \alpha \leq 1$ and $P \in \Psi$, $\alpha P \in \Psi$; it exhibits *non-decreasing returns to scale* (NDRS) if, for $\alpha \geq 1$ and $P \in \Psi$, $\alpha P \in \Psi$. Especially, a convex production set exhibits non-increasing returns to scale, but note that the converse is not true.

Recall that, given a production set Ψ with the scalar output y, the production function g can also be defined for $x \in \mathbb{R}_+^p$:

$$g(x) = \sup\{y \mid (x, y) \in \Psi\}.$$

It may be defined via the input set and the output set as well:

$$g(x) = \sup\{y \mid x \in X(y)\} = \sup\{y \mid y \in Y(x)\}.$$

We can interpret returns to scale in terms of the production function g:

- Constant returns to scale: For all $t \geq 0$, $g(tx) = tg(x)$.

- Non-decreasing returns to scale: For all $t \geq 1$, $g(tx) \geq tg(x)$

- Non-increasing returns to scale: For all $t \geq 1$, $g(tx) \leq tg(x)$.

How can one evaluate the performance of a given pair of input and output? When the outputs are scalars, one may do it with input x and output y by computing $g(x) - y$ or $y/g(x)$. This is usually inadequate though when multiple inputs or outputs are incorporated. In order to tackle this problem *radial*

Statistical Tools in Finance and Insurance (2005)

efficiency measures have been proposed. For a given input-output point (x_0, y_0), its input efficiency is defined as

$$\theta^{\mathrm{IN}}(x_0, y_0) = \inf\{\theta \mid \theta x_0 \in X(y_0)\}.$$

The efficient level of input corresponding to the output level y_0 is then given by

$$x^{\partial}(y_0) = \theta^{\mathrm{IN}}(x_0, y_0) x_0.$$

Note that $x^{\partial}(y_0)$ is the intersection of $\partial X(y_0)$ and the ray θx_0, $\theta > 0$. The output efficiency score $\theta^{\mathrm{OUT}}(x_0, y_0)$ can be defined similarly:

$$\theta^{\mathrm{OUT}}(x_0, y_0) = \sup\{\theta \mid \theta y_0 \in Y(x_0)\}.$$

The efficient level of output corresponding to the input level x_0 is given by

$$y^{\partial}(x_0) = \theta^{\mathrm{OUT}}(x_0, y_0) y_0.$$

5.2 Nonparametric Hull Methods

5.2.1 An Overview

The production set Ψ and the production function g is usually unknown, but typically a sample of production units or decision making units (DMU's) is available instead:

$$\mathcal{X} = \{(x_i, y_i), i = 1, \ldots, n\}.$$

The problem of productivity analysis is to estimate Ψ or g from the data $\mathcal{X}$.

The most popular nonparametric method is based on data envelopment analysis (DEA), which is an extension of Farrel (1957)'s idea and was popularized in a modern operational research fashion by Charnes, Cooper, and Rhodes (1978). Deprins, Simar, and Tulkens (1984), who extended the DEA to the more flexible class of production sets, introduced the free disposal hull (FDH) model dropping convexity assumption for the production set.

Statistical properties of these hull methods such as consistency and minimax rate of convergence have been studied in the literature. Park (2001) and Simar and Wilson (1999) provide insightful reviews on the statistical inference of existing nonparametric frontier models.

Statistical Tools in Finance and Insurance (2005)

5.2.2   Data Envelopment Analysis

The data envelopment analysis (DEA) of the observed sample $\mathcal{X}$ is defined as the smallest free disposable and convex set containing $\mathcal{X}$:

$$\widehat{\Psi}_{\text{DEA}} \quad = \quad \{(x,y) \in \mathbb{R}_+^{p+q} \,|\, x \geq \sum_{i=1}^{n} \gamma_i x_i, \ y \leq \sum_{i=1}^{n} \gamma_i y_i$$

$$\text{for some } (\gamma_1, \ldots, \gamma_n) \text{ such that } \sum_{i=1}^{n} \gamma_i = 1, \gamma_i \geq 0 \ \forall i\},$$

where the inequalities between vectors are understood componentwise. A set Ψ is *free disposable*, if for given $(x,y) \in \Psi$ all (x',y') with $x' \geq x$ and $y' \leq y$ belong to Ψ. Free disposability is generally assumed in economic modeling. The efficiency scores for a given input-output level (x_0, y_0) are estimated by :

$$\widehat{\theta}^{\text{IN}}(x_0, y_0) \quad = \quad \min\{\theta > 0 \,|\, (\theta x_0, y_0) \in \widehat{\Psi}_{\text{DEA}}\};$$

$$\widehat{\theta}^{\text{OUT}}(x_0, y_0) \quad = \quad \max\{\theta > 0 \,|\, (x_0, \theta y_0) \in \widehat{\Psi}_{\text{DEA}}\}.$$

The efficient levels for a given level (x_0, y_0) are estimated by

$$\widehat{x^\partial}(y_0) = \widehat{\theta}^{\text{IN}}(x_0, y_0) x_0; \quad \widehat{y^\partial}(x_0) = \widehat{\theta}^{\text{OUT}}(x_0, y_0) y_0.$$

The consistency and the convergence rate of this radial efficiency score with multidimensional inputs and outputs were established analytically by Kneip, Park, and Simar (1998). For $p = 1$ and $q = 1$, Gijbels, Mammen, Park, and Simar (1999) obtained its limit distribution depending on the curvature of the frontier and the density at boundary. Jeong and Park (2002) extended this result to higher dimension.

5.2.3   Free Disposal Hull

The free disposal hull (FDH) of the observed sample $\mathcal{X}$ is defined as the smallest free disposable set containing $\mathcal{X}$:

$$\widehat{\Psi}_{\text{FDH}} = \{(x,y) \in \mathbb{R}_+^{p+q} \,|\, x \geq x_i, \ y \leq y_i, \ i = 1, \ldots, n\}.$$

One define the estimators of efficiency scores by FDH in a similar way to DEA. By a straightforward calculation we obtain their closed forms for a given input-

Statistical Tools in Finance and Insurance (2005)

output level (x_0, y_0) :

$$\widehat{\theta}^{\text{IN}}(x_0, y_0) = \min_{i:y \leq y_i} \max_{1 \leq j \leq p} \frac{x_i^j}{x_0^j};$$

$$\widehat{\theta}^{\text{OUT}}(x_0, y_0) = \max_{i:x \geq x_i} \min_{1 \leq k \leq q} \frac{y_i^k}{y_0^k},$$

where v^j is the jth component of a vector v. The efficient levels for a given level (x_0, y_0) are estimated by the same way as those for DEA. Park, Simar, and Weiner (1999) showed that the limit distribution of the FDH estimator with multivariate setup is a Weibull distribution depending upon the slope of the frontier and the density at boundary.

5.3 DEA in Practice : Insurance Agencies

The quantlet **deahull** for DEA has the following syntax :

```
dea = deahull (X, Y)
      computes the input and output efficiency score for each DMU
      based on DEA
```

with input variables: **X** is a $n \times p$ matrix, p dimensional inputs of n DMU's, and **Y** is $n \times q$ matrix, q dimensional outputs of n DMU's.

The list **dea** contains the output variables : **dea.effscore** is a n dimensional vector containing the input efficiency scores, $\widehat{\theta}^{\text{IN}}$, of n DMU's, and **dea.efflevel** is $n \times p$ matrix containing the estimates of efficient level of inputs, $\widehat{x^\partial}$, of n DMU's.

To illustrate how DEA works, consider an example from the empirical study by Scheel (1999). He studied the efficiency of 63 agencies of a German insurance company by the DEA method. The input $(X \in \mathbb{R}_+^4)$ and output $(Y \in \mathbb{R}_+^2)$ variables were as follows:

X_1 : Number of clients of Type A

X_2 : Number of clients of Type B

Statistical Tools in Finance and Insurance (2005)

Table 5.1: Summary Statistics for 63 agencies of a German insurance company

	Minimum	Maximum	Mean	Median	Std.Error
X1	42	572	225.54	197	131.73
X2	55	481	184.44	141	110.28
X3	0	140	19.762	10	26.012
X4	73756	693820	258670	206170	160150
Y1	2	70	22.762	16	16.608
Y2	696	33075	7886.7	6038	7208

X_3 : Number of clients of Type C

X_4 : Potential new premiums in EURO

Y_1 : Number of new contracts

Y_2 : Sum of new premiums in EURO

The two output variables are typical for insurance agencies. Summary statistics for this data are given in Table 5.1. The included XploRe code reads the data **agency.dat**, creates the input variables for the quantlet **deahull**, runs the quantlet, and lists the efficiency and the efficient level of inputs for each agency as a result, see Table 5.2 and Table 5.3.

The input efficient scores **effscore** are useful as a benchmark in comparing the 63 DMU's, the insurance agencies. The list of efficient level of inputs provides with a 'goal' inputs for each agency to be efficient.

5.4   FDH in Practice : Manufacturing Industry

The quantlet **fdhull** for FDH has the following syntax:

```
eff = fdhull (X, Y)
      computes the input and output efficiency score for each DMU
      based on FDH
```

Härdle, W. and Jeong, S. (2005): Nonparametric Productivity Analysis

Statistical Tools in Finance and Insurance (2005)

Table 5.2: Efficiency score of the 63 DMU's insurance Agencies

	Effeciency score
1	0.38392
2	0.49063
3	0.86449
.	.
.	.
.	.
62	0.79892
63	1

■STFagency.xpl

Table 5.3: Efficiency level of the 63 DMU's insurance Agencies

		efflevel		
1	52.981	92.909	3.8392	108960
2	81.444	60.838	2.4531	76895
3	131.4	72.617	2.5935	96070
.	.	.	.	.
.	.	.	.	.
.	.	.	.	.
62	66.311	87.083	1.5978	111710
63	108	257	0	299910

■STFagency.xpl

with input variables: X is $n \times p$ matrix. p dimensional inputs of n DMU's, and Y is $n \times q$ matrix, q dimensional outputs of n DMU's.

The arbitrary name **eff** is used to indicate the output variable: **eff** which is the $n \times 2$ matrix containing the input and output efficiency scores of n DMU's.

In order to illustrate how this quantlet works the Manufacturing Industry Productivity Database from the National Bureau of Economic Research, USA is considered. This database is available in the internet [http://www.nber.org]

Statistical Tools in Finance and Insurance (2005)

Table 5.4: Summary Statistics for Manufacturing Industry Productivity Database (NBER, USA)

	Minimum	Maximum	Mean	Median	Std.Error
X1	0.8	500.5	37.833	21	54.929
X2	18.5	145130	4313	1957.2	10771
X3	0.5	3807.8	139.96	49.7	362
X4	15.8	64590	2962.8	1234.7	6271.1
Y	34.1	56311	3820.2	1858.5	6392

with a description of the database. It contains annual industry-level data on output, employment, payroll and other input costs, investment, capital stocks, and various industry-specific price indexes from 1958 on hundreds of manufacturing industries (indexed by 4 digits numbers) in the United States. We chose the data from 1996 (458 DMU's), for example, with 4 input variables ($p = 4$) and 1 output variable ($q = 1$) along with the study of Park, Simar, and Weiner (1999) :

X_1 : Total employment

X_2 : Total cost of materials

X_3 : Cost of electric and fuels

X_4 : Total real capital stock

Y : Total value added

Summary statistics for this data are given in Table 5.4. The included XploRe code reads the MS-Excel data file **nber96.csv**, creates the vector of the ID for DMU's and the input variables for the quantlet **fdhull**, runs the quantlet, and lists the efficiency scores, $\hat{\theta}^{IN}(\cdot, \cdot)$ and $1/\hat{\theta}^{OUT}(\cdot, \cdot)$, with ID of industries.

¿From Table 5.5, we see that the DMU indexed by 2015 is efficient in both input and output oriented aspect. This means the DMU is one of the vertices of the free disposal hull generated by the 458 observations. On the other hand, the DMU 2298 seems to be good in input efficiency but poor in output efficiency.

With these efficiency scores we can estimate the efficient level of input (or output) by multiplying (or dividing) the score to the corresponding observation.

Statistical Tools in Finance and Insurance (2005)

Table 5.5: FDH efficiency scores of the 458 observations

fdh			
1	2011	0.93803	0.97451
2	2013	0.64321	0.70255
3	2015	1	1
.	.	.	.
.	.	.	.
.	.	.	.
75	2298	0.95588	0.47239
.	.	.	.
.	.	.	.
.	.	.	.
458	3999	0.98961	0.94792

■STFnber96.xpl

For example, DMU 2013 had inputs $X_1 = 88.1$, $X_2 = 14925$, $X_3 = 250$, $X_4 = 4365.1$, and output $Y = 5954.2$ with input efficiency 0.64321 and with output efficiency 0.70255. Then the efficient level of input for this firm is given by $X_1 = 56.667$, $X_2 = 9600$, $X_3 = 160.8$, and $X_4 = 2807.7$. And in the aspect of output efficiency this firm should have increased their output by $Y = 4183.1$ with the observed level of inputs to be considered as output technically efficient.

Statistical Tools in Finance and Insurance (2005)

Bibliography

Charnes, A., Cooper, W. W, and Rhodes, E. (1978). Measuring the Inefficiency of Decision Making Units, *European Journal of Operational Research* **2**, 429 444.

Deprins, D., Simar, L. and Tulkens, H. (1984). Measuring Labor Inefficiency in Post Offices, in *The Performance of Public Enterprizes: Concepts and Measurements*, 243–267.

Färe, R., Grosskopf, S. and Lovell, C. A. K. (1985). *The Measurement of Efficiency of Production*. Kluwer-Nijhoff.

Färe, R., Grosskopf, S. and Lovell, C. A. K. (1994). *Production Frontiers*, Cambridge University Press.

Farrell, M. J. (1957). The Measurement of Productivity Efficiency, *Journal of the Royal Statistical Society, Ser. A* **120**, 253–281.

Gijbels, I., Mammen. E., Park, B. U. and Simar, L. (1999). On Estimation of Monotone and Concave Frontier Functions. *Journal of the American Statistical Association* **94**, 220 228.

Jeong, S. and Park, B. U. (2002). Limit Distributions Convex Hull Estimators of Boundaries. submitted.

Kneip, A., Park, B. U. and Simar, L. (1998). A Note on the Convergence of Nonparametric DEA Efficiency Measures. *Econometric Theory* **14**, 783–793.

Park, B. U. (2001). On Nonparametric Estimation of Data Edges, *Journal of the Korean Statistical Society* **30**, 2, 265–280.

Park, B. U., Simar, L. and Weiner, Ch. (1999). The FDH Estimator for Productivity Efficiency Scores : Asymptotic Properties, *Econometric Theory* **16**, 855–877.

Statistical Tools in Finance and Insurance (2005)

Scheel, H. (1999). Continuity of the BCC efficiency measure, in: Westermann (ed.), *Data Envelopment Analysis in the public and private service sector*, Gabler, Wiesbaden.

Shephard, R. W. (1970). *Theory of Cost and Production Function*, Princeton University Press.

Simar. L. and Wilson, P. W. (1999). Statistical Inference in Nonparametric Frontier Models : The State of the Art, Discussion Paper # 9904, Institute de Statistique, Université Catholique de Louvain, Louvain-la-Neuve, Belgium.

Integrable e-lements for Statistics Education*

Wolfgang Härdle, Sigbert Klinke and Uwe Ziegenhagen

Humboldt-Universität zu Berlin
Center for Applied Statistics and Economics

Abstract

Without doubt modern education in statistics must involve practical, computer-based data analysis but the question arises whether and how computational elements should be integrated into the canon of methodological education. Should the student see and study high-level programming code right at the beginning of his or her studies? Which technology can be presented during class and which computational elements can re-occur (at increasing level of complexity) during the different courses?

In this paper we address these questions and discuss where e-techniques have their limits in statistics education.

Keywords   electronic books, hypertext, e-supported teaching, statistical software

JEL codes   I21, C19,

1 Introduction

Since the entering of modern computing equipment into schools and universities there have been increased efforts to use computers not only for research and numerical computations

---

but also for the education of students. While traditional textbooks on statistics are usually restricted to small examples, computers offer great opportunities to enrich the teaching of statistics by the means of explaining animations or on-the-fly computations of large real-world datasets.

But each new technology does not only hold opportunities and advantages there may also be hazards or risks. And a sentence stated by John Tukey in 1965 should be taken into consideration: "Each new generation of computers offers us new possiblities, at a time when we are far from using most of the possibilities offered by those already obsolete."

We will outline our thoughts about the integration of electronic and computational elements into statistics along the courses tought at the School for Business and Economics of Humboldt-Universität zu Berlin.

Students following different course programs are taught at the Institute for Statistics and Econometrics at Humboldt-Universität zu Berlin, German Business Administration and Economics students as well as students from international bachelor or master classes or the math department.

After finishing the bachelor level, where statistics is taught in a two-term lecture three hours a week, the students who choose statistics as one main subject are required to take a course on multivariate data analysis, that enables them to understand the basics of probability theory and to analyze high-dimensional data by the means of cluster-, principal component- and factor analysis.

On the basis of this lecture the student has further options: While *Non- and Semiparametric Modelling* focuses on nonparametric density estimation and regression the student learns in the computational statistics courses the applied data analysis with SPSS and real world data. Supplementary to the *Computerbased Statistics* courses are the *XploRe Introductory Course* and the *Numerics Introductory Course* focussing on the practical work with XploRe respectively the numerical details of selected algorithms.

Statistics of financial markets is the third specialisation taught at Humboldt-Universität, covered are theoretical and practical aspects of option pricing, risk-management and time series modelling.
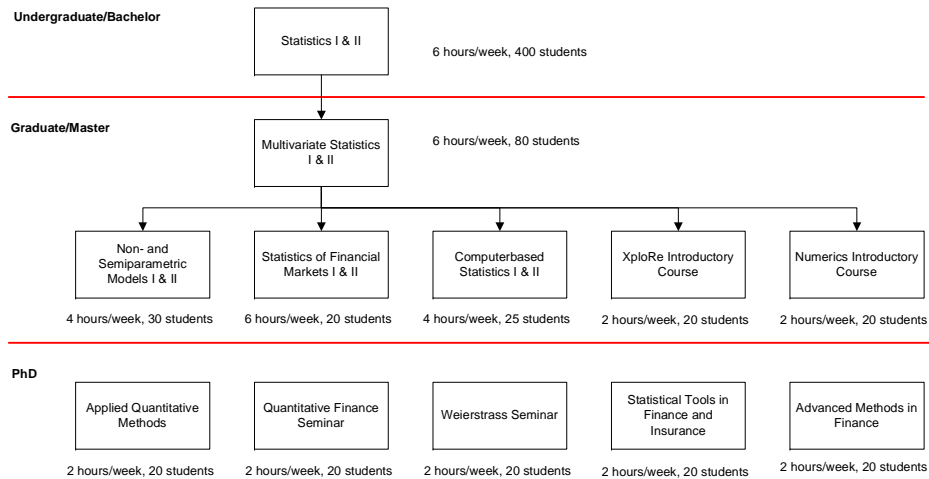
Figure 1: Typical one year cycle of statistics courses in Berlin

At the master level there are currently five courses also aimed at PhD students: In *Applied Quantitative Methods* and the *Weierstrass Seminar* recent developments of mathematical statistics and data analysis are discussed, *Advanced Methods in Finance*, the *Quantitative Finance Seminar* and *Statistical Tools in Finance and Insurance* deal with the field of financial mathematics.

2 Traditional and Modern Teaching Material

The main components of MM*Stat (Müller, M., Rönz, B., Ziegenhagen, U., 2000), that was developed to support the undergraduate teaching of statistics, are *lecture units*, *additional information* and *(non)interactive examples*, presented in a style similar to filing cards. This structure, called *MD*Booklet*, has also been used for *Numerical Methods in Statistics* and *Finance Introductory Course* (http://www.quantlet.com/mdstat/products.html). The lecture units contain common and well-known topics as basic concepts of statistics, basics of probability calculus and sampling theory.

Each "lecture" filing card provides the basic concepts of the methods considered together with definitions, important formulas and graphics. "Fully explained", "Enhanced" and "Interactive" examples show how these formulas can be applied. The first explain standard classbook examples in detail while the latter discuss important aspects of more advanced examples. The interactive examples are one of the main features of MM*Stat. By using a combination of embedded Java-based *XploRe Quantlet Client* (Borak, S.,
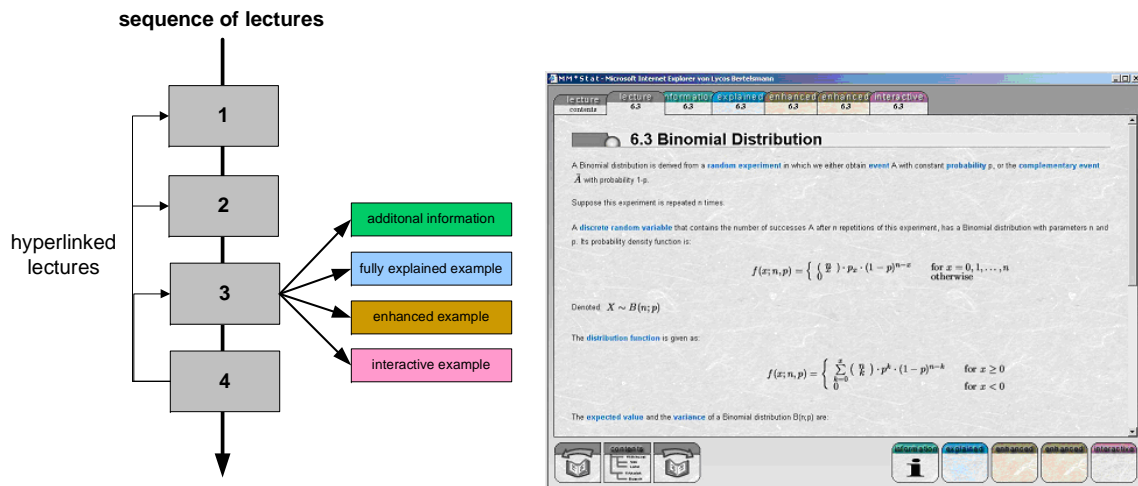
Figure 2: MM*Stat: Layer Architecture and Screenshot

Härdle, W., Lehmann, H., 2005) and a remote or locally installed *XploRe Quantlet Server* the student can compute distribution functions, histograms and test results. Compared with classical textbooks where effects of parameter changes can only be presented as sequence of graphics or tables, the user can experience the results of different settings in realtime.

The repetition of concepts and methods introduced earlier has been implemented in two different ways: To see or recall the definition of a used term, a glossary has been written and important words within the different lectures inside MM*Stat are linked to this glossary. Furthermore the student can complete multiple choice questions by selecting radio buttons, the results are given instantly.
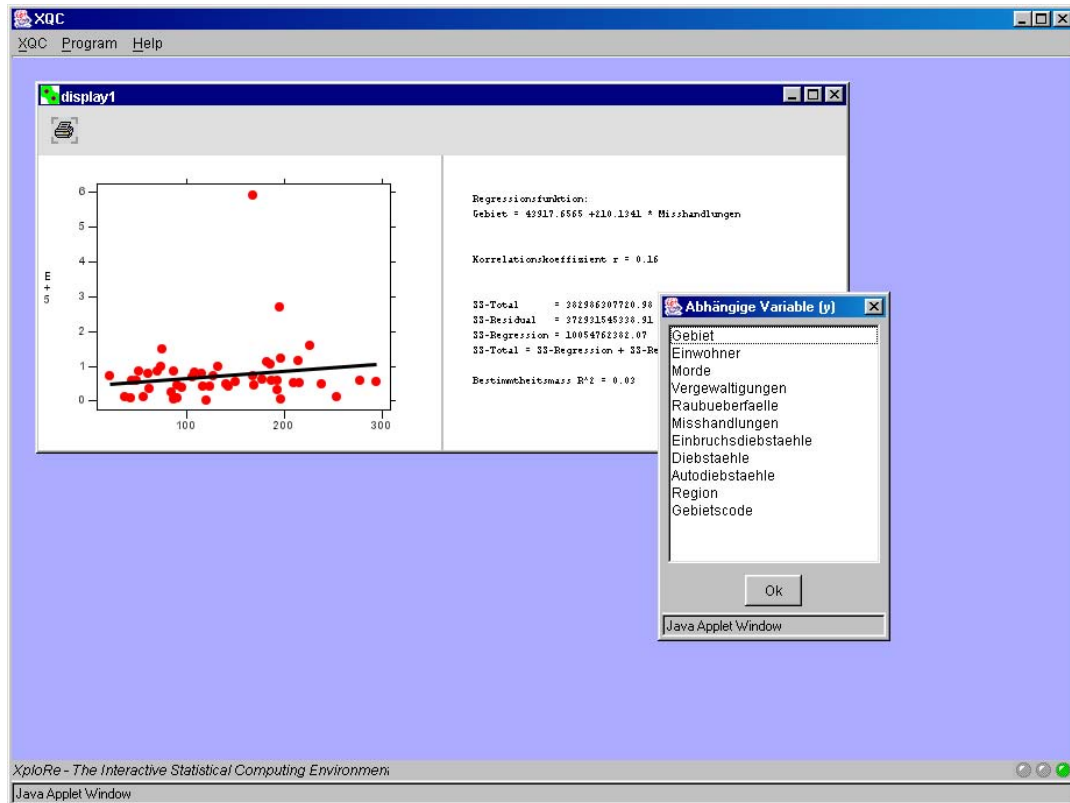
Figure 3: Example for linear regression in MM*Stat

The target group of MM*Stat, the undergraduate students, have different levels of literacy concerning computers, while some are firm in common office software tools others have little experience. So it was decided to hide the source code of the interactive examples from the user, to lower the inhibition threshold of using the software. Figure 3 shows a typical example for linear regression, which allows the user to choose the dependent and independent variable and returns the regression plot and the parameters of the model.

The combination of XploRe Quantlet Server and Client has not only been used in MM*Stat, furthermore it is used in the DoSS@d system (Mori, Y., Yamamoto, Y. and Yadohisa, H., 2003) and the books published by the members of the authors' institute (http://www.xplore-stat.de/ebooks/ebooks.html) also use this framework, although in a different way.

As mentioned the source code for the interactive examples was hidden from the user in MM*Stat, from the target group of the advanced books, graduate and PhD students on the one, researchers and professional on the other hand, we assume a certain level of computational literacy. In each book respectively the corresponding slides selected pages contain links to HTML-pages that contain the source code of the example as well as links to two different implementations of this example.

These two implementations differ in that sense that the 'edit' page calls the XploRe Quantlet Client in editor mode, which means that the user can manipulate the source

5
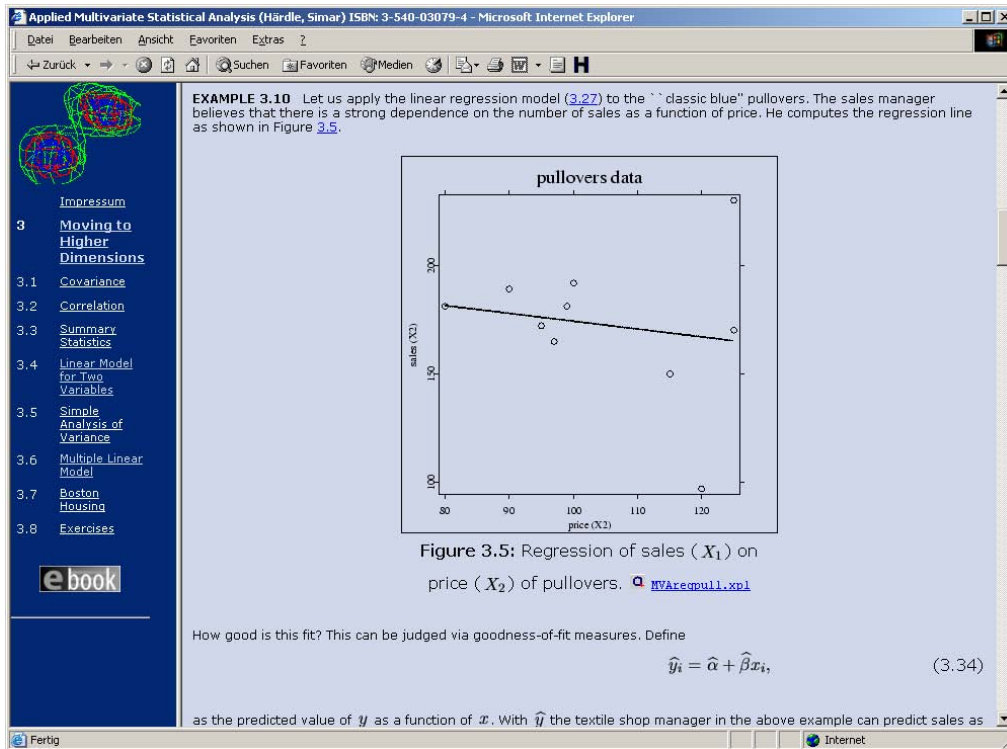
Figure 4: Applied Multivariate Statistics: HTML page with link to an example

code of the example. From MM*Stat they vary from the textual component. While at undergraduate level regression is taught using a notation with sums, the more abstract matrix-based approach is used for graduate students.

Figure 5 depicts two screenshort from the book *Applied Multivariate Analysis* by W. Härdle and L. Simar, on the left hand side the 'run' version of the quantlet, on the right the editable 'edit' version.

3 XploRe, Xplore Quantlet Server and Yxilon

Besides SPSS which is used for the computational statistics courses the software package XploRe is used for educating students at Humboldt-Universität. Developed in cooperation with MD*Tech (http://www.mdtech.de), XploRe is a full-featured statistical programming language. Using a matrix-oriented approach in combination with a C-style syntax a great variety of diverse statistical problems can be handled conveniently.

6
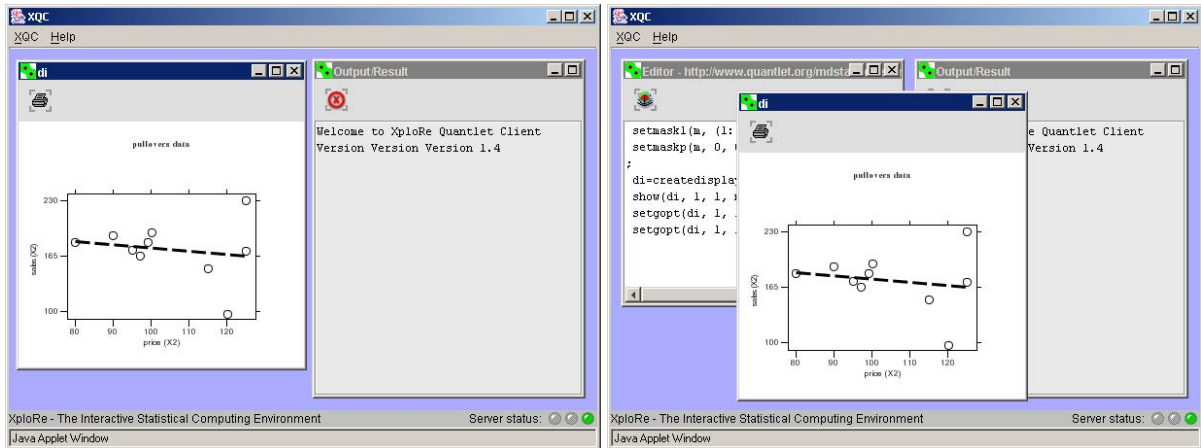
Figure 5: Applied Multivariate Statistics: *execute* and *edit* versions of an example

The implementation of XploRe incorporates ideas from the requirements of statistical software mentioned by Chambers and Lang (1999): usage from multiple front-ends, extensibility on language and native core level, interactive graphics and the inclusion of existing code (`C`, `Fortran`).

With Yxilon (Guril, Y., Klinke, S., Ziegenhagen, U., 2005) MD*Tech and Humboldt-Universität zu Berlin agreed on a new research project with the aim to supply a freely available statistical environment and to allow the implementation of recent developments in software technology while remaining fully compatible with XploRe.

The most dramatic change compared with XploRe or other packages as *R* or *Jasp* (Fujiwara, T., Ikunori K. et al. 2000) is the way the code is executed: Besides the interpretation by a so called runtime environment the idea is to compile directly to C++ and Java. Further changes are:

- published under Free-BSD license

- fully independent from machine architecture by using platform-independent protocols and software frameworks

- improved modularity to allow convenient exchange of single components

- integration of dynamic data sources as databases and webservices

- increased computing performance simplifying the language definitions

- improved integration into standard office and business software

The software architecture of Yxilon is shown in Figure 9. Via a graphical or non-graphical user interface the user accesses the system. All objects and information are held in the object database, that communicates with the other parts as the parser or runtime
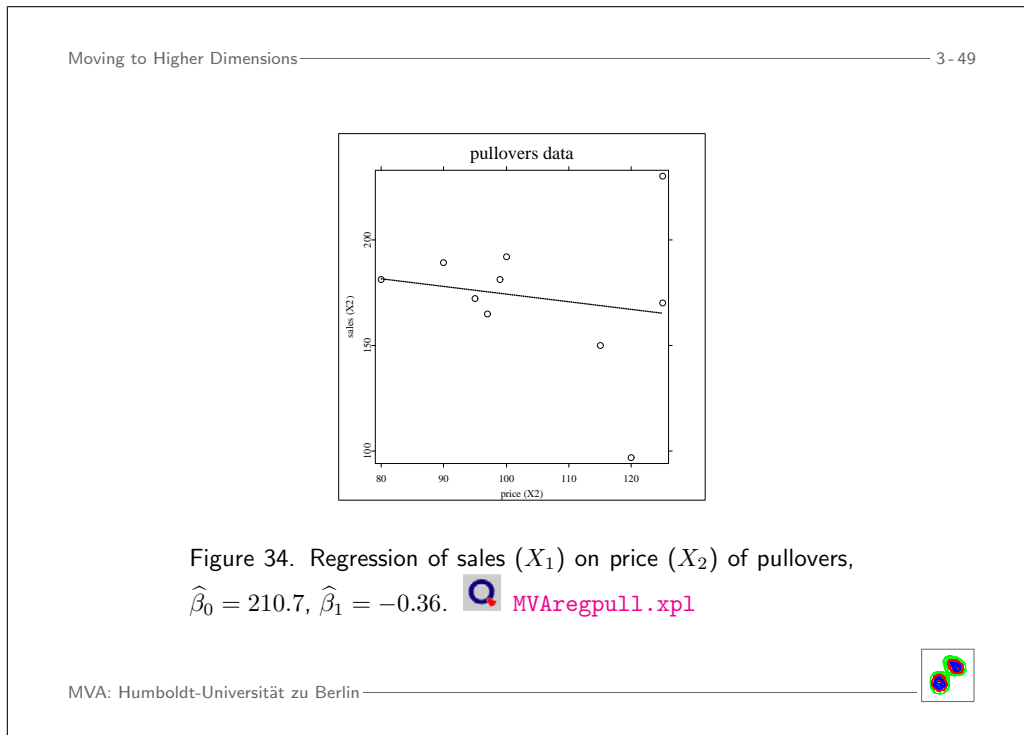
Figure 6: Applied Multivariate Statistics: slide with link to an example

environment. The parser takes the sourcecode provided by the user and converts them to either Java/C++ code or a binary format directly readable by the runtime engine. The advantage of source conversion is that the high-level, interpreted XploRe code, which is slower than binary code, can be compiled by the Java or C++ compiler to fast running machine code. Especially for computing-intensive applications as bootstrapping and simulation we expect significant improvements concerning execution time.
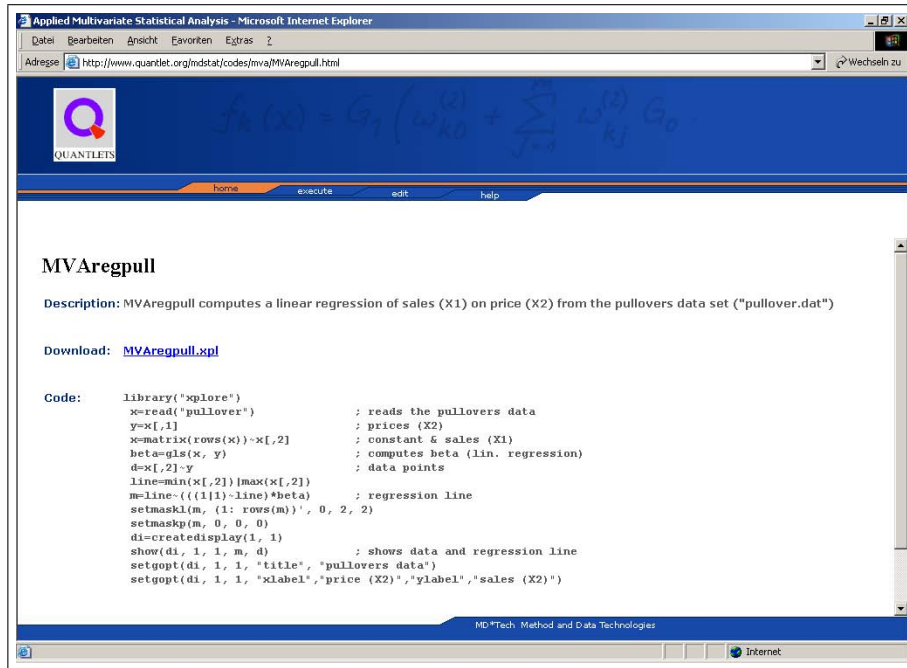
Figure 7: Applied Multivariate Statistics: homepage of the linked example
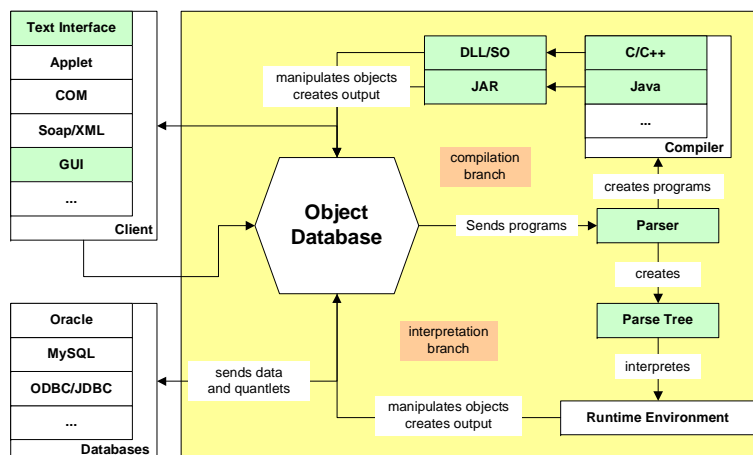
Figure 8: Architecture of the Yxilon Framework (green components in development)
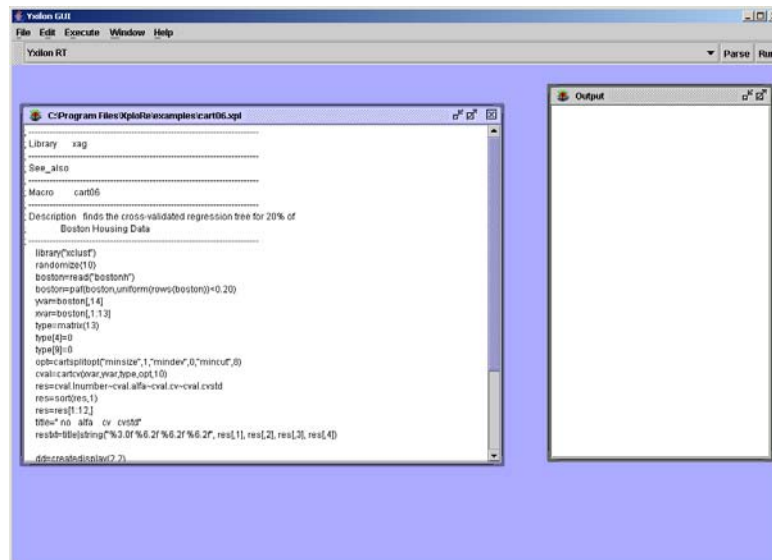
Figure 9: Screenshot of the Yxilon graphical user interface

4 Limits of e-lements in statistics education

Electronic and computational elements in education can be limited from a variety of different reasons as the technical implementation or the educational approach.

MM*Stat relies on a complex framework of HTML, Cascading Style Sheets, Java and JavaScript. While HTML and Java were already well-established standards, JavaScript and Cascading Style Sheets were implemented in different ways for each browser, so extensive work for each major webbrowser had to be invested.

From the educational point of view each use of electronic media should be questioned carefully. An analysis by Brandes, U. (2004) showed no significant improvements in the results of those students who used MM*Stat in comparison to those who used classical textbooks. The educational limits also lie in the way the information is presented. The classical way of using a textbook or blackboards also offers the opportunity to include on-the-fly notes from either the teacher or the student. Electronic teaching solutions either do not offer this capability at all or require significant technical ressources.

References

Borak, S., Härdle, W., Lehmann, H. (2005), Working with the XQC, IN: Statistical Tools for Finance and Insurance, editors: Cizek, P., Härdle, W., Weron, R., Springer Verlag

Brandes, U. (2004). Statistische Bewertung und Analyse der Klausurergebnisse Statistik

(Statistical analysis of exam results in statistics undergraduate courses), diploma thesis, Humboldt-Universität zu Berlin

Chambers, J. and Lang, D. T. (1999). Omegahat – A Component-based Statistical Computing Environment, Proceedings of the 52nd Session of the ISI, Helsinki, Finland.

Fujiwara, T., Ikunori K., Nakano, J., Yoshikazu, Y. (2000). A Statistical Package Based on Pnuts, In: COMPSTAT. Proceedings in Computational Statistics, Physica Verlag

Guril, Y., Klinke, S., Ziegenhagen, U. (2005). Yxilon – a Modular Open-Source Statistical Programming Language, In: Proceedings of the 55th Session of the International Statistical Institute (ISI), Sydney, Australia

Mori, Y., Yamamoto, Y. and Yadohisa, H. (2003). Data-oriented Learning System of Statistics based on Analysis Scenario/Story (DoLStat). Bulletin of the International Statistical Institute (ISI).

Müller, M., Rönz, B., Ziegenhagen, U.(2000). The Multimedia Project MM*Stat for Teaching Statistics, In: COMPSTAT. Proceedings in Computational Statistics, Physica Verlag

10 Predicting Bankruptcy with Support Vector Machines

Wolfgang Härdle, Rouslan Moro, and Dorothea Schäfer

The purpose of this work is to introduce one of the most promising among recently developed statistical techniques – the support vector machine (SVM) – to corporate bankruptcy analysis. An SVM is implemented for analysing such predictors as financial ratios. A method of adapting it to default probability estimation is proposed. A survey of practically applied methods is given. This work shows that support vector machines are capable of extracting useful information from financial data, although extensive data sets are required in order to fully utilize their classification power.

The support vector machine is a classification method that is based on statistical learning theory. It has already been successfully applied to optical character recognition, early medical diagnostics, and text classification. One application where SVMs outperformed other methods is electric load prediction (EUNITE, 2001), another one is optical character recognition (Vapnik, 1995). SVMs produce better classification results than parametric methods and such a popular and widely used nonparametric technique as neural networks, which is deemed to be one of the most accurate. In contrast to the latter they have very attractive properties. They give a single solution characterized by the global minimum of the optimized functional and not multiple solutions associated with the local minima as in the case of neural networks. Moreover, SVMs do not rely so heavily on heuristics, i.e. an arbitrary choice of the model and have a more flexible structure.

10.1   Bankruptcy analysis methodology

Although the early works in bankruptcy analysis were published already in the 19th century (Dev, 1974), statistical techniques were not introduced to it until the publications of Beaver (1966) and Altman (1968). Demand from financial institutions for investment risk estimation stimulated subsequent research. However, despite substantial interest, the accuracy of corporate default predictions was much lower than in the private loan sector, largely due to a small number of corporate bankruptcies.

Meanwhile, the situation in bankruptcy analysis has changed dramatically. Larger data sets with the median number of failing companies exceeding 1000 have become available. 20 years ago the median was around 40 companies and statistically significant inferences could not often be reached. The spread of computer technologies and advances in statistical learning techniques have allowed the identification of more complex data structures. Basic methods are no longer adequate for analysing expanded data sets. A demand for advanced methods of controlling and measuring default risks has rapidly increased in anticipation of the New Basel Capital Accord adoption (BCBS, 2003). The Accord emphasises the importance of risk management and encourages improvements in financial institutions' risk assessment capabilities.

In order to estimate investment risks one needs to evaluate the default probability (PD) for a company. Each company is described by a set of variables (predictors) x, such as financial ratios, and its class y that can be either $y = -1$ ('successful') or $y = 1$ ('bankrupt'). Initially, an unknown classifier function $f : x \to y$ is estimated on a training set of companies (x_i, y_i), $i = 1, ..., n$. The training set represents the data for companies which are known to have survived or gone bankrupt. Finally, f is applied to computing default probabilities (PD) that can be uniquely translated into a company rating.

The importance of financial ratios for company analysis has been known for more than a century. Among the first researchers applying financial ratios for bankruptcy prediction were Ramser (1931), Fitzpatrick (1932) and Winakor and Smith (1935). However, it was not until the publications of Beaver (1966) and Altman (1968) and the introduction of univariate and multivariate discriminant analysis that the systematic application of statistics to bankruptcy analysis began. Altman's linear Z-score model became the standard for a decade to come and is still widely used today due to its simplicity. However, its assumption of equal normal distributions for both failing and successful companies with the same covariance matrix has been justly criticized. This approach was further developed by Deakin (1972) and Altman et al. (1977).

Later on, the center of research shifted towards the logit and probit models. The original works of Martin (1977) and Ohlson (1980) were followed by (Wiginton, 1980), (Zavgren, 1983) and (Zmijewski, 1984). Among other statistical methods applied to bankruptcy analysis there are the gambler's ruin model (Wilcox, 1971), option pricing theory (Merton, 1974), recursive partitioning (Frydman et al., 1985), neural networks (Tam and Kiang, 1992) and rough sets (Dimitras et al., 1999) to name a few.

There are three main types of models used in bankruptcy analysis. The first one is structural or parametric models, e.g. the option pricing model, logit and probit regressions, discriminant analysis. They assume that the relationship between the input and output parameters can be described *a priori*. Besides their fixed structure these models are fully determined by a set of parameters. The solution requires the estimation of these parameters on a training set.

Although structural models provide a very clear interpretation of modelled processes, they have a rigid structure and are not flexible enough to capture information from the data. The non-structural or nonparametric models (e.g. neural networks or genetic algorithms) are more flexible in describing data. They do not impose very strict limitations on the classifier function but usually do not provide a clear interpretation either.

Between the structural and non-structural models lies the class of semiparametric models. These models, like the RiskCalc private company rating model developed by Moody's, are based on an underlying structural model but all or some predictors enter this structural model after a nonparametric transformation. In recent years the area of research has shifted towards non-structural and semi-parametric models since they are more flexible and better suited for practical purposes than purely structural ones.

Statistical models for corporate default prediction are of practical importance. For example, corporate bond ratings published regularly by rating agencies such as Moody's or S&P strictly correspond to company default probabilities estimated to a great extent statistically. Moody's RiskCalc model is basically a probit regression estimation of the cumulative default probability over a number of years using a linear combination of non-parametrically transformed predictors (Falkenstein, 2000). These non-linear transformations f_1, f_2, ..., f_d are estimated on univariate models. As a result, the original probit model:

$$E[y_{i,t}|x_{i,t}] \quad = \quad \Phi\left(\beta_1 x_{i1,t} + \beta_2 x_{i2,t} + ... + \beta_d x_{id,t}\right), \qquad (10.1)$$

is converted into:

$$E[y_{i,t}|x_{i,t}] \quad = \quad \Phi\{\beta_1 f_1(x_{i1,t}) + \beta_2 f_2(x_{i2,t}) + ... + \beta_d f_d(x_{id,t})\}, \quad (10.2)$$

where $y_{i,t}$ is the cumulative default probability within the prediction horizon for company i at time t. Although modifications of traditional methods like probit analysis extend their applicability, it is more desirable to base our methodology on general ideas of statistical learning theory without making many restrictive assumptions.

The ideal classification machine applying a classifying function f from the available set of functions $\mathcal{F}$ is based on the so called expected risk minimization principle. The expected risk

$$R(f) = \int \frac{1}{2} |f(x) - y| \, dP(x, y), \tag{10.3}$$

is estimated under the distribution $P(x, y)$, which is assumed to be known. This is, however, never true in practical applications and the distribution should also be estimated from the training set (x_i, y_i), $i = 1, 2, ..., n$, leading to an ill-posed problem (Tikhonov and Arsenin, 1977).

In most methods applied today in statistical packages this problem is solved by implementing another principle, namely the principle of the empirical risk minimization, i.e. risk minimization over the training set of companies, even when the training set is not representative. The empirical risk defined as:

$$\hat{R}(f) = \frac{1}{n} \sum_{i=1}^{n} \frac{1}{2} |f(x_i) - y_i|, \tag{10.4}$$

is nothing else but an average value of loss over the training set, while the expected risk is the expected value of loss under the true probability measure. The loss for i.i.d. observations is given by:

$$\frac{1}{2} |f(x) - y| = \left\{ \begin{array}{ll} 0, & \text{if classification is correct,} \\ 1, & \text{if classification is wrong.} \end{array} \right.$$

The solutions to the problems of expected and empirical risk minimization:

$$f_{opt} \quad = \quad \arg \min_{f \in \mathcal{F}} R(f), \tag{10.5}$$

$$\hat{f}_n \quad = \quad \arg \min_{f \in \mathcal{F}} \hat{R}(f), \tag{10.6}$$
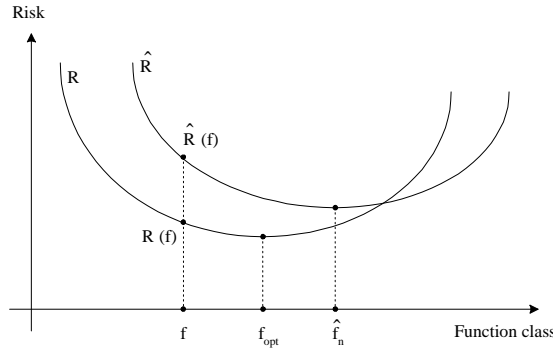
Figure 10.1: The minima f_{opt} and $\hat{f}_n$ of the expected (R) and empirical $(\hat{R})$ risk functions generally do not coincide.

generally do not coincide (Figure 10.1), although they converge to each other as $n \to \infty$ if $\mathcal{F}$ is not too large.

We cannot minimize expected risk directly since the distribution $P(x, y)$ is unknown. However, according to statistical learning theory (Vapnik, 1995), it is possible to estimate the Vapnik-Chervonenkis (VC) bound that holds with a certain probability $1 - \eta$:

$$R\left(f\right) \leq \hat{R}\left(f\right) + \phi\left(\frac{h}{n}, \frac{\ln(\eta)}{n}\right). \tag{10.7}$$

For a linear indicator function $g(x) = \operatorname{sign}(x^\top w + b)$:

$$\phi\left(\frac{h}{n}, \frac{\ln(\eta)}{n}\right) = \sqrt{\frac{h\left(\ln \frac{2n}{h}\right) - \ln \frac{\eta}{4}}{n}}, \tag{10.8}$$

where h is the VC dimension.

The VC dimension of the function set $\mathcal{F}$ in a d-dimensional space is h if some function $f \in \mathcal{F}$ can shatter h objects $\left\{x_i \in \mathbb{R}^d, i = 1, ..., h\right\}$, in all 2^h possible configurations and no set $\left\{x_j \in \mathbb{R}^d, j = 1, ..., q\right\}$, exists where $q > h$ that satisfies this property. For example, three points on a plane $(d = 2)$ can be shattered by linear indicator functions in $2^h = 2^3 = 8$ ways, whereas 4 points cannot be
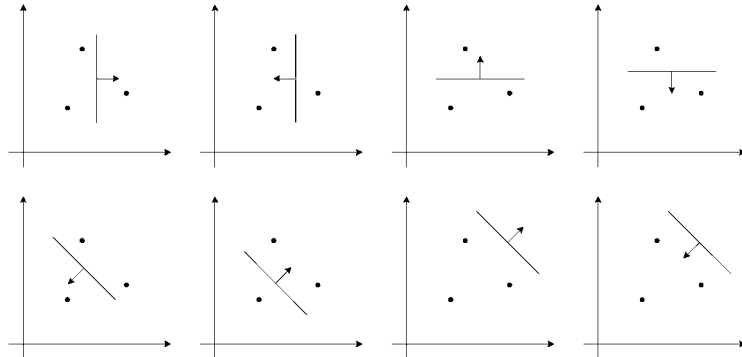
Figure 10.2: Eight possible ways of shattering 3 points on the plane with a linear indicator function.

shattered in $2^q = 2^4 = 16$ ways. Thus, the VC dimension of the set of linear indicator functions in a two-dimensional space is three, see Figure 10.2.

The expression for the VC bound (10.7) is a regularized functional where the VC dimension h is a parameter controlling complexity of the classifier function. The term $\phi\left(\frac{h}{n}, \frac{\ln(\eta)}{n}\right)$ introduces a penalty for the excessive complexity of a classifier function. There is a trade-off between the number of classification errors on the training set and the complexity of the classifier function. If the complexity were not controlled, it would be possible to find such a classifier function that would make no classification errors on the training set notwithstanding how low its generalization ability would be.

10.2   Importance of risk classification in practice

In most countries only a small percentage of firms has been rated to date. The lack of rated firms is mainly due to two factors. Firstly, an external rating is an extremely costly procedure. Secondly, until the recent past most banks decided on their loans to small and medium sized firms (SME) without asking for the client's rating figure or applying an own rating procedure to estimate the client's default risk. At best, banks based their decision on rough scoring models. At worst, the credit decision was completely left to the loan officer.

Table 10.1: Rating grades and risk premia. Source: (Damodaran, 2002) and (Füser, 2002)

Rating Class (S&P)	One year PD (%)	Risk Premia (%)
AAA	0.01	0.75
AA	0.02 − 0.04	1.00
A+	0.05	1.50
A	0.08	1.80
A-	0.11	2.00
BBB	0.15 − 0.40	2.25
BB	0.65 − 1.95	3.50
B+	3.20	4.75
B	7.00	6.50
B-	13.00	8.00
CCC	> 13	10.00
CC		11.50
C		12.70
D		14.00

Since learning to know its own risk is costly and, until recently, the lending procedure of banks failed to set the right incentives, small and medium sized firms shied away from rating. However, the regulations are about to change the environment for borrowing and lending decisions. With the implementation of the New Basel Capital Accord (Basel II) scheduled for the end of 2006 not only firms that issue debt securities on the market are in need of rating but also any ordinary firm that applies for a bank loan. If no external rating is available, banks have to employ an internal rating system and deduce each client's specific risk class. Moreover, Basel II puts pressure on firms and banks from two sides.

First, banks have to demand risk premia in accordance to the specific borrower's default probability. Table 10.1 presents an example of how individual risk classes map into risk premiums (Damodaran, 2002) and (Füser, 2002). For small US-firms a one-year default probability of 0.11% results in a spread of 2%. Of course, the mapping used by lenders will be different if the firm type or the country in which the bank is located changes. However, in any case future loan pricing has to follow the basic rule. The higher the firm's default risk is the more risk premium the bank has to charge.

Table 10.2: Rating grades and capital requirements. Source: (Damodaran, 2002) and (Füser, 2002). The figures in the last column were estimated by the authors for a loan to an SME with a turnover of 5 million euros with a maturity of 2.5 years using the data from column 2 and the recommendations of the Basel Committee on Banking Supervision (BCBS, 2003).

Rating Class (S&P)	One-year PD (%)	Capital Requirements (%) (Basel I)	Capital Requirements (%) (Basel II)
AAA	0.01	8.00	0.63
AA	0.02 − 0.04	8.00	0.93 − 1.40
A+	0.05	8.00	1.60
A	0.08	8.00	2.12
A-	0.11	8.00	2.55
BBB	0.15 − 0.40	8.00	3.05 − 5.17
BB	0.65 − 1.95	8.00	6.50 − 9.97
B+	3.20	8.00	11.90
B	7.00	8.00	16.70
B-	13.00	8.00	22.89
CCC	> 13	8.00	> 22.89
CC		8.00	
C		8.00	
D		8.00	

Second, Basel II requires banks to hold client-specific equity buffers. The magnitudes of these buffers are determined by a risk weight function defined by the Basel Committee and a solvability coefficient (8%). The function maps default probabilities into risk weights. Table 10.2 illustrates the change in the capital requirements per unit of a loan induced by switching from Basel I to Basel II. Apart from basic risk determinants such as default probability (PD), maturity and loss given default (LGD) the risk weights depend also on the type of the loan (retail loan, loan to an SME, mortgages, etc.) and the annual turnover. Table 10.2 refers to an SME loan and assumes that the borrower's annual turnover is 5 million EUR (BCBS, 2003). Since the lock-in of the bank's equity affects the provision costs of the loan, it is likely that these costs will be handed over directly to an individual borrower.

Basel II will affect any firm that is in need for external finance. As both the risk premium and the credit costs are determined by the default risk, the firms' rating will have a deeper economic impact on banks as well as on firms themselves than ever before. Thus in the wake of Basel II the choice of the right rating method is of crucial importance. To avoid friction of a large magnitude the employed method must meet certain conditions. On the one hand, the rating procedure must keep the amount of misclassifications as low as possible. On the other, it must be as simple as possible and, if employed by the borrower, also provide some guidance to him on how to improve his own rating.

SVMs have the potential to satisfy both demands. First, the procedure is easy to implement so that any firm could generate its own rating information. Second, the method is suitable for estimating a unique default probability for each firm. Third, the rating estimation done by an SVM is transparent and does not depend on heuristics or expert judgements. This property implies objectivity and a high degree of robustness against user changes. Moreover, an appropriately trained SVM enables the firm to detect the specific impact of all rating determinants on the overall classification. This property would enable the firm to find out prior to negotiations what drawbacks it has and how to overcome its problems. Overall, SVMs employed in the internal rating systems of banks will improve the transparency and accuracy of the system. Both improvements may help firms and banks to adapt to the Basel II framework more easily.

10.3 Lagrangian formulation of the SVM

Having introduced some elements of statistical learning and demonstrated the potential of SVMs for company rating we can now give a Lagrangian formulation of an SVM for the linear classification problem and generalize this approach to a nonlinear case.

In the linear case the following inequalities hold for all n points of the training set:

$$x_i^\top w + b \geq 1 - \xi_i \quad \text{for} \quad y_i = 1,$$
$$x_i^\top w + b \leq -1 + \xi_i \quad \text{for} \quad y_i = -1,$$
$$\xi_i \geq 0,$$
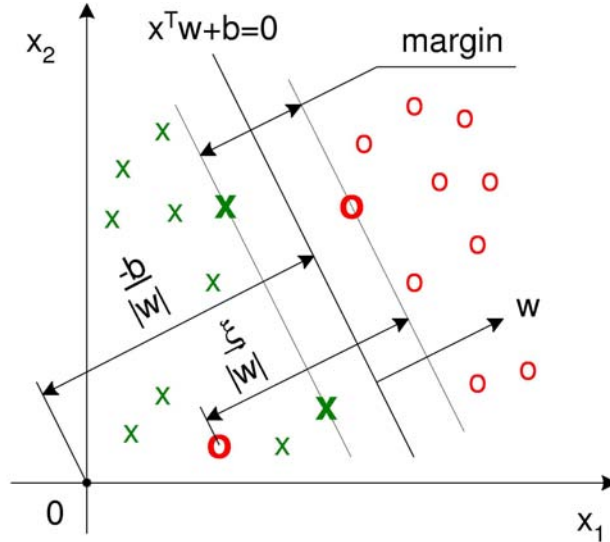
Figure 10.3: The separating hyperplane $x^\top w + b = 0$ and the margin in a non-separable case.

which can be combined into two constraints:

$$y_i(x_i^\top w + b) \geq 1 - \xi_i \tag{10.9}$$

$$\xi_i \geq 0. \tag{10.10}$$

The basic idea of the SVM classification is to find such a separating hyperplane that corresponds to the largest possible margin between the points of different classes, see Figure 10.3. Some penalty for misclassification must also be introduced. The classification error ξ_i is related to the distance from a misclassified point x_i to the canonical hyperplane bounding its class. If $\xi_i > 0$, an error in separating the two sets occurs. The objective function corresponding to penalized margin maximization is formulated as:

$$\frac{1}{2}\left\| w \right\|^2 + C\left(\sum_{i=1}^{n}\xi_i\right)^v, \tag{10.11}$$

where the parameter C characterizes the generalization ability of the machine and $\upsilon \geq 1$ is a positive integer controlling the sensitivity of the machine to outliers. The conditional minimization of the objective function with constraint (10.9) and (10.10) provides the highest possible margin in the case when classification errors are inevitable due to the linearity of the separating hyperplane. Under such a formulation the problem is convex. One can show that margin maximization reduces the VC dimension.

The Lagrange functional for the primal problem for $\upsilon = 1$ is:

$$L_P = \frac{1}{2}\left\| w \right\|^2 + C \sum_{i=1}^{n} \xi_i - \sum_{i=1}^{n} \alpha_i \{y_i \left(x_i^\top w + b\right) - 1 + \xi_i\} - \sum_{i=1}^{n} \mu_i \xi_i, \ (10.12)$$

where $\alpha_i \geq 0$ and $\mu_i \geq 0$ are Lagrange multipliers. The primal problem is formulated as:

$$\min_{w_k,b,\xi_i} \max_{\alpha_i} L_P.$$

After substituting the Karush-Kuhn-Tucker conditions (Gale et al., 1951) into the primal Lagrangian, we derive the dual Lagrangian as:

$$L_D = \sum_{i=1}^{n} \alpha_i - \frac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{n} \alpha_i \alpha_j y_i y_j x_i^\top x_j, \qquad\qquad (10.13)$$

and the dual problem is posed as:

$$\max_{\alpha_i} L_D,$$

subject to:

$$0 \leq \alpha_i \leq C,$$
$$\sum_{i=1}^{n} \alpha_i y_i = 0.$$

Those points i for which the equation $y_i(x_i^\top w + b) \leq 1$ holds are called support vectors. After training the support vector machine and deriving Lagrange multipliers (they are equal to 0 for non-support vectors) one can classify a company described by the vector of parameters x using the classification rule:

$$g(x) = \mathrm{sign}\left(x^\top w + b\right), \qquad\qquad (10.14)$$

where $w = \sum_{i=1}^{n} \alpha_i y_i x_i$ and $b = \frac{1}{2}\left(x_{+1} + x_{-1}\right)w$. x_{+1} and x_{-1} are two support vectors belonging to different classes for which $y(x^\top w + b) = 1$. The value of the classification function (the score of a company) can be computed as

$$f(x) = x^\top w + b. \tag{10.15}$$

Each value of $f(x)$ uniquely corresponds to a default probability (PD).

The SVMs can also be easily generalized to the nonlinear case. It is worth noting that all the training vectors appear in the dual Lagrangian formulation only as scalar products. This means that we can apply kernels to transform all the data into a high dimensional Hilbert feature space and use linear algorithms there:

$$\Psi : \mathbb{R}^d \mapsto \mathbb{H}. \tag{10.16}$$

If a kernel function K exists such that $K(x_i, x_j) = \Psi(x_i)^\top \Psi(x_j)$, then it can be used without knowing the transformation Ψ explicitly. A necessary and sufficient condition for a symmetric function $K(x_i, x_j)$ to be a kernel is given by Mercer's (1909) theorem. It requires positive definiteness, i.e. for any data set $x_1, ..., x_n$ and any real numbers $\lambda_1, ..., \lambda_n$ the function K must satisfy

$$\sum_{i=1}^{n}\sum_{j=1}^{n} \lambda_i \lambda_j K(x_i, x_j) \geq 0. \tag{10.17}$$

Some examples of kernel functions are:

- $K(x_i, x_j) = e^{-\|x_i - x_j\|/2\sigma^2}$ – the isotropic Gaussian kernel;

- $K(x_i, x_j) = e^{-(x_i - x_j)^\top r^{-2}\Sigma^{-1}(x_i - x_j)/2}$ – the stationary Gaussian kernel with an anisotropic radial basis; we will apply this kernel in our study taking Σ equal to the variance matrix of the training set; r is a constant;

- $K(x_i, x_j) = (x_i^\top x_j + 1)^P$ – the polynomial kernel;

- $K(x_i, x_j) = \tanh(kx_i^\top x_j - \delta)$ – the hyperbolic tangent kernel.

10.4   Description of data

For our study we selected the largest bankrupt companies with the capitalization of no less than \$1 billion that filed for protection against creditors under

Chapter 11 of the US Bankruptcy Code in 2001–2002 after the stock marked crash of 2000. We excluded a few companies due to incomplete data, leaving us with 42 companies. They were matched with 42 surviving companies with the closest capitalizations and the same US industry classification codes available through the Division of Corporate Finance of the Securities and Exchange Commission (SEC, 2004).

From the selected 84 companies 28 belonged to various manufacturing industries, 20 to telecom and IT industries, 8 to energy industries, 4 to retail industries, 6 to air transportation industries, 6 to miscellaneous service industries, 6 to food production and processing industries and 6 to construction and construction material industries. For each company the following information was collected from the annual reports for 1998–1999, i.e. 3 years prior to defaults of bankrupt companies (SEC, 2004): (i) S – sales; (ii) COGS – cost of goods sold; (iii) EBIT – earnings before interest and taxes, in most cases equal to the operating income; (iv) Int – interest payments; (v) NI – net income (loss); (vi) Cash – cash and cash equivalents; (vii) Inv – inventories; (viii) CA – current assets; (ix) TA – total assets; (x) CL – current liabilities; (xi) STD – current maturities of the long-term debt; (xii) TD – total debt; (xiii) TL – total liabilities; (xiv) Bankr – bankruptcy (1 if a company went bankrupt, -1 otherwise).

The information about the industry was summarized in the following dummy variables: (i) Indprod – manufacturing industries; (ii) Indtelc – telecom and IT industries; (iii) Indenerg – energy industries; (iv) Indret – retail industries; (v) Indair – air transportation industries; (vi) Indserv – miscellaneous service industries; (vii) Indfood – food production and processing industries; (viii) Indconst – construction and construction material industries.

Based on these financial indicators the following four groups of financial ratios were constructed and used in our study: (i) profit measures: EBIT/TA, NI/TA, EBIT/S; (ii) leverage ratios: EBIT/Int, TD/TA, TL/TA; (iii) liquidity ratios: QA/CL, Cash/TA, WC/TA, CA/CL and STD/TD, where QA is quick assets and WC is working capital; (iv) activity or turnover ratios: S/TA, Inv/COGS.

10.5   Computational results

The most significant predictors suggested by the discriminant analysis belong to profit and leverage ratios. To demonstrate the ability of an SVM to extract information from the data, we will chose two ratios from these groups: NI/TA from the profitability ratios and TL/TA from the leverage ratios. The SVMs,

Table 10.3: Descriptive statistics for the companies. All data except SIZE = $\log{(\mathrm{TA})}$ and ratios are given in billions of dollars.

Variable	Min	Max	Mean	Std. Dev.
TA	0.367	91.072	8.122	13.602
CA	0.051	10.324	1.657	1.887
CL	0.000	17.209	1.599	2.562
TL	0.115	36.437	4.880	6.537
CASH	0.000	1.714	0.192	0.333
INVENT	0.000	7.101	0.533	1.114
LTD	0.000	13.128	1.826	2.516
STD	0.000	5.015	0.198	0.641
SALES	0.036	37.120	5.016	7.141
COGS	0.028	26.381	3.486	4.771
EBIT	-2.214	29.128	0.822	3.346
INT	-0.137	0.966	0.144	0.185
NI	-2.022	4.013	0.161	0.628
EBIT/TA	-0.493	1.157	0.072	0.002
NI/TA	-0.599	0.186	-0.003	0.110
EBIT/S	-2.464	36.186	0.435	3.978
EBIT/INT	-16.897	486.945	15.094	68.968
TD/TA	0.000	1.123	0.338	0.236
TL/TA	0.270	1.463	0.706	0.214
SIZE	12.813	18.327	15.070	1.257
QA/CL	-4.003	259.814	4.209	28.433
CASH/TA	0.000	0.203	0.034	0.041
WC/TA	-0.258	0.540	0.093	0.132
CA/CL	0.041	2001.963	25.729	219.568
STD/TD	0.000	0.874	0.082	0.129
S/TA	0.002	5.559	1.008	0.914
INV/COGS	0.000	252.687	3.253	27.555

besides their Lagrangian formulation, can differ in two aspects: (i) their capacity that is controlled by the coefficient C in (10.12) and (ii) the complexity of classifier functions controlled in our case by the anisotropic radial basis in the Gaussian kernel transformation.

Triangles and squares in Figures 10.4–10.7 represent successful and failing companies from the training set, respectively. The intensity of the gray background corresponds to different score values f. The darker the area, the higher the score and the greater is the probability of default. Most successful companies lying in the bright area have positive profitability and a reasonable leverage TL/TA of around 0.4, which makes economic sense.

Figure 10.4 presents the classification results for an SVM using locally near linear classifier functions (the anisotropic radial basis is $100\Sigma^{1/2}$) with the capacity fixed at $C = 1$. The discriminating rule in this case can be approximated by a linear combination of predictors and is similar to that suggested by discriminant analysis, although the coefficients of the predictors may be different.

If the complexity of classifying functions increases (the radial basis goes down to $2\Sigma^{1/2}$) as illustrated in Figure 10.5, we get a more detailed picture. Now the areas of successful and failing companies become localized. If the radial basis is decreased further down to $0.5\Sigma^{1/2}$ (Figure 10.6), the SVM will try to track each observation. The complexity in this case is too high for the given data set.

Figure 10.7 demonstrates the effects of high capacities ($C = 300$) on the classification results. As capacity is growing, the SVM localizes only one cluster of successful companies. The area outside this cluster is associated with approximately equally high score values.

Thus, besides estimating the scores for companies the SVM also managed to learn that there always exists a cluster of successful companies, while the cluster for bankrupt companies vanishes when the capacity is high, i.e. a company must possess certain characteristics in order to be successful and failing companies can be located elsewhere. This result was obtained without using any additional knowledge besides that contained in the training set.

The calibration of the model or estimation of the mapping $f \rightarrow$ PD can be illustrated by the following example (the SVM with the radial basis $2\Sigma^{1/2}$ and capacity $C = 1$ will be applied). We can set three rating grades: safe, neutral and risky which correspond to the values of the score $f < -0.0115$, $-0.0115 < f < 0.0115$ and $f > 0.0115$, respectively, and calculate the total number of companies and the number of failing companies in each of the three groups. If the training set were representative of the whole population of companies, the ratio of failing to all companies in a group would give the estimated probability of default. Figure 10.8 shows the power (Lorenz) curve (Lorenz, 1905) – the cumulative default rate as a function of the percentile

Figure 10.4: Ratings of companies in two dimensions. The case of a low complexity of classifier functions, the radial basis is $100\Sigma^{1/2}$, the capacity is fixed at $C = 1$.

STFsvm01.xpl

of companies sorted according to their score – for the training set of companies. For the abovementioned three rating grades we derive $\text{PD}_{\text{safe}} = 0.24$, $\text{PD}_{\text{neutral}} = 0.50$ and $\text{PD}_{\text{risky}} = 0.76$.

If a sufficient number of observations is available, the model can also be calibrated for finer rating grades such as AAA or BB by adjusting the score values separating the groups of companies so that the estimated default probabilities within each group equal to those of the corresponding rating grades. Note, that we are calibrating the model on the grid determined by $\text{grad}(f) = 0$ or $\text{grad}(\hat{\text{PD}}) = 0$ and not on the orthogonal grid as in the Moody's RiskCalc model. In other words, we do not make a restrictive assumption of an independent influence of predictors as in the latter model. This can be important since,

Probability of Default
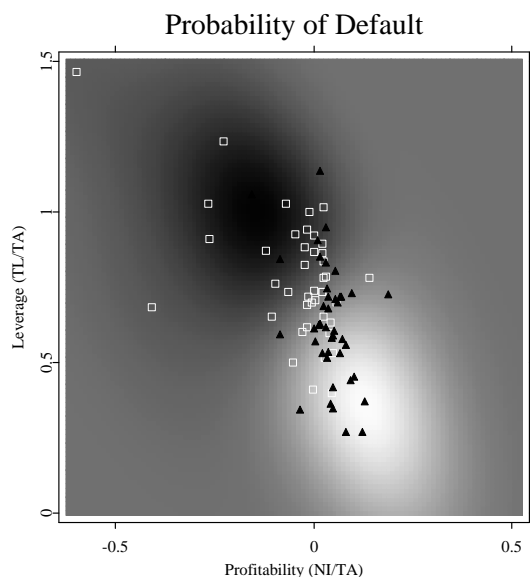
Figure 10.5: Ratings of companies in two dimensions; the case of an average complexity of classifier functions, the radial basis is $2\Sigma^{1/2}$, the capacity is fixed at $C = 1$.

STFsvm02.xpl

for example, the same decrease in profitability will have different consequences for high and low leveraged firms.

For multidimensional classification the results cannot be easily visualized. In this case we will use the cross-validation technique to compute the percentage of correct classifications and compare it with that for the discriminant analysis (DA). Note that both most widely used methods – the discriminant analysis and logit regression – choose only one significant at the 5% level predictor (NI/TA) when forward selection is used. Cross-validation has the following stages. One company is taken out of the sample and the SVM is trained on the remaining companies. Then the class of the out-of-the-sample company is evaluated by the SVM. This procedure is repeated for all the companies and the percentage of correct classifications is calculated.
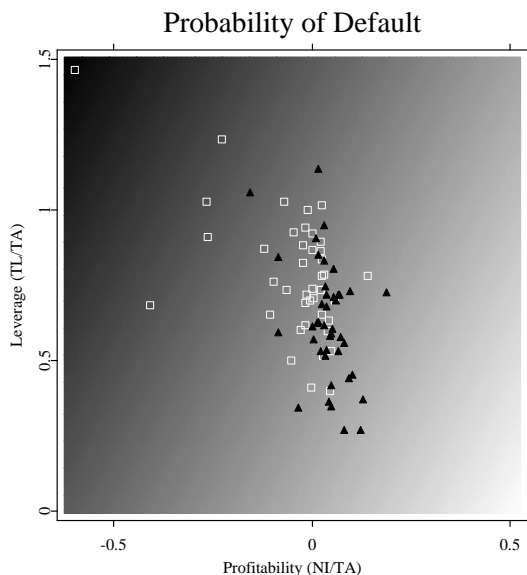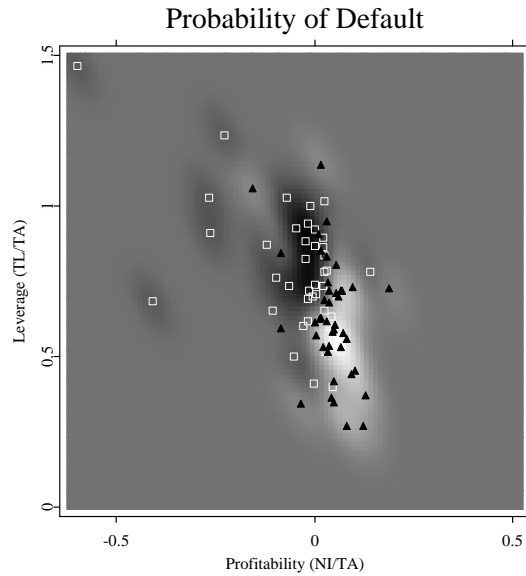
Figure 10.6: Ratings of companies in two dimensions; the case of an excessively high complexity of classifier functions, the radial basis is $0.5\Sigma^{1/2}$, the capacity is fixed at $C = 1$.

STFsvm03.xpl

The best percentage of correctly cross-validated companies (all available ratios were used as predictors) is higher for the SVM than for the discriminant analysis (62% vs. 60%). However, the difference is not significant at the 5% level. This indicates that the linear function might be considered as an optimal classifier for the number of observations in the data set we have. As for the direction vector of the separating hyperplane, it can be estimated differently by the SVM and DA without affecting much the accuracy since the correlation of underlying predictors is high.

Cluster center locations, as they were estimated using cluster analysis, are presented in Table 10.4. The results of the cluster analysis indicate that two clusters are likely to correspond to successful and failing companies. Note the
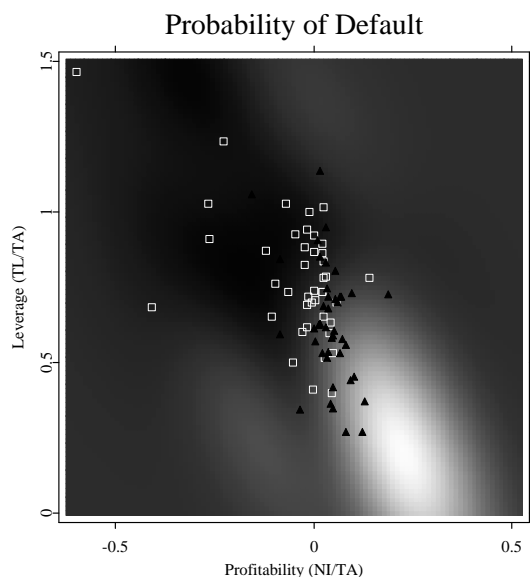
Figure 10.7: Ratings of companies in two dimensions; the case of a high capacity ($C = 300$). The radial basis is fixed at $2\Sigma^{1/2}$.

STFsvm04.xpl

substantial differences in the interest coverage ratios, NI/TA, EBIT/TA and TL/TA between the clusters.

10.6   Conclusions

As we have shown, SVMs are capable of extracting information from real life economic data. Moreover, they give an opportunity to obtain the results not very obvious at first glance. They are easily adjusted with only few parameters. This makes them particularly well suited as an underlying technique for company rating and investment risk assessment methods applied by financial institutions.
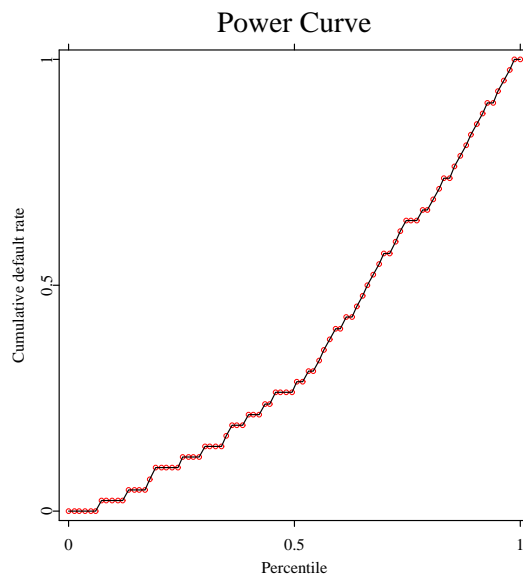
Figure 10.8: Power (Lorenz) curve (Lorenz, 1905) – the cumulative default rate
              as a function of the percentile of companies sorted according to
              their score – for the training set of companies. An SVM is applied
              with the radial basis $2\Sigma^{1/2}$ and capacity $C = 1$.

STFsvm05.xpl

SVMs are also based on very few restrictive assumptions and can reveal effects
overlooked by many other methods. They have been able to produce accurate
classification results in other areas and can become an option of choice for
company rating. However, in order to create a practically valuable methodology
one needs to combine an SVM with an extensive data set of companies and
turn to alternative formulations of SVMs better suited for processing large
data sets. Overall, we have a valuable tool for company rating that can answer
the requirements of the new capital regulations.

Table 10.4: Cluster centre locations. There are 19 members in class {-1} –
successful companies, and 65 members in class {1} – failing com-
panies.

Cluster	{-1}	{1}
EBIT/TA	0.263	0.015
NI/TA	0.078	-0.027
EBIT/S	0.313	-0.040
EBIT/INT	13.223	1.012
TD/TA	0.200	0.379
TL/TA	0.549	0.752
SIZE	15.104	15.059
QA/CL	1.108	1.361
CASH/TA	0.047	0.030
WC/TA	0.126	0.083
CA/CL	1.879	1.813
STD/TD	0.144	0.061
S/TA	1.178	0.959
INV/COGS	0.173	0.155

Bibliography

Altman, E., (1968). Financial Ratios, Discriminant Analysis and the Prediction of Corporate Bankruptcy, *The Journal of Finance*, September: 589-609.

Altman, E., Haldeman, R. and Narayanan, P., (1977). ZETA Analysis: a New Model to Identify Bankruptcy Risk of Corporations, *Journal of Banking and Finance*, June: 29-54.

Basel Committee on Banking Supervision (2003). The New Basel Capital Accord, third consultative paper, *http://www.bis.org/bcbs/cp3full.pdf*.

Beaver, W., (1966). Financial Ratios as Predictors of Failures. Empirical Research in Accounting: Selected Studies, *Journal of Accounting Research*, supplement to vol. 5: 71-111.

Damodaran, A., (2002). *Investment Valuation*, second ed., John Wiley & Sons, New York, NY.

Deakin, E., (1972). A Discriminant Analysis of Predictors of Business Failure, *Journal of Accounting Research*, Spring: 167-179.

Dev, S., (1974). Ratio Analysis and the Prediction of Company Failure in Ebits, Credits, Finance and Profits, ed. H.C. Edy and B.S. Yamey, Sweet and Maxwell, London: 61-74.

Dimitras, A., Slowinski, R., Susmaga, R. and Zopounidis, C., (1999). Business Failure Prediction Using Rough Sets, *European Journal of Operational Research*, number 114: 263-280.

EUNITE, (2001). Electricity load forecast competition of the EUropean Network on Intelligent TEchnologies for Smart Adaptive Systems, *http://neuron.tuke.sk/competition/*.

Falkenstein, E., (2000). *RiskCalc for Private Companies: Moody's Default Model*, Moody's Investors Service.

Fitzpatrick, P., (2000). *A Comparison of the Ratios of Successful Industrial Enterprises with Those of Failed Companies*, The Accounting Publishing Company.

Frydman, H., Altman, E. and Kao, D.-L., (1985). Introducing Recursive Partitioning for Financial Classification: The Case of Financial Distress, *The Journal of Finance*, **40**: 269-291.

Füser, K., (2002). Basel II – was muß der Mittelstand tun?, *http://www.ey.com/global/download.nsf/Germany/Mittelstandsrating/ $file/Mittelstandsrating.pdf.*

Härdle, W. and Simar, L. (2003). *Applied Multivariate Statistical Analysis*, Springer Verlag.

Gale, D., Kuhn, H.W. and Tucker, A.W., (1951). Linear Programming and the Theory of Games, *Activity Analysis of Production and Allocation*, ed. T.C. Koopmans, John Wiley & Sons, New York, NY: 317-329.

Lorenz, M.O., (1905). Methods for Measuring the Concentration of Wealth, *Journal of American Statistical Association*, **9**: 209-219.

Martin, D., (1977). Early Warning of Bank Failure: A Logit Regression Approach, *Journal of Banking and Finance*, number 1: 249-276.

Mercer, J., (1909). Functions of Positive and Negative Type and Their Connection with the Theory of Integral Equations, *Philosophical Transactions of the Royal Society of London*, **209**: 415-446.

Merton, R., (1974). On the Pricing of Corporate Debt: The Risk Structure of Interest Rates, *The Journal of Finance*, **29**: 449-470.

Ohlson, J., (1980). Financial Ratios and the Probabilistic Prediction of Bankruptcy, *Journal of Accounting Research*, Spring: 109-131.

Ramser, J. and Foster, L., (1931). *A Demonstration of Ratio Analysis. Bulletin No. 40*, University of Illinois, Bureau of Business Research, Urbana, Illinois.

Division of Corporate Finance of the Securities and Exchange Commission, (2004). Standard industrial classification (SIC) code list, *http://www.sec.gov/info/edgar/siccodes.htm.*

Securities and Exchange Commission, (2004). Archive of historical documents, *http://www.sec.gov/cgi-bin/srch-edgar.*

Tam, K. and Kiang, M., (1992). Managerial Application of Neural Networks: the Case of Bank Failure Prediction, *Management Science*, **38**: 926-947.

Tikhonov, A.N. and Arsenin, V.Y., (1977). *Solution of Ill-posed Problems*, W.H. Winston, Washington, DC.

Vapnik, V., (1995). *The Nature of Statistical Learning Theory*, Springer Verlag, New York, NY.

Wiginton, J., (1980). A Note on the Comparison of Logit and Discriminant Models of Consumer Credit Behaviour, *Journal of Financial and Quantitative Analysis*, **15**: 757-770.

Wilcox, A., (1971). A Simple Theory of Financial Ratios as Predictors of Failure, *Journal of Accounting Research*: 389-395.

Winakor, A. and Smith, R., (1935). *Changes in the Financial Structure of Unsuccessful Industrial Corporations. Bulletin No. 51*, University of Illinois, Bureau of Business Research, Urbana, Illinois.

Zavgren, C., (1983). The Prediction of Corporate Failure: The State of the Art, *Journal of Accounting Literature*, number 2: 1-38.

Zmijewski, M., (1984). Methodological Issues Related to the Estimation of Financial Distress Prediction Models, *Journal of Accounting Research*, **20**: 59-82.

FORSCHUNG

**Wolfgang Härdle
Harald Uhlig**

ÖKONOMISCHES

Sonderforschungsbereich 649

RISIKO

Das Thema »Ökonomisches Risiko« ist hochaktuell, denn wir leben in einer von Unsicherheit geprägten Welt: Unwägbarkeiten der Zukunft beeinflussen Entscheidungen der Gegenwart; Finanzmarktschwankungen bedrohen das in Immobilien, Rentenfonds und Aktien angelegte Vermögen der Haushalte; Unvorhergesehene Innovationen zerstören die Marktpositionen etablierter Firmen; Konjunkturbewegungen gefährden Arbeitsplätze; Wachstumsrisiken stellen die Wirtschaftspolitik vor immer wieder neue Herausforderungen; Die globalen Verflechtungen der Volkswirtschaften führen zur internationalen Übertragung nationaler Wirtschaftskrisen. Diese Unsicherheiten sind Ausprägung eines fundamentalen Phänomens, das man als »ökonomisches Risiko« bezeichnet. Es ist allgegenwärtiger Bestandteil ökonomischer Entscheidungen und Zusammenhänge. Das tiefere Verständnis ökonomischen Risikos ist von fundamentaler Bedeutung, will man die wirtschaftliche Lage von Individuen, Firmen oder der ganzen Nation verbessern.

Diese zentrale Bedeutung des ökonomischen Risikos ist in der bisherigen Forschung nicht genügend berücksichtigt worden. Zwar haben viele Disziplinen (z.B. Entscheidungstheorie der Haushalte, Mathematik und Ökonomie der Finanzmärkte, Statistik und Ökonometrie, Makroökonomie) im Rahmen ihrer Analysen auch ökonomische Risiken berücksichtigt und dabei verschiedene Herangehensweisen entwickelt. Die Grenzen zwischen diesen Disziplinen hat die Entwicklung einer umfassenden und gemeinsamen Sichtweise auf dieses Schlüsselthema allerdings bisher verhindert. Nur durch eine solche Sichtweise lassen sich aber eine Reihe drängender und disziplinübergreifender Fragen beantworten. Die Bündelung der verschiedenen Herangehensweisen, die Entwicklung einer transdisziplinären Perspektive auf das Thema »Ökonomisches Risiko« und die dadurch ermöglichte Beantwortung der gestellten Fragen ist daher das Ziel dieses Sonderforschungsbereichs (SFB 649).

I Forschungsfragen
Die grundlegenden und projektübergreifenden Fragestellungen des SFB 649 sind:

1. Was sind die wesentlichen ökonomischen Risiken und welche Konsequenzen haben sie? Welche gesamtwirtschaftlichen Risiken gibt es? Welchen zusätzlichen, individuellen Risiken sind Haushalte und Firmen ausgesetzt?

2. Wie wird und wie soll man mit diesen Risiken umgehen? Wie lassen sich diese Risiken verteilen und beherrschen?

3. Welchen Preis haben Versicherungen gegen diese Risiken? Welche Grenzen der Versicherbarkeit gibt es?

Aus diesen grundlegenden Fragestellungen ergeben sich spezifische Fragen, die in einer Reihe von Projekten behandelt werden.

*Logo des SFB 649:
»Ökonomisches Risiko«*
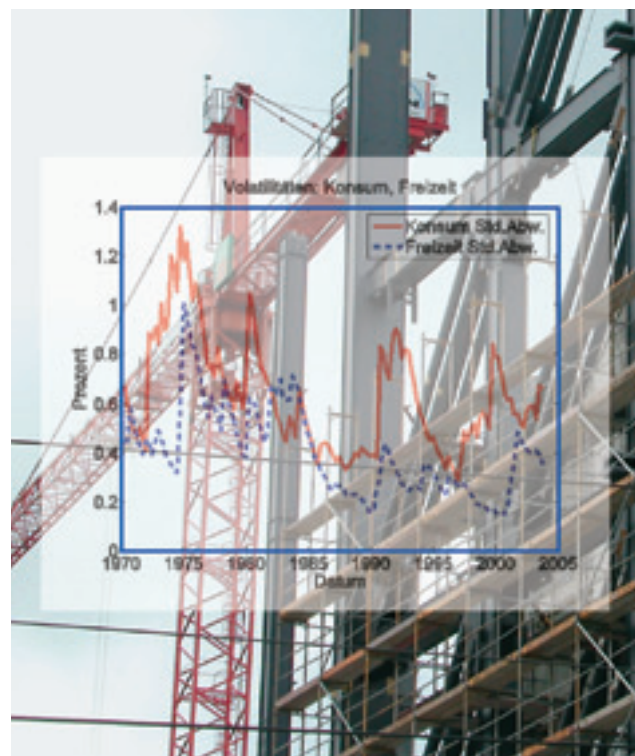
Abb. 1
Die Projektgruppe C forscht über Makroökonomische Risiken (Konjunktur, Wachstum, Zinsen, Finanzmärkte), die für die Gesamtwirtschaft nicht vermeidbar sind. Daher ist z.B. – wie die Grafik zeigt – die Analyse der Volatilität zielführend.

1. Gesamtwirtschaftliche Risiken machen sich als makroökonomische Fluktuationen, ausgelöst durch makroökonomische Schocks bemerkbar. Was sind die wichtigsten Faktoren makroökonomischer Fluktuationen und der Fluktuationen auf Finanzmärkten? Welche dynamischen Folgewirkungen makroökonomischer Schocks lassen sich empirisch auf makroökonomischer Ebene feststellen? Welche Konsequenzen ergeben sich dabei für individuelle Haushalte aus Vermögensrisiken, etwa dem Risiko der Schwankungen von Immobilienpreisen, und Humankapitalrisiken aufgrund von Arbeitslosigkeit? Welche Konsequenzen ergeben sich für Firmen aufgrund der Risiken auf Absatzmärkten? Welche Risiken ergeben sich aus dem Wechselspiel von Entscheidungen, die unter ungleich verteilten oder unvollständig verarbeiteten Informationen getroffen werden müssen?

2. Wie verhalten sich Haushalte angesichts zukünftiger Risiken und wie können Haushalte durch geeignete Portfoliowahl mit diesen Risiken umgehen? Welche Risikoaufteilung sollte zwischen Management und Firmeninhabern gewählt werden? Welche Rolle spielen Finanzmärkte bei der Verteilung von Risiken? Wie können gesamtwirtschaftliche Risiken durch geeignete wirtschaftspolitische Maßnahmen der Geld- und Fiskalpolitik kontrolliert werden?

3. Wie werden Risikofaktoren auf Finanzmärkten bewertet und welche Preise ergeben sich daraus für Versicherungen gegen diese Risiken? Welche Grenzen der Versicherbarkeit gibt es auf der Ebene

WIRTSCHAFTSWISSENSCHAFT

Härdle, W., Uhlig, H., Ökonomisches Risiko

des Managements und der Haushalte? Welche Grenzen sind der Versicherbarkeit internationaler und globaler Risiken gesetzt?

Die Projektgruppen werden sich diesen Fragen aus drei Hauptrichtungen nähern:

Projektgruppe A:
Individuelle und vertragliche Antworten auf Risiken

Erschlossen werden neue Zugänge zum Verständnis des individuellen Entscheidungsverhaltens von Haushalten und Firmen. Die Forscher untersuchen, wie die Entscheidungsregeln von Haushalten und Firmen Risiken und Unsicherheit berücksichtigen, und wie die Interaktion individueller Entscheidungen Risiko erst erzeugt. Sie berechnen Entscheidungsregeln für Haushalte zum Umgang mit Ungewissheit im Allgemeinen und untersuchen ihr Entscheidungsverhalten experimentell bezüglich des Risikos des Umgangs miteinander. Sie erforschen die Strategien zur Risikoverteilung in Firmen und Organisationen, insbesondere zwischen Management und Firmeninhabern und in Reaktion auf Absatzmarktrisiken.

Projektgruppe B:
Finanzmärkte und Risikobewertung

Finanzmärkte sind das Paradebeispiel für Märkte zur Risikoallokation. Die Forscher werden daher neue statistische Verfahren zur Analyse der Risikobewertung auf Finanzmärkten entwickeln und auf eine Reihe wichtiger Finanzmärkte anwenden. Die Vielzahl von Finanzmarktdaten werden sie nutzen, um Risikobewertungsfaktoren mit generalisierten partiell-linearen Modellen, dynamischen semiparametrischen Faktormodellen und strukturadaptiven Methoden zu schätzen. Die Analyse dieser Faktoren werden sie auf das Problem der Unternehmensbewertung aufgrund allgemeiner Risiken und aufgrund von Markenwertrisiken sowie auf Wertschwankungen für Immobilien anwen-

den. Diese empirisch-statistischen Projekte ergänzen sie durch die theoretische Erforschung der Implikationen des Insiderhandels und ungleich verteilter Informationen auf Finanzmärkten.

Projektgruppe C:
Makroökonomische Risiken

Gesamtwirtschaftliche, makroökonomische Risiken sind Risiken, die insgesamt unvermeidbar sind. Die Forscher extrahieren daher die wichtigsten makroökonomischen Faktoren, ermitteln ihre Bedeutung für Arbeits- und Finanzmärkte, erforschen ihre Konsequenzen für kurz- und langfristige Schwankungen des Wachstums sowie die internationale Übertragung nationaler Faktoren und ziehen wirtschaftspolitische Konsequenzen. Sie werden makroökonomische Risikofaktoren, ihre Allokation und ihr wirtschaftspolitisches Management untersuchen und dafür neue Zeitreihenmethoden entwickeln. Arbeitsmarktrisiken und dem Risiko der Arbeitslosigkeit werden sie besondere Aufmerksamkeit widmen. Sie werden internationale Risiken und konjunkturpolitische Konsequenzen und dabei insbesondere die Risiken der EU-Osterweiterung für die europäische Geldpolitik untersuchen. Schließlich werden sie langfristige Wachstumsrisiken aus wirtschaftshistorischer und theoretischer Sicht erforschen.

II Stärkung der internationalen Konkurrenzfähigkeit in den Wirtschaftswissenschaften

Die Struktur der internationalen Forschung im Bereich der Wirtschaftswissenschaften hat sich in den letzten zwanzig Jahren grundlegend geändert. Quantitative, numerisch-theoretische und empirisch-datenorientierte Forschung, Publikation dieser Forschung in international führenden Zeitschriften, und die Ausbildung der Studenten sowie die Entscheidungsfindung in privaten und öffentlichen Institutionen auf der Basis dieses Wissens ist inzwischen internationaler Standard. Diesen Entwicklungen ist in der Bundesrepublik bisher aber nur ungenügend Rechnung getragen worden.

Aus der ›Vogelperspektive‹ betrachtet, fließen tatsächlich große Summen in die Wirtschaftsforschung in Deutschland, und zwar (neben Einzelfördermaßnahmen und die in Universitäten als Dienstaufgabe betriebene Forschung) vor allem in die Wirtschaftsforschungsinstitute (DIW in Berlin, HWWA in Hamburg, CES-ifo in München, IFW in Kiel, IWH in Halle, RWI in Essen, ZEW in Mannheim). Die Arbeiten dort waren aber lange Zeit vor allem am Beratungsbedarf politischer Institutionen ausgerichtet. Die aktive Teilnahme am internationalen Fortschritt der Forschung auf diesem Fachgebiet, nachweisbar durch Publikationen in führenden Fachzeitschriften, hat dort in der Vergan-

Härdle, W., Uhlig, H., Ökonomisches Risiko

genheit nur eine untergeordnete Rolle gespielt. Die Wirtschaftsforschungsinstitute unternehmen seit neuester Zeit deutliche Anstrengungen, um diese Ausrichtung zu ändern. Diese Umbauprozesse werden aber einige Zeit in Anspruch nehmen.

Dagegen ist die oben definierte, international ausgerichtete Forschung in eher geringerem Umfang gefördert worden. Von über 300 Sonderforschungsberei-

chen, die die Deutsche Forschungsgemeinschaft (DFG) im Jahr 2003 gefördert hat, haben sich nur zwei vorrangig mit wirtschaftswissenschaftlichen Fragen auseinandergesetzt, der SFB 504: »Rationalitätskonzepte, Entscheidungsverhalten und ökonomische Modellierung« in Mannheim und der 2003 ausgelaufene SFB 373: »Quantifikation und Simulation ökonomischer Prozesse« in Berlin. Von 80 Max-Planck-Instituten desselben Jahres befasst sich lediglich das Max-

---

SFB 649: Ökonomisches Risiko

Seit Januar 2005 hat der neue SFB seine Forschungen zum Thema »Ökonomisches Risiko« aufgenommen. Besonders stolz ist man darauf, dass dies einer von nur drei Sonderforschungsbereichen in Deutschland ist, der sich vor allem wirtschaftswissenschaftlichen Fragen widmet. Die größte Gruppe der beteiligten Forscher kommt von der Wirtschaftswissenschaftlichen Fakultät der Humboldt-Universität. Mit dabei sind Mathematiker, Statistiker und Wirtschaftswissenschaftler der Humboldt-Universität zu Berlin, der TU Berlin, der FU Berlin, des Weierstraß-Instituts, der Universität Potsdam und des DIW. Insgesamt werden 18 Teilprojekte gefördert. Dabei werden etwa 30 Stellen für Doktoranden eingerichtet, die diese Projektthemen erforschen werden. Erstmalig in Deutschland soll ein Finanz- und Wirtschaftsdatenzentrum (FEDC) eingerichtet werden, das die Forscher und Gäste des SFB in ihrer Arbeit unterstützen wird wie auch die gefundenen Ergebnisse der Öffentlichkeit zugänglich macht.

Sprecher:
Prof. Harald Uhlig, Ph.D., Professor für Wirtschaftspolitik an der Humboldt-Universität zu Berlin.
Stellvertreter: *Prof. Dr. Wolfgang Härdle*, Professor für Statistik an der Humboldt-Universität zu Berlin.

Teilprojektleiter:
Dr. Imke Brüggemann (FU Berlin): Quantitative Analyse der Geldpolitik im erweiterten Europa (C6).
Dr. Ralf Brüggemann (Humboldt-Universität): Einheitswurzel- und Kointegrationsmethodik (C2).
Prof. Michael C. Burda, Ph.D. (Humboldt-Universität): Risiken der Erwerbstätigkeit: Humankapital, Entlohnung und Arbeitslosigkeit (C7).
Prof. Dominique Demougin, Ph.D. (Humboldt-Universität): Optimale Risikoverteilung in Organisationen (A4).
Dr. Ingolf Dittmann (Humboldt-Universität): Messung von Unternehmenswert und Risikoprämien (B4).
Prof. Dr. Wolfgang Härdle (Humboldt-Universität): Dynamische semiparametrische Modellierung (B1).

Prof. Dr. Lutz Hildebrandt (Humboldt-Universität): Markenbewertung und Bewertung von Markenstrategien (B2).
Prof. Dr. Franz Hubert (Humboldt-Universität): Immobilienbewertung und Investition (B3).
Prof. Dr. Peter Imkeller (Humboldt-Universität): Stochastische Analysis von Finanzmärkten mit heterogener Information (B6).
Prof. Dr. Dorothea Kübler (TU Berlin): Strategisches Risiko in experimentellen Spielen (A6).
Juniorprofessor Bartosz Mackowiak, Ph.D. (Humboldt-Universität): Internationale und makroökonomische Risiken, ihre Ursachen und ihre konjunkturpolitische Beherrschbarkeit (C3).
Prof. Ernst Maug, Ph.D. (Humboldt-Universität): Risiko und Design von Management-Vergütungsverträgen (A1); Messung von Unternehmenswert und Risikoprämien (B4).
Dr. Markus Reiß (Weierstraß-Institut): Stochastische Optimierung für ökonomische Modelle unter Berücksichtigung von Zeitverzögerungseffekten (C4).
Prof. Dr. Albrecht Ritschl (Humboldt-Universität): Gesamtwirtschaftliches Risiko in langfristiger Perspektive (C5).
Prof. Dr. Alexander Schied (TU Berlin): Optimierung dynamischer Konsumströme unter Ungewissheit (A3).
Prof. Dr. Vladimir Spokoiny (Weierstraß-Institut): Strukturadaptive Datenanalyse (B5).
Dr. Carsten Trenkler (Humboldt-Universität): Einheitswurzel- und Kointegrationsmethodik (C2).
Prof. Harald Uhlig, Ph.D. (Humboldt-Universität): Makroökonomische Risiken: Faktoren, Kapitalmärkte und wirtschaftspolitische Konsequenzen (C1).
Prof. Axel Werwatz, Ph.D. (DIW und Universität Potsdam): Immobilienbewertung und Investition (B3).

Fördereinrichtung:
Deutsche Forschungsgemeinschaft (DFG)

Förderzeitraum:
01/2005 – 12/2008 (erste Förderperiode)

Härdle, W., Uhlig, H., Ökonomisches Risiko

Die Projektgruppe B des SFB 649 befasst sich mit Finanz-
märkten und Risikobewertung. *Die Grafik zeigt den möglichen
Verlauf eines Aktienkurses mit Zielkorridor (roter Bereich),
der für die Gestaltung eines Optionspreises relevant ist.
Im Hintergrund: Bulle und Bär vor der Frankfurter Börse.
(Foto: Deutsche Börse)*

Planck-Institut in Bonn mit der Erforschung von wirt-
schaftswissenschaftlichen Themen.

Dies hat eine Reihe von Konsequenzen. Zunächst sind
die Konsequenzen für Deutschland als Wissenschafts-
standort für die Wirtschaftswissenschaften zu nennen.
So ist Deutschland mit einem Anteil am weltweiten
Publikationsoutput von durchschnittlich 2% in den
international führenden wirtschaftswissenschaftlichen
Zeitschriften relativ zu seiner ökonomischen Bedeu-
tung extrem schwach vertreten. In einem internatio-
nalen Ranking (Kalaitzidakis, Manuneas und Stengos,
2001) tauchen lediglich zwei Fachbereiche (Univer-
sität Bonn und Humboldt-Universität zu Berlin) unter
den 100 weltweit führenden Fachbereichen der Volks-
wirtschaft auf. Die Gesamtpublikationsaktivitäten
deutscher Wirtschaftswissenschaftler in führenden
Zeitschriften werden von kleineren Ländern wie den
Niederlanden oder Belgien übertroffen und fallen
gegenüber den USA weit zurück.

Daraus ergeben sich Konsequenzen für die Qualität
der Ausbildung, für die Qualität der wissenschaftli-
chen Politikgestaltung und für das Management deut-
scher Firmen. Die besten Universitäten sind die Uni-
versitäten mit den besten Wissenschaftlern, wie man
an der Nachfrage der Studenten nach Studienplätzen
an z.B. den Universitäten Stanford, Harvard oder Prin-
ceton unschwer erkennen kann. Hier liegt Deutschland
deutlich zurück. Dies betrifft nicht nur den Bereich
»Volkswirtschaft«, sondern auch die »Betriebswirt-
schaft«: der Zugang zu den besten MBA-Programmen
in den USA ist häufig stark umkämpft, diese Program-
me sind der Lieferant des Topmanagements weltweit
führender Wirtschaftsunternehmen. Auch deutsche
Unternehmen rekrutieren ihr Topmanagement zuneh-
mend aus den MBA-Programmen der USA.

In internationalen Organisationen wie dem Internatio-
nal Monetary Fund (IMF), der Organisation für wirt-
schaftliche Zusammenarbeit und Entwicklung (OECD)
oder der Weltbank finden sich zunehmend seltener
Ökonomen, die in Deutschland promoviert haben. Die
Bundesbank muss sich im Rahmen des Systems
europäischer Zentralbanken neu aufstellen und ihre
eigenen, international konkurrenzfähigen Forschungs-
aktivitäten verstärken. Eigene Forschungsabteilungen
der Wirtschafts- und Finanzministerien treten wissen-

Härdle, W., Uhlig, H., Ökonomisches Risiko
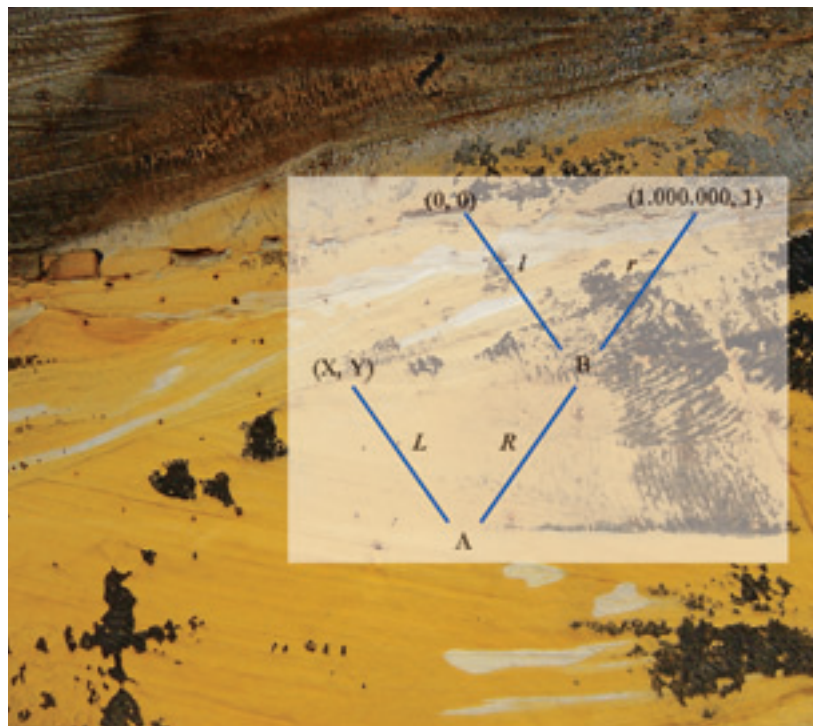
Abb. 3
Im Projektbereich A des SFB 649 werden »Individuelle und vertragliche Antworten auf Risiken« erforscht, z.B. mittels experimenteller Spiele. Strategische Unsicherheit tritt hier beispielsweise auf bei Unsicherheit über das Verhalten anderer.

schaftlich nicht in Erscheinung und sind personell deutlich unterbesetzt, um dies neben dem politischen Tagesgeschäft noch leisten zu können. Der Kontakt zum Fortschritt in den Wirtschaftswissenschaften ist daher in Deutschland ungenügend, da der Beratung von außen keine ausreichende Anzahl von Experten gegenüberstehen, die durch aktive Teilnahme am Wissenschaftsgeschehen die gleiche Sprache sprechen und die Beratungsinhalte entsprechend einschätzen können.

Die Wirtschaft in Deutschland wächst seit Jahren zu langsam. Die Finanzpolitik hat Mühe, die vom Stabilitäts- und Wachstumspakt vorgeschriebenen Verschuldungsgrenzen einzuhalten. Die Arbeitslosigkeit hat sich auf ein zu hohes Niveau eingependelt. Diese Entwicklungen bedrohen den Lebensstandard einer ganzen Nation. Es wäre sicher vermessen, in der Vernachlässigung der international orientierten, wirtschaftswissenschaftlichen Forschung die Ursache für die deutschen Wachstumsprobleme zu suchen. Es ist heute absolut notwendig, dass durch Investitionen in diesen Wissenschaftsbereich eine langfristige Per-

spektive zur Verbesserung dieses Zustandes gewährleistet wird. Mit dem SFB »Ökonomisches Risiko« wird diese Perspektive eröffnet.

III Prädestiniert als »Sitz« des neuen SFB 649: Berlin und die Humboldt-Universität

Der Berliner Raum bietet eine Reihe von wichtigen Voraussetzungen für die Einrichtung des SFB. Die Konzentration von Wissenschaftlern, die verschiedene Probleme des ökonomischen Risikos untersuchen, ist enorm. Neben den drei großen Universitäten gibt es in Berlin weitere Forschungseinrichtungen wie das Weierstraß-Institut oder das Deutsche Institut für Wirtschaftsforschung (DIW), die sich am SFB 649 beteiligen. Das Masterprogramm MEMS an der Humboldt-Universität und ein Master-Studiengang »Statistik« (www.stat.de) an der FU Berlin und der Humboldt-Unversität existieren bereits seit einigen Jahren.

Darüber hinaus hat sich Berlin in den letzten Jahren als Zentrum für angewandte Mathematik in Deutschland etabliert. Diese Stellung wurde durch die Gründung des Forschungszentrums »Mathematik für Schlüsseltechnologien« bestätigt und ausgebaut. Obwohl keine thematische Überlappung zwischen dem Forschungszentrum und dem neuen Sonderforschungsbereich besteht, wird der SFB 649 von den mathematischen Methoden, die im Forschungszentrum entwickelt werden, profitieren.

Berlin ist eine der attraktivsten wissenschaftlichen Orte, um Wissenschaftler, Postdoktoranden und längerfristige Besucher aus der ganzen Welt anzuziehen. Insbesondere sind das »Quantitative Finance Seminar«, das »Schumpeter Seminar« und das »Mathematische Statistik Seminar« zu wichtigen Foren des internationalen wissenschaftlichen Austausches geworden. Gleiches gilt für die regelmäßig stattfindenden »CASE Distinguished Lecture Series« (www.case.hu-berlin.de), »Hermann-Otto-Hirschfeld Lectures« und die »Schumpeter Lecture Series«.

Die internationale Ausstrahlung wird zudem durch das MEMS-Programm (Masters in Economics and Management Science), das deutsch-französische Doppeldiplom in Statistik und Ökonomie, das »Berlin Doctoral Program« und zahlreiche Sokrates-Programme mit Universitäten benachbarter Länder (u.a. Prag, Breslau) ergänzt.

Die Wirtschaftswissenschaftliche Fakultät der Humboldt-Universität zu Berlin zählt zu den publikationsstärksten wirtschaftswissenschaftlichen Institutionen in Deutschland. Die zentrale Lage bietet vielfache

Härdle, W., Uhlig, H., Ökonomisches Risiko

Möglichkeiten zur Vernetzung mit insbesondere wirtschaftspolitischen Institutionen. Das Hauptstadt-Netzwerk der wirtschaftswissenschaftlichen Forschung und Beratung an anderen Instituten der Humboldt-Universität, inklusive der neuen »Humboldt-Viadrina School of Governance«, an den anderen Universitäten, dem DIW, dem »Wissenschaftszentrum für Sozialforschung«, der neu gegründeten »European School of Management and Technology« sowie der »Hertie School of Governance« wird so durch ein wichtiges Glied an wichtiger Stelle ergänzt.

Durch die Einrichtung eines Finanz- und Wirtschaftsdatenzentrums (FEDC) wird eine Vorreiterfunktion für die wirtschaftswissenschaftliche Forschung für andere deutsche Fakultäten geliefert. Die hier gewonnenen Erfahrungen und der dadurch erzeugte Forschungsschub wird anderen wirtschaftswissenschaftlichen Fakultäten die Möglichkeit und die zusätzliche Motivation geben, ebenfalls verstärkt an den wichtigen, internationalen Forschungsaktivitäten teilzunehmen, und so den kritischen Rückstand Deutschlands aufzuholen.

Die Humboldt-Universität zu Berlin sieht die interdisziplinäre Kooperation zwischen der wirtschaftswissenschaftlichen Fakultät und dem Institut für Mathematik als Struktur, die auch langfristig unterstützt wird. Das interdisziplinäre Zentrum CASE – Center for Applied Statistics and Economics (www.case.hu-berlin.de) – verstärkt diese Zusammenarbeit der Institute in den quantitativen Wirtschaftswissenschaften. Es organisiert in diesem Zusammenhang z.B. die »Case Distinguished Lecture Series« u.a. in Zusammenarbeit mit der Deutschen Bank und der Volksbank Berlin.

Innerhalb der wirtschaftswissenschaftlichen Fakultät ist das Schumpeter Institut (www.wiwi.hu-berlin.de/wpol/schumpeter/) ein aktives Forum für den wissenschaftlichen Austausch. Unter den Aktivitäten sind vor allem ein wöchentliches Seminar sowie von Gastwissenschaftlern gegebene Minikurse auf Doktorandenniveau zu nennen.

Die Fähigkeit der Wirtschaftswissenschaftlichen Fakultät und des Instituts für Mathematik der Humboldt-Universität zu Berlin, wissenschaftlichen Nachwuchs von der gewünschten Qualität auszubilden und am Markt zu platzieren, ist in der Vergangenheit bereits im Rahmen des SFB 373: »Quantifikation und Simulation ökonomischer Prozesse« unter Beweis gestellt worden. Der jetzt angelaufene SFB 649: »Ökonomisches Risiko« plant, mindestens gleichermaßen erfolgreich zu sein.

Prof. Dr. Wolfgang Härdle
Jg 1953. Nach Promotion (Universität Heidelberg) und Habilitation (Universität Bonn) war W. Härdle zunächst Visiting Professor (1989–90), dann Professeur Ordinaire (C4), CORE, an der Université Catholique de Louvain (1990-92). Seit 1992 Professor (C4) an der Wirtschaftswissenschaftlichen Fakultät der Humboldt-Universität zu Berlin, Fachgebiet Statistik; Sprecher des abgeschlossenen SFB 373: »Quantifikation und Simulation ökonomischer Prozesse« (1/1994–12/2003); seit 1/2005 stellvertretender Sprecher des SFB 649: »Ökonomisches Risiko«.

Kontakt
Humboldt-Universität zu Berlin Wirtschaftswissenschaftliche Fakultät Spandauer Str. 1 D–10178 Berlin Tel.: +49 30-2093–5630 Fax: +49 30-2093–5649 E-Mail: haerdle@wiwi. hu-berlin.de http://ise.wiwi.hu-berlin.de

Prof. Harald Uhlig, Ph.D.
Jg. 1961. Nach Promotion (University of Minnesota) war H. Uhlig von 1990–1994 Assistant Professor an der Princeton University, NJ (USA); 1994–2000 Research Professor in Macroeconomics am CentER der Universität Tilburg, 1999–2000 Forschungsjahr an der Stanford University, Dept. of Economics; seit 2000 Professor für Wirtschaftspolitik an der Humboldt-Universität zu Berlin; Forschungsprofessor an der Deutschen Bundesbank seit 2004. Sprecher des SFB 649: »Ökonomisches Risiko« seit 1/2005.

Kontakt
Humboldt-Universität zu Berlin Wirtschaftswissenschaftliche Fakultät Spandauer Str. 1 D–10178 Berlin Tel.: +49 30-2093–5926 Fax: +49 30-2093–5934 E-Mail: uhlig@wiwi. hu-berlin.de http://www.wiwi. hu-berlin.de/wpol/

SFB 649 online:
http://sfb649.wiwi. hu-berlin.de/

Härdle, W., Uhlig, H., Ökonomisches Risiko

JOURNAL OF
Economic Behavior
& Organization

Transactions that did not happen and their influence on prices[☆]

Alan Kirman[a,*], Rainer Schulz[b,c],
Wolfgang Härdle[c], Axel Werwatz[d,c]

[a] *GREQAM, EHESS, IUF, 2 Rue de la Vieille Charité, 13002 Marseille, France*
[b] *University of Aberdeen Business School, Edward Wright Building, Aberdeen AB24 3QY, Scotland, UK*
[c] *Center for Applied Statistics and Economics, Humboldt-Universität zu Berlin,
Spandauer Straße 1, 10178 Berlin, Germany*
[d] *German Institute for Economic Research (DIW), Königin-Luise-Straße 5, 14195 Berlin, Germany*

Abstract

This paper studies data from the wholesale fruit and vegetables market in Marseille. We have details of counteroffers to the prices that were proposed by the seller even when no transaction took place. We analyse the evolution of prices and the relation between the final price struck and the proposals of the two parties. Periods with no buyer refusals, of offers or bargaining with no transaction will lead to a revision of the seller's first price. More importantly the sharing of the surplus moves in the buyer's favour during the day. These theoretical predictions are confirmed by our data set.
© 2004 Elsevier B.V. All rights reserved.

* Corresponding author: Tel.: +33 491140751; fax: +33 491100227.
E-mail address: kirman@ehess.univ-mrs.fr (A. Kirman).

1. Introduction

Bargaining literature presents several problems when one confronts it with empirical facts. In many markets, unlike the situation in the Rubinstein (1982) framework, haggling does, in fact, take place. It is almost never the case that we have details of the intermediate offers and counteroffers that make up such haggling, and this is particularly true of those cases in which bargaining took place but no bargain was struck. Data on final transactions does not reveal the process that arrived at the final bargain. This would be of no importance if the offers and counteroffers that are made during the bargaining process had no influence on the final outcome. However, if we regard them as revealing information, then they would seem potentially influential.

In this paper we present evidence from the wholesale fruit and vegetable market in Marseille, and we analyse the data for four products. This data contains details of the prices asked and the counteroffers made, even when no transaction took place. This situation is difficult to fit into the existing literature. What happens in reality is that both buyer and seller get information about "the state of the market" as the day goes on. Each price proposed conveys information to the buyer, and each counteroffer, or refusal, provides information for the seller.

In a market with perishable goods or in one where buyers can be separated according to how long they remain in the market, one might expect to see declining prices over the day. Although in a world where waiting is not costly, the prices at the end of the day should be the same as at the beginning if the appropriate strategies were adopted, there are various explanations for the fact that, in reality, descending prices are frequently observed. The literature on the "declining price phenomenon" in auctions, Ashenfelter (1989), McAfee and Vincent (1993), Pezanis-Christou (1997), bears witness to this. However, curiously in Härdle and Kirman (1995), we found that this feature was essentially absent from the Marseille fish market.

To get a clearer picture of the evolution of prices in this market, one could think of relating the analysis to the bargaining under two-sided uncertainty literature. In that case, both buyer and seller have information that is unknown to the other, and it is typically assumed that the information in question is the "value" of the object over which bargaining is taking place. In our case the information that is being implicitly conveyed is the aggregate quantity available and hence which prices may obtain. Thus, a refusal of a price by a buyer may reflect his having found a cheaper price elsewhere or his lack of fear of not finding what he wants elsewhere. This information should be useful to the seller who may then modify the price he has set. Similarly, if several buyers in succession accept the price asked immediately, then the seller might reasonably infer that this product is in short supply and then modify his price accordingly. In other words, the behaviour of the buyer is a signal about his reservation price that, in turn, reflects his ideas about the distribution of prices available. The problem is that the situation we have described does not correspond to the standard literature since several different partners are involved successively. Thus the offers and counteroffers in one matching influence those made in a subsequent matching, but the new partner does not have the same information as if he were there from the outset.

If all agents of each type, buyer or seller, were identical, the fact that the market unfolds in this fashion would have no significance. However, the problem would be of a different nature since each seller would be aware of the state of the market from the outset. As it is, sellers obtain their stocks from different sources, they have certain clients who are loyal and do not go elsewhere, and the clients have different demands. Thus, although the price charged to each particular buyer is specific to the seller involved, that price is influenced by the general state of the market. What price I, as a buyer, can obtain today will depend on what the seller I visit has had to pay and his view of the quantity available on the market today. The maximum price that I am willing to accept, my reservation price, will depend on the offers that I have obtained previously and my view of the quantity available. Each time there is a meeting, information is conveyed, and the price asked in the next period may be influenced by that information. Since our data is for one seller, we cannot track the reactions of buyers as they move from one seller to another. However, we can observe the reactions of our seller to each meeting. The purpose of this paper is to examine those reactions.

2. Our data

We have data for 15 trading days in March of the year 1983. In total, we have 3960 observations for four kinds of fruits and vegetables. These are leeks, domestic tomatoes, imported tomatoes, and oranges. The oranges and the imported tomatoes are from Spain, whereas the leeks and the domestic tomatoes are from France. For the purpose of collecting data, every trading day is divided into a time grid. From 7 to 9 o'clock, the distance between grid points is 2 min 30 s. Thus, there are 48 time intervals with equal length. From 9 to 12 o'clock, there are 18 intervals of 10 min length. Fig. 1 depicts the way in which a meeting between our seller and an arbitrary buyer on the wholesale fruit and vegetable market in Marseille might unfold.

The seller makes an initial price offer p_1^* that is either accepted or rejected. If the buyer accepts, a *transaction occurs with no bargaining* (TNB) at the offered price. Thus, the transaction price p is equal to p_1^*. If the buyer rejects he may either make a counteroffer c or walk away. In the latter case, the meeting ends without a deal being struck, and we *only observe the seller's offer price* (OO). If he makes a counteroffer, then both parties either agree on a transaction price somewhere in the closed interval defined by c and p_1^* or stop the negotiation. In the wholesale fruit and vegetable market in Marseille, this final round of bargaining usually takes the form of the seller making a second and final price offer p_2^*, given c and p_1^*. If the buyer accepts then we observe three prices (that are p_1^*, c, and p_2^*) and refer to this outcome as a *transaction with bargaining* (TWB). In that case, the transaction price p is equal to p_2^*. In the other case we observe the seller's initial *offer and the buyer's counteroffer* (OC), but no transaction occurs.

Panel A of Table 1 shows the relative frequencies of these different categories of encounters for the different sorts of vegetables, cumulated over all days. Panel B shows the distribution of the categories given that a contact happened. Whereas the frequencies for the two types of tomatoes and oranges are quite similar, the behavior on the leeks market is different. For the former, no transaction occurs for about 5% of contacts. For the latter,
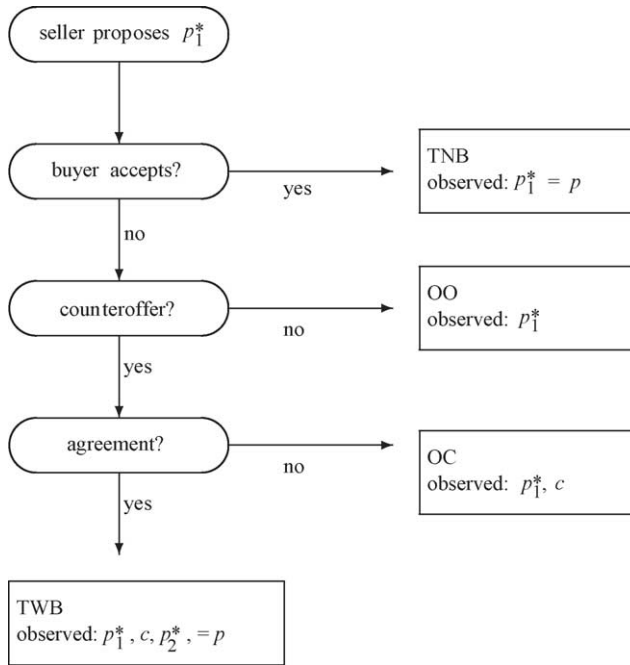
Fig. 1. How a meeting between seller and buyer may evolve. The abbreviations are transaction with no bargaining (TNB), offer only (OO), offer and counteroffer (OC), and transaction with bargaining (TWB). p_1^* is the seller's initial offer price, c the buyer's counteroffer, p_2^* the seller's second offer price and p the transaction price.

about 30% of all contacts end without a transaction. Thus, it was more difficult to strike a deal for leek than for the other products.

In this paper we use our data to shed light on the following questions: (1) How did the prices offered by the seller evolve during the day? (2) How did the sellers' prices react to a successful transaction or to the failure to transact? (3) How did buyers' counteroffers evolve over the day? (4) What was the influence of the difference between the seller's price and that of the buyer on the probability of a transaction? (5) How did the transaction price relate to the seller's and buyer's price, and did this vary over the day?

3. Related theoretical approaches

As has been mentioned, at first sight the problem posed here might be thought of as corresponding to a bargaining situation with two-sided uncertainty. The price that a seller is willing to accept is unknown to the buyer, and the price that a buyer is willing to pay is unknown to the seller at each point in time. This setting seems to be related to that in which each player has a private reservation value as described by Myerson and Satterthwaite (1983). However, for their results they require that the reservation values be drawn from

Table 1
Whole database: relative frequency of the categories for the different sorts of vegetables (%)

| | No contact | Contact | | | |
| | | OO | OC | Transaction | |
				TNB	TWB
Panel A: Relative frequencies of different categories					
Leeks	41.11	11.62	5.96	12.22	29.09
Domestic tomatoes	58.59	1.72	0.51	11.01	28.18
Imported tomatoes	51.72	1.11	0.61	11.92	34.65
Oranges	53.03	1.11	0.91	5.35	39.60
Panel B: Relative frequencies conditional on a contact					
Leeks		19.73	10.12	20.75	49.40
Domestic tomatoes		4.15	1.22	26.59	68.05
Imported tomatoes		2.30	1.26	24.69	71.76
Oranges		2.37	1.94	11.40	84.30

Notes: Every trading day is divided into time grids. From 7 to 9 o'clock, the length of the grid is 2 min 30 s. From 9 to 12 o'clock, the length of the grid is 10 min. The categories describe the way in which a meeting between seller and buyer in a time grid might unfold. If no buyer shows up, no contact happened. If a buyer shows up, there are four categories: offer only (OO), offer and counteroffer (OC), transaction with no bargaining (TNB), and transaction with bargaining (TWB); see also Fig. 1.

distributions that are known to each of the partners, even though the particular values are not known. In our case the values are under constant revision. Each meeting corresponds to the situation involved in Myerson and Satterthwaite's analysis except for the fact that the distribution of each player's reservation value has been modified by his experience. Thus, at any point in time there is no reason to believe that the distributions of the two bargainers will coincide.

The question then becomes one of what happens when both partners have formed an idea of the state of the market. In this case one would expect to find a convergence to a conventional bargaining problem. A second problem arises in our particular case. Learning is accompanied by transactions; the process is one of non-tatonnement. Thus the reservation price of the parties will be conditioned not only by their knowledge of the market situation but also by their sales or purchases up to that point.

Given these difficulties we tried, as a first approach, to examine a certain number of plausible hypotheses and to establish a certain number of stylised facts that would be consistent with the theory for simpler situations.

The central theme of this paper is the influence of bargaining on the probability of transactions and on the behaviour of buyers and sellers. A model developed by Grossman and Perry (1986) is related to our analysis. They developed a model in which a buyer bargains with a seller and where information about the reservation prices is incomplete. In their case the distribution of the buyer's reservation price is known but the realisation is not. The seller's reservation price is assumed to be public knowledge. The negotiation involves one indivisible unit of a good. The game is structured as follows: (i) at t the seller makes an initial offer; (ii) at $t + 1$ the buyer accepts or makes a counteroffer; (iii) at $t + 2$ the seller accepts the counteroffer or makes a new offer; (iv) this process continues until an offer is

accepted. At each point in time the seller has a prior distribution on the buyer's reservation price, $F(x)$, and in the Grossman Perry model this distribution is common knowledge. In their model an action for a player is an acceptance of the previous offer or a new counteroffer. There is a discount factor Δ such that if the expected value for the player of playing the game today is $V(a, b)$, then the value tomorrow (if there no change in the support of the distribution) is $\Delta V(a, b)$. In our case the value of waiting is the value of playing with the next buyer or seller.

The basis result shown by Grossman and Perry is that one will observe a series of offers and counteroffers terminating in an offer that is accepted. The essential feature of the model that tends to negotiations actually being observed is the asymmetric information. Negotiations reveals the information about the buyer's reservation price.

In our case the result of Grossman and Perry does not apply directly since the seller's reservation price is not zero but is what he considers to be the value of playing with someone else. Their model does, however, provide us with insight as to how the process develops between two agents and suggests that one should see offers that are taken, counteroffers that are accepted, and new offers that are accepted. The distribution $F(x)$ is revised at time t, and we must then ask the question as to how the observations up to a certain point affect the revision. The difficulty about models such as that of Grossman and Perry is that they require full-blown game theoretic reasoning. In what follows we will specify the behaviour of our seller with a more limited reasoning. After making some observations about how the buyer's reservation prices will shift over time we will present a model that represents the seller's choice of his initial price and then a revision process, and finally we will consider what the predicted sign of the effect of various types of observation should be on the parameter governing the revision process.

3.1. Buyer's behaviour

An important question here is how buyers revise their reservation prices. They are searching in the market, and given the distribution of prices that they observe, they will modify their reservation prices. Can we reduce our situation to that analysed in standard search models? There the situation is clear. There is a distribution of sellers' prices amongst which buyers search and, in the standard version, the rule is simple. Consider that price which is such that the gain from an additional search is just equal to the cost of making that search; call this the *reservation price*. Then search until one finds a price that is less than or equal to this value. Formally if the cumulative density of prices is given by $F(p)$ and the cost is c, then the reservation price is given by R such that

$$\int_0^R (R - p) \, dF(p) = \int_0^R F(p) \, dp = c.$$

The first problem is that F is assumed to be known and to be fixed during the search. This is important since as Gastwirth (1976) points out the mis-specification of F may lead to a highly inefficient level of search. As he notes, if someone based a reservation price on a uniform distribution while the distribution was in fact triangular on the same support, he

would, in expectation, spend twice as much and would search five times more often than if he were correctly informed.

In our case the buyer does not know the true distribution, and what is more, that the distribution is changing over time. To clarify ideas we can adopt an approach suggested by Rothschild (1974) in a classic paper. Since prices have to be in monetary units and they are evidently bounded, it is reasonable to assume that there is a finite number (say n) of them. The prices $p_1, p_2, \ldots, p_n$ are arranged in ascending order, and the distribution of prices is then a multinomial distribution Π where

$$\Pi \in \Delta = \left\{ (p_1, p_2, \ldots, p_n) \in \mathbb{R}^n | p_i \geqslant 0, \sum p_i = 1 \right\}.$$

Now suppose that the buyer does not know Π but has a prior $F(\cdot)$ over Δ. He then updates his prior using Bayes' rule. He does this using the information available at each time. His information consists in the number of times N_i he has observed the ith price, so this information can be represented by the vector $N = (N_1, \ldots, N_n)$, or more conveniently, one can think of the vector λ in the simplex Δ where

$$\lambda_i = \frac{N_i}{\sum_{j=1}^n N_j}.$$

This vector is updated in the obvious way as an observation is made. Under reasonable conditions, such a process will converge to the true distribution. There are technical problems with the initial prior (see Rothschild) but if the number of observations becomes sufficiently large this can be ignored. Yet in the case in question, updating is basically limited to 1 day. Buyers are faced each day with a distribution that is not the same as on the previous days, so they are obliged to start with a relatively uninformative prior.

One answer to this problem would be to think of a situation where, on each day, the price in the outside market where the sellers obtain their stocks is either high or low. The problem of the buyer then becomes deciding whether he is sampling from a distribution with mean μ_h or mean μ_l where $\mu_h > \mu_l$.

Early in the day he is faced with a classical problem in economics: that of choosing between profiting from the opportunity he is currently faced with or gaining more information by searching again. The gain in information about the distribution is offset by the potential lost opportunity if the seller no longer has stocks available.

Here again, the situation is not simple since the distribution is constantly being modified in two ways. First, some sellers are no longer active after they have sold all their stock, and second, as the day goes on they, in turn, modify their prices.

The situation for each buyer and seller is then reduced to two steps. Once he observes the prices or reactions, he has to decide which is the prevalent distribution, and then he has to decide on his bargaining behaviour. This will be true even when the distribution of seller's prices is more degenerate. How will the buyers' reservation prices evolve over the day? Those who have high reservation prices to start with will disappear from the market since they will accept offers except from very high price setters. At each successive round those who will be left in the market are buyers with initially low reservation prices faced with

higher priced sellers. However, any given seller will be faced with a distribution that is being truncated from the top. Even though some buyers maybe revising their prices upwards, the distribution as a whole will be shifting down.

3.2. The timing of negotiations

Two things are important to note here. The information provided by our seller suggests that the larger buyers operate early in the morning, and it is the case that there are proportionally more sales without negotiation in the earlier period of the market. This would be consistent with the evidence provided by Weisbuch et al. (2000) that large buyers typically do not "shop around" and become "locked in" to particular sellers. Many economic explanations can be given for this fact, such as the existence of implicit contracts or preferential treatment. The argument used in the paper referred to was one of simple reinforcement learning. However, with one exception there is little evidence that volume decreases over the day, and we will come back to this later. It might seem to be the case that as the market unwinds, the negotiators would arrive at a one-shot sub game perfect equilibrium and that this would remove counteroffers at the end of the day.

With the exception of local tomatoes, which were never sold after 9 h 30 min, transactions continued until the official closing time. In the case of perishable goods a phenomenon that one might expect to see is a version of the ultimatum game. However no obvious distinction can be made between the behaviour of the market for oranges and that for imported tomatoes, for example.

Consider first of all the prices proposed by sellers as time passes. As can be seen from Fig. 2, three different patterns emerge. The prices proposed for oranges and imported tomatoes are practically constant over the day; no trend is statistically detectable. The prices of leeks decline steadily over the day, whilst local tomatoes on average are constant till 8 h 30 min and then drop sharply and become constant again at 9 h.

How might one set about explaining these differences? A first observation is that those products with constant prices are both imported, having been bought on a large foreign wholesale market at a price that is essentially common knowledge so that each seller has a fairly precise idea of what the other sellers have paid. The state of the market in terms of the quantities available is thus not difficult to estimate. The only difficulty would come from demand side uncertainty. The stylised fact here would then be that prices for oranges and imported tomatoes would be fairly constant across the market whilst those for leeks and domestic tomatoes would vary at least at the beginning of the day. Two-sided learning is taking place on the market for domestic products whilst in the other markets, the learning is much more limited.

3.3. Seller's behaviour

For the reasons given we assume that the seller will modify his idea of the distribution of buyer's prices over the day. Given his view of this distribution he will choose the best price to offer to the buyer. Once he sees the reaction of the buyer he will revise his view of the distribution by truncating it at his first offer price. Of course if the buyer had a reservation price above the first offer price he would have accepted.
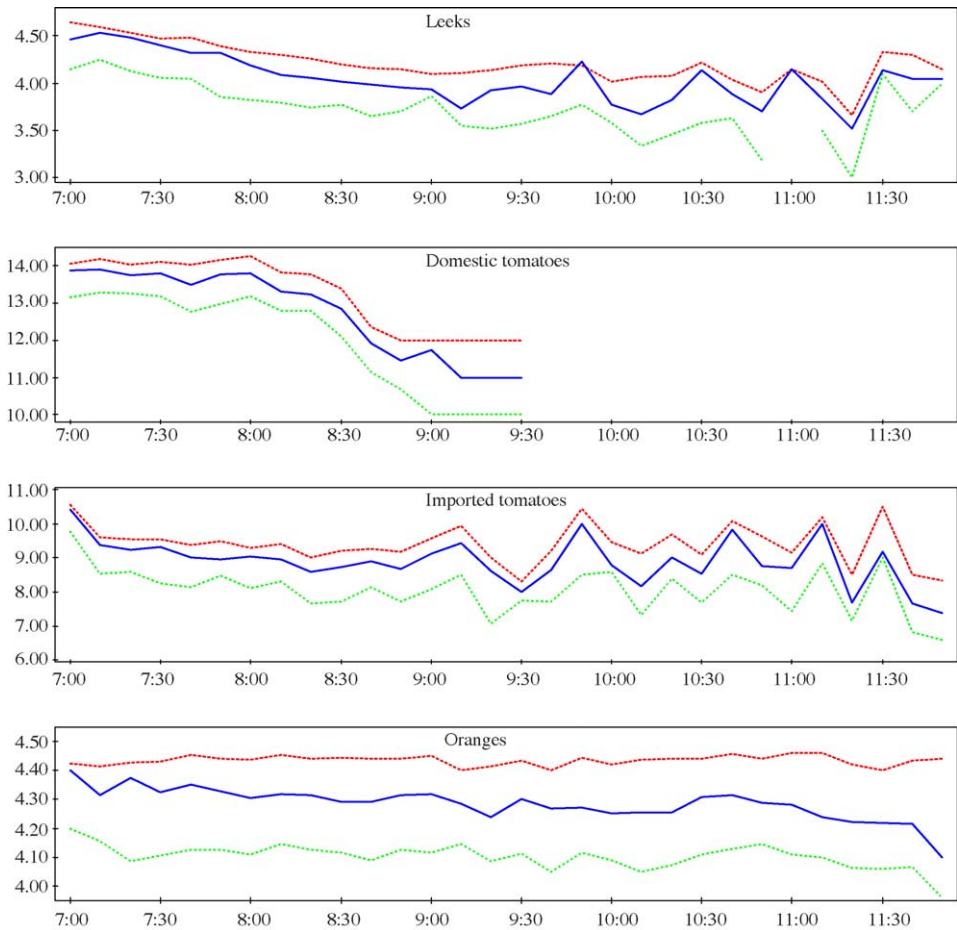
Fig. 2. Average offer (upper dashed), transaction price (middle solid), and counteroffer (lower dotted) for the four kinds of fruits and vegetables.

Assume that the seller's guess about the distribution of x (that is the reservation value of the buyer) is given by the following family of density functions

$$f(x) = \begin{cases} f_b + 2\dfrac{b-x}{b-a}\left(\dfrac{1}{b-a} - f_b\right) & \text{for } x \in [a, b], \\ 0 & \text{else}, \end{cases}$$

where

$$f_b \in \left[0, \frac{1}{b-a}\right],$$

is just $f(b)$. It is easy to check that the density function is a linear non-increasing function in x. If we set $f_b = 1/(b - a)$, then we obtain the density of the Uniform distribution. For $f_b = 0$, we obtain the density of the Triangular distribution. The corresponding distribution function is

$$F(x) = \begin{cases} 0 & \text{for } x < a, \\ 1 - \left[f_b(b - x) + \{1 - f_b(b - a)\} \left(\dfrac{b - x}{b - a} \right)^2 \right] & \text{for } x \in [a, b], \\ 1 & \text{for } x > b. \end{cases}$$

The seller chooses the offer price to maximize his expected gain

$$p^* = \arg\max(p - v)\{1 - F(p)\}. \tag{1}$$

Here, $v \in (a, b)$ denotes the reservation valuation of the seller. The first term is just his gain, given that he can strike the deal. The second term is the probability that the reservation value x of the buyer is above p. The offer price is implicitly given by $1 - F(p^*) = (p^* - v)f(p^*)$. For the second order condition

$$-2f(p^*) - (p^* - v)f_x(p^*) < 0, \tag{2}$$

to be fulfilled, it is sufficient that

$$p^* \leqslant \tfrac{1}{2}b + v, \tag{3}$$

holds. It can be shown that for our family of density functions, this inequality is always satisfied. How does the optimal offer price change when f_b increases? Total differentiation of the first order condition yields

$$\frac{\mathrm{d}p^*}{\mathrm{d}f_b} = \frac{F_{f_b}(p^*) + (p^* - v)f_{f_b}(p^*)}{-2f(p^*) - (p^* - v)f_x(p^*)} > 0. \tag{4}$$

For our family of density functions a higher f_b is accompanied by a lower probability mass of low reservation values and a higher probability mass for higher reservation values. A higher probability of good events makes the seller set a higher offer price. If we calculate the optimal price for the density with the lowest mass on high reservation values (the Triangular distribution), we obtain

$$p^*|_{f_b=0} = \tfrac{1}{3}b + 2v \, .$$

If we calculate the optimal price instead for the distribution that puts the largest mass on high reservation values (the Uniform distribution), we obtain

$$p^*|_{f_b=\frac{1}{b-a}} = \tfrac{1}{2}b + v.$$

Using (4), we obtain that

$$p^* \in \left[\frac{b + 2v}{3}, \frac{b + v}{2} \right], \tag{5}$$

where p^* is increasing in f_b. It is not difficult to prove that the seller always prefers higher values of f_b because this increases his expected gain. The increase is $-(p^* - v)F_{f_b}(p^*) > 0$ (using the first order condition) with

$$F_{f_b}(p^*) = -(b - p^*)\left(1 - \frac{b - p^*}{b - a}\right) < 0.$$

After observing the counteroffer c, the seller updates his guess about the distribution of buyer's reservation value x. Given the seller's first offer p^* and the buyer's counter offer c, the seller knows that $x \in [c, p^*]$. The conditional density is

$$f(x|c \leqslant x \leqslant p^*) = \begin{cases} D^{-1}\left\{ f_b + 2\dfrac{b - x}{b - a}\left(\dfrac{1}{b - a} - f_b\right)\right\} & \text{for } x \in [c, p^*], \\ 0 & \text{else}, \end{cases}$$

where the denominator is $D = F(p^*) - F(c)$. It is possible to rewrite the above given conditional density as

$$f(x|c \leqslant x \leqslant p^*) = \begin{cases} f_{p^*} + 2\dfrac{p^* - x}{p^* - c}\left(\dfrac{1}{p^* - c} - f_{p^*}\right) & \text{for } x \in [c, p^*], \\ 0 & \text{else}, \end{cases} \tag{6}$$

with $f_{p^*} = f(p^*|c \leqslant x \leqslant p^*)$. In that case, all results that we have derived for the optimal offer will also hold for the second offer given the family of density functions (6).

4. Empirical results

In this section, we study the movement of the seller's first and second price offers, p_1^* and p_2^*, over the course of the trading day. How does the price setting behavior of the seller evolve during the day in the light of his previous experiences? In particular, does he react to periods of inactivity, refusals, or counteroffers? Throughout we will use the theoretical results derived in the previous section to interpret the empirical results. Specifically, we employ the following variables empirically to describe the events occurring up to and at a particular point in time t during any given trading day as an (ex-post) *bargaining power index*

$$I_t \stackrel{\text{def}}{=} \frac{p_t - c_t}{p_{1,t}^* - c_t}, \tag{7}$$

the *relative change in the initial offer price*

$$\Pi_{\text{offer},t} \stackrel{\text{def}}{=} \frac{p^*_{1,t} - p^*_{1,t-1-\tau}}{p^*_{1,t-1-\tau}} \,, \tag{8}$$

the *relative deviation of the offer from the counteroffer*

$$\Delta_t \stackrel{\text{def}}{=} \frac{p^*_{1,t} - c_t}{p^*_{1,t}} \geqslant 0 \,, \tag{9}$$

the *number of succeeding intervals with no contact* (τ_t), and dummy variables describing the outcome at t ($d_{i,t}$). The ex-post measure of bargaining power I_t is observed if, at time t, the seller meets a buyer, and the two negotiate and strike a deal (outcome TWB in Fig. 1). It relates the difference between transaction price p (recall from the discussion of Fig. 1 that $p = p^*_2$ in this situation) and counteroffer c to the difference between the first offer p^*_1 and the counteroffer c. The index takes on values in the interval $I \in [0, 1]$ and attains its maximum if $p = p^*_1$ (when the seller prevails) and its minimum if $p = c$ (the buyer has his way). The relative change in the initial offer price Π is simply the percentage change in p^*_1 relative to the last time period ($t - 1 - \tau$, with $\tau \geq 0$) when p^*_1 was observed. The relative deviation of the offer from the counteroffer Δ measures how far apart seller and buyer have been initially if bargaining occurred.

4.1. Seller's first price

Starting with the seller's initial price offer p^*_1, the data shows no or very little movement during most days for oranges and domestic and imported tomatoes while in the market for leeks p^*_1 fluctuates during all 15 trading days, usually moving downwards over time as can be seen from the upper panel in Fig. 3. For this reason, we will mainly focus on the latter market in the remainder of the analysis.

To describe parsimoniously the (detrended) behavior of p^*_1 over time we use the "inflation"-type measure $\Pi_{\text{offer},t}$ defined in (8). Table 2 shows that, in most cases, $\Pi_{\text{offer},t}$ is equal to zero. Indeed, for imported tomatoes and oranges, there are only seven occasions of non-zero price movements during all 15 days.

We use linear regression to explain the behavior of $\Pi_{\text{offer},t}$ by variables capturing the preceding history of inactivity, negotiations and transactions. In the spirit of the theory of Section 3.3, the experience accumulated throughout the day leads the seller to update f_b,

Table 2
Summary statistics for the relative changes in current offer prices with respect to previously observed offer prices for the different products; see also Eq. (8).

	Mean	Std.	Number of observations		
			Total	With $\Pi = 0$	With $\Pi < 0$
Leeks	−0.37	1.35	568	497	65
Domestic tomatoes	−0.02	0.49	395	390	4
Imported tomatoes	0.00	0.64	462	455	4
Oranges	0.00	0.85	448	441	4
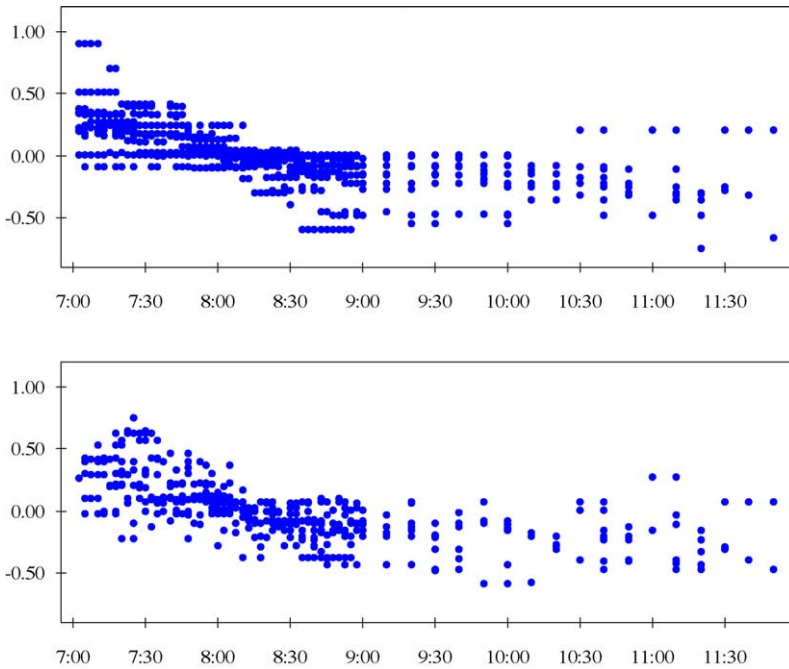
Fig. 3. Behaviour of initial offer prices (upper panel) and transaction prices (lower panel) for leeks over time as deviations from their daily average.

the parameter governing the shape of the distribution of buyers' reservation values that he assumes he is facing. For instance, periods of inactivity or series of non-cooperative buyers should lead him to decrease f_b and, by virtue of (4), lead to reductions in $p_{1,t}^*$. In the empirical model we use the number of preceding intervals with no contact, τ, to capture this effect, while $\Delta_{t-1-\tau_\Delta}$ and $I_{t-1-\tau_I}$ measure the influence of the last occurrences (prior to time t) of unsuccessful and successful negotiations, respectively (where the $t-1-\tau_\Delta$ and $t-1-\tau_I$ subscripts accommodate the possibility that these last negotiations may have occurred prior to period $t-1-\tau$). We also included dummy variables representing the outcome of the most recent meeting but only retained $d_{\mathrm{OO},t-1-\tau}$ (where the "OO" subscript employs the notation of Fig. 1 to indicate an "offer only" meeting). Thus, we arrived at the following regression model:

$$\Pi_{\mathrm{offer},t,\tau} = \beta_0 + \beta_1 \tau + \beta_2 \Delta_{t-1-\tau_\Delta} + \beta_3 \tilde{I}_{t-1-\tau_I} + \beta_4 d_{\mathrm{OO},t-1-\tau} + \varepsilon. \tag{10}$$

This specification focuses on the influence of the more recent events on the current change in the seller's offer price. Estimates of the coefficients $\beta_0, \ldots, \beta_4$ are given in Table 3.

The negative estimate of β_0 reflects the fact that, on average, p_1^* moves downward. The negative sign of the coefficient of τ implies that this downward tendency is accelerated by periods of no contact. Similarly, the further apart the seller and buyer have been during previous negotiations (i.e. the larger $\Delta_{t-1-\tau_\Delta}$), the greater the seller's subsequent reduction in the initial offer price. On the other hand, the positive estimate of β_3 implies that preceding

Table 3
Explaining the behavior of the relative change in the initial offer price for leeks with a linear regression

	Coefficient	*t*-Statistic	*P*-value
τ	−0.2418	−3.64	0.000
$\tilde{I}_{t-1-\tau_I}$	0.5620	3.42	0.001
$\Delta_{t-1-\tau_\Delta}$	−2.3689	−1.84	0.067
$d_{OO,t-1-\tau}$	−0.3205	1.91	0.058
Constant	−0.2955	−1.63	0.104
Diagnostics			
R^2	0.0622	$\bar{R}^2$	0.0547
F-statistic	8.3	*P*-value (*F*-statistic)	0.0000

Notes: 506 observations are included. Explanatory variables are the number τ of succeeding intervals with no contact, the transformed bargaining power index $\tilde{I}_{t-1-\tau_I}$ that measures the relative deviation between transaction, initial and counteroffer in the previously observed interval with bargaining, the relative deviation $\Delta_{t-1-\tau_\Delta}$ of the offer from the counteroffer in the previously observed interval with counteroffer, and $d_{OO,t-1-\tau}$ indicates whether the most recent meeting was an offer only.

negotiations resulting in a transaction price close to the seller's initial offer (i.e., if $\tilde{I}_{t-1-\tau_I}$ is large) slow down the seller's tendency to lower p_1^*. Finally, the estimated negative effect of $d_{OO,t-1-\tau}$ is evidence that an "offer only" meeting in the most recent past is prompting the seller to lower the offer price.

Summing up, the estimates in Table 3 suggest that the speed with which the seller reduces p_1^* during any given day is increased (decreased) if previous periods brought "bad news" ("good news"). In terms of the notation of the theory of Section 3.3, "bad news" translates into a downward revision of f_b, the parameter governing the shape of the distribution of buyers' reservation values. Given this downward revision of f_b, Eq. (4) shows that the rational response of the seller is to lower p_1^*.

4.2. Seller's second price

Turning to the seller's second price p_2^*, the lower panel in Fig. 3 shows a downward trend over time for leeks similarly to the one observed for the initial offer price, yet when we estimated a regression model analogous to (10) for the change in transaction prices, all coefficients were insignificant, except the one for $\tilde{I}_{t-1-\tau_I}$ which had the "wrong" (i.e. positive) sign. We conclude from these (unreported) results that there is a significant impact of the most recent past on the seller's initial offer price but not on his second price.

4.3. Bargaining index

From the discussion of Fig. 1 it follows that we observe the seller's second price only if a transaction occurs, in which case $p = p_2^*$. Hence, the bargaining index of Eq. (7) might be written in terms of p_2^* as

$$I_t = \frac{p_{2,t}^* - c_t}{p_{1,t}^* - c_t}. \tag{11}$$

Table 4 gives summary statistics for the bargaining index for each of the four products. For each product, sample mean and median are greater or equal to 0.5. Hence, averaged over all

Table 4
Summary statistics for the realizations of the bargaining power index for the different products

	Mean	Std.	Lower quartile	Median	Upper quartile	Number of observations
Leeks	0.67	0.35	0.50	0.71	1.00	406
Domestic tomatoes	0.66	0.31	0.50	0.50	1.00	387
Imported tomatoes	0.67	0.35	0.50	0.67	1.00	460
Oranges	0.57	0.25	0.40	0.60	0.75	442

Notes: The bargaining index measures the relative deviation between transaction price and counteroffer with respect to the difference between initial offer price and counteroffer; see also Eq. (7). The transaction price is missing for eight observations of transactions with bargaining, and we can calculate the index only for 1695 observations.

negotiations and regardless of day or time of day, the transaction price (that is seller's second price) tends to be closer to the seller's initial price offer than to the buyer's counteroffer.

The index distributions for each product have mass points at 0, 0.5 and 1. Hence, in most negotiations seller and buyer either meet exactly half way between $p_{1,t}^*$ and c_t or end up at either extreme.

How does I_t evolve during the average trading day? Pooling the data of all four products, Fig. 4 shows a kernel smooth of the index with confidence intervals at the 95% level (Härdle, 1990; Härdle et al., 2000). Hence, on average, the index declines during the course of the day with episodes of a particularly rapid decline at the beginning and end of the trading day.

How can this apparent decline in the seller's bargaining power be explained? Since our data provides plenty of information about the single seller but very little information about the many buyers he faces, we will try to explain the movement in I_t by movements in $p_{2,t}^*$. This line of reasoning is aided by the fact that in the data the denominator of I_t is fairly constant over time; see Fig. 2. We may therefore set $p_{1,t}^* - c_t = k$ and rewrite I_t as

$$I_t = 1 - \frac{p_{1,t}^* - p_{2,t}^*}{k}. \tag{12}$$

Given that the buyer makes a counteroffer c_t, it is reasonable that the second offer price will be in the interval $[c_t, p_{1,t}^*]$. If we assume that the buyers prepared to pay higher prices
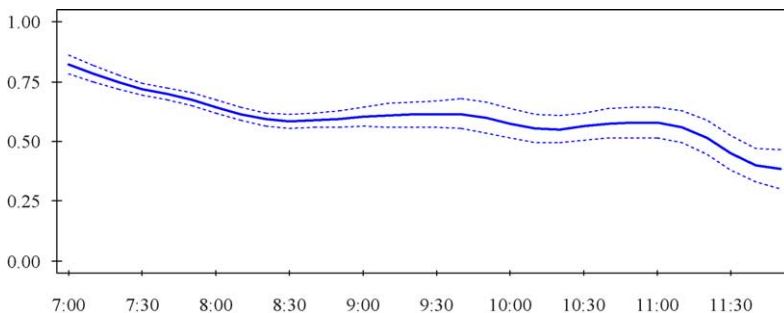
Fig. 4. Smoothed bargaining power index. The bargaining index measures the relative deviation between transaction price and counteroffer with respect to the difference between initial offer price and counteroffer; see also Eq. (7). Confidence intervals at the 95% level.

leave the market early and that the ultimatum game effect comes into play, then the model of Section 3.3 implies that the seller's guess about $f_{p_1^*}$ decreases during the course of day. If this results in an increase of $(p_{1,t}^* - p_{2,t}^*)$, then the index must be decreasing during the day. However, given our family of density functions, simulating the index reveals that the index *increases* over time given a constant reservation valuation v. To derive a decreasing index over time, we must allow furthermore that the reservation valuation of the seller decreases over time. It is important to mention that a decreasing v_t suffices to generate a downward sloping index curve.

5. Conclusion

In this paper we have analysed the effect of the bargaining process on prices. We have observations not only on the prices of transaction but also on offers and counteroffers even when no transactions took place. We first suggested a theoretical approach to the overall problem faced by the seller that we observed and then suggested a simple model to explain how the seller revised his beliefs over the day as he was faced with different buyers. The model predicted that proposed prices would decline over the day. More importantly we constructed an index of the seller's bargaining power, and the model predicted that this would decline over the day. The data from the wholesale fruit and vegetable market in Marseille was consistent with these predictions.

Appendix A

A.1. Additional tables and graphics

The following tables report the relative frequencies of the contact categories for the two segments of the observation grid. The first segment is from 7 to 9 o'clock and has 48 observations per product and day, whereas the second segment is from 9 to 12 o'clock and has only 18 observations. Table A.1 shows relative frequencies of the contact categories

Fig. A.1. Product leeks. Average number of contacts (upper solid) and negotiations (lower dashed) during the course of the day.

Table A.1
Observations from 7 to 9 o'clock: relative frequency of the categories for the different sorts of vegetables (%)

| | No contact | Contact | | | |
| | | | | Transaction | |
		OO	OC	TNB	TWB
Panel A: Relative frequencies of all five categories					
Leeks	34.44	13.61	6.81	14.17	30.97
Domestic tomatoes	31.99	2.86	0.84	18.18	46.13
Imported tomatoes	55.28	1.39	0.69	12.92	29.72
Oranges	57.64	1.53	1.11	6.67	33.06
Panel B: Relative frequencies conditional on a contact					
Leeks		20.76	10.38	21.61	47.25
Domestic tomatoes		4.21	1.24	26.73	67.82
Imported tomatoes		3.11	1.55	28.88	66.46
Oranges		3.61	2.62	15.74	78.03

Notes: The categories describe the way in which a meeting between seller and buyer in a time grid might unfold. If no buyer shows up, no contact happened. If a buyer shows up, there are four categories: offer only (OO), offer and counteroffer (OC), transaction with no bargaining (TNB), and transaction with bargaining (TWB).

Table A.2
Observations from 9 to 12 o'clock: relative frequency of the categories for the different sorts of vegetables (%)

| | No contact | Contact | | | |
| | | | | Transaction | |
		OO	OC	TNB	TWB
Panel A: Relative frequencies of different categories					
Leeks	59.22	6.67	3.53	6.67	23.92
Domestic tomatoes	98.98	0.00	0.00	0.00	1.02
Imported tomatoes	41.96	0.39	0.39	9.41	47.84
Oranges	39.61	0.00	0.39	1.96	58.04
Panel B: Relative frequencies conditional on a contact					
Leeks		16.35	8.65	16.35	58.65
Domestic tomatoes		0.00	0.00	0.00	100.00
Imported tomatoes		0.68	0.68	16.22	82.43
Oranges		0.00	0.65	3.25	96.10

Notes: The categories describe the way in which a meeting between seller and buyer in a time grid might unfold. If no buyer shows up, no contact happened. If a buyer shows up, there are four categories: offer only (OO), offer and counteroffer (OC), transaction with no bargaining (TNB), and transaction with bargaining (TWB).

for the first segment. Table A.2 shows relative frequencies of the contact categories for the second segment.

Another aspect is the evolution of the negotiations during the course of the day; see Figs. A.1–A.4. With respect to the relative frequency of meetings with counteroffer, the products are quite similar; see Figs. A.5–A.8 .
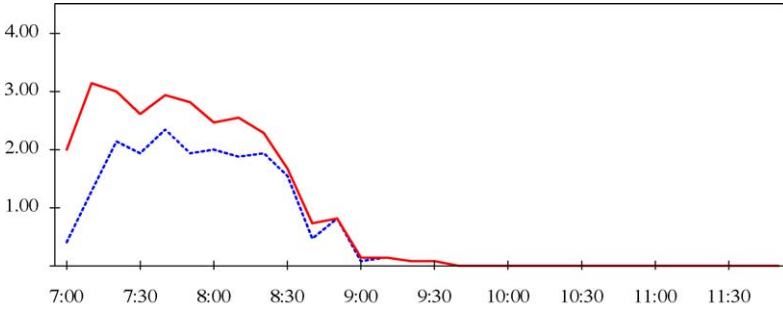
Fig. A.2. Product domestic tomatoes. Average number of contacts (upper solid) and negotiations (lower dashed) during the course of the day.
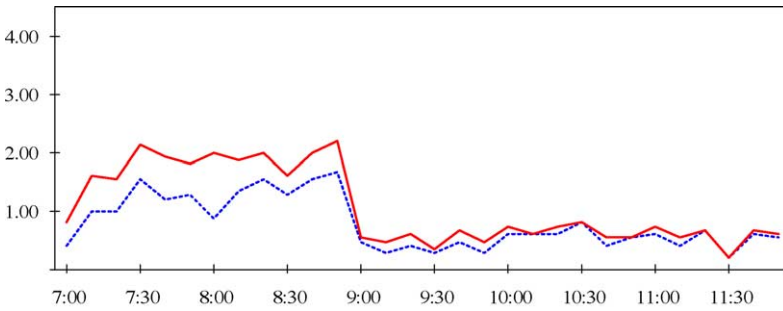
Fig. A.3. Product imported tomatoes. Average number of contacts (upper solid) and negotiations (lower dashed) during the course of the day.
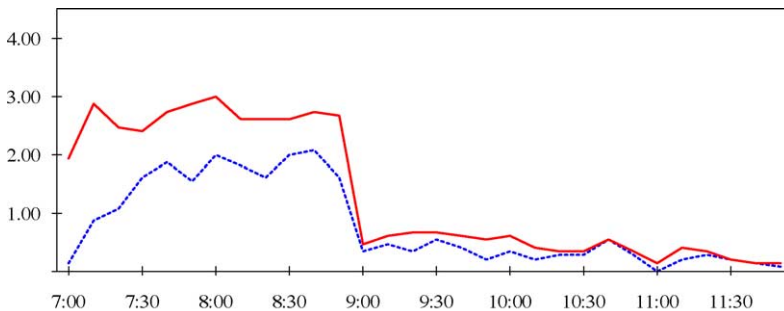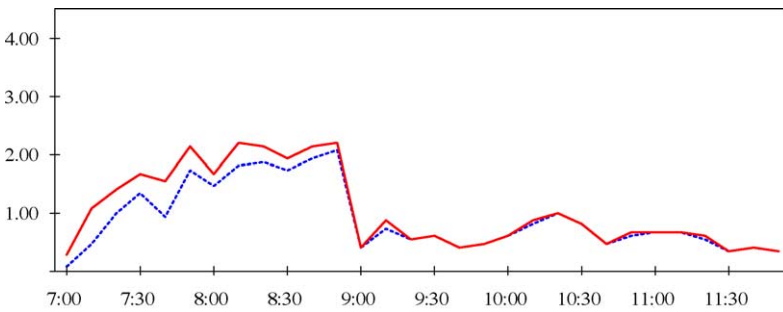

Fig. A.4. Product oranges. Average number of contacts (upper solid) and negotiations (lower dashed) during the course of the day.
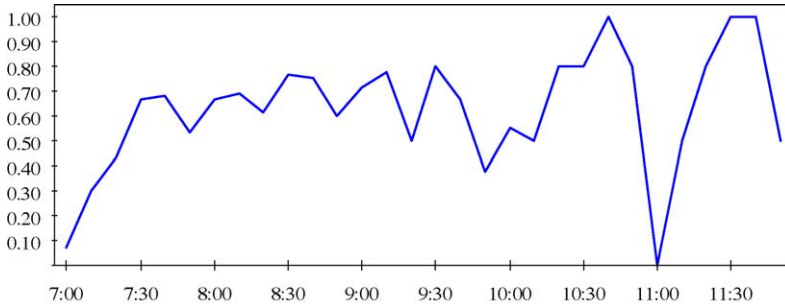
Fig. A.5. Product leeks. Fraction of meetings with counteroffers of buyers.
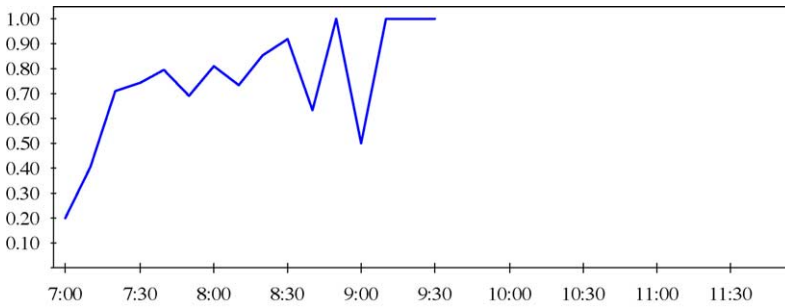
Fig. A.6. Product domestic tomatoes. Fraction of meetings with counteroffers of buyers.
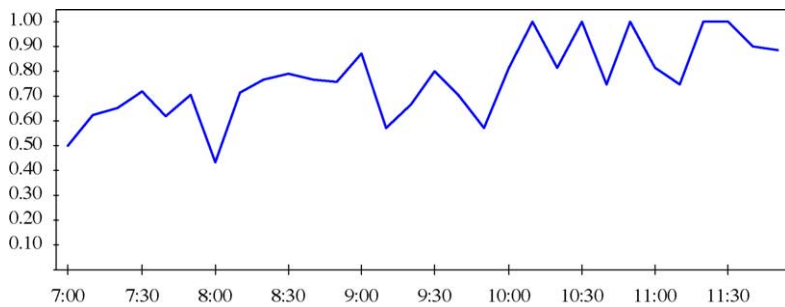
Fig. A.7. Product imported tomatoes. Fraction of meetings with counteroffers of buyers.
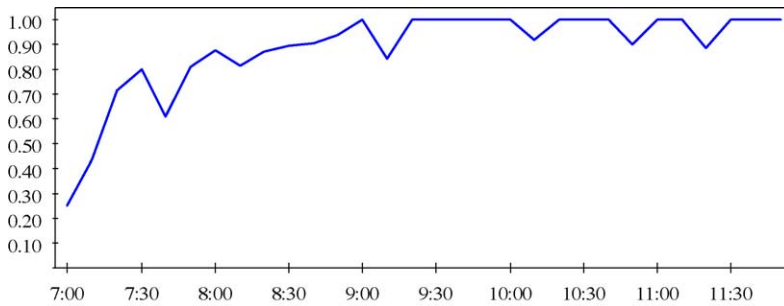
Fig. A.8. Product oranges. Fraction of meetings with counteroffers of buyers.

Table A.3 shows a contingency table for leeks.

Figs. A.9–A.12 show frequencies of the bargaining power index

$$I_t \overset{\text{def}}{\equiv} \frac{p_t - c_t}{p^*_{1,t} - c_t} \ . \tag{A.1}$$

The bargaining index measures the relative deviation between transaction price and counteroffer with respect to the difference between initial offer price and counteroffer.

A.2. Second order condition in 3.3.1

To see that

$$p^* \leqslant \tfrac{1}{2}b + v, \tag{A.2}$$

is sufficient, one just has to calculate the second order condition

$$-2f(p^*) - (p^* - v)f_x(p^*) < 0,$$

for our family of density functions. It is easy to see that the second order condition is fulfilled if (A.2) holds. To prove that the inequality for the optimal offer price always holds, write

Table A.3

Contingency table for leeks: relative price changes in initial offer prices and categories

			Transaction		Total
	OO	OC	TNB	TWB	
$\Pi < 0$	7.83	18.64	14.88	9.38	11.15
$\Pi = 0$	79.13	81.36	82.64	89.58	85.25
$\Pi > 0$	13.04	0.00	2.48	1.04	3.60
Total	100.00	100.00	100.00	100.00	100.00

Notes: If a buyer shows up, there are four categories: offer only (OO), offer and counteroffer (OC), transaction with no bargaining (TNB), and transaction with bargaining (TWB). The relative change Π is calculated with the current offer price and the previously observed offer price. Chi-square test of independence statistic with 6 degrees of freedom is 44.02, Pr = 0.000.
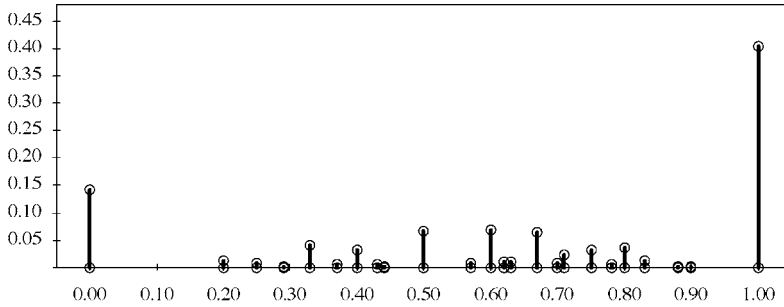
Fig. A.9. Frequency of the bargaining power index (A.1) for leeks. The bargaining index measures the relative deviation between transaction price and counteroffer with respect to the difference between initial offer price and counteroffer.
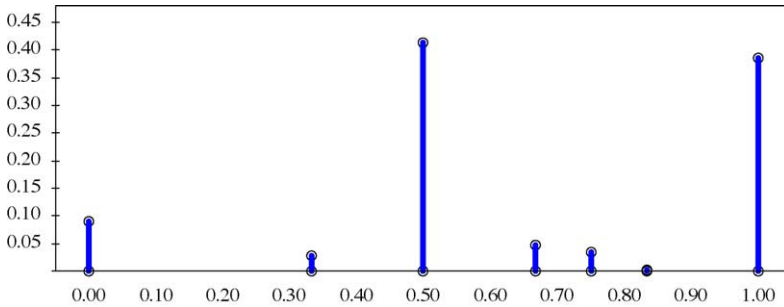
Fig. A.10. Frequency of the bargaining power index (A.1) for domestic tomatoes. The bargaining index measures the relative deviation between transaction price and counteroffer with respect to the difference between initial offer price and counteroffer.
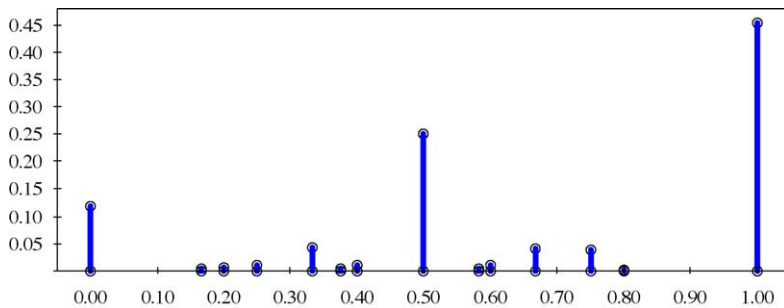
Fig. A.11. Frequency of the bargaining power index (A.1) for imported tomatoes. The bargaining index measures the relative deviation between transaction price and counteroffer with respect to the difference between initial offer price and counteroffer.
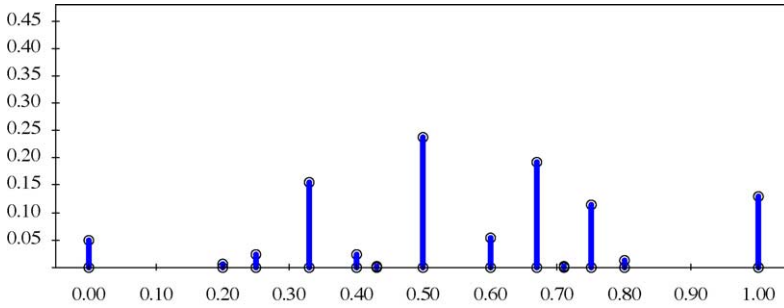
Fig. A.12. Frequency of the bargaining power index (A.1) for oranges. The bargaining index measures the relative deviation between transaction price and counteroffer with respect to the difference between initial offer price and counteroffer.
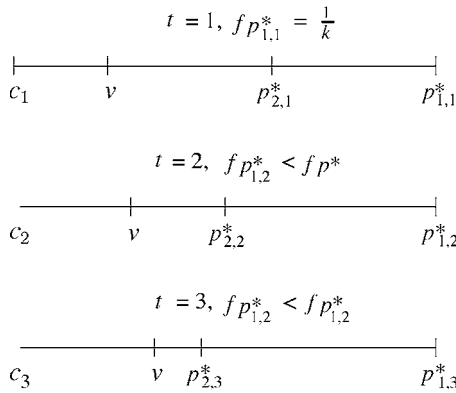
Fig. A.13. Relation between first and second price given a decreasing density f of buyers with high reservation prices. p_1^* is the seller's initial offer price, c the buyer's counteroffer, v the seller's constant reservation valuation, and p_2^* the transaction price. The figure illustrates the special case where the bargaining index goes down.
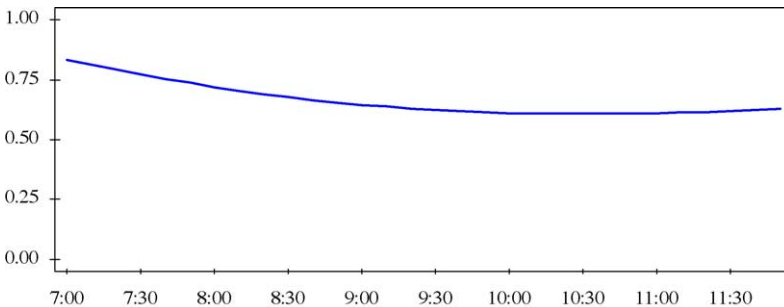
Fig. A.14. Simulated bargaining power index, where both f (that is the density of buyers with maximal reservation price) and seller's reservation valuation v decrease over time.

the first order condition as

$$f_b\{b + v - 2p^*\} + \frac{1 - f_b(b - a)}{(b - a)^2}\{(b + v - 2p^*)^2 - (p^* - v)^2\} = 0 \ .$$

Now, let us as assume firstly that

$$f_b \in \left(0, \frac{1}{b - a}\right) \ .$$

We show that no p with

$$p \in \left(\frac{b + v}{2}, b\right],$$

is a solution of the first order condition: when p is in this range, it follows immediately that the expression in the first curly brackets of the first order condition is negative. To have a solution, the expression in the second curly brackets must be positive. However, at the lower bound $(b + v)/2$ this expression is clearly negative. At the upper bound b it is zero. Is it possible that the expression will be positive for some intermediate values? In that case, the expression must be equal to zero at least once, but we easily see that the solution for

$$(b + v - 2p)^2 = (p - v)^2 \ \Leftrightarrow \ -(b + v - 2p) = p - v,$$

is $p = b$ (recalling that the expression on the RHS is non-positive and the expression on the LHS is non-negative). Now, what happens if $f_b = 1/(b - a)$? In that case, it is easy to see that $p^* = (b + v)/2$. Eventually, for $f_b = 0$ we have the two solutions b and $(b + 2v)/3$, where only the latter fulfils the second order condition. Thus, inequality (A.2) holds for p^*.

A.3. Total differential

To derive the sign of

$$\frac{\mathrm{d}p^*}{\mathrm{d}f_b} = \frac{F_{f_b}(p^*) + (p^* - v)f_{f_b}(p^*)}{-2f(p^*) - (p^* - v)f_x(p^*)} > 0 \ ,$$

note that the denominator is negative due to the second order condition. For the numerator we obtain after some manipulations the expression

$$-\left\{\left(1 - \frac{b - p^*}{b - a}\right)(b + v - 2p^*) + \frac{b - p^*}{b - a}(p^* - v)\right\} < 0 \ .$$

The term in squared brackets is positive: the optimal price is above a and less than b, so that the fraction is positive and less than one. We know that $b + v - 2p^* \geqslant 0$. The optimal offer price must always be above the reservation value v of the seller.

A.4. Conditional density function

First, we extend $b - x = (b - p) + (p - x)$ in the original variant of the conditional density, and we obtain after deleting all equivalent terms that all we have to show is

$$\frac{1 - (b - a)f_b}{(b - a)^2} = \frac{D\{1 - (p^* - c)f_{p^*}\}}{(p^* - c)^2}.$$

We show the equivalence of these two expressions by substituting for $D = F(p^*) - F(c)$ and by using

$$-2(b - p^*)(p^* - c) + (b - c)^2 - (b - p^*)^2 = (p - c)^2.$$

A.5. Bargaining power index

Fig. A.13 clarifies the argument for a decreasing $p_{1,t}^* - p_{2,t}^*$ through time. The total length of each interval is k. The reservation price v of the seller stays constant. We know that the first price goes down, and thus $p_{1,1}^* > p_{1,2}^* > p_{1,3}^*$. The first interval shows p_2^* for the upper bound of $f_{p_1^*}$. The second interval shows p_2^* for a lower value of $f_{p_1^*}$. In that case, we know with

$$\frac{\mathrm{d}p^*}{\mathrm{d}f_b} = \frac{F_{f_b}(p^*) + (p^* - v)f_{f_b}(p^*)}{-2f(p^*) - (p^* - v)f_x(p^*)} > 0,$$

and

$$p^* \in \left[\frac{b + 2v}{3}, \frac{b + v}{2}\right],$$

that p_2^* will be lower than $(p_1^* + v)/2$. If the difference between first price and v does not shrink too much, then the difference between first price and second price increases. The same lines of reasoning can be used for the third interval.

Fig. A.14 shows a simulated index where the reservation valuation decreases with a rate of 4.5% between succeeding intervals. Here, we have set $a = 0, b = 2, v_0 = 1.5$ and $k = 0.75$. The density of f_b starts with the density of the Uniform distribution (that is $f_b = 0.5$) and decreases over time, reaching at the end the density of the Triangular distribution with $f_b = 0$.

References

Ashenfelter, O.C., 1989. How auctions work for wine and art. Journal of Economic Perspectives 3 (3), 23–36.

Gastwirth, J.L., 1976. On probabilistic models of consumer search for information. Quarterly Journal of Economics 90, 38–50.

Grossman, S.J., Perry, M., 1986. Sequential bargaining under asymmetric information. Journal of Economic Theory 39, 120–154.

Härdle, W., 1990. Applied Nonparametric Regression. Cambridge University Press, Cambridge.

Härdle, W., Kirman, A., 1995. Nonclassical demand. A model-free examination of price-quantity relations in the Marseille fish market. Journal of Econometrics 67, 227–257.

Härdle, W., Klinke, S., Müller, M. (Eds.), 2000. XploRe – Learning Guide. Springer, Berlin, New York.

McAfee, R.P., Vincent, D., 1993. The declining price anomaly. Journal of Economic Theory 60, 191–212.

Myerson, R.B., Satterthwaite, M.A., 1983. Efficient mechanisms for bilateral trading. Journal of Economic Theory 29, 265–281.

Pezanis-Christou, P., 1997. Sequential auctions with supply uncertainty. Working Paper. Laboratory for Experimental Economics, University of Bonn.

Rothschild, M., 1974. Searching for the lowest price when the distribution of prices is unknown. Journal of Political Economy 82, 689–711.

Rubinstein, A., 1982. Perfect equilibrium in a bargaining model. Econometrica 50, 97–109.

Weisbuch, G., Kirman, A., Herreiner, D., 2000. Market organization and trading relationships. Economic Journal 110, 411–436.

Empirical Likelihood-based Dimension Reduction Inference for Linear Error-in-Responses Models with Validation Study

BY QIHUA WANG

Academy of Mathematics and System Science, Chinese Academy of Science
Beijing 100080, P.R. China and
Heilongjiang University, Harbin, P.R. China

AND

WOLFGANG HÄRDLE
Center for Applied Statistics and Economics, Humboldt-Universität zu Berlin
10178 Berlin, Germany

Abstract

In this paper, linear errors-in-response models are considered in the presence of validation data on the responses. A semiparametric dimension reduction technique is employed to obtain an asymptotic normality estimator of β, estimated empirical loglikelihoods and adjusted empirical loglikelihoods for the vector of regression coefficients and linear combinations of the regression coefficients, respectively. The estimated empirical log-likelihoods are shown to be asymptotically distributed as weighted sums of independent χ_1^2 and the adjusted empirical loglikelihoods are proved to be asymptotically distributed as standard chi-squares, respectively. A simulation study is conducted to compare the proposed methods in terms of coverage accuracies and average lengths of the confidence intervals.

Key Words: Confidence intervals; Error-in-response; Validation data.

1   Introduction

Let Y be a scalar response variable and X be a p-variate explanatory variable in regression. We consider the linear model

$$Y = \mathbf{X}^T \beta + e, \tag{1.1}$$

where $\beta = (\beta_1, \ldots, \beta_p)^T$ is a $p \times 1$ vector of unknown regression coefficients, ϵ is a random statistical error, and given X, the error $e = Y - X^T\beta$ are identically independently distributed.

The regression problems where some of the predictors are mearsured with error have been paid considerable attention and extensively studied. Excellent introductions to the area were provided by Fuller (1987). The problem with error-in-response has received less attention, mainly because the additive error model is assumed, and hence the problem can be handled with standard methodology (Buonaccorsi, 1996). In some practical problem, the exact measurement of Y is difficult, time consuming or expensive, but instead one can observe a surrogate $\tilde{Y}$ for Y, which is more easily obtained using some relatively easy methods or techniques. Generally, the relationship between the surrogate variables $\tilde{Y}$ and the true variables Y can be rather complicated compared to the classical additive error structure usually assumed (Fuller,1987). The additive error model is often not appropriate, and some authors (e.g., Buonaccorsi, 1996; Carroll and Stefenski, 1990; Pepe, 1992) have considered more complex measurement error models for either regression or the response. However, the resulting inferences could be sensitive to the assumed model. Actually, in many practical settings, it is even difficult to specify the relationship between true variables and their surrogated variables. That is, the most realistic situation may be that no error model structure or distribution assumption of true variables given the surrogate variables is specified. However, this situation presents serious difficulties towards obtaining correct statistical analysis. Biases caused by

1

measurement errors could be difficult to access accurately without extra observations and information. One of solutions is to use the help of validation data. Some examples where validation data are available can be found in Wittes, Lakatos and Probstfied (1989), Duncan and Hill (1985) and Pepe (1992) among others. With help of validation data, some statisticians developed statistical inference techniques based on primary observations without specifying any error structure and the distribution assumption of the true variable given the surrogate variable (See, for example, Stefanski and Carroll (1987), Carroll and Wand (1991), Pepe and Fleming (1991), Pepe (1992), Pepe *et al* (1994), Reilly and Pepe (1995), Sepanski and Lee (1995), Wang (1999,2000) and Wang and Rao (2002) among others).

For model (1.1), we consider settings where no error equation or distribution assumption of Y given $\widetilde{Y}$ is specified, but some validation data are available to relate Y and $\widetilde{Y}$. With help of validation data, we define the estimator of β and develop empirical likelihood inference for β and its linear combinations. To use the surrogate $\widetilde{Y}'s$, let us rewrite the model (1.1) such that $\widetilde{Y}$ is related to X. Notice that $\widetilde{Y}$ and X provide useful information in predicting the unknown Y. We rewrite the model (1.1) as

$$u(Z) = \mathbf{X}^T \beta + \epsilon \qquad (1.2)$$

where $Z = (\widetilde{Y}, \mathbf{X})$, $u(z) = E[Y|Z = z]$ and $\epsilon = e - (Y - u(Z))$. If $u(\cdot)$ was a known function , (1.2) is then a standard statistical model and hence standard statistical inference approaches such as the least square and empirical likelihood due to Owen (1988, 1991) for linear model can be applied to inference for β or linear combinations of β from the primary data. Usually, $u(\cdot)$ is unknown. Hence, the LSE and empirical log-likelihood functions contain unknown $u(\cdot)$. A natural method is to replace $u(\cdot)$ in the LSE and empirical log-likelihood functions by an estimator of $u(\cdot)$ and define a final estimator of β and estimated empirical log-likelihood functions. Here, we estimate $u(\cdot)$ by kernel regression approach. But, such a method would require a

2

large validation data set, which is difficult or expensive to obtain in order to be feasible because of the use of kernel regression with $p + 1$ dimension explanatory variables Z. That is, "curse of dimension" will limit the usefulness of this method. This motivates us to use dimension reduction technique by assuming

$$u(z) = m(\alpha^T z), \tag{1.3}$$

where $m(\cdot)$ is an unknown function and α is a $(p + 1) \times 1$ vector of unknown parameter. This assumption is reasonable in many applications. It applies at least to generalized linear models and can be expected to apply in a variety of situation. α can be first estimated by sliced inverse regression (SIR) techniques (see, e.g., Li (1991), Duan and Li (1991) and Zhu and Fang (1996)). Then, we can estimate $u(\cdot)$ by the kernel regression with univariate explanatory variable with validation data. We will prove that the resulting estimator of β is asymptotically normal and the estimated empirical log-likelihood functions for β and its linear combinations are asymptotically weighted sums of independent χ_1^2 variables with unknown weights. As a result, they cannot be applied directly to construct confidence regions for β. To overcome this difficulty, several different methods may be used. In the first method, the unknown weights are estimated consistently so that the distribution of the estimated weighted sums of chi-squared variables can be calculated from the data. In the second method, the estimated empirical loglikelihood functions are adjusted so that the resulting adjusted empirical loglikelihood functions are asymptotically distributed as standard chi-squares.

This paper is organized as follows. In Section 2, we define a modified LSE with asymptotic normality. In Section 3, we define an estimated empirical loglikelihood and an adjusted empirical loglikelihood for β, and show that the estimated empirical loglikelihood is asymptotically distributed as a weighted sum of indepedent χ^2 and the adjusted empirical loglikelihood is asymptotically distributed as a standard chi-square. In Section 4, we define an estimated empirical loglikelihood and an adjusted

3

empirical loglikelihood for linear combinations of β and state their asymptotic results similar to those in Section 4. In Section 5, we report some Monte Carlo simulation results for the finite sample performance of the proposed approaches. The appendix presents the proofs of the main results.

2 Estimation

Assume we have a validation data set containing n independent and identically distributed observations of $\{(\widetilde{Y}_i, Y_i, \mathbf{X}_i)_{i=1}^n\}$ and a primary data set containing N independent and identically distributed observations of $\{(\widetilde{Y}_j, \mathbf{X}_j)_{j=n+1}^{n+N}\}$. The primary data set is independent of the validation data set. If $u(\cdot)$ was known, the LSE for β with the primary data can be defined to be

$$
\widetilde{\beta}_N = \left(\sum_{j=n+1}^{n+N} \mathbf{X}_j \mathbf{X}_j^T \right)^{-1} \sum_{j=n+1}^{n+N} \mathbf{X}_j u(Z_j),
$$

Usually, $u(\cdot)$ is unknown. Naturally, we may define a final estimator of β by replacing $u(\cdot)$ in $\widetilde{\beta}_N$ with its estimator. Next, we define an estimator of $u(\cdot)$ based on the dimension reduction model (1.3).

Denote $\mathbf{X} = (X_1, X_2, \cdots, X_p)$, $R(Y) = (R_1(Y), \cdots, R_{p+1}(Y))^T = (E[\widetilde{Y}|Y], E[X_1|Y], \cdots, E[X_p|Y])^T$, $\Lambda = Cov(R(Y)) = Cov(E[Z|Y])$. Denote by Z_{ij} the jth component of Z_i for $i = 1, 2, \cdots, n$ and $j = 1, 2, \cdots, p+1$. Let

$$
R_{nj}^*(y) = \frac{1}{nh_{1,n}} \sum_{i=1}^n Z_{ij} K_1\left(\frac{y - Y_i}{h_{1,n}} \right), j = 1, 2, \cdots, p+1
$$

and

$$
\widehat{f}_n(y) = \frac{1}{nh_{1,n}} \sum_{i=1}^n K_1\left(\frac{y - Y_i}{h_{1,n}} \right),
$$

where $K_1(\cdot)$ is a kernel function and $h_{1,n}$ is a bandwidth. For each fixed $b > 0$, let

$$
\begin{aligned}
\widehat{f}_{nb}(y) &= max(\widehat{f}_n(y), b) \\
\widehat{R}_{nb}(y) &= \left(\frac{R_{nj}^*(y)}{\widehat{f}_{nb}(y)} \right)_{(p+1) \times 1}
\end{aligned}
$$

4

and

$$\widehat{\Lambda}_n = \frac{1}{n} \sum_{j=1}^n (\widehat{R}_{nb}(Y_j))(\widehat{R}_{nb}(Y_j))^T - \left(\frac{1}{n} \sum_{j=1}^n \widehat{R}_{nb}(Y_j) \right) \left(\frac{1}{n} \sum_{j=1}^n \widehat{R}_{nb}(Y_j) \right)^T$$

Let α_n be the eigenvector corresponding to the maximum eigenvalue of $\widehat{\Lambda}_n$. By Zhu and Fang (1996), we can estimate α by α_n. Then, $u(z) = m(\alpha^T z)$ can be estimated by

$$\widehat{u}_n(z) = \frac{\sum_{i=1}^n K_2 \left(\frac{\alpha_n^T (z - Z_i)}{h_{2,n}} \right) Y_i}{\sum_{i=1}^n K_2 \left(\frac{\alpha_n^T (z - Z_i)}{h_{2,n}} \right)} \tag{2.1}$$

where $h_{2,n}$ is a bandwidth and $K_2(\cdot)$ is a kernel function. To avoid technical difficulties due to small values in the denominator of $\widehat{u}_n(\cdot)$, we define a truncation version of $\widehat{u}_n(\cdot)$.

Let $\widehat{f}_{n,z}(z) = (n_2 h_{2,n})^{-1} \sum_{i=1}^n K_2 \left(\frac{\alpha_n(z - Z_i)}{h_{2,n}} \right)$ and $\widehat{f}_{\tau_n, z}(z) = \max\{\widehat{f}_{n,z}(z), \tau_n\}$ for some positive constant sequence τ_n tending to zero. The truncated version of $\widehat{u}_n(z)$ is then defined by

$$\widehat{u}_{\tau_n}(z) = \frac{\widehat{u}_n(z) \widehat{f}_{n,z}(z)}{\widehat{f}_{\tau_n, z}(z)}.$$

We then can define a final estimator of β, $\widehat{\beta}_{nN}$ say, by replacing $u(\cdot)$ in $\widetilde{\beta}_N$ with $\widehat{u}_{\tau_n}(x)$. That is,

$$\widehat{\beta}_{n,N} = \left(\sum_{j=n+1}^{n+N} \mathbf{X}_j \mathbf{X}_j^T \right)^{-1} \sum_{j=n+1}^{n+N} \mathbf{X}_j \widehat{u}_{\tau_n}(Z_j).$$

Let $\Sigma = E\mathbf{X}\mathbf{X}^T$ and $V_1(\beta) = E[(u(Z) - \mathbf{X}^T \beta)^2 \mathbf{X}\mathbf{X}^T] + \lambda E[(Y - u(Z))^2 \mathbf{X}\mathbf{X}^T]$, where $\lambda = \frac{N}{n}$.

THEOREM 2.1 Under all the assumptions listed in the Appendix, we have

$$\sqrt{N}(\widehat{\beta}_{n,N} - \beta) \xrightarrow{\mathcal{L}} N(0, V(\beta)),$$

where $V(\beta) = \Sigma^{-1} V_1(\beta) \Sigma^{-1}$.

REMARK 2.1 The first term in the asymptotic covariance of $\widehat{\beta}_{n,N}$ is the contribution of the primary data, the Fisher information for β in the primary sample by

the regression relationship between $u(z)$ and X. The second term represents the extra cost due to estimation of unknown $u(Z)$.

REMARL 2.2 The asymptotic covariance of $\widehat{\beta}_{n,N}$ can be estimated consistently by

$$V_{n,N} = \Sigma_{n,N}^{-1} \widehat{V}_1(\widehat{\beta}_{n,N}) \Sigma_{n,N}^{-1}$$

where

$$\Sigma_{n,N} = \frac{1}{N} \sum_{j=n+1}^{n+N} \mathbf{X}_j \mathbf{X}_j^T$$

and

$$\widehat{V}_1(\widehat{\beta}_{n,N}) = \frac{1}{N} \sum_{j=n+1}^{n+N} [(\widehat{u}_{\gamma_n}(Z_j) - \mathbf{X}_j^T \widehat{\beta}_{n,N})^2 \mathbf{X}_j \mathbf{X}_j^T] + \frac{N}{n^2} \sum_{i=1}^{n} [(Y_i - \widehat{u}_{\gamma_n}(Z_i))^2 \mathbf{X}_i \mathbf{X}_i^T].$$

REMARK 2.3 To use information sufficiently, one may define the estimator of β to be

$$\widetilde{\beta}_{n,N} = \widetilde{\Sigma}_{n,N}^{-1} \widetilde{A}_{n,N},$$

where $\widetilde{\Sigma}_{n,N} = \frac{1}{n} \sum_{i=1}^{n} \mathbf{X}_i \mathbf{X}_i^T + \frac{1}{N} \sum_{k=n+1}^{n+N} \mathbf{X}_k \mathbf{X}_k^T$ and $\widetilde{A}_{n,N} = \frac{1}{n} \sum_{i=1}^{n} \mathbf{X}_i Y_i + \frac{1}{N} \sum_{k=n+1}^{n+N} \mathbf{X}_k u_{\tau_n}(Z_k)$. In most applications, however, the primary data set is large relative to the validation data, i.e., λ is large. For example, in the nurses health study described by Rosner *et al* (1989), $\lambda = 517.6$. In such cases, there is little information about β in the validation data set, and there will be little difference between $\widetilde{\beta}_{n,N}$ and $\widehat{\beta}_{n,N}$. On the other hand, It is much simpler to consider $\widehat{\beta}_{n,N}$. For these reasons, we consider $\widehat{\beta}_{n,N}$ only.

3 Estimated and adjusted empirical likelihood for β

We first give some motivations for defining an estimated empirical likelihood. Let $A_j(\beta) = \mathbf{X}_j(u(Z_j) - \mathbf{X}_j^T \beta)$. Then, we have $EA_j(\beta) = 0$. Let F_p be the distribution function which assigns probability p_j at point $(\mathbf{X}_j, \widetilde{Y}_j)$, respectively, for $j = n+$

$1, \cdots, n + N$. Then, we may define the empirical loglikelihood, evaluated at β, as

$$l_N(\beta) = -2 \max \sum_{j=n+1}^{n+N} \log(Np_j),$$

where the maximum is over $\sum_{j=n+1}^{n+N} p_j A_j(\beta) = 0$ and $\sum_{j=n+1}^{n+N} p_j = 1$. If β is the true parameter, then $l_N(\beta)$ can be shown to be asymptotically distributed as a standard χ^2 with p degrees of freedom. However, this result cannot be used to make inference about β because $l_N(\beta)$ contains the unknown terms $u(Z_j)$, and hence β is not identifiable. Naturally, we replace $u(\cdot)$ in $l_N(\beta)$ by an estimator of it and define an estimated empirical log-likelihood, $\hat{l}_N(\beta)$. Here, we replace $u(\cdot)$ in $l_N(\beta)$ with $\hat{u}_{\tau_n}(\cdot)$ and define an estimated empirical log-likelihood by

$$\hat{l}_N(\beta) = -2 \max_{\sum_{j=n+1}^{n+N} p_j \hat{A}_j(\beta)=0} \sum_{j=n+1}^{n+N} (NP_j). \tag{3.1}$$

where $\hat{A}_j(\beta)$ is $A_j(\beta)$ with $u(\cdot)$ replaced with $\hat{u}_{\tau_n}(\cdot)$.

By using the Lagrange multiplier method, the optimal values of $p'_j s$ satisfying (3.1) can be shown to be

$$p_j = \frac{1}{N} \frac{1}{1 + \lambda^T \hat{A}_j(\beta)},$$

where λ is the solution of the equation

$$\frac{1}{N} \sum_{j=n+1}^{n+N} \frac{\hat{A}_j(\beta)}{1 + \lambda^T \hat{A}_j(\beta)} = 0.$$

This yields

$$\hat{l}_{n,N}(\beta) = 2 \sum_{j=n+1}^{n+N} \log\{1 + \lambda^T(\hat{u}_{\tau_n}(Z_j) - \mathbf{X}_j^T \beta)\}.$$

Let $V_0(\beta) = E[\mathbf{X}\mathbf{X}^T(u(Z) - \mathbf{X}^T\beta)^2]$.

THEOREM 3.1 Under the assumptions given in the Appendix, if β is the true parameter, we have

$$\hat{l}_{n,N}(\beta) \xrightarrow{\mathcal{L}} \sum_{i=1}^{p} w_i \chi^2_{1,i},$$

where $\chi^2_{1,i}$ for $1 \leq i \leq p$ are independent χ^2_1 variables and $w_1, w_2, \cdots, w_p$ are the eigenvalues of $D(\beta) = V_0^{-1}(\beta) V_1(\beta)$ with $V_1(\beta)$ defined in Theorem 2.1.

7

To apply Theorem 3.1 to construct a confidence region or inteval for β, we must estimate the unkown weights w_i consistently. By the "plug in" method, $V_1(\beta)$ and $V_0(\beta)$ can be estimated consistently by $\widehat{V}_1(\widehat{\beta}_{n,N})$, which is defined in Section 2, and

$$\widehat{V}_0(\widehat{\beta}_{n,N}) = N^{-1} \sum_{j=n+1}^{n+N} [\mathbf{X}_j \mathbf{X}_j^T (\widehat{u}_{\tau_n}(Z_j) - \mathbf{X}_j^T \widehat{\beta}_{n,N})^2]$$

respectively. This implies that the eigenvalues of $\widehat{D}(\widehat{\beta}_{n,N}) = \widehat{V}_0^{-1}(\widehat{\beta}_{n,N})\widehat{V}_1(\widehat{\beta}_{n,N})$, $\widehat{w}_i$ say, estimate w_i consistently for $i = 1, 2, \cdots, p$. Let $\widehat{c}_\alpha$ be the $1 - \alpha$ quantile of the conditional distribution of the weighted sum $\widehat{S} = \widehat{w}_1 \chi_{1,1}^2 + \cdots + \widehat{w}_p \chi_{1,p}^2$ given the data. Then the confidence region for β with asymptotically correct coverage probability $1 - \alpha$ can be defined to be

$$\widehat{I}_\alpha(\widetilde{\beta}) = \{\widetilde{\beta} : \widehat{l}_{n,N}(\widetilde{\beta}) \leq \widehat{c}_\alpha\}.$$

In practice, the conditional distribution of the weighted sum $\widehat{S}$ given the data $\{(\mathbf{X}_i, Y_i, \widetilde{Y}_i)_{i=1}^n\}$ and $\{(\mathbf{X}_j, \widetilde{Y}_j)_{j=n+1}^{n+N}\}$ can be obtained using Monte Carlo simulations by repeatedly generating independent samples $\chi_{1,1}^2, \cdots, \chi_{1,d}^2$ from the χ_1^2 distribution.

In the absence of measurement error, $D(\beta)$ reduces to an identity matrix so that $w_i = 1$ for $1 \leq i \leq p$ and Theorem 3.1 reduces to Owen's (1991) result that the empirical loglikelihood is asymptotically χ_1^2. Next, we define an adjusted empirical log-likelihood whose asymptotic distribution is a standard chi-square.

Let

$$\widehat{H}(\beta) = \{\sum_{j=n+1}^{n+N} \widehat{A}_j(\beta)\}\{\sum_{j=n+1}^{n+N} \widehat{A}_j(\beta)\}^T.$$

By examing the asymptotic expansion of $\widehat{l}_{n,N}(\beta)$, we define an adjusted empirical loglikelihood by

$$\widehat{l}_{ad}(\beta) = \widehat{r}(\beta)\widehat{l}_{n,N}(\beta), \tag{3.1}$$

which can be proved to be asymptotically χ_d^2, where

$$\widehat{r}(\beta) = \frac{\text{tr}(\widehat{V}^{-1}(\beta)\widehat{H}(\beta))}{\text{tr}(\widehat{V}_0^{-1}(\beta)\widehat{H}(\beta))}.$$

8

THEOREM 3.2. Under the regularity conditions given in the appendix, if β is the true value of the parameter, we have

(a) as $n \to \infty$

$$\hat{l}_{ad}(\beta) \xrightarrow{\mathcal{L}} \chi_p^2,$$

where χ_p^2 is a standard chi-square random variable with p degrees of freedom.

(b)

$$\lim_{n \to \infty} P(\beta \in \hat{I}_{ad,\alpha}(\tilde{\beta})) = 1 - \alpha,$$

where $\hat{I}_{ad,\alpha}(\tilde{\beta}) = \{\tilde{\beta} : \hat{l}_{ad}(\tilde{\beta}) \leq \chi_{p,\alpha}^2\}$ with $\chi_{p,\alpha}^2$ the $1-\alpha$ quantile of the χ_p^2 distribution for $0 < \alpha < 1$.

4 Estimated and adjusted empirical likelihoods for linear combinations of β

In practice, statisticians are often confronted with the problem of constructing confidence intervals or regions for a particular regression coefficient, a subvector of β or linear combinations of β. To address this problem, we develop empirical likelihood method to make inference for a vector of linear combinations $\theta = C\beta$ of β, where $C = (C_1, C_2)$, C_1 is a $k \times k$ matrix and C_2 is a $k \times (p-k)$ matrix. For example, θ is the subvector of the first k regression coefficients if $C_1 = I_k$ and $C_2 = 0$. If $k = 1$, then θ reduces to a single linear combination, which includes an individual regression coefficients and the mean response at a given X level as special cases. Without loss of generality, we assume that C_1^{-1} exists.

Let $\gamma = (\theta^T, \beta_{0(k)}^T)^T$, where $\beta_{0(k)}$ denotes the column subvector of the last $p-k$ elements of β. Let $\mathbf{X}_i = (\mathbf{X}_{i1}^T, \mathbf{X}_{i2}^T)^T$, where $\mathbf{X}_{i1}$ and $\mathbf{X}_{i2}$ are $k \times 1$ and $(p-k) \times 1$ subvectors. Let $\widetilde{\mathbf{X}}_i = (\widetilde{X}_{i1}^T, \widetilde{X}_{i2}^T)^T = (\mathbf{X}_{i1}^T C_1^{-1}, \mathbf{X}_{i2}^T - \mathbf{X}_{i1}^T C_1^{-1} C_2)^T$. Then, model (1.2) reduces to

$$u(Z_j) = \widetilde{\mathbf{X}}_j^T \gamma + \epsilon, j = n+1, \cdots, n+N.$$

9

If $u(\cdot)$ was known, the LSE of γ can be defined to be $\widetilde{\gamma}_{n,N} = (\sum_{j=n+1}^{n+N} \widetilde{\mathbf{X}}_j \widetilde{\mathbf{X}}_j^T)^{-1}$ $(\sum_{j=n+1}^{n+N} \widetilde{\mathbf{X}}_j u(Z_j))$. Let $\widetilde{\beta}_{n(k)}$ denote the subvector of the last $p-k$ elements of $\widetilde{\gamma}_{n,N}$. We have

$$E\{\widetilde{X}_{j1}(u(Z_j) - \widetilde{X}_{j1}^T\theta - \widetilde{X}_{j2}^T\widetilde{\beta}_{n(k)}\} = 0, j = n+1, \cdots, n+N.$$

Let $\widehat{\gamma}_{n,N}$ be $\widetilde{\gamma}_{n,N}$ with $u(\cdot)$ replaced by $\widehat{u}_{\tau_n}(\cdot)$. Let $\widehat{\beta}_{n(k)}$ denote the subvector of the last $p-k$ elements of $\widehat{\gamma}_{n,N}$. Similar to the previous section, for a given θ, we introduce the following auxiliary variables

$$\widehat{W}_j(\theta) = \widetilde{X}_{j1}(\widehat{u}_{\tau_n}(Z_j) - \widetilde{X}_{j1}^T\theta - \widetilde{X}_{j2}^T\widehat{\beta}_{n(k)})$$

and define an estimated empirical log-likelihood function

$$\widetilde{l}_{n,N}(\theta) = 2 \sum_{j=n+1}^{n+N} \log(1 + \zeta^T\widehat{W}_j(\theta)),$$

where ζ is the solution of the following equation

$$\sum_{j=n+1}^{n+N} \frac{\widehat{W}_j(\theta)}{1 + \zeta^T\widehat{W}_j(\theta)} = 0.$$

Let $\widetilde{\mathbf{X}} = (\widetilde{X}_1^T, \widetilde{X}_2^T)^T$ and

$$
\begin{aligned}
K &= E(\widetilde{X}_1 \widetilde{X}_2^T) \\
P &= E(\widetilde{X}_2 \widetilde{X}_2^T) \\
\eta &= \widetilde{X}_1 - E(\widetilde{X}_1 \widetilde{\mathbf{X}}^T)\{E[\widetilde{\mathbf{X}}\widetilde{\mathbf{X}}^T]\}^{-1}\widetilde{\mathbf{X}} \\
&\quad + E[\widetilde{X}_1 \widetilde{X}_1^T]\{E(\widetilde{X}_1 \widetilde{X}_1^T) - K^T P^{-1} K\}^{-1}(\widetilde{X}_1 - K^T P^{-1} \widetilde{X}_2), \\
V_0^*(\theta) &= E[\widetilde{X}_1 \widetilde{X}_1^T(u(Z) - \widetilde{X}_1^T\theta - \widetilde{X}_2^T\beta_{(k)})^2], \\
V_1^*(\theta) &= E[(u(Z) - \widetilde{X}_1^T\theta - \widetilde{X}_2^T\beta_{(k)})^2\eta\eta^T]] + \lambda E[(Y - u(Z))^2\eta\eta^T]
\end{aligned}
$$

THEOREM 4.1 Under the assumptions listed in the Appendix, we have

$$\widetilde{l}_{n,N}(\theta) \xrightarrow{\mathcal{L}} \sum_{i=1}^{k} \widetilde{w}_{1,i}\chi_{1,i},$$

10

where $\widetilde{w}_{1,i}, 1 \le i \le k$ are the eigenvalues of $\widetilde{V}_0^{-1}(\theta)\widetilde{V}_1(\theta)$ and $\chi_{1,i}^2$ is independent standard χ_1^2 variables for $i = 1, 2, \cdots, k$.

Let

$$
\begin{aligned}
K_N &= \frac{1}{N} \sum_{j=n+1}^{n+N} \widetilde{X}_{j1} \widetilde{X}_{j2}^T \\
P_N &= \frac{1}{N} \sum_{j=n+1}^{n+N} \widetilde{X}_{j2} \widetilde{X}_{j2}^T \\
\eta_j &= \widetilde{X}_{j1} - \left(\frac{1}{N} \sum_{j=n+1}^{n+N} \widetilde{X}_{j1} \widetilde{\mathbf{X}}_j^T \right) \left(\frac{1}{N} \sum_{j=n+1}^{n+N} \widetilde{\mathbf{X}}_j \widetilde{\mathbf{X}}_j^T \right)^{-1} \widetilde{\mathbf{X}}_j \\
&\quad + \left(\frac{1}{N} \sum_{j=n+1}^{n+N} \widetilde{X}_{j1} \widetilde{X}_{j1}^T \right) \left(\frac{1}{N} \sum_{j=n+1}^{n+N} \widetilde{X}_{j1} \widetilde{X}_{j1}^T - K_N^T P_N^{-1} K_N \right)^{-1} \left(\widetilde{X}_{j1} - K_N^T P_N^{-1} \widetilde{X}_{j2} \right), \\
\widehat{V}_0^*(\theta) &= \frac{1}{N} \sum_{j=n+1}^{n+N} \widetilde{\mathbf{X}}_j \widetilde{\mathbf{X}}_j^T (\widehat{u}_{\tau_n}(Z_j) - \widetilde{X}_{j1}^T \theta - \widetilde{X}_{j2}^T \widehat{\beta}_{(k)})^2 \\
\widehat{V}_1^*(\theta) &= \frac{1}{N} \sum_{j=N+1}^{n+N} [(\widehat{u}_{\tau_n}(Z_j) - \widetilde{X}_{j1}^T \theta - \widetilde{X}_{j2}^T \widehat{\beta}_{(k)})^2 \widehat{\eta}_j \widehat{\eta}_j^T] \\
&\quad + \frac{N}{n^2} \sum_{i=1}^{n} [(Y_i - \widehat{u}_{\tau_n}(Z_i))^2 \widehat{\eta}_i \widehat{\eta}_i^T]
\end{aligned}
$$

and

$$
\widetilde{H}(\theta) = \left(\frac{1}{\sqrt{N}} \sum_{j=n+1}^{n+N} W_j(\theta) \right) \left(\frac{1}{\sqrt{N}} \sum_{j=n+1}^{n+N} W_j(\theta) \right)^T.
$$

An adjusted empirical loglikelihood is then defined by

$$
\widetilde{l}_{ad}(\theta) = \widetilde{r}(\theta) \widetilde{l}_{n,N}(\theta),
$$

where

$$
\widetilde{r}_n(\theta) = \frac{\text{tr}(\widehat{V}_1^{*-1}(\theta) \widetilde{H}(\theta))}{\text{tr}(\widehat{V}_0^{*-1}(\theta) \widetilde{H}(\theta))}.
$$

THEOREM 4.2 Assume the assumptions listed in the Appendix, $\widetilde{l}_{ad}(\theta)$ is asymptotically standard chi-square distributed with k degrees of freedom.

5   Simulation Studies

We conducted simulation to better understand the finite-sample performances of the proposed inferential procedures.

11

In our simulation studies, we consider the two cases: $p = 1$ and $p = 2$. For the case: $p = 1$, The surrogates $\tilde{Y}$ were generated as the standard normal random variables. The linear model considered was $Y = \mathbf{X}^T \beta + e$, where $\beta = 2.3$ and $\mathbf{X}$ was generated from a standard exponential distribution, while e given $Z = (\mathbf{X}, \tilde{Y})$ was normally distributed with mean $(\alpha^T Z)^2 - E(\alpha^T Z)^2 - \mathbf{X}^T \beta + E\mathbf{X}^T \beta$ and variance $\sigma^2 = 1$. We let $\alpha = (1.23, 0.32)^T$, and we estimate α using α_n given in Section 2. The simulation were run with validation data and primary data sizes of $(n, N) = (10, 30), (30, 90), (60, 180), (10, 50), (30, 150)$ and $(60, 300)$. The bandwidths $h_{1,n} = n^{-\frac{15}{96}}$ and $h_{2,n} = n^{-\frac{2}{5}}$, and the kernel functions $K_1(\cdot)$ and $K_2(\cdot)$ are taken to be

$$K_1(u) = \begin{cases} -\frac{15}{8}u^2 + \frac{9}{8}, & -1 \le u \le 1 \\ 0, & \text{otherwise} \end{cases}$$

and

$$K_2(u) = \begin{cases} \frac{15}{16}(1 - 2u^2 + u^4), & -1 \le u \le 1 \\ 0, & \text{otherwise} \end{cases}$$

b_n and τ_n were taken to be $n^{-\frac{1}{24}}$ respectively. We calculated the coverage probabilities and the average lengths of the confidence intervals, with nominal level 0.90 and 0.95, respectively, by using 5000 simulation runs. The simulation results are presented in Tables 1 and 2.

For the case: $p = 2$. The surrogates $\tilde{Y}$ were generated as the standard normal random variables. Consider the linear model (1.1) with $\beta = (-1.24, 3.10)^T$, where X was generated from a standard bivariate normal distribution with correlation coefficient $\rho = 0.62$; while e given $Z = (X^T, \tilde{Y})^T$ was normally distributed with mean $(\alpha^T Z)^2 - E(\alpha^T Z)^2 - \mathbf{X}^T \beta + E(\mathbf{X}^T \beta)$ and variance $\sigma^2 = 1$. We let $\alpha = (0.25, -1.31, 1.85)^T$, and we estimate α using α_n given in Section 2. The simulation were run with the same validation data and primary data sizes as in the case: $p = 1$. Also, the bandwidths $h_{1,n}, h_{2,n}, K_1(\cdot), K_2(\cdot), b_n$ and τ_n were taken to be the same as in the case: $p = 1$. We calculated the coverage probabilities of the confidence intervals, with nominal level 0.90 and 0.95, respectively, by using 5000 simulation runs. The simulation results are reported in Tables 3.

6   Appendix

(C.u): $u(\cdot)$ has bounded partial derivatives of order two.

(C.X): $E[X_{ir}^4] < \infty, r = 1, 2, \cdots, p$

(C.$\tilde{Y}$): $E|\tilde{Y}|^4 < \infty$.

(C.Y): $\sup_{z \in \mathcal{Z}} E[Y^2|Z = z] < \infty$

(C.e)i: $E[e|Z] = 0$

    ii: $\sup_{z \in \mathcal{Z}} E[e^2|Z = z] < \infty$.

(C.Σ): $E[XX^\tau]$ is a positive definite matrix.

(C.K_1)i: $K_1(\cdot)$ is symmetric about 0 with bounded support $[-1, 1]$

    ii: $\int_{-1}^1 K_1(u)\, du = 1$ and $\int_{-1}^1 u^i K_1(u)\, du = 0, i = 1, 2, 3$

(C.$h_{1,n}$): As $n \to \infty, h_{1,n} \sim n^{-c_1}$, $b \sim n^{-c_2}$ with positive numbers c_1 and c_2 satisfying that $\frac{1}{8} + \frac{c_2}{4} < c_1 < \frac{1}{4} - c_2$, and the notation "$\sim$" means that two quatities have the same coverage order.

(C.K_2): $K_2(\cdot)$ is a bounded nonnegative kernel function of order one with bounded support.

(C.$h_{2,n}$)i: $nh_{2,n}^{\frac{3}{2}}\tau_n \to \infty$,

    ii: $\frac{nh_{2,n}^3}{\tau_n} \to 0$

(C.f)i: $\sqrt{N}E\{|u(Z)I[f_{\mathcal{Z}}(Z) \leq \tau_n]\} \longrightarrow 0$

    ii: $f_{\mathcal{Z}}(z)$ has bounded partial derivatives of order one.

(C.Nn): $\frac{N}{n} \to \lambda$, where λ is a nonnegative constant.

(C.R^*): $R_i^*(y)$ for $i = 1, 2, \cdots, p + 2$, and $f_Y(y)$ are 3-times differentable and their third derivatives satisfy the following conditions: there exists a neighborhood of the orign, say U_1, and a constant $c > 0$ such that for any $u \in U$

$$
\begin{aligned}
|f_Y^{(3)}(y + u) - f_Y^{(3)}(u)| &\leq c|u|, \\
|R_i^{*(3)}(y + u) - R_i^{*(3)}(u)| &\leq c|u|, i = 1, 2, \cdots, p
\end{aligned}
$$

13

(C.R)i: For pair $1 \leq i, l \leq p+2$ and for any $u \in U$

$$|R_i(y+u)R_l(y+u) - R_i(y)R_l(y)| \leq c|u|$$

ii: $\sqrt{n}ER_i(Y)R_l(Y)I[f_Y(Y) < b] = o(1)$ as $n \to \infty$, for $1 \leq i, l \leq d$, where $I[\cdot]$ is the indicator function and b satisfies $(C.h_{1,n})$.

REMARK: (C.X), (C.Y), $(C.K_1)$, $(C.h_{1,n})$, $(C.R^*)$, (C.R) are used in Zhu and Fang (1997) to obtain the result $\widehat{\alpha}_n - \alpha = O_p(n^{-\frac{1}{2}})$. The remaining conditions are standard.

Sketch of Proof of Theorem 2.1. Clearly

$$\widehat{\beta}_{n,N} - \beta = \Sigma_{n,N}^{-1} A_{n,N} \tag{A.1}$$

and

$$\Sigma_{n,N} \xrightarrow{p} \Sigma, \tag{A.2}$$

where $\Sigma_{n,N}$ and Σ are defined in Section 2 and

$$A_{n,N} = \frac{1}{N} \sum_{k=n+1}^{n+N} \mathbf{X}_k(\widehat{u}_{\tau_n}(Z_k) - \mathbf{X}_k^T \beta)$$

$A_{n,N}$ can be decomposed as

$$
\begin{aligned}
A_{n,N} &= \frac{1}{N} \sum_{k=n+1}^{n+N} \mathbf{X}_k(u(Z_k) - \mathbf{X}_k^T \beta) \\
&\quad + \frac{1}{nN} \sum_{k=n+1}^{n+N} \sum_{i=1}^{n} \frac{\mathbf{X}_k(Y_i - u(Z_k)) K_2\left(\frac{\alpha^T(Z_i - Z_k)}{h_{2,n}}\right)}{h_{2,n} f_Z(Z_k)} + o_p(N^{-\frac{1}{2}})
\end{aligned}
\tag{A.3}
$$

By (A.3), using arguments similar to Wang (1999), it can be shown

$$\sqrt{N} A_{n,N} \xrightarrow{\mathcal{L}} N(0, V_1(\beta)) \tag{A.4}$$

(A.1), (A.2) and (A.4) together prove Theorem 2.1.

Sketch of Proof of Theorem 3.1. It can be proved that

$$
\begin{cases}
\max_{n+1 \leq j \leq n+N} \widehat{A}_j(\beta) &= o_p(N^{\frac{1}{2}}), \\
\frac{1}{N} \sum_{j=n+1}^{n+N} \widehat{A}_j(\beta) \widehat{A}_j(\beta) &= O_p(1) \\
\lambda &= O_p(N^{-\frac{1}{2}}).
\end{cases}
\tag{A.5}
$$

14

By (A.5), standard arguments can be used to prove

$$\widehat{l}_{n,N}(\beta) = \{\frac{1}{\sqrt{N}} \sum_{j=n+1}^{n+N} \widehat{A}_j(\beta)\}^T \widehat{V}_0^{-1}(\beta) \{\frac{1}{\sqrt{N}} \sum_{j=n+1}^{n+N} \widehat{A}_j(\beta)\} + o_p(1). \qquad (A.6)$$

It can be shown that $\widehat{V}_0(\beta) \xrightarrow{p} V_0(\beta)$ by the fact $\widehat{\alpha}_n - \alpha = O_p(n^{-\frac{1}{2}})$ and $\frac{1}{\sqrt{N}} \sum_{j=n+1}^{n+N} \widehat{A}_j(\beta) = O_p(1)$, which can be obtained from (A.4). This proves

$$\widehat{l}_{n,N}(\beta) = \{\frac{1}{\sqrt{N}} V^{-\frac{1}{2}}(\beta) \sum_{j=n+1}^{n+N} \widehat{A}_j(\beta)\}^T D(\beta) \{\frac{1}{\sqrt{N}} V^{-\frac{1}{2}}(\beta) \sum_{j=n+1}^{n+N} \widehat{A}(\beta)\} + o_p(1) \quad (A.7)$$

where $D(\beta) = V^{\frac{1}{2}}(\beta) V_0^{-1}(\beta) V^{\frac{1}{2}}(\beta)$.

Using arguments similar to Wang and Rao (2002), we can prove Theorem 3.1 by (A.6) and (A.7).

Sketch of Proof of Theorem 3.2

It can be shown that

$$\widehat{r}(\beta) = O_p(1). \qquad (A.8)$$

This together with (A.6) proves

$$\widehat{l}_{ad}(\beta) = \{\frac{1}{\sqrt{N}} \sum_{j=n+1}^{n+N} \widehat{A}_j(\beta)\} \widehat{V}_1^{-1}(\beta) \{\frac{1}{\sqrt{N}} \sum_{j=n+1}^{n+N} \widehat{A}_j(\beta)\} + o_p(1).$$

By (A.4) and the fact $\widehat{V}_1(\beta) \xrightarrow{p} V_1(\beta)$, Theorem 3.2 (i) is then proved. Theorem 3.2 (ii) is a direct result of (i).

Sketch of Proof of Theorem 4.1. Similar to (A.6), we have

$$\widetilde{l}_{n,N}(\theta) = \{\frac{1}{\sqrt{N}} \sum_{j=n+1}^{n+N} \widehat{W}_j(\theta)\} \widehat{V}_0^{*-1}(\theta) \{\frac{1}{\sqrt{N}} \sum_{j=n+1}^{n+N} \widehat{W}_j(\theta)\} + o_p(1), \qquad (A.9)$$

where $\widehat{V}_0^*(\theta)$ is defined in Section 4.

Observe that

$$\begin{aligned} \frac{1}{\sqrt{N}} \sum_{j=n+1}^{n+N} \widehat{W}_j(\theta) &= \frac{1}{\sqrt{N}} \sum_{j=n+1}^{n+N} \widetilde{X}_{j1}(\widehat{u}_{\tau_n}(Z_j) - \widetilde{X}_j^T \gamma) \\ &\quad + \frac{1}{\sqrt{N}} \sum_{j=n+1}^{n+N} \widetilde{X}_{j1} \widetilde{X}_j^T (\gamma - \widehat{\gamma}_n) \\ &\quad + \frac{1}{\sqrt{N}} \sum_{j=n+1}^{n+N} \widetilde{X}_{j1} \widetilde{X}_{j1}^T (\widehat{\theta}_n - \theta) \\ &:= T_{N1} + T_{N2} + T_{N3} \end{aligned} \qquad (A.10)$$

15

Standard arguments can be used to get

$$T_{N2} = -\{E[\widetilde{X}_{j1}\widetilde{\mathbf{X}}_j^T]\}\{E\widetilde{\mathbf{X}}_j\widetilde{\mathbf{X}}_j^T\}\{\frac{1}{\sqrt{N}}\sum_{j=n+1}^{n+N}\widetilde{\mathbf{X}}_j(u_{\tau_n}(Z_j)-\widetilde{\mathbf{X}}_j^T\gamma)\} + o_p(1) \quad (A.11)$$

and

$$
\begin{aligned}
T_{N3} &= E\{\widetilde{X}_{j1}\widetilde{X}_{j1}^T\}\{E(\widetilde{X}_{j1}\widetilde{X}_{j1}^T)-K^TP^{-1}K\}^{-1} \\
&\times \left[\frac{1}{\sqrt{N}}\sum_{j=n+1}^{n+N}(\widetilde{X}_{j1}-K^TP^{-1}\widetilde{X}_{j2})(\widehat{u}_{\tau_n}(Z_j)-\widetilde{\mathbf{X}}_j^T\gamma)\right] + o_p(1).
\end{aligned}
\quad (A.12)
$$

It follows from (A.10)—(A.12)

$$\frac{1}{\sqrt{N}}\sum_{j=n+1}^{n+N}\widehat{W}_j(\theta) = \frac{1}{\sqrt{N}}\sum_{j=n+1}^{n+N}\eta_j(\widehat{u}_{\tau_n}(Z_j)-\widetilde{\mathbf{X}}_j^T\gamma) + o_p(1) \xrightarrow{\mathcal{L}} N(0,V_1^*(\theta)), \quad (A.13)$$

where $V_1^*(\theta)$ is as defined in Section 4. Note that $\widehat{V}_0^*(\theta) \xrightarrow{p} V_0^*(\theta)$, where $V_0^*(\theta)$ is defined in Section 4. This together with (A.13) proves Theorem 3.1 by arguments similar to Wang and Rao (2002).

Sketch of Proof of Theorem 4.2 The arguments are similar to that of Theorem 4.2 from Theorem 4.1.

16

REFERENCES

Buonaccorsi, J.P. (1996). Measurement error in the response in the general linear model. J. Amer. Statist. Assoc., 91(434), 633-642.

Carroll R.J. and Stefanski, L.A. (1990). Approximate quasi-likelihood estimation in models with surrogate predictors. J. Amer. Statist. Assoc., 85,652-663.

Carroll, R.J., Knickerbocker, R. K. and Wang, C. Y.(1995) Dimension reduction in a semiparametric regression model with errors in covariates. The Annals of Statistics 23 161-181.

Carroll,R.J. and Wand, M.P.(1991). Semiparametric estimation in logistic measure error models. J. Roy. Statist.soc., Ser B **53** 652-663.

Duan,N and Li, K.C.(1991). Slicing regression: a link-free regression method. Ann. Statist. 19, 505-530.

Duncan, G. and Hill, D.(1985). An investigations of the extent and consequences of measurement error in labor-economics survey data. Journal of Labor Economics 3 508-532.

Fuller, W.A.(1987). Measurement error models. New York, John Wiley & Sons, Inc.

Li, K. C. (1991). Sliced inverse regression for dimension reduction (with discussion). J. Amer. Statist. Assoc. 86 337-342.

Owen, A. (1988). Empirical likelihood ratio confidence intervals for single functional. *Biometrika* **75**, 237-249.

Owen, A.(1991). Empirical likelihood for linear models. *Ann. Statist.* **19**, 1725-1747.

Pepe, M.S.(1992). Inference using surrogate outcome data and a validation sample. Biometrika 79 355-365.

Pepe, M.S. and Fleming, T.R.(1991). A general nonparametric method for dealing with errors in missing or surrogate covariate data. J. Amer. Statist. Assoc. 86 108-113.

Pepe, M. S., Reilly, M. and Fleming, T. R. (1994). Auxiliary outcome data and the mean score method. J. Statist. Plan. Inference 42   137-160.

Reilly, M. and Pepe, M. S. (1995). A mean score method for missing and auxiliary covariate data in regression models. Biometrika 82, 299-314.

Rosner, B., Willett, W.C. and Spiegelman, D. (1989). Correction of logistic regression relative risk estimates and confidence intervals for systematic within-person measurement error. Statist. Med, 8, 1075-1093

17

Stefanski, L. A.(1987) and Carrol, R.J.(1987). Conditional scores and optimal scores for generalized linear measurement error models. Biometrika 74 703-716.

Sepanski, J.H. and Lee, L.F.(1995). Semiparametric Estimation of nonlinear error-in-variables models with validation study. J. Nonparametric Statistics 4 365-394.

Wang, Q.H.(1999). Estimation of partial linear error-in-variables model. Jourmal of Multivariate Analysis 69 30-64.

Wang, Q.H. (2000). Estimation of linear error-in-covariables models with validation data under random censorship. Journal of Multivariate Analysis 74, 245-266.

Wang, Q.H. and Rao,J.N.K.(2002). Empirical likelihood-based in linear errors-in-covariables models with validation data. Biometrika 89,

Wittes, J., Lakatos, E. and Probstfied, J.(1989). Surrogate endpoints in clinical trails: Cardiovascular diseases. Statist. Med. 8, 415-425.

Zhu, L.X. and Fang, K.T.(1996). Asymptotics for kernel estimator of sliced inverse regression. Ann. Statist., 24, 1053-1068.

Nonparametric state price density estimation using constrained least squares and the bootstrap

Adonis Yatchew[a],[*],[1], Wolfgang Härdle[b],[2]

[a]*Department of Economics, University of Toronto, 150 St. George St., Toronto, Canada M5S 3G7*
[b]*Humboldt-Universität zu Berlin, Center for Applied Statistics and Economics, Germany*

Abstract

The economic theory of option pricing imposes constraints on the structure of call functions and state price densities. Except in a few polar cases, it does *not* prescribe functional forms. This paper proposes a nonparametric estimator of option pricing models which incorporates various restrictions (such as monotonicity and convexity) within a single least squares procedure. The bootstrap is used to produce confidence intervals for the call function and its first two derivatives and to calibrate a residual regression test of shape constraints. We apply the techniques to option pricing data on the DAX.
© 2005 Elsevier B.V. All rights reserved.

JEL classification: C14; G12

Keywords: Options; State price density; Nonparametric least squares; Bootstrap

---
[*]Corresponding author. Tel.: +1 416 978 7128; fax: +1 416 978 6713.

E-mail addresses: Yatchew@chass.utoronto.ca (A. Yatchew), Haerdle@wiwi.hu-berlin.de
(W. Härdle).

1. State price density estimation

1.1. Parametric or Nonparametric?

Option price data have characteristics which are both nonparametric and parametric in nature. The economic theory of option pricing predicts that the price of a call option should be a monotone decreasing convex function of the strike price. It also predicts that the state price density (SPD) which is proportional to the second derivative of the call function, is a valid density function over future values of the underlying asset price, and hence must be nonnegative and integrate to one. Except in a few polar cases, the theory does *not* prescribe specific functional forms. (Indeed the volatility smile is an example of a clear violation of the lognormal *parametric* specification implied by Black–Scholes.) All this points to a nonparametric approach to estimation of the call function and its derivatives.

On the other hand, multiple transactions are typically observed at a finite vector of strike prices. Thus, one could argue that the model for the option price—as a function of the strike price (other variables held constant)—is intrinsically parametric. Indeed given sufficient data, one can obtain a good estimate of the call function by simply taking the mean transactions price at each strike price. Unfortunately, even with large data-sets, accurate estimation of the call function at a finite number of points does not assure good estimates of its first and second derivatives, should they exist. To incorporate smoothness and curvature properties, one can select a parametric family which is differentiable in the strike price, and impose constraints on coefficients. Such an approach, however, risks specification failures.

Fortunately, nonparametric regression provides a good reservoir of candidates for flexible estimation. Indeed, a number of authors have used nonparametric or semiparametric techniques in the estimation or testing of derivative asset models. Among them are Aït-Sahalia (1996), Jackwerth and Rubinstein (1996), Ghysels et al. (1997), Aït-Sahalia and Lo (2000), Broadie et al. (2000a, b), Garcia and Gencay (2000), Cont (2001), Cont and da Fonseca (2002), Daglish (2003), Härdle et al. (2002) and Aït-Sahalia and Duarte (2003).

In earlier work, Yatchew and Bos (1997) showed how nonparametric least squares can easily incorporate constraints such as monotonicity, concavity, additive separability, homotheticity and other implications of economic theory. Indeed, there is a growing literature on the imposition and testing of curvature properties on nonparametric estimators. For references see Yatchew (2003, Chapter 6).

In the current paper, we combine shape restrictions with nonparametric regression to estimate the call price function and the SPD within a single least squares procedure. Constraints include smoothness of various order derivatives, monotonicity and convexity of the call function and integration to one of the SPD. Confidence intervals and test procedures may be implemented using bootstrap methods. In addition to providing simulation results we apply the procedures to option data on the DAX index for January 4, 1999.

As an initial illustration of the benefits of smooth constrained estimation, particularly when estimating derivatives, we have generated 20 independent

transactions prices at each of 25 strike prices. Details of the data generating mechanism are contained in Section 3 below. The top panel of Fig. 1A depicts all 500 observations and the 'true' call function. As is typical in market data, the variance decreases as the option price declines. The second panel depicts the estimated call function obtained by taking the mean transactions price at each of the 25 strike prices. The bottom panel depicts our smooth constrained estimate. Both estimates lie close to the true function.

Fig. 1B contains estimates of the first derivative. The upper panel depicts first-order divided differences of the point means, (these are the slopes of the lines joining the consecutive means in the middle panel of Fig. 1A). By the mean value theorem, they should provide a reasonable estimate of the true first derivative near the point of approximation. But as can be seen, the estimate deteriorates rapidly as one moves to the left and the variance in transactions prices increases. The bottom panel depicts the first derivative of the proposed smooth constrained estimate which by comparison is close to the first derivative of the true call function.

Fig. 1C illustrates estimates of the second derivative of the call function. The upper panel depicts second-order divided differences of the point means. (These are the slopes of the lines joining consecutive points in the top panel of Fig. 1B.) The estimates gyrate wildly around the true second derivative. The lower panel depicts the second derivative of the smooth constrained estimate which tracks the true function reasonably well (note the change in scale of the vertical axis).

A number of practical advantages ensue from the procedures we propose. First, various combinations of constraints can be incorporated in a natural way within a single least squares procedure. Second, our 'smoothing' parameter has an intuitive interpretation since it measures the smoothness of the class of functions over which estimation is taking place by using a (Sobolev) norm. If one wants to impose smoothness on higher order derivatives, this can be done by a simple modification to the norm. Third, call functions and SPDs can be estimated on an hour-by-hour, day-by-day or 'moving window' basis, and changes in shape can be tracked and tested. Fourth, our procedures readily accommodate heteroskedasticity and time series structure in the residuals.

In the following, we consider two types of generating mechanisms for the 'x' variable. In the first, x is drawn from a continuous distribution as would be the case if one were estimating the call function as a function of 'moneyness'. In the second, x is drawn from a discrete distribution at a finite set of strike prices as depicted in the upper panel of Fig. 1A.

The paper is organized as follows. The remainder of this section summarizes some relevant financial theory and establishes notation. Section 2 outlines the estimator as well as inference procedures. Section 3 describes the results of simulations and estimation using DAX index options data. Section 4 contains our conclusions. An Appendix contains proofs.

1.2. Financial market theory

Implicit in the prices of traded financial assets are Arrow-Debreu prices or in a continuous setting, the SPD. These are elementary building blocks for understanding
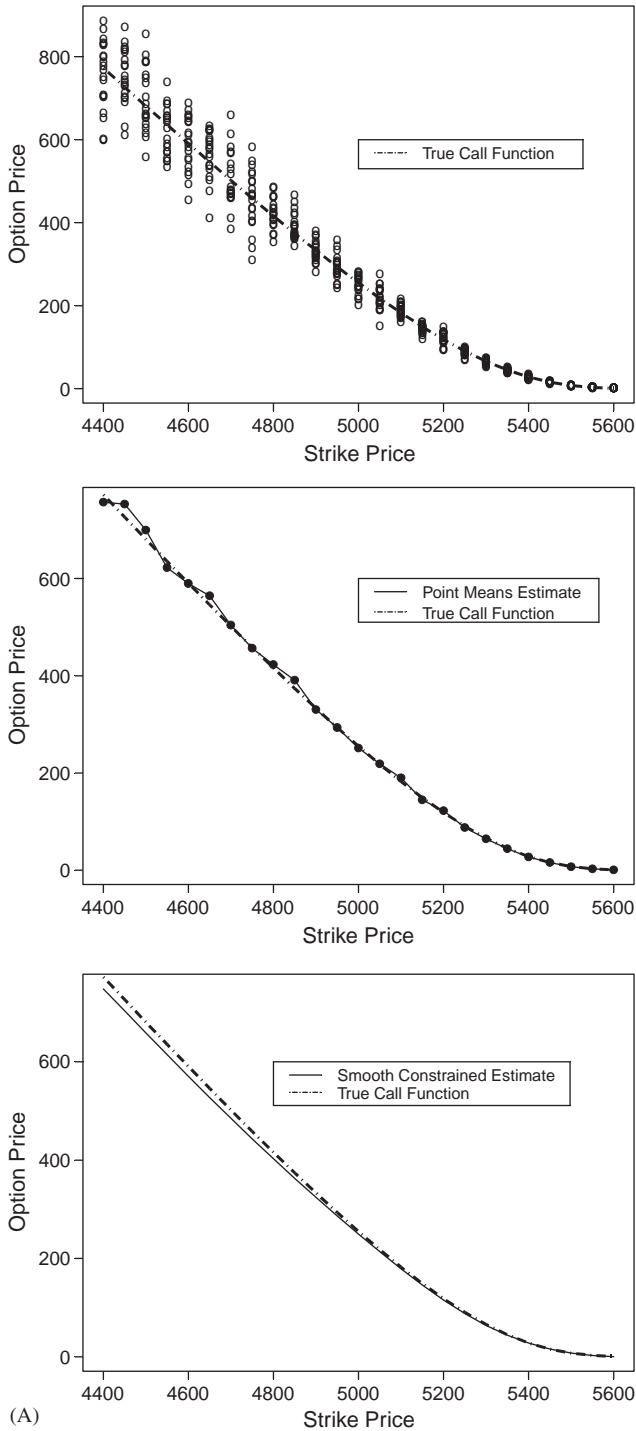
Fig. 1. (A) Data and estimated call function; (B) estimated first derivative; (C) estimated SPDs.
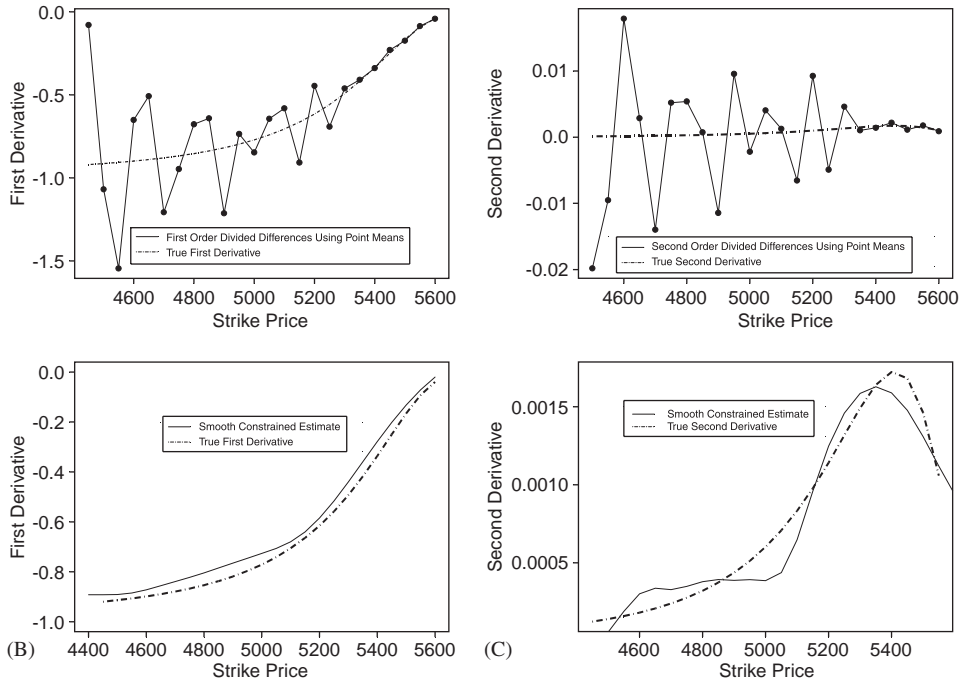
Fig. 1. (*Continued*)

markets under uncertainty. The existence and characterization of SPDs has been studied by Black and Scholes (1973), Merton (1973), Rubinstein (1976) and Lucas (1978) amongst many others. Under the assumption of noarbitrage, the SPD is usually called the risk neutral density because if one assumes that all investors are risk neutral, then the expected return on all assets must equal the risk free rate of interest. Cox and Ross (1976) showed that under this assumption, the Black–Scholes equation follows immediately. Other approaches have been proposed by Derman and Kani (1994) and Barle and Cakici (1998).

Let x be the strike price for a call option which will expire at time T. Let t be the current time, r the interest rate, τ the time to expiry and δ the dividend yield. Let P_t and P_T denote prices of the underlying asset at times t and T, respectively. Then the call pricing function at time t is given by

$$C(x, \tau, P_t, r, \delta) = e^{-r\tau} \int_0^\infty \max(P_T - x, 0) f^*(P_T | P_t, \tau, r, \delta) \, dP_T, \tag{1}$$

where the function f^* is the SPD. It assigns probabilities to various values of the asset at time of expiration given the current asset price, the time to expiry, the current risk-free interest rate and the corresponding dividend yield of the asset. As stated earlier, the call function is monotone decreasing and convex in x.

Breeden and Litzenberger (1978) show that the second derivative of the call pricing function with respect to the strike price is related to the SPD by

$$f^*(P_T|P_t, \tau, r, \delta) = e^{r\tau} \frac{\partial^2 C(x, \tau, P_t, r, \delta)}{\partial x^2}. \tag{2}$$

We will focus on data over a sufficiently brief time span so that we may take the time to maturity, the underlying asset price, the interest rate and dividend yield to be roughly constant. Our central objective will be to estimate the call function subject to monotonicity and convexity constraints.

We will use the following notational conventions. For an arbitrary vector z and matrices A, B we will use z_s, A_{st}, $[Az]_s$ and $[AB]_{st}$ to denote elements. Occasionally, we will need to refer to sub-matrices of a matrix. In this case we will adopt the notation $A[a:b, c:d]$ to refer to those elements which are in rows a through b and columns c through d. Given a function f, we will denote derivatives using bracketed superscripts, e.g., $f^{(1)}, f^{(2)}$.

2. Constrained nonparametric procedures

2.1. Nonparametric least squares

We begin with constrained nonparametric least squares estimation of a function of one variable on the interval $[a, b]$. Given data $(x_1, y_1), \ldots, (x_n, y_n)$, let $x = (x_1, \ldots, x_n)$ and $y = (y_1, \ldots, y_n)$. (The x_i will be strike price or 'moneyness' and y_i the option price.) With mild abuse of notation we will use x and y to denote the variable in question in addition to the vector of observations on that variable. Our model is given by

$$y_i = m(x_i) + \varepsilon_i \quad i = 1, \ldots, n. \tag{3}$$

We will assume that the regression function m is four times differentiable, which in a nonparametric setting, will ensure consistent and smooth estimates of the function and its first and second derivatives. (Other orders of differentiation can readily be accommodated using the framework below.) We will assume that the x_i lie in the interval $[a, b]$. (For example, if the x variable is 'moneyness' then $[a, b]$ would typically be the interval $[0.8, 1.2]$.) We will also assume the residuals ε_i are independent and heteroskedastic, but this assumption can be weakened. Let Σ be the diagonal matrix with diagonal values $\sigma^2(x_1), \ldots, \sigma^2(x_n)$.

Let C^4 be the space of four times continuously differentiable scalar functions, i.e., $C^4 = \{m : [a, b] \to \mathbb{R}^1 | m^{(i)} \in C^0, i = 0, 1, \ldots, 4\}$, where C^0 is the set of continuous functions on $[a, b]$. On the space C^4 define the norm, $\|m\|_{\infty,4} = \sum_{i=0}^{4} \max_x |m^{(i)}|$ in which case C^4 is a complete, normed, linear space, i.e., a Banach space. Given the inner product:

$$\langle m_1, m_2 \rangle_{\text{Sob}} = \sum_{i=0}^{4} \int_a^b m_1^{(i)} m_2^{(i)} \, dx \tag{4}$$

with corresponding norm:

$$\|m\|_{\text{Sob}} = \left[\sum_{i=0}^{4} \int_{a}^{b} [m^{(i)}]^2 \, dx \right]^{1/2} \tag{5}$$

define the Sobolev space $\mathcal{H}^4$ as the completion of $\{m \in C^4\}$ with respect to $\|m\|_{\text{Sob}}$. We are interested in the following optimization problem:

$$\min_{m} \quad \frac{1}{n} [y - m(x)]^{\text{T}} \Sigma^{-1} [y - m(x)]$$
$$\text{s.t.} \quad \|m\|_{\text{Sob}}^2 \leqslant L \tag{6}$$

which imposes a smoothness condition with smoothing parameter L. By varying this parameter, we control the smoothness of the ball of functions over which estimation is taking place.

Using techniques well-known in the spline function literature, it can be shown that the infinite dimensional optimization problem (6) can be replaced by a finite dimensional optimization problem as we outline below, (see e.g., Wahba, 1990 or Yatchew and Bos, 1997).

Since the evaluation of functions at a specific point is a linear operator, by the Riesz Representation Theorem, given a point x_o there exists a function r_{x_o} in $\mathcal{H}^4$ called a *representor* such that $\langle r_{x_o}, h \rangle_{\text{Sob}} = h(x_o)$ for any $h \in \mathcal{H}^4$. Let $r_{x_1}, \dots, r_{x_n}$ be the representors of function evaluation at $x_1, \dots, x_n$, respectively. (Details of the calculation of representors may be found in Yatchew and Bos (1997) or Yatchew (2003, Appendix D).) Let R be the $n \times n$ representor matrix whose columns (and rows) equal the representors evaluated at $x_1, \dots, x_n$; that is, $R_{ij} = \langle r_{x_i}, r_{x_j} \rangle_{\text{Sob}} = r_{x_i}(x_j) = r_{x_j}(x_i)$. If one solves:

$$\min_{c} \quad \frac{1}{n} [y - Rc]^{\text{T}} \Sigma^{-1} [y - Rc]$$
$$\text{s.t.} \quad c^{\text{T}} Rc \leqslant L, \tag{7}$$

where c is an $n \times 1$ vector, then the minimum value is equal to that obtained by solving (6). Furthermore, there exists a solution to (7) of the form:

$$\hat{m}(x) = \sum_{i=1}^{n} \hat{c}_i r_{x_i}(x), \tag{8}$$

where $\hat{c} = (\hat{c}_1, \dots, \hat{c}_n)^{\text{T}}$ solves (7). First and second derivatives may be estimated by differentiating (8):

$$\hat{m}^{(1)}(x) = \sum_{i=1}^{n} \hat{c}_i r_{x_i}^{(1)}(x) \tag{9}$$

and

$$\hat{m}^{(2)}(x) = \sum_{i=1}^{n} \hat{c}_i r_{x_i}^{(2)}(x). \tag{10}$$

Define $R^{(1)}$ to be the $n \times n$ matrix whose columns (and rows) are the first derivatives of the representors evaluated at $x_1, \ldots, x_n$. Define $R^{(2)}$ in a similar fashion. Then the estimates of the call function and its first two derivatives at the vector of observed strike prices are given by $\hat{m}(x) = R\hat{c}$, $\hat{m}^{(1)}(x) = R^{(1)}\hat{c}$ and $\hat{m}^{(2)}(x) = R^{(2)}\hat{c}$ where $x = (x_1, \ldots, x_n)$.

Proposition 1. *Suppose one is given data* $(x_1, y_1), \ldots, (x_n, y_n)$ *where* $y_i = m(x_i) + \varepsilon_i$, *the* ε_i *are independently distributed with* $Var(\varepsilon_i) = \sigma^2(x_i) = \sigma_i^2$ *and* $0 < \sigma^2(x) < K$ *for some* K. *The* x_i *are i.i.d. with continuous density* $p(x)$ *on* $[a, b]$ *bounded away from zero. Let* $\hat{m}$ *satisfy* (6). *Then* $\sup_{x \in [a,b]} |\hat{m}^{(s)} - m^{(s)}| \xrightarrow{p} 0$ *for* $s = 0, 1, 2$ *and* $1/n \sum (m(x_i) - \hat{m}(x_i))^2 = O_P(n^{-8/9})$.

The result establishes the consistency of the estimator and its first two derivatives. It also establishes the rate of convergence of the estimator which equals the optimal rate for four-times differentiable nonparametric functions of one variable (see Stone, 1980, 1982). The rate of convergence will be useful for implementing residual regression tests of hypotheses on m. One can replace the true variances in Eq. (6) with consistent estimates, or even with ones, though doing the latter would likely reduce efficiency.

2.2. Imposition of constraints

Optimization problem (7) allows easy incorporation of various restrictions. Suppose one wants to impose the constraint that $\hat{m}$ is monotone decreasing at each x_i. Then one restricts the first derivative (9) to be negative at these points. To impose convexity, one can require the second derivative (10) to be positive. Then the quadratic optimization problem (7) can be supplemented with the monotonicity constraints:

$$R^{(1)}c \leqslant 0 \tag{11}$$

and the convexity constraints:

$$R^{(2)}c \geqslant 0. \tag{12}$$

Suppose we solve (7) subject to monotonicity and convexity constraints (11) and (12). Then the conclusions of Proposition 1 continue to hold as long as the true function m is *strictly* monotone and convex (see Yatchew and Bos, 1997). This is because, as sample size grows, and the estimate of the function and its first two derivatives approach their true counterparts, the smoothness constraint alone will be sufficient to ensure that the estimated function will be monotone and convex. That is, the monotonicity and convexity constraints will eventually become non-binding.

Next we turn to imposing unimodality and 'tail' constraints on the SPD. Suppose the current asset price lies in the interval $[x_{i^*}, x_{i^*+1}]$ and that one wants to impose the constraint that the SPD is unimodal with the mode in this same interval. Since the SPD is proportional to the second derivative $m^{(2)}$, one needs to impose constraints on *its* derivative, that is on $m^{(3)}$. Define $R^{(3)}$ to be the $n \times n$ matrix whose columns (and rows) are the third derivatives of the representors evaluated at $x_1, \ldots, x_n$. Then

one imposes the constraints:

$$R^{(3)}_{i^* \times n}[1:i^*, 1:n]c_{n \times 1} \geqslant 0 \quad R^{(3)}_{(n-i^*) \times n}[i^*+1:n, \ 1:n]c_{n \times 1} \leqslant 0. \tag{13}$$

The first set of i^* inequalities ensures that the SPD has a positive derivative at strike prices below the current asset price, the remaining $n - i^*$ inequalities provide for a negative derivative at higher strike prices.

A weaker set of constraints on the SPD involves requiring that the density declines as one moves out in the tails. In the empirical work below, we typically impose such 'tail constraints'. These constraints are particularly beneficial when estimating the left tail of the SPD from call function data since the variance of the residual is larger there (see, e.g., Fig. 1A, top panel) and there are typically fewer observations at strike prices that are far from the current value of the index.

Finally, since the integral under the density cannot exceed one, we have:

$$\mathrm{e}^{r\tau} \sum_{i=1}^{n} c_i[r^{(1)}_{x_i}(b) - r^{(1)}_{x_i}(a)] \leqslant 1, \tag{14}$$

where the exponential term reflects the proportionality factor relating the second derivative to the SPD as in Eq. (2).

2.3. Testing monotonicity and convexity

Suppose one wants to test monotonicity and convexity. The residual regression test considered by Fan and Li (1996) and Zheng (1996) can be adapted to produce a test of these properties. (See also Yatchew (2003), Chapter 6.) The basic idea underlying the procedure is that one takes the residuals from the 'restricted regression' which imposes the constraints to be tested, then performs a kernel regression of these residuals on the explanatory variable x to see whether there is anything left to be explained. If so, then the null hypothesis is rejected.

Proposition 2. *Suppose m is strictly monotone and convex and $\hat{m}$ is the smooth constrained estimator obtained by solving* (7) *subject to monotonicity and convexity constraints* (11) *and* (12). *Let $\lambda n \to \infty$, $\lambda^{1/2}n^{1-8/9} \to 0$. Let K be a kernel function (such as the normal, uniform or triangular kernel), and define*

$$\mathrm{U} = \frac{1}{n} \sum_i (y_i - \hat{m}(x_i)) \left[\frac{1}{\lambda n} \sum_{j \neq i}(y_j - \hat{m}(x_j))K\left(\frac{x_j - x_i}{\lambda}\right) \right] \tag{15a}$$

then

$$\lambda^{1/2}n\mathrm{U} \xrightarrow{D} \mathrm{N}\left(0, 2\int \sigma^4(x)p^2(x) \int K^2(u)\right). \tag{15b}$$

Let the estimated variance of U be given by

$$\hat{\sigma}^2_{\mathrm{U}} = \frac{2}{\lambda^2 n^4} \sum_i \sum_{j \neq i}(y_i - \hat{m}(x_i))^2(y_j - \hat{m}(x_j))^2 K^2\left(\frac{x_j - x_i}{\lambda}\right). \tag{15c}$$

Then, $\lambda n^2 \hat{\sigma}^2_{\mathrm{U}} \xrightarrow{p} 2\int \sigma^4(x)p^2(x) \int K^2(u)$. Hence $\mathrm{U}/\hat{\sigma}_{\mathrm{U}} \xrightarrow{D} \mathrm{N}(0,1)$.

The test described in Proposition 2, may be performed using the indicated asymptotic normal approximation. Alternatively, it may be implemented using the bootstrap as we describe below. It is consistent against nonmonotone or nonconvex alternatives.

2.4. Multiple observations

As we indicated in the introduction, option price data often consist of multiple observations at a finite vector of strike prices. We will need to modify our set-up to incorporate this characteristic. Let $X = (X_1, \ldots, X_k)$ be the vector of k distinct strike prices. (In Fig. 1A, there are $k = 25$ distinct strike prices with 20 observations at each price so that $n = 500$.) We will assume that the vector X is in increasing order. Let $\sigma^2(X_1), \ldots, \sigma^2(X_k)$ be the residual variances at each of the distinct strike prices. Let B be the $n \times k$ matrix such that:

$$B_{ij} = 1 \quad \text{if } x_i = X_j$$
$$= 0 \quad \text{otherwise.} \tag{16}$$

We may now rewrite (6) as

$$\min_m \quad \tfrac{1}{n} \left[y_{n \times 1} - B_{n \times k} m(X)_{k \times 1} \right]^{\mathrm{T}} \Sigma_{n \times n}^{-1} \left[y_{n \times 1} - B_{n \times k} m(X)_{k \times 1} \right]$$
$$\text{s.t.} \quad \|m\|_{\mathrm{Sob}}^2 \leqslant L. \tag{17}$$

Noting that the representer matrix R is in this case $k \times k$, the analogue to (7) becomes:

$$\min_c \quad \tfrac{1}{n} [y - BRc]^{\mathrm{T}} \Sigma^{-1} [y - BRc]$$
$$\text{s.t.} \quad c^{\mathrm{T}} Rc \leqslant L, \tag{18}$$

where c is a $k \times 1$ vector. Monotonicity and convexity constraints (11) and (12) can be added noting that $R^{(1)}$ and $R^{(2)}$ are now also $k \times k$ matrices.

Even if the number of distinct strike prices k does not increase, the call function can be estimated consistently at $X_1, \ldots, X_k$. However, as was pointed out by a referee, this does not assure that estimates of derivatives are estimated consistently. Indeed, derivatives cannot be consistently estimated at a point without an accumulation of observations in the neighborhood of the point, though of course a sufficient condition for consistency of the first two derivatives is that the function $m(x)$ is a linear combination of the representors $r_{X_1}, \ldots, r_{X_k}$.

Proposition 3. *Suppose one is given data* $(x_1, y_1), \ldots, (x_n, y_n)$, *where* $y_i = m(x_i) + \varepsilon_i$, *the* ε_i *are independently distributed with* $Var(\varepsilon_i) = \sigma_i^2$ *and* x_i *are sampled from a discrete distribution whose support is* $X_1, \ldots, X_k$ *with corresponding probabilities* $\pi_1, \ldots, \pi_k$. *Suppose* m *lies strictly inside the ball of functions* $\|m\|_{\mathrm{Sob}}^2 < L$ *and* m *is strictly monotone decreasing and strictly convex and is a linear combination of the representors* $r_{X_1}, \ldots, r_{X_k}$. *Let* $\bar{y}(X) = \bar{y}(X_1, \ldots, X_k)$ *be the* k-*dimensional vector of averaged values of* y *at the* k *distinct values of* X. *Let* $\hat{c}$ *minimize* (18) *with the added*

constraints (11) *and* (12) *and define* $\hat{m}(X) = \hat{m}(X_1, \ldots, X_k) = R\hat{c}$, $\hat{m}^{(1)}(X) = R^{(1)}\hat{c}$ *and* $\hat{m}^{(2)}(X) = R^{(2)}\hat{c}$. *Let* $\Omega/n = Var(\bar{y}(X))$ *be the* $k \times k$ *diagonal matrix of variances of the point mean estimators, i.e.,* $\Omega_{jj} = \sigma^2(X_j)/\pi_j$. *Then* $\lim_{n\to\infty} Prob\lfloor \hat{m}(X) = \bar{y}(X)\rfloor = 1$, $\hat{m}(X) \overset{p}{\to} m(X)$, $\hat{m}^{(1)}(X) \overset{p}{\to} m^{(1)}(X)$ *and* $\hat{m}^{(2)}(X) \overset{p}{\to} m^{(2)}(X)$. *Furthermore,*

$$n^{1/2}(\hat{m}(X) - m(X)) \overset{D}{\to} N(0, \Omega), \tag{19}$$

$$n^{1/2}(\hat{c} - c) \overset{D}{\to} N(0, R^{-1}\Omega R^{-1^{\mathrm{T}}}),$$

$$n^{1/2}(\hat{m}^{(1)}(X) - m^{(1)}(X)) \overset{D}{\to} N(0, R^{(1)} R^{-1}\Omega R^{-1^{\mathrm{T}}} R^{(1)^{\mathrm{T}}}),$$

$$n^{1/2}(\hat{m}^{(2)}(X) - m^{(2)}(X)) \overset{D}{\to} N(0, R^{(2)} R^{-1}\Omega R^{-1^{\mathrm{T}}} R^{(2)^{\mathrm{T}}}). \tag{20}$$

Proposition 3 states that as data accumulate at each strike price, the inequalities implied by the smoothness, monotonicity and convexity constraints eventually become nonbinding, the estimator becomes identical to the point mean estimator and the call function m is estimated consistently at the observed strike prices. Moreover, because the true call function is here assumed to be a linear combination of the representors at the observed strike prices, the first and second derivatives are also estimated consistently.

Proposition 3 provides for asymptotic scalar and vector confidence regions of the call function, its first derivative and the SPD. For example, if one is interested in confidence intervals at strike price X_j, the asymptotic pivots are:

$$\tau^{(0)} = n^{1/2} \frac{(\hat{m}(X_j) - m(X_j))}{[\hat{\Omega}]_{jj}^{1/2}},$$

$$\tau^{(1)} = n^{1/2} \frac{(\hat{m}^{(1)}(X_j) - m^{(1)}(X_j))}{[R^{(1)} R^{-1}\hat{\Omega} R^{-1^{\mathrm{T}}} R^{(1)^{\mathrm{T}}}]_{jj}^{1/2}},$$

$$\tau^{(2)} = n^{1/2} \frac{(\hat{m}^{(2)}(X_j) - m^{(2)}(X_j))}{[R^{(2)} R^{-1}\hat{\Omega} R^{-1^{\mathrm{T}}} R^{(2)^{\mathrm{T}}}]_{jj}^{1/2}}. \tag{21}$$

We note that the $\sigma^2(X_j)$ and the π_j may be estimated using the sample variance and sample proportion of observations at each distinct strike price.

2.5. Bootstrap procedures

Percentile and percentile-t procedures are commonly used for constructing confidence intervals. The latter are often found to be more accurate when the statistic is an asymptotic pivot (see Hall (1992) for extensive arguments in support of this proposition). On the other hand, percentile methods might be better when the asymptotic approximation to the distribution of the pivot is poor as a result of small sample size or slow convergence. Table 1 contains an algorithm for constructing percentile confidence intervals for the call function and its first two derivatives at X_j. As there are multiple observations at each strike price, we can accommodate

Table 1
Percentile confidence intervals for $m(X_j)$, $m^{(1)}(X_j)$ and $m^{(2)}(X_j)$

1.		Calculate $\hat{c}$ and $\hat{m}$ by solving (18) subject to (11) and (12). Calculate the estimate residuals $\hat{\varepsilon}_1, \ldots, \hat{\varepsilon}_n$.
2.	(a)	Construct a bootstrap data-set $(x_i, y_1^B), \ldots, (x_n, y_n^B)$ where $y_i^B = \hat{m}(x_i) + \varepsilon_i^B$ and ε_i^B is obtained by sampling from $\hat{\varepsilon}_1, \ldots, \hat{\varepsilon}_n$ using the wild bootstrap.
	(b)	Using the bootstrap data-set obtain $\hat{c}^B$ by solving (18) subject to (11) and (12). Calculate and save $\hat{m}^B(X_j)$, $\hat{m}^{(1)B}(X_j)$ and $\hat{m}^{(2)B}(X_j)$.
3.		Repeat step 2 multiple times.
4.		To obtain 95% point-wise confidence intervals for $m(X_j)$, $m^{(1)}(X_j)$ and $m^{(2)}(X_j)$ obtain 0.025 and 0.975 quantiles of the corresponding bootstrap estimates.

Table 2
Bootstrap residual regression test of monotonicity and convexity

1.		Calculate $\hat{c}$ and $\hat{m}$ by solving (18) subject to (11) and (12). Save the estimates of the regression function $\hat{m}(x_1), \ldots, \hat{m}(x_n)$ and the residuals $\hat{\varepsilon}_1, \ldots, \hat{\varepsilon}_n$.
2.		Calculate U, $\hat{\sigma}_U$, and $U/\hat{\sigma}_U$ as in Proposition 2.
3.	(a)	Sample using the wild bootstrap from $\hat{\varepsilon}_1, \ldots, \hat{\varepsilon}_n$ to obtain $\hat{\varepsilon}_1^B, \ldots, \hat{\varepsilon}_n^B$ and construct a bootstrap data set $(y_1^B, x_1), \ldots, (y_n^B, x_n)$, where $y_i^B = \hat{m}(x_i) + \hat{\varepsilon}_i^B$.
	(b)	Using the bootstrap data set, estimate the model under the null and calculate $\hat{\sigma}_U^B$, U^B, and $U^B/\hat{\sigma}_U^B$.
	(c)	Repeat steps (a) and (b) multiple times, each time saving the value of the standardized test statistic $U^B/\hat{\sigma}_U^B$. Define the bootstrap critical value for a 5 percent significance level test to be the 95th percentile of the $U^B/\hat{\sigma}_U^B$.
4.		Compare $U/\hat{\sigma}_U$, the actual value of the statistic, with the bootstrap critical value.

heteroskedasticity by resampling from the estimated residuals at each strike price or we can use the wild bootstrap (see Wu, 1986 or Härdle, 1990). The procedures are applicable with the obvious modifications for a general confidence level α. Algorithms for constructing percentile-t confidence intervals may be constructed with modest additional effort. Table 2 summarizes the bootstrap algorithm for implementing the test in Proposition 2.

3. Numerical results

3.1. Simulations

In order to solve the various constrained optimization problems described in this paper, we used GAMS—general algebraic modeling system (see Brooke et al., 1992) which is a general package for solving a broad range of linear, nonlinear, integer, and other optimization problems subject to constraints.

In their simulations, Aït-Sahalia and Duarte (2003) calibrate their model using characteristics of the S&P options market. We calibrate our experiments using DAX

options in January 1999 which expire in February of that year. At that time the DAX index was in the vicinity of 5000. The 25 distinct strike prices range over the interval 4400–5600 in 50 unit increments. We set the short term interest rate r to 3.5%, the dividend yield δ to 2%, the time to maturity τ to 0.15 and the current price (value) of the index S to 5100. We assume the volatility smile function is linear in the strike price, i.e., $\sigma = 0.4 - 0.00025 * (x - 4400)$. Let $F = S \exp((r - \delta)\tau)$ be the forward price. Define $d_1 = (\log(F/x) + \sigma^2\tau/2)/\sigma\sqrt{\tau}$ and $d_2 = d_1 - \sigma\sqrt{\tau}$. Then the 'true' call function is given by $m(x) = \exp(-r\tau)(F\Phi(d_1) - x\Phi(d_2))$, where Φ is the standard normal cumulative distribution function. At each strike price, the residual standard deviation is set to 10% of the option price. For a given observation with strike price x_i, the 'observed' option price is given by $y_i = m(x_i) + 0.1m(x_i)\varepsilon_i$, where the ε_i are i.i.d. standard normal.

We have already seen the 'true' call function, its first derivative and SPD plotted in Figs. 1A–C above. In each of the simulations below we assume there are 20 observations at each of the 25 strike prices for a total of 500 observations. Figs. 2A–C illustrate the impact of constraints on estimation. The 'smooth' estimator imposes only the Sobolev constraint as in Eqs. (6) and (7) with the degree of smoothness determined by applying cross-validation. Monotonicity and convexity constraints are imposed using Eqs. (11) and (12). Finally, we impose 'tail' constraints similar to (13) which require the estimated SPD to be nondecreasing over the lowest five strike prices and nonincreasing over the highest five. One of the purposes is to improve the estimator of the SPD at the boundaries. In each case the '90% point-wise intervals' contain the central 90% of estimates from 100 replications. The '90% uniform intervals' are obtained by taking the central 99.6% of the estimates at each of the 25 distinct strike prices ($0.996^{25} \simeq 0.90$).

As may be seen in Fig. 2A, the improvement in estimation of the call function resulting from adding constraints is barely discernible. Fig. 2B illustrates the impacts on estimation of the first derivative. The benefits of adding shape constraints are clearly evident (note that the vertical scale narrows as one moves down the figure). The most dramatic impact of the constraints is on estimation of the SPD as may be seen in Fig. 2C. Smoothness alone produces a very broad band of estimates, so much so that the true SPD looks quite flat. Adding monotonicity and convexity improves the estimates substantially, though they are quite imprecise at low strike prices. This is in part due to the much larger variance there. The 'tail' constraints alleviate this problem.

Consistent with Fig. 2, the average MSE of estimating the SPD falls dramatically—indeed by three orders of magnitude—when smoothness, monotonicity, convexity and tail constraints are imposed. Even supplementing the smoothness constraint with monotonicity, convexity and unimodality reduces the MSE of the SPD estimator by an order of magnitude.

Next we turn to confidence interval estimation. With multiple observations at each strike price and under the assumptions of Proposition 3, asymptotic confidence intervals can be estimated using (21). For bootstrap intervals we use the percentile method outlined in Table 1. In each case we performed 100 bootstrap draws. For the call function, the asymptotic confidence intervals are reasonable though somewhat
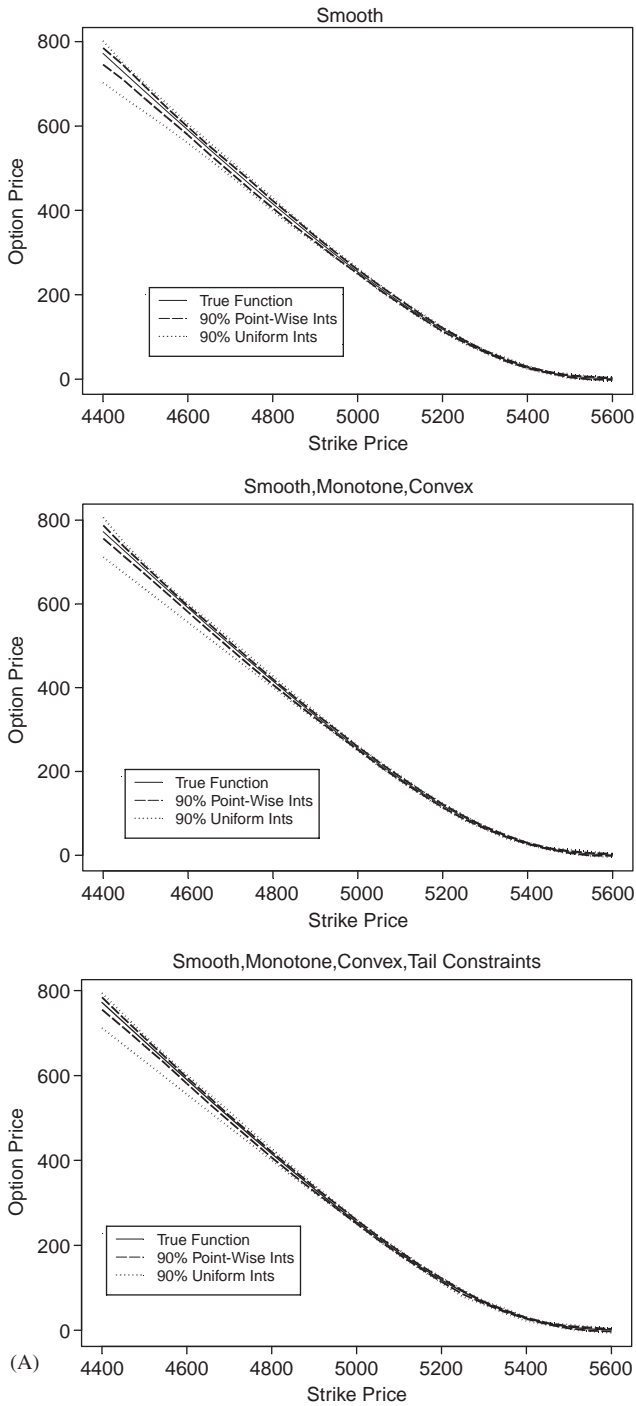
Fig. 2. (A) Effects of constraints on call function estimates; (B) effects of constraints on first derivative estimates; (C) effects of constraints on SPD estimates.
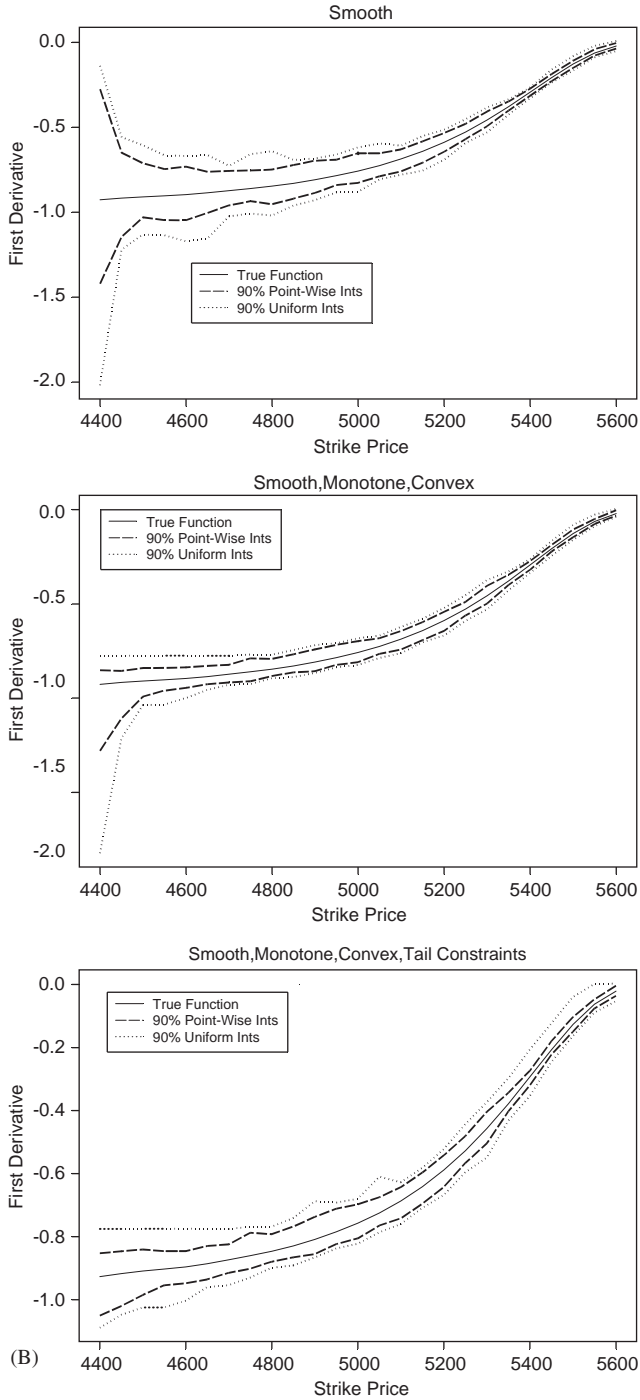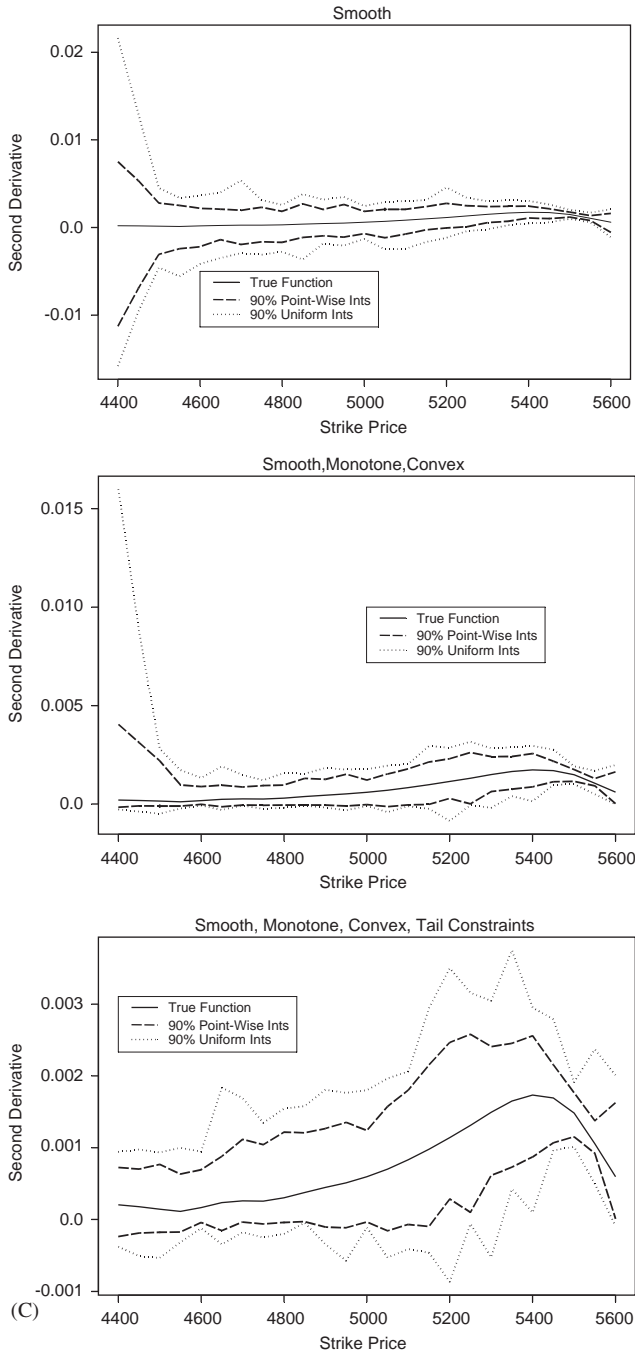
Fig. 2. (*Continued*)

Fig. 2. (*Continued*)

Table 3
Bootstrap vs asymptotic confidence interval coverage frequencies

	Nominal coverage probability		
	---	---	---
	0.99	0.95	0.90
	Coverage frequency (averaged over X's)		
Call function			
Bootstrap	0.97	0.93	0.89
Asymptotic	0.92	0.85	0.78
First derivative			
Bootstrap	0.97	0.94	0.89
Asymptotic	1.00	1.00	1.00
State price density			
Bootstrap	0.98	0.96	0.92
Asymptotic	1.00	1.00	1.00

Smoothing parameter selected using cross-validation.

broader than the bootstrap intervals. However, for the first derivative and particularly the SPD estimates, confidence intervals based on the asymptotic approximation are extremely poor relative to the bootstrap intervals. The underlying reason is that the asymptotic approximation—just as the point mean and divided difference estimators in Figs. 1A–C does not incorporate smoothness constraints.

Next we turn to the accuracy of the bootstrap intervals. We produced 500 samples and in each case obtained 100 bootstrap re-samples. The model that was estimated was the fully constrained version with smoothness, monotonicity, convexity and 'tail' constraints. The smoothness constraint was again determined through cross-validation. We calculated bootstrap and asymptotic confidence intervals. Coverage frequencies are summarized in Table 3. When estimating the call function itself, the average width of asymptotic and bootstrap confidence intervals are of comparable magnitude though the bootstrap yields more accurate coverage probabilities. When estimating the first derivative or SPD, asymptotic intervals are several orders of magnitude wider than bootstrap intervals.

In previous work, Li and Wang (1998), considered asymptotic and bootstrap versions of a residual regression test of a parametric null against a nonparametric alternative. They found that the actual mean of the asymptotically $N(0,1)$ test statistic to be below zero and that the distribution was better approximated by the bootstrap. (For details, see Li and Wang (1998, page 155, Tables 1–2).) We performed simulations of the residual regression test of monotonicity and convexity in Proposition 2, implementing both the asymptotic and (wild) bootstrap versions of the procedure (see Table 2). We too found that the mean of the test statistic was below zero. For example, with 25 distinct strike prices and the data generating mechanism described at the beginning of this section, the mean and variance of the test statistic $U/\hat{\sigma}_U$ of Proposition 2 were -0.27 and 0.97, respectively, resulting in rejection probabilities below nominal values under the null. Rejection probabilities under the null hypothesis are contained in Table 4.

Table 4
Residual regression test rejection probabilities: H_o is true

Nominal	Asymptotic	Bootstrap
0.01	0.002	0.012
0.05	0.016	0.030
0.10	0.042	0.082
0.20	0.104	0.168
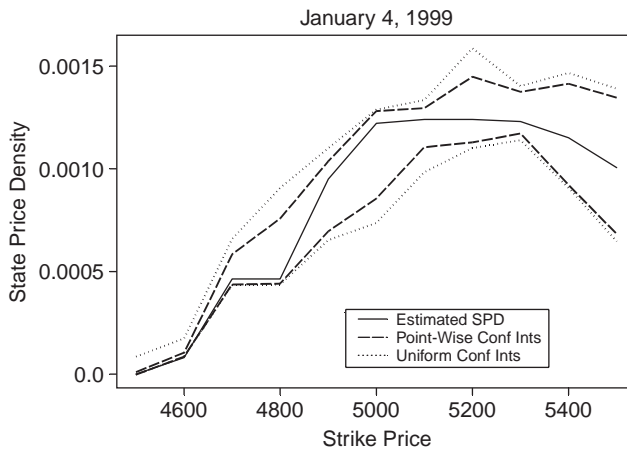
Fig. 3. SPD using DAX index data.

3.2. Applications to DAX index option data

In this section we use the tools we have described to analyze data on DAX index options for Monday, January 4, 1999. On that day, the DAX closed at 5290.

We restricted our estimates to call options which trade at 100 point intervals between 4500 and 5500 and expire at the end of February. Some trades do indeed take place outside this range of strike prices, but there are few of them. Our data contained a total of 902 observations. The smoothness parameter was selected using cross-validation.

Fig. 3 illustrates the estimated SPD along with 90% point-wise bootstrap confidence intervals for the January fourth data. In addition we have included uniform confidence bounds for the SPD by taking the central 99% of the estimates at each of the 11 distinct strike prices ($0.99^{11} \simeq 0.90$). With the DAX closing at 5290, one can expect that there is some probability mass beyond the 5500 level. Options at 5500 averaged 22Euro suggesting that the market assigned a positive probability to the DAX exceeding 5500 at time of expiration of the options. Indeed our estimated SPD integrates to about 0.8. We also performed a test of the constraints using the procedure outlined in Proposition 2. The resulting statistic, which is asymptotically

standard normal under the null, was -0.43 suggesting that the monotonicity, convexity and 'tail' constraints would appear to be consistent with the data.

4. Conclusions

In this paper, we propose a nonparametric least squares estimator for modeling option prices and state price densities as a function of the strike price. The estimator readily incorporates various constraints such as monotonicity and convexity. We outline a 'residual regression' type test of these properties. We also propose bootstrap procedures for constructing confidence intervals around the estimated call function and its derivatives.

Our methods can be readily extended in at least two ways. First, time series structure can be introduced into the residuals without complicating the optimization problem in Eqs. (6) and (7). Second, given appropriate data, the call function can be estimated as a function of additional explanatory variables, such as the current asset price and the time to expiry. The nonparametric least squares problem in (7) remains as given, but the representor matrix R must now be calculated using representors which are functions of several variables. This can be readily accommodated using a generalization of the Sobolev inner product in (4). The resulting multivariate representors are simply products of the univariate representors used in this paper. The details of these kinds of multivariate nonparametric regression procedures may be found in Yatchew and Bos (1997).

Acknowledgements

Appendix. Proof of Propositions

Proof of Proposition 1. By Yatchew and Bos (1997), Theorem 3.1.2. $\hat{m}$ converges to m in mean squared error at the indicated rate of convergence. By Theorem 2.3 in Yatchew and Bos (1997), all functions in the estimating set have derivatives up to order 3 uniformly bounded in sup-norm. Combining these two results ensures that $\hat{m}$, $\hat{m}^{(1)}$ and $\hat{m}^{(2)}$ convergence in sup-norm. $\quad\square$

Proof of Proposition 2. See Yatchew and Bos (1997) Theorem 5.1.1. $\quad\square$

Proof of Proposition 3. First note that $R^{-1}\bar{y}(X)$ solves the unconstrained version of optimization problem (18):

$$\min_c \tfrac{1}{n}[y - BRc]^{\mathrm{T}}\Sigma^{-1}[y - BRc].$$

Now let $\hat{c}$ minimize (18) subject to (11) and (12). Since $m(x)$ is a linear combination of the representors at $X = (X_1, \ldots, X_k)$, $\hat{m}(x) = \sum_{j=1}^{k} \hat{c}_j r_{x_j}(x)$ and its first two derivatives consistently estimate their true counterparts. Hence, in addition, $\hat{m}(X)^{\mathrm{T}} R^{-1} \hat{m}(X) \xrightarrow{P} m(X)^{\mathrm{T}} R^{-1} m(X) < L$ and as sample size increases, the smoothness constraint becomes nonbinding in probability. Furthermore, because the first and second derivatives are estimated consistently, strict monotonicity and convexity implies that these constraints also become nonbinding in probability. Thus $\lim_{n \to \infty} Prob\lfloor \hat{m}(X) = \bar{y}(X) \rfloor = 1$.

Using conventional central limit theorems, $n^{1/2}(\bar{y}(X) - m(X)) \xrightarrow{D} \mathrm{N}(0, \Omega)$ and Eq. (20) follows immediately. □

References

Aït-Sahalia, Y., 1996. Nonparametric pricing of interest rate derivative securities. Econometrica 64, 527–560.

Aït-Sahalia, Y., Duarte, J., 2003. Nonparametric option pricing under shape restrictions. Journal of Econometrics 116, 9–47.

Aït-Sahalia, Y., Lo, A., 2000. Nonparametric risk management and implied risk aversion. Journal of Econometrics 94, 9–51.

Barle, S., Cakici, N., 1998. How to grow a smiling tree. Journal Financial Engineering 7, 127–146.

Black, F., Scholes, M., 1973. The pricing of options and corporate liabilities. Journal of Political Economy 81, 637–659.

Breeden, D., Litzenberger, R., 1978. Prices of state-contingent claims implicit in option prices. Journal of Business 51, 621–651.

Broadie, M., Detemple, J., Ghysels, E., Torrés, O., 2000a. Nonparametric estimation of American options' exercise boundaries and call prices. Journal of Economic Dynamics and Control 24, 1829–1857.

Broadie, M., Detemple, J., Ghysels, E., Torrés, O., 2000b. American options with stochastic dividends and volatility: a nonparametric investigation. Journal of Econometrics 94, 53–92.

Brooke, A., Kendrick, D., Meeraus, A., 1992. GAMS. Scientific Press, Redwood City, CA.

Cont, R., 2001. Empirical properties of asset returns, stylized facts and statistical issues. Quantitative Finance 1, 223–236.

Cont, R., da Fonseca, J., 2002. Dynamics of implied volatility surfaces. Quantitative Finance 2, 45–60.

Cox, J., Ross, S., 1976. The valuation of options for alternative stochastic processes. Journal of Financial Economics 3, 145–166.

Daglish, T., 2003. A pricing and hedging comparison of parametric and non-parametric approaches for American index options. Journal of Financial Econometrics 1 (3), 327–364.

Derman, E., Kani, I., 1994. Riding on the smile. Risk 7, 32–39.

Fan, Y., Li, Q., 1996. Consistent model specification tests: omitted variables and semiparametric functional forms. Econometrica 64, 865–890.

Garcia, R., Gencay, R., 2000. Pricing and hedging derivative securities with neural networks and a homogeneity hint. Journal of Econometrics 94, 93–115.

Ghysels, E., Patilea, V., Renault, E., Torres, O., 1997. Nonparametric methods and option pricing, core. Discussion paper, 9775.

Hall, P., 1992. The Bootstrap and Edgeworth Expansion. Springer, New York.

Härdle, W., 1990. Applied Nonparametric Regression, Econometric Society Monograph Series, vol. 19. Cambridge University Press, Cambridge.

Härdle, W., Kleinow, T., Stahl, G. (Eds.), 2002. Applied Quantitative Finance: Theory and Computational Tools. Springer, Berlin.

Jackwerth, J., Rubinstein, M., 1996. Recovering probability distributions from option prices. Journal of Finance 51, 1611–1632.

Li, Q., Wang, S., 1998. A simple consistent bootstrap test for a parametric regression function. Journal of Econometrics 87, 145–165.

Lucas, R., 1978. Asset prices in an exchange economy. Econometrica 46, 1429–1446.

Merton, R., 1973. Rational theory of option pricing. Bell Journal of Economics and Management Science 4, 141–183.

Rubinstein, M., 1976. The valuation of uncertain income streams and the pricing of options. Bell Journal of Economics 7, 407–425.

Stone, C., 1980. Optimal rates of convergence for nonparametric estimators. Annals of Statistics 8, 1348–1360.

Stone, C., 1982. Optimal global rates of convergence for nonparametric regression. Annals of Statistics 10, 1040–1053.

Wahba, G., 1990. Spline Models for Observational Data, CBMS-NSF Regional Conference Series in Applied Mathematics, #59, Society for Industrial and Applied Mathematics, USA.

Wu, C.-F.J., 1986. Jackknife, bootstrap and other resampling methods in regression analysis (with discussion). Annals of Statistics 14, 1261–1350.

Yatchew, A., 2003. Semiparametric Regression for the Applied Econometrician. Cambridge University Press, New York.

Yatchew, A.J., Bos, L., 1997. Nonparametric regression and testing in economic models. Journal of Quantitative Economics 13, 81–131 (see also www.economics.utoronto.ca/yatchew/).

Zheng, J., 1996. A consistent test of functional form via nonparametric estimation techniques. Journal of Econometrics 75, 263–289.

On the Difficulty to Design Arabic E-learning System in Statistics *

Taleb Ahmad - Wolfgang Härdle - Julius Mungo

Humboldt-Universität zu Berlin

Department of Statistics

Spandauer Str. 1

D-10178 Berlin

March 22, 2006

Abstract

In this paper, we present a case study, which describe the development of the Statistic e-learning-course in Arabic language –"Arabic MM*STAT". The basic frame for this E-book, the system MM*STAT was developed at the School for Business and Economics of Humboldt-Universität zu Berlin. Arabic MM*STAT uses a HTML - based filing card structure. We discuss the difficulties of the implementation of such a system in to the standard WWW formats and present the solutions needed for Arab educational institutions and the Arabic user. Those solutions are consistent with the Arabic language, and include the modern trend in the e-learning environment.

Keywords: electronic books, Arabtex, MM*STAT, Statistical software.

JEL Classification I21, C19

1   Introduction

E-Learning is the modern model of education, transferring the classical text books to an electronic medium. It has become an important stream of education in present and has a promising future. E-Learning uses net work technology to manage, design, deliver, select, and extend classical learning. Our approach, based on the MM*STAT, provides a flexible e-learning Statistic environment, created at the institute of Statistics and Economics of the Humboldt Universität zu Berlin. It provides e-learning of Statistics in many languages that include English, German, French, Polish, Czech, Spanish, Italian ; see **http://www.md-stat.com**. Our main aim in this project is to extend e-learning concept to include the Arabic language. Originally the MD*Book is a flexible user interface tool for generating user interactive e-learning documents such as MM*STAT. This tool uses the LaTeX format file to enable compilation into PDF, Java, HTML. We aim to use the LaTeX file format for creating an Arabic text. As we will show this remains a problem, basically caused by the incompatibility of the standard Latex2html with the phenomenons of the Arabic typography. The learning system (Arabic MM*STAT) has been implemented in the website: **http://www.md-stat.com**. In what follows, we present in section 2 an Arabic MM*STAT description. And discuss in section 3 the difficulties to design the Arabic MM*STAT which involves the Arabic style and the technical problems. In section 4 we summarize our work, given a direction for the future on Arabic LaTeX.

2   Arabic MM*STAT

Arabic MM*STAT is a Web-based learning platform that provides a fully integrated student environment to learn statistics. MM*STAT allows an interactive introduction into the world of statistics. The goal of Arabic MM*STAT designer is :

1. To contribute to the evolution of Statistics e-learning by the development of an efficient system.

2. To create an Arabic MM*STAT, because there were not enough Arabic statistic e-learning platforms. The purpose of this study is to develop a platform that will be used completely in Arabic, together with other languages such as English and German.

We know that the Arabic language is different from English and other languages in a number of respects:

1. Arabic language is written from right to left.

2. It is possible to form hundreds of words from one root word (al- Fedaghi and al-Sadoun, 1990).

3. In Arabic the definite article and prepositions, are not separated from their following word by a space, on another hand some words have different meanings but have the same written form.

4. The peculiar morphology of Arabic might render methods used for English text retrieval inappropriate. For example, the English phrase "and she wrote it" comprising of four words would be written in Arabic as one word "wakatabathu" (wa=and, kataba=wrote, t=she,hu=it). Moukdad, H., Andrew, L., (2001). Generally, the difficulty with someone who has translated a software product into another language to get a realistic picture of the challenges. Not only are there vocabularies the translator might not know, but they also must adapt sentence structure, significance, and colloquialisms to convey equivalent meaning.

3   Difficulties to design, an Arabic MM*STAT

We now discuss some practical and technical problems. The development of MM*Stat by (Müller, M., Rönz, B., Ziegenhagen, U., 2000) has been the first step in the transition from traditional textbooks to integrated learning environments. This project is with multiple languages and our goal is to apply this project in Arabic language. To do this, one needs Arabic Windows XP version, as a tool to write the Arabic code. We then have to translate our articles in Arabic Microsoft Word, and save the files with unicode (file.unicode). Another problem is the direction. Writing in Arabic is from right to left but the direction in HTML appears from left to right, which technical problem for working in Arabic. What is the solution for this problem? To solve this problem we use the code $dir = "rtl"$ in HTML file at every paragraph and link - Figure - Table and the statistics forms, then the direction of the writing changes from right to left.

MD*Book Software environment requires LaTeX. Another problem is the LaTeX Format. Arab text does not work with LaTeX . We aim to finding a solution to this problem in our project. There was a project for Arabtex in LaTeX presented by Prof. Klaus Lagally - Universitt Stuttgart. This project was a good step but it is not enough, because the letter in Arab text in this project was written with English, not with Arabic and this project did not have the ability to reduce Statistics on an Arabic Website. This project uses only PDF, not HTML. To summarize, the goal of my work is to create an electronic book in Arabic using LaTeX source code and the MD*Book tool.

Arabic MM*STAT Realisation

Figure 1 presents Arabic MM*STAT - Cover Page. Figure 2 presents Arabic MM*STAT - Content Page. The user can enter the courses topics via a list of contents and can go to any desired topic or course chapter. By hovering the mouse pointer over a selected course file for a few seconds will appear to identify the course topic.

Figure 1: Arabic MM*STAT - Cover Page
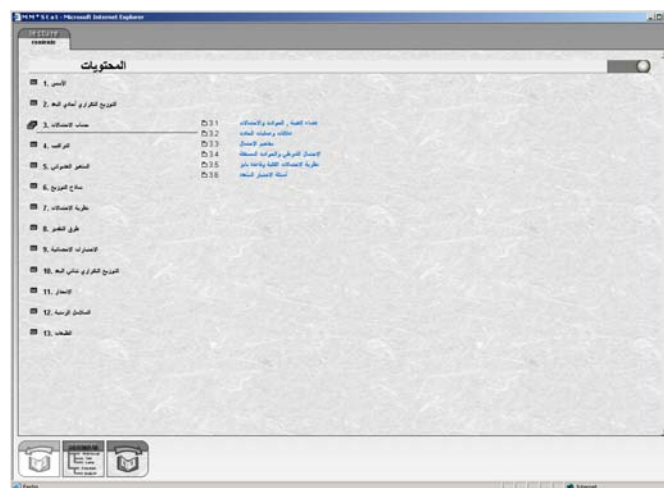
Figure 2: Arabic MM*STAT - Content Page

Figure 3 shows a graphical user interface (GUI) of a statistic topic in Arabic MM*STAT

and Figure 4 shows the corresponding English version.
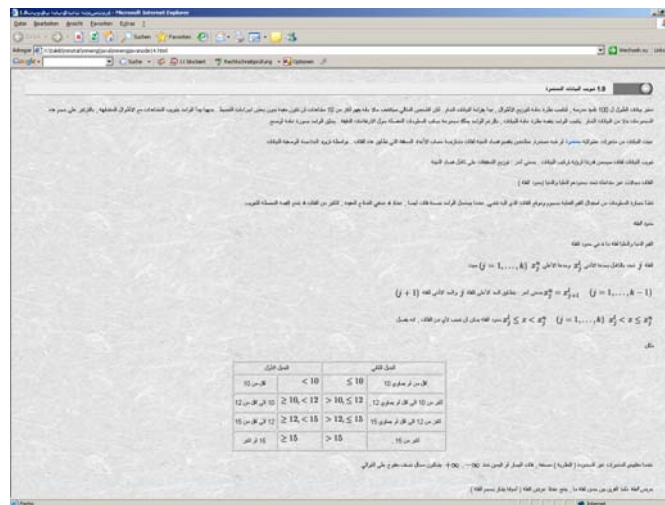
Figure 3: Arabic MM*Stat: Layer Architecture and Screenshot
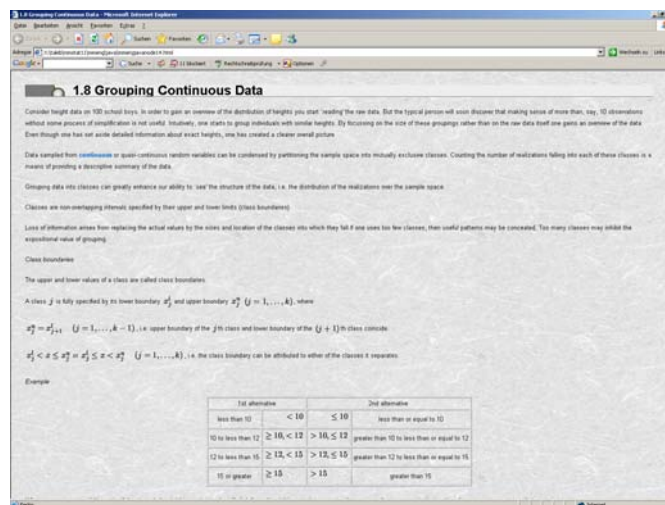
Figure 4: English MM*Stat: Layer Architecture and Screenshot

The Arabic MM*STAT Course has five components structures; the lecture, information, explained, enhanced and interactive. Each lecture gives the basic concepts of the general statistical theory, definitions, formulae, mathematical proofs. Arabic MM*Stat provides the students to compute distribution functions, graphics and derive results for statistical tests. The students or anyone interested in statistics can interactively learn about basic concepts of statistics at any time and anywhere.

4 Conclusion

This is the possibility of creating an e-learning system with Arabic MM*STAT. And discussed the problems encountered, both in the context of the Arabic language itself and the technical problems, which is the nonexistence of a possibility to use LaTeX code in Arabic. Our next goal is to explore possible solutions to the problem of Arabic LaTeX. Actually the goal has to do with XML.

References

AbdulJaleel, N., Larkey, L., *English to Arabic Transliteration for Information Retrieval: A Statistical Approach*, Center for Intelligent Information Retrieval Computer Science, University of Massachusetts.

Al-Fedaghi, S and H. Al-Sadoun. Morphological compression of Arabic text. Information Processing Management 26(2): 30316., 1990.

Bates, T., *Forum on National strategies for E-learning*, IIEP Virtual Institute, 21 October to 15 November 2002.

Jemni, M., Bahattab, A., *A collaborative approach for development of arabic courses for E-learning, a case study of Tunisian-Saudi Arabian experience*.

Hammad, S., AL-ayyoub,A.,Sarie,T., *combining existing E-learning components towards an ivle*, Arab Open University, Amman.

Härdle,W., Klinke,S., Ziegenhagen,U., *eLearning - A Critical Review*, Humboldt-Universität zu Berlin, January 5, 2006.

Lagally, K., *ArabTeX : a System for Typesetting Arabic : User Manual Version 3.00*, Universitt Stuttgart, 1993.

Mungo, J., *e-learning/e-teaching An Implementation and Evaluation of a Finance Introductory Course*, Humboldt-Universität zu Berlin, 2004.

Moukdad, H., Andrew, L., Information Retrieval from Full-Text Arabic Databases,Can Search Engines Designed for English Do the Job?,Canada, 2001.

Müller, M., Rönz, B., Ziegenhagen, U., *The Multimedia Project MM*Stat for Teaching Statistics, In COMPSTAT. Proceedings in Computational Statistics*, Physica Verlag,(2000).

Maegaard, B., Choukri, K., Mokbel, C., Yaseen, M., *Language Technology for Arabic*, NEMLAR, Center for Sprogteknolo gi. University of Copenhagen, July 2005.

Zahran, S., *Promotion of Development-Oriented ICT Applications for E-Learning*, Western Asia Preparatory Conference for the World Summit on the Information Society (WSIS), Beirut, 4-6 February 2003.

15 Spatial Statistics

Pavel Čížek, Wolfgang Härdle and Jürgen Symanzik

15.1 Introduction

This chapter deals with the analysis of spatial data. Such data can be the structure of biological cells, the distribution of plants and animals in a geographic region, the occurrence of diseases in a county or state, economic data in different administrative districts, climate data such as temperature or precipitation over geographic regions, and the distribution of galaxies in space. Spatial data often are not independent. Temperature and precipitation measurements at two locations that are 10 km apart will be more similar than such measurements at two locations that are 1000 km or even 100 km apart. Infectious diseases often occur in spatial clusters. One region of a country may encounter hundreds of cases while another region may encounter only very few cases. Thus, spatial data analysis/spatial statistics deals with the quantitative study of phenomena that are located in some two- or higher-dimensional space.

There are three main types of spatial data: spatial point patterns (we speak of spatial point processes when we refer to the underlying stochastic processes that result in observable point patterns), geostatistical data (also called spatially continuous data), and lattice data (also called area data). If there is an additional temporal component, we speak of spatio-temporal data.

Throughout this chapter, we will be dealing with a two-dimensional area of interest, called D. We assume that we have n sample locations $\mathbf{x}_i$ that are usually located in D, but some may occasionally be located outside of D. We make n observations $Z(\mathbf{x}_i)$ at these locations and want to describe the underlying spatial process $Z(\mathbf{x})$.

For spatial point patterns, the point locations in D are random and their number n is unknown in advance. We are primarily interested in the observed locations $\mathbf{x}_i \in D$ that represent particular point events, such as the

occurrence of a disease, a species habitat, or the location of trees. A typical question for spatial point patterns is whether the observed locations are clustered, randomly distributed, or regularly distributed. In this case, we simply assume that $Z(\mathbf{x}_i) = 1$ for all $\mathbf{x}_i \in D$, i.e., there is no other data being recorded than the spatial locations of the n point events. Sometimes, additional variables such as the time of the occurrence of an event or physical measurements such as height, diameter, and crown-defoliation of a tree are measured. Then, $Z(\mathbf{x}_i)$ represents a random variable (or a random vector) at location $\mathbf{x}_i \in D$. Details on spatial point patterns can be found, for example, in Diggle (2003).

For geostatistical data, D is a fixed subset of the two-dimensional space. We observe $Z(\mathbf{x}_i)$ that represents a random variable (or a random vector) at location $\mathbf{x}_i \in D$. We are usually not interested in the locations $\mathbf{x}_i$ at which the sample measurements were taken. In fact, we assume the locations have been randomly chosen or are predetermined by towns or measurement stations. Of interest are only the univariate or multivariate measurements $Z(\mathbf{x}_i)$, such as amount of precipitation or number and concentration of different air pollutants measured at a particular measurement station. Because the number of locations n where actual measurements were taken is sparse, we usually want to estimate the values for the variable(s) of interest at locations where no actual measurements were taken. The term spatially continuous data relates to the fact that measurements at least theoretically can be taken at any location $\mathbf{x}_i \in D$. However, time and cost constraints usually only allow measurements to be taken at a limited number of locations. Details on geostatistical data can be found, for example, in Isaaks and Srivastava (1989) and Wackernagel (1998).

For lattice data, D is a fixed subset of the two-dimensional space that consists of countably many sub-areas of regular or irregular shape. We observe $Z(\mathbf{x}_i)$ that represents a random variable (or a random vector) at sub-area $\mathbf{x}_i \in D$. Lattice data represents spatial data that has been aggregated to areal units. Instead of knowing the exact locations where a disease occurred, or the exact residence addresses and income of all employees, we may only know the total number of occurrences (or the incidence rate) of a disease for different health service areas, or the average annual income of employees in different countries of the European Community. Statistical variables associated with areas often do not vary continuously over space. For example, the tax rate on goods might be 6% in one state and 8% in the next state. Often (but not always), the given spatial areas are considered to be the only spatial locations at which the variables of interest can be measured. We typically want to describe and model the observed spatial pattern and determine possible explanations for this pattern, for example, which other factors might be associated with higher

disease rates in many of the urban health service areas.

In addition to the specific references cited above, the reader is referred to Bailey and Gatrell (1995), Ripley (1981), or Cressie (1993) for comprehensive overviews dealing with all three main types of spatial data and beyond. This chapter closely follows Chapter 14 in Venables and Ripley (1999) and, therefore, the notation used in Ripley (1981). The XploRe quantlets used in this chapter have been adapted from the S-Plus and underlying C code that accompanies Venables and Ripley (1999), with kind permission from Brian D. Ripley. The XploRe quantlets support techniques for spatial point patterns and geostatistical data. In the electronic version of this chapter, the **Q** symbol underneath each figure provides a direct link to the XploRe code that was used to create the figure. At this point, the XploRe quantlib spatial does not support any techniques for lattice data.

In Section 15.2, we discuss techniques for the analysis of geostatistical data. In Section 15.3, we discuss techniques for the analysis of spatial point processes. We finish with a short discussion in Section 15.4.

15.2   Analysis of Geostatistical Data

In this section, we discuss techniques for the analysis of geostatistical data, in particular spatial interpolation, smoothing, kriging, correlograms, and variograms. We assume that we have n fixed sampling locations $\mathbf{x}_i \in D$ and observe $Z(\mathbf{x}_i)$. Our goal is to predict the spatial process $Z(\mathbf{x})$ for any $\mathbf{x} \in D$ or, rather, the mean value $E(Z(\mathbf{x}))$ for this spatial process for any $\mathbf{x} \in D$. A simple, but often unrealistic, assumption is that $\{Z(\mathbf{x}) \mid \mathbf{x} \in D\}$ are independent and the distributions of $Z(\mathbf{x})$ only differ in the mean but otherwise are identical. More realistically, we should assume some degree of spatial correlation and therefore incorporate spatial dependence into our model. This leads to first order effects and second order effects. First order effects relate to variation in the mean value of a process in space, i.e., a global (or large scale) trend. Second order effects relate to the spatial correlation structure, i.e., these are local (or small scale) effects.

Our examples in this section are based on the `topo` dataset, introduced in Davis (1973), Table 6.4, and further discussed in Ripley (1981, p. 58–72). This dataset consists of $n = 52$ spatial locations with measurements of topographic elevation, measured in feet above sea level. The area D is an approximate square with side lengths of about 305 feet. The coordinates have been labelled in 50 feet units, such that the actual spatial locations $\mathbf{x}_i$ fall into the area $(0.2, 6.3) \times (0.1, 6.2)$, where the first represent East-West

coordinates and the second represent North-South coordinates. For convenience, the origin $(0, 0)$ is assumed to be located in the Southwest corner of the area D. In all of our calculations and plots, we assume that the area of interest actually is the square area $(0, 6.5) \times (0, 6.5)$. The examples in this section show how to do computations and produce graphics similar to those in Sections 14.1 and 14.2 in Venables and Ripley (1999), using XploRe.

15.2.1   Trend Surfaces

A trend surface analysis is a simple approach to model the global first order effect in the mean value of a spatially continuous process. First, we assume that $\{Z(\mathbf{x}) \mid \mathbf{x} \in D\}$ are independent. Let us further express the spatial coordinates as $\mathbf{x} = (x, y)$. Often, the expected value $E(Z(\mathbf{x})) = f(\mathbf{x})$ of a spatial process is expressed as a trend surface, i.e., a polynomial regression surface of the form

$$f(\mathbf{x}) = \sum_{0 \leq r+s \leq p} a_{rs} x^r y^s,$$

where the parameter p is called the order (or the degree) of the trend surface. For example, a quadratic trend surface is of the form

$$f(\mathbf{x}) = a_{00} + a_{10}x + a_{01}y + a_{20}x^2 + a_{11}xy + a_{02}y^2$$

and has $P = 6$ coefficients. In general, a least squares (LS) trend surface of degree p has $P = (p+1)(p+2)/2$ coefficients that are chosen to minimize

$$\sum_{i=1}^{n} (Z(\mathbf{x}_i) - f(\mathbf{x}_i))^2,$$

where $\mathbf{x}_i \in D, i = 1, \ldots, n$, are the locations at which the n sample measurements were taken.

Figure 15.1 shows trend surfaces of degree 2, 3, 4, and 6 fitted to the `topo.dat` dataset. The 52 spatial locations are overlaid as small red crosses on the trend surfaces. This figure looks similar to Figure 14.1 in Venables and Ripley (1999). When fitting a trend surface of degree $p = 6$, $P = 28$ coefficients have to be estimated from only 52 spatial locations. As a consequence, problems with the visual appearance of the trend surface of degree 6 can be noticed due to extrapolation near the edges.

Another problem with fitting trend surfaces is that often the spatial locations are not regularly spaced. Instead, the spatial locations are often more dense where the surface is high (e.g., for air pollution, we want to determine the spatial locations where the pollution level might be critical for human health
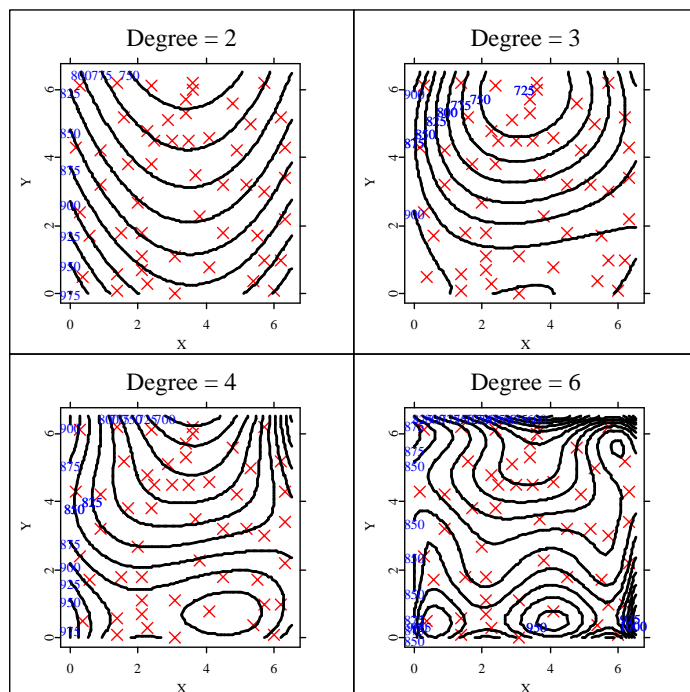
Figure 15.1: Least squares trend surfaces of degree 2, 3, 4, and 6 for the topo.dat dataset.

XCSspa01.xpl

and thus we take additional measurements in regions with elevated pollution levels). It becomes necessary to consider the spatial correlation of the errors as well. Venables and Ripley (1999) suggested the model

$$Z(\mathbf{x}) = \mathbf{f}(\mathbf{x})^\top \boldsymbol{\beta} + \varepsilon(\mathbf{x}),$$

where $\mathbf{f}(\mathbf{x}) = (f_1(\mathbf{x}), \ldots, f_P(\mathbf{x}))^\top$ is a parametrized trend term as above, $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_P)^\top$ is a parameter vector, and $\varepsilon(\mathbf{x})$ is a zero-mean spatial stochastic process of errors. In our previous example of a quadratic trend surface, we have $\mathbf{f}(\mathbf{x}) = (1, x, y, x^2, xy, y^2)^\top$ and $\boldsymbol{\beta} = (a_{00}, a_{10}, a_{01}, a_{20}, a_{11}, a_{02})^\top$. It is further assumed that $\varepsilon(\mathbf{x})$ possesses second moments and has the covariance function

$$C(\mathbf{x}, \mathbf{y}) = Cov(\varepsilon(\mathbf{x}), \varepsilon(\mathbf{y})),$$

where $\mathbf{x}, \mathbf{y} \in D$. If we assume that we obtain measurements at n spatial locations $\mathbf{x}_1, \ldots, \mathbf{x}_n \in D$, we can define the model

$$\mathbf{Z} = \mathbf{F}\boldsymbol{\beta} + \boldsymbol{\varepsilon},$$

where

$$\mathbf{Z} = \begin{bmatrix} Z(\mathbf{x}_1) \\ \vdots \\ Z(\mathbf{x}_n) \end{bmatrix}, \quad \mathbf{F} = \begin{bmatrix} \mathbf{f}(\mathbf{x}_1)^{\top} \\ \vdots \\ \mathbf{f}(\mathbf{x}_n)^{\top} \end{bmatrix}, \quad \text{and} \quad \boldsymbol{\varepsilon} = \begin{bmatrix} \varepsilon(\mathbf{x}_1) \\ \vdots \\ \varepsilon(\mathbf{x}_n) \end{bmatrix}.$$

If we further define $K = [C(\mathbf{x}_i, \mathbf{x}_j)]$ and assume that K has full rank, we can estimate $\boldsymbol{\beta}$ by generalized least squares (GLS), i.e., by minimizing

$$(\mathbf{Z} - \mathbf{F}\boldsymbol{\beta})^{\top} K^{-1} (\mathbf{Z} - \mathbf{F}\boldsymbol{\beta}).$$

Additional mathematical details regarding the underlying calculations can be found in Venables and Ripley (1999, p. 436) and in Ripley (1981, p. 47). The resulting GLS trend surface of degree 2 is shown in the upper right of Figure 15.2 and it is contrasted to the ordinary least squares trend surface of the same degree in the upper left (a replicate from Figure 15.1). This figure looks similar to Figure 14.5 in Venables and Ripley (1999). The choice of the covariance function C that is needed for these calculations is discussed in Section 15.2.3.

15.2.2   Kriging

Kriging is a technique developed in the early 1960s by Matheron and his school in the mining industry. The term kriging is named after the South African mining engineer D. G. Krige. When we speak of kriging, we basically mean an optimal prediction of unobserved values of the spatial process

$$Z(\mathbf{x}) = \mathbf{f}(\mathbf{x})^{\top} \boldsymbol{\beta} + \varepsilon(\mathbf{x}),$$

as introduced before.

Several forms of kriging exist. In its most general form, called universal kriging, the process $Z(\mathbf{x})$ is fitted by generalized least squares, predicting the value of the functional term as well as the value of the error term at location $\mathbf{x} \in D$ and taking their sum. In contrast to the generalized least squares approach discussed in Section 15.2.1, it is no longer assumed that $\varepsilon(\mathbf{x})$ is a zero-mean process of errors.
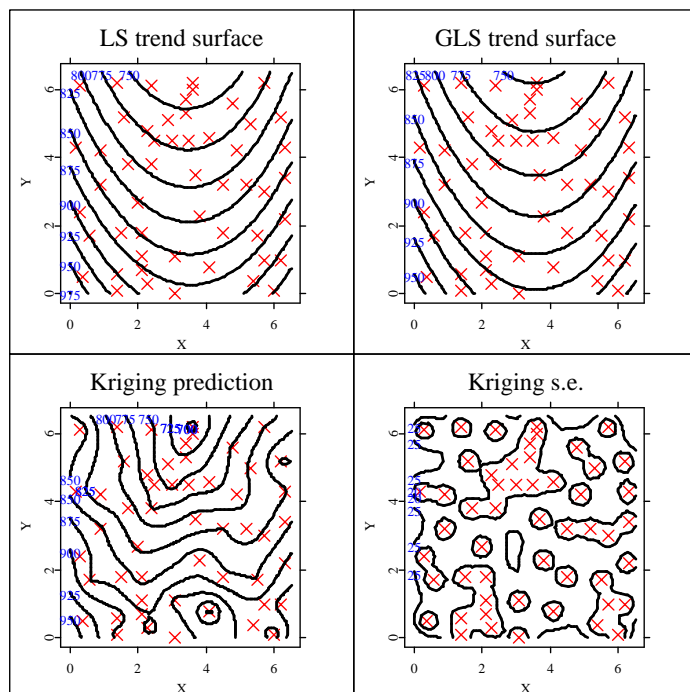
Figure 15.2: Least squares trend surface (upper left), generalized least squares trend surface (upper right), kriging surface (lower left), and kriging standard errors (lower right) for the `topo.dat` dataset.

Q XCSspa02.xpl

A simplification of universal kriging is (ordinary) kriging where the trend surface is assumed to be of degree 0, i.e., the process $Z(\mathbf{x})$ has constant mean $E(Z(\mathbf{x})) = \mu$ for all $\mathbf{x} \in D$, with $\mu \in \mathbb{R}$ unknown. To meet this condition, the process $Z(\mathbf{x})$ often is first detrended as described in Section 15.2.1 and the spatial residuals are used for the next analysis steps.

A further simplification of ordinary kriging is simple kriging where it is assumed that $\mu \in \mathbb{R}$ is known in advance and thus does not have to be estimated from the data. Other forms of kriging, such as robust kriging, disjunctive kriging, block kriging, and co-kriging have been discussed in the literature.

Details can be found in Bailey and Gatrell (1995) or Cressie (1993) for example.

The computational steps for universal kriging and the calculation of the error variances have been discussed in Venables and Ripley (1999, p. 439), with additional mathematical details provided in Ripley (1981, p. 47–50). Universal kriging predictions (based on a trend surface of degree 2) and kriging standard errors are displayed in the lower left and lower right of Figure 15.2. There are considerable differences in the appearances of the kriging surface and the previously introduced least squares trend surface (upper left) and the generalized least squares trend surface (upper right). This figure looks similar to Figure 14.5 in Venables and Ripley (1999). The choice of the covariance function C that is needed for these calculations is discussed in Section 15.2.3.

15.2.3   Correlogram and Variogram

In the previous sections, we have made the often unrealistic assumption that the covariance function

$$C(\mathbf{x}, \mathbf{y}) = Cov(\varepsilon(\mathbf{x}), \varepsilon(\mathbf{y}))$$

is known. However, if C is unknown, we have to make additional assumptions for the underlying spatial process that will allow us to estimate C.

One common assumption for second order effects is that the spatial process $\{Z(\mathbf{x}) \mid \mathbf{x} \in D\}$ exhibits stationarity (also called homogeneity). This means that the statistical properties of $Z(\mathbf{x})$ are independent of the absolute locations in D, i.e., $E(Z(\mathbf{x})) = \mu$ and $Var(Z(\mathbf{x})) = \sigma^2$ are constant in D. This also implies that $Cov(Z(\mathbf{x}), Z(\mathbf{y}))$ for $\mathbf{x} \neq \mathbf{y}$ depends only on the relative locations (i.e., distance and direction) of $\mathbf{x}$ and $\mathbf{y}$, but it does not depend on their absolute locations in D. Thus, for any $\mathbf{x}, \mathbf{y} \in D$ with $\mathbf{x} + \mathbf{r} = \mathbf{y}$, it holds that $Cov(Z(\mathbf{x}), Z(\mathbf{y})) = C(\mathbf{r})$.

Moreover, we call the spatial process $\{Z(\mathbf{x}) \mid \mathbf{x} \in D\}$ isotropic if it is stationary and $Cov(Z(\mathbf{x}), Z(\mathbf{y}))$ depends only on the distance r between $\mathbf{x}$ and $\mathbf{y}$, but it does not depend on the direction in which they are separated. If mean, variance, or covariance differ over D, the spatial process exhibits nonstationarity (also called heterogeneity).

A somewhat weaker assumption than stationarity (or even isotropy) is intrinsic stationarity where we assume that $E(Z(\mathbf{x} + \mathbf{r}) - Z(\mathbf{x})) = 0$ and $Var(Z(\mathbf{x} + \mathbf{r}) - Z(\mathbf{x})) = 2\gamma(\mathbf{r})$ for any separation vector $\mathbf{r}$. Thus, we have constant variance in the differences between values at locations that are separated by a given distance and direction $\mathbf{r}$. The quantity $2\gamma(\mathbf{r})$ is called the

variogram while the quantity $\gamma(\mathbf{r})$ is called the semi-variogram. Often, the prefix "semi" is omitted and we speak of a variogram although we mean a semi-variogram. The factor 2 has the effect that $\gamma(\mathbf{r}) = \sigma^2$ (and not $2\sigma^2$) for large separation vectors $\mathbf{r}$, i.e., $\lim\limits_{||\mathbf{r}||\to\infty} \gamma(\mathbf{r}) = \sigma^2$.

If we assume isotropy, the covariance function

$$Cov(Z(\mathbf{x}), Z(\mathbf{y})) = c(d(\mathbf{x}, \mathbf{y})) = c(r)$$

becomes a function c that entirely depends on the Euclidean distance $r = ||\mathbf{r}|| = d(\mathbf{x}, \mathbf{y})$ between $\mathbf{x}$ and $\mathbf{y}$. Common choices for theoretical covariance functions c are discussed further below.

Given that $c(0) > 0$, we define

$$\rho(r) = \frac{c(r)}{c(0)} = \frac{c(r)}{\sigma^2}$$

and call this function the correlogram (also known as autocorrelation function) at distance r. It should be noted that

$$\sigma^2 = Var(Z(\mathbf{x})) = Cov(Z(\mathbf{x}), Z(\mathbf{x})) = c(d(\mathbf{x}, \mathbf{x})) = c(0).$$

Apparently, it must hold that $\rho(0) = 1$. However, there is no reason why $\lim\limits_{r\to 0+} \rho(r)$ could not be less than one for very small distances r. If this is the case, we speak of a nugget effect that represents a microscale variation of the spatial process. A possible explanation of such an effect are measurement errors, i.e., if multiple measurements at location $\mathbf{x}$ were made, they would all be slightly different. Extensions of theoretical covariance functions that include a nugget effect can be found in the previously cited literature. We will include a nugget effect in some of our examples later in this section, though, and ask the reader to carefully check what happens for very small distances r.

One other relationship between semi-variogram γ and covariance function C (for stationary processes) should be noted. It holds that

$$
\begin{aligned}
\gamma(\mathbf{r}) \;&=\; \frac{1}{2} Var(Z(\mathbf{x} + \mathbf{r}) - Z(\mathbf{x})) \\
&=\; \frac{1}{2}\left\{ Var(Z(\mathbf{x} + \mathbf{r})) + Var(Z(\mathbf{x})) - 2C(Z(\mathbf{x} + \mathbf{r}), Z(\mathbf{x})) \right\} \\
&=\; \sigma^2 - C(\mathbf{r}).
\end{aligned}
$$

While this holds for the theoretical relationship, it should be noted that this does not hold in general for the empirical functions, i.e., commonly we have $\hat{\gamma}(\mathbf{r}) \neq \hat{C}(0) - \hat{C}(\mathbf{r})$, where $\hat{\gamma}$ and $\hat{C}$ relate to the estimates of γ and C.

Before we consider theoretical covariance functions, we take a closer look at empirical correlograms and empirical variograms for isotropic processes. For both of these, we first divide the range of the data into m bins $[a_i, b_i)$ that are usually equally wide. For an empirical correlogram, for each bin i, we then determine the covariance $\hat{c}$ for those pairs of sample locations $\mathbf{x}, \mathbf{y} \in D$ where $d(\mathbf{x}, \mathbf{y}) \in [a_i, b_i)$. We finally divide by the overall variance to obtain $\hat{\rho}$ for bin i. Similarly, for an empirical variogram, for each bin i, we determine the average squared difference between $Z(\mathbf{x})$ and $Z(\mathbf{y})$ for those pairs of sample locations $\mathbf{x}, \mathbf{y} \in D$ where $d(\mathbf{x}, \mathbf{y}) \in [a_i, b_i)$. More formally,

$$\hat{c}(r_i) = \frac{1}{n(r_i)} \sum_{d(\mathbf{x}_i, \mathbf{x}_j) \in [a_i, b_i)} (Z(\mathbf{x}_i) - \overline{x})(Z(\mathbf{x}_j) - \overline{x}),$$

where $\overline{x} = 1/n \sum_{i=1}^{n} Z(\mathbf{x}_i)$ is the mean of all observed sample values and $n(r_i)$ is the number of the pairs of locations $\mathbf{x}_i, \mathbf{x}_j \in D$ that have a separation distance that falls into the interval $[a_i, b_i)$. r_i for which $\hat{c}$ is calculated is usually chosen as the mean of all separation distances that fall into bin i (although variations in the literature also choose $r_i = (a_i + b_i)/2$, i.e., the midpoint of the interval $[a_i, b_i)$). Similarly,

$$2\hat{\gamma}(r_i) = \frac{1}{n(r_i)} \sum_{d(\mathbf{x}_i, \mathbf{x}_j) \in [a_i, b_i)} (Z(\mathbf{x}_i) - Z(\mathbf{x}_j))^2.$$

A common way to suppress unreliable estimates of $\hat{\rho}$ and $\hat{\gamma}$ is to use the calculated results only for bins i that contain at least six pairs of sample locations.

An empirical correlogram and an empirical variogram for the residuals of the `topo.dat` dataset, based on a least squares quadratic trend surface as described in Section 15.2.1, are shown in Figure 15.3. This figure looks similar to Figure 14.6 in Venables and Ripley (1999).

General requirements for theoretical covariance functions for general spatial processes are symmetry and non-negative definiteness. For stationary processes, or even isotropic processes, exponential, Gaussian, and spherical families of covariance functions meet the general requirements and are frequently fitted to the data. An exponential covariance function has the form

$$c(r) = \sigma^2 \exp\left(-\frac{r}{d}\right).$$

An exponential covariance function has been overlaid on the correlogram
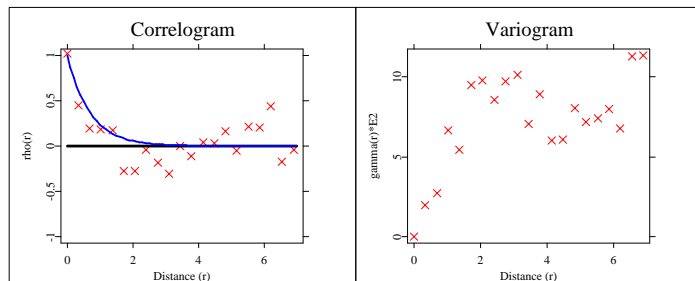
Figure 15.3: Empirical correlogram (left) and empirical variogram (right) for the residuals of the `topo.dat` dataset after fitting a least squares quadratic trend surface.

XCSspa03.xpl

(left) in Figure 15.3. A Gaussian covariance function has the form

$$c(r) = \sigma^2 \exp\left(-\frac{r2}{d^2}\right).$$

A spherical covariance function in two dimensions has the form

$$c(r) = \sigma^2 \left\{ 1 - \frac{2}{\pi} \left(\frac{r}{d} \sqrt{1 - \frac{r^2}{d^2}} + \sin^{-1} \frac{r}{d} \right) \right\} \boldsymbol{I}_{(r \in [0,d])}$$

and a spherical covariance function in three dimensions (but also valid as a covariance function in two dimensions) has the form

$$c(r) = \sigma^2 \left(1 - \frac{3r}{2d} + \frac{r^3}{2d^3} \right) \boldsymbol{I}_{(r \in [0,d])},$$

where $\boldsymbol{I}_{(r \in [0,d])}$ is the indicator function that takes value 1 if $r \in [0, d]$ and 0 otherwise. Within XploRe, the first version of these spherical covariance functions is used in case the optional parameter "D" in the XploRe quantlet call is omitted or takes the value 2, whereas the second version is used for all other values of "D".

In the formulas above, the fixed distance $d = d(\mathbf{x}, \mathbf{y})$ is called the range and basically represents the lag distance beyond which $Z(\mathbf{x})$ and $Z(\mathbf{y})$ are uncorrelated. In this context, σ^2 is called the sill and represents the variance of the process $Z(\mathbf{x})$ as defined above.
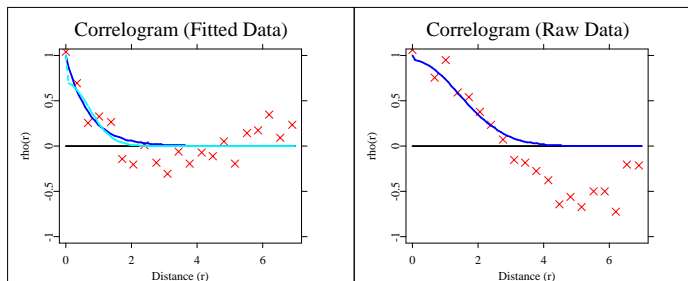
Figure 15.4: Empirical correlograms for the `topo.dat` dataset, based on residuals from a quadratic trend surface (left) and the raw data (right). An exponential covariance function (solid blue line) and a Gaussian covariance function (dashed cyan line) have been fitted to the empirical correlogram of the residuals (left) while a Gaussian covariance function has been fitted to the empirical correlogram of the raw data (right).

XCSspa04.xpl

Figure 15.4 shows two empirical correlograms with overlaid covariance functions. In the left plot, we look at an empirical correlogram for the residuals from a quadratic trend surface. An exponential covariance function [1] (solid blue line) and a Gaussian covariance function with some nugget effect [2] (dashed cyan line) have been overlaid. In the right plot, we look at an empirical correlogram for the raw data. A Gaussian covariance function [3] with a small nugget effect and wide range has been overlaid. This figure looks similar to Figure 14.7 in Venables and Ripley (1999).

The effect on the kriging surface and kriging standard errors when using [1] as the covariance function has already been shown in Figure 15.2 in the bottom row. In Figure 15.5, we see the effect on the kriging surface and kriging standard errors when using covariance function [2] in the top row and when using covariance function [3] in the bottom row. While all three kriging surfaces look similar in the center of our region of interest D, we can notice considerable differences on the edges of D. However, considerable differences can be noticed in the structure of the kriging standard errors in the center of D, depending on which of the three covariance functions has been selected. Figure 15.5 looks similar to Figure 14.8 in Venables and Ripley (1999).
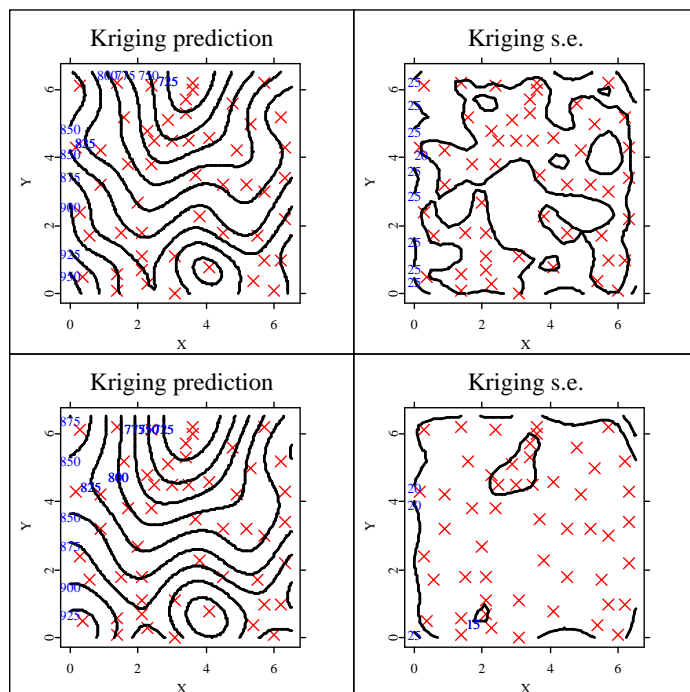
Figure 15.5: Kriging predictions (left) and kriging standard errors (right) for the residuals of the `topo.dat` dataset using covariance function [2] (top) and for the raw data without fitting a trend surface using covariance function [3] (bottom).

XCSspa05.xpl

15.3   Spatial Point Process Analysis

In this section, we discuss techniques for the analysis of spatial point patterns. A spatial point pattern is a set of randomly observed locations $\mathbf{x}_i$ that are more or less irregularly distributed in the area of interest D. One of the questions we are interested in when analyzing a spatial point pattern is whether the observed locations $\mathbf{x}_i$ are completely randomly distributed in D or whether there exists some spatial pattern, e.g., clustering or regularity.

Our examples in this section are based on the `pines` dataset, introduced

in Strand (1972) and further discussed in Ripley (1981, p. 172–175). This dataset consists of $n = 71$ spatial locations, representing pine saplings in a 10×10-meter square. The spatial locations of the pines are shown in the upper left plot of Figure 15.6. The examples in this section show how to do computations and produce graphics similar to those in Section 14.3 in Venables and Ripley (1999), using XploRe.

One way to describe a spatial point process is to describe the number of events that are occurring in arbitrary sub-areas D_i of D, i.e.,

$$\{Z(D_i) \mid D_i \subseteq D\},$$

where $Z(D_i)$ is a random variable that represents the number of events in sub-area D_i. Obviously, this means that $E(Z(D_i))$ and any higher-order moments depend on the size(s) of the sub-areas that are involved, which is of little practical use. Instead, we characterize spatial point patterns based on the limiting behavior of these quantities per unit area. First order properties are described as intensity $\lambda(\mathbf{x})$ that represents the mean number of events per unit area at point $\mathbf{x}$, i.e.,

$$\lambda(\mathbf{x}) = \lim_{dx \to 0} \frac{E(Z(\mathbf{dx}))}{dx},$$

where $\mathbf{dx}$ is a small area around point $\mathbf{x}$ and $dx = ||\mathbf{dx}||$ is the size of area $\mathbf{dx}$. For a stationary point process, we have $\lambda(\mathbf{x}) = \lambda$ for all $\mathbf{x} \in D$, i.e., the intensity is constant over D. An obvious estimate of λ then is $\hat{\lambda} = n/A$, where n is the number of observed points in D and $A = ||D||$ is the size of area D.

A point process that is used to model a completely random appearance of the point pattern in D is the homogeneous Poisson process. Given such a process, we speak of complete spatial randomness (CSR). For the process $\{Z(D_i) \mid D_i \subseteq D\}$, it holds that $Z(D_i)$ and $Z(D_j)$ are independent for any choices of $D_i, D_j \in D$. The probability distribution of $Z(D_i)$ is a Poisson distribution with mean value λA_i where $A_i = ||D_i||$ is the size of area D_i. The corresponding probability density function (pdf) of $Z(D_i)$ is

$$f_{Z(D_i)}(z) = \frac{(\lambda A_i)^z}{z!} e^{-\lambda A_i},$$

where λ represents the constant intensity over D. CSR usually represents the null hypotheses when we want to assess whether an observed point pattern shows clustering or regularity instead.

The second moment of a point process can be specified via the K function, where $\lambda K(r)$ is the expected number of points within distance r of any of the
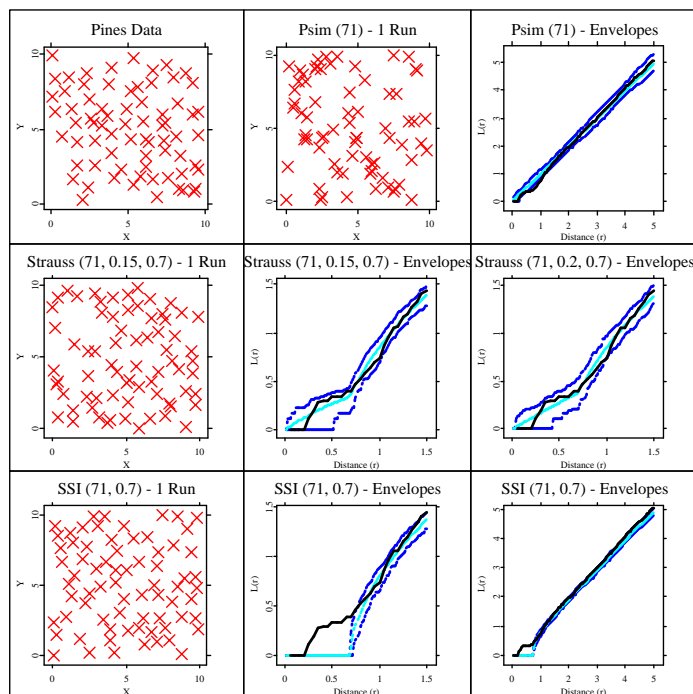
Figure 15.6: Observed point pattern for the `pines.dat` dataset (upper left), and comparisons with a CSR process (upper middle and upper right), two Strauss processes (middle row), and a SSI process (bottom row). In the plots titled "Envelopes", the dark (blue) lines represent the lower and upper simulation envelopes and the light (cyan) line the average of the simulations runs of the specified point process. The solid black line shows the observed $\hat{L}(r)$ for the `pines.dat` dataset.

Q XCSspa06.xpl

points of the observed point pattern. For the homogeneous Poisson process, it can be shown that $K(r) = \pi r^2$. If we obtain values that are significantly greater than this value at a distance r_0 or significantly less than this value at a distance r_0, this would be an indicator of clustering or regularity, respectively, at this distance.

Instead of working with $K(r)$, one often switches to $L(r) = \sqrt{K(r)/\pi}$ (or $\tilde{L}(r) = \sqrt{K(r)/\pi} - r$). For the homogeneous Poisson process, $L(r)$ therefore would be a straight line with slope 1 that goes through the origin (or $\tilde{L}(r) = 0$ for all distances r).

Let us define r_{ij} as the Euclidean distance between the ith and j^{th} of the observed point locations $\mathbf{x}_i$ and $\mathbf{x}_j$ and let $\boldsymbol{I}_{(r_{ij} \in [0,r])}$ be the indicator function that takes value 1 if $0 \leq r_{ij} \leq r$ and 0 otherwise. A possible estimator of $K(r)$ then is

$$\hat{K}(r) = \frac{A}{n^2} \sum_{i,j;i \neq j} \boldsymbol{I}_{(r_{ij} \in [0,r])}.$$

This estimator reveals one problem for points that are close to the edge of D. The summation excludes pairs of points for which the second point is outside of D and thus is unobservable. Therefore, we need some edge-correction. Let us consider a circle centered on point $\mathbf{x}_i$, passing through point $\mathbf{x}_j$. Let w_{ij} be the proportion of the circumference of the circle that lies within D, i.e., this represents the conditional probability that a point is observed in D, given that it is within distance r_{ij} from the ith point. Therefore, an edge-corrected estimator of $K(r)$ is

$$\hat{K}(r) = \frac{A}{n^2} \sum_{i,j;i \neq j} \frac{\boldsymbol{I}_{(r_{ij} \in [0,r])}}{w_{ij}}.$$

An edge-corrected estimator of $L(r)$ can be obtained as $\hat{L}(r) = \sqrt{\hat{K}(r)/\pi}$ (or $\hat{\tilde{L}}(r) = \sqrt{\hat{K}(r)/\pi} - r$).

A graphical assessment for CSR is to plot $\hat{L}(r)$ (or $\hat{\tilde{L}}(r)$) versus r. Any major departure from a straight line at distance r_0 then indicates clustering (if above the straight line) or regularity (if below the straight line), respectively. Apparently, an empirical plot based on observed point locations will rarely produce a straight line, even if the observed pattern originates from a homogeneous Poisson process. Therefore, we work with simulation envelopes of the assumed process.

To produce simulation envelopes, we simulate the assumed point process s times, say 100 times, in our examples. At each distance r, the minimum, average, and maximum, i.e., $l(r) = \min_{i=1,\ldots,s} \hat{L}_i(r)$, $a(r) = \operatorname*{avg}_{i=1,\ldots,s} \hat{L}_i(r)$, and $u(r) = \max_{i=1,\ldots,s} \hat{L}_i(r)$ are calculated, where $\hat{L}_i(r)$ is the estimate of $L(r)$ in the ith simulation run. We call $l(r)$ and $u(r)$ the lower and upper simulation

envelopes of the simulated process. If we treat $\hat{L}(r)$ as a random variable, then we can approximate that

$$P(\hat{L}(r) > u(r)) = P(\hat{L}(r) < l(r)) = \frac{1}{s+1}.$$

Thus, for $s = 100$ independent simulation runs, we obtain a two-sided test at the approximate 2% significance level that rejects the null hypothesis of CSR if the observed $\hat{L}(r)$ falls above the upper or below the lower simulation envelope at some distance r.

We first make $s = 100$ independent simulation runs of a Poisson process with $n = 71$ observations that inhibit the same spatial domain as the original data. We draw the result of the first simulation run in the upper center plot in Figure 15.6. It is expected that for a Poisson process, we visually get the impression that points are clustered in some sub-areas of D while other sub-areas of D only contain few points.

The upper right plot in Figure 15.6 shows the lower and upper simulation envelopes as well as the average of the $s = 100$ simulation runs. The observed $\hat{L}(r)$ for the `pines.dat` dataset (the solid black line) falls below the lower simulation envelope for distances $r \in [0.7, 1.2]$ approximately. This means, our observed process is significantly different from a homogeneous Poisson process. Instead, we have some regularity for distances from around 0.7 to 1.2 meters.

Common departures from CSR towards clustering processes, such as heterogeneous Poisson processes, Cox processes, and Poisson cluster processes, can be found in the previously cited literature on spatial point patterns. Here, we take a closer look at two processes that are used as models for regularity, i.e., Strauss processes and sequential spatial inhibition processes.

Many regular point patterns can be described by simple inhibition rules. For example, the cell centers of animal cells cannot be located closer than the diameter of the cells, or two plants in close proximity are less likely to survive and grow to maximum height (although this is not impossible) than plants that are further apart. A general class of spatial point patterns that describes such processes are Markov point processes. A subclass of such processes are pairwise interaction point processes with Strauss processes as special cases.

In a Strauss process, points in D that are less than some distance $R > 0$ apart are called neighbors. The joint density function for n point locations $\mathbf{x}_1, \ldots, \mathbf{x}_n \in D$ which contains m distinct pairs of neighbors is specified as

$$f(\mathbf{x}_1, \ldots, \mathbf{x}_n) = ab^n c^m,$$

where $a > 0$ is a normalizing constant, $b > 0$ reflects the intensity of the

process, and c with $0 \leq c \leq 1$ describes the interactions between neighbors. Special cases for c are $c = 1$ which gives a homogeneous Poisson process with intensity b and $c = 0$ which results in a simple inhibition process that contains no events at a distance less than or equal to R. All other cases with $0 < c < 1$ represent some form of regularity. The smaller c and the larger m, the less likely it becomes that another point occurs in proximity to these n points.

In the plots in the middle row of Figure 15.6, we consider two Strauss processes as possible alternatives to CSR. The middle left plot shows one simulation run of a Strauss process with $R = 0.7$ (i.e., the distance at which regularity starts based on the CSR simulation envelopes) and $c = 0.15$. In the middle center plot, we show the lower and upper simulation envelopes as well as the average of the $s = 100$ simulation runs of a Strauss process with these parameters. In the middle right plot, we show the lower and upper simulation envelopes as well as the average of the $s = 100$ simulation runs of another Strauss process with $R = 0.7$ and $c = 0.2$. In both plots, the observed $\hat{L}(r)$ for the `pines.dat` dataset (the solid black line) falls within the lower and upper simulation envelopes of these Strauss processes. Thus, both of these Strauss processes are possible models for the observed point pattern.

Just for illustrative purposes, we also look at $s = 100$ simulation runs of Matern's sequential spatial inhibition (SSI) process in the bottom row of Figure 15.6. The bottom left plot shows one simulation run of a SSI process that prohibits any two points to occur at a distance less than $R = 0.7$. In the bottom center plot and bottom right plot, we show the lower and upper simulation envelopes as well as the average of the $s = 100$ simulation runs of an SSI process with this parameter. While the observed $\hat{L}(r)$ for the `pines.dat` dataset (the solid black line) falls within the lower and upper simulation envelopes of this SSI process for larger values of r, say $r > 2$, shown in the bottom right plot, the observed $\hat{L}(r)$ falls clearly above the upper simulation envelope of this SSI process. This means that a process that requires a minimum distance of $R = 0.7$ between any two points is not a valid model for the observed point pattern. Pine saplings are less likely (than what we would expect under CSR) to occur in proximity of $R = 0.7$ meters or less, but it is not impossible for them to occur in such a close proximity.

15.4 Discussion

In this chapter, we have discussed some basic methods for the analysis of spatial data, in particular for the analysis of geostatistical data and for spatial point patterns. Each of the figures has been produced in XploRe, using quantlets from the XploRe quantlib spatial. A quick glance at the literature introduced in Section 15.1 reveals that the methods discussed here represent only a small fraction of the methods commonly used for spatial data. However, many additional methods can be directly implemented in XploRe as they only require linear algebra and matrix operations. Additional point processes can be simulated by following simple rules. In fact, Symanzik et al. (1998) provided examples of additional user-written spatial functions in XploRe for the analysis of spatial data. Thus, this chapter hopefully has provided some insights how to conduct an analysis of spatial data within XploRe.

15.5 Acknowledgements

Bibliography

Bailey, T. C. and Gatrell, A. C. (1995). *Interactive Spatial Data Analysis*, Longman/Wiley, Burnt Mill, UK/New York, NY.

Cressie, N. A. C. (1993). *Statistics for Spatial Data (Revised Edition)*, Wiley, New York, NY.

Davis, J. C. (1973). *Statistics and Data Analysis in Geology*, Wiley, New York, NY.

Diggle, P. J. (2003). *Statistical Analysis of Spatial Point Patterns (Second Edition)*, Arnold/Oxford University Press, London/New York, NY.

Isaaks, E. H. and Srivastava, R. M. (1989). *An Introduction to Applied Geostatistics*, Oxford University Press, New York, NY.

Ripley, B. D. (1981). *Spatial Statistics*, Wiley, New York, NY.

Strand, L. (1972). A Model for Stand Growth, *IUFRO Third Conference Advisory Group of Forest Statisticians*, Institut National de la Recherche Agronomique (INRA), Paris, 207–216.

Symanzik, J., Kötter, T., Schmelzer, S., Klinke, S., Cook, D. and Swayne, D. F. (1998). Spatial Data Analysis in the Dynamically Linked Arc-View/XGobi/XploRe Environment, *Computing Science and Statistics*, 29(1):561–569.

Venables, W. N. and Ripley, B. D. (1999). *Modern Applied Statistics with S-Plus (Third Edition)*, Springer, New York, NY.

Wackernagel, H. (1998). *Multivariate Geostatistics (Second, Completely Revised Edition)*, Springer, Berlin.

e-Learning Statistics – A Selective Review

Wolfgang Härdle[1], Sigbert Klinke[2], and Uwe Ziegenhagen[3]

[1] Institute for Statistics and Econometrics, Center for Applied Statistics and Economics `haerdle@wiwi.hu-berlin.de`
[2] Institute for Statistics and Econometrics, `sigbert@wiwi.hu-berlin.de`
[3] Institute for Statistics and Econometrics, Center for Applied Statistics and Economics `ziegenhagen@wiwi.hu-berlin.de`

Modern computing equipment is present at schools and universities at all levels of education. [4] In the statistical sciences computers offer great opportunities to enrich the learning process by the means of e.g. animations, software integration or on-the-fly computations.

A personal review of different e-learning platforms for statistics is done in this paper. This review reveals facts that could be taken into account for future e-learning platforms in statistics. One of the most striking discoveries of our analysis is that students of statistics actually do *not* use electronic media in the desired frequency and actually rely more on print media such as books, copies of slides, etc.

Keywords

e-learning, electronic books, hypertext courseware, statistical software

1 Introduction

There has been a plethora of approaches to find the 'holy grail' of e-learning statistics, satisfying the needs of teachers and students as well. In fact the demand of students and teachers for e-learning must be correctly balanced as the following example illustrates.

The aim of a student, for example, is to get immediate and fast access to media which aid him or her in the acquisition of knowledge to pass the exam. The teacher on the other hand is interested in improving his classes and providing the student with knowledge for sophisticated data analyses under

---

2   Wolfgang Härdle, Sigbert Klinke, and Uwe Ziegenhagen

different conditions. These aims are generally conflicting and hardly mappable into a single e-learning system.

In addition each course level requires its special e-learning architecture since the degree of computer literacy and the willingness to perform interactive statistical learning tasks is, of course, a function of this level. At ISE/CASE the education is based on the following elements: introductory statistics and probability theory in the first year, multivariate statistics and first computational steps in the second year and specialisation in a theoretical or applied field (e.g. finance, insurance) and research seminars in the third year. An overview about the course structure is given in Table 1.

Table 1. Overview of the course structure at ISE

Degree	Group	Class	e-Media
Bachelor	Introduction	Statistics I & II	MM*Stat
			Q & A
	Multivariate Statistics	Multivariate Statistics I	e-stat
		XploRe Introductory Course	XploRe
	Applied Statistics	Computerbased Statistics I & II	eBooks
		Data Mining/Statistical Learning	Excel
		Numerical Introductory Course	
	Privatissimum	Privatissimum	
		Bachelor Thesis	
Master	Multivariate Statistics	Multivariate Statistics I	e-stat
		XploRe Introductory Course	XploRe
	Statistics of Fin. Markets	Statistics of Fin. Markets I & II	R
	Advanced Statistics	Multivariate Statistics II	Matlab
		Non- and Semiparametrics I & II	MD*Booklets
		Numerical Introductory Course	
		Applied Quantitative Methods	
	Privatissimum	Privatissimum	
		Master Thesis	
PhD	Financial Statistics	Quantitative Finance Seminar	XploRe
		Adv. Stat. Methods in Finance	eBooks
		Mathematical Statistics Seminar	
		Statistical Tools in Finance and Insurance	

During the last ten years we have developed several systems on our own, in addition we have been involved in a variety of e-learning projects. Based on our experience with these projects we observed that:

1. Interactivity is not appreciated or understood.
2. HTML pages and screenshots of interactive examples are often printed out for take home studies.
3. Discussion groups are rarely used.
4. Feedback through uploading of e.g. homework is often refused.

This might all be a consequence of technical issues (internet browser, operating system), we do believe though that e-learning in statistics should not invest too much optimism but rather provide a thoughtful set of really integrable e-learning examples.

This paper is organized as follows. In Section 2 we describe a variety of e-learning systems, in Section 3 we report our main findings. Section 4 concludes the findings with a recommendation on future e-learning architectures.

2 Modern e-learning materials

2.1 MM*Stat

A first-year student learning statistics needs to be introduced into descriptive statistics and the basic concepts of probability theory and estimation. The highschool training puts him on a level of knowledge which typically incorporates the judgement of roulette or lottery outcomes but not the testing of hypotheses and the graphical display of data. The idea of MM*Stat (`http://www.quantlet.com/mdstat/mmstat.html`) was to directly pick up the highschool knowledge and then to go on with basic concepts of graphical data analysis.

The screenshot in Figure 1 shows screenshots for the Arab and English edition of MM*Stat and depicts the HTML-based filing card structure: Each 'lecture' filing card provides basic concepts of methods considered together with definitions, important formulas and graphics. 'Fully explained', 'Enhanced' and 'Interactive' examples are linked to the lecture card and show how the formulas can be applied. This *MD*Booklet* structure was implemented manually for the German edition, for other languages (see Table 2) an automatic conversion utility [KW2002] had been developed on the basis of LaTeX2HTML.

Fig. 1. Screenshot of MM*Stat, Arab and English version

4   Wolfgang Härdle, Sigbert Klinke, and Uwe Ziegenhagen

Table 2. Languages covered by MM*Stat

Arab (in development)	Czech
Dutch	English
French	German
Italian	Portuguese
Slovenian	Spanish
Turkish	

For the integration of interactive examples a combination of an embedded XploRe Quantlet Client [BHL2005] and a local or remote XploRe Quantlet Server (XQS) is used. Thus MM*Stat enables the student to compute distribution functions, draw histograms or derive results for statistical tests in the web browser.

In our opinion MM*Stat provides certain advantages compared with textbooks. On the one hand printed books are usually restricted to small datasets since it is exhausting to compute e.g. a regression line for a dataset of 100 value pairs, on the other hand the effects of of a parameter change can only be depicted as a sequence of graphs and tables. With MM*Stat, real-world datasets can be used, effects of parameter changes are shown in real-time.

One of the cornerstones of teaching, the repetition of earlier introduced concepts and methods is implemented in two different ways: To see or recall the definition of a used term, a glossary has been written, important words within the different lectures are linked to this glossary. In addition each chapter contains multiple choice questions, via JavaScript these questions are evaluated on-the-fly, the results are shown to the student.

2.2 Electronic Books

To generate an added value for print media the XQS technology used for MM*Stat has been successfully applied to books, which have been designed for 2nd year and above students. In the current implementation each book (see `http://www.xplore-stat.de/ebooks/ebooks.html`), its HTML version and the corresponding set of slides contain references to web pages on a webserver. These HTML pages show the XploRe source code and allow to run the example in a Java applet. In addition to MM*Stat there is an additional mode which allows the student to completely modify the source code of the example.

Figure 2 depicts two screenshots from the book *Applied Multivariate Analysis* [HS2003]. On the left hand side the 'run' version of the quantlet is shown, on the right the 'edit' version.
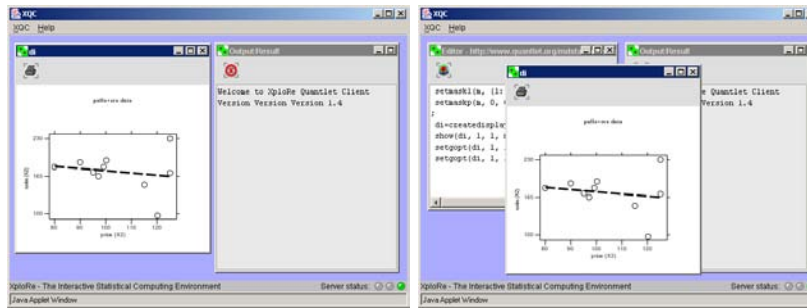
Fig. 2. *execute* and *edit* versions of an 'Applied Multivariate Statistics' example
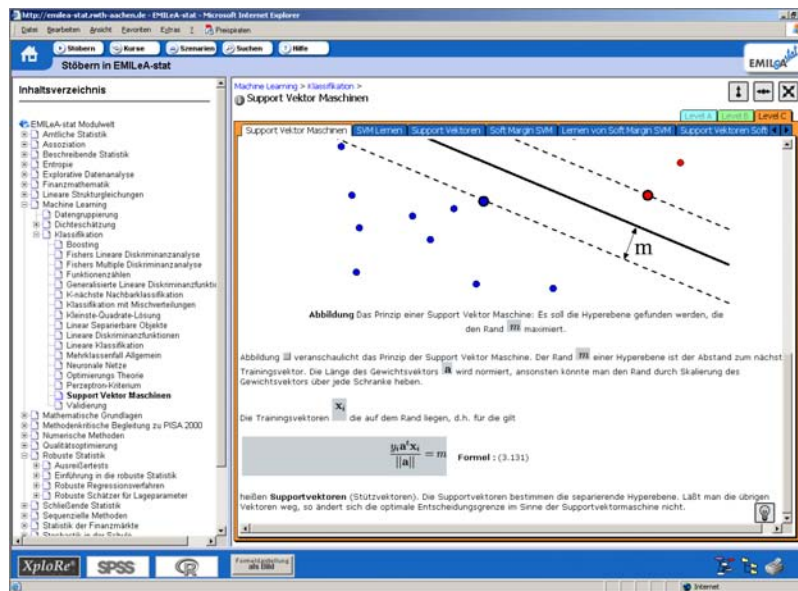
Fig. 3. e-stat: navigation toolbar and SVM example

2.3 e-stat

The aim of the e-stat project (`http://emilea-stat.rwth-aachen.de/`) was to deliver a significant contribution for the improvement of the practical statistical education. The content modules were provided as XML by the project partners, a consortium of nine German universities. The XML code was then equipped with dedicated views and scenarios. The targeted user

6   Wolfgang Härdle, Sigbert Klinke, and Uwe Ziegenhagen

groups were not only students but also people with a general interest in statistics.

The various views and scenarios, which are linked to real-world applications include [CCK2002]:

- Method-based: Learning of methods along a predefined way of modules with terms, examples and exercises
- Problem-based: Based on a simplified description of the problem, a consulting component then proposes a suitable solution
- View-based: Problem description from specific areas serve as examples to deal with the underlying terms

To provide suitable information for different groups of users the e-stat system offers three levels (introductory, applied, advanced) of abstraction with increasing degrees of difficulty. Figure 3 shows a screenshot of the e-stat web site: On the left there is a list of all available modules which are then displayed in the right frame.

2.4 Q&A
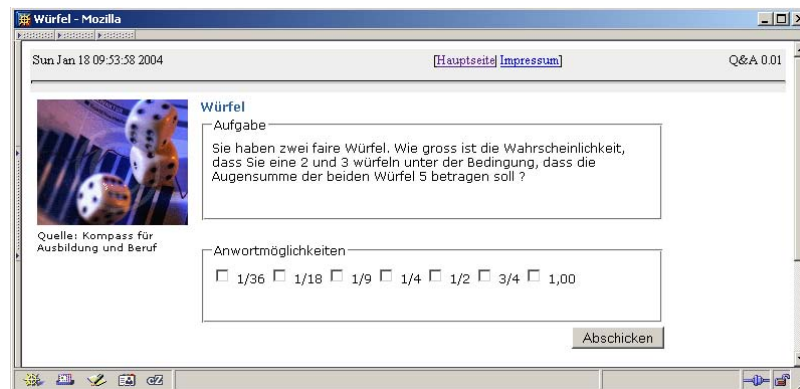
Fig. 4. A 'simple' Q&A exercise: Throwing two dices.

The Q&A system [K2004] provides an interactive environment for the exam preparation in undergraduate courses of statistics. The platform handles two types of exercises: 'simple' and 'variable' exercises. For both types of exercises a problem is presented to the student with multiple choice answers, see Figure 4). The homepage of the project is `http://stirner.wiwi.hu-berlin.de/qa/`.

Q&A allows to check for common student mistakes. Let us consider for example a simple exercise, shown in Figure 4. Assume two dices are thrown.

What is the probability that one dice shows '2' and the other '3' under the condition that the sum of both dices is '5'?

Typically the students make two mistakes. They do not recognize that they have a) to compute a conditional probability and b) take into account that there are two elementary events.

Depending on the answers the student chooses, one can trace back what kind of mistakes they usually make. We categorized the answers of 215 students in seven classes:

1. a) wrong, b) wrong 23%,
2. a) wrong, b) correct 39%,
3. a) correct, b) wrong 6%,
4. a) correct, b) correct 16%,
5. not answered at all/other answers 11%,
6. an answer, which belongs to the other version 5%,

As a reply to the first four groups of errors different web pages with hints on the correct solution are given.

2.5 Moodle
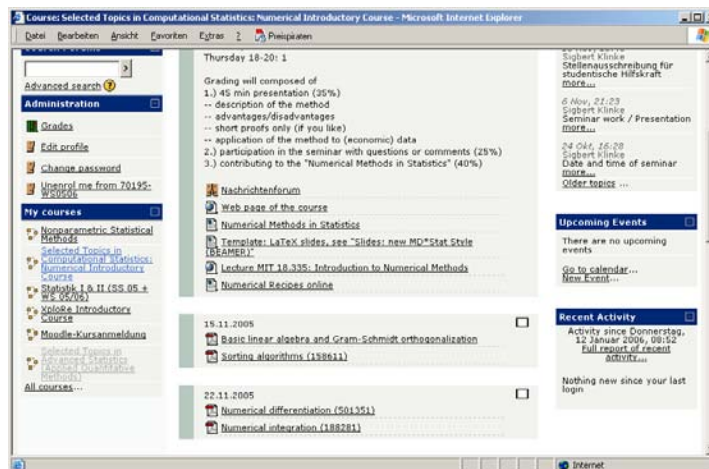
Fig. 5. Numerical Introductory Course in the Moodle system

Moodle [K2005] is a course management system (CMS) - a free, Open Source software package designed using sound pedagogical principles, to help lecturers create effective online learning communities. Moodle has a large and diverse user community with over 75,000 registered users at http://moodle.org, covering 70 languages in 138 countries.

8   Wolfgang Härdle, Sigbert Klinke, and Uwe Ziegenhagen

At Humboldt-Universität zu Berlin the Moodle system is maintained by the 'Multimedia Lern- and Lehrzentrum' from the computing center of the university. At ISE/CASE the work with Moodle started at the beginning of 2005; since that time it has been used for several courses at Bachelor and Master level.

Moodle allows to

1. structure lectures according to a set of topics or a time table,
2. maintain a class forum and forums related to each topic or date of the time table,
3. group students to solve different tasks or exercises,
4. upload and manage documents for lecturers, e.g. lecture slides, and students, e.g. solutions for exercises and
5. generate online exercises with direct evaluation.

The Moodle system, a screenshot is depicted in Figure 5, provides many more features, yet only a few have been used by the ISE. The first question that each lecturer should answer is: Why should one of their student use such a system? Our answer is that we have moved the teaching materials, e.g. slides, information about the grading process from the web pages of the institute to the Moodle system. Thus to download the formulary, the exercise collections or the slides the students have to enter the Moodle system. To receive an account the students have to provide a valid e-mail address, this allows us, together with the email features of Moodle, to make important announcements to the whole class.

2.6 Other Packages

Although spreadsheet software such as Excel may not be the optimal choice for data analysis [MW2005], they are suitable tools for the statistics education. They are available on almost every computer and the majority of the students knows how to operate them. At ISE Excel sheets are mainly used in the education of undergraduate students, e.g. to visualize the Central Limit Theorem or parameter changes of distribution functions. The built-in functions are easily accessible, without knowing a high-level programming language such as Java or C++ formulas may be edited or graphical user interface components such as sliders, radiobuttons and pull-down menus be used. Figure 6 displays a screenshot of a spreadsheet to visualize eight different probability density functions, via sliders the parameters can be modified.
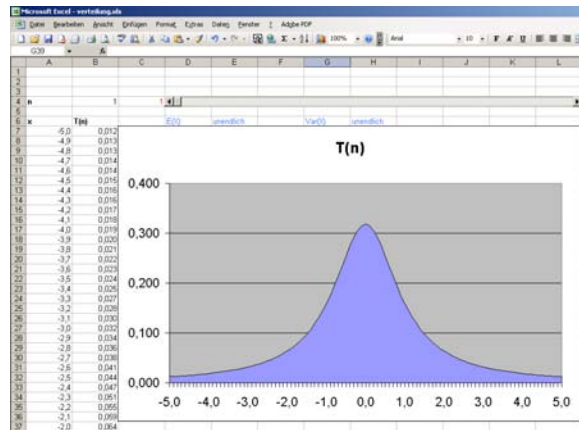
Fig. 6. Visualisation of parameter changes with Excel

DoLStat@d [MYY2003] has been developed at Okayama University, Japan by Yuichi Mori and colleagues. This web based learning system, available on-line at `http://mo161.soci.ous.ac.jp/@d/DoLStat/index.html`, provides various courses which are classified into major categories such as 'General statistics', 'Research field' and 'Statistical method'. Each course contains real world data with their analysis stories, see Figure 7, which are from a data-oriented statistical database system. The provided data sets are designed for the learning purpose of the course and are ordered in educational perspective.

'Neue Statistik' (`http://www.neuestatistik.de/`) is a multimedial teaching platform for statistics for German universities. The aim [G2002] was to replace the formal and mathematical way in the education of undergraduate students by a problem-oriented and practical approach. This includes the usage of animations, diagrams and video sequences. Part of the system is the free teaching software 'Statistiklabor', which is based on R.

3 Evaluation

There are four different aspects in the work with e-learning environments, the technical, the content and the attitudes from lecturers and students.

For MM*Stat the issues arose mostly from the technical side. When MM*Stat was developed, the two leading web browsers were the partly incompatible Netscape 4 and Internet Explorer 5, therefore the work focused on these two programs. Incompatibilities with newer browsers such as Internet Explorer 6 and Mozilla Firebird prevented the successful usage of MM*Stat

10   Wolfgang Härdle, Sigbert Klinke, and Uwe Ziegenhagen
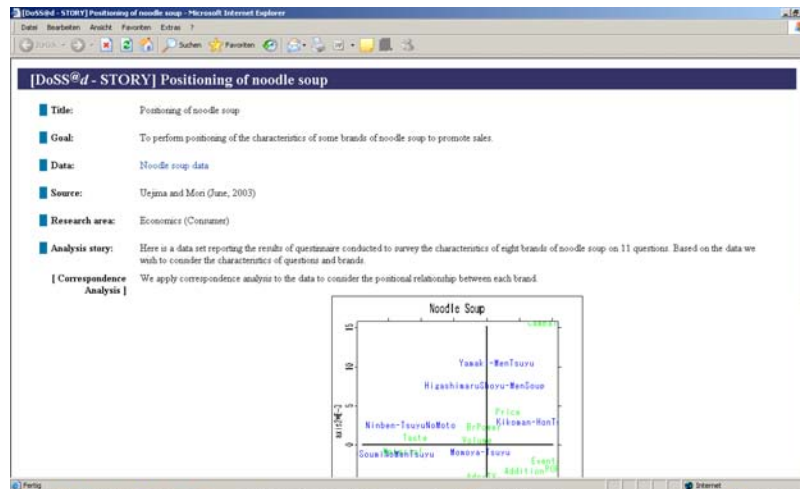
Fig. 7. Screenshot of a DoLStat story

inside the lectures and exercise classes. For languages such as Arab the La-
TeX/LaTeX2HTML software does not provide enough support, here a huge
load of manual work is necessary.

A statistical analysis of questionnaires among students revealed that
MM*Stat partially replaced traditional media such as books for a signifi-
cant proportion of students but did not lead to better grades.

The electronic books are HTML versions of printed books, enriched with
interactive examples. These examples require the Java Runtime Environment
which is an additional barrier for the usage. There were no further adjustments
paying tribute to special requirements of electronic media. The quality of the
content largely depends on the authors. Since the students prefer to have
a printed copy of the book, the eBooks mainly serve as easily searchable
encyclopedia.

The e-stat project mainly suffered from technical and content issues. The
complicated structure of modules requires a certain amount of knowledge,
especially of XML. Furthermore the different modules are diverse, in their
quality and amount. For some modules there are extensive descriptions at all
levels, for a large proportion of modules there are only information at one of
three levels.

The basic idea of Q&A was that the student invests time on thinking about
the correct answer for an exercise, however an analysis of the generated log
files revealed an opposite behavior. Instead of carefully selecting an answer,
the student usually click on the first solution, then the second,. . . until he or
she has found the correct solution. Q&A also requires a significant amount

of work from the lecturer. For each common error a web page with hints and comments has to be written.

Specialized systems such as Moodle may provide advantages especially in terms of options for the multiple choice questions. However, the use of a standard software limits the flexibility in comparison with self-developed solutions, e.g. in the design of the answer pages.

Moodle differs from the other environments since its main purpose is to manage contents for lectures and to provide forums, chats and content storage for various types of lectures. However the complexity of Moodle requires an initial investment from the lecturer as they are forced to structure their lectures, either on topics or a time schedule. Furthermore they have to investigate which of the offered functionality is useful for their classes, as not each class may, e.g. require a student forum.

The use of electronic media for courses also requires adjustments in the way how students learn, they only accept such a technology if they have real advantages. From our experience students rarely use forums; even in lectures with a large audience, e.g. the introductory courses in statistics with nearly 500 students, more than 99% of the entries came from the lecturer himself, a test with anonymous accounts provided same results. Most students prefer to send their questions via email or consult the lecturer personally, but even the number of emails was considerably low. In our opinion the students are afraid of asking a 'foolish' question and rather prefer the direct interaction from face to face. The use of forums even seems to cause, especially when anonymous accounts or webmailer account are enabled, a tendency to use inappropriate language. Several forums had to be closed at ISE due to insulting postings from the students.

4 Conclusion

A few years ago, e-learning in statistics was considered to be an absolute *must have* for the modern education of students. The experience has shown that the goals, which are connected with e-learning elements, are different for students, lecturers and developers. The developers were interested in the technically most sophisticated solutions, the lecturers in the proliferation of their materials and the students in an easy way to achieve enough knowledge to pass the exam.

Nowadays we see e-learning as valuable support tool to aid the learning process, but the problems which may arise in the work with e-learning tools are still manyfold. Besides technical problems such as the limitation to specific web browsers there are pedagogical issues and side effects as well.

12   Wolfgang Härdle, Sigbert Klinke, and Uwe Ziegenhagen

Therefore our conclusion is that:

1. e-learning cannot replace the interaction of student, teacher and blackboard
2. e-learning tools can only be successful if they satisfy the need of all participants of the system

The requirements for excellent electronic media in education are manyfold: Robust and reliable technology, high-quality contents and the willingness to adjust the own behavior from both, the students and the lecturers.

References

[HS2003]   Härdle, W., Simar, L. (2003): Applied multivariate statistical analysis, Springer Verlag, Heidelberg

[BHL2005]   Borak, S., Härdle, W., Lehmann, H. (2005): Working with the XQC. In: Cizek, P., Härdle, W., Weron, R. (eds) Statistical Tools for Finance and Insurance, Springer Verlag, Heidelberg

[CCK2002]   Cramer, E., Cramer, C., Kamps, U. (2002): e-stat: A web-based learning environment in applied statistics. In: Härdle, W., Rönz, B. (eds) COMP-STAT. Proceedings in Computational Statistics - 15th Symposium held in Berlin, Physika Verlag, Heidelberg, p. 309-315

[G2002]   Grune, C. (2002): Neue Statistik, Talk at Learntec 2002, Karlsruhe

[K2004]   Klinke, S. (2004): Q&A - Variable multiple choice exercises with commented answers. In: Antoch, J. (ed.) COMPSTAT. Proceedings in Computational Statistics - 16th Symposium held in Prague, Physika Verlag, Heidelberg, p. 1299-1304

[KW2002]   Klinke, S., Witzel, R. (2002): MD*Book online - A tool for creating interactive documents. In: COMPSTAT. Proceedings in Computational Statistics - 15th Symposium held in Berlin, Physika Verlag, Heidelberg, p. 98-99

[MW2005]   McCullogh, B.D., Wilson, B. (2005): On the accuracy of statistical procedures in Microsoft Excel 2003. In: Computational Statistics and Data Analysis, Volume 49, Number 4, p. 1244-1253

[MYY2003]   Mori, Y., Yamamoto, Y. and Yadohisa, H. (2003): Data-oriented Learning System of Statistics based on Analysis Scenario/Story (DoLStat). Bulletin of the International Statistical Institute (ISI).

[K2005]   Kristöfl, R. (2005): Evaluation von Lernplattformen: Verfahren, Ergebnisse und Empfehlungen, Report of the Austrian ministry for Education, Science and Culture, `http://moodle.de/mod/resource/view.php?id=706`

5 Partially Linear Models

Wolfgang Härdle and Hua Liang

5.1 Introduction

Partially linear models (PLM) are regression models in which the response depends on some covariates linearly but on other covariates nonparametrically. PLMs generalize standard linear regression techniques and are special cases of additive models. This chapter covers the basic results and explains how PLMs are applied in the biometric practice. More specifically, we are mainly concerned with least squares estimators of the linear parameter while the nonparametric part is estimated by e.g. kernel regression, spline approximation, piecewise polynomial and local polynomial techniques. When the model is heteroscedastic, the variance functions are approximated by weighted least squares estimators. Numerous examples illustrate the implementation in practice.

PLMs are defined by

$$Y = X^\top \beta + g(T) + \varepsilon, \tag{5.1}$$

where X and T are d-dimensional and scalar regressors, β is a vector of unknown parameters, $g(\cdot)$ an unknown smooth function and ε an error term with mean zero conditional on X and T.

The PLM is a special form of the additive regression models Hastie and Tibshrani (1990); Stone (1985), which allows easier interpretation of the effect of each variables and may be preferable to a completely nonparametric regression since the well-known reason "curse of dimensionality". On the other hand, PLMs are more flexible than the standard linear models since they combine both parametric and nonparametric components.

Several methods have been proposed to estimate PLMs. Suppose there are n observations $\{X_i, T_i, Y_i\}_{i=1}^n$. Engle, Granger, Rice and Weiss (1986), Heck-

man (1986) and Rice (1986) used spline smoothing and defined estimators of β and g as the solution of

$$\arg\min_{\beta,g} \frac{1}{n} \sum_{i=1}^{n} \{Y_i - X_i^\top \beta - g(T_i)\}^2 + \lambda \int \{g''(u)\}^2 du. \tag{5.2}$$

Speckman (1988) estimated the nonparametric component by $\mathcal{W}\gamma$, where $\mathcal{W}$ is a $(n \times q)-$matrix of full rank and γ is an additional parameter. A PLM may be rewritten in a matrix form

$$Y = X\beta + \mathcal{W}\gamma + \varepsilon. \tag{5.3}$$

The estimator of β based on (5.3) is

$$\widehat{\beta}_S = \{X^\top(I - P_{\mathcal{W}})X\}^{-1}\{X^\top(I - P_{\mathcal{W}})Y\}, \tag{5.4}$$

where $P_{\mathcal{W}} = \mathcal{W}(\mathcal{W}^\top \mathcal{W})^{-1}\mathcal{W}^\top$ is a projection matrix and I is a $d-$order identity matrix. Green, Jennison and Seheult (1985) proposed another class of estimates

$$\widehat{\beta}_{\mathrm{GJS}} = \{X^\top(I - \mathcal{W}_h)X)\}^{-1}\{X^\top(I - \mathcal{W}_h)Y)\}$$

by replacing $\mathcal{W}$ in (5.4) by another smoother operator $\mathcal{W}_h$. Chen (1988) proposed a piecewise polynomial to approximate nonparametric function and then derived the least squares estimator which is the same form as (5.4). Recently Härdle, Liang and Gao (2000) have systematically summarized the different approaches to PLM estimation.

No matter which regression method is used for the nonparametric part, the forms of the estimators of β may always be written as

$$\{X^\top(I - W)X\}^{-1}\{X^\top(I - W)Y\},$$

where W is a projection operation. The estimators are asymptotically normal under appropriate assumptions.

The next section will be concerned with several nonparametric fit methods for $g(t)$ because of their popularity, beauty and importance in nonparametric statistics. In Section 5.4, the Framingham heart study data are investigated for illustrating the theory and the proposed statistical techniques.

5.2 Estimation and Nonparametric Fits

As stated in the previous section, different ways to approximate the non-parametric part yield the corresponding estimators of β. The popular non-parametric methods includes kernel regression, local polynomial, piecewise polynomial and smoothing spline. Related works are referred to Wand and Jones (1995), Eubank (1988), and Fan and Gijbels (1996). Härdle (1990) gives an extensive discussion of various nonparametric statistical methods based on the kernel estimator. This section mainly mentions the estimation procedure for β when one adapts these nonparametric methods and explains how to use XploRe quantlets to calculate the estimates.

5.2.1 Kernel Regression

Let $K(\cdot)$ be a kernel function satisfying certain conditions and h_n be a bandwidth parameter. The weight function is defined as:

$$\omega_{ni}(t) = K\Big(\frac{t - T_i}{h_n}\Big)\Big/ \sum_{j=1}^{n} K\Big(\frac{t - T_j}{h_n}\Big).$$

Let $g_n(t, \beta) = \sum_{i=1}^{n} \omega_{ni}(t)(Y_i - X_i^\top \beta)$ for a given β. Substitute $g_n(T_i, \beta)$ into (5.1) and use least square criterion. Then the least squares estimator of β is obtained as

$$\widehat{\beta}_{\mathrm{KR}} = (\widetilde{\mathbf{X}}^\top \widetilde{\mathbf{X}})^{-1} \widetilde{\mathbf{X}}^\top \widetilde{\mathbf{Y}},$$

where $\widetilde{\mathbf{X}}^\top = (\widetilde{X}_1, \ldots, \widetilde{X}_n)$ with $\widetilde{X}_j = X_j - \sum_{i=1}^{n} \omega_{ni}(T_j)X_i$ and $\widetilde{\mathbf{Y}}^\top = (\widetilde{Y}_1, \ldots, \widetilde{Y}_n)$ with $\widetilde{Y}_j = Y_j - \sum_{i=1}^{n} \omega_{ni}(T_j)Y_i$. The nonparametric part $g(t)$ is estimated by:

$$\widehat{g}_n(t) = \sum_{i=1}^{n} \omega_{ni}(t)(Y_i - X_i^\top \widehat{\beta}_{\mathrm{KR}}).$$

When $\varepsilon_1, \ldots, \varepsilon_n$ are identically distributed, their common variance σ^2 may be estimated by $\widehat{\sigma}_n^2 = (\widetilde{\mathbf{Y}} - \widetilde{\mathbf{X}}\widehat{\beta}_{\mathrm{KR}})^\top (\widetilde{\mathbf{Y}} - \widetilde{\mathbf{X}}\widehat{\beta}_{\mathrm{KR}})$.

For detailed discussion on asymptotic theories of these estimators we refer to Härdle, Liang and Gao (2000) and Speckman (1988). A main result on the estimator $\widehat{\beta}_{\mathrm{KR}}$ is:

THEOREM 5.1 *Suppose (i)* $\sup_{0 \leq t \leq 1} E(\|X\|^3|t) < \infty$ *and* $\Sigma = Cov\{X - E(X|T)\}$ *is a positive definite matrix. (ii)* $g(t)$ *and* $E(x_{ij}|t)$ *are Lipschitz continuous; and (iii) the bandwidth* $h \approx \lambda n^{-1/5}$ *for some* $0 < \lambda < \infty$*. Then*

$$\sqrt{n}(\widehat{\beta}_{\mathrm{KR}} - \beta) \overset{\mathcal{L}}{\longrightarrow} N(0, \sigma^2 \Sigma^{-1}).$$

In XploRe the quantlet `plmk` calculates the estimates $\widehat{\beta}_{\mathrm{KR}}$, $\widehat{\sigma}_n^2$ and $\widehat{g}_n(t)$. Its syntax is the following:

```
plmest=plmk(x,t,y,h)
```

 `plmk.xpl`
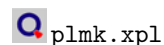
Input parameters:
`x`: the linear regressors
`t`: represents the non-linear regressors
`y`: the response
`h`: determines the bandwidth

Output parameters:
`plmest.hbeat`: estimate the parameter of X
`plmest.hsigma`: estimate the variance of the error
`plmest.hg`: estimate the nonparametric part

5.2.2   Local Polynomial

The kernel regression (or local constant) can be improved by using local linear, more generally, local polynomial smoothers since they have appealing asymptotic bias and variance terms that are not adversely affected at the boundary, Fan and Gijbels (1996).

Suppose that the $(p+1)$-th derivative of $g(t)$ at the point t_0 exists. We then approximate the unknown regression function $g(t)$ locally by a polynomial of order p. A Taylor expansion gives, for t in a neighborhood of t_0,

Figure 5.1: The simulation results for nonparametric function via quantlet `plmk`. Real data (thin) and the fitting (thick)

$\blacksquare$ `XCSplm01.xpl`

$$
\begin{aligned}
g(t) &\approx g(t_0) + g'(t_0)(t - t_0) + \frac{g^{(2)}(t_0)}{2!}(t - t_0)^2 + \cdots + \frac{g^{(p)}(t_0)}{p!}(t - t_0)^p \\
&\stackrel{\text{def}}{=} \sum_{j=0}^{p} \alpha_j (t - t_0)^j.
\end{aligned} \tag{5.5}
$$

To estimate β and $g(t)$, we first estimate α_j as the functions of β, denoted as $\alpha_j(\beta)$, by minimizing

$$
\sum_{i=1}^{n} \left\{ Y_i - X_i^\top \beta - \sum_{j=0}^{p} \alpha_j (T_i - t_0)^j \right\}^2 K_h(T_i - t_0), \tag{5.6}
$$

where h is a bandwidth controlling the size of the local neighborhood, and $K_h(\cdot) = K(\cdot/h)/h$ with K a kernel function. Minimize

$$\sum_{i=1}^{n} \left\{ Y_i - X_i^{\top}\beta - \sum_{j=0}^{p} \alpha_j(\beta)(T_i - t_0)^j \right\}^2. \tag{5.7}$$

Denote the solution of (5.7) by β_n. Let $\alpha_j(\beta_n)$ be the estimate of α_j, and denote by $\widehat{\alpha}_{jn}$ $j = 0, \ldots, p$. It is clear from the Taylor expansion in (5.5) that $\nu!\widehat{\alpha}_{jn}$ is an estimator of $g^{(j)}(t_0)$ for $j = 0, \ldots, p$. To estimate the entire function $g^{(j)}(\cdot)$ we solve the above weighted least squares problem for all points t_0 in the domain of interest.

It is more convenient to work with matrix notation. Denote by $\mathbf{Z}$ the design matrix of T in problem (5.6). That is,

$$\mathbf{Z} = \begin{pmatrix} 1 & (T_1 - t_0) & \ldots & (T_1 - t_0)^p \\ \vdots & \vdots & \vdots & \vdots \\ 1 & (T_n - t_0) & \ldots & (T_n - t_0)^p \end{pmatrix}.$$

Set $\mathbf{Y} = (Y_1, \cdots, Y_n)^{\top}$ and $\alpha(\beta) = (\alpha_0(\beta), \cdots, \alpha_p(\beta))^{\top}$. Let $\mathbf{W}$ be the $n \times n$ diagonal matrix of weights: $\mathbf{W} = \mathrm{diag}\{K_h(T_i - t_0)\}$. The weighted least squares problems (5.6) and (5.7) can be rewritten as

$$\min_{\beta} (\mathbf{Y} - \mathbf{X}\beta - \mathbf{Z}\alpha)^{\top}\mathbf{W}(\mathbf{Y} - \mathbf{X}\beta - \mathbf{Z}\alpha),$$

$$\min_{\alpha} \{\mathbf{Y} - \mathbf{X}\beta - \mathbf{Z}\alpha(\beta)\}^{\top}\{\mathbf{Y} - \mathbf{X}\beta - \mathbf{Z}\alpha(\beta)\},$$

with $\alpha(\beta) = (\alpha_0(\beta), \ldots, \alpha_p(\beta))^{\top}$. The solution vectors are provided by weighted least squares and are given by

$$\widehat{\beta}_{\mathrm{LP}} = [\mathbf{X}^{\top}\{\mathbf{I} - \mathbf{Z}(\mathbf{Z}^{\top}\mathbf{W}\mathbf{Z})^{-1}\mathbf{Z}^{\top}\mathbf{W}\}\mathbf{X}]^{-1}\mathbf{X}^{\top}\{\mathbf{I} - \mathbf{Z}(\mathbf{Z}^{\top}\mathbf{W}\mathbf{Z})^{-1}\mathbf{Z}^{\top}\mathbf{W}\}\mathbf{Y}$$

$$\widehat{\alpha} = (\mathbf{Z}^{\top}\mathbf{W}\mathbf{Z})^{-1}\mathbf{Z}^{\top}\mathbf{W}(\mathbf{Y} - \mathbf{X}\widehat{\beta}_{\mathrm{LP}})$$

Theoretically the asymptotic normality is still valid under the conditions similarly to those of Theorem 5.1. More detailed theoretical discussions are referred to Hamilton and Truong (1997).

The quantlet `plmp` is assigned to handle the calculation of $\widehat{\beta}_{\mathrm{LP}}$ and $\widehat{\alpha}$. Its syntax is similar to that of the quantlet `plmk`:

```
plmest=plmp(x,t,y,h,{p})
```

where x, t, y, h are the same as in the quantlet `plmk`. p is the local polynomial order. The default value is $p = 1$, meaning the local linear estimator.

As a consequence, the estimate of the parameter equals

$$(1.2019, 1.2986, 1.3968)$$

and the estimates of the nonparametric function is shown in Figure 5.2. There exist obvious differences between these results from the quantlet `plmk` and `plmp`. More specifically, the results for parametric and nonparametric estimation from the quantlet `plmp` are preferable to those from the quantlet `plmk`.
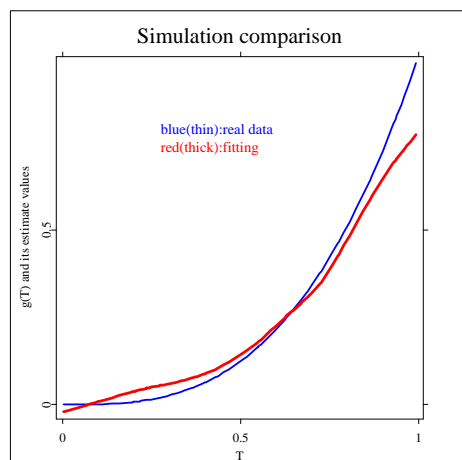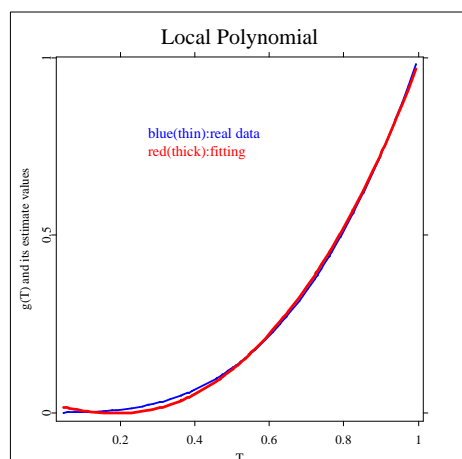

Figure 5.2: The simulation results for nonparametric function via quantlet `plmp`. Real data (thin) and the fitting (thick).

XCSplm02.xpl

5.2.3   Piecewise Polynomial

We assume g is Hölder continuous smooth of order $p = (m + r)$, that is, let r and m denote nonnegative real constants $0 < r \le 1$, m is nonnegative integer such that

$$|g^{(m)}(t') - g^{(m)}(t)| < M|t' - t|^r, \text{ for } t, t' \in [0, 1].$$

Piecewise polynomial approximation for the function $g(\cdot)$ on $[0, 1]$ is defined as follows. Given a positive M_n, divide $[0, 1]$ in M_n intervals with equal length $1/M_n$. The estimator has the form of a piecewise polynomial of degree m based on the M_n intervals, where the $(m + 1)M_n$ coefficients are chosen by the method of least squares on the basis of the data. The basic principle is concisely stated as follows.

Let $I_{n\nu}(t)$ be the indicator function of the ν-th interval, and d_ν be the midpoint of the ν-th interval, so that $I_{n\nu}(t) = 1$ or 0 according to $t \in [(\nu - 1)/M_n, \nu/M_n)$ for $\nu = 1, \ldots, M_n$ and $[1 - 1/M_n, 1]$ or not. $P_{n\nu}(t)$ be the m-order Taylor expansion of $g(t)$ at the point d_ν. Denote

$$P_{n\nu}(t) = \sum_{j=0}^{m} a_{ju} t^j \text{ for } t \text{ in the } \nu\text{-th interval.}$$

Consider the piecewise polynomial approximation of g of degree m given by

$$g_n^*(t) = \sum_{\nu=1}^{M_n} I_\nu(t) P_{n\nu}(t).$$

Suppose we have n observed data $(X_1, T_1, Y_1), \ldots, (X_n, T_n, Y_n)$. Denote

$$\mathbf{Z} = \begin{pmatrix} I_{n1}(T_1) & \ldots & I_{n1}(T_1)T_1^m & \ldots & I_{nM_n}(T_1) & \ldots & I_{nM_n}(T_1)T_1^m \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ I_{n1}(T_n) & \ldots & I_{n1}(T_n)T_n^m & \ldots & I_{nM_n}(T_n) & \ldots & I_{nM_n}(T_n)T_n^m \end{pmatrix}$$

and

$$\eta_{\mathbf{g}} = (a_{01}, \ldots, a_{m1}, a_{02}, \ldots, a_{m2}, \ldots, a_{0M_n}, \ldots, a_{mM_n})^\top.$$

Then
$$\begin{pmatrix} g_n^*(T_1) \\ \vdots \\ g_n^*(T_n) \end{pmatrix} = \begin{pmatrix} \sum_{u=1}^{M_n} I_{nu}(T_1) P_{n\nu}(T_1) \\ \vdots \\ \sum_{u=1}^{M_n} I_{nu}(T_n) P_{n\nu}(T_n) \end{pmatrix} = \mathbf{Z}\eta_{\mathbf{g}}.$$

Hence we need to find β and η_g to minimize

$$(\mathbf{Y} - \mathbf{X}\beta - \mathbf{Z}\eta_\mathbf{g})^\top (\mathbf{Y} - \mathbf{X}\beta - \mathbf{Z}\eta_\mathbf{g}).$$

Suppose that the solution of minimization problem exists. The estimators of β and η_g are

$$\widehat{\beta}_{\mathrm{PP}} = \{\mathbf{X}^\top(\mathbf{I} - \mathbf{P})\mathbf{X}\}^{-1}\mathbf{X}^\top(\mathbf{I} - \mathbf{P})\mathbf{Y}$$

and $\eta_{ng} = \mathbf{A}(\mathbf{Y} - \mathbf{X}\widehat{\beta}_{\mathrm{PP}})$, where $\mathbf{A} = (\mathbf{Z}^\top\mathbf{Z})^{-1}\mathbf{Z}^\top$ and $\mathbf{P} = \mathbf{Z}\mathbf{A}$. The estimate of $g(t)$ may be described

$$g_n(t) = z(\mathbf{Z}^\top\mathbf{Z})^{-1}\mathbf{Z}^\top(\mathbf{Y} - \mathbf{X}\widehat{\beta}_{\mathrm{PP}})$$

for a suitable z.

THEOREM 5.2 *There exist positive definite matrices Σ_{00} and Σ_{01} such that both $Cov(X|t) - \Sigma_{00}$ and $\Sigma_{01} - Cov(X|t)$ are nonnegative definite for all $t \in [0,1]$. Suppose that $\lim_{n\to\infty} n^{-\lambda}M_n = 0$ for some $\lambda \in (0,1)$ and $\lim_{n\to\infty} \sqrt{n}M_n^{-p} = 0$. Then $\sqrt{n}(\widehat{\beta}_{\mathrm{PP}} - \beta) \xrightarrow{\mathcal{L}} N(0, \sigma^2\Sigma^{-1})$.*

The quantlet `plmp` evaluates the estimates $\widehat{\beta}_{\mathrm{PP}}$ and $g_n(t)$ stated above. Its syntax is similar to those of the two previous quantlets:

```
plmest=plmp(x,t,y,m,mn)
```

where `m` and `mn` represent m and M_n, respectively. We now use the quantlet `plmp` to investigate the example considered in the quantlet `plmk`. We assume $m = 2$ and $M_n = 5$ and compute the related estimates via the quantlet `plmp`. The implementation works as follows.

 XCSplm03.xpl

The result for parameter β is `plmest.hbeta=` $(1.2, 1.2999, 1.3988)^\top$. Alternatively the estimates for nonparametric part are also given.

5.2.4   Least Square Spline

This subsection introduces least squares splines. We only state its algorithm rather than the theory, which can be found in Eubank (1988) for an overall discussion.

Suppose that g has $m-1$ absolutely continuous derivatives and m-th derivative that is square integrable and satisfies $\int_0^1 \{g^{(m)}(t)\}^2 dt < C$ for a specified $C > 0$. Via a Taylor expansion, the PLM can be rewritten as

$$Y = X^\top \beta + \sum_{j=1}^{m} \alpha_j T^{j-1} + Rem(T) + \varepsilon,$$

where $Rem(s) = (m-1)!^{-1} \int_0^1 \{g^{(m)}(t)(t-s)_+^{m-1}\}^2 \, dt$. By using a quadrature rule, $Rem(s)$ can be approximated by a sum of the form $\sum_{j=1}^{k} d_j (t-t_j)_+^{m-1}$ for some set of coefficients $d_1, \ldots, d_k$ and points $0 < t_1, \ldots, < t_k < 1$. Take a basis $V_1(t) = 1$, $V_2(t) = t$, ..., $V_m(t) = t^{m-1}$, $V_{m+1}(t) = (t-t_1)^{m-1}$, ..., $V_{m+k}(t) = (t-t_k)^{m-1}$ and set

$$\eta = (\alpha_1, \ldots, \alpha_m, d_1, \ldots, d_k) \stackrel{\text{def}}{=} (\eta_1, \ldots, \eta_{m+k})^\top$$

.

The least squares spline estimator is to minimize

$$\arg \min_{\beta, \eta} \frac{1}{n} \sum_{i=1}^{n} \left\{ Y_i - X_i^\top \beta - \sum_{j=1}^{m+k} \eta_j V_j(T_i) \right\}^2 .$$

Conveniently with matrix notation, denote $\mathbf{Z} = (Z_{ij})$ with $Z_{ij} = \{V_j(T_i)\}$ for $i = 1, \ldots, n$ and $j = 1, \ldots, m+k$ and $\mathbf{X} = (X_1, \ldots, X_n)^\top$. The least squares spline estimator is equivalent to the solution of the minimizing problem

$$(\mathbf{Y} - \mathbf{X}\beta - \mathbf{Z}\eta)^\top (\mathbf{Y} - \mathbf{X}\beta - \mathbf{Z}\eta).$$

If the problem has an unique solution, its form is the same as $(\widehat{\beta}_{\text{PP}}, \eta_{ng})$ in the subsection about piecewise polynomial. Otherwise, we may use a ridge estimator idea to modify the estimator. `plmls` is concerned with implementation of the above algorithm in XploRe.

```
plmest=plmls(x,t,y,m,knots)
```

 XCSplm04.xpl
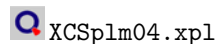
Input parameters:
x: $n \times d$ matrix of the linear design points
t: $n \times 1$ vector of the non-linear design points
y: $n \times 1$ vector of the response variables
m: the order of spline
knots: $k \times 1$ vector of knot sequence knots

Output parameters:
plmest.hbeat: $d \times 1$ vector of the estimate of the parameter
plmest.hg: the estimate of the nonparametric part

5.3   Heteroscedastic Cases

When the variance function given covariates (X, T) is non-constant, the estimators of β proposed in former section is inefficient. The strategy of overcoming this drawback is to use weighted least squares estimation. Three cases will be briefly discussed. Let $\{(Y_i, X_i, T_i), i = 1, \ldots, n\}$ denote a sequence of random samples from

$$Y_i = X_i^\top \beta + g(T_i) + \sigma_i \xi_i, i = 1, \ldots, n, \tag{5.8}$$

where X_i, T_i are the same as those in model (5.1). ξ_i are i.i.d. with mean 0 and variance 1, and σ_i^2 are some functions, whose concrete forms will be discussed later.

In general, the least squares estimator $\widehat{\beta}_{\mathrm{LS}}$ is modified to a weighted least squares estimator

$$\beta_W = \left(\sum_{i=1}^{n} \gamma_i \widetilde{X}_i \widetilde{X}_i^\top \right)^{-1} \left(\sum_{i=1}^{n} \gamma_i \widetilde{X}_i \widetilde{Y}_i \right) \tag{5.9}$$

for some weight γ_i, $i = 1, \ldots, n$. In our model (5.8) we take $\gamma_i = 1/\sigma_i^2$. In principle the weights γ_i (or σ_i^2) are unknown and must be estimated. Let

$\{\widehat{\gamma}_i, i = 1, \ldots, n\}$ be a sequence of estimators of γ. One may define an estimator of β by substituting γ_i in (5.9) by $\widehat{\gamma}_i$. Let

$$\widehat{\beta}_{\mathrm{WLS}} = \left(\sum_{i=1}^{n} \widehat{\gamma}_i \widetilde{X}_i \widetilde{X}_i^{\top} \right)^{-1} \left(\sum_{i=1}^{n} \widehat{\gamma}_i \widetilde{X}_i \widetilde{Y}_i \right)$$

be the estimator of β.

Under suitable conditions, the estimator $\widehat{\beta}_{\mathrm{WLS}}$ is asymptotically equivalent to that supposed the function σ_i^2 to be known. Therefore $\widehat{\beta}_{\mathrm{WLS}}$ is more efficient than the estimators given in the previous section. The following subsections present three variance functions and construct their estimators. Three non-parametric heteroscedastic structures will be studied. In the remainder of this section, $H(\cdot)$ is always assumed to be unknown Lipschitz continuous.

5.3.1   Variance Is a Function of Exogenous Variables

Suppose $\sigma_i^2 = H(W_i)$, where $\{W_i; i = 1, \ldots, n\}$ are design points, which are assumed to be independent of ξ_i and (X_i, T_i) and defined on $[0, 1]$ in the case where $\{W_i; i = 1, \ldots, n\}$ are random design points. Let $\widehat{\beta}_{\mathrm{LS}}$ and $\widehat{g}_n(\cdot)$ be initial estimators of β and $g(\cdot)$, for example, given by kernel regression in Section 5.2.1. Define

$$\widehat{H}_n(w) = \sum_{j=1}^{n} \widetilde{W}_{nj}(w)\{Y_j - X_j^{\top}\widehat{\beta}_{\mathrm{LS}} - \widehat{g}_n(T_i)\}^2$$

as the estimator of $H(w)$, where $\{\widetilde{W}_{nj}(t); i = 1, \ldots, n\}$ is a sequence of weight functions satisfying appropriate assumptions. Then let $\widehat{\sigma}_{ni}^2 = H_n(W_i)$.

Quantlet `plmhetexog` performs the weighted least squares estimate of the parameter. In the procedure of estimating the variance function, the estimate given by `plmk` is taken as the primary one.

Q `XCSplm05.xpl`

5.3.2 Variance Is an Unknown Function of T

Suppose that the variance σ_i^2 is a function of the design points T_i, i.e., $\sigma_i^2 = H(T_i)$, with $H(\cdot)$ an unknown Lipschitz continuous function. Similarly to subsection 5.3.1, we define the estimator of $H(\cdot)$ as

$$\widehat{H}_n(t) = \sum_{j=1}^{n} \widetilde{W}_{nj}(t)\{Y_j - X_j^\top \widehat{\beta}_{\mathrm{LS}} - \widehat{g}_n(T_i)\}^2.$$

Quantlet `plmhett` calculates the weighted least squares estimate of the parameter in this case. In the procedure of estimating the variance function, the estimate given by `plmk` is taken as the primary one.

```
plmest=plmhett(x,t,y,h,h1)
```

 XCSplm06.xpl

5.3.3 Variance Is a Function of the Mean

We consider the model (5.8) with $\sigma_i^2 = H\{X_i^\top \beta + g(T_i)\}$, which means that the variance is an unknown function of the mean response.

Since $H(\cdot)$ is assumed to be completely unknown, the standard method is to get information about $H(\cdot)$ by replication, i.e., we consider the following "improved" partially linear heteroscedastic model

$$Y_{ij} = X_i^\top \beta + g(T_i) + \sigma_i \xi_{ij}, \quad j = 1, \ldots, m_i; \quad i = 1, \ldots, n,$$

where Y_{ij} is the response of the j-th replicate at the design point (X_i, T_i), ξ_{ij} are i.i.d. with mean 0 and variance 1, β, $g(\cdot)$ and (X_i, T_i) are the same as before.

We compute the predicted value $X_i^\top \widehat{\beta}_{\mathrm{LS}} + \widehat{g}_n(T_i)$ by fitting the least squares estimator $\widehat{\beta}_{\mathrm{LS}}$ and nonparametric estimator $\widehat{g}_n(T_i)$ to the data and the residuals $Y_{ij} - \{X_i^\top \widehat{\beta}_{\mathrm{LS}} + \widehat{g}_n(T_i)\}$, and estimate σ_i^2 by

$$\widehat{\sigma}_i^2 = \frac{1}{m_i} \sum_{j=1}^{m_i} [Y_{ij} - \{X_i^\top \widehat{\beta}_{\mathrm{LS}} + \widehat{g}_n(T_i)\}]^2,$$

where each m_i is unbounded.

Quantlet `plmhetmean` executes the above algorithm in XploRe. For calculation simplicity, we use the same replicate in practice. The estimate given by `plmk` is taken as the primary one.

```
plmest=plmhetmean(mn,x,t,y,h)
```
Q XCSplm07.xpl

5.4   Real Data Examples

In this section we analyze the well known Framingham data set and illustrate the calculation results when using the quantlets introduced in Section 5.2.

EXAMPLE 5.1 *We use the data from the* Framingham Heart Study *which consists of a series of exams taken two years apart, to illustrate one of the applications of PLM in biometrics. There are* 1615 *men, aged between* 31 *to* 65, *in this data set. The outcome* Y *represents systolic blood pressure (SBP). Covariates employed in this example are patient's age (T) and the serum cholesterol level (X). Empirical study indicates that SBP linearly depends upon the serum cholesterol level but nonlinearly on age. For this reason, we apply PLM to investigate the function relationship between* Y *and* (T, X).

Specifically, we estimate β and $g(\cdot)$ in the model

$$Y_i = X_i\beta + g(T_i) + \varepsilon_i, \quad i = 1, \cdots, 1615.$$

For nonparametric fitting, we use a Nadaraya-Watson weight function with quartic kernel

$$(15/16)(1 - u^2)^2 \boldsymbol{I}(|u| \leq 1)$$

and choose the bandwidth using cross-validation.

The estimated value of the linear parameter equals to 10.617, and the estimate of $g(T)$ is given in Figure 5.3. The figure shows that with the age increasing, SBP increases but looks like a straight line. The older the age, the higher the SBP is.
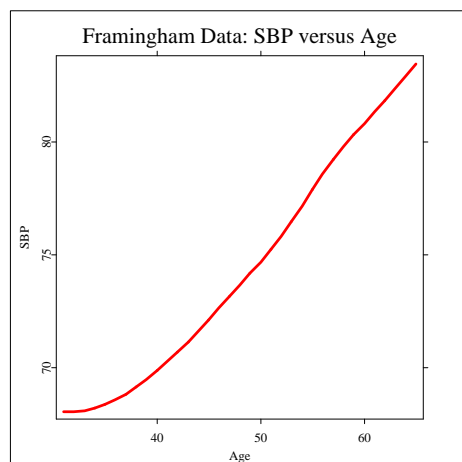
Figure 5.3: Relationship SBP and serum cholesterol level in Framingham Heart Study.

Q XCSplm08.xpl

EXAMPLE 5.2 *This is an example of using PLM to analyze the* NHANES Cancer data. *This data set is a cohort study originally consisting of* 8596 *women, who were interviewed about their nutrition habits and when later examined for evidence of cancer. We restrict attention to a sub-cohort of* 3145 *women aged* $25 - 50$ *who have no missing data the variables of interest. The outcome* Y *is saturated fat, while the predictors include age, body mass index (BMI), protein and vitamin A and B intaken. Again it is believable that* Y *depends as in* (5.2) *nonlinearly on age but linear upon other dummy variables.*

In this example we give an illustration of the `plmls` for the real data. We select $m = 3$ and the knots at $(35, 46)$. As a consequence, the estimates of linear parameters are $(-0.162, 0.317, -0.00002, -0.0047)$, and the nonparametric estimated are shown in Figure 5.4. The curve of the nonparametric part in this data set is completely different from that of the above example and looks like arch-shape. The pattern reaches to maximum point at about age 35.

We also run other quantlets for these two data sets. We found that the estimates of nonparametric parts from different quantlets have similar shapes,
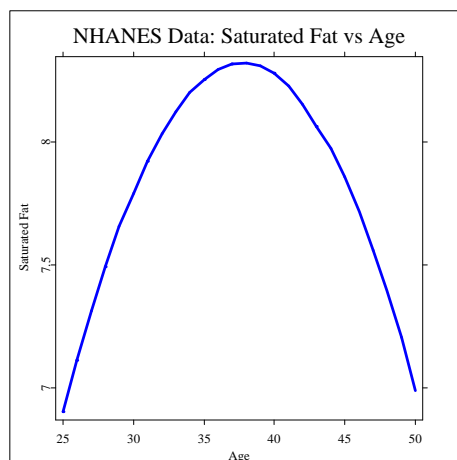
Figure 5.4: NHANES regression of saturated fat on age.

Q XCSnhanes.xpl

although differences in the magnitude of the estimates from different estimation methods are visible.

Bibliography

de Boor, C. (1978). *A Practical Guide to Splines*. New York:Springer-Verlag.

Chen, H.(1988). Convergence rates for parametric components in a partly linear model. *Annals of Statistics*, **16**, 136-146.

Engle, R. F., Granger, C. W. J., Rice, J. and Weiss, A. (1986). Semiparametric estimates of the relation between weather and electricity sales. *J. Amer. Stat. Assoc.*, **81**, 310-320.

Eubank, R. L. (1988). *Spline Smoothing and Nonparametric Regression*. New York: Marcel Dekker.

Fan, J. and Gijbels, I. (1996). *Local Polynomial Modelling and Its Applications*. Vol. 66 of *Monographs on Statistics and Applied Probability*, Chapman and Hall, New York.

Green, P., Jennison, C. and Seheult, A. (1985). Analysis of field experiments by least squares smoothing. *Journal of the Royal Statistical Society, Series B*, **47**, 299-315.

Härdle, W. (1990). *Applied Nonparametric Regression.* Cambridge University Press, New York.

Härdle, W., Liang, H. and Gao, J. T. (2000). *Partially Linear Models.* Springer-Physica-Verlag, Heidelberg.

Hastie, T. J. and Tibshirani, R. J. (1990). *Generalized Additive Models.* Vol. 43 of *Monographs on Statistics and Applied Probability*, Chapman and Hall, London.

Hamilton, S. A. and Truong, Y. K. (1997). Local linear estimation in partly linear models. *Journal of Multivariate Analysis*, **60**, 1-19.

Heckman, N.E. (1986). Spline smoothing in partly linear models. *Journal of the Royal Statistical Society, Series B*, **48**, 244-248.

Rice, J.(1986). Convergence rates for partially splined models. *Statistics & Probability Letters*, **4**, 203-208.

Robinson, P.M.(1988). Root-*n*-consistent semiparametric regression. *Econometrica*, **56**, 931-954.

Speckman, P. (1988). Kernel smoothing in partial linear models. *Journal of the Royal Statistical Society, Series B*, **50**, 413-436.

Stone, J. C. (1985). Additive regression and other nonparametric models. *Annals of Statistics*, **13**, 689-705.

Wand, M.P. and Jones, M. C.(1995). *Kernel Smoothing.* Vol. 60 of *Monographs on Statistics and Applied Probability*, Chapman and Hall, London.

Color Harmonization in Car Manufacturing Processes

A. Andriyashin, M. Benko, W. Härdle, R. Timofeev and U. Ziegenhagen

Humboldt-Universität zu Berlin, School for Business and Economics
Institute for Statistics and Econometrics, Unter den Linden 6
10099 Berlin, Germany

SUMMARY

One of the major cost factors in car manufacturing is the painting of body and other parts such as wing or bonnet. Surprisingly, the painting may be even more expensive than the body itself. From this point of view it is clear that car manufacturers need to observe the painting process carefully to avoid any deviations from the desired result. Especially for metallic colors where the shining is based on microscopic aluminium particles, customers tend to be very sensitive towards a difference in the light reflection of different parts of the car.

The following study, carried out in close cooperation with a partner from car industry, combines classical tests and nonparametric smoothing techniques to detect trends in the process of car painting. The localized versions motivated by t-test, Mann-Kendall, Cox-Stuart and a change point test are employed in this study. Suitable parameter settings and the properties of the proposed tests are studied by simulations based on resampling methods borrowed from nonparametric smoothing.

The aim of the analysis is to find a reliable technical solution which avoids any interaction from a human side. Copyright © 2000 John Wiley & Sons, Ltd.

KEY WORDS: Smoothing, Resampling, Nonparametric Regression, Trend detection

1. Introduction

Car painting is a complex combination of different layers of base coat, color and protective finishing coat. The setup for the painting process requires the optimal adjustment of a variety of different parameters such as humidity, temperature and the consistence of the lacquer itself. Even small changes to one of these parameter may influence the adhesion of the particles and their reflection behavior.

This applied paper arose from close cooperation with a partner from industry. Together we have set up a series of tests that combine approaches from local, nonparametric smoothing and classical statistical process control with the main focus to detect trends in the underlying production process. The aim is not do to detect when the process leaves a certain confidence band, the aim is to use statistical methods to detect changes before the process is going out of control.

The available dataset contains 7100 cars painted in one of 20 colors that had been collected during a three-week production run in August 2004 by specially trained personnel who used specialized devices that measured the color in terms of L, a and b values.

These values describe specific colors in the so called L*a*b color space [8], standardized by the *Commission Internationale de l'Eclairage* in 1976. While there are different schemes for the representation of colors such as RGB, CMYK and HSB, L*a*b has the advantage to include all colors visible to the human eye, moreover RGB and CMYK color spaces are subspaces of L*a*b.

In L*a*b color space each specific color is represented by a three-dimensional vector of luminance (the gray level), a and b values for the red-green and blue-yellow components of the color respectively. Since each lacquer has different properties concerning the level of refraction data have been collected at five different angles for both wing and bonnet. At the angles 15° and 110° the dataset showed huge variations. As we assume that the handling of the measuring device was difficult at these angles they have not been considered during the analysis.

δ	L*	a*	b*
15°	98.94	1.58	4.43
25°	81.35	1.17	4.20
45°	50.98	0.73	2.58
75°	32.30	0.48	1.29
110°	26.11	0.21	0.87

Table I. L*a*b values for a 'silver' car body measured at five different angles.

Table I shows an example for measurements of 'silver' colored cars. It is worth mentioning that non-metallic colors show less differences in the values across the measurement angles.

Clearly each color needs to be analyzed separately, for simplicity we focus on a 'silver' color in the further analysis. This laquer accounts for approximately 60% of all cars and is especially difficult to handle in the sense that even smallest differences in the refraction behavior may be detected by customers. Furthermore the analysis focuses on the L*-values, since this component plays the dominant role in gray and silver laquers.
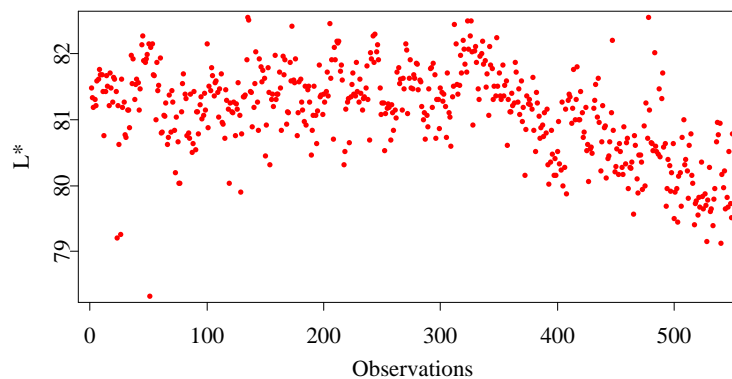
Figure 1. Data sequence of 550 L* observations, angle 25°

Figure 1 depicts the L*-values for a sequence of 550 silver colored cars, measured at a unique point on the bodies. Visually one might consider observation number 350 as the starting point for a trend. The x-axis shows the index of the observation. According to our partners there are no interdependencies between the lacquers in the production process.

The paper has the following structure. Section 2 gives an overview of proposed tests, introduces a localization framework and comments on the evaluation of test results. In section 3 the behavior of proposed tests is analyzed using simulation study. Concluding remarks are given in section 4.

2. Tests

The first test considered here is motivated by standard slope coefficient test (t-test) in the linear regression model. Clearly linearity cannot be assumed globally for the whole dataset, see figure 1). Therefore a local version of this test, an approach borrowed from smoothing methods in nonparametric regression is proposed, for details see [2]. As it will be seen in the following text, the localization of this test depends on the window size (also called bandwidth or smoothing parameter). In order to relax this dependence, a modification (called 'slope'-test) is proposed, where for each point the significance of the local slope (first derivative of the mean function) is tested using three different window sizes.

Both of these tests focus on the significance of the slope. Focusing directly on the detection of jumps (discontinuity point), local versions of Cox-Stuart, change point and Mann-Kendall test were used in addition.

The following model is assumed:

$$y_i = m(x_i) + \varepsilon_i \quad E(\varepsilon_i) = 0, \text{var}(\varepsilon_i) < \infty \tag{1}$$

with the conditional mean function $E(y_i|x_i) = m(x_i)$ and the error term ε_i. y_i denotes the dependent variable measured at the point x_i.

Strictly speaking, in the car painting application the data were measured and collected car-by-car without considering the exact time point of measurement.
Hence the following notation would be sufficient:

$$y_i = m_i + \varepsilon_i \quad E(\varepsilon_i) = 0, \text{var}(\varepsilon_i) < \infty \tag{2}$$

with the mean function $E(y_i) = m_i$ and the error term ε_i. y_i denotes the observed variable (L* value). However notation in the form of (1) is more convenient for introducing the localization framework. Note that the equation (2) can be represented in the form of (1) simply by setting $x_i = i$.

2.1. t-test

The basic idea behind the t-test is to test whether the slope β_1 of the linear model:

$$m(x) = \beta_0 + \beta_1 x$$

is significantly different from zero.

4   ANDRIYASHIN, BENKO, HÄRDLE, TIMOFEEV, ZIEGENHAGEN

Estimators of the linear regression line are obtained by minimizing the sum of squares:

$$(\widehat{\beta_0}, \widehat{\beta_1}) = \arg \min_{\beta_0, \beta_1} \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i)^2. \tag{3}$$

Under the null hypothesis $\beta_1 = 0$ the test statistics follows the t-distribution with $n-2$ degrees of freedom.

Strictly speaking the t-distribution of the test statistics relies on normality assumption of the dependent variables y_i, however it is known that the t-test is relatively robust towards slight violation of this assumption.

Local t-test   While formula (3) shows the global setup for t-test, local versions allowing to relax the assumptions on the conditional mean function m in (2) are employed for all tests in this study.

Referring to [1] for technical details, one can motivate the localized setup as follows. Instead of assuming the linearity of m globally, i.e. assuming $m(x) = \beta_0 + \beta_1 x$, $x \in [x_0, x_n]$, one point x is fixed and one essentially assumes that m can be approximated sufficiently well in the small neighborhood (window) around x. Alternatively this approach can be motivated by allowing the coefficients β_0, β_1 to change, depending on the point x, i.e. $\beta_{0,x}, \beta_{1,x}$.

In the car painting application the localization is performed on the set of (design) points $x_l = x_0 + l\Delta$, $l = 1, \ldots, n/L$. For each point x_l the $\beta_{0,x_l}, \beta_{1,x_l}$ are calculated using observations from window of L cases only. More formally, (3) is modified to:

$$\left(\widehat{\beta_{0,x_l}}, \widehat{\beta_{1,x_l}}\right) = \arg \min_{\beta_0, \beta_1} \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i)^2 \mathrm{I}(x_i \in L_l) \tag{4}$$

where I denotes the indicator function and $L_l = [x_0 + (l-1) \cdot \Delta, x_0 + (l-1) \cdot \Delta + L]$, $l = 1 \ldots {}^{n}/_{L}$. Please note that the car painting application is a quality control problem, therefore only the left neighborhood (history w.r.t. point x_l) can be used to signal a significant trend, see construction of the interval L_l above. In the classical local smoothing the localization window is typically centered around the point x_l.

The approach can be generalized by additional weighting of the points in the localization window, e.g. the weights can be set smaller for observations that are far away from point x_l and consequently higher for points close to x_l. This can be done by introduction of the kernel function $K(\cdot)$, see [2] for details. The estimation procedure can be rewritten as:

$$\left(\widehat{\beta_{0,x_l}}, \widehat{\beta_{1,x_l}}\right) = \arg \min_{\beta_0, \beta_1} \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i)^2 K\left(\frac{x_i - x_l}{h}\right) \frac{1}{h}. \tag{5}$$

Typically the kernel function is a symmetric (in classical regression problems) density function with compact support, Gaussian probability density function is also a popular choice. The parameter h is a so called bandwidth and can be regarded as the half of the localization window.

It is a well known fact that the bandwidth h has a crucial influence on the result. Large values of h yield a smooth but possibly highly biased estimator while small values of h lead to less biased but also less smooth results. For broader discussion of the local polynomial smoothing properties and optimal bandwidth choice refer to [1].

Note that t-test relies on the assumption that the errors (ε_i) are not correlated. To verify this condition Durbin-Watson test for three different windows with different localization windows was conducted. At the 95% confidence level the hypothesis of autocorrelation of the errors was rejected.

2.2. Slope Test

In order to reduce the influence of the bandwidth selection problem, a modification of t-test where the slope is tested at each point x_l using three different bandwidths (widths of the localization window) simultaneously is proposed. The idea of using different smoothing parameters in the testing problems, to authors best knowledge was first proposed by [7] and referred to as 'significant trace'.

For simplicity reasons this procedure is presented using the first setup of the test from equation (4), i.e. uniform kernel function (identity function). The generalization to an arbitrary kernel function is straightforward.

The slope test refines the idea of t-test by using 3 different window sizes for each point x_l with lengths $1/3 \cdot L$, $2/3 \cdot L$ and L, see figure 2. For each of these three intervals denoted by L_1, L_2 and L_3 the t-statistics is computed, and a significant trend for the window L is imputed if all the three test procedures show a significant deviation from zero simultaneously. Due to the simultaneousness of three procedures an adjustment of the significance level α in the form of Bonferroni correction is required: $\alpha_i = \alpha/3$ where α is the (overall) desired confidence level for a test (e.g. 5%) and α_i is the confidence level of each of three t-subtests.
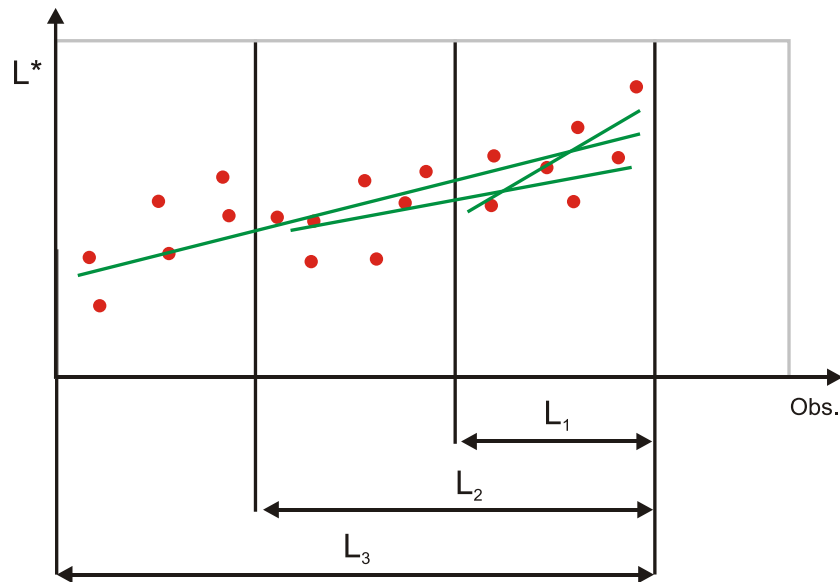
Figure 2. Basic idea of slope test

6   ANDRIYASHIN, BENKO, HÄRDLE, TIMOFEEV, ZIEGENHAGEN

2.3. Change Point Test

The following three tests (change point, Mann-Kendall and Cox-Stuart) will be introduced only in the global setup, however all of them are localized in the same way as the t-test described above.

While t-test and slope test detect trends focusing on the slope of the underlying process, these tests may be inappropriate in situations where the change in the production process is not reflected by a change of the slope but instead by a sudden jump (discontinuity point). Therefore a test that examines the means of subgroups inside the dataset is added to the portfolio of tests.

The change point test processes the data in the following way: The observations are divided into two intervals L_1 and L_2. The number of observation in L_1 is denoted by n_1, the number of observations in L_2 by n_2.

For each of the two intervals the sample mean is calculated:

$$\widehat{\mu_j} = n_j^{-1} \sum_{i=1}^{n_j} y_i \text{ for } y_i \in L_j, \ j \in \{1, 2\}.$$

The null hypothesis of equality of the true means $\mu_1 = \mu_2$ versus the alternative hypothesis $\mu_1 \neq \mu_2$ in the intervals is tested. The resulting test statistics V has an asymptotic χ^2 distribution with one degree of freedom.

$$\text{V} = \frac{n_1 \cdot n_2}{n_1 + n_2} \frac{(\widehat{\mu_1} - \widehat{\mu_2})^2}{\tilde{\sigma}^2} \xrightarrow{\mathcal{L}} \chi^2(1)$$

The unknown standard error of the residuals $\tilde{\sigma}$ is estimated by the sample variance of the residuals from nonparametric regression (see [5]). The residual $r_{i,h}$ equals the difference of the observed value y_i and the estimated value $\widehat{m_h}(x_i)$.

$$r_{i,h} = y_i - \widehat{m_h}(x_i) \tag{6}$$

with $\widehat{m_h}(x_i)$ denoting the Nadaraya-Watson estimator:

$$\widehat{m_h}(x_i) = \frac{\sum_j K\left(\frac{x_j - x_i}{h}\right) y_j}{\sum_j K\left(\frac{x_j - x_i}{h}\right)}$$

with kernel function K. The bandwidth h is computed using cross-validation, for details refer to [5].

2.4. Mann-Kendall Test

This classical approach for trend detection compares the differences of subsequent observation and makes a statement on the trend using the proportion of positive differences among all computed differences. The test statistics is given by:

$$S = \sum_{i=2}^{n} \sum_{j=1}^{i-1} \text{sign}(y_i - y_j) \tag{7}$$

Tabulated critical values are given in [6], for $n \to \infty$ the distribution of the test statistics converges to the standard normal distribution.

$$S^* = \frac{S}{\sqrt{\frac{n(n-1)(2n+5)}{18}}} \xrightarrow{\mathcal{L}} N(0,1) \tag{8}$$

2.5. Cox-Stuart Test

While the Mann-Kendall test analyzes the differences of subsequent observations, the Cox-Stuart test uses two partitions of the underlying sample to detect trends. From the sequence of observations $y_1, \ldots, y_n$ the dividing observation m is calculated as following:

$$m = \begin{cases} \frac{n}{2} & \text{if } n \text{ is even} \\ \frac{n+1}{2} & \text{if } n \text{ is odd} \end{cases} \tag{9}$$

Using m one can set up the two groups and calculate the differences d_i of the observations from the two partitions.

$$d_i = (y_{i+m} - y_m) \text{ for } \begin{cases} 1, \ldots, m & \text{if } n \text{ is even} \\ 1, \ldots, m-1 & \text{if } n \text{ is odd} \end{cases} \tag{10}$$

The test statistics T is then obtained as the number of all positive d_i, where w denotes the number of y_i that are unequal to zero. The null hypothesis H_0 of 'no trend' is rejected if:

$$t < {}^1\!/_2 \left(l - z_{1-\alpha/2}\sqrt{w} \right) \text{ or } t > 1 - {}^1\!/_2 \left(l - z_{1-\alpha/2}\sqrt{w} \right) \tag{11}$$

with l being the number of non-negative d_i and $z_{1-\alpha/2}$ as the $1 - {}^\alpha\!/_2$ quantile of standard normal distribution.

2.6. Evaluation of Test Results

For the evaluation of the test results the following heuristics is proposed – a persistent trend should cause not just a single signal but a series of subsequent signals. To formalize this idea a *summation measure* is proposed. While a test generates a signal with a respective inverse p-value (i.e. $1-p$ where p is the p-value), the summation measure sums up significant (inverse) p-values sequentially. Significant means that $1 - p \geq 1 - \alpha$ or equivalently $p \leq \alpha$ where α is the confidence level.

As an example of such setup consider an example for t-test on figure 3 where a sequence of 550 observations is used. The window size L is 75 observations, the stepsize Δ of this window is set to 1, the significance level is 5%.

The lower part of figure 3 contains the observations and a slope line for each significant test, these lines are colored with red.

At this point one has to think carefully about the treatment of blanks (non-significant p-values) in sequences of p-values. What if there is a cluster of 10 significant p-values with a single insignificant one *inside* – should then the summation measure stop at every first absence of signal or be more flexible allowing some noise inside? If so, to what extent? The other question is what the minimum number of consequent significant p-values is to form a trend message, because one or two indications may be false alarms?

8   ANDRIYASHIN, BENKO, HÄRDLE, TIMOFEEV, ZIEGENHAGEN
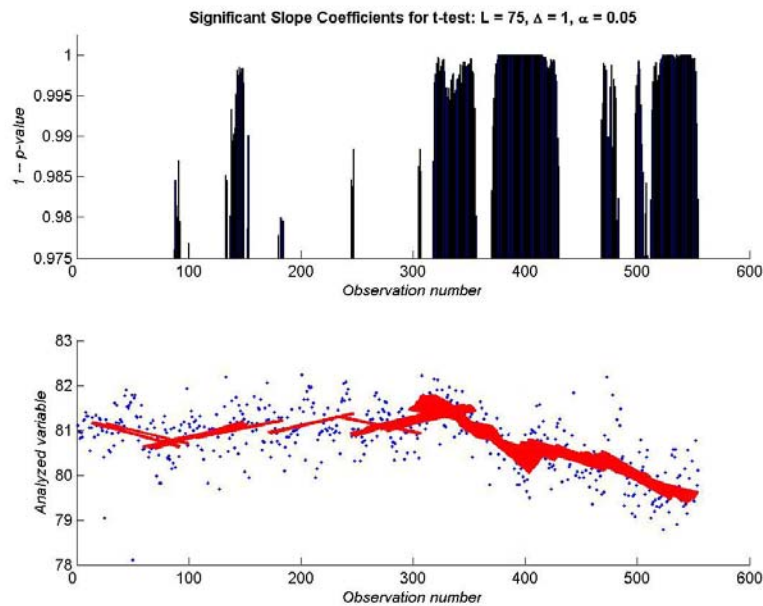
Figure 3. t-test results of the dataset

Therefore two additional parameters are introduced to address this issue. Let τ denote the minimum number of subsequent signals (significant p-values) that is eligible for creation of non-zero summation measure and κ be the maximum number of tolerated blanks (insignificant p-values).

Although the general idea of aggregation of the information from the individual test is now available, a critical value to indicate a persistent trend is clearly needed. It is worth pointing out that this value is certainly dependent on both τ and κ.

Let us refer to an example of on figure 4 where summation measure is computed with $\kappa = 3$ and $\tau = 10$.

Looking at the summed inverse p-values depicted in figure 4 (the dashed line at the upper part), one can assume the existence of a persistent trend if the height of the sum breaks a certain limit, a threshold value η. Clearly the heuristical approach of the summation method makes it difficult to determine the critical value. After discussion with industry the following rule of thumb was used for the threshold value – 50% of the maximum value of the summation measure. In practice this value is surely inappropriate since the maximum value of the sum is not known in advance, whereas it serves quite well for the estimation of internal test localizing parameters (see below).

Figure 4 shows an example of the threshold value. Summed values that exceed 50% threshold are marked using vertical red bars on figure 5.
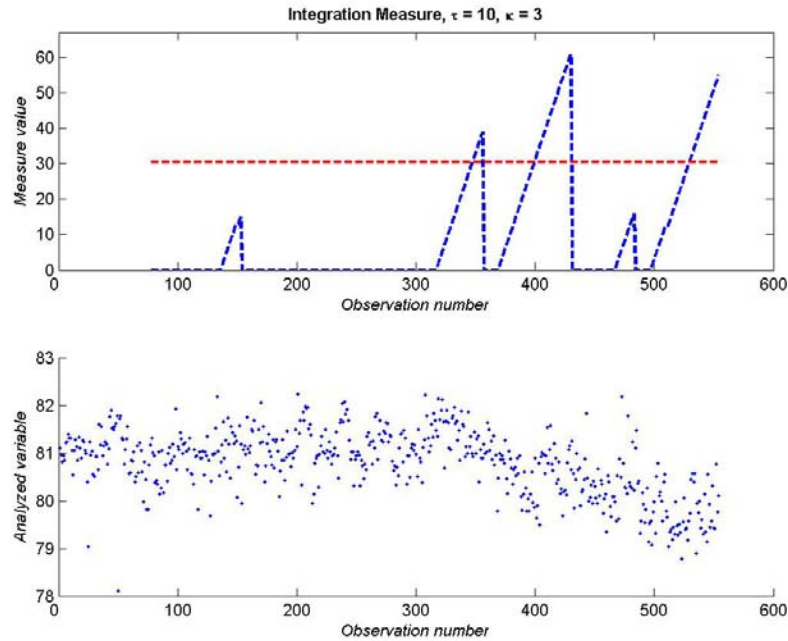
Figure 4. Summation measure for the t-test results from figure 3

3. Simulation

The main goal of this section is to perform simulations, essentially to study the detection power of the available tests. Our analysis is based on pseudo-residual resampling. In addition the simulations have also been used to determine the internal test parameters.

For the comparison of the different tests simulated datasets are used. The basic structure of each dataset consists of 550 observations, the first 300 observations have no trend (slope = 0), beginning from observation number 301 an artificial trend with slope of angle γ is added. The quality of performed test is measured by the delay - the number of observations after observation 300 until the trend has been detected.

One can argue that the residuals ε_i, see (2), can be modeled as normally i.i.d. However, it will be shown in the following paragraph that this does not coincide with the data set observed in the car painting application.

The residuals ε_i, see (6), are estimated by $r_{i,h}$, i.e. as the differences between original and smoothed data, see figure 6.

Figure 7 shows the nonparametric estimate of the density estimated from $r_{i,h}$, $i = 1, \ldots, N$ and the normal density estimated from the same sample. The nonparametric estimate is obtained by the kernel density estimator, see [2]. To be able to compare these parametric and nonparametric estimates the same quality the same level of bias needs to be achieved. Hence the normal density is additionally smoothed with the same settings for kernel and
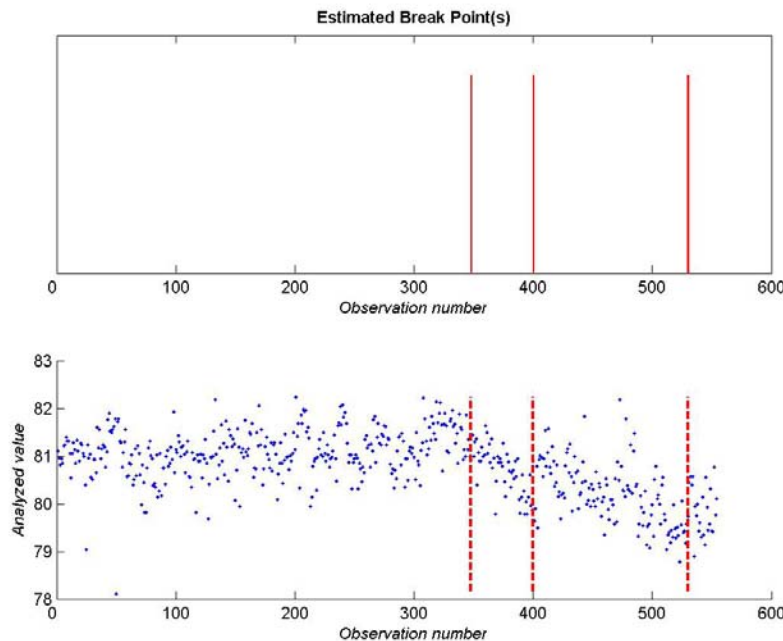
Figure 5. Generated signals for the t-test results from figure 3 using the summation measure.

bandwidth as in the nonparametric estimate (for further discussion see [4]). Figure 7 shows the clear difference between these two densities.

Using the Jarque-Bera and Kolmogorov-Smirnov tests, the residuals were examined for normality. Since both tests rejected the normality hypothesis at the 95% confidence level, it is inappropriate to generate simulated data using normally distributed residuals. Although one could use a broader class of parametric density functions to fit the residuals, it was decided to employ a resampling method to obtain datasets with the same random residual structure as the original data.

Simulated data noise was constructed by sampling with replacement from the estimated residuals $r_{i,h}$. The drawn residual was returned back to the pool so that there is a positive probability of choosing the same residual more than one time.

During the simulations three angles were employed to create artificial datasets: $\tan \gamma = 0.002$ ($0.1145°$, see figure 8), $\tan \gamma = 0.005$ ($0.2864°$) and $\tan \gamma = 0.05$ ($2.8624°$).

Table II summarizes the simulation results for internal test parameter estimation. For each set of parameters 50 simulations were performed. Let mean delay be defined as an average delay for 5 performed tests: t-test, slope test, change point test, Mann-Kendall test and Cox-Stuart test. For $\tan \gamma = 0.05$ (Tables IV) the parameters L and κ had no significant influence on the mean delay, bigger values for τ gave smaller average delays. For the two small angles $\tan \gamma = 0.002$ and $\tan \gamma = 0.005$ (Tables II and III) some of the tests gave inconsistent results, that can be traced back to the dependance of the threshold value η from κ and Δ.
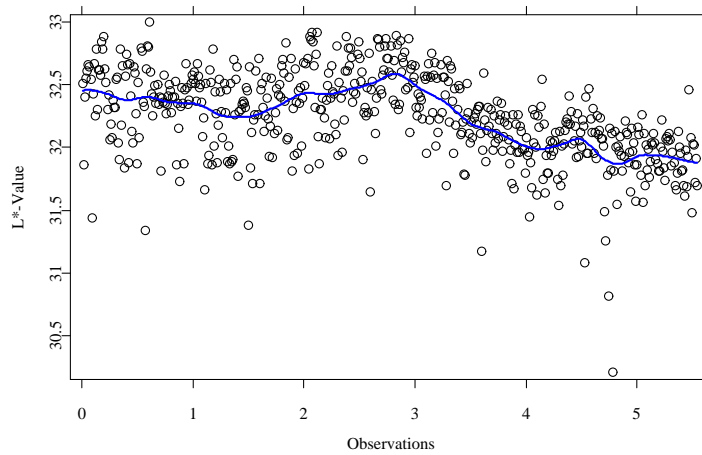
Figure 6. Original data (black points) and smoothed curve (blue solid curve).
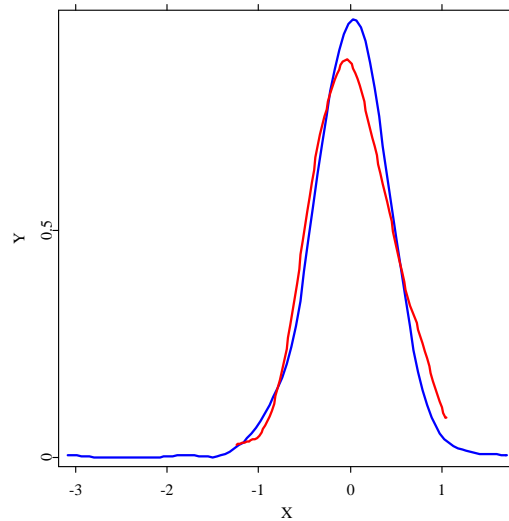
Figure 7. Kernel density estimation for the original dataset

Since for each test a big number of simulations was conducted, aggregated results of test efficiency are provided in Figure 9. Each boxplot was built independently for each test over all performed simulations across different parameter values γ, L, κ, τ.

Apart from mean delay evidence, it is of great interest to investigate the behaviour of proposed tests in respect to false alarms. As before 50 simulations for each of five tests were conducted, where the average false alarm statistics for a given set of parameters was calculated.

Recall that in our simulations first 300 observations contain no trend while the last 250
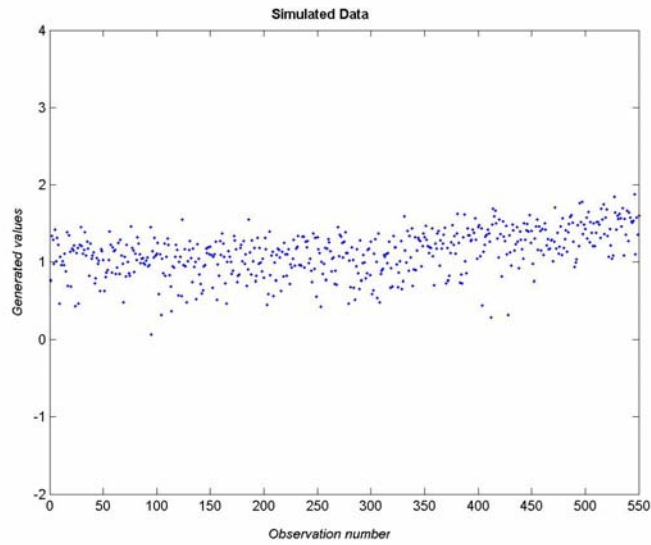
Figure 8. Simulated data with angle $\tan \gamma = 0.002$

		Mean delay (5 tests)					
	L	25		50		75	
	κ	3	5	3	5	3	5
τ							
3		110	140	120	150	185	175
5		204	128	145	150	100	125
10		Failed	Failed	Failed	Failed	120	140

Table II. Test efficiency for simulated data with $\tan \gamma = 0.002$

		Mean delay (5 tests)					
	L	25		50		75	
	κ	3	5	3	5	3	5
τ							
3		130	170	150	120	215	225
5		160	125	100	115	145	145
10		Failed	Failed	130	120	90	95

Table III. Test efficiency for simulated data with $\tan \gamma = 0.005$

observations constitute a trend with a given slope γ. Any signal before the observation number 300 generated by a procedure is acknowledged as a false alarm. Note, however, that there is no

	Mean delay (5 tests)					
L	25		50		75	
κ	3	5	3	5	3	5
τ						
3	110	140	120	150	185	175
5	204	128	145	150	100	125
10	Failed	Failed	Failed	Failed	120	140

Table IV. Test efficiency for simulated data with $\tan\gamma = 0.05$
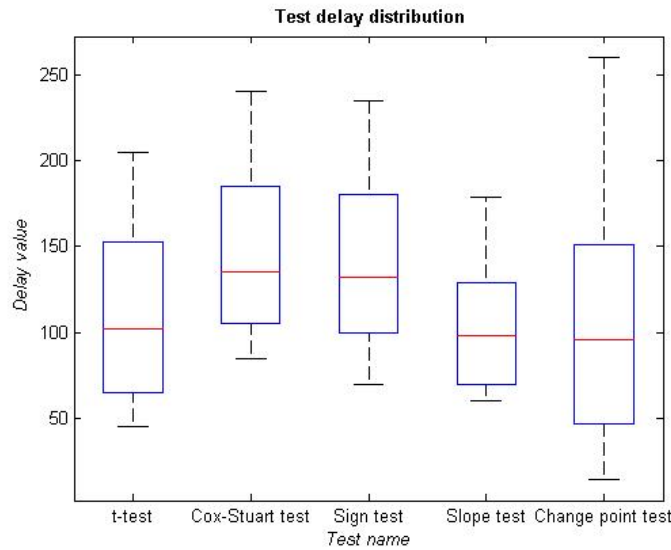
Figure 9. Boxplots for test delays of five performed tests

further need to compute this statistics for different slope angles since by definition false alarms can only occur before the beginning of the trend. In the table below, an averaged number of false alarms is provided, where averaging was performed over the number of simulations and the number of tests.

The main conclusion one can draw from this table is that proper adjusted integration measure parameters (τ and κ) can effectively avoid false alarms. In case when $\tau = 1$ the integration measure is in fact no longer employed. Enough big τ and κ make it possible to eliminate single false signals and therefore the proposed procedures work effectively in this regard.

		Mean number of false alarms					
	L	25		50		75	
	κ	3	5	3	5	3	5
τ							
1		2.3	1.9	2.1	1.7	1.1	0.5
2		1.2	0.6	1.0	0.4	0.3	0
3		0	0	0	0	0	0
5		0	0	0	0	0	0
10		0	0	0	0	0	0

Table V. False alarm statistics for simulated data

4. Conclusions

The best performance has been shown by t-test, change-point test and slope test, therefore we recommended the simultaneous use of these three tests to our partner. The detection power of Mann-Kendall and Cox-Stuart test was below the power of the first three tests so this pair was not recommended for usage in a given situation.

The t-test showed good results in most of the simulation runs, also it is fast to compute. For outliers and small window sizes L it proved to be inefficient, because it is based on the linear regression model.

The change point tests showed the best results on average. The drawback of this test is that it is slow and difficult in computation, since $\hat{\sigma}$ has to be computed from the data and it tends to generate single false signals, therefore the test parameters need to be adjusted carefully. But due to the employment of summation measure, single false signals are not transformed into false alarms. Therefore the majority of false alarms could be avoided.

All tests showed a dependence on the setup for L, κ and τ, values for $L = 75$, $\kappa = 5$ and $\tau = 3$ showed good results.

Further research has to provide suitable ranges for the threshold value, the used 50% of the maximum summation measure value was only employed as initial proposal.

Robust estimation of dimension reduction space

P. Čížek[a],[*], W. Härdle[b],[1]

[a]*Department of Econometrics & OR, Tilburg University, P.O. Box 90153, 5000 LE, Tilburg, The Netherlands*
[b]*Institut für Statistik und Ökonometrie, Humboldt-Universität zu Berlin, Spandauer Str. 1, D-10178 Berlin, Germany*

Abstract

Most dimension reduction methods based on nonparametric smoothing are highly sensitive to outliers and to data coming from heavy-tailed distributions. Two recently proposed methods, minimum average variance estimation and outer product of gradients, can be and are made robust in such a way that preserves all advantages of the original approach. Their extension based on the local one-step M-estimators is sufficiently robust to outliers and data from heavy-tailed distributions, it is relatively easy to implement, and surprisingly, it performs as well as the original methods when applied to normally distributed data.
© 2005 Elsevier B.V. All rights reserved.

Keywords: Dimension reduction; L- and M-estimation; Nonparametric regression

1. Introduction

In regression, we aim to estimate the regression function, which describes the relationship between a dependent variable $y \in \mathbb{R}$ and explanatory variables $X \in \mathbb{R}^p$. This relationship can be, without prior knowledge and with full generality, modelled nonparametrically, but an increasing number of explanatory variables makes nonparametric estimation suffer from the curse of dimensionality (Bellman, 1961). There are two main approaches to deal with a large number of explanatory variables: to assume a simpler form of the regression function (e.g., its additivity) or to reduce the dimension of the space of explanatory variables. The latter, more general approach received a lot of attention recently; see Li (1991) and Xia et al. (2002), for instance. Our first aim is to study the latter approach and examine its sensitivity to heavy-tailed and contaminated data, which can adversely influence both parametric and nonparametric estimation methods (see Čížek, 2004; Sakata and White, 1998, for evidence in financial data). Further, we propose robust and computationally feasible modifications of Xia et al. (2002)'s methods and study their behavior by means of Monte Carlo experiments.

A dimension-reduction (DR) regression model can be written as

$$y = g\left(B_0^\top X\right) + \varepsilon, \tag{1}$$

where g is an unknown smooth link function, B_0 represents a $p \times D$ orthogonal matrix, $D \leqslant p$, and $E(\varepsilon|X) = 0$ almost surely. Hence, to explain the dependent variable y, the p-dimensional explanatory variable space for X can be reduced to a D-dimensional space given by B_0, which we assume to be unique in this context (for $D = p$, the standard nonparametric regression model results). The vectors of B_0 are called directions in this context. DR methods aim to find the dimension D of the DR space and the matrix B_0 defining this space.

Recently, Xia et al. (2002) proposed the minimum average variance estimator (MAVE), which improves in several aspects over other existing estimators, such as sliced inverse regression (SIR) by Li (1991). First, MAVE does not need undersmoothing when estimating the link function g to achieve a faster rate of convergence. Second, MAVE can be applied to many models including time series data and readily extended to other related problems such as classification (Antoniadis et al., 2003) and functional data analysis (Amato et al., 2005). Finally, the MAVE approach renders generalizations of some other nonparametric methods; for example, Xia's outer product of gradients (OPG) estimator extends the average derivative estimator (ADE) of Härdle and Stoker (1989) to multi-index models.

Despite many features, MAVE does not seem to be robust to outliers in the dependent variable y since it is based on local least-squares estimation (for evidence, see Rousseeuw and Leroy, 2003, in parametric and Čížek, 2004, in nonparametric regression). This concern was expressed by many authors in a discussion of MAVE (e.g., Critchley, 2002; Cui and Li, 2002; Čížek et al., 2002), and in a broader sense of violation of distributional assumptions, is common to many nonparametric DR techniques. See for instance comments of Cook and Weisberg (1991) and Härdle and Tsybakov (1991) to the SIR method (Li, 1991) or comments of He and Fung (2001) concerning the work of Cook and Yin (2001). In the case of SIR, at least its sensitivity to outliers in the space of explanatory variables X (so-called leverage points) was remedied by Gather et al. (2001). All of their concerns about the robustness of DR methods are further amplified by the fact that analyzed data are typically highly dimensional, and as such, are difficult to check and clean. Therefore, because of the many advantages that MAVE possess, we address its low robustness to outlying observations and propose ways to improve it without affecting main strengths of MAVE. Additionally, we also employ and generalize OPG because, despite being inferior to MAVE, it provides an easy-to-implement and fast-to-compute method that could prove preferable in some applications (especially if dimension D is expected to be small).

The rest of the paper is organized as follows. In Section 2, we describe both the MAVE and OPG methods and discuss their sensitivity to outliers. Robust enhancements of these methods are proposed in Section 3. Finally, we compare all methods by means of simulations in Section 4.

2. Estimation of dimension reduction space

In this section, we first present MAVE and OPG (Sections 2.1 and 2.2) as well as a procedure for determining the effective DR dimension by means of cross-validation (CV) (Section 2.3). At the end of the section, we will motivate our concerns about robustness of these methods (Section 2.4).

2.1. The MAVE method

Let d represent the working dimension, $1 \leqslant d \leqslant p$. For a given number d of directions B in model (1), Xia et al. (2002) proposed to estimate B by minimizing the unexplained variance $E\{y - E(y|X)\}^2 = E\{y - g(B^\top X)\}^2$, where the unknown function g is locally approximated by a linear function; that is, $g(B^\top X_0) \approx a_0 + b_0^\top B^\top (X - X_0)$ around some X_0. The novel feature of MAVE is that one minimizes simultaneously with respect to directions B and coefficients a_0 and b_0 of the local linear approximation. Hence, given a sample $(X_i, y_i)_{i=1}^n$ from (X, y), MAVE estimates B by minimizing

$$\min_{\substack{B : B^\top B = I_d \\ a_j, b_j, j=1,\dots,n}} \sum_{i=1}^n \sum_{j=1}^n \left[y_i - \left\{ a_j + b_j^\top B^\top (X_i - X_j) \right\} \right]^2 w_{ij}, \tag{2}$$

where w_{ij} are weights describing the local character of linear approximation. Initially, weights at any point X_0 are given by a multidimensional kernel function K_h, where h refers to a bandwidth: $w_{i0} = K_h(X_i - X_0)\left\{\sum_{i=1}^n K_h(X_i - X_0)\right\}^{-1}$; $i = 1, \dots, n$. Additionally, once we have an estimate $\hat{B}$ of the DR space, it is possible to iterate using weights based on

distances in the reduced space: $w_{i0} = K_h \left\{ \hat{B}^\top (X_i - X_0) \right\} \left[\sum_{i=1}^n K_h \left\{ \hat{B}^\top (X_i - X_0) \right\} \right]^{-1}$. Iterating until convergence results in a refined MAVE, which is the estimator we understand in the rest of the paper under MAVE.

Xia et al. (2002) also proposed a nontrivial iterative estimation procedure based on repeating two simpler optimizations of (2): one with respect to a_j, b_j given an estimate $\hat{B}$ and another with respect to B given estimates $\hat{a}_j$, $\hat{b}_j$. This computational approach greatly simplifies and speeds up estimation.

2.2. The OPG method

Based on the MAVE approach, Xia et al. (2002) also generalized ADE of Härdle and Stoker (1989) to multi-index models. Instead of using a moment condition for the gradient of the regression function g in model (1), $E\{\nabla g(X)\} = 0$, the average OPG is used: $\Sigma = E\left\{\nabla g(X) \nabla^\top g(X)\right\}$. It can be shown that the DR matrix B consists of the d eigenvectors corresponding to the d largest eigenvalues of Σ. Thus, recalling that local linear regression solves at point X_j

$$\min_{a_j, b_j} \sum_{i=1}^n \left[y_i - \left\{ a_j + b_j^\top (X_i - X_j) \right\} \right]^2 w_{ij}, \tag{3}$$

we can estimate Σ by $\hat{\Sigma} = (1/n) \sum_{i=1}^n \hat{b}_j \hat{b}_j^\top$, where $\hat{b}_j$ are estimates of b_j from (3). Hence, OPG consists of estimating $\hat{\Sigma}$ and determining its d eigenvectors with largest eigenvalues. The choice of weights w_{ij} can be done in the same way as for MAVE.

Although OPG does not exhibit the convergence rate of MAVE, it is easy to implement, fast to compute, and can be flexibly combined with robust estimation methods as shown in Section 3. Moreover, our simulations show that it can perform as well as MAVE if just one or two directions, $d \leqslant 2$, are of interest.

2.3. Dimension of dimension-reduction space

The described methods can estimate the DR space for a pre-specified dimension d. To determine d, Xia et al. (2002) extend the CV approach of Yao and Tong (1994) and estimate d by $\hat{d} = \arg \min_{0 \leqslant d \leqslant p} \text{CV}(d)$, where

$$\text{CV}(d) = \sum_{j=1}^n \left[y_j - \sum_{i=1, i \neq j}^n \frac{y_i K_h \left\{ \hat{B}^\top (X_i - X_j) \right\}}{\sum_{i=1, i \neq j}^n K_h \left\{ \hat{B}^\top (X_i - X_j) \right\}} \right]^2 \tag{4}$$

for $d > 0$ and $\text{CV}(0) = (1/n) \sum_{i=1}^n (y_i - \bar{y})^2$ to incorporate the possibility of y and X being independent. Note that this CV criterion can be also used to select bandwidth h, which results in two-dimensional CV over d and h. Although we use this time-demanding strategy in our simulations, in practice it is possible to consistently select bandwidth h^* in the DR space for the largest dimension $d = p$ and employ this bandwidth for all other dimensions.

2.4. Robustness of dimension reduction

The complexity of highly dimensional data processed by means of DR methods requires estimation methodology robust to data contamination, which can arise from miscoding or heterogeneity not captured or presumed in a model. In nonparametric regression, even data coming from a heavy-tailed distribution can exhibit effects similar to data contamination. Since both MAVE, OPG, and CV are based on least-squares criteria, their sensitivity to outlying observations can be rather high. Here, we discuss possible effects of a single outlying observation on the estimation results, that is on $\hat{a}_j$, $\hat{b}_j$, and $\hat{B}$, and on the CV selection of bandwidth h and dimension d. At the end of the section, we demonstrate the effects of an outlier in a simple example.

Considering OPG and local parametric regression (3), the estimates $\hat{a}_j$ and $\hat{b}_j$ are just a linear combination of y_i values. Since the weights w_{ij} are independent of y_i for a given bandwidth h, even a single outlying value y_i, $|y_i| \to \infty$, can arbitrarily change the estimated coefficients $\hat{a}_j$ and $\hat{b}_j$ around X_j if h is sufficiently large. This effect then influences the matrix $\hat{\Sigma} = (1/n) \sum_{i=1}^n \hat{b}_j \hat{b}_j^\top$. In the case of MAVE defined by (2), the situation is more complicated since the local linear approximation of the link function given by a_j and b_j can adjust simultaneously with directions B. In general,

Table 1
Optimal CV bandwidth for dimension d and a data set with one additive outlier c

Simulated data with an additive outlier

Dimension d	2	4	6	8	10
$c = 0$	0.28	0.61	1.00	1.00	1.65
$c = 50$	0.37	0.61	0.78	1.00	2.12
$c = 100$	0.47	1.00	0.78	2.12	2.12
$c = 150$	1.28	0.78	0.78	2.12	2.72
$c = 250$	1.00	0.78	0.61	0.47	0.61
$c = 500$	0.08	0.61	0.37	0.47	0.61

Table 2
Optimal CV dimension d for a data set with one additive outlier c

Simulated data with an additive outlier

Method	$c = 0$	$c = 50$	$c = 100$	$c = 150$	$c = 250$	$c = 500$
OPG	1	2	5	1	1	1
MAVE	2	2	3	2	1	1

it is not possible to explicitly state which parameters will be affected and how. However, it is likely that the effect of an outlier will vary with the working dimension d.

In addition, nonparametric estimation depends on an auxiliary parameter, bandwidth h, and its choice—done here by CV—is crucial for the performance of a method. As argued in Ronchetti et al. (1997), an outlier can significantly bias results of the least-squares-based CV. For OPG (3), bandwidth h is chosen generally too small: the CV criterion (4) is minimized when the outlier affects as small a number of observations in its neighborhood as possible. This occurs when bandwidth h is small. For MAVE (2), the situation is again complicated by the fact that the outlier can be "isolated" not only by a small bandwidth, but possibly by a specific choice of directions B as well. Furthermore, since CV is also used to determine the dimension of the DR space, an outlier can adversely affect the estimation of dimension D as well.

To exemplify the influence of a single outlying observation on the bandwidth and dimension selection, we generated a random sample of 100 observations from the following nonlinear model:

$$ y_i = \left(X_i^\top b_1 \right)^2 - \left(0.5 + X_i^\top b_2 \right)^2 + 15 \cos \left(X_i^\top b_3 \right) + 0.5\varepsilon_i, $$

where the random variables X_i and error term ε_i come from the standard normal distribution. The choice of directions, $\beta_1 = (1, 2, 3, 0, 0, 0, 0, 0, 0, 0)/\sqrt{14}$, $\beta_2 = (-2, 1, 0, 1, 0, 0, 0, 0, 0, 0)/\sqrt{6}$, and $\beta_3 = (0, 0, 0, 0, 0, 0, 0, 1, 1, 1)/\sqrt{3}$, implies that the data space has dimension $p = 10$ and the effective DR has dimension $D = 3$. Additionally, we included one observation that has value y_i increased by c ranging from 0 to 500. We first applied OPG since it allows to determine the local linear approximation (3) and the corresponding bandwidth h separately from directions B. For various values of outlier c and dimension d, the optimal bandwidth h_{opt} was chosen by CV, see Table 1 (line $c = 0$ corresponds to data without the outlier). Although it does not change monotonically with c, there is a general trend of h_{opt} being smaller with increasing outlier value c, especially for larger dimensions d. Using the cross-validated bandwidths, we also estimated dimension d as a function of outlier value c. The results of CV based on MAVE and OPG estimates are summarized in Table 2. OPG seems rather sensitive to this outlier because the estimated dimension varies between 1 and 5 for $c = 0, \ldots, 500$. MAVE results in more stable estimates, which are however still influenced by the outlier position.

3. Robust dimension reduction

Both MAVE and OPG seem to be sensitive to data contamination. Our aim is thus to propose robust enhancements of MAVE and OPG that should preserve their present qualities, increase their robustness, and be computationally feasible. We discuss first general ways of making MAVE and OPG more robust (Section 3.1). Next, we address computational issues and propose robust MAVE and OPG that are computationally feasible (Section 3.2). Finally, we adapt the CV procedure mentioned in Section 2.3 for robust estimation (Section 3.3).

Within this section, we concentrate on protecting the MAVE and OPG estimates against outlying observations, but not against leverage points, which have only limited influence on the estimation. This can be seen, for example, from the MAVE definition (2): the DR matrix B plays a role only for observations, where the slope b_j of the local linear approximation is nonzero. In the case of a single outlier separated from other observations, this is always true. In the case of a cluster of outliers, that is a group of outlying observations closer to each other than the bandwidth h in use, the slope b_j will be also relatively close to zero if we suppose outliers are just a random noise. One can of course argue that a cluster of outliers can have such values of the dependent variable y that they have a particular structure and significantly nonzero slope b_j (i.e., there is a regression relationship). Then it is however rather likely that they are a valid group of distant observations and they should not be excluded from estimation even though this may lead to a larger DR space.

3.1. Robust MAVE and OPG

There are many robust alternatives to least squares in linear regression. Using them in nonparametric regression adds a requirement on an easy and fast implementation, which excludes many so-called high-breakdown point methods (Rousseeuw and Leroy, 2003), and on the other hand, eliminates need for robustness against leverage points to some extent. During the last few decades, local L- and M-estimators have become particularly popular and well studied; see Boente and Fraiman (1994), Čížek (2004), and Fan and Jiang (1999), Härdle and Tsybakov (1988), respectively.

The application of local L- or M-estimation to MAVE and OPG theoretically reduces to replacing the squared residuals in (2) and (3) by a general convex function ρ_σ of residuals, where σ represents the variance of residuals. Whereas L-estimators do not require the knowledge of σ, in the case of M-estimators, robust choices of ρ_σ depend on an estimate $\hat{\sigma}$ of residual variance. In parametric estimation (Hampel et al., 1986), σ is typically treated as a nuisance parameter or is set proportional to the median absolute deviation (MAD). To reflect local character of estimation given by w_{ij} in (2) and (3) and to minimize computational cost, we propose to estimate the variance of residuals by the weighted MAD with weights w_{ij}. Specifically, we estimate the residual variance at point X_0 by

$$\hat{\sigma}_0(y) = 1.4826 \, \text{wmed}_0 \left\{ |y_i - \text{wmed}_0(y)| \right\},$$

where $\text{wmed}_0(y)$ represents the weighted median at X_0,

$$\text{wmed}_0(y) = \min_{k=1,\ldots,n} \left\{ y_{(k)} \left| \sum_{i=1}^n \frac{K_h(X_i - X_0)}{\sum_{i=1}^n K_h(X_i - X_0)} \cdot I\left(y_i \leqslant y_{(k)}\right) \geqslant 0.5 \right. \right\},$$

and $y_{(k)}$ is the kth order statistics of y_i, $i = 1, \ldots, n$.

In the case of OPG, this means that one applies a local polynomial L- or M-estimator with a given function ρ_σ and a variance estimate $\hat{\sigma}$,

$$\min_{a_j, b_j} \sum_{i=1}^n \rho_{\hat{\sigma}} \left[y_i - \left\{ a_j + b_j^\top (X_i - X_j) \right\} \right] w_{ij}, \tag{5}$$

and then obtains the DR space B as d eigenvectors of $\hat{\Sigma} = (1/n) \sum_{i=1}^n \hat{b}_j \hat{b}_j^\top$ with largest eigenvalues.

In the case of MAVE, having a general objective function $\rho_{\hat{\sigma}}$ leads to

$$\min_{\substack{B:B^\top B = I_d \\ a_j, b_j, j=1,\ldots,n}} \sum_{i=1}^n \sum_{j=1}^n \rho_{\hat{\sigma}} \left[y_i - \left\{ a_j + b_j^\top B^\top (X_i - X_j) \right\} \right] w_{ij}. \tag{6}$$

Although (6) cannot be minimized using the algorithm proposed by Xia et al. (2002), the optimization can be still carried out by repeated estimation of (6) with respect to a_j, b_j given an estimate $\hat{B}$ and with respect to B given estimates $\hat{a}_j, \hat{b}_j (j = 1, \ldots, n)$. The first step is just the already mentioned local L- or M-estimation analogous to (5). To facilite the second step—estimation of B, let us observe that (6) can be reformulated as follows. For $B = (\beta_1, \ldots, \beta_d)$ and given $(a_j, b_j^\top) = (a_j, b_{j1}, \ldots, b_{jd})$, (6) is equivalent to

$$\min_{B:B^\top B=I_d} \sum_{i=1}^n \sum_{j=1}^n \rho_{\hat{\sigma}} \left[y_i - \left\{ a_j + \sum_{k=1}^d \beta_k^\top b_{jk} \left(X_i - X_j \right) \right\} \right] w_{ij}. \tag{7}$$

This represents a standard regression problem with n^2 observations and pd variables, which can be estimated by standard parametric methods. Although simulations show that estimating MAVE this way leads to slightly better estimates than the original algorithm, the size of regression (7) will be enormous as the sample size increases, which will hinder computation. For example, there are very fast algorithms available for computing least squares and L_1 regression in large data sets, see Koenker and Portnoy (1997), and even in these two special cases computation becomes 10–20 times slower than the original algorithm for samples of just 100 observations. This disqualifies such an algorithm from practical use.

3.2. One-step estimation

To be able to employ the fast original MAVE algorithm, robustness has to be achieved only by modifying weights w_{ij} in (2). To achieve this, we propose to use one-step M-estimators as discussed in Fan and Jiang (1999) and Welsh and Ronchetti (2002): the (iteratively) reweighted least squares approach. First using an initial highly robust estimate $\hat{\beta}_0 = \left\{ \hat{B}_0, \hat{a}_{0j}, \hat{b}_{0j} \right\}$, we construct weights w_{ij}^* such that the objective function (2) is equivalent to (6) at $\hat{\beta}_0$: $w_{ij}^* = w_{ij} \rho_{\hat{\sigma}}(r_{0i}) / r_{0i}^2$, where $r_{0i} = y_i - \left\{ \hat{a}_{0j} + \hat{b}_{0j}^\top \hat{B}_0^\top (X_i - X_j) \right\}$. Next, we perform the original least-squares-based algorithm using the constructed weights w_{ij}^*. Contrary to Fan and Jiang (1999), we use the L_1 regression as the initial robust estimator, which guarantees robustness against outliers (Giloni and Simonoff, 2005). Protecting against leverage points is not necessary since estimation is done in a local window given by the bandwidth and kernel function.

3.3. Robust cross-validation

The robust estimation of the DR matrix B_0 is not sufficient if dimension d is not known. As indicated by Ronchetti et al. (1997) and the example in Section 2.4, using the CV criterion (4) can lead to a bias in dimension estimation (and bandwidth selection) even if a robust estimator of B_0 is used. Now, due to the local nature of nonparametric regression, we have to "protect" CV primarily against outliers in the y-direction. In this context, the L_1 estimator is highly robust and the same should apply to CV based on L_1 rather than L_2 norm (Giloni and Simonoff, 2005). Thus, we propose to use instead of (4) the L_1 CV of Wang and Scott (1994),

$$\mathrm{CV}(d) = \sum_{j=1}^n \left| y_j - \sum_{i=1, i \neq j}^n \frac{y_i K_h \left\{ \hat{B}^\top (X_i - X_j) \right\}}{\sum_{i=1, i \neq j}^n K_h \left\{ \hat{B}^\top (X_i - X_j) \right\}} \right|, \tag{8}$$

to determine both the optimal bandwidth and dimension. This procedure is further referred to as CVA instead of CV, which is used only for the L_2 CV.

4. Simulations

To study the finite sample performance of MAVE, OPG, and their modifications proposed in Section 3, we perform a set of Monte Carlo simulations under various distributional assumptions. The employed data-generating model is introduced in Section 4.1. Next, we compare the performance of all methods in estimating the directions of the DR space (Section 4.2). Finally, we examine the performance of the CV and CVA dimension estimation (Section 4.3).

4.1. Simulation models

Throughout this section, we consider the data-generating model

$$y_i = \left(X_i^\top \beta_1\right)^2 - \left(0.5 + X_i^\top \beta_2\right)^2 + 15 \cos\left(X_i^\top \beta_3\right) + 0.5\varepsilon_i, \tag{9}$$

where the vector X_i of explanatory variables has the standard normal distribution in $\mathbb{R}^{10}$ and $\beta_1 = (1, 2, 3, 0, 0, 0, 0, 0, 0, 0)/\sqrt{14}$, $\beta_2 = (-2, 1, 0, 1, 0, 0, 0, 0, 0, 0)/\sqrt{6}$, and $\beta_3 = (0, 0, 0, 0, 0, 0, 0, 1, 1, 1)/\sqrt{3}$. The effective DR space given by $B_0 = (\beta_1, \beta_2, \beta_3)$ has thus dimension $D = 3$. To compare the robustness properties of all estimators, we use three error distributions.

(1) The standard normal errors, $\varepsilon_i \sim N(0, 1)$, generate Gaussian data without any outlying observations.
(2) The Student distributed errors, $\varepsilon_i \sim t_1$, with one degree of freedom simulate data from a heavy-tailed distribution.
(3) The contaminated errors, $\varepsilon_i \sim 0.95N(0, 1) + 0.05U(-600, 600)$, represent (normal) data containing approximately 5% of outlying observations, which are symmetrically distributed around zero.
(4) The asymmetrically contaminated errors, $\varepsilon_i \sim 0.95N(0, 1) + 0.05U(0, 600)$, represent (normal) data containing approximately 5% of outlying observations.

For the sake of brevity, we refer to these three cases as NORMAL, STUDENT, OUTLIERS, and AOUTLIERS, respectively. For all simulations from (9), we use sample size $n = 100$ and 500 repetitions (we observed that the results for larger samples sizes, such as $n = 200$, are qualitatively the same as for $n = 100$). Nonparametric smoothing employs the Gaussian kernel in all cases. Bandwidth is cross-validated using CVA for the proposed robust methods and using both CV and CVA for the original methods.

Let us note that we compare the methods using the same distance measure of the estimated space $\hat{B}$ and the true space $B_0 = (\beta_1, \beta_2, \beta_3)$ as Xia et al. (2002): $m\left(\hat{B}, B_0\right) = \left\|\left(I - B_0 B_0^\top\right)\hat{B}\right\|$ for $d < D = 3$ and $m\left(\hat{B}, B_0\right) = \left\|\left(I - \hat{B}\hat{B}^\top\right)B_0\right\|$ for $d \geqslant D = 3$ ($D = 3$ is the true dimension of the reduced space used in our simulations, whereas d denotes the dimension used for estimation).

4.2. Estimation of dimension reduction space

MAVE, OPG, and their modifications are now compared by means of simulations. The estimators include MAVE defined in (6) and OPG defined in (5) using the following functions $\rho_{\hat{\sigma}}$:

1. $\rho_{\hat{\sigma}}(x) = x^2$ (least squares),
2. $\rho_{\hat{\sigma}}(x) = |x|$ (least absolute deviation),
3. $\rho_{\hat{\sigma}}(x) = \int \text{sgn}(x) \min(|x|, \hat{\sigma})\, dx$ (M-estimate with the Huber function),
4. $\rho_{\hat{\sigma}}(x) = \int \text{sgn}(x) \max\left\{0, \min\left(|x|, \hat{\sigma}\right) - \max\left(0, |x| - 2\hat{\sigma}\right)\right\}\, dx$ (M-estimate with the Hampel function).

We refer to these functions as LS, L1, HUBER, and HAMPEL. The first choice corresponds to standard MAVE and OPG. The second one relies on the local L-estimation that, in the case of MAVE, has to be performed by the slow algorithm based on alternating formulations (6) and (7), see Section 3.1. The last two choices represent MAVE and OPG relying on the local one-step M-estimation as described in Section 3.2.

Before discussing the estimation results, let us recall our concerns about computational speed of each method that motivated the use of sub-optimal, but fast OPG and precludes practical use of MAVE based on the L_1 estimator. Given model (9), Table 3 summarizes computational times for all methods relative to OPG–LS (note that OPG–LS was typically estimated in 0.5–2 s depending on the computer in use). The results are to some extent implementation-specific, which does not allow to directly compare optimized least-square methods and unoptimized robust variants. Nevertheless, it is clear that OPG can be computed much faster than MAVE (if the algorithm is properly optimized) and that the general algorithm from Section 3.1 used for MAVE–L1 is too slow for real applications.

Let us first deal with the results concerning OPG and its modifications presented in Table 4. For data NORMAL, all modifications outperform the original method (OPG–LS–CV). It is interesting to note that, in the case of OPG–LS,

Table 3
Average computational times for each method at sample size $n = 100$ and dimension $d = 3$ relative to standard OPG

Computational times				
	LS	L_1	HUBER	HAMPEL
OPG	1.0	4.6	10.4	12.3
MAVE	7.5	298.7	13.8	14.5

Table 4
Median errors of OPG estimates for dimension $d = 3$

Simulated data NORMAL, STUDENT, and OUTLIERS					
Data	Method	$m_0\left(\hat{\beta}_1\right)$	$m_0\left(\hat{\beta}_2\right)$	$m_0\left(\hat{\beta}_3\right)$	$m_0\left(\hat{B}\right)$
NORMAL	OPG LS CV	0.005	0.137	0.235	0.310
	OPG LS CVA	0.004	0.132	0.232	0.299
	OPG L1	0.004	0.132	0.207	0.278
	OPG HUBER	0.004	0.130	0.209	0.285
	OPG HAMPEL	0.005	0.147	0.211	0.310
STUDENT	OPG LS CV	0.065	0.483	0.588	0.929
	OPG LS CVA	0.062	0.461	0.554	0.914
	OPG L1	0.018	0.351	0.459	0.825
	OPG HUBER	0.016	0.338	0.458	0.798
	OPG HAMPEL	0.015	0.309	0.439	0.714
OUTLIERS	OPG LS CV	0.623	0.600	0.603	0.971
	OPG LS CVA	0.663	0.608	0.592	0.971
	OPG L1	0.013	0.302	0.445	0.722
	OPG HUBER	0.007	0.183	0.340	0.476
	OPG HAMPEL	0.006	0.164	0.285	0.385
AOUTLIERS	OPG LS CV	0.607	0.620	0.570	0.973
	OPG LS CVA	0.602	0.607	0.572	0.974
	OPG L1	0.016	0.359	0.606	0.906
	OPG HUBER	0.009	0.266	0.588	0.860
	OPG HAMPEL	0.008	0.201	0.484	0.674

the CVA criterion performed better than CV. This might be due to a relatively small number of observations relative to the dimension of the data, but our feeling is that this is typical rather than exceptional for many DR applications. Nevertheless, even the comparison of OPG–LS–CVA, OPG–HUBER, and OPG–HAMPEL does not reveal significant differences in the performance of these methods. For data STUDENT, all robust versions of OPG provide similar results, whereas the estimates by original OPG–LS exhibit rather large errors, especially in the first direction β_1. For data OUTLIERS and AOUTLIERS, there is a large difference between nonrobust OPG–LS and the robust modifications of OPG, which documents sensitivity of the original OPG estimator to outliers. Although the performance of all robust estimators is relatively similar, OPG–HAMPEL is best due to its full-rejection feature (observations with too large residuals get zero weight).

The simulation results for MAVE are summarized in Table 5. For data NORMAL, we again observe the positive influence of CVA on the original MAVE. All estimates except for MAVE–HAMPEL perform almost equally well. MAVE–HAMPEL provides worst estimates since it fully rejects some data points, which surprisingly did not matter in the case of OPG. For data STUDENT, OUTLIERS and AOUTLIERS, all robust versions of MAVE by far out-perform the original MAVE, which exhibit rather large errors in all directions. Similarly to OPG, MAVE–HAMPEL is best due to the full rejection of extreme observations; this effect is rather pronounced in data OUTLIERS and AOUTLIERS.

Table 5
Median errors of MAVE estimates for dimension $d = 3$

Simulated data NORMAL, STUDENT, and OUTLIERS

Data	Method	$m_0\left(\hat{\beta}_1\right)$	$m_0\left(\hat{\beta}_2\right)$	$m_0\left(\hat{\beta}_3\right)$	$m_0\left(\hat{B}\right)$
NORMAL	MAVE LS CV	0.007	0.090	0.095	0.171
	MAVE LS CVA	0.004	0.073	0.079	0.134
	MAVE L1	0.005	0.056	0.092	0.134
	MAVE HUBER	0.006	0.104	0.102	0.185
	MAVE HAMPEL	0.009	0.133	0.152	0.270
STUDENT	MAVE LS CV	0.278	0.352	0.532	0.900
	MAVE LS CVA	0.220	0.389	0.502	0.892
	MAVE L1	0.024	0.304	0.425	0.758
	MAVE HUBER	0.034	0.289	0.394	0.677
	MAVE HAMPEL	0.031	0.244	0.344	0.579
OUTLIERS	MAVE LS CV	0.717	0.703	0.723	0.982
	MAVE LS CVA	0.721	0.674	0.714	0.982
	MAVE L1	0.025	0.129	0.222	0.370
	MAVE HUBER	0.018	0.225	0.207	0.432
	MAVE HAMPEL	0.010	0.165	0.187	0.333
AOUTLIERS	MAVE LS CV	0.716	0.698	0.714	0.979
	MAVE LS CVA	0.709	0.687	0.706	0.979
	MAVE L1	0.027	0.192	0.385	0.751
	MAVE HUBER	0.027	0.357	0.480	0.851
	MAVE HAMPEL	0.013	0.231	0.277	0.504

All presented results clearly document the need for and advantages of the proposed robust modifications of MAVE and OPG. Comparing the results for both methods, they have rather similar structure, but MAVE always outperforms OPG when considering the estimation of the whole DR space. On the other hand, OPG seems to be very good in identifying the first and to some extent also the second direction. Let us note that results can be further improved by adjusting function $\rho_{\hat{\sigma}}$, the choice of which was typical rather than optimal.

4.3. Cross-validation simulations

The estimation of the DR space considered in the previous section is now complemented by a study on the estimation of the effective dimension D. Simulating again from model (9), we estimated the DR dimension $\hat{d}$ ($D = 3$) using MAVE and OPG with $\rho_{\hat{\sigma}}$-functions LS and HAMPEL and the CV and CVA criteria (4) and (8), respectively. Note that the design of model (9) makes the identification of the third direction β_3 difficult given the rather small sample size $n = 100$. Therefore, we accept not only estimate $\hat{d} = 3$, but $\hat{d} = 2$ as appropriate. Moreover, as a referee noted, overestimating the DR dimension does not lead to the loss of information because the estimated DR space is larger than the true one. Hence, we consider here also slight overestimation, $\hat{d} = 4$, as acceptable.

Results for all models are summarized in Tables 6 and 7 for MAVE and OPG, respectively. Judging all methods by the number of simulated data sets for which the estimated dimension $\hat{d}$ falls between two and four, MAVE always performs similarly or better than OPG. For the original methods, the CVA criterion is preferable to CV as in Section 4.2. A more interesting result is that MAVE–HAMPEL used with CVA outperformed the original MAVE and OPG not only in data OUTLIERS and STUDENT, but also in the case of data NORMAL. The only case where MAVE–LS is preferable is in the number of data set for which the estimated dimension $\hat{d}$ equals 3 in data NORMAL. This could be partially accounted for by the relatively small sample size. Finally, notice that the use of a robust DR method such as MAVE–HAMPEL does not suffice to identify the dimension of the DR space: a robust CV criterion such as CVA has to be used as well.

Table 6
Estimates of the DR dimension by MAVE variants using L_2 (CV) and L_1 (CVA) cross validation

Simulated data NORMAL, STUDENT, and OUTLIER

Data	Estimation	CV	$\hat{d} = 1$	$\hat{d} = 2$	$\hat{d} = 3$	$\hat{d} = 4$	$\hat{d} \geqslant 5$
NORMAL	MAVE LS	CV	13	34	52	1	0
	MAVE LS	CVA	6	35	59	0	0
	MAVE HAMPEL	CV	6	53	40	1	0
	MAVE HAMPEL	CVA	6	54	40	0	0
STUDENT	MAVE LS	CV	24	35	30	8	3
	MAVE LS	CVA	15	32	38	11	4
	MAVE HAMPEL	CV	18	39	27	8	8
	MAVE HAMPEL	CVA	16	56	26	2	0
OUTLIERS	MAVE LS	CV	0	1	14	20	65
	MAVE LS	CVA	6	5	19	15	55
	MAVE HAMPEL	CV	2	2	5	8	83
	MAVE HAMPEL	CVA	7	51	30	4	8

Entries represent the fractions of samples (in percent) out of 500 with estimated dimension $\hat{d}$.

Table 7
Estimates of the DR dimension by OPG variants using L_2 (CV) and L_1 (CVA) cross validation

Simulated data NORMAL, STUDENT, and OUTLIER

Data	Estimation	CV	$\hat{d} = 1$	$\hat{d} = 2$	$\hat{d} = 3$	$\hat{d} = 4$	$\hat{d} \geqslant 5$
NORMAL	OPG LS	CV	7	44	46	2	1
	OPG LS	CVA	8	57	33	1	1
	OPG HAMPEL	CV	11	45	42	0	2
	OPG HAMPEL	CVA	10	50	38	2	0
STUDENT	OPG LS	CV	19	35	27	12	7
	OPG LS	CVA	11	30	34	17	8
	OPG HAMPEL	CV	13	31	35	14	7
	OPG HAMPEL	CVA	22	48	27	3	0
OUTLIERS	OPG LS	CV	2	1	13	13	71
	OPG LS	CVA	15	7	15	7	56
	OPG HAMPEL	CV	2	3	3	12	80
	OPG HAMPEL	CVA	9	40	29	7	15

Entries represent the fraction of samples (in percent) out of 500 with estimated dimension d.

5. Conclusion

We proposed robust enhancements of MAVE and OPG that can perform equally well as the original methods under 'normal' data, are robust to outliers and heavy-tailed distributions, and are easy to implement. Should we pick up one method for a general use, MAVE–HUBER seems to be the most suitable candidate as (i) MAVE–LS is not robust, (ii) MAVE–L1 is slow to compute, see Section 3.1, and (iii) MAVE–HAMPEL does not perform so well for normal data. On the other hand, the question of a CV criterion for robust nonparametric regression is largely open and requires further research.

References

Amato, U., Antoniadis, A., De Feis, I., 2005. Dimension reduction in function regression with applications. Comput. Statist. Data Anal., in press.

Antoniadis, A., Lambert-Lacroix, S., Leblanc, F., 2003. Effective dimension reduction methods for tumor classification using gene expression data. Bioinformatics 19 (5), 563–570.

Bellman, R.E., 1961. Adaptive Control Processes. Princeton University Press, Princeton.

Boente, G., Fraiman, R., 1994. Local L-estimators for nonparametric regression under dependence. J. Nonparametric Statist. 4, 91–101.

Cook, R.D., Weisberg, S., 1991. Sliced inverse regression for dimension reduction: comment. J. Amer. Statist. Assoc. 86, 328–332.

Cook, R.D., Yin, X., 2001. Dimension reduction and visualization in discriminant analysis. Austral. New Zealand J. Statist. 43, 147–177.

Critchley, F., 2002. Discussion on the paper 'An adaptive estimation of dimension reduction space'. J. Roy. Statist. Soc., Ser. B. 64, 392.

Cui, H., Li, G., 2002. Discussion on the paper 'An adaptive estimation of dimension reduction space'. J. Roy. Statist. Soc., Ser. B 64, 394.

Čížek, P., 2004. Smoothed local L-estimation with an application. In: Hubert, M., Pison, G., Struyf, A., Van Aelst, S. (Eds.), Theory and Applications of Recent Robust Methods. Birkhäuser, Basel, pp. 59–70.

Čížek, P., Härdle, W., Yang, L., 2002. Discussion on the paper 'An adaptive estimation of dimension reduction space'. J. Roy. Statist. Soc., Ser B. 64, 397.

Fan, J., Jiang, J., 1999. Variable bandwidth and one-step local M-estimator. Sci. China Ser. A 29, 1–15.

Gather, U., Hilker, T., Becker, C., 2001. A robustified version of sliced inverse regression. In: Fernholz, L.T., Morgenthaler, S., Stahel, W. (Eds.), Statistics in Genetics and in the Environmental Sciences. Birkhäuser, Basel, pp. 147–157.

Giloni, A., Simonoff, J.S., 2005. The conditional breakdown properties of least absolute value local polynomial estimators. Nonparametric Statist. 17 (1), 15–30.

Hampel, F.R., Ronchetti, E.M., Rousseeuw, P.J., Stahel, W.A., 1986. Robust Statistics, The Approach Based on Influence Function. Wiley, New York.

Härdle, W., Stoker, T.M., 1989. Investigating smooth multiple regression by method of average derivatives. J. Amer. Statist. Assoc. 84, 986–995.

Härdle, W., Tsybakov, A.B., 1988. Robust nonparametric regression with simultaneous scale curve estimation. Ann. Statist. 16, 120–135.

Härdle, W., Tsybakov, A.B., 1991. Sliced inverse regression for dimension reduction: comment. J. Amer. Statist. Assoc. 86, 333–335.

He, X., Fung, W.K., 2001. Discussion on the paper 'Dimension reduction and visualization in discriminant analysis'. Austral. New Zealand J. Statist. 43, 190–193.

Koenker, R., Portnoy, S., 1997. The Gaussian hare and the Laplacian tortoise: computability of squared-error vs. absolute-error estimators. Statist. Sci. 12, 279–300.

Li, K.C., 1991. Sliced inverse regression for dimension reduction. J. Amer. Statist. Assoc. 86, 316–342.

Rousseeuw, P.J., Leroy, A.M., 2003. Robust Regression and Outlier Detection. Wiley, New York.

Ronchetti, E., Field, C., Blanchard, W., 1997. Robust linear model selection by cross-validation. J. Amer. Statist. Assoc. 92, 1017–1023.

Sakata, S., White, H., 1998. High breakdown point conditional dispersion estimation with application to S&P 500 daily returns volatility. Econometrica 66 (3), 529–567.

Wang, F.T., Scott, D.W., 1994. The L_1 method for robust nonparametric regression. J. Amer. Statist. Assoc. 89, 65–76.

Welsh, A.H., Ronchetti, E., 2002. A journey in single steps: robust one-step M-estimation in linear regression. J. Statist. Plann. Inference 103, 287–310.

Xia, Y., Tong, H., Li, W.K., Zhu, L.-X., 2002. An adaptive estimation of dimension reduction space. J. Roy. Statist. Soc., Ser. B 64, 363–410.

Yao, Q., Tong, H., 1994. On subset selection in nonparametric stochastic regression. Statistica Sinica 4, 51–70.

On the appropriateness
of inappropriate VaR models

By Wolfgang Härdle, Zdeněk Hlávka and Gerhard Stahl*

Summary The Value-at-Risk calculation reduces the dimensionality of the risk factor space. The main reasons for such simplifications are, e.g., technical efficiency, the logic and statistical appropriateness of the model. In Chapter 2 we present three simple mappings: the mapping on the market index, the principal components model and the model with equally correlated risk factors. The comparison of these models in Chapter 3 is based on the literature on the verification of weather forecasts (Murphy and Winkler, 1992; Murphy, 1997). Some considerations on the quantitative analysis are presented in the fourth chapter. In the last chapter, we present empirical analysis of the DAX data using XploRe.

Zusammenfassung Die Berechnung des VaR führt zur Reduktion der Dimension des Raumes der Risikofaktoren. Die vorzunehmenden Vereinfachungen resultieren aus unterschiedlichen Beweggründen, z.B. technische Effizienz, Sachlogik der Ergebnisse und statistische Adäquanz des Modells. Im Kapitel 2 stellen wir drei gängige Mappingverfahren vor: das Marktindexmodell, das Hauptkomponentenmodell und das Modell mit gleichkorrelierten Risikofaktoren. Impulse für Methoden zum Vergleich dieser Modelle im Kapitel 3 kamen vor allem aus der Literatur zur Praxis der Beurteilung von Wetterprognosen (Murphy und Winkler, 1992; Murphy, 1997). Umfangreiche Überlegungen zu einer quantitativen Analyse werden im vierten Kapitel dieser Arbeit vorgestellt. Die empirische Analyse der DAX Daten wird abschließend mit XploRe durchgeführt.

1. Introduction

The well-known G.E.P. Box's remark *all models are wrong, but some are useful* describes one of the main problems arising in each application of Value-at-Risk models: the 'useful model' should be as simple as possible. This requires appropriate selection of variables in order to minimize the bias and the variance of the model simultaneously.

Value-at-Risk (VaR), as a standard measure of market risk, predicts the possible loss of a certain portfolio at a specified level α. Mathematically, VaR is the quantile at level α of the random future portfolio value changes. The task of calculating VaR for high dimensional portfolios requires simple and fast algorithms. In the following, we will present some very simple common methods for VaR calculation and we will investigate their appropriateness.

Our desire to choose a simple model leads to dimension reduction techniques applied either in the space of risk-factors or the distributions of the corresponding random variables. In practice, the – from RiskMetrics by J.P.Morgan/Reuters originating – term *mapping* is used in both cases. The importance of mapping procedures is overwhelming. Many VaR papers

focus on one-dimensional identically distributed random variables with an insufficient perspective towards the practically more relevant multivariate situation. The question of the interplay between dimensionality- and model reduction and VaR precision is not asked and, consequently, not answered satisfactorily.

Corresponding to the complicated character of the VaR models, the commonly applied simplifications have various motivations. The most important ones are: the technical efficiency, the interpretability of the results, and the statistical appropriateness of the model.

Most often, the superior technical performance of the implemented model is in the focus of the implementation. A lower dimensional model allowing simple and precise handling of the numerical calculations and, at the same time, describing the important dynamic factors, can serve this purpose very well. Furthermore, the choice of the risk-factors depends also on availability of the up-to-date quality approved input parameters such as, for example, market data. Many institutes use the data matrix provided by the RiskMetrics group and containing about 400 risk-factors.

Although the, by mapping simplified, system cannot overcome the inherent rank defficiencies, it can still, at least in a statistical sense, reflect the covariance structure. A carefully parameterized model reduces rank defficiency in the estimated covariance matrix (Davé and Stahl, 1998). This is the case especially for the multicollinearity, i. e., the situation often encountered in practice with a very large number of variables observed during a small time window. The dimension reduction techniques remove the insignificant correlations and can be used for clear and concise summarizing reports, see Reimers and Zerbs (1998).

The VaR models are applied not only as a measure of risk but also as a means of fulfilling the regulatory office requirements. Special care must be exhibited in the case of capital reserves for open positions since their size depends strongly on the model, see Huschens (1998). In the case of Maximum Loss risk measures, Studer (1997) has shown that the dimension is nonlinearly related to the calculated risk.

The following example considers portfolios consisting of long or short positions. The market value ν_s at time s of the portfolio Π_t is given as

$$\nu_s(\Lambda_t) = \Lambda_t^\top \mathbf{P}_s,$$

where the vector $\Lambda_t \in \mathbb{R}^n$ denotes the nominal volumes at time t and $\mathbf{P}_s$ denotes the corresponding vector of prices. A portfolio such as $\{VW, DC\}$ with $\Lambda_t = (100, 200)^\top$ and $p_s = (40, 60)^\top$ would lead to the value $\nu_s = 16\ 000$ for $s = 16.6.2005$. The stochastic models for the risk of the value changes for given price $\mathbf{P}_t = p_t$ at time t,

$$\nu_{t+h}(\Lambda_t) - \nu_t(\Lambda_t) = \Lambda_t^\top (\mathbf{P}_{t+h} - p_t) = (\Lambda_t p_t)^\top \frac{(\mathbf{P}_{t+h} - p_t)}{p_t} \qquad (1)$$

$$= w_t^\top \mathbf{R}_{t+h}, \qquad (2)$$

are not based on the price process, but instead on the vector $\mathbf{R}_t$ of the (discrete) returns. Note that both the multiplication and division by the vector p_t on the right hand side of (1) are componentwise. The exposure vector, $w_t = (w_{1t}, \ldots, w_{nt})^\top$, is defined as a componentwise product of Λ_t and p_t. In the above example, the exposure is $w_t = (4\,000, 12\,000)^\top$.

The *conditional loss distribution*

$$\mathcal{L}(L_{t+h} \mid \mathcal{H}_t) \tag{3}$$

of the possible losses of the portfolio $L_{t+h} = -w_t^\top \mathbf{R}_{t+h}$ in a time horizon h, is the object of our interest. In the following, we will set $h = 1$ (one trading day). The conditioning information $\mathcal{H}_t$, mostly defined by a moving window of fixed length N, consists in the simplest cases only of the observed market prices.

Assuming that the distribution of $\mathbf{R}_{t+h}$ is known, we could assess the distribution of losses (3) via Monte Carlo simulations since the changes of the portfolio value are a known function of the vector

$$L_{t+h} = \Psi_t(\mathbf{R}_{t+h}) . \tag{4}$$

In some situations, one can derive explicit solutions that lead to evaluation of an integral involving the loss distribution.

A portfolio is called linear if the function Ψ_t in (4) is linear. For linear portfolios, the returns $\mathbf{R}_t$ are often parametrized by elliptical distributions with

$$E(\mathbf{R}_t) = 0$$

(RiskMetrics by J.P.Morgan/Reuters, 1996; Litterman and Winkelmann, 1998). For definiton of elliptical distribution refer to Härdle and Simar (2003). In this situation, the predictive distributions (3) for linear portfolios can be evaluated analytically. The dynamics is mostly analyzed using (I)GARCH- or White Noise models, Gouriéroux (1997).

In a Gaussian model, e. g., a White Noise model with $N(0, \sigma_{it}^2)$ distributed innovations, the VaR of the i-th asset at level α is given by

$$VaR_i(\alpha, h, t) = w_{it} q_\alpha \sigma_{ith} ,$$

where $\sigma_{ith}^2 = E(R_{i,t+h}^2)$ and q_α denotes the α–quantile of the standard normal distribution. Denoting by

$$\mathbf{VaR} = (VaR_1, \cdots, VaR_n)^\top$$

the VaR-vector (at level α and horizon h), the VaR of a portfolio Π_t can be expressed as

$$\mathrm{VaR}(\Pi_t) = \mathrm{VaR}(\Pi_t, \alpha, h) = \sqrt{\mathbf{VaR}^\top C_t \mathbf{VaR}}, \tag{5}$$

where C_t is the *correlation matrix* of $\mathbf{R}_{t+h}$ corrected by the signs of our positions, i. e., $c_{ijt} = \rho_{ijt} \operatorname{sign}(w_{it}) \operatorname{sign}(w_{jt})$. The vector $\mathbf{VaR}$ provides

insight into possible concentrations into single risk factors. Hence, the expression (5) is to be preferred over the similar expression

$$VaR(\Pi_t) = \sqrt{w_t^\top \Sigma_t w_t} \,.$$

In the next section, we present three common mappings. The third section is devoted to the diagnostics. In the fourth section, we will present our considerations related to the quantitative analysis. The empirical analysis in the last section has been performed with XploRe.

2. Mappings

The mappings that are used in practice, are mostly based in the space of the portfolio variables and depend on projections (principal components) or regression methods. A prominent example is the mapping of a portfolio onto the market index. Applications based on the simplification of the parameter space are in practice not very common. RiskMetrics by J.P.Morgan/Reuters (1996) provides an important example: the sparse modelling of IGARCH models using only one smoothing parameter for all risk factors.

A VaR model of a linear portfolio based on (5) would require estimation of altogether $n(n+1)/2$ parameters of the covariance matrix. The amount n of the portfolio variables is mostly too large to consider every variable as a risk factor in the VaR model of the general market risk. Considering the fact that realistic values of n are usually much larger than 5 000 and, in certain cases, reach values as high as 500 000, one tries to keep the number of risk factors as small as possible, i. e.,

$$\dim(\tilde{\mathbf{R}}) = d \ll n.$$

In the case of the RiskMetrics data matrix, the number of risk factors, d, is approximately 400. Models with $d > 5\,000$ are rarely used in practice.

In order to apply a mapping, we have to select a vector of risk factors:

$$\tilde{\mathbf{R}} = (\tilde{R}_1, \cdots, \tilde{R}_d)^\top. \tag{6}$$

The choice is usually done empirically. Formally, the mapping can be described using a function $\mathcal{M}$ as follows:

$$\mathcal{M} : \mathbf{R} \longrightarrow \mathcal{M}(\mathbf{R}) = \tilde{\mathbf{R}}.$$

For a given $\tilde{\mathbf{R}}$ we have to derive a function ψ such that the distribution of the variable $\tilde{L}_t = \psi(\tilde{\mathbf{R}})$ approximates the distribution of the portfolio losses $L_t = \Psi(\mathbf{R})$ as good as possible. The function ψ maps the exposures w_t implicitly on the risk factors. In the ideal situation, with similarly defined $\tilde{\mathcal{H}}_t$, we would have

$$\mathcal{L}(L_{t+h} \mid \mathcal{H}_t) = \mathcal{L}(\tilde{L}_{t+h} \mid \tilde{\mathcal{H}}_t)$$

or
$$VaR(\Pi_t) = VaR(\tilde{\Pi}_t).$$
In the simplest case, $\mathcal{M}$ is an inclusion mapping or the risk factors $\tilde{R}_i$ are functions of $\mathbf{R}$. For example, $\tilde{\mathbf{R}}$ could contain a subset of the elements of $\mathbf{R}$ or a linear combination of $\mathbf{R}$.

The common mapping procedures are mostly based on one meta portfolio characterized by the set of all available instruments, without taking into account the characteristics of the specific portfolio, e. g., by using the w_t in the construction of the mapping. In the following, we explore the index model published by RiskMetrics by J.P.Morgan/Reuters (1996), the mapping on synthetic indices (principal components), and we suggest a new method based on a simplified correlation matrix.

2.1. REGRESSION APPROACH: INDEX MODEL. Among all mappings, the market index model is most commonly applied in the practice, Jorion (1997). Cuthbertson (1996) provides an overview of the common variants of the CAPM model as well as the factor models, and he gives a critical assessment of the corresponding model assumptions. The use of the beta factors is based on the assumed relation between the returns of an asset and the returns of a corresponding index, e. g., the DAX. The risk of having an open position in a certain asset is explained using the risk of the index by mapping the position on the index.

Let R_{it} denote the return of the i-th asset and R_{mt} the return of the index. The index model is then specified by the following equations:
$$R_{it} = \beta_i R_{mt} + \varepsilon_{it}, \tag{7}$$
where $E(\varepsilon_{it}) = 0$, $E(\varepsilon_{it}^2) = \sigma_{\varepsilon,it}^2$, $E(\varepsilon_{it}R_{mt}) = 0$, and $E(\varepsilon_{it}\varepsilon_{jt}) = 0$.

These equations are valid for times $t = 1,\ldots,T$ and ε_{it} is called the residual (or idiosyncratic) variable. The above model specification defines seemingly unrelated regression equations. The consequence is that, in this case, the parameters do not have to be estimated jointly.

The variance of R_{it} can be decomposed as
$$\sigma_{it}^2 = \beta_i^2 \sigma_{mt}^2 + \sigma_{\varepsilon,it}^2, \tag{8}$$
and for the covariance we have
$$\sigma_{ijt}^2 = \beta_i \beta_j \sigma_{mt}^2. \tag{9}$$
Under the assumptions of the model (7) and using (8) and (9) we have for the variance of (3) that

$$
\begin{aligned}
w_t^\top \Sigma_t w_t &= \sum_{j,i=1}^{n} \sigma_{it}\sigma_{jt}\rho_{ijt}w_{it}w_{jt} \\
&= \sum_{j,i=1}^{n} \beta_i\beta_j\sigma_{mt}^2 w_{it}w_{jt} + \sum_{i=1}^{n} \sigma_{\varepsilon,it}^2 w_{it}^2 \\
&= \sigma_{mt}^2 w_t^\top \beta\beta^\top w_t + w_t^\top D_{\varepsilon_t} w_t, \tag{10}
\end{aligned}
$$

where $\beta = (\beta_1, \cdots, \beta_n)^\top$ denotes the vector of betas. The diagonal matrix $D_{\varepsilon_t} = \mathrm{diag}(\sigma_{\varepsilon,1}^2, \cdots, \sigma_{\varepsilon,n}^2)$ denotes the covariance matrix of the vector ε. Equation (10) motivates the following approximations of the covariance matrix Σ_t:

$$\Sigma_{\mathcal{D}} = \sigma_m^2 \beta\beta^\top + D_\varepsilon \,, \tag{11}$$

$$\Sigma_\beta = \sigma_m^2 \beta\beta^\top \,. \tag{12}$$

The quality of the approximation strongly depends on the validity of (7). The first term in the sum in (10) allows us to interpret the approximation (12) as a mapping. In this case of a mapping on an index, Equation (7) becomes:

$$\mathcal{M}(w_t) = \beta^\top w_t \quad \text{and} \quad \mathcal{M}(\mathbf{R}) = \{R_m\}\,,$$

$$\sigma_{mt}^2 w_t^\top \beta\beta^\top w_t + w_t^\top D_{\varepsilon_t} w_t = \sigma_{mt}^2 \mathcal{M}^2(w_t) + w_t^\top D_{\varepsilon_t} w_t \,.$$

The advantages of this mapping are obvious: on one hand, we have to estimate only $n + 1$, respectively $2n + 1$ parameters $(\dim(\beta) = n, \dim(\varepsilon) = n,$ and $\sigma_m^2)$ of the matrix Σ_β or, respectively, $\Sigma_{\mathcal{D}}$ (instead of $n(n+1)/2$ parameters of Σ_t) and, on the other hand, the use of the risk factor R_m leads to a substantial dimension reduction of the VaR model. These mappings use approximations of $VaR(\Pi_t)$ that follow from Equations (11) and (12).

$$VaR_{\mathcal{D}}\{\mathcal{M}(\Pi_t)\} = q_\alpha \sqrt{w_t^\top \Sigma_{\mathcal{D}} w_t}$$

$$= \sqrt{\mathbf{VaR}_{\mathcal{M}}^\top \left(\mathbf{1}_{(n\times n)} + B^{-1} D_{\varepsilon_t} B^{-1}\right) \mathbf{VaR}_{\mathcal{M}}}\,, \quad (13)$$

$$VaR_\beta\{\mathcal{M}(\Pi_t)\} = q_\alpha \sqrt{w_t^\top \Sigma_\beta w_t} = \sqrt{\mathbf{VaR}_{\mathcal{M}}^\top \mathbf{1}_{(n\times n)} \mathbf{VaR}_{\mathcal{M}}}\,, \quad (14)$$

where the i-th element of $\mathbf{VaR}_{\mathcal{M}}$ denotes the VaR of the i-th asset mapped on the index R_m:

$$\mathbf{VaR}_{\mathcal{M}} = VaR_m \beta\,.$$

B denotes the diagonal matrix $\mathrm{diag}(|\beta_1|, \cdots, |\beta_n|)$ and $\mathbf{1}_{(n\times n)}$ is an $(n \times n)$ matrix of ones. The relations (13) and (14) show clearly how the mappings (11) and (12) simplify the VaR calculation:

1. In (13), C in (5) is replaced by $\mathbf{1}_{(n\times n)} + B^{-1} D_{\varepsilon_t} B^{-1}$.
2. In (14), C in (5) is replaced by $\mathbf{1}_{(n\times n)}$.
3. In (13) and (14), $\mathbf{VaR}$ in (5) is replaced by $\mathbf{VaR}_{\mathcal{M}}$.

Obviously, the effects of the above simplifications vary and their consequences cannot be, in general, predicted. The square VaR of a linear portfolio is a quadratic form of w_t and Σ_t:

$$VaR^2 = w_t^\top \Sigma_t w_t = \mathbf{VaR}^\top C_t \mathbf{VaR}\,.$$

Obviously,

$$\mathbf{VaR}_{\mathcal{M}}^\top C_t \mathbf{VaR}_{\mathcal{M}} \leq \mathbf{VaR}^\top C_t \mathbf{VaR} \leq \mathbf{VaR}^\top \mathbf{1}_{(n\times n)} \mathbf{VaR}\,.$$

The effect of a mapping cannot be, in general, fully determined even for portfolios consisting only from long positions. For portfolios, consisting from both long and short positions, is it completely impossible. The substitution of C_t by $\mathbf{1}_{(n \times n)}$ in (5) is conservative, whereas the substitution of **VaR** by $\mathbf{VaR}_{\mathcal{M}}$ tends to underestimate the actual risk. The elimination of long and short positions mapped on an index (quasi a stochastic netting) leads to problems that might arise when singular covariance matrices are used (Davé and Stahl, 1998). So, $\mathrm{rank}(\mathbf{1}_{(n \times n)}) = 1$ implies that there exists portfolio with exposure $\tilde{w}$ such that

$$\tilde{w}^\top \, \mathbf{1}_{(n \times n)} \, \tilde{w} = 0 \, .$$

Thus the VaRs corresponding to $\tilde{w}$ are equal to zero. Hence, the empirical analyses assessing the validity of the beta factor models are not accessible.

2.2. MAPPING ON PRINCIPAL COMPONENTS. The application of principal components is a well known procedure (Jolliffe, 1986; Christensen, 1991; Härdle and Simar, 2003). The orthogonalization of the variance matrix Σ of the vector **R** leads to

$$\Sigma = \Gamma \Lambda \Gamma^\top \, ,$$

where $\Lambda = \mathrm{diag}(\lambda_1, \cdots, \lambda_n)$ is the diagonal matrix of the eigenvalues and Γ is the matrix consisting of eigenvectors of Σ. Under the usual regularity assumptions $(\lambda_i > 0)$, **R** can be expressed as

$$\mathbf{R} = \Gamma Y.$$

The elements $(Y_1, \cdots, Y_n)$ of the vectors Y are uncorrelated and are called the principal components of **R**. The total variability of **R** is equal to the trace of Σ:

$$\mathrm{trace}(\Sigma) = \mathrm{trace}(\Gamma \Lambda \Gamma^\top) = \sum_{i=1}^{n} \lambda_i.$$

Thus, the principal components with the largest variances (eigenvalues) explain most of the total variability of **R** and, in this sense, can be considered as the most important influence factors. Ordering the eigenvalues so that $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_n$ and denoting by $\Gamma_{(i)}$ the i-th column of the corresponding Γ, the approximation

$$\mathbf{R} \approx \sum_{i=1}^{k} \Gamma_{(i)} Y_i \tag{15}$$

with $k < n$ explains

$$\xi_k = \frac{\sum_{i=1}^{k} \lambda_i}{\sum_{i=1}^{n} \lambda_i} \tag{16}$$

of the total variability of **R**. Here, the risk factors $\tilde{\mathbf{R}} = \mathcal{M}(\mathbf{R})$ are the principal components $Y_1, \ldots, Y_k$. Typically, the dimension k is chosen as the minimal number of the principal components that guarantees that $\xi_k > 1 - \xi^*$

where $0 < \xi^* < 1$. In a linear, normally distributed case, we have immediately for the Value-at-Risk of the mapped portfolio at level α that

$$VaR = \Phi^{-1}(\alpha)\sqrt{w_t^\top[\Gamma_{(1)} \cdots \Gamma_{(k)}]\Lambda[\Gamma_{(1)} \cdots \Gamma_{(k)}]^\top w_t} \, . \tag{17}$$

2.3. MAPPING OF THE PARAMETER SPACE: EQUALLY CORRELATED RISK FACTORS. The mappings presented in this section simplify the structure of the parameter space of the stochastic model of the risk factors

$$\Sigma_t = \mathrm{diag}(\sigma_1, \ldots, \sigma_n) \, C_t \, \mathrm{diag}(\sigma_1, \ldots, \sigma_n)$$

by the following restriction:

$$C_t = (\rho_{ijt})_{i,j=1,\cdots,n} \text{ with constant } \rho_{ijt} = \rho \in [-1, 1] \, . \tag{18}$$

Such models are currently applied in the area of credit risk, but similar ideas are also well-known in the portfolio theory, see Elton and Gruber (1995).

The estimation of the restricted correlation matrix can be based on the following random effects model:

$$R_{it}/\sigma_i = \alpha_t + \varepsilon_{it},$$

where $E\alpha_t = E\varepsilon_{it} = ER_{it} = 0$, $ER_{it}^2 = \sigma_i^2$, $E\alpha_t^2 = \sigma_\alpha^2$, $E\varepsilon_{it}^2 = \sigma_\varepsilon^2$, $E\alpha_s\varepsilon_{it} = 0$. For the variance and covariance of the returns we have

$$Var(R_{it}) = \sigma_i^2(\sigma_\alpha^2 + \sigma_\varepsilon^2) = \sigma_i^2$$
$$Cov(R_{it}, R_{jt}) = \sigma_i\sigma_j\sigma_\alpha^2.$$

We define $R_{it}^* = R_{it}/\sigma_i$. A suitable estimator of σ_i^2 is, for example, the empirical variance of the returns R_{it}, $t = 1, \ldots, T$. Next, we define $\bar{R}_t^* = \sum_{i=1}^n R_{it}^*$ and $\bar{\varepsilon}_t = \sum_{i=1}^n \varepsilon_{it}$. Obviously,

$$R_{it}^* - \bar{R}_t^* = \varepsilon_{it} - \bar{\varepsilon}_t \, ,$$

and we apply the following estimator:

$$\hat{\sigma}_\varepsilon^2 = \frac{\sum_{i=1}^n \sum_{t=1}^T (R_{it}^* - \bar{R}_t^*)^2}{n(T-1)} \, ,$$

$$\hat{\sigma}_\alpha^2 = 1 - \hat{\sigma}_\varepsilon^2 \, .$$

The standard deviation in the denominator of R_{it}^* is estimated by the empirical standard deviation.

The correlation matrix of the returns is then approximated by

$$\hat{\Sigma} = \mathrm{diag}(\sigma_1, \ldots, \sigma_n) \begin{pmatrix} 1 & \sigma_\alpha^2 & \cdots & \sigma_\alpha^2 \\ \sigma_\alpha^2 & 1 & \cdots & \sigma_\alpha^2 \\ & & \vdots & \\ \sigma_\alpha^2 & \cdots & \sigma_\alpha^2 & 1 \end{pmatrix} \mathrm{diag}(\sigma_1, \ldots, \sigma_n).$$

Comparing the advantages and disadvantages of the mappings presented in this section, it can be seen, that each of the mappings tries to reach, from its own perspective, a compromise between the precision of the VaR estimation, the simplicity of interpretation, the complexity with concern on the application of the procedure in practice, and the simplicity of its technical implementation.

3. DIAGNOSTICS

Value-at-Risk models are a special case of statistical prediction models, since the estimated distributions of the risk factors $\tilde{\mathbf{R}}_{t+h}$, describing our uncertainty about the future of the market, allow to access the conditional loss distribution

$$\mathcal{L}(L_{t+h} \mid \mathcal{H}_t)$$

in a horizon h. The empirical basis for judgements on the quality of the predictions of a VaR model is a series of pairs,

$$\{P_t, \nu_t\}_{i=1}^N, \tag{19}$$

consisting of the prediction, P_t, and the corresponding realization, ν_t, see Diebold *et al.* (1998).

The shorthand notation P_t denotes either the complete estimated predictive distribution or some derived parameter (e.g., VaR, tail-VaR, standard deviation), the letter P stands for 'prediction'. The use of the symbol $\widehat{P}_t$ in some situation stresses the fact that the distribution we work with is estimated.

The symbol ν_t denotes the corresponding change of the value of the portfolio Π_t in a fixed time horizon, here fixed as one day, $h = 1$. We assume that the observations ν_t are realizations of random variables N_t. We distinguish between l_t and ν_t: the variable L_t refers to the space of the risk factors, i.e., $\mathcal{M}(\mathbf{R})$, whereas N_t refers to the space of the portfolio variables.

3.1. TOOLS. The area of the probability forecasts is one of the less explored areas of mathematical statistics. Substantial part of the theoretical foundations stem from Dawid. His summary papers (Dawid, 1986; Dawid, 1997) provide extensive overview of the topic. The motivations came from the literature on the evaluation of the weather forecasts. The first publication comes from 1884 (Gilbert, 1884; Peirce 1884), the relevant literature begins with Brier's papers in fifties and advances steadily since sixties mainly through Murphy and Winkler (1992) and Murphy (1997).

The above mentioned applied papers focus on discrete random variables. The probability scale $[0, 1]$ is cut into k categories, further l denotes the number of the events, e.g., for weather forecasts $l = 2$ for $E = \{$rain, no rain$\}$. Let us assume that we have N probability forecasts $\{P_{ti}\}_{t=1}^N$ as well as the corresponding events $\{\nu_t\}_{t=1}^N$. Denoting by P_{ti}, $i = 1, \cdots, l$, the

probabilities predicted for t and by E_{ti} a zero-one variable which is equal to 1 if at time t we observe event from the category i ($\nu_t = i$) and 0 otherwise. Using this notation, the Brier score of a forecast system is defined as

$$\text{BS} = \frac{1}{N} \sum_{i=1}^{l} \sum_{t=1}^{N} (P_{ti} - E_{ti})^2. \tag{20}$$

In order to assess the quality of a forecast system in more detail, Murphy and Winkler (1992) start, in the context of discrete random variables, from Equation (19), ignoring the information of the time dependency, by considering the joint relative frequencies

$$h(p_i, e_j) = \frac{N_{ij}}{N}, \tag{21}$$

where N_{ij} denotes the number of times when the event e_j, $j = 1, \ldots, l$, was predicted with the probability p_i, $i = 1, \ldots, k$.

The factorization of (21)

$$h(p, e) = h(e \mid p)\, h(p) \tag{22}$$
$$= h(p \mid e)\, h(e) \tag{23}$$

into conditional and marginal frequency distributions is the basis for definitions of additional indices and plots.

IMPORTANT INDICES AND PLOTS. The Brier Score can be interpreted as a Mean Squared Error (MSE) because of its following representation:

$$\text{BS} = \sum_{i=1}^{k} \sum_{j=1}^{l} (p_i - e_j)^2 h(p_i, e_j) \tag{24}$$
$$= (\mu_p - \mu_e)^2 + \sigma_p^2 + \sigma_e^2 - 2\sigma_p \sigma_e \rho_{pe}. \tag{25}$$

The Skill Score is defined as a coefficient of determination of the predictive probabilities seen as a fitted value in the regression model for the events. It can be calculated as a square of the correlation coefficient between the predictions P_{ti} and the zero-one variables E_{ti}.

Another evaluation of the quality of the forecasts results from the comparison of

$$p_i \quad \text{and} \quad \frac{1}{l} \sum_{j=1}^{l} N_{ij}. \tag{26}$$

In a perfect forecast system, the probabilities p_i would correspond to the observed relative frequencies. If this is the case, we say that the forecasts are calibrated. In the calibration plot, where $\frac{1}{N} \sum_{j=1}^{k} N_{ij}$ is plotted against p_i, the points should ideally lie on the diagonal of the unit square.

The forecast method can discriminate between the events if the conditional distributions $h(p|e = 0)$ and $h(p|e = 1)$ are significantly different.

Typically, the distribution of $h(p|e = 0)$ should be concentrated close to zero and the distribution of $h(p|e = 1)$ should lie close to 1. The more different the two distributions are, the better the forecast can discriminate between the two possible values of event e. Some common measures include difference in means or variances of the two distributions. However, most informative are discrimination plots that display the two distributions graphically.

The joint distribution $h(p, e)$ leads to many other measures of calibration, discrimination, refinement, resolution, bias or skill of the forecast procedures. From the many choices we have decided to use the Brier Score and the Skill Score, that can be interpreted as the MSE and the coefficient of determination, respectively (Murphy and Winkler, 1992).

For appropriate VaR forecasts, the realizations of the variables $F_{L_t}(N_t)$ should not be distinguishable from independent draws from uniform $[0,1]$ distributions, where $F_{L_t}(x)$ denotes the distribution function of P_t. The P-P plot of the transformed observations

$$\{\hat{F}_{L_t}(\nu_t)\}_{t=1}^{N} \tag{27}$$

is called the (absolute) empirical calibration curve.

The weather and VaR forecasts usually differ only in one detail: in VaR models we focus on varying events (quantiles or forecast intervals) with fixed probabilities, by contrast in weather forecasts we obtain probability forecasts of fixed events.

The following indices and plots are motivated by their focus on the events, more precisely on the intervals $(-\infty, VaR]$. To this end, (19) has to be accordingly modified. For an interval forecast based on (5), the model is calibrated if the series

$$\left\{ s_t^\alpha = \frac{\nu_t}{\sigma_t} \right\}_{t=1}^{N} \tag{28}$$

cannot be distinguished from a Gaussian White Noise process. Here, σ_t denotes an estimate of the scale of ν_t. The Quantile-Quantile plots (Q-Q plots) are a convenient graphical device. An important numerical parameter is here

$$\kappa = \hat{\sigma}_{S^\alpha} \tag{29}$$

that attempts to quantify the amount of under- and overestimation of the risk forecasted by the model. It is easy to see that in the ideal case, if σ_t reflects perfectly the standard deviation of the process ν_t, the parameter κ would be close to one. Small values of the parameter κ indicate that the model overestimates and large values of the parameter suggest that the model underestimates the true risk.

The empirical calibration curve, i.e., the P-P plot of the time series (27), is a diagnostic plot for checking whether the uniform $U[0,1]$ distribution is fitting the marginal distributions of the process $\{F_{L_t}(N_t)\}_{t=1}^{N}$.

For the calibrated model, the points in the plot should concentrate close to the diagonal of the unit square. Furthermore, the symmetry properties of the plotted distributions can be easily evaluated. In the context of the

284   WOLFGANG HÄRDLE *et al.*

VaR models, the main disadvantage of this plot is its focus on the centre of
the distribution (Wilk and Gnanadesikan, 1968).

The application of a Q-Q plot as a diagnostics tool for calibration is more
subjective since, in this situation, we do not have any fixed reference points
(such as the diagonal of the unit square for the P-P plots). We can assess the
symmetry and distribution of the process, but the calibration itself cannot
be assessed without further specification of the model such as, e. g., (35).
The advantage of Q-Q plots lies in its focus on the deviations in the tails of
the distribution (Wilk and Gnanadesikan, 1968).

The absolute and relative empirical calibration curves investigate only
the appropriateness of the marginal distribution of the process L_t. They
allow to verify the appropriateness of the choice of the risk factors, but do
not allow to draw any conclusion on the temporal dependency structures.
The heterogeneity- and independence properties are investigated by the
means of time plots. The time series of the indicator function of the VaR
exceedances at level α

$$\{I\{\nu_t > VaR_{t-1}(\alpha)\}\}_{t=1}^N \qquad (30)$$

allows to discover clusters that indicate time dependency of the realizations
s^α. Plotting the time series

$$\{\hat{P}_t, \nu_t\}_{t=1}^N, \qquad (31)$$

the inhomogeneity can be discovered visually. χ^2 statistics allow to identify
periods in which the forecasts were not independent.

4. Considerations on the quantitative analysis

The following empirical analysis compares the presented models on simu-
lated portfolios by the means of the diagnostic tools, evaluating the forecast
quality and presented in the previous section. We start with the description
of the data set and the design of the simulations.

4.1. Design of the study.

The data set. The empirical basis of the following study is provided by
the daily discrete returns

$$r_t = \frac{p_{t+1} - p_t}{p_t} \qquad (32)$$

obtained from the following 18 German assets and the German market in-
dex DAX: Allianz, BASF, Bayer, BMW, Commerzbank, DAX, Deutsche
Bank, Degussa-Huels, Dresdner Bank, Hoechst, Hypovereinsbank, Karstadt,
Lufthansa, Linde, MAN, Mannesmann, Münchner Rück, Preussag, RWE.

Financial time series typically contain holes corresponding to the national
holidays. It is common practice of the data providers to replace the missing

values by the values from the previous day. This procedure results, according to the definition in (32), in zero returns. This happens to approximately 5% of the data. Thus, the data set could be described as a mixture of two distributions, for example,

$$0.95 \; N(0, \Sigma) + 0.05 \; \delta_0 \,,$$

where δ_0 is the distribution degenerated in zero. In order to get rid of the possible influence of this kind of model misspecification on the evaluation of the quality of forecasts, these values were removed from the time series.

The time series begins on the January 1, 1997 and ends on June 18, 1999.

SIMULATION OF THE PORTFOLIO. In order to simulate the exposures, w_t, as realistically as possible, we have set limits – similarly to the capacity load of VaR limits – controlling the behaviour of the simulations:

$$w_{t+1}^\top = \frac{w_t^\top}{\sigma_{\Pi_t}} \{600 \; \mathrm{diag}(1/6 + U_1, \cdots, 1/6 + U_n)\} \,,$$

where U_n is a n-dimensional uniform distribution and σ_{Π_t} is the volatility of the portfolio. This way of simulating of the portfolio weights guarantees random changes and, at the same time, incorporates natural bounds on the volatility of the portfolio. Notice that if the portfolio volatility increases, the values of w_t decreases in order to keep the VaR under a specified bound. Simple calculation shows that, if the volatility of the returns does not change, the volatility of this simulated portfolio with varying weights should lie close to 400.

The reasoning behind this approach is based on the fact that the VaR models are, aside to the risk measurement, applied also for risk control. The practical implementations are usually based on some limits on the VaR. It has been observed in practice that these limits exhibit constant capacity over the time within certain bounds; this is the motivation and justification for the above assumption.

Due to the simulated portfolio weights that keep automatically the portfolio losses under control, we further assume that the value of the portfolio stays within the fixed interval

$$N_t \in [-50\,000, 50\,000] \,.$$

This interval is split into disjoint sets A_i:

$$[-50\,000, 50\,000] = \bigcup_{i=1}^{5} A_i \,.$$

In our study, the intervals

$$A_1 = [-50\,000, -500), A_2 = [-500, -320), A_3 = [-320, 320),$$

$$A_4 = [320, 500) \text{ and } A_5 = [500, 50\,000]$$

are used for the evaluation of the probability forecasts, i. e., the probabilities of the portfolio changes falling into these intervals are compared with the true events. The intervals A_1–A_5 were selected with respect to the simulated volatility so that the forecasted probabilities span the interval $(0, 1)$ as regularly as possible. This property is important for the methods for the verification of probability forecasts described in Section 5.2. The relative frequencies of the forecasting probabilities resulting from this choice of intervals A_1–A_5 are plotted in Figure 7.

THE MODELS. As shown in Sections 1 and 2, the predictive portfolio loss distribution
$$\mathcal{L}(L_{t+h} \mid \mathcal{H}_t)$$
using one class of models, here based on the variance-covariance structure, can lead to different results depending on the used mapping. The following empirical analysis evaluates and compares the forecasting quality of models M_1–M_6, defined as follows:

1. The model M_1 uses all 18 risk variables:
$$\mathcal{L}(\mathbf{R}_{t+1} \mid \mathcal{H}_t) = N_{18}(0, \Sigma_t). \tag{33}$$

   The VaR calculation is based on (5).
2. The model M_2 is based on the classical beta factor mapping according to (12). The VaR calculation is based on Formula (14).
3. The extended beta factor model, M_3, uses (11). The VaR calculation is based on (13).
4. The model M_4 is based on the principal components (15), the parameter ξ_k in (16) was fixed at 80%. The VaR calculation is based on (17).
5. The model M_5 is a modification of M_4 such that the weighted returns $(w_{1t}R_{1t}, \cdots, w_{nt}R_{nt})$ instead of $\mathbf{R}_t$ are used for the principal components analysis.
6. The model M_6 uses the simple parametrization, as described in (18). All off-diagonal elements of the correlation matrix are identical, $\rho_{ij} = \rho$ for all $i \neq j$.

The evaluation of the risk is, for all considered models, based on the empirical covariance matrix calculated from the observed returns. After the simplifications provided by the respective models, the risk is evaluated for the simulated portfolios.

5. EMPIRICAL ANALYSIS

The main focus of our investigation is the evaluation of the quality of the forecasts provided by models M_1–M_6.

The VaR forecasts of these models are based on the assumption of a multivariate Gaussian White Noise process as a stochastic model for the returns $\mathbf{R}_t$ of the risk factors.

It is known that this process is not able to capture adequately some common characteristics of financial time series such as, e. g., fat tails, leptokurtosis, changes in volatility.

Since the national (see, for example, section 'Law/Gesetze' at www.bafin.de) as well as international regulatory norms (Basel Committee on Banking Supervision, 2005) practically forbid the applications of exponantially weighted observations which enter in the applications of GARCH or IGARCH models, we did not consider these models in this study (Härdle and Stahl, 2000). Hence, we do not present any analysis on how well does such stochastic model capture the dynamics of the process.

QQ plot of Mahalanobis radii
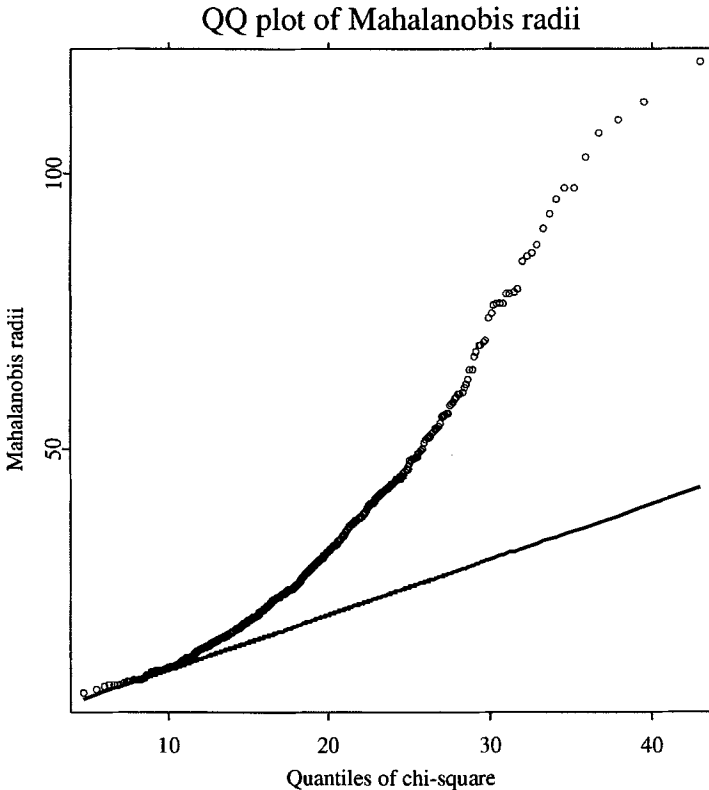
FIGURE 1. Q-Q plot of the squared Mahalanobis radii for the observed returns r_t of the risk factors. The straight line displays the expected values under normality.

The verification of whether the observations $\{\mathbf{r_t}\}_{t=1}^{T}$ stem from the p-dimensional normal $N_p(0, \Sigma)$ distribution, is carried out by means of a Q-Q plot. Assuming that the data are $N(0, \Sigma)$ distributed, we have

$$\mathbf{r}_t^\top \Sigma^{-1} \mathbf{r}_t \sim \chi^2(p).$$

The sorted observed Mahalanobis radii $\mathbf{r}_t^\top \Sigma^{-1} \mathbf{r}_t$ are then compared with the quantiles of the $\chi^2(p)$ distribution in a Q-Q plot. The Q-Q plot displayed in

288   WOLFGANG HÄRDLE *et al.*

Figure 1 deviates largely from the reference line. We conclude that the data hardly come from a multivariate normal distribution. It is now of crucial importance whether this fact influences the quality of the forecasts or not.
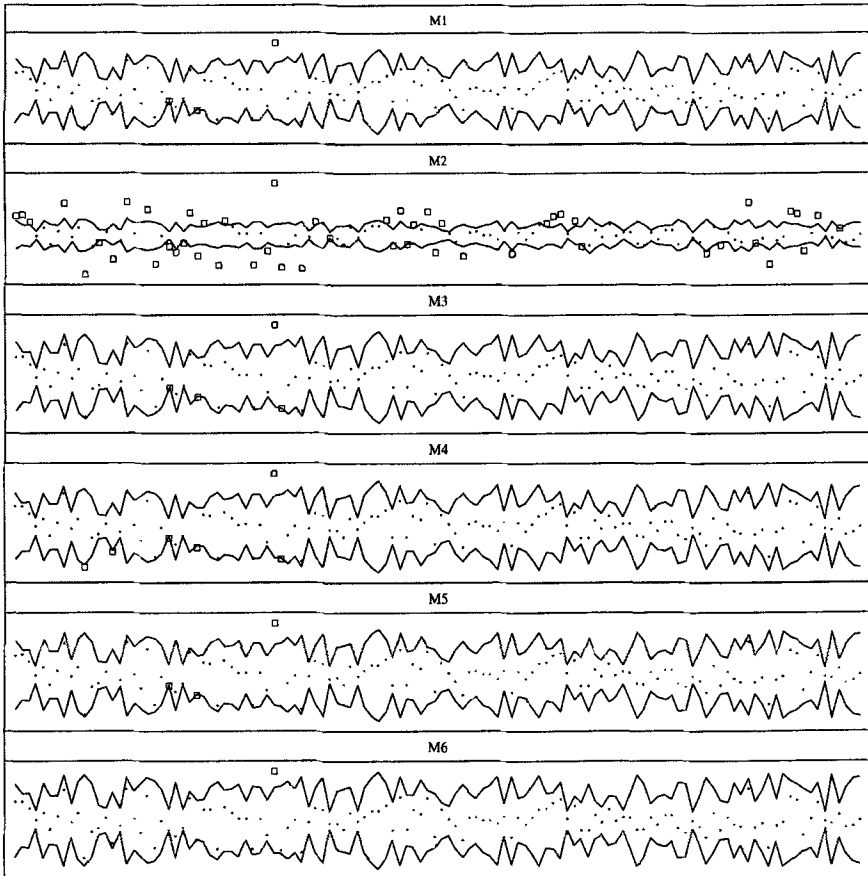
FIGURE 2. Time series of the changes of the value of the portfolio and the VaR forecasts for 249 trading days. The VaR forecasts at level 99% and 1% are plotted as lines. The value changes of the portfolio are plotted as dot if they lie between the two lines, the values of the portfolio falling outside the predicted VaR region are denoted by squares. The plots corresponding to models M_1–M_6 are displayed from top to bottom.

Figure 2 shows the time series

$$\{P_t, \nu_t\}_{i=1}^{T}$$

of the VaR forecasts at level $\alpha = 99\%$ and $\alpha = 1\%$ and the value changes of the portfolio for $T = 249$ trading days for models M_1–M_6. Here, P_t denotes the VaR predictions and ν_t denotes the observed changes of the portfolio value. The exceedances, i. e., the value changes falling outside the predicted

VaR bounds, are marked by squares. With the probability of exceedance set to 2%, we should on average observe 4.98 exceedances.

The forecasting quality of the model M_2 with more than 50 exceedances is obviously insufficient. The VaR bounds for models M_1, M_3, M_4, and M_5 are very similar and the number of exceedances varies between 3 and 6, reasonably close to the expected value 4.98. In our simulation, model M_6 seems to overestimate the risk, leading to only one exceedance.

The daily practice of VaR modelling has shown that exploratory diagnostic tools for assessing the quality of the forecasts of the VaR models are sufficient both for sub and full portfolios. Hence, in the following, we will omit the formal statistical inference presented in Davé and Stahl (1998). We will explore the qualities of the models by means of graphical and descriptive statistical tools.

	M_1	M_2	M_3	M_4	M_5	M_6
κ	1.07	2.32	1.06	1.13	1.07	1.07
Brier score	0.15	0.21	0.15	0.16	0.15	0.16
Skill score	0.35	0.15	0.35	0.34	0.35	0.35

TABLE 1. Parameter κ, Brier score and Skill score for the evaluation of the quality of models M_1–M_6. A good model should exhibit κ close to 1, small value of Brier score, and large value of the Skill score.
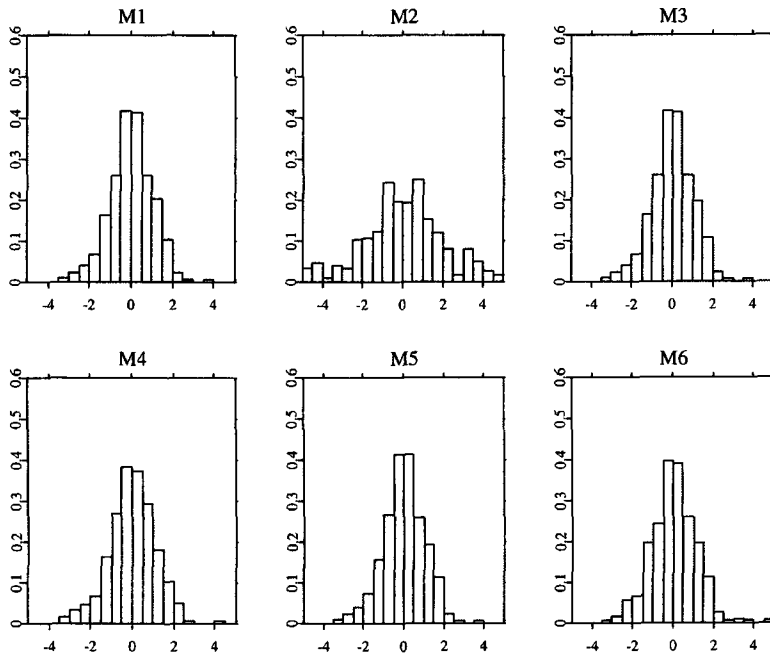

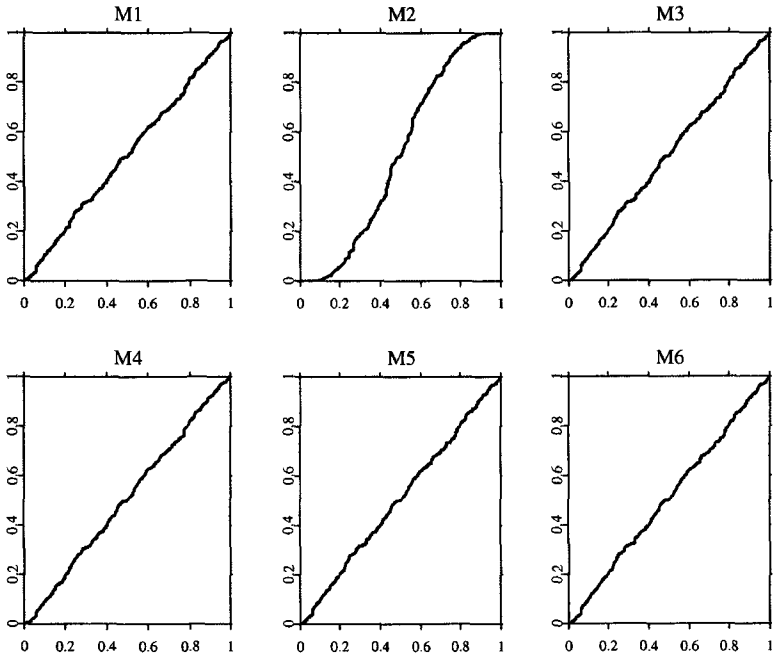FIGURE 3. Histograms of the variable s_t^α, defined in (28), for models M_1–M_6.

FIGURE 4. P-P plots for the variable s_t^α, defined in (28), for models M_1–M_6.
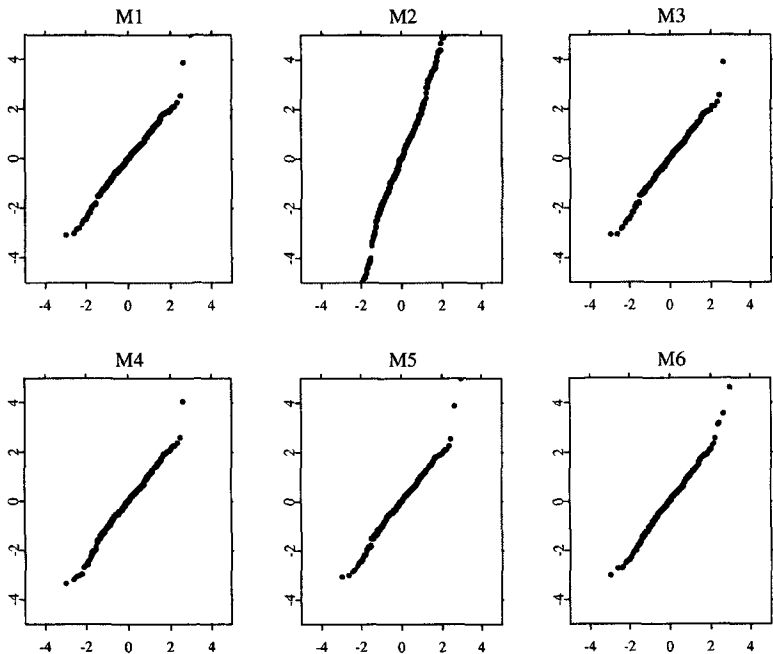

FIGURE 5. Q-Q plots for the variable s_t^α, defined in (28), for models M_1–M_6.

The graphical analysis of the forecast quality of the VaR models can be based, apart of the timeplot in Figure 2, on various plots of the empirical calibration curves (Dawid, 1984). One approach is based on the standardization of the value, L_t, of the portfolio at time t. Since the random variable

$$S_t^\alpha = \frac{L_t}{VaR_{t-1}} \tag{34}$$

has a normal distribution, we obtain that asymptotically

$$\mathcal{L}(L_t / \widehat{VaR}_{t-1}) \approx N(0, 2.33^{-2}), \tag{35}$$

since $\widehat{VaR}_{t-1}$ is in the Gaussian model defined as $2.33 \, \hat\sigma_{t-1}$. Note also that the inequality $S_t^\alpha < -1$ characterizes the exceedances of the VaR forecast that can be easily identified in the corresponding timeplots. The validity of the VaR model can be verified by checking whether $\{F_{L_t}(N_t)\}_{t=1}^T$ and $\{S_t^\alpha\}_{t=1}^T$ are White Noise processes.

In Figures 3–5, we plot the histograms, the P-P plots and the Q-Q plots for the variables s_t^α defined in (28). Note, that in the above Gaussian context, $2.33 \, s_t^\alpha$ is equal to the above defined S_t^α. All histograms in Figure 3, apart of model M_2, look similar to the standard normal distribution. The histograms for the remaining models, M_1 and M_3–M_6 look almost identical and we can say that all mappings give comparable results. The same conclusion can be derived from the P-P and Q-Q plots in Figures 4 and 5. Again, model M_2 gives much worse results than the remaining models which look again very similar and seem to satisfy our assumptions.

The parameter κ, defined in (29), is tabulated for the six models in Table 1. All models tend to underestimate the risk. Again, the model M_2 gives unacceptable results. The level of the risk underestimation for the remaining models is much lower and close to one.

5.1. EXCEEDANCES. The time series of the indicator function of the VaR exceedances (30) at level $\alpha = 80\%$ is plotted in Figure 6. Theoretically, the number of exceedances should lie close to 20%. Visual inspection of such time plots should point out potential time inhomogeneities. Figure 6 is complemented by Table 2 that lists the percentages of the exceedances for each model and for each quarter of the year separately. Most of the time, the percentages are moving rather close ($\pm 6\%$) to the expected 20%. The two exceptions are model M_2 and first quarter of 1999. The bad results for model M_2 are consistent with the results of the previous analyses. The bad behaviour of all models in the beginning of 1999 indicates some model heterogeneities or volatility changes in this period.

5.2. VERIFICATION OF PROBABILITY FORECASTS. The probability forecasts, analyzed in this section, relate to the probabilities of the portfolio changes falling into intervals A_1–A_5, described in Section 4.1. From the
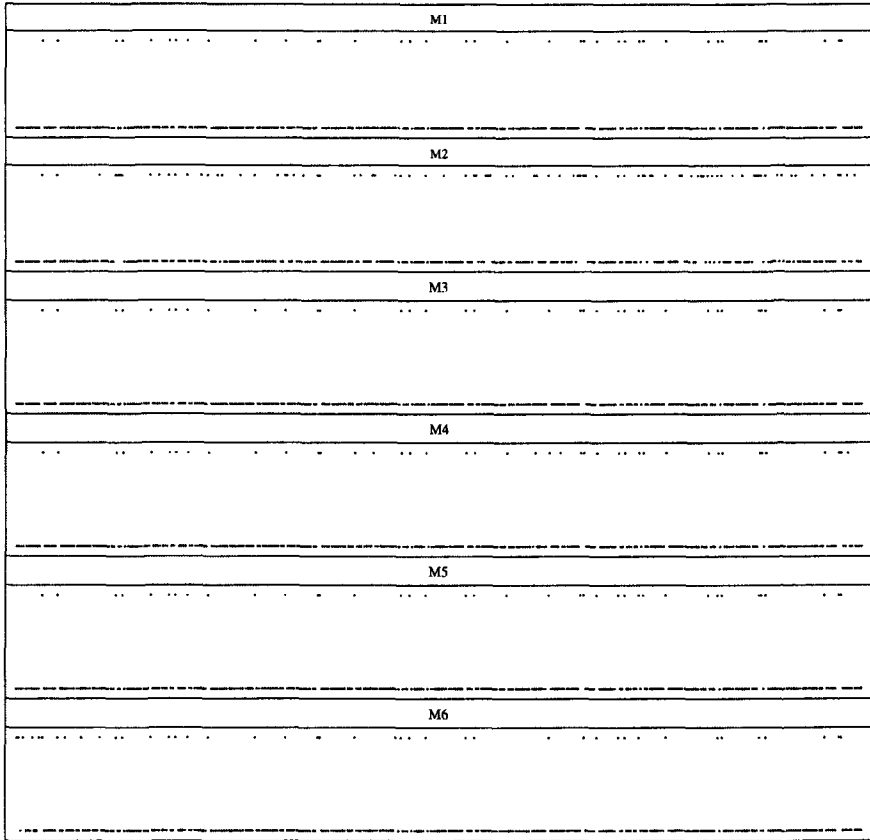
FIGURE 6. The time series of the VaR exceedances at level $\alpha = 0.8$ for models M_1–M_6.

simulation, we have both the predicted probability that the change of the portfolio falls into one of these intervals and we also know the true realization.

The Brier score and Skill score for the models M_1–M_6 are given in the second and third row of Table 1. From this point of view, we can say that the best models are the full model M_1 and models M_3 and M_5. Models M_4 and M_6 are just a bit worse. Model M_2 shows, once again, worst behaviour.

The relative frequencies of the forecasting probabilities are plotted in the first and in the third row of plots in Figure 7. The intervals A_1–A_5 were selected so that, in this simulations, the forecasting probabilities cover the interval $(0,1)$ as uniformly as possible. We observe that only model M_2 behaves differently. In comparison to other models, its forecasting probabilities are much larger. Clearly, this is caused by its underestimation of the risk, observed already in the previous analyses, see Figure 2.

The second and the fourth row of plots in Figure 7 show calibration plots for all six models, i. e., each plot displays the forecasting probabilities on the x-axis and the corresponding relative frequencies of the success on the

	M_1	M_2	M_3	M_4	M_5	M_6
Q1 1998	0.14	0.22	0.14	0.14	0.14	0.14
	(1.01)	(0.40)	(1.01)	(1.01)	(1.01)	(1.01)
Q2 1998	0.17	0.37	0.17	0.17	0.17	0.17
	(0.52)	(2.97)	(0.52)	(0.52)	(0.52)	(0.52)
Q3 1998	0.26	0.52	0.24	0.26	0.26	0.26
	(1.05)	(5.73)	(0.78)	(1.05)	(1.05)	(1.05)
Q4 1998	0.14	0.44	0.14	0.14	0.14	0.14
	(1.01)	(4.34)	(1.01)	(1.01)	(1.01)	(1.01)
Q1 1999	0.39	0.70	0.39	0.39	0.39	0.39
	(3.21)	(8.85)	(3.21)	(3.21)	(3.21)	(3.21)
Q2 1999	0.17	0.63	0.17	0.17	0.17	0.17
	(0.43)	(7.00)	(0.40)	(0.43)	(0.43)	(0.43)
χ^2	13.9	188.0	13.4	13.9	13.9	13.9
p-value	0.020	0.000	0.026	0.020	0.020	0.020

TABLE 2. Relative frequencies of the exceedances of the VaR at level $\alpha = 80\%$ for each quart in our data set for models M_1–M_6. The contributions to the χ^2 statistic are given in the parentheses. The last row gives the p-values for the test of the hypothesis $H_0 : p = 0.2$ against the alternative $H_1 : p \neq 0.2$.

y-axis. Clearly, a perfectly calibrated model should lie very close to the diagonal of the unit square. In this case, all models (apart of M_2) provide well calibrated forecasts. Model M_2 underestimates the forecasting probabilities.

In Figure 8, we show the discrimination plots. One curve in each plot is the relative frequency of the forecasting probabilities conditioned on a success, while the other one conditiones on a failure. For good forecasts, one would like to predict success with high probability if it really occures and with low probability if it does not occur. Hence, for a good model, the two curves should be very far from each other. Again, we observe that the behaviour of model M_2 is worse than the behaviour of the other models.

5.3. CONCLUSION. The results of the empirical analyses suggest that all mappings, apart of M_2, lead to results comparable with the full model M_1. The model M_5 using the weighted principal component analysis gives, as expected, slightly better results that model M_4 that uses principal components method without taking the weights into consideration. Surprisingly, the model M_6 seems to give slightly better results than both models based on the principal components.

Our conclusion is that model M_2 should not be used in practice. The remaining models give comparable results. The comparison of the models was based on the indices given in Tables 1 and 2 whereas the graphical methods (Figures 1–8) help to understand why one method is better than the other one.
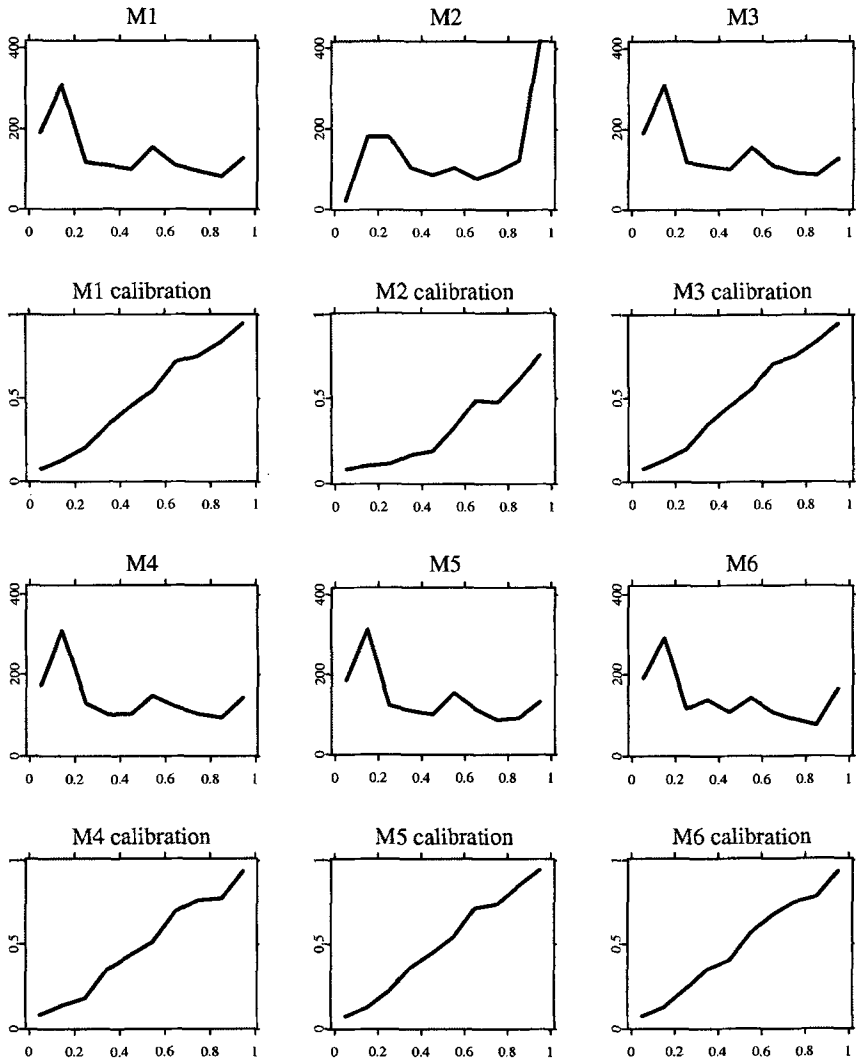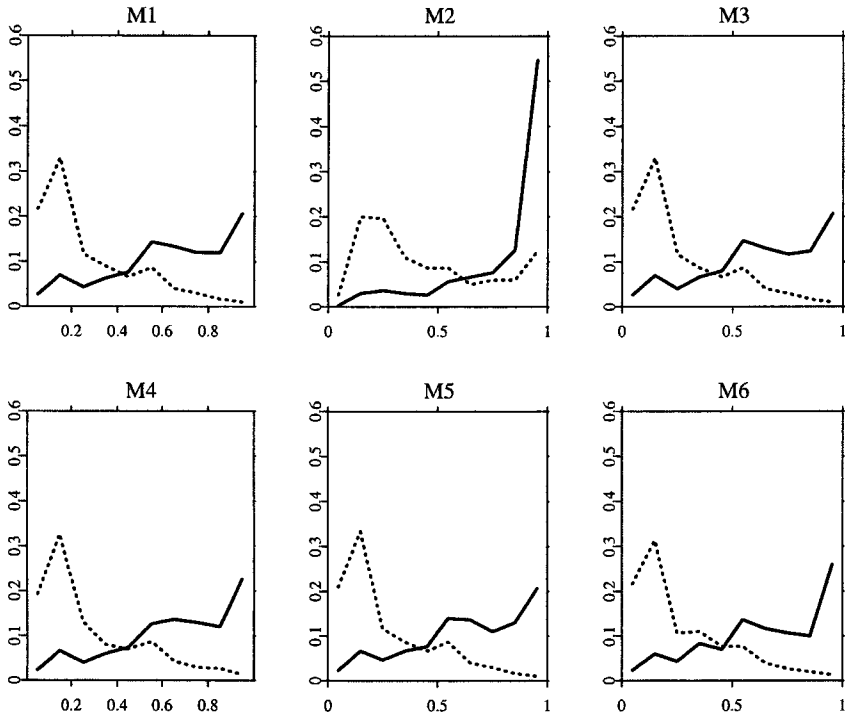
FIGURE 7. The relative frequencies of the forecast probabilities and the discrete calibration curves for models M_1–M_6.

FIGURE 8. Discrimination curves for models M_1–M_6.

REFERENCES

BASEL COMMITTEE ON BANKING SUPERVISION (2005). International convergence of capital measurement and capital standards, a revised framework (November 2005 Revision). Bank of International Settlements, Basel.

CHATFIELD, C. (1995). Model uncertainty, data mining and statistical inference. *Journal of the Royal Statistical Society, Series A* **158** 419–466.

CHATFIELD, C. (1996). Model uncertainty and forecast accuracy. *Journal of Forecasting* **15** 495–466.

CHRISTENSEN, R. (1991). *Linear Models for Multivariate Time Series and Spatial Data*. Springer, Heidelberg.

CUTHBERTSON, K. (1996). *Quantitative Financial Economics*. Wiley, New York.

DAVÉ, R. D., STAHL, G. (1998). On the accuracy of VaR estimates based on the variance-covariance approach. In *Risk Measurement, Econometrics and Neural Networks* (G. Bol, G. Nakhaeizadeh, K. Vollmer, eds.), 189–232. Physica, Heidelberg.

DAWID, A. P. (1984). The prequential approach. *Journal of the Royal Statistical Society, Series A* **147** 278–292.

DAWID, A. P. (1986). Probability Forecasting. In *Encyclopedia of Statistical Sciences* (S. Kotz, N. L. Johnson, C. B. Read, eds.), vol. 7, pp. 210-218. Wiley, New York.

DAWID, A. P. (1997). Prequential Analysis. In *Encyclopedia of Statistical Sciences Update Volume 1* (S. Kotz, C. B. Read, D. L. Banks, eds.), pp. 464-470. Wiley, New York.

DIEBOLD, F. X., GUNTHER, T. A., TAY A. S. (1998). Evaluating density forecasts with applications to financial risk management. *International Economic Review* **39** 863–883.

ELTON, E. J., GRUBER, M. J. (1995). *Modern Portfolio Theory and Investment Analysis.* 5th ed., Wiley, New York.

GILBERT, G. F. (1884). Finley's tornado predictions. *American Meteorological Journal* **1** 166–172.

GOURIÉROUX, C. (1997). *ARCH Models and Financial Applications.* Springer, New York.

HÄRDLE, W., STAHL, G. (2000). Backtesting beyond VaR. In *Measuring Risk in Complex Stochastic Systems, Lecture Notes in Statistics 147* (J. Franke, W. Härdle, G. Stahl, eds.), 119–130. Springer, Heidelberg.

HÄRDLE, W., SIMAR, L. (2003). *Applied Multivariate Statistical Analysis.* Springer, Berlin.

HENDRY, D. F. (1995). *Dynamic Econometrics.* Oxford University Press, Oxford.

HSIAO, C. (1986). *Analysis of Panel Data.* Cambridge University Press, Cambridge.

HUSCHENS, S. (1998). Messung des besonderen Kursrisikos durch Varianzzerlegung. *Kredit und Kapital* **4** 567–591.

JOLLIFFE, I. T. (1986). *Principal Component Analysis.* Springer, New York.

JORION, P. (1997). *Value At Risk.* IRWIN, Chicago.

J. P. MORGAN/REUTERS (1996). RiskMetrics – technical document, 4th ed. Morgan Guaranty Trust Company, New York.

KREININ, A., MERKOULOVITCH, L., ROSEN, D., ZERBS, M. (1998). Principal component analysis in quasi Monte Carlo simulation. *ALGO Research* **1** 21–29.

LITTERMAN, R., WINKELMANN, K. (1998). *Estimating Covariance Matrices.* Risk Management Series, Goldman Sachs.

MURPHY, A. H. (1997). Forecast verfication. In *Economic Value of Weather and Climate Forecasts* (R. W. Katz, A. H. Murphy, eds.), Cambridge University Press, Cambridge.

MURPHY, A. H., WINKLER, R. L. (1992). Diagnostic verification of probability forecasts. *International Journal of Forecasting* **7** 435–455.

PEIRCE, C. S. (1884). The numerical measure of the success of predictions. *Science* **4** 453–454.

REIMERS, M., ZERBS, M. (1998). Dimension reduction by asset blocks. *ALGO Research* **1** 43–55.

RIDDER, T. (1998). Basics of statistical VaR-estimation. In *Risk Measurement, Econometrics and Neural Networks* (G. Bol, G. Nakhaeizadeh, K. Vollmer, eds.), 161–188. Physica, Heidelberg.

RIDDER, T., STAHL, G. (2000). Flexibles oder starres Cash-Flow Mapping?. In *Handbuch Risikomanagement* (L. Johanning, B. Rudolph, Hrsg.), 269–288. Uhlenbruch, Bad Soden.

STAHL, G., TRABER, U. (2000). Backtesting in Action. In *Kreditrisikomanagement* (A. Oehler, Hrsg.), Schäffer-Poeschl, Stuttgart.

STUDER, G. (1997). Maximum Loss for Measurement of Market Risk. PhD Thesis, ETH Zürich.

WILK, M. B., GNANADESIKAN, R. (1968). Probability plotting for the analysis of data. *Biometrika* **55** 1-17.

Wolfgang Härdle
C.A.S.E.
Wirtschaftswiss. Fakultät
Institut für Statistik und Ökonometrie
Humboldt-Universität zu Berlin
Spandauer Str. 1
10178 Berlin
Email: haerdle@wiwi.hu-berlin.de

Zdeněk Hlávka
Charles University in Prague
Faculty of Mathematics and Physics
Department of Statistics
Sokolovská 83
18675 Prague
Email: hlavka@wiwi.hu-berlin.de

Gerhard Stahl
BaFin, Bundesanstalt für
Finanzdienstleistungsaufsicht
Graurheindorfer Str. 108
53117 Bonn

Semiparametric Diffusion Estimation and Application to a Stock Market Index

Wolfgang Härdle[1], Torsten Kleinow[1], Alexander Korostelev[2],
Camille Logeay[3], Eckhard Platen[4]

[1] Humboldt-Universität zu Berlin, Institut für Statistik und Ökonometrie, Spandauer
Strasse 1, 10178 Berlin, Germany, E-mail: kleinow@wiwi.hu-berlin.de
[2] Wayne State University, Detroit, USA
[3] Deutsches Institut für Wirtschaftsforschung, Berlin, Germany
[4] University of Technology Sydney, School of Finance and Economics, PO Box 123
Broadway, Sydney, NSW 2007, Australia

This version: April 26, 2001

Abstract

The analysis of diffusion processes in financial models is crucially dependent
on the form of the drift and diffusion coefficient functions. A methodology is
proposed for estimating and testing coefficient functions for ergodic diffusions
that are not directly observable. It is based on semiparametric and nonpara-
metric estimates. The testing is performed via the wild bootstrap resampling
technique. The method is illustrated on S&P 500 index data.

JEL classification: C51, C52, G22

Keywords: Diffusion, Identification, Continuous-time financial models, Semi-
parametric methods, Kernel smoothing, Bootstrap

1 Introduction

The analysis of time series and diffusion models with stochastic variance and
covariance has been very intense in the last decade. In particular, the research on
stochastic volatility models in finance has been driven in several directions, see
Ghysels, Harvey & Renault (1996) and Frey (1997) for a survey. One direction
is related to modern continuous time finance, see Genon-Catalot, Jeantheau
& Larédo (2000), where one has realized the necessity to relate empirically
observed volatility or implied volatility data to theoretically derived stochastic
volatility models. Another direction focuses on time series models for asset
price and volatility which provide a well established framework for testing and
modeling nonlinear variance structures in finance, see Engle & Bollerslev (1986).

As shown by Nelson (1990) and Duan (1997), standard time series models,
such as ARCH, GARCH etc., converge for vanishing time step size between
observations towards corresponding diffusion models. These diffusion models
are fully characterized by their drift and diffusion coefficient functions. The
particular choice of the time discretization is not essential for a discrete time
approximation of a diffusion process as long as the maximum time step size
is small enough. Since real data are very frequently observed it is natural to

interpret these as values of a discretely observed diffusion process or those of a corresponding discrete time approximation of a diffusion. The main statistical task in such a diffusion approach is the identification of the underlying drift and diffusion coefficient functions.

To handle this task in standard situations, a well developed statistical theory is now available. Parameter estimation methods for discretely observed stationary diffusion processes have been derived, for instance, by Bibby & Sørensen (1995), Hansen & Scheinkman (1995), Ait-Sahalia (1996), Gallant & Tauchen (1996) and Kessler & Soerensen (1999). Hansen, Scheinkman & Touzi (1998) consider a nonparametric method based on the spectral decomposition of the conditional expectation operator to identify the drift and diffusion coefficients. Jiang & Knight (1997) propose nonparametric estimators for the drift and diffusion coefficient, that are based on the approximation of the local time process and the estimation of the marginal density of the diffusion. A crucial assumptions for most of the developed estimation techniques is that a stationary diffusion is directly observed. Unfortunately, in reality this assumption is often not fulfilled. For instance, Genon-Catalot et al. (2000) consider a stochastic volatility model where the observed stock price depends on a hidden volatility process, which is itself a stationary diffusion.

We study here a case where a non-stationary diffusion process, an index, is observed. To be able to apply methods that rely on ergodicity we express the observed process as the product of an ergodic process and a smooth function of time. This smooth function is interpreted as average growth of the index. Due to the unknown impact of the average growth on the observed data, the ergodic part of our model is not directly observable. The proposed methodology combines recently developed nonparametric and parametric methods in order to estimate and probe the drift and diffusion coefficients of the ergodic process.

To illustrate our methodology we concentrate here on the empirical analysis of a particular stock market index, the S&P 500. The statistical analysis of stock prices, exchange rates etc. is similar but not in the focus of this paper. We concentrate here on the case where an index is modeled by a scalar diffusion process.

The framework of Platen (2000) fully characterizes a financial market by the specification of the different denominations of the, so called, best benchmark portfolio. The stock index and the index benchmarked stock prices can be interpreted as denominations of the best benchmark portfolio. As a consequence, exchange prices are ratios of corresponding denominations of the best benchmark portfolio. Furthermore, this portfolio represents the optimal growth portfolio, see Karatzas & Shreve (1998). A well diversified market index, as the S&P 500, comes close to the optimal growth portfolio. For this reason, the inference for the index is also the first step in the statistical analysis of an exchange rate or stock price.

We assume that an appropriately normalized index process $X = \{X(t), t \geq 0\}$ can be interpreted as an ergodic process. Based on this assumption we focus on the inference of this normalized process X instead of the index $S = \{S(t), t \geq 0\}$ itself. This allows us to direct our attention towards the identification of the drift and diffusion coefficient functions of an ergodic diffusion. In Figure 1 we plot the S&P 500 index S with daily data from 1977 to 1997 together with an average index $\bar{S} = \{\bar{S}(t), t \geq 0\}$. Such an average index $\bar{S}$ can be obtained in different ways. For instance, it could be exogenously given by a

function of economic and financial quantities, i.e. inflation rate, growth rate of the domestic product, interest rate, etc.. It could also be derived by a kernel smoothing procedure, with an appropriate bandwidth or filter length h. This is the choice which we will study in this paper.

We construct the normalized index X by dividing the original index S by the above described average index $\bar{S}$, that is

$$X(t) = \frac{S(t)}{\bar{S}(t)} \tag{1.1}$$

for $t \geq 0$. The resulting normalized index X, derived via a kernel smoother, is shown in Figure 1. Its path resembles that of a stationary diffusion process. Note in the middle of our plot the sudden decline caused by the 1987 crash, which we do not remove from our sample.
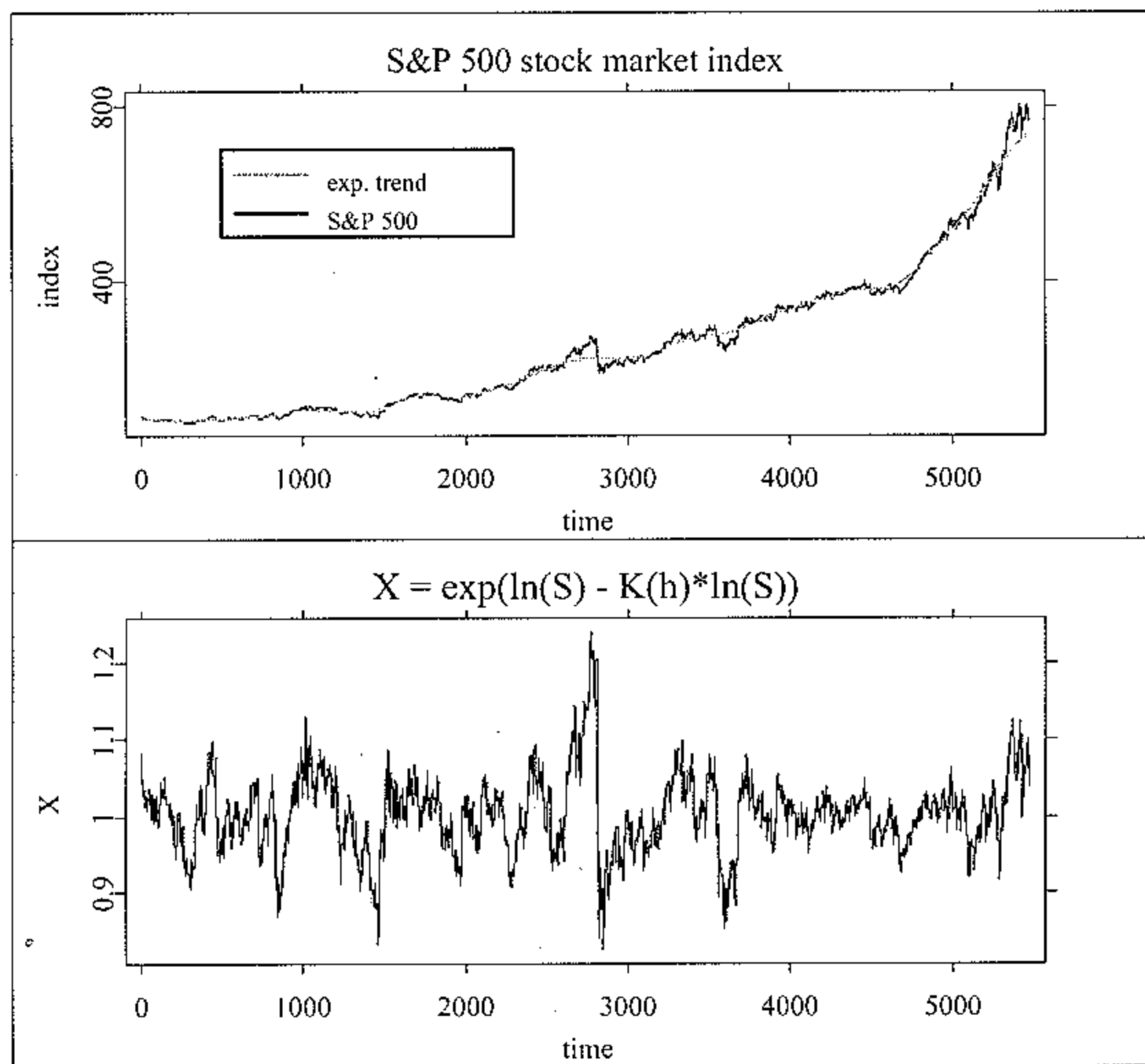

Figure 1: S&P 500 index S, average index $\bar{S}$ and normalized index X

We assume for the value of the index $S(t)$ at time t a representation of the form

$$S(t) = S(0)Z(t)\exp\left\{\int_0^t \eta(s)ds\right\} \tag{1.2}$$

for $t \geq 0$. Here $\eta(t)$ is interpreted as the deterministic, time dependent growth rate of the index at time t. Furthermore $Z(t)$ denotes the value of a positive ergodic diffusion process Z at time t, that means, Z solves the Itô stochastic differential equation (SDE)

$$dZ(t) = m\{Z(t)\}dt + \sigma\{Z(t)\}dW(t) \qquad (1.3)$$

for $t \geq 0$. Here $W = \{W(t), t \geq 0\}$ denotes a standard Wiener process and $m\{.\}$ and $\sigma\{.\}$ are the drift and diffusion coefficient functions. Due to the factor $S(0)$ we assume that Z is stable about 1, which models a mean reverting behavior. On the other hand, $Z(t)$ has to be positive for all $t \geq 0$.

To make our parametric model specific we may choose for Z a square root process, that is positive and stable about an equilibrium reference level. The square root process is also known as the Cox-Ingersoll-Ross (CIR) process, see Cox, Ingersoll & Ross (1985). The functional form (1.2) that models the index is a special case of the minimal market model (MMM) proposed in Platen (2000).

Another parametric model arises if we choose

$$Z(t) = \exp\{U(t)\} \qquad (1.4)$$

with an Ornstein-Uhlenbeck process $U = \{U(t), t \geq 0\}$. This leads us to the exponential of an Ornstein-Uhlenbeck process as index model. Such a model has been used, for instance, in Föllmer & Schweizer (1993), Platen & Rebolledo (1996) and Fleming & Sheu (1999).

To compute the average index $\bar{S}$ in (1.1), we apply a kernel smoother to the logarithm of S and then calculate $\bar{S}$ as the exponential of this smoothed process. This removes the average deterministic growth in (1.2). For the analysis of the resulting normalized index we have to take into account that the residuals $\ln S - \ln \bar{S}$ are corrupted by the smoother. This will be shown in detail later on. It means, that the normalized index X is not a diffusion and in particular it does not equal the diffusion Z. For this reason we cannot directly apply estimation methods for discretely observed diffusions. From the statistical point of view we are faced with a nonparametric regression model with error terms that are not independent and identically distributed but are the discrete observations of a diffusion process. The analysis of these error terms and the clarification of their relationship to Z is a main task in this paper.

We remark, that the index process S is itself a diffusion. When Z is specified according to (1.4) with an Ornstein-Uhlenbeck process U, Itô's formula yields the representation

$$dS(t) = \{\eta(t) - \beta\}S(t)dt + \gamma \exp\left\{\int_0^t \eta(s)ds\right\} dW(t) \qquad (1.5)$$

for $t \geq 0$. The parameters β, γ and η cannot be easily separated in this representation. For this reason, we develop a statistical methodology for models that are based on the representation (1.2).

In Section 2.1 we introduce the parametric model for Z. The kernel smoothing and the computation of X is described in Section 2.2. The choice of the kernel and bandwidth and its influence on the average index is discussed in Section 2.3 together with the corresponding parameter estimation methods. Section 2.4 introduces discrete time approximations of diffusion processes and Section

2.5 describes the nonparametric estimation of drift and diffusion coefficient functions. In Section 2.6, a parametric model is tested versus a purely nonparametric alternative. This test is carried out by the bootstrap technique described in Appendix A.3. In Section 3 and 4 we apply the introduced methodology to S&P 500 data and also in a simulation study.

We remark that the proposed methodology applies directly to situations, where normalized data can be modeled by an ergodic diffusion process. Emphasis is here given to the case of an Ornstein-Uhlenbeck process, and results on the influence of the kernel smoother are included for this case.

2 Statistical Methodology for a Normalized Diffusion

2.1 Parametric Models

As discussed in the introduction, one can, in principle, use various parametric ergodic diffusion models. Let us mention two examples. Both of them have mean reverting drift coefficients. In the case where the squared diffusion coefficient has the form

$$\sigma^2(z) = \nu^2 z, \quad z > 0, \tag{2.1}$$

with a positive constant ν, we obtain in (1.2) a square root process Z. Here we assume that Z satisfies the SDE

$$dZ(t) = \{\psi - \varphi Z(t)\}dt + \nu\sqrt{Z(t)}dW(t) \tag{2.2}$$

for $t \geq 0$ and with $\psi > \nu^2/2, \varphi > 0$. Note that a stationary and ergodic solution of (2.2) exists with the expected value $\mu_\infty = E[Z] = \psi/\varphi > 0$. Since the ratio $Z(t)/\mu_\infty$ is again a square root process and any constant term can be absorbed by $S(0)$ in (1.2), one can for simplicity assume that $\mu_\infty = E[Z] = 1$. This choice leads us to the SDE

$$dZ(t) = \varphi\{1 - Z(t)\}dt + \nu\sqrt{Z(t)}dW(t). \tag{2.3}$$

for $t \geq 0$.

We obtain a second example for an ergodic diffusion by defining Z as in (1.4), where U denotes the well-known Ornstein-Uhlenbeck process with

$$dU(t) = -\beta U(t)dt + \gamma dW(t). \tag{2.4}$$

for $t \geq 0$. Since U fluctuates about its reference level 0 and is ergodic, Z as given in (1.4) is an ergodic, positive diffusion process fluctuating about 1.

2.2 Kernel Smoothing

Denote by K_h a smoother with a kernel K and a bandwidth h. The smoothing of any process is denoted by a convolution operator $(*)$. As mentioned before, the normalized index $X(t)$ in (1.1) can be defined by the exponential of the difference of

$$L(t) = \ln\{Z(t)\} \tag{2.5}$$

and its smoother $(K_h * L)(t)$, that is:

$$X(t) = \exp\{ \ln S(t) - (K_h * \ln S)(t) \} = \exp\{ L(t) - (K_h * L)(t) \}. \quad (2.6)$$

Equation (2.6) holds if we neglect the difference between the accumulated deterministic growth rate $\int_0^t \eta(s)ds$ in (1.2) and its smoother, this means:

$$\int_0^t \eta(s)\,ds - \left(K_h * \int_0^{\cdot} \eta(s)\,ds \right)(t) \approx 0.$$

Here we arrive at a delicate point of our study. If we want to remove efficiently the deterministic growth rate in (1.2), then the value h should be chosen relatively small. Indeed, smaller values for h reduce the bias. On the other hand, the smaller the value of h is chosen, the more X is corrupted by $K_h * L$ in (2.6).

The smoother $K_h * L$ is differentiable for differentiable kernels K and thus of bounded variation. Due to the smoothing procedure $K_h * L$ involves also future information about L. Thus X is not a diffusion process. For this reason, we cannot treat $\ln X(t)$ in (2.6) as the logarithm of a square root process or as an Ornstein-Uhlenbeck process. A more detailed analysis of X has therefore to be performed. This is the objective of the next section. However note, in the case when $\tilde{S}$ is obtained exogenously and not by a smoothing procedure, X might still be a diffusion.

2.3   Estimation of Parameters

In this section we assume that the only observations available are those of

$$\ln X(t) = L(t) - (K_h * L)(t) \quad (2.7)$$

in (2.6) and that L is the Ornstein-Uhlenbeck process U given in (2.4). The estimation problem that we now consider is that for the parameters β and γ in (2.4). In principle the value of γ can be restored from the quadratic variation of either $Z(t)$ or $L(t)$. For differentiable kernels K_h in (2.6), the process $(K_h * L)(t)$ is also differentiable. For this reason it holds that

$$\lim_{\Delta t \to 0} \sum_{i=1}^{n} \{ \ln X(i\Delta t) - \ln X(i\Delta t - \Delta t) \}^2 \overset{L^2}{=} \int_0^T d<L>_t \quad (2.8)$$

for $n = T/\Delta t$. Here $d<L>_t$ denotes the differential of the quadratic variation of the process L at time t. Empirical results confirm that the quadratic variation is not sensitive to the choice of h. For more details on that see Table 1. The following formula provides a stable estimate of γ^2 in the form

$$
\begin{aligned}
\hat{\gamma}^2 &= T^{-1} \sum_{i=1}^{n} \{ \ln X(i\Delta t) - \ln X(i\Delta t - \Delta t) \}^2 \\
&\approx T^{-1} \int_0^T d<L>_t .
\end{aligned}
\quad (2.9)
$$

To estimate the speed of adjustment parameter β in (2.4) we could use the well-known form of the stationary variance of the Ornstein-Uhlenbeck process L. Along with (2.9) this would result in a first estimator of β with

$$\hat{\beta}_1 = \hat{\gamma}^2 / (2\,Var[L]) . \quad (2.10)$$

Unfortunately, the substitution of $Var[L]$ by $Var[\ln X]$ makes $\hat{\beta}_1$ strongly dependent on h. Indeed, the variance

$$Var[\ln X] = Var[L - K_h * L] \tag{2.11}$$

increases as h grows, and only for very large values of h we can expect that $Var[\ln X] \approx Var[L]$.

It is not just the variance of the random process $\ln X$ that changes with h. Also its autocorrelation function depends on the bandwidth h. The correlation between the values of the process $\ln X$, distant by a constant time length $\tau > 0$, diminishes as h decreases. For this reason we propose a selection method for h based on the simultaneous estimation of β from the variance and from the autocorrelation function of the process $\ln X$. The idea is simple, if for each value of h there are two different estimates of the same parameter β, then the best choice of h is considered to be that, which brings these estimates as close as possible to each other.

The autocorrelation function $\rho^{(L)}(\tau)$ of the Ornstein-Uhlenbeck process L equals

$$\rho^{(L)}(\tau) = e^{-\beta\tau} \tag{2.12}$$

for $\tau > 0$. Thus, β represents the absolute value of the slope of this function at zero. Hence another estimate of β from the observations of L would be

$$\hat{\beta}_2 = \left| \frac{\partial^+}{\partial\tau}\rho^{(L)}(\tau) \right|\Big|_{\tau=0}, \tag{2.13}$$

where

$$\frac{\partial^+}{\partial\tau}\rho(\tau) = \lim_{s\to 0,\, s>0} \frac{\rho(\tau + s) - \rho(\tau)}{s}$$

for $\tau \geq 0$ denotes the right hand derivative of ρ with respect to τ.

Unfortunately, the estimator in (2.13) is not feasible since L is not observed. In Appendix A.1 we show for the process $\ln X$ that its stationary variance is asymptotically

$$Var[\ln X] = \frac{\gamma^2}{2\beta}\left(1 - \frac{c_K}{\beta h} + O(h^{-2})\right) \quad \text{as} \ \ h \to \infty, \tag{2.14}$$

where the constant c_K depends on the kernel K. Furthermore, we prove in Appendix A.1 for the autocorrelation function $\rho_h^{(\ln X)}(\tau)$ of $\ln X$ the asymptotics

$$
\begin{aligned}
\rho_h^{(\ln X)}(\tau) &= Corr\big[\ln X(\tau); \ln X(0)\big] \\
&= \frac{e^{-\beta\tau} - \frac{c_K}{\beta h} + O(\tau h^{-2})}{1 - \frac{c_K}{\beta h} + O(h^{-2})}, \qquad \tau \geq 0
\end{aligned} \tag{2.15}
$$

as $h \to \infty$ with the same constant c_K as in (2.14). In Appendix A.1 this constant is calculated for the rectangle and the Epanechnikov kernels.

It follows from equation (2.14) that the first-order approximation of the stationary variance of $\ln X$ is

$$Var[\ln X] \approx \left\{1 - \frac{c_K}{\beta h}\right\} Var[L]. \tag{2.16}$$

By (2.15), the slope of the autocorrelation of $\ln X$ at zero is asymptotically

$$\left| \frac{\partial^+}{\partial \tau} \rho_h^{(\ln X)}(\tau) \right|\Big|_{\tau=0} \approx \frac{\beta}{1 - \frac{c_K}{\beta h}} \tag{2.17}$$

as $h \to \infty$. Thus this slope is steeper than that of ρ^L at $\tau = 0$.

The immediate consequence of (2.14) and (2.15) is that the formulas (2.10) and (2.13) for $\hat{\beta}_1$ and $\hat{\beta}_2$, respectively, have to be modified if the process $\ln X$ rather than L is observed. In Appendix A.1 we show that the correct modification is provided by the expressions

$$\hat{\beta}_1(h) = \frac{\hat{\gamma}^2}{2 Var[\ln X]} - \frac{c_K}{h} \tag{2.18}$$

and

$$\hat{\beta}_2(h) = \left| \frac{\partial^+}{\partial \tau} \rho_h^{(\ln X)}(\tau) \right|\Big|_{\tau=0} - \frac{c_K}{h}, \tag{2.19}$$

respectively. Finally, our method for the selection of h is based on the following balance equation

$$\hat{\beta}_1(h) = \hat{\beta}_2(h) \tag{2.20}$$

which equals both estimates.

After h is chosen, we need to restore the process L, which is needed in the remaining nonparametric and parametric analysis. From (2.6), proceeding formally, one arrives at the following iterative formula:

$$
\begin{aligned}
L = \ln X + K_h * L &= \ln X + K_h * (\ln X + K_h * L) \\
&= \ldots \\
&= \ln X + K_h * \ln X + K_h * K_h * \ln X + \ldots. \tag{2.21}
\end{aligned}
$$

The justification for the restoration formula (2.21) comes from the fact that if one neglects the boundary effects, the smoothing operator K_h is a contracting operator in L_2, as shown in Appendix A.1. In the practical application of (2.21), we rely on the fact that the smoother of the original process L is close to the smoother of $L - K_h * L$. In practice, only one or two convolutions are meaningful. After the restoration process is completed, the parameter β can be estimated directly from L by (2.10).

We were able to establish in this paper the above correction terms for the Ornstein-Uhlenbeck process. One could, in principle, estimate parameters also under the assumption that X itself is a square root process or another ergodic diffusion. However, if the average index $\bar{S}$ is calculated via a smoothing procedure, a similar bandwidth selection method has to be developed. At that stage this is left for future research.

2.4 Discrete Time Approximation of a Diffusion

We interpret the process $Z(t)$, which appears in (1.2), as a positive ergodic scalar diffusion process that is the solution of the SDE (1.3). The drift m :

$[0, \infty) \mapsto (-\infty, \infty)$ and the diffusion coefficient $\sigma : [0, \infty) \mapsto [0, \infty)$ in (1.3) are assumed to be sufficiently regular, such that a unique solution of (1.3) exists.

For the existence of an ergodic solution of (1.3) the drift and the diffusion coefficient must satisfy some ergodicity conditions, see for instance Bibby & Sørensen (1995). The most important condition is that the stationary Kolmogorov forward equation

$$\frac{1}{2} \frac{\partial}{\partial z} \left\{ \sigma^2(z) p_0(z) \right\} - m(z) p_0(z) = 0$$

must have a solution $p_0(z)$ which is then up to a constant the stationary probability density.

In particular, the above conditions hold for the exponential of an Ornstein-Uhlenbeck process with a log normal stationary distribution and for the square root process which has a Gamma-distribution as stationary distribution.

Let us assume that the diffusion process Z is observed at discrete times $t_i = i\Delta$, $i = 1, 2, \ldots$, with a time step size $\Delta > 0$. Here we suppose that Δ is small or, more precisely, will tend to zero asymptotically. Under rather weak assumptions, see Kloeden & Platen (1999), on the functions m and σ^2, it can be shown that the Euler approximation

$$Z^\Delta(t) = Z^\Delta(0) + \int_0^t m\{Z^\Delta(t_{i_s})\} ds \; + \int_0^t \sigma\{Z^\Delta(t_{i_s})\} dW(s) \qquad (2.22)$$

with $t_{i_s} = \max\{t_i, \, t_i \leq s\}$, converges in a mean square sense to Z as $\Delta \to 0$, i.e.,

$$\lim_{\Delta \to 0} E \left[\sup_{0 \leq t \leq T} |Z^\Delta(t) - Z(t)|^2 \right] = 0, \quad T > 0. \qquad (2.23)$$

From now on, we assume that a discrete time approximation Z^Δ exists in the form of (2.22), and that the convergence property (2.23) holds. For the purposes of this paper, Δ will always be considered to be small enough so that one can substitute Z by Z^Δ without any major error in our interpretation of the observed data. The concrete choice of Δ does not matter since all the relevant information about the model is contained in the drift m and diffusion coefficient σ. As the estimates of moments of higher order terms in the case of the daily observed S&P 500 show, the step size Δ corresponding to the given daily observations is small enough so that the difference between $Z^\Delta(t_i)$ and $Z(t_i)$ is indeed negligible. Thus we interpret the increments of the observed data as those of the Euler approximation (2.22), that is

$$Z^\Delta(t_{i+1}) - Z^\Delta(t_i) = m\{Z^\Delta(t_i)\}\Delta \; + \sigma\{Z^\Delta(t_i)\}\{W(t_{i+1}) - W(t_i)\} \quad (2.24)$$

for $i = 0, 1, \ldots$. Note that the observations $(Z^\Delta(t_i))$, $i = 0, 1, \ldots$ form a state dependent time series.

2.5 Nonparametric Estimation

For two consecutive observations $Z^\Delta(t_i)$ and $Z^\Delta(t_{i+1})$ of the above time series we define the increment

$$Y_i = Z^\Delta(t_{i+1}) - Z^\Delta(t_i) = m\{Z^\Delta(t_i)\}\Delta + \sigma\{Z^\Delta(t_i)\}\sqrt{\Delta}\varepsilon_i \qquad (2.25)$$

with independent standard Gaussian random variables

$$\varepsilon_i = \frac{W(t_{i+1}) - W(t_i)}{\sqrt{\Delta}} \sim \mathcal{N}(0, 1).$$

It is now possible to use the increments Y_i to estimate the functions $m\{.\}$ and $\sigma^2\{.\}$. In particular, we apply the local linear nonparametric method explained in Fan & Gijbels (1996) and Härdle & Tsybakov (1997). To justify the use of this method, the time series $Z^\Delta(t_i)$ has to be ergodic and to meet some technical conditions, see Härdle & Tsybakov (1997). The square root process and the Ornstein-Uhlenbeck process are examples, where the discrete time approximation (2.25) satisfies these conditions.

In this nonparametric framework we obtain the drift function estimator

$$\hat{m}_{h_1}(z) = \frac{1}{\Delta} \hat{\beta}_0(z) \tag{2.26}$$

with

$$
\begin{aligned}
\hat{\beta}(z) &= \begin{pmatrix} \hat{\beta}_0(z) \\ \hat{\beta}_1(z) \end{pmatrix} \\
&= \operatorname{argmin}_{b_0, b_1} \left(\sum_{i=1}^{n} \left\{ Y_i - b_0 - b_1(Z^\Delta(t_i) - z) \right\}^2 K_{h_1}(z - Z^\Delta(t_i)) \right).
\end{aligned}
$$

The bandwidth $h_1 > 0$ is chosen with respect to the Silvermans rule of thumb, see Härdle (1990). The Gaussian density function is used as the kernel.

We apply for the squared diffusion function $\sigma^2(z)$ a two-step estimation. First we compute from the above drift function estimator (2.26) the values $\hat{m}_{h_1}\{Z^\Delta(t_i)\}$. In the second step we use the squared diffusion function estimator

$$\hat{\sigma}^2_{h_1}(z) = \frac{1}{\Delta} \hat{\delta}_0(z) \tag{2.27}$$

with

$$
\begin{aligned}
\hat{\delta}(z) &= \begin{pmatrix} \hat{\delta}_0(z) \\ \hat{\delta}_1(z) \end{pmatrix} \\
&= \operatorname{argmin} \Bigg(\sum_{i=1}^{n} \left\{ (Y_i - \Delta \hat{m}_{h_1}\{Z^\Delta(t_i)\})^2 \right. \\
&\qquad\qquad \left. -\delta_0 - \delta_1(Z^\Delta(t_i) - z) \right\}^2 K_{h_1}(z - Z^\Delta(t_i)) \Bigg)
\end{aligned}
$$

and bandwidth $h_1 > 0$. For details and the asymptotic properties of these estimators and the construction of corresponding confidence bands we refer to Appendix A.2 and A.3.

The application of the above methodology to the data Z^Δ yields corresponding nonparametric estimates with an estimated squared diffusion coefficient function as well as a drift coefficient function along with corresponding confidence bands as we demonstrate below.

2.6 Testing the Parametric Model

We construct tests to compare the nonparametric estimates of Section 2.5 for $m\{.\}$ and $\sigma^2\{.\}$ to parametric forms, for example the coefficients of an Ornstein-Uhlenbeck process.

To derive the null hypotheses in the case when Z is the exponential of an Ornstein Uhlenbeck process, we apply Itô's formula to $Z(t) = \exp\{U(t)\}$. Here U satisfies (2.4) and one obtains

$$
\begin{aligned}
dZ(t) &= d(\exp\{U(t)\}) \\
&= Z(t)\left\{-\beta \ln Z(t) + \frac{1}{2}\gamma^2\right\}dt + \gamma Z(t)dW(t) \quad (2.28)
\end{aligned}
$$

for $t \geq 0$. The null hypotheses of the tests are therefore

$$
H_0(m) : m(z) = z\left\{-\beta \ln z + \frac{1}{2}\gamma^2\right\}
$$

and

$$
H_0(\sigma^2) : \sigma^2(z) = \gamma^2 z^2,
$$

while the alternative is nonparametric.

We construct confidence bands with the bootstrap method. The idea is to bootstrap the original discrete time series and estimate each time the drift and squared diffusion coefficients nonparametrically as described in Section 2.5. With these estimates one can then construct confidence bands for the two functions.

We choose the bootstrap method because it leads to better coverage probabilities than, for instance, a Gaussian approximation. In Neumann & Kreiss (1998) it was shown for a time series similar to (2.25) that the coverage probability is of order $O(n^{-q})$ for some $q > 0$, where n is the number of observations. A Gaussian approximation, see Hall (1985), leads to a coverage probability of order $O(1/\ln(n))$.

The bootstrap method is described in Appendix A.3. The asymptotic results for the $(1 - \alpha)$ confidence bands $KB(m)$ and $(KB(\sigma^2))$, that is

$$
P\{m(z) \in KB(m)\} \to 1 - \alpha
$$

and

$$
P\{\sigma^2(z) \in KB(\sigma^2)\} \to 1 - \alpha
$$

respectively, are proved in Franke, Kreiss, Mammen & Neumann (1998).

3 Empirical Analysis of the S&P 500

We apply the methods introduced in Section 2 to daily observations of the S&P 500 index from 31.12.1976 to 31.12.1997 (5479 observations). The data are obtained from Thomson Financial Datastream.

For the kernel smoothing of S we choose the Epanechnikov kernel. The constant c_K that appears in the correction terms in (2.18) and (2.19) are known for this particular kernel, see Appendix A.1.

As already mentioned in Section 2.3 the estimates for the parameter γ calculated from formula (2.9) are small relative to 1 and do not change significantly with h. Table 1 shows the estimated values for different values of h. The variance of the process X is also shown in that table. The small variance and the fact that X is stable about 1 justifies to concentrate on the case of a geometric Ornstein-Uhlenbeck process defined by (1.4) and (2.4).

h	200	250	300	350	400
$Var(X)$	0.0018303	0.0023465	0.0029246	0.0035622	0.0042183
$\hat{\gamma}$	0.0090593	0.0090703	0.0090849	0.0090991	0.0091103

Table 1: Estimated values for γ and the estimated variance of X for different bandwidths h.

The next step in our analysis is the choice of h. Due to the long range of observations we apply a flexible bandwidth to the data. This flexible bandwidth was calculated by splitting the data in overlapping subintervals of different lengths and calculating an optimal fixed bandwidth for every subinterval. The bandwidth is chosen to be optimal with respect to the balance equation (2.20). To get a continuous optimal bandwidth function $h_{opt}(t)$ we interpolate the resulting values. The function $t \mapsto h_{opt}(t)$ is shown in Figure 2. The fi-
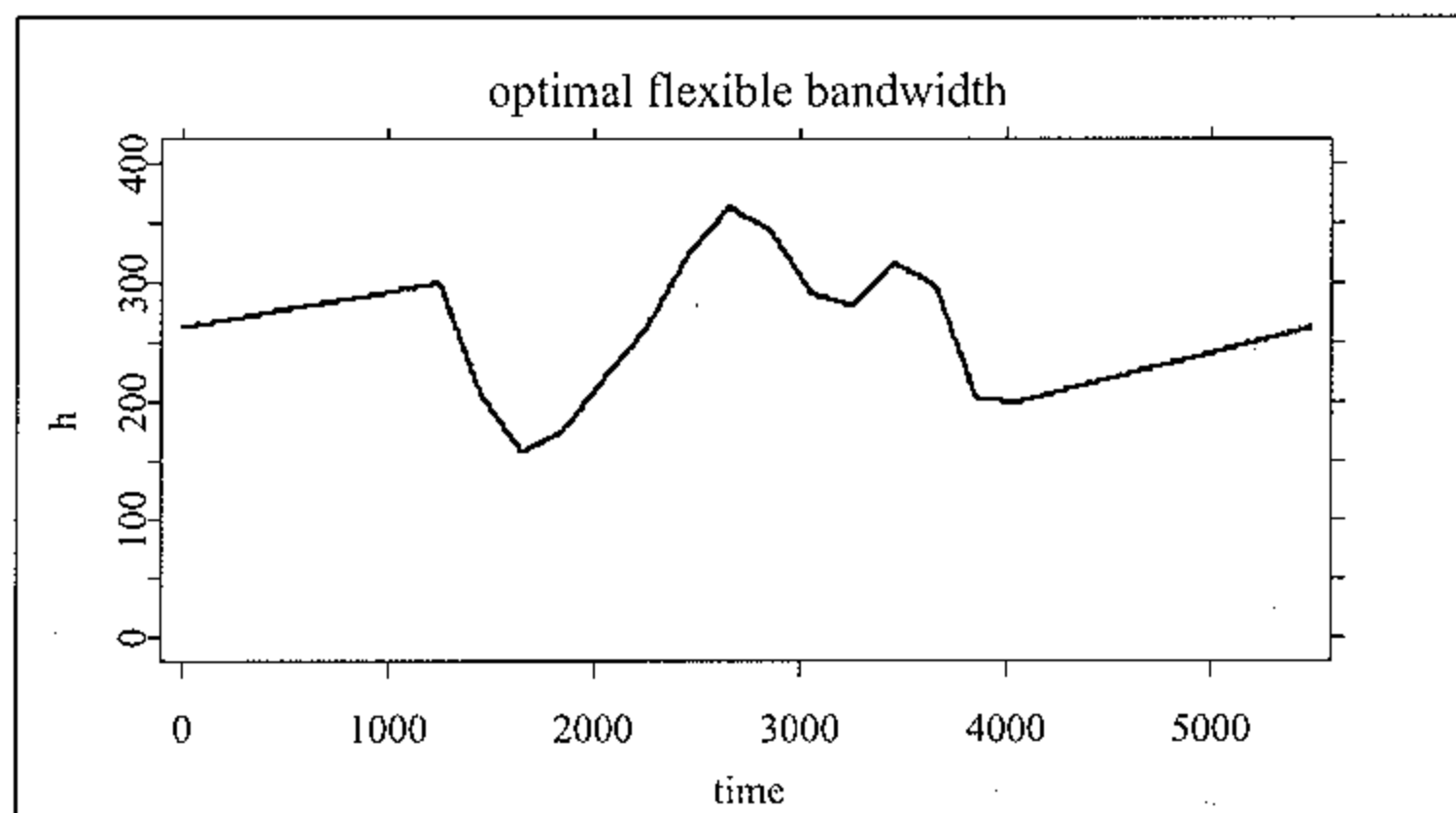

Figure 2: The optimal flexible bandwidth $h_{opt}(t)$.

nal values for β are $\hat{\beta}_1(h_{opt}) = 0.010352$ and $\hat{\beta}_2(h_{opt}) = 0.0089721$ and the ratio is $\hat{\beta}_1(h_{opt})/\hat{\beta}_1(h_{opt}) = 1.1538$. For fixed bandwidths in the range of h_{opt} this ratio is given in Table 2. All these ratios are larger than those for h_{opt} which justifies the use of the flexible bandwidth. The estimated value for γ is $\hat{\gamma}(h_{opt}) = 0.0091033$.

Now we are in the position to restore the path of the process Z and to

h	200	225	250	275	300
$\hat{\beta}_1(h)$	0.017913	0.01569	0.013826	0.012197	0.010776
$\hat{\beta}_2(h)$	0.01328	0.012098	0.011092	0.0098626	0.0086047
$\hat{\beta}_1(h)/\hat{\beta}_2(h)$	1.3489	1.2969	1.2465	1.2367	1.2523

Table 2: Estimated values for β for different fixed bandwidths h.

estimate the parameters. We get the following estimates from the restored path

$$\hat{\beta}_1 = 0.01003, \qquad \hat{\beta}_2 = 0.0093294,$$
$$\hat{\beta}_1/\hat{\beta}_2 = 1.0751, \qquad \hat{\gamma} = 0.0092454.$$

To finish the empirical analysis we apply the test procedure described in Section 2.6. Figure 3 shows the nonparametric estimates of the drift and squared diffusion coefficient of the restored process Z together with the 90% confidence bands. The almost straight lines show the parametric estimates with respect to the estimated values of the restored process Z. The vertical lines enclose the interval where 99% of the observed data reside.
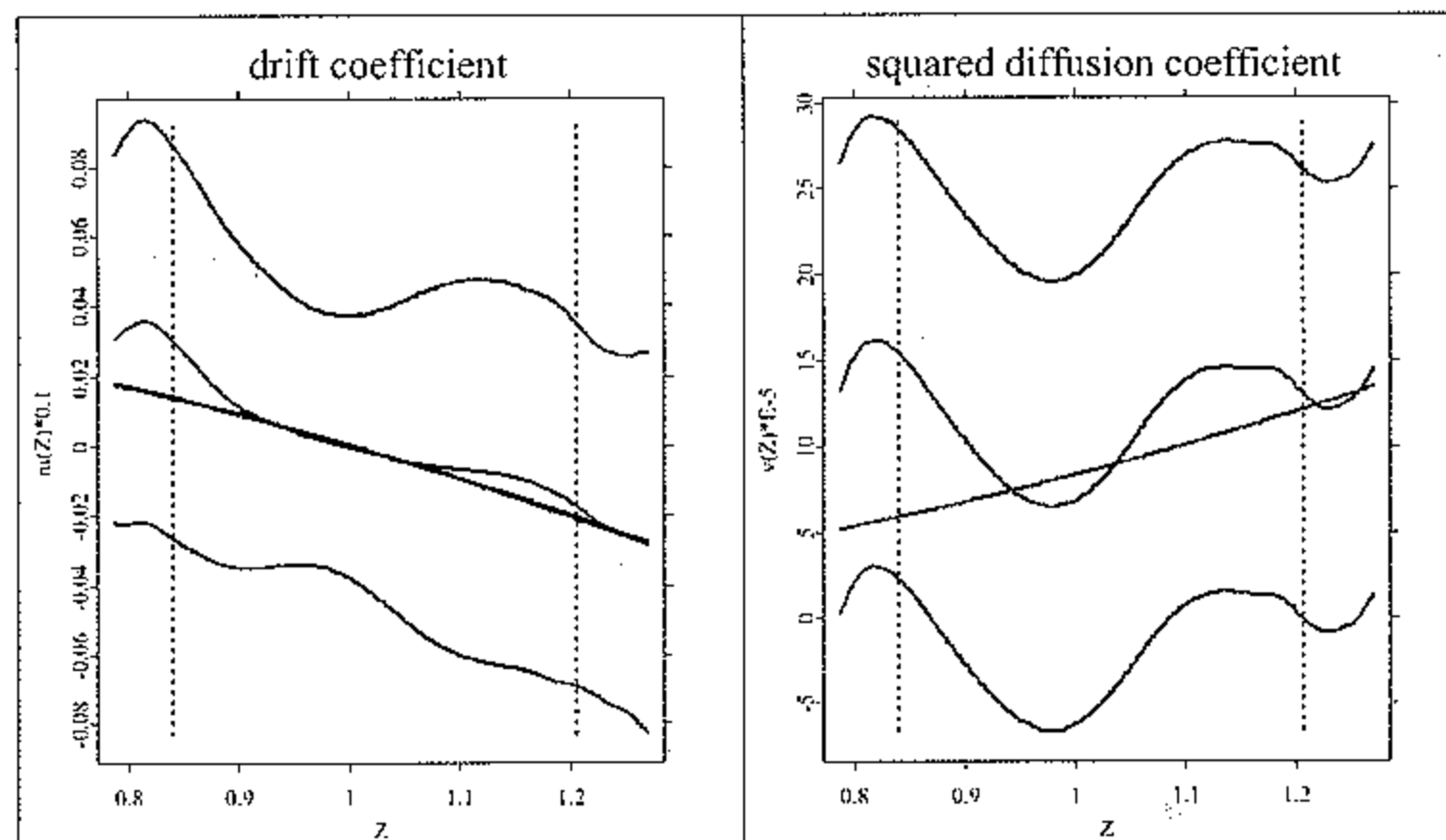

Figure 3: Nonparametric and parametric estimates of the drift $m(.)$ and squared diffusion coefficient $\sigma^2(.)$ with 90% confidence bands.

Both parametric functions are surely inside the confidence bands. Thus the null hypothesis of the geometric Ornstein–Uhlenbeck process cannot be rejected.

4   Simulation Study

We perform now a simulation study by applying the estimation methods introduced in Section 2.3 to simulated trajectories of the Ornstein-Uhlenbeck process U. The drift and diffusion parameters β and γ in (2.4) are estimated directly

from the observations of U as well as from the residual of a kernel smoothing procedure.

It is well known that the transition probability of an Ornstein-Uhlenbeck process is normal with conditional mean

$$E[U_{t+\Delta}|U_t = u] = ue^{-\beta\Delta}$$

and conditional variance

$$Var(U_{t+\Delta}|U_t = u) = \frac{\gamma^2}{-2\beta}\left(e^{-2\beta\Delta} - 1\right).$$

Using this Gaussian transition probability we simulate 100 paths of the process U with time step size $\Delta = 1$. The true parameters are set to $\beta = 0.01$ and $\gamma = 0.01$, which correspond approximately to the empirical estimates for the S&P 500 index in Section 3.

For the analysis of the directly observed process U we apply three estimators for the speed of adjustment parameter β. Besides $\hat{\beta}_1$ and $\hat{\beta}_2$ introduced in (2.10) and (2.13), we use also the estimator

$$\hat{\beta}_3 = -\frac{1}{\Delta}\ln\frac{\sum_{i=1}^{n}U_{i-1}U_i}{\sum_{i=1}^{n}U_{i-1}^2}, \tag{4.1}$$

which is based on martingal estimating functions and was proposed in Bibby & Sørensen (1995). It is easy to see, that $\hat{\beta}_3$ is related to the autocorrelation function of U. For details about this estimator and the theory of martingal estimating functions we refer to Bibby & Sørensen (1995) and the references therein. The diffusion coefficient is estimated via the slope of the quadratic variation, similarly as in (2.9).

The first row of Table 3 shows the means of the corresponding estimated values. In the second row the variance of the estimates are shown. We emphasize that the results are based on a directly observed simulated diffusion.

$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_3$	γ	$\hat{\beta}_1/\hat{\beta}_2$	opt h
0.01070	0.01018	0.01028	0.00996		
4.589e-06	3.695e-06	4.541e-06	1.059e-08		
0.00964	0.00967		0.00996	0.99435	295.0
7.161e-06	6.258e-06		1.067e-08	0.00248	

Table 3: Estimated parameters.

Furthermore, we simulate the logarithm of the index in (1.2) $\ln S$ as the sum of a linear function and U. In a second step we calculate $\ln X$ as in (2.6) with the Epanechnikov kernel, see Appendix A.1. We then estimate from the simulated data the parameters β and γ of U by the methods in (2.18), (2.19) and (2.9). This gives us an idea about the fixed sample behavior of these estimation methods when the residuals of a kernel smoothing procedure are observed instead of those of an Ornstein-Uhlenbeck process itself.

The estimated values calculated from the simulated trajectories of $\ln X$ are shown in the third and fourth row of Tables 3. The results clearly demonstrate

that the correction terms in (2.18) and (2.19) are necessary to obtain a reasonable estimated values. In the situation considered here, the correction terms equal each other and have approximately the value $c_k/h \approx 0.0061$, see Appendix A.1. Since the correction terms for $\hat{\beta}_3$ are not considered, we have not to report them in Table 3.

The table also shows the mean and the variance of the ratio $\hat{\beta}_1/\hat{\beta}_2$ used to select the bandwidth h, see (2.20). The mean of the selected bandwidth h, which brings this ratio as close as possible to one, is given in the last column.

The second part of the simulation study treats the bootstrap procedure. We apply the bootstrap methodology as introduced in Appendix A.3 to a simulated path of an Ornstein-Uhlenbeck process U following the dynamics in (2.4) with parameters $\beta = 0.01$, $\gamma = 0.01$ and $\Delta = 1$. The values of the parameters are reasonable with respect to the empirical results for the S&P 500. The number of observations is 5000 and the number of the bootstrapped series for the confidence bands is 160. The two plots in Figure 4 show the nonparametric estimators for the drift and squared diffusion coefficient together with their 90% confidence bands constructed by the bootstrap procedure. The plots also show the true parametric functions for the drift and diffusion coefficient. The dotted vertical lines are the empirical 0.005 and 0.995 quantiles of the stationary distribution of $\exp(U)$. If we only consider the range between these quantities, i.e. the range where 99% of the data reside, then both of the parametric functions remain inside the confidence bands. This means, the null hypotheses $H_0(m)$ and $H_0(\sigma^2)$ as in Section 2.6 cannot be rejected for data in this range.
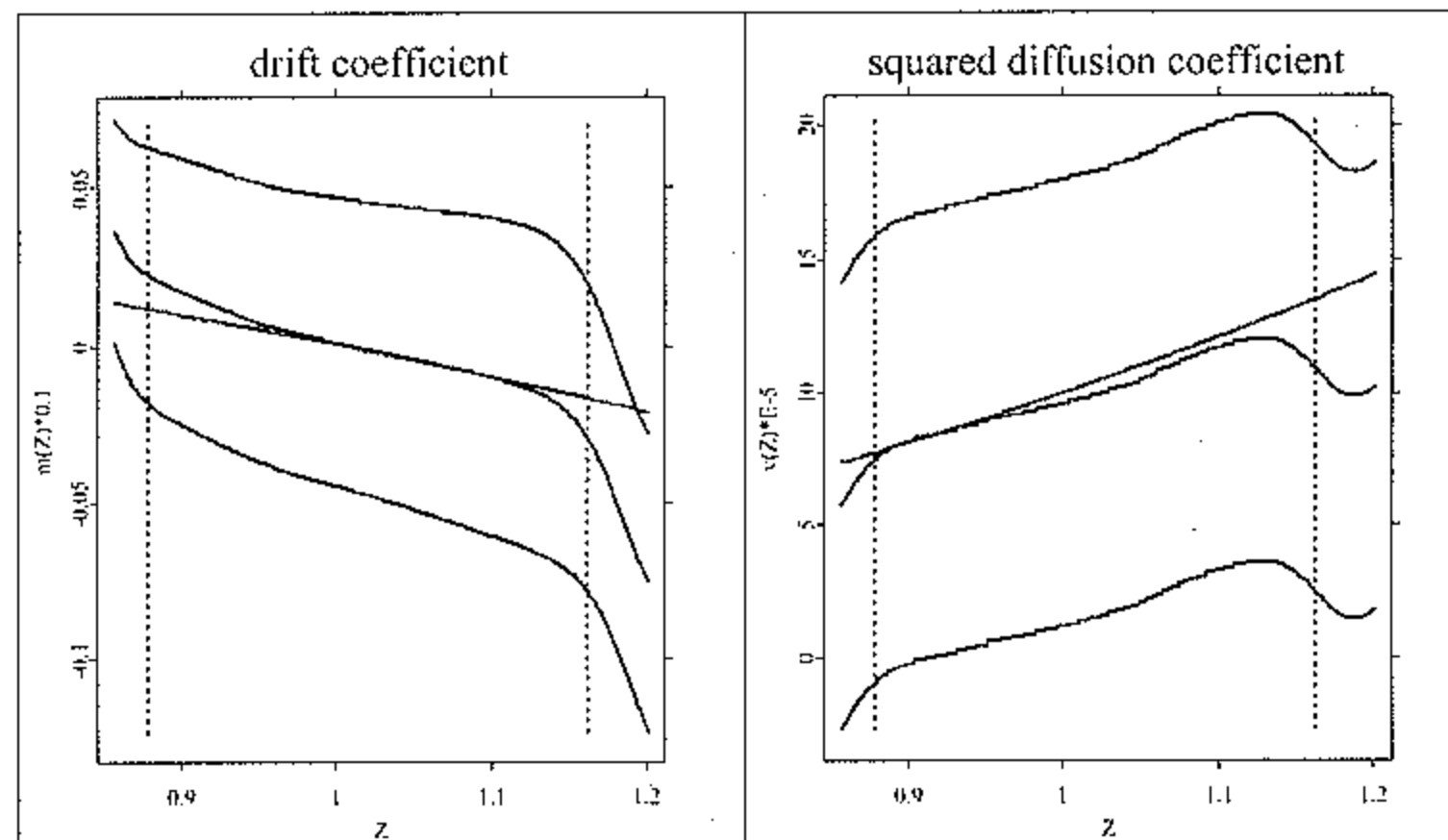

Figure 4: Nonparametric estimates for the drift and squared diffusion coefficients of a simulated geometric Ornstein-Uhlenbeck process, confidence bands and true functions.

5 Conclusion

In this paper we modeled an index as the product of an ergodic diffusion and a deterministic growth process. In the first part we proposed a methodology that allows us to separate the estimation of the average growth of the index and that of the parameters of the ergodic diffusion. The general methodology was carried out for the particular case of an Ornstein-Uhlenbeck process. A challenge for future research is to establish similar estimation methods for parameters of a square root process and other ergodic diffusions. The derivation of the corresponding correction terms will be the key problem in such an approach.

In the second part of the paper we developed a semiparametric testing procedure for the drift and diffusion coefficient functions of an ergodic diffusion. The test is based on a comparison of the parametric forms of these functions to their nonparametric estimators. Finally, an empirical analysis of the S&P 500 stock market index and a simulation study completed the paper.

A Appendix

A.1 Parameter Estimation

Let $U(t)$ be the Ornstein-Uhlenbeck process satisfying (2.4). Introduce the autocovariance of the process $U - K_h * U$ as

$$Cov(\tau) = COV[(U - K_h * U)(\tau); \ (U - K_h * U)(0)], \quad 0 < \tau << h.$$

Let

$$\rho_h(\tau) = \frac{Cov(\tau)}{Cov(0)}$$

be the autocorrelation function of $U - K_h * U$.

Proposition 1. (i) *If $K(u)$ is the rectangle kernel, i.e., $K(u) = (1/2)\,\mathbb{I}(|u| \leq 1)$, then as $h \to \infty$ we have that*

$$Cov(\tau) \ = \ \frac{\gamma^2}{2\beta}\Big(e^{-\beta\tau} - \frac{1}{\beta h} - \frac{1}{2\,\beta h}\,(\tau/h) + O(h^{-2}) \Big). \tag{A.1}$$

(ii) If $K(u) = (3/4)(1 - u^2)\,\mathbb{I}(|u| \leq 1)$ is the Epanechnikov kernel, then

$$Cov(\tau) \ = \ \frac{\gamma^2}{2\beta}\Big(e^{-\beta\tau} - \frac{1.8}{\beta h} - \frac{1.5}{\beta h}\,(\tau/h)^2 + O(h^{-2}) \Big). \tag{A.2}$$

Proposition 2. *Under the assumptions of Proposition 1, up to the terms of the magnitude $O(h^{-2})$, the following equation holds for β:*

$$\frac{\partial^+}{\partial \tau}\rho_h(\tau)\Big|_{\tau=0} = -\frac{\beta}{1 - c_K/(\beta h)}.$$

The solution of this equation is approximately

$$\hat{\beta}_2 = \Big|\frac{\partial^+}{\partial \tau}\rho_h(\tau)\Big|\Big|_{\tau=0} - \frac{c_K}{h} + O(h^{-2})$$

where $c_K = 1$ *if K is the rectangle kernel, and $c_K = 1.8$ if K is the Epanechnikov kernel.*

To the proof of Proposition 1. (i) Integrate the both sides of (2.4) with the rectangle kernel. The integration results in

$$-\frac{1}{2\beta h}\{U(t+h) - U(t-h)\} = (K_h * U)(t) - \frac{\gamma}{2\beta h}\{W(t+h) - W(t-h)\},$$

or

$$(K_h * U)(t) = \frac{\gamma}{2\beta h}\big(W(t+h) - W(t-h)\big) - \frac{1}{2\beta h}\big(U(t+h) - U(t-h)\big). \,(\text{A.3})$$

The autocovariance function equals

$$\begin{aligned}
Cov(\tau) \;=\; & E[U(\tau)U(0)] - E[U(\tau)\,(K_h * U)(0)] \\
& - E[U(0)\,(K_h * U)(\tau)] + E[(K_h * U)(\tau)\,(K_h * U)(0)]. \,(\text{A.4})
\end{aligned}$$

The first term in the latter formula is $E[U(\tau)U(0)] = \{\gamma^2/(2\beta)\}e^{-\beta\tau}$. With the help of (A.3) and the explicit representation

$$U(t) = \int_{-\infty}^{t} \exp\{-\beta(t-s)\}\,\gamma\,dW(s),$$

one finds by direct calculation that each of the negative terms on the right-hand side of (A.4) contributes

$$-\frac{\gamma^2}{(2\beta^2 \, h)} + O(h^{-2}),$$

while the covariance of $K_h * U$ adds up to

$$\frac{\gamma^2(2h - \tau)}{(2\beta h)^2} + O(h^{-2}).$$

Combining these results, we arrive at (A.1).

(ii) Integrating (2.4), we find as in (A.3) that

$$(K_h * U)(t) = -\frac{3}{2\beta h^3} \int_{t-h}^{t+h} (s - t)U(s)ds + \frac{\gamma}{\beta}\,(K_h * W)(t). \qquad (\text{A.5})$$

It is straightforward to verify that the variance of the first term on the right-hand side of (A.5) has the magnitude $O(h^{-2})$ for h large. This term is negligible as compared to the second one. As in part (i), we obtain

$$\begin{aligned}
Cov(\tau) \;=\;& \frac{\gamma^2}{2\beta}e^{-\beta\tau} - 2\left(\frac{3\gamma^2}{4\beta^2 h} + O(h^{-2})\right) + \frac{\gamma^2}{\beta^2}\,E\big[(K_h * W)(\tau)\,(K_h * W)(0)\big] \\
\;=\;& \frac{\gamma^2}{2\beta}e^{-\beta\tau} - \frac{3\gamma^2}{2\beta^2 h} + \frac{\gamma^2}{\beta^2}\left(\frac{3}{5} - \frac{3}{4}\left(\frac{\tau}{h}\right)^2\right) + O(h^{-2}). \qquad (\text{A.6})
\end{aligned}$$

This proves (A.2). □

To the proof of Proposition 2. The slope of the autocorrelation at zero follows from (A.1) and (A.2). Let

$$A = \left|\frac{\partial^+}{\partial\tau}\rho_h(\tau)\right|\bigg|_{\tau=0},$$

then for $\hat{\beta}_2$ the quadratic equation $\hat{\beta}_2^2 \, h - \hat{\beta}_2 \, hA + c_K A = 0$ holds with root $\hat{\beta}_2 = A - c_K/h + O(h^{-2})$. $\qquad\square$

Proposition 3. *Let C_0 be a space of continuous function with finite support. Define K as the rectangle or Epanechnikov kernel. Then the operator K_h is a contracting operator on the space $L_2 \bigcap C_0$ with the L_2-norm.*

To the proof of Proposition 3. The Fourier transformation for the rectangle kernel is $\tilde{K}(z) = (\sin z)/z$, and for the Epanechnikov kernel is $\tilde{K}(z) = 3(\sin z - z \cos z)/z^3$ with unique maximum value 1 at $z = 0$. Thus, for the n-th iterative convolution, $\|K^n\|_2 \to 0$ as $n \to \infty$. This confirms the result. $\qquad\square$

A.2   Asymptotic Properties of LLP

The asymptotic properties of local polynomial estimates are studied in Fan & Gijbels (1996) and Härdle, Klinke & Müller (1999). Under some smoothness conditions with bandwidth $h_1 = k_0/n^{1/5}$ for a constant $k_0 > 0$, the results applied to our case provide the following formulas on the asymptotic normality

$$n^{2/5} \{\hat{m}_{h_1}(z) - m(z)\} \xrightarrow{\mathcal{D}} N\left(\frac{k_0^2}{2}\mu_2(K) \, \Delta \, m''(z), \, \frac{\|K\|_2^2 \, \Delta \, \sigma^2(z)}{k_0 p_0(z)} \right) \qquad (A.7)$$

$$n^{2/5} \{\hat{\sigma}_{h_1}^2(z) - \sigma^2(z)\} \xrightarrow{\mathcal{D}} N\left(\frac{k_0^2}{2}\mu_2(K) \left(\Delta \left(\sigma^2(z)\right)'' + 2(\Delta m'(z))^2 \right), \, \frac{2\Delta^2 \sigma^4(z) \|K\|_2^2}{k_0 p_0(z)} \right). \qquad (A.8)$$

Here $\mu_2(K)$ is the second moment of the kernel K and $\|K\|_2$ is its L_2 norm. Furthermore $p_0(z)$ denotes the stationary density of Z as given in Section 2.4.

A.3   The Bootstrap Procedure

The confidence bands in Section 2.6 for the nonparametric estimator are constructed by the following bootstrap method :

1. Choose a bandwidth g, which is larger than the optimal h_1 in order to have oversmoothing. Estimate then nonparametrically $m(.)$ and $\sigma^2(.)$ and obtain the residual estimated errors :

$$\hat{\varepsilon}_i = \frac{Y_i - \Delta \hat{m}_g\{Z^\Delta(t_i)\}}{\sqrt{\Delta}\hat{\sigma}_g\{Z^\Delta(t_i)\}}.$$

Since we make the assumption that the ε_i has zero-mean, we subtract the sample mean of $\hat{\varepsilon}_i$.

2. Replicate N times the series of the $(\hat{\varepsilon}_i)$ with wild bootstrap obtaining $(\varepsilon_i^{*,n})$ for $n = 1, \ldots, N$ and build N new bootstrapped series $(Z^{*,n})$:

$$Z_1^{*,n} = Z^\Delta(t_1)$$

$$Z_{i+1}^{*,n} - Z_i^{*,n} = \Delta \hat{m}_g(Z_i^{*,n}) + \sqrt{\Delta}\hat{\sigma}_g(Z_i^{*,n})\varepsilon_i^{*,n}.$$

Estimate again $m(z)$ and $\sigma^2(z)$ for each of the N bootstrapped series with bandwidth h_1.

3. Build the statistics:

$$T_m^* = \sup_z \frac{|\hat{m}_{h_1}^{*,n}(z) - \hat{m}_{h_1}(z)|}{\hat{\sigma}_{h_1}^{*,n}(z)}$$

and

$$T_\sigma^* = \sup_z |(\hat{\sigma}^2)_{h_1}^{*,n}(z) - \hat{\sigma}_{h_1}^2(z)|$$

4. Form the $(1-\alpha)$ confidence bands KB

$$KB(m(.)) = \left[\hat{m}_{h_1}(z) - \hat{\sigma}_{h_1}(z)t_{m,1-\alpha/2} \,,\, \hat{m}_{h_1}(z) + \hat{\sigma}_{h_1}(z)t_{m,\alpha/2}\right]$$

and

$$KB(\sigma^2(.)) = \left[\hat{\sigma}_{h_1}^2(z) - t_{\sigma,1-\alpha/2} \,,\, \hat{\sigma}_{h_1}^2(z) + t_{\sigma,\alpha/2}\right]$$

where $t_{m,\alpha}$ and $t_{\sigma,\alpha}$ denote the empirical α-quantile of T_m^* and T_σ^*, respectively.

Acknowledgments

References

Ait-Sahalia, Y. (1996), 'Nonparametric pricing of interest rate derivative securities.', *Econometrica* **64**(3), 527–560.

Bibby, B. M. & Sørensen, M. (1995), 'Martingale estimation functions for discretely observed diffusion processes', *Bernoulli* **1**(1/2), 17 – 40.

Cox, J., Ingersoll, J. & Ross, S. (1985), 'A theory of term structure of interest rates', *Econometrica* **2**(53), 385–407.

Duan, J.-C. (1997), 'Augmented GARCH(p,q) process and its diffusion limit.', *J. Econom.* **79**(1), 97–127.

Engle, R. F. & Bollerslev, T. (1986), 'Modelling the persistence of conditional variances', *Econometric Rev.* **5**(1), 1–87. With comments a reply by the authors.

Fan, J. & Gijbels, I. (1996), *Local Polynomial Modelling and its Applications - Theory and Methodologies*, Chapman and Hall, New York.

Fleming, W. H. & Sheu, S.-J. (1999), 'Optimal long term growth rate of expected utility of wealth.', *Ann. Appl. Probab.* **9**(3), 871–903.

Föllmer, H. & Schweizer, M. (1993), 'A microeconomic approach to diffusion models for stock prices.', *Math. Finance* **3**(1), 1–23.

Franke, J., Kreiss, J.-P., Mammen, E. & Neumann, M. (1998), Properties of the nonparametric autoregressive bootstrap. Discussion paper, SFB 373, No. 54/98, Humboldt University Berlin.

Frey, R. (1997), 'Derivative asset analysis in models with level-dependent and stochastic volatility.', *CWI Q.* **10**(1), 1–34.

Gallant, A. & Tauchen, G. (1996), 'Which moments to match?', *Econometric Theory* **12**(4).

Genon-Catalot, V., Jeantheau, T. & Larédo, C. (2000), 'Stochastic volatility models as hidden markov models and statistical applications', *Bernoulli* **6**(6).

Ghysels, E., Harvey, A. & Renault, E. (1996), *Statistical Methods in Finance*, Vol. 14 of *Handbook of Statistics*, North-Holland, chapter Stochastic Volatitlity, pp. 119–191.

Hall, P. (1985), 'Resampling a coverage pattern', *Stoch. Proc. Appl.* **20**, 231–246.

Hansen, L. P. & Scheinkman, J. A. (1995), 'Back to the future: Generating moment implications for continuous-time Markov processes.', *Econometrica* **63**(4), 767–804.

Hansen, L. P., Scheinkman, J. A. & Touzi, N. (1998), 'Spectral methods for identifying scalar diffusions.', *J. Econom.* **86**(1), 1–32.

Härdle, W. (1990), *Applied Nonparametric Regression*, number 19 *in* 'Econometric Society Monographs', Cambridge University Press.

Härdle, W., Klinke, S. & Müller, M. (1999), *XploRe -The Statistical Computing Environment*, Springer Verlag, New York.

Härdle, W. & Tsybakov, A. (1997), 'Local polynomial estimators of the volatility function in nonparametric autoregression', *Journal of Econometrics* **81**, 223–242.

Jiang, G. & Knight, J. (1997), 'A nonparametric approach to the estimation of duffusion processes, with an application to a short-term interest rate model', *Econometric Theory* **13**(5).

Karatzas, I. & Shreve, S. E. (1998), *Methods of Mathematical Finance*, Vol. 39 of *Applications of Mathematics, Stochastic Modelling and Applied Probability*, Springer Verlag New York.

Kessler, M. & Soerensen, M. (1999), 'Estimating equations based on eigenfunctions for a discretely observed diffusion process.', *Bernoulli* **5**(2), 299–314.

Kloeden, P. E. & Platen, E. (1999), *Numerical Solution of Stochastic Differential Equations*, Vol. 23 of *Applications of Mathematics*, Springer Verlag Berlin Heidelberg.

Nelson, D. B. (1990), 'ARCH models as diffusion approximations', *J. Econ.* **45**(1/2), 7–38.

Neumann, M. & Kreiss, J.-P. (1998), 'Regression-type inference in nonparametric autoregression', *Ann. Statist.* **26**, 1570–1613.

Platen, E. (2000), Risk premia and financial modelling without measure transformation. University of Technology Sydney, School of Finance & and Economics and Department of Mathematical Sciences.

Platen, E. & Rebolledo, R. (1996), 'Principles for modelling financial markets.', *J. Appl. Probab.* **33**(3), 601–613.

Semi-parametric estimation of partially linear single-index models

Yingcun Xia[a],*, Wolfgang Härdle[b]

[a]*Department of Statistics and Applied Probability, National University of Singapore, Singapore*
[b]*CASE—Center for Applied Statistics and Economics, Humboldt-Universität zu Berlin, Berlin, Germany*

Abstract

One of the most difficult problems in applications of semi-parametric partially linear single-index models (PLSIM) is the choice of pilot estimators and complexity parameters which may result in radically different estimators. Pilot estimators are often assumed to be root-n consistent, although they are not given in a constructible way. Complexity parameters, such as a smoothing bandwidth are constrained to a certain speed, which is rarely determinable in practical situations.

In this paper, efficient, constructible and practicable estimators of PLSIMs are designed with applications to time series. The proposed technique answers two questions from Carroll et al. [Generalized partially linear single-index models, J. Amer. Statist. Assoc. 92 (1997) 477–489]: no root-n pilot estimator for the single-index part of the model is needed and complexity parameters can be selected at the optimal smoothing rate. The asymptotic distribution is derived and the corresponding algorithm is easily implemented. Examples from real data sets (credit-scoring and environmental statistics) illustrate the technique and the proposed methodology of minimum average variance estimation (MAVE).
© 2005 Elsevier Inc. All rights reserved.

AMS 1991 subject classification: 62G07; 62G20; 62F10

Keywords: Asymptotic distribution; Generalized partially linear model; Local linear smoother; Optimal consistency rate; Single-index model

1. Introduction

Although the presence of nonlinearities in statistical data analysis is often modelled with non- and semi-parametric methods, there are still few noncritical semi-parametric techniques. One

---

* Corresponding author. Fax: +65 6872 2523.

E-mail address: staxyc@stat.nus.edu.sg (Y. Xia).

argument that has been advanced is that—despite a reduction in dimensionality—the practical estimation still depends heavily on pilot estimators and complexity parameters. Another argument against finely tuned semi-parametrics is that mathematical tools for inferential decisions and software implementations are either missing or not readily accessible. The purpose of this paper is to show that such critiques may be refuted even for the very flexible class of *partially linear single-index models* (PLSIM):

$$y = \beta_0^T Z + g(\theta_0^T X) + \varepsilon, \tag{1.1}$$

where $E(\varepsilon|X, Z) = 0$ almost surely, β_0 and θ_0 are unknown parameters, $g(\cdot)$ is an unknown link function. The PLSIM (1.1) was first analyzed by Carroll et al. [4] and contains the single-index models ($\beta_0 \equiv 0$), generalized partially linear models (X one-dimensional and y observed logits) and generalized linear models ($\beta_0 \equiv 0$ and g known). The advantage of the PLSIM lies in its generality and its flexibility. The wide spread application of PLSIMs though is somewhat obstructed by the facts described above: necessity of pilot estimators for θ_0 and complexity parameters such as bandwidths (to estimate the link function g). In this paper, we further assume $\|\theta_0\| = 1$ and that the first entry of θ_0 is positive for model identification; see [26].

The issue of the order of magnitude of the complexity parameter was addressed in [4, Eq. (18), p. 483]. The convenience of a root-n pilot estimator for θ_0 was employed in [7] but was found to severely influence the final estimate. In practical application, these two important questions will be addressed in this paper: we will show that a simple multi-dimensional kernel estimator suffices to ensure root-n consistency of the parametric parts of (1.1) and that no under-smoothing is required for the proposed algorithm.

One motivation of our work comes from credit scoring and the study of nonlinear effects in retail banking. Another motivation comes from the analysis of circulatory and respiratory problems in Hong Kong and the study of the complicated effect of weather conditions on the health problems. *Credit scoring* methods are designed to asses risk measures for potential borrowers, companies, etc. Typically, the scoring is reduced to a classification or (quadratic) discriminant analysis problem, see [1,10]. The credit data set of Müller and Rönz [17] consists of 6180 cases with 8 metric variables ($x_2, \ldots, x_9$) and 15 categorical explanatory variables ($x_{10}, \ldots, x_{24}$). Co-variates x_3, x_4, x_5 are of the most interest and represent, respectively, duration of payment, amount of loan and customer's age. The response variable y was 0 if customers paid their installments without problem, or 1 otherwise. There were 513 cases with a y value of 1. A scatterplot matrix of the observations (x_2, x_3, x_4, x_5) is given in Fig. 1.

The distribution of the variable y (black points in Fig. 1) shows a clear nonlinear structure and speaks therefore against a linear discriminant analysis. A logit model

$$\text{logit}\{P(y = 1|X, Z)\} = \beta_0^T Z + \theta_0^T X \tag{1.2}$$

(also of linear structure) shows clear nonlinearity in the residuals, see [17]. Here X denotes the vector of metric variables and Z the vector of categorical variables. Müller and Rönz [17] therefore applied a partially linear approach as in [22] by replacing one linear term in (1.2) operating on the metric variable x_5 by a nonparametric function $g(x_5)$ as shown in Fig. 2.

We partition the range of x_4 (or x_5) into 50 intervals with equal lengths. We cluster the observations with x_4 (or x_5) in the same interval as one class. We calculate the relative frequencies $\hat{p}$ for $y = 1$. In Fig. 2, the variable x_4 (or x_5) is plotted against the logit($\hat{p}$) $= \log(\hat{p}/(1 - \hat{p}))$. Using bootstrap, the nonlinearity was tested and found to be significant. The question of how to integrate further nonlinear influences by the other metric variables was analyzed in [17] at a
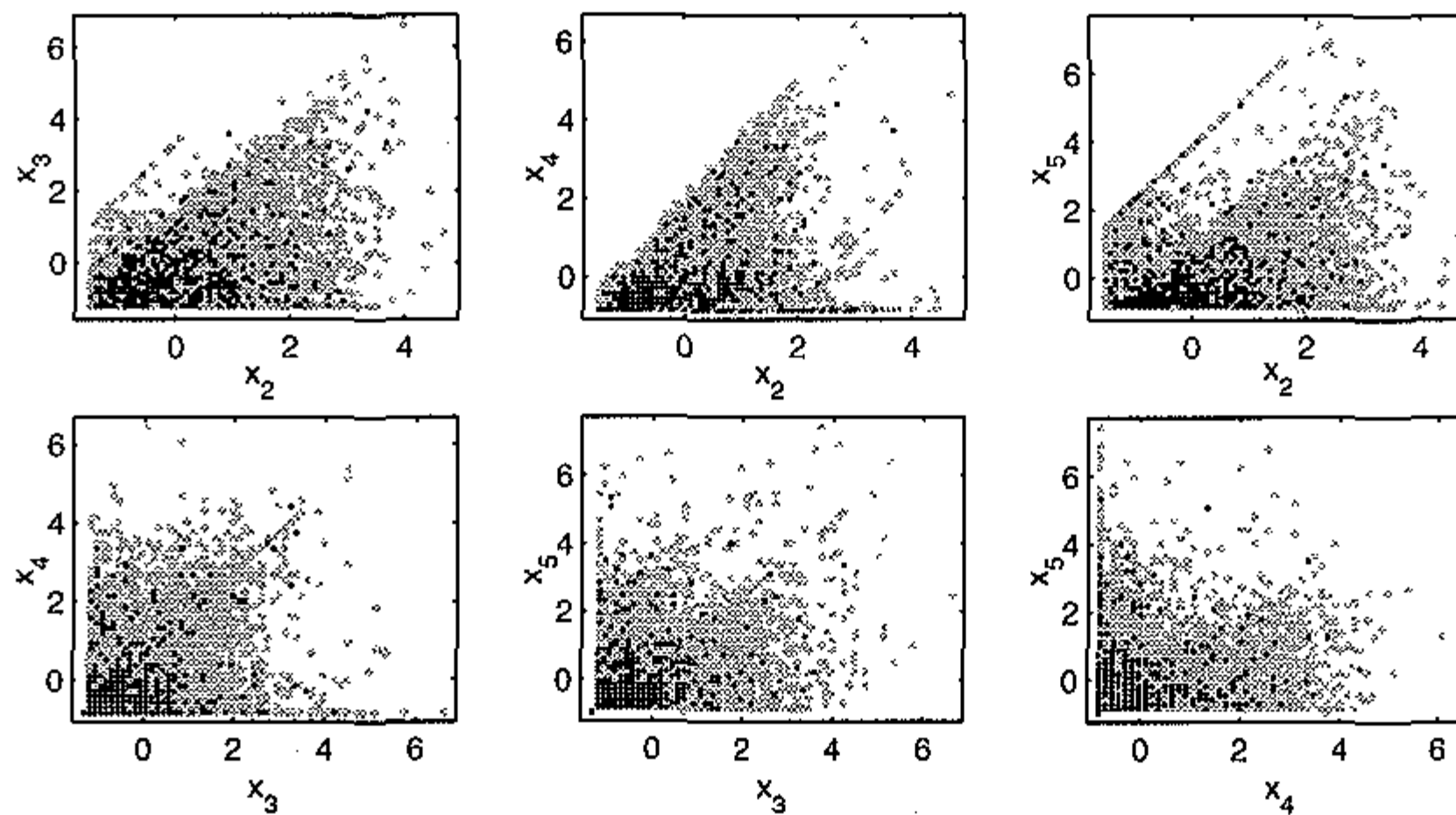
Fig. 1. Scatterplots: variables x_2 to x_5, observations corresponding to $y = 1$ are emphasized in black.
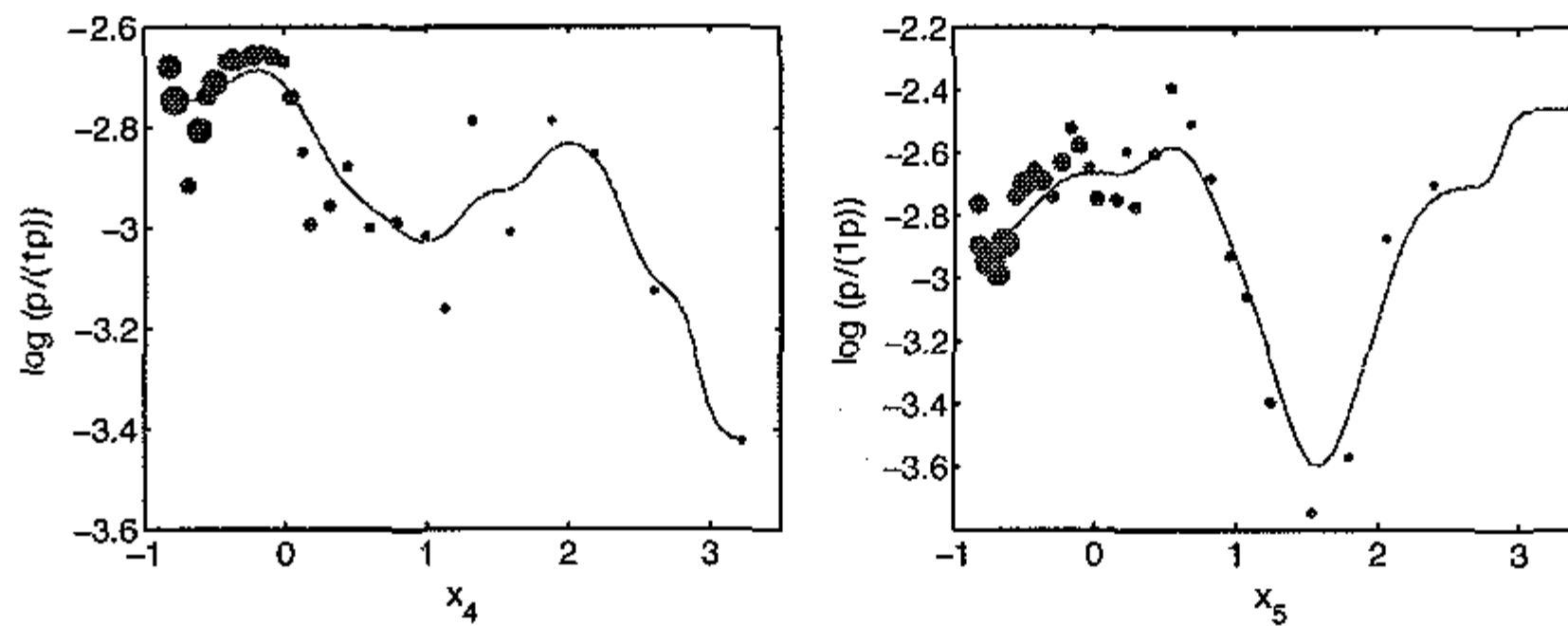

Fig. 2. Marginal dependency. Thicker bullets correspond to more observations in a class. The lines are local linear smoothers.

multi-dimensional kernel regression (e.g. on (x_4, x_5), see Fig. 5.6 in their article) and found to be too difficult to implement due to the high-dimensional kernel smoothing. The technique that we develop here will make it possible to overcome the dimensionality issue and indicate nonlinear influences on the logits via the PLSIM. Based on these discussion, we can consider fit the relation between $Y = \log(p/(1 - p))$ and (X, Z) by

$$y = \beta^T \tilde{Z} + g(\theta_1 x_4 + \theta_2 x_5) + \varepsilon, \tag{1.3}$$

where $\tilde{Z} = (x_2, x_3, x_6, \ldots, x_{24})$.

The other motivation of this research comes from the investigation of the number of daily hospital admissions of patients suffering from the circulatory and respiratory (CR) problems in Hong Kong from 1994–1996. There is a jump in the numbers at the beginning of 1995 due to the additional hospital beds released to accommodate CR patients from the beginning of 1995. We remove this jump by a simple kernel smoothing over time and denote the remaining time series by y_t. The pollutants and weather conditions might cause the CR problems. The pollutants include sulfur dioxide (x_{1t}, in $\mu g\, m^{-3}$), nitrogen dioxide (x_{2t}, in $\mu g\, m^{-3}$), respirable suspended particulates (x_{3t}, in $\mu g\, m^{-3}$) and ozone (x_{4t}, in $\mu g\, m^{-3}$), and weather conditions include temperature (x_{5t}, in °C)
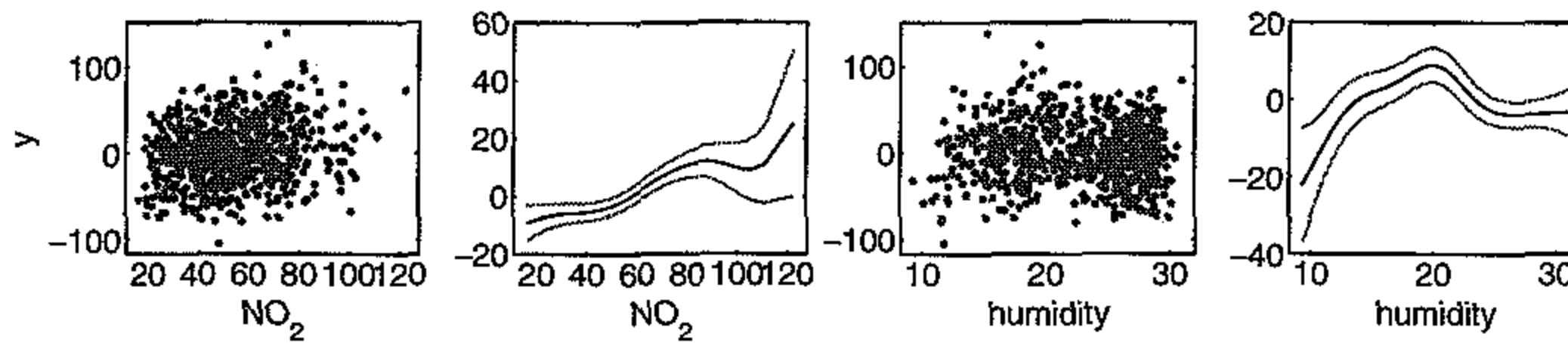
Fig. 3. The first and third panels are the plots of daily y against NO_2 and humidity, respectively. In the second and fourth panels, the central lines are kernel smoothers of y on NO_2 and humidity, respectively, the upper and lower lines are the corresponding 95% pointwise confidence intervals.

and relative humidity (x_{6t}, in %). It is obvious that the higher the levels of air pollutants are, the stronger they tend to cause health problems. Furthermore, simple kernel smoothing suggests that we can approximate the relations between y_t and the pollution levels linearly; see Fig. 3. However, for the other covariates such as temperature and humidity, the relations are unknown and might be nonlinear. Fig. 3 is simple regression analyses based on kernel smoothing. The relation of y_t with NO_2 is almost linear, but the relation of y_t with humidity is nonlinear and hard to explain. To explore the relation between y_t and air pollutants and weather conditions, we may consider the following model:

$$y_t = \beta^T Z_t + g(\theta^T X_t) + \varepsilon_t, \tag{1.4}$$

where Z_t consists of levels of pollutants and their lagged variables, and X_t consists of weather conditions and their lagged variables.

A nonparametric estimation method can be evaluated in the following four aspects: (1) computational cost, (2) efficiency, (3) conditions necessary for the consistency, and (4) the range of the bandwidth. Most nonparametric methods for the estimation of model (1.1) or its special cases are criticized in one or another aspects. The methods in [4,8,13,23,24] include complicated optimization techniques and no simple algorithm is available up to now. The method of Li [14] required symmetric distribution of the covariate; Härdle and Stoker [9] and Hristache et al. [12] required that $|Eg'(\theta_0^T X)|$ is away from 0. If these conditions are violated, their methods cannot obtain useful estimators. The method of Härdle and Stoker [9] and the method of Hristache et al. [11,12] are not asymptotically efficient in the semi-parametric sense. Most of the methods mentioned above require a bandwidth that is much smaller than the data-driven bandwidth in order to allow the estimator of the parameters to achieve root-n consistency, i.e. under-smoothing the link function is needed; see, [4,6,9,11,12] among others. More discussions on the selection of bandwidth for the partially linear model can be found in [15].

In this paper, we present the minimum average variance estimation (MAVE) method that will provide a remedy to these four weak points.

2. Estimation method and asymptotic results

The basic algorithm for estimating the parameters in (1.1) is based on observing that

$$(\beta_0, \theta_0) = \arg\min_{\beta,\theta} E \left[y - \{\beta^T Z + g(\theta^T X)\} \right]^2 \tag{2.1}$$

subject to $\theta^T \theta = 1$. By conditioning on $\xi = \theta^T X$, we see that (2.1) equals $E_\xi \sigma^2_{\beta,\theta}(\xi)$ where

$$\sigma^2_{\beta,\theta}(\xi) = E\left[(y - \{\beta^T Z + g(\xi)\})^2 | \theta^T X = \xi\right].$$

It follows that

$$E\left[y - \{\beta^T Z + g(\theta^T X)\}\right]^2 = E_\xi \sigma^2_{\beta,\theta}(\theta^T X).$$

Therefore, minimization (2.1) is equivalent to,

$$(\beta_0, \theta_0) = \arg\min_{\beta,\theta} E_\xi \sigma^2_{\beta,\theta}(\xi) \tag{2.2}$$

subject to $\theta^T \theta = 1$. Let $\{(X_i, Z_i, y_i) \, i = 1, 2, \ldots, n\}$ be a sample from (X, Z, y). The conditional expectation in (2.2) is now approximated by the sample analog. For X_i close to x, we have the following local linear approximation:

$$y_i - \beta_0^T Z - g(\theta_0^T X_i) \approx y_i - \beta_0^T Z_i - g(\theta_0^T x) - g'(\theta_0^T x) X_{i0}^T \theta_0,$$

where $X_{i0} = X_i - x$. Following the idea of local linear smoothing [5], we may estimate $\sigma^2_{\beta,\theta}(\theta^T x)$ by

$$\hat{\sigma}^2_{\beta,\theta}(\theta^T x) = \min_{a,d} \sum_{i=1}^{n} \left\{y_i - \beta^T Z_i - a - d X_{i0}^T \theta\right\}^2 w_{i0}. \tag{2.3}$$

Here, $w_{i0} \geqslant 0, i = 1, 2, \ldots, n$, are some weights with $\sum_{i=1}^{n} w_{i0} = 1$, typically centering at x. Let $X_{ij} = X_i - X_j$. By (2.2) and (2.3), our estimation procedure is to minimize

$$\frac{1}{n} \sum_{j=1}^{n} G(\theta^T X_j) I_n(X_j) \sum_{i=1}^{n} \left\{y_i - \beta^T Z_i - a_j - d_j X_{ij}^T \theta\right\}^2 w_{ij} \tag{2.4}$$

with respect to (a_j, d_j) and (β, θ), where $G(\cdot)$ is another weight function that controls the contribution of (X_j, Z_j, y_j) to the estimation of β and θ. For example, when the model is assumed to be heteroscedastic and $Var(y|X, Z) = V(\theta_0^T X)$, then $G(.) = V(.)$; see [4,7]. $I_n(x)$ is employed here for technical purpose to handle the boundary points. It is given in the next section. See also [7]. For simplicity, we can take $I_n(\cdot) = 1$ in practice. If the noise term is negligible, the true parameter (β_0, θ_0) is a solution to the minimization problem in (2.4) with $w_{ij} \to 0$ if $X_i - X_j \to 0$. This observation can give us an intuition about why minimizing (2.4) can generate a consistent estimator. We call the estimation procedure the MAVE (conditional) method. The main difference between our method and those in [4,7] is that our method estimate all the parameters and the nonparametric function by minimizing a single common loss function as in (2.4) while those in the two papers estimated nonparametric functions and parameters by minimizing two different loss functions, respectively. This difference is the reason that under-smoothing is not necessary in our method. Similar spirit appeared in [7] for the single-index model, where they minimized a common loss function with respect to both smoothing parameter h and the single-index, but employed another one for the nonparametric function.

Minimizing (2.4) can be decomposed to two typical quadratic programming problems by fixing $(a_j, d_j), j = 1, \ldots, n$ and (β, θ) alternatively. They can be solved easily with simple analytic

expressions. Given (β, θ), we have from minimizing (2.4)

$$\begin{pmatrix} a_j \\ d_j \end{pmatrix} = \left\{ \sum_{i=1}^{n} w_{ij} \begin{pmatrix} 1 \\ X_{ij}^T \theta \end{pmatrix} \begin{pmatrix} 1 \\ X_{ij}^T \theta \end{pmatrix}^T \right\}^{-1} \sum_{i=1}^{n} w_{ij} \begin{pmatrix} 1 \\ X_{ij}^T \theta \end{pmatrix} (y_i - \beta^T Z_i). \tag{2.5}$$

Given (a_j, d_j), we have from minimizing (2.4)

$$\begin{pmatrix} \beta \\ \theta \end{pmatrix} = \left\{ \sum_{j=1}^{n} G(\theta^T X_j) I_n(X_j) \sum_{i=1}^{n} w_{ij} \begin{pmatrix} Z_i \\ d_j X_{ij} \end{pmatrix} \begin{pmatrix} Z_i \\ d_j X_{ij} \end{pmatrix}^T \right\}^{-1}$$

$$\times \sum_{j=1}^{n} G(\theta^T X_j) I_n(X_j) \sum_{i=1}^{n} w_{ij} \begin{pmatrix} Z_i \\ d_j X_{ij} \end{pmatrix} (y_i - a_j) \tag{2.6}$$

and standardize $\theta := sgn_1\, \theta/|\theta|$, where sgn_1 is the sign of the first entry in θ. Here and later, $|\gamma| = (\gamma^T \gamma)^{1/2}$ for any vector γ. The minimization in (2.4) can be solved by iterations between (2.5) and (2.6). We call the above procedure the PLSIM algorithm.

The choice of the weights w_{ij} plays an important role in different estimation methods. See [11,12,25]. In this paper, we use two sets of weights. Suppose $H(\cdot)$ and $K(\cdot)$ are a p-variate symmetric density function and a univariate symmetric density function, respectively. The first set of weights is $w_{ij} = H_{b,i}(X_j)/\sum_{\ell=1}^{n} H_{b,\ell}(X_j)$, where $H_{b,i}(X_j) = b^{-p} H(X_{ij}/b)$ and b is a bandwidth. This is a multivariate dimensional kernel weight. For this kind of weights, we set $I_n(x) = 1$ if $n^{-1} \sum_{\ell=1}^{n} H_{b,\ell}(x) > c_0$; 0 otherwise for some constant $c_0 > 0$. Iterating (2.5) and (2.6) until convergence, denote the estimators (i.e., the final values) of θ and β by $\tilde{\theta}$ and $\tilde{\beta}$, respectively. Because of the so-called "curse of dimensionality", the estimation based on this kind of weights is not efficient. However, the multivariate kernel weight can help us to find consistent initial estimators for θ_0 and β_0. We then use single-index kernel weights $w_{i,j}^{\theta} = K_{h,i}^{\theta}(\theta^T X_j)/\sum_{\ell=1}^{n} K_{h,\ell}^{\theta}(\theta^T X_j)$, where $K_{h,i}^{\theta}(v) = h^{-1} K\{(\theta^T X_i - v)/h\}$, h is the bandwidth and θ is the previous estimate of θ_0. Here, we take $I_n(x) = 1$ if $\hat{f}_\theta(\theta^T x) > c_0$; 0 otherwise, where $\hat{f}_\theta(v) = n^{-1} \sum_{\ell=1}^{n} K_{h,\ell}^{\theta}(v)$. Iterating (2.5) and (2.6) until convergence, denote the estimators (i.e. the final values) of θ and β by $\hat{\theta}$ and $\hat{\beta}$, respectively. After obtaining estimates $\hat{\theta}$ and $\hat{\beta}$, we can then estimate $g(v)$ by the solution of a_j in (2.5) with $\theta^T X_j$ replaced by v, denote the estimate by $\hat{g}(v)$.

Lemma 1. *Let $\tilde{\beta}$ and $\tilde{\theta}$ be the estimators based on the multi-dimensional kernel weight. Suppose that (C1)–(C6) in Section 6 hold, $b \to 0$ and $nb^{p+2}/\log n \to \infty$. If we start the estimation procedure with θ such that $\theta^T \theta_0 \neq 0$, then*

$$\tilde{\theta} - \theta_0 = o_P(1), \quad \tilde{\beta} - \beta_0 = o_P(1).$$

This lemma suggests that we can obtain consistent initial estimators of θ_0 and β_0 using multi-dimensional kernel. Although these initial estimators have a slow consistent rate, they suffice for us to obtain root-n consistent estimators eventually. If the design of X is symmetric, the method of Li [14] can also be used to get consistent initial estimators. There are other numerical methods available for the choice of the initial values. See, for example, [2,20]. However, those methods cannot guarantee the consistent properties for the initial estimators. The consistency is needed for the technical purpose to get the following theorem. Let $\mu_\theta(x) = E(X|\theta^T X = \theta^T x)$,

$v_\theta(x) = E(Z|\theta^T X = \theta^T x)$, and for $k = 0$ and 2,

$$W_k = E\left\{ G(\theta_0^T X) I(f_{\theta_0}(\theta_0^T X) > c_0) \begin{pmatrix} Z - v_{\theta_0}(X) \\ g'(\theta_0^T X)\{X - \mu_{\theta_0}(X)\} \end{pmatrix} \right.$$
$$\left. \times \begin{pmatrix} Z - v_{\theta_0}(X) \\ g'(\theta_0^T X)\{X - \mu_{\theta_0}(X)\} \end{pmatrix}^T |\varepsilon|^k \right\}.$$

Theorem 1. *Let $(\hat{\beta}, \hat{\theta})$ be the estimators based on the single-index kernel weight starting with $(\beta, \theta) = (\tilde{\beta}, \tilde{\theta})$. Suppose (C1)–(C6) in Section 6 hold, $h \sim n^{-\delta}$ with $1/6 < \delta < 1/4$ and that $E\{\varepsilon_i|(X_j, Z_j, y_j), j < i\} = 0$ almost surely. Then*

$$n^{1/2}\begin{pmatrix} \hat{\beta} - \beta_0 \\ \hat{\theta} - \theta_0 \end{pmatrix} \xrightarrow{D} N(0, W_0^- W_2 W_0^-),$$

where W_0^- is the Moore–Penrose inverse of W_0. If further the density function $f_{\theta_0}(v)$ of $\theta_0^T X$ is positive and the derivative of $E(\varepsilon^2|\theta_0^T X = v)$ exists, then

$$(nh)^{1/2}\{\hat{g}(v) - g(v) - \frac{1}{2}\kappa_2 g''(v)h^2\} \xrightarrow{D} N(0, f_{\theta_0}^{-1}(v)\int (K(v))^2 \, dv E(\varepsilon^2|\theta_0^T X = v)),$$

where $\kappa_2 = \int K(v)v^2 \, dv$.

If $E\{\varepsilon_i|(X_j, Z_j, y_j), j < i\} \neq 0$, then the asymptotic normal distribution still holds, but the covariance matrix in the distribution depends on the structure of the stochastic process of the observations. If $E(\varepsilon^2|X, Z) = \sigma^2$ is constant, then the asymptotic distribution of $(\hat{\beta}, \hat{\theta})$ is the same as that obtained by Carroll et al. [4]. They further showed that their estimator is efficient in the semi-parametric sense if the conditional density of Y given X and Z belongs to the exponential distribution family. Therefore, our estimator is also efficient in the semi-parametric sense under the same conditions. If Theorem 1 is used for statistical inference about the parameters, we need to have a consistent estimator for the covariance matrix. Here, we provide such a consistent estimator for W_k as

$$\hat{W}_k = n^{-1}\sum_{i=1}^n G(\hat{\theta}^T X_i) I(\hat{f}_{\hat{\theta}}(\hat{\theta}^T X_i) > c_0)\begin{pmatrix} Z_i - \hat{v}_{\hat{\theta}}(X_i) \\ d_i\{X_i - \hat{\mu}_{\hat{\theta}}(X_i)\} \end{pmatrix}$$
$$\times \begin{pmatrix} Z_i - \hat{v}_{\hat{\theta}}(X_i) \\ d_i(\hat{\theta}^T X)\{X_i - \hat{\mu}_{\hat{\theta}}(X_i)\} \end{pmatrix}^T (Y_i - \hat{\beta}^T Z_i - a_i)^k,$$

where $\hat{\mu}_{\hat{\theta}}(x) = n^{-1}\sum_{i=1}^n K_{h,i}^{\hat{\theta}}(X_i)X_i/\hat{f}_{\hat{\theta}}(\hat{\theta}^T x)$, $\hat{v}_{\hat{\theta}}(x) = n^{-1}\sum_{i=1}^n K_{h,i}^{\hat{\theta}}(X_i)Z_i/\hat{f}_{\hat{\theta}}(\hat{\theta}^T x)$ and $\hat{f}_{\hat{\theta}}(v)$ is defined above and $k = 0$ and 2.

Bandwidth selection is always an important issue for nonparametric smoothing. One of the advantages of our method is that we do not need under-smoothing the link function. Therefore, most commonly used bandwidth selection methods can be employed here. Consider estimation of g at the final step of the iterations. For a given function $w(\cdot)$ with compact support, minimizing the asymptotic weighted mean-squared error with weight $f_{\theta_0}(\cdot)w(\cdot)$ yields the optimal global

bandwidth

$$
h_0 = \left\{ \frac{\sigma^2 \int w(u)\,du \int (K(u))^2\,du}{\kappa_2^2 \int g''(u) f_{\theta_0}(u) w(u)\,du} \right\}^{1/5} n^{-1/5}.
$$

See also the discussion in Carroll et al. [4]. Both the cross-validation bandwidth selection method and the plug-in method [21] can be used to obtain bandwidths that are asymptotically consistent of h_0. Here, we give the details for the cross-validation bandwidth selection method. For any θ and β, let

$$
CV_0(b) = n^{-1} \sum_{i=1}^{n} \{Y_i - \beta^T Z_i - \tilde{a}_i^{\backslash i}\}^2 \quad \text{and} \quad CV(h) = n^{-1} \sum_{i=1}^{n} \{Y_i - \beta^T Z_i - \hat{a}_i^{\backslash i}\}^2,
$$

where $\tilde{a}_i^{\backslash i}$ and $\hat{a}_i^{\backslash i}$ are, respectively, given in

$$
\begin{pmatrix} \tilde{a}_i^{\backslash i} \\ \tilde{d}_i^{\backslash i} \end{pmatrix} = \left\{ \sum_{\substack{j=1 \\ j \neq i}}^{n} w_{ji} \begin{pmatrix} 1 \\ X_{ji}^T \theta \end{pmatrix} \begin{pmatrix} 1 \\ X_{ji}^T \theta \end{pmatrix}^T \right\}^{-1} \sum_{\substack{j=1 \\ j \neq i}}^{n} w_{ji} \begin{pmatrix} 1 \\ X_{ji}^T \theta \end{pmatrix} \{Y_j - \beta^T Z_j\}
$$

and

$$
\begin{pmatrix} \hat{a}_i^{\backslash i} \\ \hat{d}_i^{\backslash i} \end{pmatrix} = \left\{ \sum_{\substack{j=1 \\ j \neq i}}^{n} w_{ji}^{\theta} \begin{pmatrix} 1 \\ X_{ji}^T \theta \end{pmatrix} \begin{pmatrix} 1 \\ X_{ji}^T \theta \end{pmatrix}^T \right\}^{-1} \sum_{\substack{j=1 \\ j \neq i}}^{n} w_{ji}^{\theta} \begin{pmatrix} 1 \\ X_{ji}^T \theta \end{pmatrix} \{Y_j - \beta^T Z_j\}.
$$

The bandwidth in each step is then $\hat{b} = \arg\min_b CV_0(b)$ for the weight function w_{ij} and $\hat{h} = \arg\min_h CV(h)$ for the weight function w_{ji}^{θ}.

3. Numerical comparisons

In this section, we first use an example to demonstrate the relation between estimation errors and the bandwidth. We then use the examples in [4,7] to check the performance of our estimation method for finite data sets. In our simulations, kernel functions $H(x) = 3(1 - |x|^2)I(|x| < 1)/4$ and $K(u) = 3(1 - u^2)I(|u| < 1)/4$ are used. A computer code for the calculation is available at http://www.stat.nus.edu.sg/~staxyc/plsi.m

Example 3.1. Consider the following model:

$$
y_t = \beta_{01} z_{1t} + \beta_{02} z_{2t} + 2 \exp\{-3(\theta_{01} x_{t-1} + \theta_{02} x_{t-2} + \theta_{03} x_{t-3})^2\} + 0.5\varepsilon_t,
$$

where $x_t = 0.4 x_{t-1} - 0.5 x_{t-2} + u_t$ with $u_t, t = 1, 2, \ldots, \overset{\text{IID}}{\sim} Uniform(-1, 1)$; z_{1t} and $z_{2t}, t = 1, 2, \ldots$ are IID as binary distribution taking values 0 and 1 with probability 0.5 each; $\varepsilon_t, t = 1, 2, \ldots, \overset{\text{IID}}{\sim} N(0, 1)$; and that $\{u_t\}, \{z_{1t}\}, \{z_{1t}\}$ and $\{\varepsilon_t\}$ are independent series. Here, $Z_t = (z_{1t}, z_{2t})^T$ and $X_t = (x_{t-1}, x_{t-2}, x_{t-3}, \ldots, x_{t-p})^T$. The true parameters are $\beta = (\beta_{01}, \beta_{02})^T = (1, 2)^T$ and $\theta = (\theta_{01}, \theta_{02}, \theta_{03}, \ldots, \theta_{0p})^T = (-2/3, 1/3, 2/3, 0, \ldots, 0)^T$. We define the estimation errors as $e_\beta = (|\hat{\beta}_1 - \beta_{01}| + |\hat{\beta}_2 - \beta_{02}|)/2$ and $e_\theta = 1 - |\hat{\theta}^T \theta_0|$ for $\hat{\beta} = (\hat{\beta}_1, \hat{\beta}_2)^T$
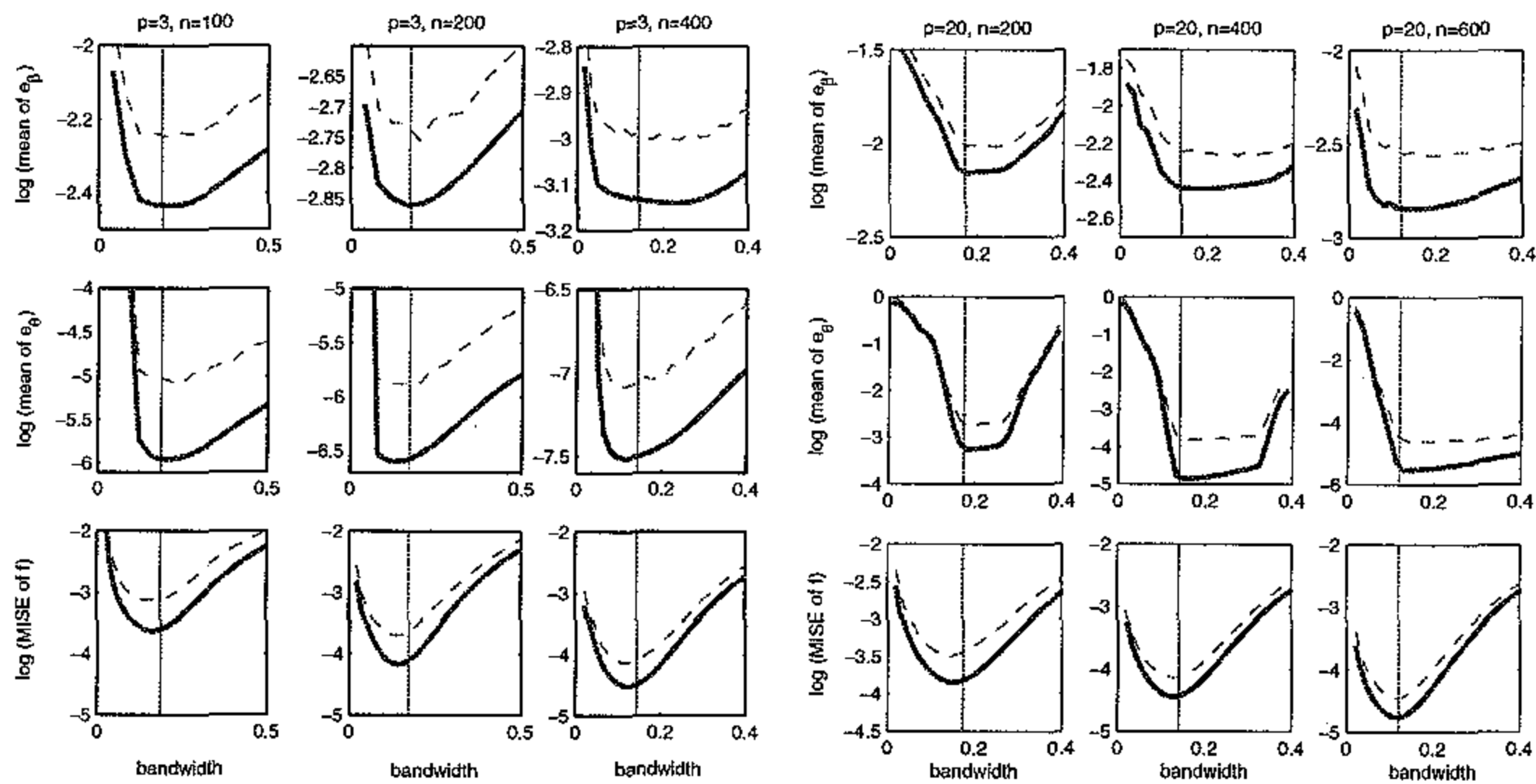
Fig. 4. Simulation results of Example 3.1. The solid lines are logarithms of the means of the estimation errors from 200 replications; the dash lines are those of the 90% quantiles of the estimation errors; the vertical lines are averages of corresponding cross-validation bandwidths.

and $\hat{\theta}$, respectively. We use the MSE for the estimation error of the link function g as $MSE = n^{-1} \sum_{i=1}^{n} \{ \hat{g}(\hat{\theta}^T X_i) - g(\theta_0^T X_i) \}^2$. With different dimension p, sample sizes and bandwidths, the logarithm of the average errors (the solid lines) and the 90% quantiles of the errors (the dash line) are shown in Fig. 4 (the number of replications is 200). The vertical lines are the corresponding average values of cross-validation bandwidths. Fig. 4 shows that the estimation procedure works quite well and stable and that the cross-validation bandwidth is applicable to the estimation of the parameters as well as the nonparametric function.

Example 3.2. Consider the following two models:

$$y = 4\{(x_1 + x_2 - 1)/\sqrt{2}\}^2 + 4 + 0.2\varepsilon, \tag{3.1}$$

$$y = \sin\{\pi((x_1 + x_2 + x_3)/\sqrt{3} - A)/(B - A)\} + \beta Z + 0.1\varepsilon, \tag{3.2}$$

where x_1, x_2, x_3 are independent uniformly distributed on $[0, 1]$, $A = 0.3912$ and $B = 1.3409$. Model (3.1) was used by Härdle et al. [7], in which $\theta_0 = (\theta_{11}, \theta_{12})^T = (1/\sqrt{2}, 1/\sqrt{2})^T$. Model (3.2) was used by Carroll et al. [4], in which $\theta_0 = (\theta_{21}, \theta_{22}, \theta_{23})^T = (1/\sqrt{3}, 1/\sqrt{3}, 1/\sqrt{3})^T$. We start the simulation for model (3.1) with $\theta = (1, 3)^T/\sqrt{10}$ and model (3.2) with $\theta = (0, 1, 2)^T/\sqrt{5}$. The cross-validation bandwidth is used. The number of replications is 100. With sample size $n = 50$, 100 and 200, respectively, the simulation results are listed in Table 1.

For model (3.1), the corresponding simulation results of $\phi = \arccos(\theta_{11})$ were 0.766(0.103), 0.792(0.084), 0.782(0.045) for $n = 50$, 100 and 200, respectively, in of Härdle et al. [7]. Our results outperform theirs. A possible reason is that minimizing the cross-validation type of residuals was used to estimate the parameters in their paper, which reduces the estimation efficiency. See [24] for details. For model (3.2), the corresponding simulation results of Carroll et al. [4] for θ_{21}, θ_{22} and θ_{23} were $(1.4e - 4)$, $(1.6e - 4)$ and $(1.3e - 4)$, respectively, when $n = 200$. Our results also improve theirs.

Table 1
Mean and mean-squared error (in parentheses) of the estimated parameters for models (3.1) and (3.2)

n	Model (3.1)			Model (3.2)			
	θ_{11}	θ_{12}	$\phi = \arccos(\theta_{11})$	θ_{21}	θ_{22}	θ_{23}	β
50	0.7117 (0.0040)	0.6965 (0.0045)	0.7746 (0.0918)	0.5793 (5.5e − 4)	0.5727 (5.7e − 4)	0.5785 (6.5e − 4)	0.2967 (1.1e − 3)
100	0.7074 (0.0015)	0.7047 (0.0015)	0.7835 (0.0541)	0.5785 (2.8e − 4)	0.5780 (2.6e − 4)	0.5748 (2.2e − 4)	0.2972 (4.7e − 4)
200	0.7071 (0.0008)	0.7059 (0.0008)	0.7845 (0.0403)	0.5776 (1.2e − 4)	0.5770 (1.3e − 4)	0.5772 (1.2e − 4)	0.2992 (2.5e − 4)

4. Real data analysis

Now we return to our real data sets in Section 1. The Epanechnikov kernel and the cross-validation bandwidths are used in the calculations below.

4.1. Credit scoring

We partition x_4 and x_5 into 20 intervals separately. All the observations in a combination of one interval of x_4 and one interval of x_5 form a group. In all the groups, the proportion $\hat{p}_i$, of customers who paid their installments without problem, and $y_i = \log(\hat{p}_i/(1 - \hat{p}_i))$ are calculated. We consider model (1.3) with all the covariates by taking $Z = (x_2, x_3, x_6, x_7, x_8, x_9, x_{10}, x_{11}, x_{12}, x_{13}, x_{14}, x_{15}, x_{16}, x_{17}, x_{18}, x_{19}, x_{20}, x_{21}, x_{22}, x_{23}, x_{24})^T$, $X = (x_4, x_5)^T$. Because $x_{10}, \ldots, x_{24}$ are categorical variables, they are denoted by dummy variables. Here, x_4 and x_5 are standardized, respectively, for ease of calculations. For simplicity, we assume $E(\varepsilon^2|X, Z) = \sigma^2$ is a constant. Applying the estimation procedure to the data set, we obtain the estimates of the parameters as listed in Table 2. See [17] for more explanations. The estimate of the unknown function is shown in the right panel of Fig. 5. The nonlinearity in x_4 and x_5, i.e. $\hat{\theta}^T X = 0.249x_4 + 0.969x_5$, is clear as shown in Fig. 5.

Based on the estimation results as shown in Fig. 5, we have the following conclusion. Recall that x_4 is the amount of loan. Fig. 2 in Section 1 shows that as the loan increases, the costumer's ability to pay the installment increases, i.e. y decreases. This seems misleading. For age x_5, Fig. 2 suggests that there is a range for the age at which the costumer has less problem to pay their loan. In comparison, the estimated model can give us a more reasonable explanation. Fig. 5 suggests that as the age and the loan, $\hat{\theta}^T X = 0.249x_4 + 0.969x_5$, increases, the ability to pay the installments decreases. However, there is an age range, in which the costumer has high ability to pay the loan.

4.2. Circulatory and respiratory (CR) problems in Hong Kong

Due to the hospital booking system, the day-of-the-week can affect y_t. We use dummy variables to describe the day of the tth observation by a 6-dimension vector $(D_{t1}, \ldots, D_{t6})$, where $D_{tk} = 1$ if the observation is taken on the kth day of a week; 0 otherwise. Together with lagged variables of pollutants and weather conditions in 1 week, we take $Z_t = (D_{t1}, \ldots, D_{t6}, x_{1,t-1}, \ldots, x_{1,t-7}, x_{2,t-1}, \ldots, x_{2,t-7}, x_{3,t-1}, \ldots, x_{3,t-7}, x_{4,t-1}, \ldots, x_{4,t-7})^T$ and $X_t = (x_{5,t-1}, \ldots, x_{5,t-7}, x_{6,t-1},$
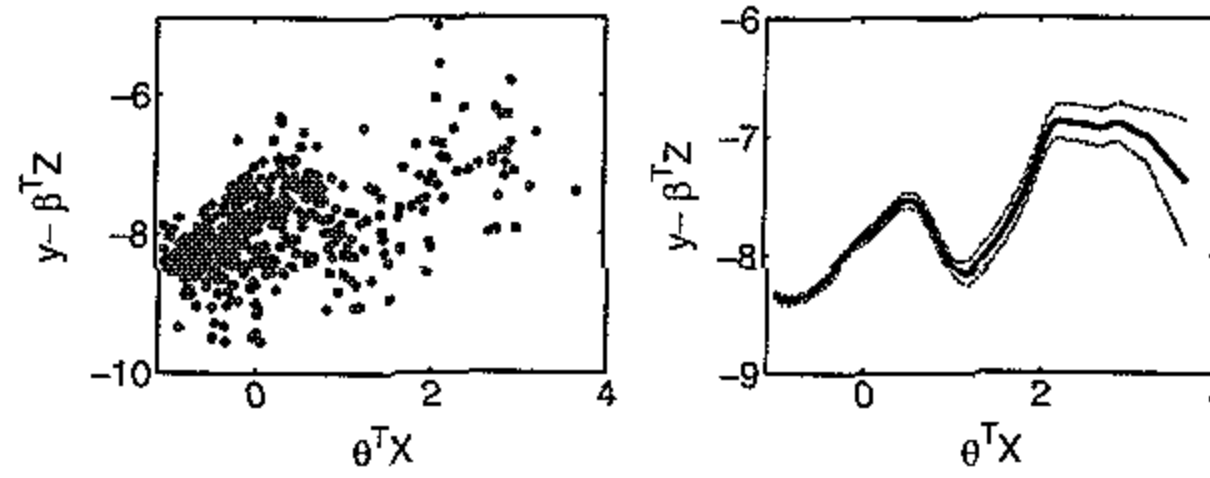
Fig. 5. Estimation results of the credit scoring data. The left panel is $y - \hat{\beta}^T Z$ plotted against $\hat{\theta}^T X$. The right panel is the estimated g and 95% symmetric pointwise confidence interval.

Table 2
Estimation results of the credit scoring data

Variable	Coeff.	SE	Variable	Coeff.	SE	Variable	Coeff.	SE
2	0.159	0.163	17#2	**−1.718**	0.472	20#3	−0.082	0.294
3	0.021	0.114	17#3	**−1.211**	0.433	20#4	0.263	0.251
6	−0.109	0.105	17#4	**1.977**	0.576	21#2	**−2.194**	0.683
7	**−0.454**	0.119	17#5	**4.715**	0.932	21#3	**−1.490**	0.363
8	0.189	0.120	17#6	**−1.327**	0.316	22#2	−1.102	0.582
9	0.032	0.091	18#2	**2.145**	0.528	22#3	−0.785	0.490
10#2	**0.817**	0.302	18#3	**1.037**	0.413	22#4	0.753	0.715
11#2	0.188	0.293	18#4	0.878	0.447	22#5	0.770	0.584
12#2	**0.635**	0.303	18#5	**1.756**	0.359	22#6	**−3.837**	0.957
13#2	**−0.815**	0.276	18#6	**1.876**	0.449	22#7	**2.253**	0.608
14#2	**1.680**	0.544	18#7	**1.770**	0.551	22#8	0.838	0.531
15#2	**1.416**	0.347	19#2	0.416	0.369	22#9	**1.441**	0.526
15#3	**2.411**	0.469	19#3	**1.287**	0.307	22#10	−1.519	1.199
15#4	**3.247**	0.520	19#4	−0.966	0.539	22#11	−0.644	0.510
15#5	**2.782**	0.617	19#5	1.343	0.673	23#2	0.087	0.350
15#6	**0.987**	0.374	19#6	**1.691**	0.465	23#3	0.787	0.499
16#2	0.214	0.476	19#7	0.992	0.539	24#2	**1.717**	0.612
16#3	0.680	0.431	19#8	**−1.170**	0.566	4	**0.249**	0.026
16#4	**1.714**	0.539	19#9	0.173	0.608	5	**0.969**	0.007
16#5	**1.218**	0.442	19#10	1.070	0.348			
16#6	**1.588**	0.465	20#2	**2.021**	0.539	$\hat{\sigma}^2 = 0.1589$		

The bold coeff. are statistically significant with "t-value" greater than 2.0.

$\ldots, x_{6,t-7})^T$ in model (1.1). Here, $x_{1,t}, \ldots, x_{6,t}$ are standardized. We further assume that $E(\varepsilon_t^2 \mid X_t, Z_t) = \sigma^2$ is a constant. By the asymptotic distribution of the parameters, we remove those covariates with smallest t-values in the estimated model one by one and re-estimate the model. Continue this procedure until all the covariates have t-values larger than 1.8. We finally obtain the following model (the values in the parentheses are the corresponding standard errors of the estimators):

$$y_t = -0.3831 D_{t1} - 0.1728 D_{t2} - 0.5636 D_{t3} - 0.7399 D_{t4} - 1.0871 D_{t5} - 1.1562 D_{t6}$$
$$\quad (0.0942) \quad\quad (0.0943) \quad\quad (0.0945) \quad\quad (0.0947) \quad\quad (0.0946) \quad\quad (0.0942)$$
$$\quad + 0.0957 x_{2,t-1} + \hat{g}(0.4257 x_{5,t-2} - 0.6079 x_{5,t-5} + 0.4974 x_{6,t-4} + 0.4492 x_{6,t-7}).$$
$$\quad (0.0287) \quad\quad (0.1576) \quad\quad (0.1104) \quad\quad (0.1745) \quad\quad (0.1475)$$

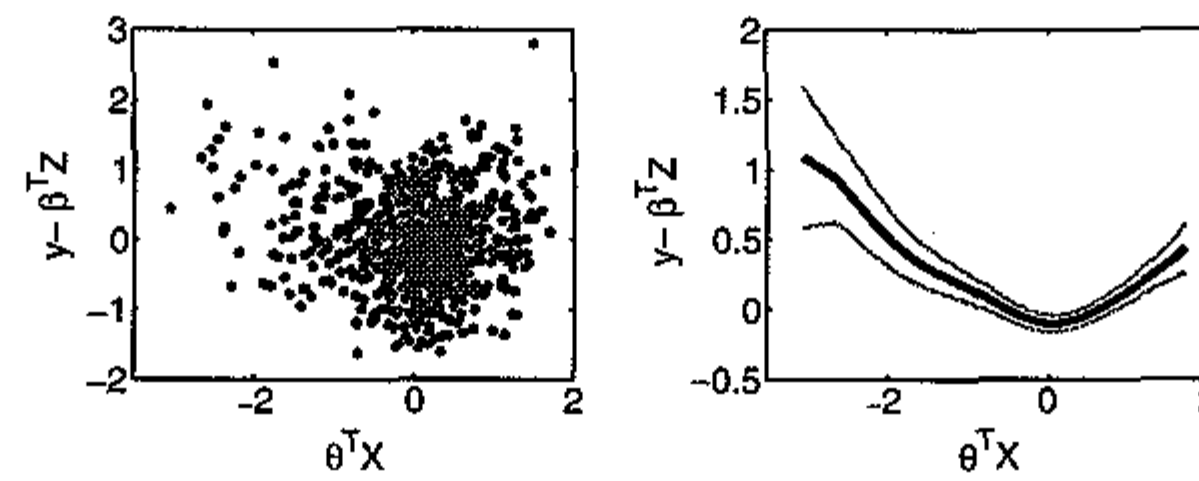The estimated link function $\hat{g}$ is shown in Fig. 6.

Fig. 6. Estimation results for the circulatory and respiratory problems in Hong Kong. The left panel is $y_t - \hat{\beta}^T Z_t$ plotted against $\hat{\theta}^T X_t$. The right panel is the estimated g and 95% symmetric pointwise confidence interval.

Based on this model, the effects of weather conditions on the CR problems are as follows. The coefficients of temperatures $x_{5,t-2}$ and $x_{5,t-5}$ forms a contrast. Together with Fig. 6, it suggests that a rapid temperature variation (rather than the temperature itself) will increase the hospital admission y_t. The coefficients of humidity $x_{6,t-4}$ and $x_{6,t-7}$ have about the same value, which can be taken as an average. Together with Fig. 6, it suggests that extreme dry or wet weather will increase the hospital admission in Hong Kong.

5. Conclusion

This paper proposes an estimation method for the partially linear single-index model. Good properties of the method has been demonstrated theoretically and by simulations. These properties are: (1) a $\sqrt{n}$-consistent pilot estimator is not needed, see Theorem 1. This solves the problems addressed in [4]; (2) convergence of the PLSIM algorithm is proved, see the proof of Theorem 1 in Section 6; (3) a wide range of bandwidth can be used including the cross-validated bandwidth. This makes the algorithm stable and frees it from the audible critique on "the necessity of uncontrollable hyperparameters"; (4) under normal noise assumptions, the estimators of the parameters are asymptotically efficient in semi-parametric sense, see [4]; and (5) the PLSIM algorithm is applicable to time series. Of course, there are some limitations to the method. Because the estimators are based on minimizing the sum of residual squares, they are not the most efficient in the semi-parametric sense when the noise is not normally distributed. Extending the approach in this paper to more complicated models such as the generalized partially linear single-index model given in [4], the simple algorithm proposed in this paper cannot be used.

Some important problems for the model and the estimation method need to be investigated in further work. These problems include variable selection and model checking. However, since the estimation has been simplified in this paper to solving simple regression problems as in (2.5) and (2.6). It is possible that those problems can be solved based on the methods in this paper.

6. Assumptions and proofs

Let $U = (X^T, Z^T)^T$. Suppose $\{(U_i, y_i), i = 1, \ldots, n\}$ is a set of observations. We make the following assumptions on the stochastic nature of the observations, the link function and the kernel functions.

(C1) The observations are a strongly mixing and stationary sequence with geometric decaying mixing rate $\alpha(k)$.

(C2) With Probability 1, X lies in a compact set $\mathcal{D}$; the marginal density functions f of X and f_θ of $\theta^T X$ for any $|\theta| = 1$ have bounded derivatives; regions $\{x : f(x) \geqslant c_0\}$ and $\{x : f_\theta(\theta^T x) > c_0\}$ for all $\theta : |\theta| = 1$ are nonempty.

(C3) For any perpendicular unit norm vectors θ and ϑ, the joint density function $f(u_1, u_2)$ of $(\theta^T X, \vartheta^T X)$ satisfies $f(u_1, u_2) < c f_{\theta^T X}(u_1) f_{\vartheta^T X}(u_2)$, where c is a constant.

(C4) g has bounded, continuous third-order derivative; the conditional expectations $E(Z|X = x)$, $E(ZZ^T|X = x)$, $E(U|\theta^T X = v)$ and $E(UU^T|\theta^T X = v)$ have bounded derivatives; $E(y^r|X = x)$, $E(|Z|^r|X = x)$, $E(|Z_\ell||Z_1||X_1 = x_1, X_\ell = x_\ell)$ and $E(|Z_\ell||Z_1| \big| \theta^T X_1$

$= u, \theta^T X_\ell = v)$ are bounded by a constant for all $\ell > 0$, x_1, x_ℓ, x, u and v, where $r > 3$.

(C5) H is a density function with bounded derivative and compact support $\{|x| \leqslant a_0\}$ for some $a_0 > 0$; K is a symmetric density function with bounded derivative and compact support $[-b_0, b_0]$ for some $b_0 > 0$ and that the Fourier transform of K is absolute integrable.

(C6) Matrix $E\{(Z - E(Z|X))(Z - E(Z|X))^T\}$ is a positive definite matrix.

The mixing rate [3] in (C1) can be relaxed to algebraic rate $\alpha(k) = O(k^{-p})$. Suppose the bandwidth $h \sim n^{-\delta}$. Then the mixing rate satisfying the following equation is sufficient:

$$\sum_{n=1}^{\infty} n^{-\{\frac{1}{2} - \frac{1}{r} - \delta(\frac{1}{2} + \frac{1}{r})\}p + 2p + 1 + \frac{1}{r} + (\frac{1}{2} + \frac{1}{r})\delta}(\log n)^{p/2} < \infty.$$

The regions with positive densities in (C2) are needed to avoid zero values of the denominator of kernel estimator of regression. There are different approaches for this purpose. See, e.g. [7,9,15]. However, their ideas are the same. We can further assume that c_0 decreases to 0 with n at a slow speed, but it makes no difference in practices. Assumption (C3) ensures successful searching for the direction θ globally. If we restrict the searching area, the assumption can be removed; see [7]. The third-order derivatives in (C4) is needed for higher-order expansion. Actually, existence of second-order derivative is sufficient for the root-n consistency. In this paper, we only employ kernel functions with compact support as in (C5). We further assume that $\kappa_2 \stackrel{\text{def}}{=} \int K(v)v^2 = 1$ and $\mathcal{H}_2 \stackrel{\text{def}}{=} \int H(U)UU^T \, dU = I_{p \times p}$; otherwise we may take $K(v) =: K(v/\sqrt{\kappa_2})/\sqrt{\kappa_2}$ and $H(U) =: H(\mathcal{H}_2^{-1/2} U)(det(\mathcal{H}_2))^{-1/2}$. (C6) is imposed for identification. Similarly, if we search for the direction θ in a small neighbor of θ_0 as in [4,7], (C6) can be removed.

6.1. Proof of the main results

The basic tools are given in Lemmas A.1–A.3. Some simple calculation results are listed in Lemmas 6.1–6.3. Based on these lemmas, Lemma 1 and Theorem 1 are proved. Let $\delta_\theta = |\theta - \theta_0|$, $\delta_\beta = |\beta - \beta_0|$ and $\delta_\gamma = \delta_\theta + \delta_\beta$. In a bounded parameter space, δ_θ, δ_β and δ_γ are bounded. Let $\delta_{pn} = \{\log n/(nb^p)\}^{1/2}$, $\tau_{pn} = b^2 + \delta_{pn}$, $\delta_n = \{\log n/(nh)\}^{1/2}$, $\tau_n = h^2 + \delta_n$ and $\delta_{0n} = (\log n/n)^{1/2}$. By the condition $h \sim n^{-\delta}$ with $1/6 < \delta < 1/4$, we have $\delta_{0n} \ll h^2 \ll h^{-1}\delta_n$ and $\delta_n \ll h$. We shall use these relations frequently in our calculations. Let $\Theta = \{\theta : |\theta| = 1\}$. Suppose A_n is a matrix. $A_n = O(a_n)$ (or $o(a_n)$) means every element in A_n is $O(a_n)$ (or $o(a_n)$) almost surely. We adopt the consistency in the sense of "almost surely" because we need to prove the convergence of the algorithm, which theoretically need infinite iterations. Let $c, c_1, c_2, \ldots$ be a set of constants. For ease of exposition, c may have different values at different places. We abbreviate $K_h(\theta^T X_{i0})$ and $H_b(X_{i0})$ as $K^\theta_{h,i}(x)$ (or $K^\theta_{h,i}$) and $H_{b,i}(x)$ (or $H_{b,i}$), respectively, in the following context. We take $G(\cdot) \equiv 1$ in the proofs for simplicity.

In the following context, we abbreviate L for any function $L(x)$, and L_θ or $L_\theta(x)$ for any function $L_\theta(\theta^T x)$. Let v_θ and μ_θ be defined as in Section 2, and

$$v = E(Z|X = x), \quad \pi = E(ZZ^T|X = x), \quad \pi_\theta = E(ZZ^T|\theta^T X = \theta^T x),$$
$$\tilde{\Sigma}_\theta = E(XX^T|\theta^T X = \theta^T x) - \mu_\theta x^T - x\mu_\theta^T + xx^T.$$

Let

$$\varsigma_0 = \frac{1}{n}\sum_{i=1}^{n} H_{b,i}, \quad S_1 = \frac{1}{n}\sum_{i=1}^{n} H_{b,i}X_{i0}, \quad S_2 = \frac{1}{n}\sum_{i=1}^{n} H_{b,i}X_{i0}X_{i0}^T,$$

$$T_1 = \frac{1}{n}\sum_{i=1}^{n} H_{b,i}Z_i, \quad T_2 = \frac{1}{n}\sum_{i=1}^{n} H_{b,i}Z_iZ_i^T, \quad C_2 = \frac{1}{n}\sum_{i=1}^{n} H_{b,i}X_{i0}Z_i^T,$$

$$E_1 = \frac{1}{n}\sum_{i=1}^{n} H_{b,i}Z_iy_i, \quad D_1 = \frac{1}{n}\sum_{i=1}^{n} H_{b,i}X_{i0}y_i, \quad W_n = \varsigma_0 S_2 - S_1 S_1^T$$

and

$$\bar{w}_{a,i}^\theta(x) = \{\theta^T S_2\theta\}H_{b,i} - \theta^T S_1 H_{b,i}\theta^T X_{i0}, \quad \bar{w}_{d,i}^\theta(x) = \varsigma_0 H_{b,i}\theta^T X_{i0} - \theta^T S_1 H_{b,i}.$$

Based on (2.4), we can obtain initial estimators of θ_0 and β_0 as follows. Choose a vector θ with norm 1 and any vector β. Let $\bar{w}_j^\theta = \theta^T W_n(X_j)\theta$ and calculate

$$\bar{a}_j^\theta = \{\bar{w}_j^\theta\}^{-1}\sum_{i=1}^{n} \bar{w}_{a,i}^\theta(X_j)\{y_i - \beta^T Z_i\}, \quad \bar{d}_j^\theta = \{\bar{w}_j^\theta\}^{-1}\sum_{i=1}^{n} \bar{w}_{d,i}^\theta(X_j)\{y_i - \beta^T Z_i\}, \quad (6.1)$$

$$\begin{pmatrix} \bar{\beta} \\ \bar{\theta} \end{pmatrix} = \{\bar{D}_n^\theta\}^- \sum_{j=1}^{n} I_n(X_j) \begin{pmatrix} E_1(X_j) - \bar{a}_j^\theta T_1(X_j) \\ \bar{d}_j^\theta D_1(X_j) - \bar{a}_j^\theta \bar{d}_j^\theta S_1(X_j) \end{pmatrix} \Big/ \varsigma_0(X_j),$$

$$\bar{\theta} := sgn_1 \bar{\theta}/|\bar{\theta}|, \quad (6.2)$$

where sgn_1 is the sign of first entry in $\bar{\theta}$ and

$$\bar{D}_n^\theta = \sum_{j=1}^{n} I_n(X_j) \begin{pmatrix} T_2(X_j) & \bar{d}_j^\theta C_2(X_j) \\ \bar{d}_j^\theta C_2^T(X_j) & (\bar{d}_j^\theta)^2 S_2(X_j) \end{pmatrix} \Big/ \varsigma_0(X_j)$$

and A^- denotes the Moore–Penrose inverse of matrix A. Repeat the calculations in (6.1) and (6.2) with (θ, β) replaced by $(\bar{\theta}, \bar{\beta})$ until convergence. Denote the final value by $(\tilde{\beta}, \tilde{\theta})$. Next, we shall improve the efficiency of the estimators using a univariate kernel. Let

$$\varsigma_k^\theta = \frac{1}{n}\sum_{i=1}^{n} K_{h,i}^\theta\{\theta^T X_{i0}\}^k, \quad k = 0, 1, 2, 3,$$

$$w_{a,i}^\theta = \varsigma_2^\theta K_{h,i}^\theta - \varsigma_1^\theta K_{h,i}^\theta \theta^T X_{i0}, \quad w_{d,i}^\theta = \varsigma_0^\theta K_{h,i}^\theta \theta^T X_{i0} - \varsigma_1^\theta K_{h,i}^\theta,$$

$$w^\theta = \frac{1}{n}\sum_{i=1}^{n} w_{a,i}^\theta, \quad S_1^\theta = \frac{1}{n}\sum_{i=1}^{n} K_{h,i}^\theta X_{i0}, \quad S_2^\theta = \frac{1}{n}\sum_{i=1}^{n} K_{h,i}^\theta X_{i0}X_{i0}^T,$$

$$T_1^\theta = \frac{1}{n} \sum_{i=1}^{n} K_{h,i}^\theta Z_i, \quad E_1^\theta = \frac{1}{n} \sum_{i=1}^{n} K_{h,i}^\theta Z_i y_i, \quad D_1^\theta = \frac{1}{n} \sum_{i=1}^{n} K_{h,i}^\theta X_{i0} y_i,$$

$$T_2^\theta = \frac{1}{n} \sum_{i=1}^{n} K_{h,i}^\theta Z_i Z_i^T, \quad C_2^\theta = \frac{1}{n} \sum_{i=1}^{n} K_{h,i}^\theta \theta^T X_{i0} Z_i^T,$$

$$S_{1,1}^\theta = \frac{1}{n} \sum_{i=1}^{n} K_{h,i}^\theta \{\theta^T X_{i0}\} X_{i0}, \quad S_{2,1}^\theta = \frac{1}{n} \sum_{i=1}^{n} K_{h,i}^\theta \{\theta^T X_{i0}\}^2 X_{i0},$$

$$S_{1,2}^\theta = \frac{1}{n} \sum_{i=1}^{n} K_{h,i}^\theta \{\theta^T X_{i0}\} X_{i0} X_{i0}^T, \quad S_3^\theta = \frac{1}{n} \sum_{i=1}^{n} K_{h,i}^\theta X_{i0} \{(\theta - \theta_0)^T X_{i0}\}^2.$$

Based on (2.4), we improve the estimators $\tilde\theta$ and $\tilde\beta$ as follows. Let $w_j^\theta = w^\theta(X_j)$. Starting with $(\theta, \beta) = (\tilde\theta, \tilde\beta)$, calculate

$$\tilde a_j^\theta = (w_j^\theta)^{-1} \sum_{i=1}^{n} w_{a,i}^\theta(X_j)\{y_i - \beta^T Z_i\}, \quad \tilde d_j^\theta = (w_j^\theta)^{-1} \sum_{i=1}^{n} w_{d,i}^\theta(X_j)\{y_i - \beta^T Z_i\}, \quad (6.3)$$

$$\begin{pmatrix} \tilde\beta \\ \tilde\theta \end{pmatrix} = (\tilde D_n^\theta)^- \sum_{j=1}^{n} I_n(X_j) \begin{pmatrix} E_1^\theta(X_j) - \tilde a_j^\theta T_1^\theta(X_j) \\ \tilde d_j^\theta D_1^\theta(X_j) - \tilde a_j^\theta \tilde d_j^\theta S_1^\theta(X_j) \end{pmatrix} \Big/ \varsigma_0^\theta(X_j),$$

$$\tilde\theta := sgn_1 \, \tilde\theta / |\tilde\theta|, \tag{6.4}$$

where sgn_1 is the sign of first entry of $\tilde\theta$ and

$$\tilde D_n^\theta = \sum_{j=1}^{n} I_n(X_j) \begin{pmatrix} T_2^\theta(X_j) & \tilde d_j^\theta C_2^\theta(X_j) \\ \tilde d_j^\theta \{C_2^\theta(X_j)\}^T & (\tilde d_j^\theta)^2 S_2^\theta(X_j) \end{pmatrix} \Big/ \varsigma_0^\theta(X_j).$$

Repeat the procedure (6.3) and (6.4) with (θ, β) replaced by $(\tilde\theta, \tilde\beta)$ until convergence. Denote the final value by $(\hat\beta, \hat\theta)$.

Let $\bar\Delta_i(x) = y_i - \bar a - \beta_0^T Z_i - \bar d X_{i0}^T \theta_0$ and $\tilde\Delta_i^\theta(x) = y_i - \tilde a - \beta_0^T Z_i - \tilde d X_{i0}^T \theta_0$. We have

$$\begin{pmatrix} \bar\beta \\ \bar\theta \end{pmatrix} = \begin{pmatrix} \beta_0 \\ \theta_0 \end{pmatrix} + \tilde D_n^-(\theta) \sum_{j=1}^{n} I_n(X_j) \sum_{i=1}^{n} H_{b,i}(X_j) \begin{pmatrix} Z_i \\ X_{ij}\bar d_j \end{pmatrix} \bar\Delta_i(X_j)/\varsigma_0(X_j), \tag{6.5}$$

$$\begin{pmatrix} \bar\beta \\ \tilde\theta \end{pmatrix} = \begin{pmatrix} \beta_0 \\ \theta_0 \end{pmatrix} + \tilde D_n^-(\theta) \sum_{j=1}^{n} I_n(X_j) \sum_{i=1}^{n} K_{h,i}^\theta(\theta^T X_j) \begin{pmatrix} Z_i \\ X_{ij}\tilde d_j \end{pmatrix} \tilde\Delta_i^\theta(X_j)/\varsigma_0^\theta(\theta^T X_j). \tag{6.6}$$

By Lemmas A.1 and A.3, we have

$$\varsigma_0 = f(x) + O(J_0 + \tau_{pn}), \quad S_1 = O(bJ_0 + b\tau_{pn}),$$

$$S_2 = f(x) I_{p \times p} b^2 + O(b^2 J_0 + b^2 \tau_{pn}), \quad T_1 = f(x)v(x) + O(J_0 + \tau_{pn}),$$

$$T_2 = f(x)\pi(x) + O(J_0 + \tau_{pn}), \quad \frac{1}{n} \sum_{i=1}^{n} H_{b,i} Z_i \varepsilon_i = O(\delta_{pn}),$$

$$\frac{1}{n}\sum_{i=1}^{n} H_{b,i}\varepsilon_i = O(\delta_{pn}), \quad \frac{1}{n}\sum_{i=1}^{n} H_{b,i}X_{i0}\{\theta^T X_{i0}\}^k\varepsilon_i = O(b^{k+1}\delta_{pn}),$$

$$\frac{1}{n}\sum_{i=1}^{n} H_{b,i}|X_{i0}|^k = O(b^k), \quad C_2 = O(bJ_0 + b^2 + b\delta_n), \tag{6.7}$$

and

$$\varsigma_0^\theta = f_\theta + O(J_\theta + \tau_n), \quad \varsigma_1^\theta = O(hJ_\theta + h^2 + h\delta_n), \quad \varsigma_2^\theta = f_\theta h^2 + O(h^2 J_\theta + h^2\tau_n),$$

$$\varsigma_3^\theta = O(h^4 + b^3 J_\theta + h^3\delta_n), \quad S_1^\theta = f_\theta\{\mu_\theta - x\} + O(J_\theta + \tau_n), \quad S_2^\theta = \tilde{\Sigma}_\theta f_\theta + O(J_\theta + \tau_n),$$

$$w^\theta = f_\theta^2 h^2 + O(h^2 J_\theta + h^2\tau_n), \quad T_1^\theta = f_\theta v_\theta + O(J_\theta + \tau_n), \quad T_2^\theta = f_\theta\pi_\theta + O(J_\theta + \tau_n),$$

$$C_2^\theta = O(hJ_\theta + h^2 + h\tau_n), \quad S_{1,1}^\theta = O(hJ_\theta + h^2 + h\tau_n), \quad S_{1,2}^\theta = O(hJ_\theta + h^2 + h\tau_n),$$

$$S_{2,1}^\theta = f_\theta\{\mu_\theta - x\}h^2 + O(h^2 J_\theta + h^2\tau_n), \quad S_3^\theta = O(\delta_\theta^2). \tag{6.8}$$

Let $\bar{a}, \bar{d}, \tilde{a}$ and $\tilde{d}$ be, respectively, the values of $\bar{a}_j, \bar{d}_j, \tilde{a}_j$ and $\tilde{d}_j$ with X_j replaced by x. For simplicity, we further assume that $f(x) > c_0$ and $f_\theta(\theta^T x) > c_0$ for all $x \in \mathcal{D}$ (otherwise, we only need to change $\mathcal{D}$ to $\{x : f(x) > c_0\}$ or $\{x : f_\theta(\theta^T x) > c_0\}$ in the proofs). Thus, $I_n(X_j) \equiv 1$ when n is sufficiently large.

Lemma 6.1. *Let* $\beta_d = \beta_0 - \beta$ *and* $\theta_d = \theta_0 - \theta$. *Suppose assumptions* (C1)–(C5) *hold. We have*

$$\bar{a} = g(\theta_0^T x) + v^T\beta_d + O(J_0 + b + \delta_{pn}),$$

$$\bar{d} = \theta^T\theta_0 g'(\theta_0^T x) + O\{(1 + b^{-1}J_0)\delta_\beta + b^{-1}\delta_{pn} + b\},$$

$$\tilde{a} = g(\theta_0^T x) + g'(\theta_0^T x)\{\mu_\theta - x\}^T\theta_d + v_\theta^T\beta_d + \tfrac{1}{2}g''(\theta_0^T x)h^2 + R_{n,3}$$
$$\quad + O(\delta_\theta^2 + J_\theta\delta_\gamma + \tau_n\delta_\gamma + h\tau_n),$$

$$\tilde{d} = g'(\theta_0^T x) + h^{-1}R_{n,4} + O\{\delta_\theta^2 + (h^{-1}J_\theta + 1 + h^{-1}\delta_n)\delta_\gamma + \tau_n\}$$

uniformly for $x \in \mathcal{D}$ *and* $\theta \in \Theta$, *where* $R_{n,3} = \{nf_\theta\}^{-1}\sum_{i=1}^{n} K_{h,i}^\theta\varepsilon_i$ *and* $R_{n,4} = \{nhf_\theta\}^{-1}\sum_{i=1}^{n} K_{h,i}^\theta\theta^T X_{i0}\varepsilon_i$.

Lemma 6.2. *Suppose assumptions* (C1)–(C5) *hold. We have*

$$\frac{1}{n}\bar{D}_n(\theta)$$
$$= \begin{pmatrix} E(ZZ^T) + O(b + \delta_{pn}) & O(b^2 + b\delta_{pn}) \\ O(b^2 + b\delta_{pn}) & (\theta^T\theta_0)^2 E\{g'(\theta_0^T X)\}^2 I_{p\times p} b^2 + O(b\delta_{pn} + b^2\delta_\beta) \end{pmatrix}$$

and

$$\frac{1}{n}\tilde{D}_n(\theta) = \begin{pmatrix} E(ZZ^T) & \tilde{C}_{12} \\ \tilde{C}_{12}^T & 2\tilde{W}_0 \end{pmatrix} + O(h^{-1}\delta_n + \delta_\gamma)$$

uniformly for $\theta \in \Theta$, *where* $\tilde{C}_{12} = E\{g'(\theta_0^T X)Z(\mu_{\theta_0}(X) - X)^T\}$ *and* $\tilde{W}_0 = E[\{g'(\theta_0^T X)\}^2 \{X - \mu_{\theta_0}(X)\}\{X - \mu_{\theta_0}(X)\}^T]$.

Lemma 6.3. *Suppose assumptions* (C1)–(C5) *hold. Then*

$$\frac{1}{n^2}\sum_{j=1}^{n}\sum_{i=1}^{n}H_{b,i}(X_j)Z_i\bar{\Delta}_i(X_j)/\varsigma_0(X_j) = E\{v(X)v^T(X)\}(\beta - \beta_0) + O(b + \delta_{pn}),$$

$$\frac{1}{n^2}\sum_{j=1}^{n}\bar{d}_j\sum_{i=1}^{n}H_{b,i}(X_j)X_{ij}\bar{\Delta}_i(X_j)/\varsigma_0(X_j) = b^2(\theta^T\theta_0)(1 - \theta^T\theta_0)E\{g'(\theta_0^T X)\}^2\theta_0$$
$$+ O(b^3 + b\delta_{pn} + b^2\delta_\beta),$$

$$\frac{1}{n^2}\sum_{j=1}^{n}\sum_{i=1}^{n}K_{h,i}^{\theta}(X_j)Z_i\tilde{\Delta}_i^{\theta}(X_j)/\varsigma_0^{\theta}(\theta^T X_j) = E\{v_\theta(X)v_\theta^T(X)\}\beta_d + \frac{1}{n}\sum_{i=1}^{n}\{Z_i - v_\theta(X_i)\}\varepsilon_i$$
$$+ O\{(\delta_\theta + h + h^{-1}\delta_n)\delta_\gamma + h\tau_n\},$$

$$\frac{1}{n^2}\sum_{j=1}^{n}\tilde{d}_j\sum_{i=1}^{n}K_{h,i}^{\theta}(X_j)X_{ij}\tilde{\Delta}_i^{\theta}(X_j)/\varsigma_0^{\theta}(\theta^T X_j)$$

$$= \tilde{W}_0\theta_d + \frac{1}{n}\sum_{i=1}^{n}g'(\theta_0^T X_i)\{\mu_{\theta_0}(X_j) - X_i\}\varepsilon_i$$
$$+ O\{(\delta_\gamma + h^{-1}\delta_n + h)\delta_\gamma + h\tau_n + h^{-1}\delta_n^2\},$$

uniformly for $\theta \in \Theta$.

Proof of Lemma 1. We shall prove that the equations in the Lemma 1 hold with probability 1. From Lemmas 6.2 and 6.3 and (6.5), we have for any β and θ with $\theta^T\theta = 1$,

$$\bar{\beta} - \beta_0 = \{E(ZZ^T)\}^{-1}E\{v(Z)v^T(Z)\}(\beta - \beta_0) + O(b + b^{-1}\delta_{pn}). \tag{6.9}$$

Note that the above equation does not depend on the choice of θ. This is because we use a multivariate kernel, i.e. we use a more general multivariate function to replace $g(\theta_0^T x)$. In the algorithm, (6.9) can be written as

$$\bar{\beta}_{k+1} - \beta_0 = \{E(ZZ^T)\}^{-1}E(v(X)v^T(X))(\bar{\beta}_k - \beta_0) + O(b + b^{-1}\delta_{pn}), \tag{6.10}$$

where the sub-index k indicates that the corresponding values are the results of the kth iteration in the algorithm; see (6.1) and (6.2). By assumption (C6), $E(ZZ^T) - E\{v(X)v^T(X)\}$ is a positive definite matrix. Note that $E\{v(X)v^T(X)\}$ is a semi-positive matrix. Hence the eigenvalues of $\{E(ZZ^T)\}^{-1}E\{v(X)v^T(X)\}$ are all less than 1. After sufficiently many steps, we have from (6.10)

$$\bar{\beta}_k - \beta_0 = O(b + b^{-1}\delta_{pn}). \tag{6.11}$$

See the proof of Theorem 1 below for more details. Therefore

$$\tilde{\beta} - \beta_0 = O(b + b^{-1}\delta_{pn}). \tag{6.12}$$

If $\theta^T\theta_0 \neq 0$, then it follows from Lemmas 6.2 and 6.3 and (6.5) that

$$\bar{\theta} - \theta_0 = (\theta^T\theta_0)^{-1}(1 - \theta^T\theta_0)\theta_0 + O(\delta_\beta + b + b^{-1}\delta_{pn}),$$

i.e. $\bar{\theta} = (\theta^T \theta_0)^{-1}\theta_0 + O(\delta_\beta + b + b^{-1}\delta_{pn})$. By (6.12), we may assume δ_β is small enough (otherwise, take $\beta = \tilde{\beta}$). Thus

$$\tilde{\theta} := sgn_1 \bar{\theta}/|\bar{\theta}| = \theta_0 + O(\delta_\beta + b + b^{-1}\delta_{pn}),$$

where sgn_1 is the sign of first entry of $\bar{\theta}$. In the algorithm, we have

$$\tilde{\theta}_{k+1} - \theta_0 = O(\delta_{\tilde{\beta}_k} + b + b^{-1}\delta_{pn}). \tag{6.13}$$

Combining (6.11) and (6.13), we have,

$$\tilde{\theta} - \theta_0 = O(b + b^{-1}\delta_{pn}). \tag{6.14}$$

The proof is completed. $\square$

Proof of Theorem 1. It follows from Lemmas 6.2 and 6.3 and (6.6) that

$$\begin{pmatrix} \tilde{\beta} - \beta_0 \\ \tilde{\theta} - \theta_0 \end{pmatrix} = \tilde{D}^- N_n + \tilde{D}^- \tilde{C} \begin{pmatrix} \beta - \beta_0 \\ \theta - \theta_0 \end{pmatrix}$$

$$+ O\{(\delta_\gamma + h + h^{-1}\delta_n)\delta_\gamma + h\tau_n + h^{-1}\delta_n^2\}, \tag{6.15}$$

where

$$\tilde{C} = \begin{pmatrix} E\{v_{\theta_0}(X)v_{\theta_0}^T(X)\} & 0 \\ 0 & \tilde{W}_0 \end{pmatrix}, \qquad \tilde{D} = \begin{pmatrix} E(ZZ^T) & \tilde{C}_{12} \\ \tilde{C}_{12}^T & 2\tilde{W}_0 \end{pmatrix},$$

$$N_n = \frac{1}{n}\sum_{i=1}^{n} \begin{pmatrix} Z_i - v(X_i|\theta_0) \\ g'(\theta_0^T X_i)\{\mu_{\theta_0}(X_j) - X_i\} \end{pmatrix} \varepsilon_i.$$

Following the proof of Lemma 1 of Xia et al. [24], we have $\tilde{C}$, $\tilde{D}$ and $W_0 = \tilde{D} - \tilde{C}$ are all semi-positive matrices with rank $p + q - 1$. Therefore, $D \overset{def}{=} (\tilde{D}^-)^{1/2}\tilde{C}(\tilde{D}^-)^{1/2}$ is a semi-positive matrix with all eigenvalues less than 1. There exist $1 > \lambda_1 \geqslant \lambda_2 \geqslant \cdots \geqslant \lambda_{p+q-1} > 0$ and orthogonal matrix Γ such that

$$D = \Gamma diag(\lambda_1, \ldots, \lambda_{p+q-1}, 0)\Gamma^T.$$

Let $(\tilde{\beta}_k, \tilde{\theta}_k)$ be the calculation results of the kth iteration in the algorithm; see (6.3) and (6.4). For any k, Eq. (6.15) holds with $(\tilde{\beta}, \tilde{\theta})$ replaced by $(\tilde{\beta}_{k+1}, \tilde{\theta}_{k+1})$ and (β, θ) by $(\tilde{\beta}_k, \tilde{\theta}_k)$. Let $\tilde{\gamma}_k = \tilde{D}^{1/2}(\tilde{\beta}_k^T - \beta_0^T, \tilde{\theta}_k^T - \theta_0^T)^T$, we have

$$\tilde{\gamma}_{k+1} = (\tilde{D}^-)^{1/2}N_n + D\tilde{\gamma}_k + O\{(\delta_{\tilde{\gamma}_k} + h + h^{-1}\delta_n)\delta_{\tilde{\gamma}_k} + h\tau_n + h^{-1}\delta_n^2\}. \tag{6.16}$$

It follows that

$$\delta_{\tilde{\gamma}_{k+1}} \leqslant \delta_{0n} + \lambda_1\delta_{\tilde{\gamma}_k} + c(\delta_{\tilde{\gamma}_k} + h + h^{-1}\delta_n)\delta_{\tilde{\gamma}_k} + c(h\tau_n + h^{-1}\delta_n^2)$$

$$= \delta_{0n} + \{\lambda_1 + c\delta_{\tilde{\gamma}_k} + c(h + h^{-1}\delta_n)\}\delta_{\tilde{\gamma}_k} + c(h\tau_n + h^{-1}\delta_n^2) \tag{6.17}$$

almost surely, where c is a constant. We can further take $c > 1$. Because $h, h^{-1}\delta_n, \tau_n, \delta_{0n} \to 0$ as $n \to \infty$, we may assume that

$$c(h + h^{-1}\delta_n) \leqslant (1 - \lambda_1)/3, \qquad \delta_{0n} + c(h\tau_n + h^{-1}\delta_n^2) \leqslant (1 - \lambda_1)^2/(9c). \tag{6.18}$$

By (6.12) and (6.14), we may assume

$$\delta_{\tilde\gamma_1} \leqslant (1-\lambda_1)/(3c). \tag{6.19}$$

Therefore, it follows from (6.17)–(6.19) that

$$\delta_{\tilde\gamma_2} \leqslant \{\lambda_1 + 2(1-\lambda_1)/3\}(1-\lambda_1)/(3c) + (1-\lambda_1)^2/(9c) = (1-\lambda_1)/(3c). \tag{6.20}$$

From (6.17), (6.18) and (6.20), we have

$$\delta_{\tilde\gamma_3} \leqslant (1-\lambda_1)/(3c).$$

Consequently, $\delta_{\tilde\gamma_k} \leqslant (1-\lambda_1)/(3c)$ for all k. Therefore we have from (6.17) that

$$\delta_{\tilde\gamma_{k+1}} \leqslant \lambda_0 \delta_{\tilde\gamma_k} + \delta_{0n} + c(h\tau_n + h^{-1}\delta_n^2)$$

almost surely, where $0 \leqslant \lambda_0 < (2+\lambda_1)/3 < 1$. It follows that

$$\delta_{\tilde\gamma_k} \leqslant \lambda_0^k \delta_{\tilde\gamma_1} + \{\delta_{0n} + c(h\tau_n + h^{-1}\delta_n^2)\} \sum_{j=1}^{\infty} \lambda_0^j = O(\delta_{0n} + h\tau_n + h^{-1}\delta_n^2)$$

for sufficiently large k. By (6.16), we have

$$\tilde D^{1/2} \begin{pmatrix} \hat\beta - \beta_0 \\ \hat\theta - \theta_0 \end{pmatrix} = (\tilde D^-)^{1/2} N_n + D\tilde D^{1/2} \begin{pmatrix} \hat\beta - \beta_0 \\ \hat\theta - \theta_0 \end{pmatrix} + O(\delta_{0n}^2 + h\tau_n + h^{-1}\delta_n^2)$$

$$= (\tilde D^-)^{1/2} N_n + D\tilde D^{1/2} \begin{pmatrix} \hat\beta - \beta_0 \\ \hat\theta - \theta_0 \end{pmatrix} + o(n^{-1/2}). \tag{6.21}$$

The facts that $n^{1/2}h^3 \to 0$ and $n^{1/2}h^{-1}\delta_n^2 \to 0$ are used in the last step above. It follows from (6.21) that

$$(\tilde D - \tilde D^{1/2} D\tilde D^{1/2}) \begin{pmatrix} \hat\beta - \beta_0 \\ \hat\theta - \theta_0 \end{pmatrix} = N_n + o(n^{-1/2})$$

or

$$W_0 \begin{pmatrix} \hat\beta - \beta_0 \\ \hat\theta - \theta_0 \end{pmatrix} = N_n + o(n^{-1/2}).$$

The first part of Theorem 1 follows from the above equation and the central limiting theorem of dependent data, see e.g. [19]. The second part follows immediately from the first part and Theorem 1 of Carroll et al. [4]. □

6.2. Proofs of Lemmas 6.1–6.3

In this subsection, we first give some basic lemmas. Based on these basic lemmas, the proofs of Lemmas 6.1–6.3 are simple algebraic calculations. We will state the main ideas of the calculations without going into the details. However, the details can be obtained on request from the authors. It can also be downloaded from http://www.stat.nus.edu.sg/~staxyc/plsi.pdf

Lemma A.1. *Suppose that $m_1(\theta, x, z)$ and $\varphi(x, z, v)$ are measurable functions with $\sup_{\theta \in \Theta} E$ $\{|m_1(\theta, X, Z)|^r\} < \infty$ for some $r > 2$ and $\sup_{x,z} |m_1(\theta, x, z) - m_1(\theta_0, x, z)| < c|\theta - \theta_0|$. Let $\varphi_i = \varphi(X_i, Z_i, y_i)$. Assume $\sup_{\theta \in \Theta, v} E(|\varphi_i|^r \,|\, \theta^T X = v) < \infty$ and $\sup_{\theta \in \Theta, u, v} E(|\varphi_i \varphi_1| \,|\, \theta^T X_1 = u, \theta^T X_i = v) < c$ for all $i > 1$. Let $g(v)$ be any function with continuous second-order derivative, $m(u, v) = g(u) - g(v) - g'(v)(u - v) - g''(v)(u - v)^2/2$ and $\zeta_i^{k, \ell} = m(\theta_0^T X_i, \theta_0^T x) z_i^k (\theta^T X_{i0})^\ell$, where z_i is any component of Z_i, $k = 0, 1$ and $\ell = 0, 1$. If (C1) hold, then*

$$\sup_{\theta \in \Theta} \left| \frac{1}{n} \sum_{i=1}^{n} \{ m_1(\theta, X_i, Z_i) - E m_1(\theta, X_i, Z_i) \} \right| = O(\delta_{0n}),$$

$$\sup_{|\theta - \theta_0| < a_n} \left| \frac{1}{n} \sum_{i=1}^{n} \{ m_1(\theta, X_i, Z_i) - m_1(\theta_0, X_i, Z_i) \} \varepsilon_i \right| = O(a_n \delta_{0n}),$$

where $a_n \to 0$ as $n \to \infty$. If further (C2)–(C5) hold, $h \sim n^{-\delta}$ with $0 < \delta < 1 - 2/r$, then

$$\sup_{x \in \mathcal{D}} \left| \frac{1}{n} \sum_{i=1}^{n} \{ H_{b,i} \varphi_i - E(H_{b,i} \varphi_i) \} \right| = O(\delta_{pn}),$$

$$\sup_{\substack{\theta \in \Theta \\ x \in \mathcal{D}}} \left| \frac{1}{n} \sum_{i=1}^{n} \{ K_{h,i}^\theta \varphi_i - E(K_{h,i}^\theta \varphi_i) \} \right| = O(\delta_n),$$

$$\sup_{\substack{|\theta - \theta_0| < a_n \\ x \in \mathcal{D}}} \left| \frac{1}{n} \sum_{i=1}^{n} \{ K_{h,i}^\theta \zeta_i^{k, \ell} - E(K_{h,i}^\theta \zeta_i^{k, \ell}) \} \right| = O\{\delta_n h^\ell (a_n^2 + h^2)\}.$$

The proof of Lemma A.1 is quite standard. The details can be found in [8,16,23].

Lemma A.2. *Let φ_i be defined in Lemma A.1 and $f(x, z, y)$ be the density function of (X, Z, y). If (C1) and (C5) hold, then*

$$\sup_{\theta \in \Theta} \left| \frac{1}{n^2} \sum_{i=1}^{n} \sum_{j=1}^{n} \left\{ K_{h,i}^\theta(X_j) \varphi_j - \int K_{h,i}^\theta(x) \varphi(x, z, y) f(x, z, y) \, dx \, dz \, dy \right\} \varepsilon_i \right| = O(\delta_n^2).$$

Proof. Let $\Delta_n(\theta)$ be the value in the absolute symbols. By the continuity of $K_{h,i}^\theta$ in θ, there are $n_1 < c n^{2p}$ points $\theta_{n,1}, \ldots, \theta_{n,n_1}$ in Θ such that $\cup_{k=1}^{n_1} \{\theta : |\theta - \theta_{n,k}| < h^2 \delta_n^2\} \supset \Theta$ and

$$\max_{1 \leqslant k \leqslant n_1} \sup_{|\theta - \theta_{n,k}| < h^2 \delta_n^2} \left| \Delta_n(\theta) - \Delta_n(\theta_{n,k}) \right| = O(\delta_n^2). \tag{6.22}$$

The Fourier transform $\phi(s) = \int \exp(isv) K(v) \, dv$ will be used in the following, where i is the imaginary unit. Thus $K(v) = \int \exp(-isv) \phi(s) \, ds$. We have

$$\Delta_n(\theta_{n,k}) = \frac{1}{n^2} h^{-1} \sum_{j=1}^{n} \sum_{i=1}^{n} \int \left[\exp\{-is\theta_{n,k}^T X_{ij}/h\} \varphi_j \right.$$

$$\left. - \int \exp\{-is\theta_{n,k}^T X_{i0}/h\} \varphi(x, z, y) f(x, z, y) \, dx \, dz \, dy \right] \phi(s) \, ds \, \varepsilon_i$$

$$= h^{-1} \int \frac{1}{n} \sum_{i=1}^{n} \exp(-\mathrm{i}s\theta_{n,k}^{T} X_i / h)\varepsilon_i \cdot \frac{1}{n} \sum_{j=1}^{n} \left[\exp(\mathrm{i}s\theta_{n,k}^{T} X_j / h)\varphi_j \right.$$

$$\left. - \int \exp(\mathrm{i}s\theta_{n,k}^{T} x / h)\varphi(x, z, y) f(x, z, y)\, dx\, dz\, dy \right] \phi(s)\, ds.$$

Similar to the proof of Lemma A.1, we have

$$\max_{1 \leqslant k \leqslant n_1} \left| \frac{1}{n} \sum_{i=1}^{n} \exp(-\mathrm{i}s\theta_{n,k}^{T} X_i / h)\varepsilon_i \right| \leqslant c_8 \delta_{0n},$$

$$\max_{1 \leqslant k \leqslant n_1} \left| \frac{1}{n} \sum_{j=1}^{n} \left[\exp(\mathrm{i}s\theta_{n,k}^{T} X_j / h)\varphi_j - \int \exp(\mathrm{i}s\theta_{n,k}^{T} x / h)\varphi(x, z, y) f(x, z, y)\, dx\, dz\, dy \right] \right|$$

$$\leqslant c_9 \delta_{0n}$$

almost surely, where c_8 and c_9 are constants which do not depend on s. Hence

$$\max_{1 \leqslant k \leqslant n_1} \left| \Delta_n(\theta_{n,k}) \right| \leqslant h^{-1} \int c_8 \delta_{0n} c_9 \delta_{0n} |\phi(s)|\, ds = O(h^{-1} \delta_{0n}^2) = O(\delta_n^2). \tag{6.23}$$

Note that

$$\sup_{\theta \in \Theta} |\Delta_n(\theta)| \leqslant \max_{1 \leqslant k \leqslant n_1} \left| \Delta_n(\theta_{n,k}) \right| + \max_{1 \leqslant k \leqslant n_1} \sup_{|\theta - \theta_{n,k}| < h^2 \delta_n^2} \left| \Delta_n(\theta) - \Delta_n(\theta_{n,k}) \right|. \tag{6.24}$$

Therefore, the second part of Lemma A.2 follows from (6.22) to (6.24). $\square$

For easy of exposition, we abuse $\mathcal{D}$ as the positive support of the $f(x)$. Let $d(x, \mathcal{D}^c) = \min_{x' \in \mathbb{R}^p - \mathcal{D}} |x - x'|$ and define bounded functions $J_0(x)$, $J_\theta(v)$ such that $J_0(x) = 0$ if $d(x, \mathbb{R}^p - \mathcal{D}) > a_0 b$ and $J_\theta(\theta^T x) = 0$ if $d(\theta^T x, \theta^T (\mathbb{R}^p - \mathcal{D})) > b_0 h$. By the definition, we have

$$\frac{1}{n} \sum_{j=1}^{n} J_0(X_j) = O(b), \qquad \frac{1}{n} \sum_{j=1}^{n} J_\theta(X_j) = O(h). \tag{6.25}$$

To cope with the boundary points, we give the following nonuniform rate of convergency.

Lemma A.3. *Suppose assumptions* (C3) *and* (C5) *hold. Then*

$$EH_b(X - x)\{\theta^T (X - x)/b\}^k \{\vartheta^T (X - x)/b\}^\ell = v_{k,\ell}^{\theta,\vartheta} f(x) + J_0(x) + O(h),$$

$$EK_h(\theta^T (X - x))\{\theta^T (X - x)/h\}^\ell = \tau_\ell f_\theta(\theta^T x) + J_\theta(x) + O(h)$$

uniformly for $\theta, \vartheta \in \Theta$ and $x \in \mathcal{D}$, where $v_{k,\ell}^{\theta,\vartheta} = \int_{\mathbb{R}^p} H(U)(\theta^T U)^k (\vartheta^T U)^\ell\, dU$ and $\tau_\ell = \int K(u) u^\ell\, du$.

Proof. We here only give the details for the first part. If $d(x, \mathcal{D}^c) > a_0 b$, we define $J_0(x) = 0$. From (C5), we have

$$\int_{\mathcal{D}} H_b(U - x)\{\theta^T (U - x)/b\}^k \{\vartheta^T (U - x)/b\}^\ell f(U)\, dU$$

$$= \int_{\mathbb{R}^p} H(U)\{\theta^T U\}^k \{\theta^T U\}^\ell f(x + hU)\, dU = v_{k,\ell}^{\theta,\vartheta} f(x) + O(h).$$

If $d(x, \mathcal{D}^c) < a_0 b$, we have by (C3)

$$J_0(x) \overset{\text{def}}{=} \int_{\mathcal{D}} H_b(U - x)\{\theta^T(U - x)/b\}^k \{\vartheta^T(U - x)/b\}^\ell f(U)\, dU$$

$$\leqslant \int_{\mathbb{R}^p} H(U)\{\theta^T U\}^k \{\vartheta^T U\}^\ell f(x + hU)\, dU = O(1).$$

Therefore, the first part of Lemma A.3 follows. $\square$

Outline of Proofs of Lemmas 6.1–6.3. The proofs are actually algebraic calculations based on Lemmas A.1–A.3. In the proofs, we need to apply Taylor expansions to the model at $\theta_0^T x$ and have

$$y_i = \beta_0^T Z_i + g(\theta_0^T x) + g'(\theta_0^T x)\theta_0^T X_{i0} + \frac{1}{2}g''(\theta_0^T x)\{\theta_0^T X_{i0}\}^2 + O(|\theta_0^T X_{i0}|^3) + \varepsilon_i$$

$$= \beta_0^T Z_i + g(\theta_0^T x) + g'(\theta_0^T x)\theta_0^T X_{i0} + \frac{1}{2}g''(\theta_0^T x)\{\theta^T X_{i0}\}^2$$

$$+ O(|\theta^T X_{i0}|^3 + \delta_\theta|\theta^T X_{i0}| + \delta_\theta^2) + \varepsilon_i.$$

We then handle each term in the expansions separately. For this purpose, we need to handle terms of summations like

$$\frac{1}{n}\sum_{i=1}^n K_h(\theta^T X_{i0})\mathbf{z}_i = E\{K_h(\theta^T X_{i0})\mathbf{z}_i\}$$

$$+\frac{1}{n}\sum_{i=1}^n \left[K_h(\theta^T X_{i0})\mathbf{z}_i - E\{K_h(\theta^T X_{i0})\mathbf{z}_i\}\right] \tag{6.26}$$

and

$$\frac{1}{n^2}\sum_{i,j=1}^n K_h(\theta^T X_{ij})\mathbf{z}_i = E\{K_h(\theta^T X_{ij})\mathbf{z}_i\}$$

$$+\frac{1}{n}\sum_{i,j=1}^n \left[K_h(\theta^T X_{ij})\mathbf{z}_i - E\{K_h(\theta^T X_{ij})\mathbf{z}_i\}\right], \tag{6.27}$$

where $\mathbf{z}_i$ is defined in Lemma A.1. For (6.26), we can apply Lemma A.1 to the second term on the right-hand side and Lemma A.3 to the first term. Similarly, we can handle the first and the second terms on the right-hand side of (6.27) by Lemmas A.3 and A.2, respectively.

Acknowledgements

References

[1] G. Arminger, D. Euache, T. Bonne, Analyzing credit risk data: a comparison of logistic discrimination, classification tree analysis, and feed forward subroles, Comput. Statist. 12 (1997) 293–310.

[2] D.M. Bates, D.G. Watts, Nonlinear Regression Analysis and Its Applications, Wiley, New York, 1988.

[3] R.C. Bradley, Basic properties of strong mixing conditions, in: E. Eberlein, M.S. Taqqu (Eds.), Dependence in Probability and Statistics: A Survey of Recent Results, Birkhauser, Boston, 1986, pp. 165–192.

[4] R.J. Carroll, J. Fan, I. Gijbels, M.P. Wand, Generalized partially linear single-index models, J. Amer. Statist. Assoc. 92 (1997) 477–489.

[5] J. Fan, I. Gijbels, Local Polynomial Modeling and Its Applications, Chapman & Hall, London, 1996.

[6] P. Hall, On projection pursuit regression, Ann. Statist. 17 (1989) 573–588.

[7] W. Härdle, P. Hall, H. Ichimura, Optimal smoothing in single-index models, Ann. Statist. 21 (1993) 157–178.

[8] W. Härdle, P. Janssen, R. Serfling, Strong uniform consistency rates for estimators of conditional functionals, Ann. Statist. 16 (1988) 1428–1449.

[9] W. Härdle, T.M. Stoker, Investigating smooth multiple regression by method of average derivatives, J. Amer. Statist. Assoc. 84 (1989) 986–995.

[10] W.E. Henley, D.J. Hand, A k-nearest neighbour classifier for assessing consumer credit risk, The Statistician 45 (1996) 77–95.

[11] M. Hristache, A. Juditsky, J. Polzehl, V. Spokoiny, Structure adaptive approach for dimension reduction, Ann. Statist. 29 (2001) 593–627.

[12] M. Hristache, A. Juditsky, V. Spokoiny, Direct estimation of the single-index coefficients in single-index models, Ann. Statist. 29 (2001) 1537–1566.

[13] H. Ichimura, L. Lee, Semiparametric least squares estimation of multiple index models: single equation estimation, in: W. Barnett, J. Powell, G. Tauchen (Eds.), Nonparametric and Semiparametric Methods in Econometrics and Statistics, Cambridge University Press, Cambridge, 1991.

[14] K.C. Li, Sliced inverse regression for dimension reduction (with discussion), Amer. Statist. Assoc. 86 (1991) 316–342.

[15] O. Linton, Second order approximation in the partially linear regression model, Econometrica 63 (1995) 1079–1112.

[16] E. Masry, Multivariate local polynomial regression for time series: uniform strong consistency and rates, J. Time Ser. Anal. 17 (1996) 571–599.

[17] M. Müller, B. Rönz, Credit scoring using semiparametric methods, in: J. Franke, W. Härdle, G. Stahl (Eds.), Measuring Risk in Complex Stochastic Systems, Springer Lecture Notes in Statistics, vol. 147, Springer, Berlin, 2000, pp. 85–102.

[19] E. Rio, The functional law of the iterated logarithm for stationary strongly mixing sequences, Ann. Probab. 23 (1995) 1188–1203.

[20] P.M. Robinson, Root-N-Consistent semiparametric regression, Econometrica 56 (1988) 931–954.

[21] D. Ruppert, J. Sheather, P.M. Wand, An effective bandwidth selector for local least squares regression, J. Amer. Statist. Assoc. 90 (1995) 1257–1270.

[22] T.A. Severini, I.G. Staniswalis, Quasi-likelihood estimation in semiparametric models, J. Amer. Statist. Assoc. 89 (1994) 501–511.

[23] Y. Xia, W.K. Li, On single-index coefficient regression models, J. Amer. Statist. Assoc. 94 (1999) 1275–1285.

[24] Y. Xia, H. Tong, W.K. Li, On extended partially linear single-index models, Biometrika 86 (1999) 831–842.

[25] Y. Xia, H. Tong, W.K. Li, L. Zhu, An adaptive estimation of dimension reduction space (with discussions), J. Roy. Statist. Soc. B. 64 (2002) 1–28.

[26] Y. Yu, D. Ruppert, Penalized spline estimation for partially linear single-index models, J. Amer. Statist. Assoc. 97 (2002) 1042–1054.

Estimation and Testing for Varying Coefficients in Additive Models with Marginal Integration

Lijian YANG   Byeong U. PARK   Lan XUE   Wolfgang HÄRDLE

September 6, 2005

Abstract

We propose marginal integration estimation and testing methods for the coefficients of varying coefficient multivariate regression model. Asymptotic distribution theory is developed for the estimation method which enjoys the same rate of convergence as univariate function estimation. For the test statistic, asymptotic normal theory is established. These theoretical results are derived under the fairly general conditions of absolute regularity (β-mixing). Application of the test procedure to the West German real GNP data reveals that a partially linear varying coefficient model is best parsimonious in fitting the data dynamics, a fact that is also confirmed with residual diagnostics.

KEY WORDS: Equivalent kernels; German real GNP; Local polynomial; Marginal integration; Rate of convergence

1 INTRODUCTION

Parametric regression analysis usually assumes that the response variable Y depends linearly on a vector $\mathbf{X}$ of predictor variables. More flexible non- and semi- parametric regression models allow the dependence to be of more general nonlinear forms. On the other hand, the appeal of simplicity and interpretation still motivates search for models that are nonparametric in nature but have special features that are appropriate for the data involved. Such are additive models (Chen and Tsay 1993a, Linton and Nielsen 1995, Masry and Tjøstheim 1995, 1997, Mammen, Linton and Nielsen 1999, Sperlich, Tjøstheim and Yang 2002), generalized additive models (Linton and Härdle 1996), partially linear models (Härdle, Liang and Gao 2000), etc.

In this paper, we consider a form of flexible nonparametric regression model proposed by Hastie and Tibshirani (1993). The following model

$$Y_i = m(\mathbf{X}_i, \mathbf{T}_i) + \sigma(\mathbf{X}_i, \mathbf{T}_i)\varepsilon_i, \ i = 1, ..., n \qquad (1)$$

where $\{\varepsilon_i\}_{i \geq 1}$ are i.i.d. white noise, each ε_i independent of $(\mathbf{X}_i, \mathbf{T}_i)$ where

$$\mathbf{X}_i = (X_{i1}, ..., X_{id})^T, \mathbf{T}_i = (T_{i1}, ..., T_{id})^T, \qquad (2)$$

is called a *varying-coefficient model* if

$$m(\mathbf{X}_i, \mathbf{T}_i) = \sum_{s=1}^{d} f_s(X_{is})T_{is}. \qquad \textbf{(Model I)}$$

In Model I, all the variables $\{X_s\}_{s=1}^{d}$ are different from each other. The model with all the variables $\{X_s\}_{s=1}^{d}$ being the same, i.e., $m(X_i, \mathbf{T}_i) = \sum_{s=1}^{d} f_s(X_i)T_{is}$, is the *functional coefficient model* of Chen and Tsay (1993b) with univariate coefficient functions. The latter is different from Model I and was fully discussed by Cai, Fan and Li (2000) and Cai, Fan and Yao (2000). Indeed, Hastie and Tibshirani (1993) fitted real data examples exclusively with the functional coefficient model. Although the name varying-coefficient model was used by Cai, Fan and Li (2000), the model they studied was the same model proposed by Chen and Tsay (1993b), except with the additional feature of a possibly non-trivial link function. Cai, Fan and Li (2000) used local maximum likelihood estimation for all coefficient functions $\{f_s\}_{s=1}^{d}$, whose computing was no more than a univariate estimation, due to the fact that all these univariate functions depend on the same variable X. The estimation method proposed for the functional coefficient model does not apply for Model I.

For Model I, the only existing estimation method was the backfitting method of Hastie and Tibshirani (1993), which has not been theoretically justified. Intuitively, inference about model (1) is no more complex than that of univariate models. In this paper, we develop a marginal integration type estimator for each varying coefficient $\{f_s\}_{s=1}^{d}$ in the case when each varying coefficient can have a different variable. Our method achieves the optimal rate of convergence for univariate function estimation, and has a simple asymptotic theory for the estimators.

As an illustration of the effectiveness of Model I, we consider a real time series data $\{Y_t\}_{t=1}^{n}$ on West German GNP in Section 5. After taking first difference and de-seasonalization, the data is considered strictly stationary, as shown by the dotted curve in Figure 4. The varying coefficient models $Y_t = f_1(Y_{t-1})Y_{t-2} + f_2(Y_{t-3})Y_{t-4} + (\text{noise})$ and $Y_t = f_1(Y_{t-3})Y_{t-2} + f_2(Y_{t-1})Y_{t-4} + (\text{noise})$ are fitted and the estimates of the functions f_1 and f_2 are plotted in Figure 2. These varying coefficient AR models have 2.81 and 2.46 times, respectively, more prediction power than the simple linear AR model. See Table 3 to find $0.00059/0.00021 = 2.81$ and $0.00059/0.00024 = 2.46$. More details about the data and the modelling procedures are found in Section 5.

Model I may be viewed as a special case of a functional coefficient model with *multivariate* coefficient functions $m(\mathbf{X}_i, \mathbf{T}_i) = \sum_{s=1}^{d} g_s(\mathbf{X}_i)T_{is}$, where $g_s(\mathbf{X}_i) = f_s(X_{is})$ for $s = 1, \ldots, d$. In this respect, it would be of interest to compare Model I with some related FAR (functional coefficient autoregressive) models. For example, for the varying coefficient model $Y_t = f_1(Y_{t-3})Y_{t-1} + f_2(Y_{t-4})Y_{t-2} + (\text{noise})$, one may consider the following FAR model for a comparison: $Y_t = f_1(Y_{t-3}, Y_{t-4})Y_{t-1} + f_2(Y_{t-3}, Y_{t-4})Y_{t-2} + (\text{noise})$. In a simulation study that is presented in Section 4, we find that the mean average squared residuals and the mean average squared prediction errors of the FAR model are larger than those of the varying coefficient model. More details on the simulation results are found in Section 4.2.

Of another special practical interests is the model that allows some of the X_s's to be the same. For this we consider the following generalization of Model I:

$$m(\mathbf{X}_i, \mathbf{T}_i) = \sum_{s=1}^{d_0} \sum_{u=1}^{r_s} f_{su}(X_{is})T_{isu}, \qquad \text{(Model II)}$$

where now the coefficient functions $f_{s1}, \ldots, f_{sr_s}$ depends on the same variable X_s. In Model II the dimension of $\mathbf{X}$ is d_0 which is less than $d = \sum_{s=1}^{d_0} r_s$, the dimension of $\mathbf{T}$, and all the variables $\{X_s\}_{s=1}^{d_0}$ are different from each other. An advantage of Model II is that it alleviates the dimensionality problem that the marginal integration method may have in fitting Model I. Furthermore, the functional coefficient model of Chen and Tsay (1993b) is a special case

3

of Model II where $d_0 = 1$. As an example of Model II, one may have

$$Y_t = c + a_1(r_t)M_t + a_2(r_t)M_t^2 + a_3(r_t)M_t^2 I_{\{M_t<0\}} + b_1(t)\tau_t + b_2(t)\tau_t^2 + \varepsilon_t, \quad t = 1, ..., n$$

in which Y_t denotes the implied volatility, r_t the interest rate, M_t the moneyness, and τ_t the maturity at time t.

Although our models consist of additive bivariate functions, they are linear in the variables $T_s(T_{su})$. One interesting question one may ask is: are some of the coefficient functions $f_s(f_{su})$ constant? If the answer is yes for some but not all, then the model is partially linear in some variables $T_s(T_{su})$; if the answer is yes to all, then the model is the classical linear regression model. Any constant $f_s(f_{su})$ can then be estimated at $1/\sqrt{n}$-rate of convergence. A formal testing procedure is proposed in Section 3 for determining the constancy of coefficient functions $f_s(f_{su})$. For the German GNP data, it is found that f_1 can be set to a constant, while f_2 can not.

We organize the paper as the follows. In Section 2, we describe marginal estimation methods for Models I and II, and derive asymptotic distribution theory of the estimators. In Section 3, a test procedure is proposed to test the hypothesis that $f_s(f_{su})$ is a constant. In Section 4 we illustrate the finite sample properties of our proposals in the estimation and testing problems. In Section 5, we apply our estimation and testing methods to the West German real GNP data. All technical assumptions and proofs are in the Appendix.

2 ESTIMATION OF VARYING COEFFICIENTS

2.1 Model I

In this section we formulate local polynomial integration estimators of the coefficient functions $\{f_s\}_{s=1}^d$ in Model I. For general background on the local polynomial method, see Stone (1977), Katkovnik (1979), Ruppert and Wand (1994), Wand and Jones (1995) and Fan and Gijbels (1996).

We assume that each ε_i is independent of the vectors $\{(\mathbf{X}_j, \mathbf{T}_j)\}_{j=1,...,i}$ for each $i = 1, ..., n$. This is sufficient for obtaining our main results on distribution theory as we assume $\{(\mathbf{X}_j, \mathbf{T}_j)\}_{j=1,...,n}$ to be strictly stationary and geometrically β-mixing in assumption A2 (see appendix.), but weaker than the usual assumption that each ε_i is independent of the vectors $\{(\mathbf{X}_j, \mathbf{T}_j)\}_{j=1,...,n}$.

Note that if there exists nontrivial linear dependence among the variables T_s with corresponding functions of X_s as coefficients, then functions f_s are unidentifiable. To be precise,

4

suppose that

$$\sum_{s=1}^{d} r_s(X_{is})T_{is} = 0, a.s.$$

for some nonzero measurable functions r_s, then the regression function m in Model I equals

$$\sum_{s=1}^{d} \{f_s(X_{is}) + r_s(X_{is})\} T_{is}$$

as well. Hence for identifiability, we assume that

$$\sum_{s=1}^{d} r_s(X_{is})T_{is} = 0 \ a.s. \implies r_s(x) \equiv 0, \ s = 1, ..., d. \tag{3}$$

The condition (3) may be considered as an analogue of linear independence between covariates in linear models. It is a sufficient condition of avoiding *concurvity* as termed by Hastie and Tibshirani (1990). The term *concurvity* in addtive models is understood as an analogue of *collinearity* in linear models. The condition is closely related to the invertibility of the matrix $\mathbf{Z}_s^T \mathbf{W}_s(\mathbf{X}_{-s})\mathbf{Z}_s$ to be defined below, see Subsection A.2 of the Appendix for more details.

Now Let $\mathbf{x} = (x_1, ..., x_d)^T \in I\!\!R^d$ be a point where we want to estimate the functions $\{f_s\}_{s=1}^{d}$. We denote by $(\mathbf{X}, \mathbf{T}) = (X_1, ..., X_d, T_1, ..., T_d)$ a generic random vector having the same distribution as $(\mathbf{X}_i, \mathbf{T}_i) = (X_{i1}, ..., X_{id}, T_{i1}, ..., T_{id})$, and define $\mathbf{X}_{-s}$ and $\mathbf{T}_{-s}$, as obtained from $\mathbf{X}$ and $\mathbf{T}$ by removing the s-th components, by

$$\mathbf{X}_{-s} = (X_1, .., X_{s-1}, X_{s+1} ..., X_d)^T, \ s = 1, ..., d,$$
$$\mathbf{T}_{-s} = (T_1, .., T_{s-1}, T_{s+1} ..., T_d)^T, \ s = 1, ..., d.$$

For a kernel function K we write $K_h(u) = K(u/h)/h$. We fit p-th order local polynomials to estimate the varying coefficients. Write $\mathbf{Y} = (Y_i)_{1 \leq i \leq n}$ and denote $p(u) = (1, u, ..., u^p)^T$. Define $\mathbf{Z}_s$ be the $n \times (p+d)$ matrix which has $\left(p\left\{(X_{is} - x_s)/h\right\}^T T_{is}, \mathbf{T}_{i,-s}^T\right)$ as its i-th row. Let $\mathbf{W}_s(\mathbf{x}_{-s}) \equiv \mathbf{W}_s(x_s, \mathbf{x}_{-s})$ be the $n \times n$ diagonal matrix defined by

$$\mathbf{W}_s(\mathbf{x}_{-s}) = \text{diag}\left\{K_h(X_{js} - x_s)L_g(\mathbf{X}_{j,-s} - \mathbf{x}_{-s})/n\right\}_{1 \leq j \leq n}$$

where $L_g(\mathbf{u}) = (g_1 \cdots g_{s-1}g_{s+1} \cdots g_d)^{-1} L(g_1^{-1}u_1, ..., g_{s-1}^{-1}u_{s-1}, g_{s+1}^{-1}u_{s+1}, ..., g_d^{-1}u_d)$, L is a $(d-1)$-variate kernel, and $g_1, ..., g_{s-1}, g_{s+1}, ..., g_d$ are bandwidths that are allowed to be different from each other. Then the first component of the minimizer $\hat{\beta}$ of the weighted sum of squares

$$\sum_{j=1}^{n}\left\{Y_j - \sum_{l=0}^{p}\beta_{sl}(X_{js} - x_s)^l T_{js} - \sum_{k \neq s}\beta_k T_{jk}\right\}^2 K_h(X_{js} - x_s)L_g(\mathbf{X}_{j,-s} - \mathbf{x}_{-s})$$

5

is given by

$$\hat{\beta}_{s0} \equiv \hat{\beta}_{s0}(\mathbf{x}_{-s}) = e_0^T \left(\mathbf{Z}_s^T \mathbf{W}_s(\mathbf{x}_{-s})\mathbf{Z}_s\right)^{-1} \mathbf{Z}_s^T \mathbf{W}_s(\mathbf{x}_{-s})\mathbf{Y}$$

where e_l is the $(p+d)$-dimensional vector whose entries are zero except the $(l+1)$-th element which equals 1.

The integration estimator of $f_s(x_s)$ is a weighted average of $\beta_{s0}(\mathbf{X}_{i,-s})$'s, i.e.

$$\hat{f}_s(x_s) = \sum_{i=1}^n w_{-s}(\mathbf{X}_{i,-s})\hat{\beta}_{s0}(\mathbf{X}_{i,-s}) / \sum_{i=1}^n w_{-s}(\mathbf{X}_{i,-s}), \tag{4}$$

where the weight function $w_{-s}(\cdot)$ has a compact support with nonempty interior, and is introduced here to avoid some technical difficulty that may arise when the density of $\mathbf{X}_{i,-s}$'s has an unbounded support. Based on (4), one can predict Y given any realization $(\mathbf{x}, \mathbf{t})$ of $(\mathbf{X}, \mathbf{T})$ by the predictor

$$\hat{m}(\mathbf{x}, \mathbf{t}) = \sum_{s=1}^d \hat{f}_s(x_s)t_s. \tag{5}$$

In the estimation procedure for f_s for a given s, we fit local constants for the other varying coefficients $f_{s'}$, $s' \neq s$. One could fit higher order local polynomials for those varying coefficients, too. The theoretical performance of the resulting estimator would be the same as the present one, however. The smoothing bias of the present estimator due to the local averaging for $f_{s'}$, $s' \neq s$ can be made negligible by choosing the bandwidth vector g of smaller order than h and using a higher-order kernel L. See the conditions for the bandwidths and the kernel L given in the Appendix. In fact, the approach of taking a smaller bandwidth g and a higher order kernel L for the directions not of interest was also adopted by Fan, Härdle and Mammen (1998). One may sacrifice some rate of convergence in order to use a lower order kernel.

Let φ, φ_{-s} and φ_s denote the densities of $\mathbf{X}$, $\mathbf{X}_{-s}$ and X_s, respectively. Define

$$b_s(x_s) = \frac{f_s^{(p+1)}(x_s) \int u^{p+1} E\left\{w_{-s}(\mathbf{X}_{-s})T_s K_s^*(u; \mathbf{T}, x_s, \mathbf{X}_{-s})\right\} du}{(p+1)! E\{w_{-s}(\mathbf{X}_{-s})\}},$$

$$\sigma_s^2(x_s) = E\left[\frac{w_{-s}^2(\mathbf{X}_{-s})}{\varphi^2(\mathbf{X})}\varphi_{-s}^2(\mathbf{X}_{-s})\sigma^2(\mathbf{X}, \mathbf{T}) \int K_s^{*2}(u; \mathbf{T}, \mathbf{X}) du \,\bigg|\, X_s = x_s\right] \frac{\varphi_s(x_s)}{E^2\{w_{-s}(\mathbf{X}_{-s})\}},$$

where K_s^* is an equivalent kernel defined at (A.7).

Theorem 1 *Under the assumptions A1-A7 given in the appendix, we have, for any $s = 1, ..., d$, as $n \to \infty$,*

$$\sqrt{nh}\left\{\hat{f}_s(x_s) - f_s(x_s) - h^{p+1}b_s(x_s)\right\} \xrightarrow{\mathcal{L}} N\left\{0, \sigma_s^2(x_s)\right\}. \tag{6}$$

6

The estimator $\hat{m}(\mathbf{x}, \mathbf{t})$ of the prediction function $m(\mathbf{x}, \mathbf{t})$ enjoys the same rate of convergence as that of a single varying coefficient, and its asymptotic parameters are easily calculated from those of the $\hat{f}_s(x_s)$'s and the value of $\mathbf{t}$, as in the following theorem

Theorem 2 *Under the assumptions A1-A7 given in the appendix, we have, for any $s \neq s'$,*

$$\text{cov}\left[\sqrt{nh}\left\{\hat{f}_s(x_s) - f_s(x_s)\right\}, \sqrt{nh}\left\{\hat{f}_{s'}(x_{s'}) - f_{s'}(x_{s'})\right\}\right] \longrightarrow 0, \tag{7}$$

as $n \to \infty$, and hence

$$\sqrt{nh}\left\{\hat{m}(\mathbf{x}, \mathbf{t}) - m(\mathbf{x}, \mathbf{t}) - h^{p+1}b_m(\mathbf{x}, \mathbf{t})\right\} \xrightarrow{\mathcal{L}} N\left\{0, \sigma_m^2(\mathbf{x}, \mathbf{t})\right\} \tag{8}$$

where $b_m(\mathbf{x}, \mathbf{t}) = \sum_{s=1}^{d} b_s(x_s)t_s$ and $\sigma_m^2(\mathbf{x}, \mathbf{t}) = \sum_{s=1}^{d} \sigma_s^2(x_s)t_s^2$.

We comment here that Theorems 1 and 2 hold only for local polynomial estimators of odd degree p, while similar results hold for p even as well. In particular, $p = 0$ corresponds to integrating the well-known Nadaraya-Watson type estimator. When an even p is used instead, the variance formula remains the same while the bias formula contains extra terms involving the derivatives of the design density.

For selecting the bandwidths, following the idea of Ruppert, Sheather and Wand (1995) in local least squares regression, several plug-in type bandwidth selectors may be developed based on the asymptotic formulas given in the above theorems. Also, the modified multifold cross-validation criterion considered by Cai, Fan and Yao (2000) may be adapted for the above estimation. Theoretical development for these bandwidth selectors is beyond the scope of the paper. Below we describe a simple plug-in selection procedure for h and $\mathbf{g}$, which is employed in our numerical study in Sections 4 and 5.

The optimal bandwidth h_{opt} which minimizes the asymptotic mean integrated squared error of $\hat{f}_s$ is given by

$$h_{\text{opt}} = \left\{\frac{\int \sigma_s^2(x_s)\,dx_s}{2n(p+1)\int b_s^2(x_s)\,dx_s}\right\}^{1/(2p+3)}$$

7

Now, $\int b_s^2 (x_s) \, dx_s$ and $\int \sigma_s^2 (x_s) \, dx_s$ can be approximated respectively by

$$
\left[(p+1)! \, n^{-1} \sum_{i=1}^{n} w_{-s}(\mathbf{X}_{i,-s}) \right]^{-2} \int \left[f_s^{(p+1)} (x_s) \int u^{p+1} n^{-1} \sum_{i=1}^{n} \left\{ w_{-s} (\mathbf{X}_{i,-s}) \right. \right.
$$
$$
\left. \left. \times T_{is} K_s^* (u, \mathbf{T}_i, x_s, \mathbf{X}_{i,-s}) \right\} du \right]^2 dx_s,
$$

$$
\left[n^{-1} \sum_{i=1}^{n} w_{-s}(\mathbf{X}_{i,-s}) \right]^{-2} n^{-1} \sum_{i=1}^{n} w_{-s}^2 (\mathbf{X}_{i,-s}) \, \varphi^{-2} (\mathbf{X}_i) \, \varphi_{-s}^2 (\mathbf{X}_{i,-s}) \, \sigma^2 (\mathbf{X}_i, \mathbf{T}_i)
$$
$$
\times \int K_s^{*2} (u, \mathbf{T}_i, \mathbf{X}_i) \, du.
$$

The unknown functions $f_s^{(p+1)} (x_s), \sigma^2(\mathbf{x}, \mathbf{t}), \varphi(\mathbf{x}), \varphi(\mathbf{x}_{-s})$ and K_s^* may be substituted with their estimators as follows.

The $(p+1)$th derivative function $f_s^{(p+1)} (x_s)$ is estimated by fitting a polynomial regression model of degree $(p+2)$:

$$
m(\mathbf{X}, \mathbf{T}) = \sum_{s=1}^{d} \sum_{k=0}^{p+2} a_{s,k} X_s^k T_s.
$$

This leads to an estimator $\hat{f}_s^{(p+1)} (x_s) = (p+1)! \, \hat{a}_{s,p+1} + (p+2)! \, \hat{a}_{s,p+2} \, x_s$. As a by-product, the mean squared residual is used as an estimator of $\sigma^2(\mathbf{x}, \mathbf{t})$. The density functions $\varphi(\mathbf{x})$ and $\varphi(\mathbf{x}_{-s})$ are estimated by

$$
\hat{\varphi}(\mathbf{x}) = \frac{1}{n} \sum_{i=1}^{n} \prod_{s=1}^{d} \frac{1}{h(\mathbf{X}, d)} \phi \left(\frac{X_{is} - x_s}{h(\mathbf{X}, d)} \right),
$$

$$
\hat{\varphi}_{-s}(\mathbf{x}_{-s}) = \frac{1}{n} \sum_{i=1}^{n} \prod_{s' \neq s}^{d} \frac{1}{h(\mathbf{X}_{-s}, d-1)} \phi \left(\frac{X_{is'} - x_{s'}}{h(\mathbf{X}_{-s}, d_2 - 1)} \right)
$$

with the standard normal density ϕ and the rule-of-the-thumb bandwidth

$$
h(\mathbf{X}, m) = \sqrt{\widehat{\mathrm{var}} (\mathbf{X})} \, \{4/(m+2)\}^{1/(m+4)} \, n^{-1/(m+4)}.
$$

According to its definition given at (A.7), the dependence of the function $K_s^* (u, \mathbf{t}, \mathbf{x})$ on u and $\mathbf{t}$ is completely known. The only unknown term $E \left(\mathbf{TT}^T | \mathbf{X} = \mathbf{x} \right)$ contained in $S_s^{-1} (\mathbf{x})$ is estimated by fitting a matrix polynomial regression

$$
E \left(\mathbf{TT}^T | \mathbf{X} = \mathbf{x} \right) = \mathbf{c} + \sum_{s=1}^{d} \sum_{k=1}^{p} \mathbf{c}_{s,k} x_s^k
$$

in which the coefficients $\mathbf{c}$ and $\mathbf{c}_{s,k}$ are $d \times d$ matrices.

For the bandwidth vector g, we note that the choice $g_1 = \cdots = g_{s-1} = g_{s+1} = \cdots = g_d = (\log n)^{-1} h^{(p+1)/q}$ with h asymptotic to $n^{-1/(2p+3)}$ satisfies the condition (A7) for Theorem 1 if q, the order of the kernel L, is greater than $(d-1)/2$. Thus one may take $g_j \equiv (\log n)^{-1} h_{\text{opt}}^{(p+1)/q}$ for $j = 1, \ldots, s-1, s+1, \ldots, d$, where h_{opt} is the optimal bandwidth obtained from the above procedure.

2.2 Model II

In this section we describe local polynomial integration estimators of the coefficient functions $\{f_{su}, 1 \le u \le r_s, 1 \le s \le d_0\}$ in Model II. For the identifiability of the functions f_{su}, we assume that

$$\sum_{s=1}^{d_0} \sum_{u=1}^{r_s} r_{su}(X_{is}) T_{isu} = 0 \; a.s. \implies r_{su}(x) \equiv 0, \; u = 1, \ldots, r_s, \; s = 1, \ldots, d_0.$$

Define $\mathbf{X}_{-s}$ and $\mathbf{x}_{-s}$ as in Section 2.1. Let $\hat{\beta}_{su0}(\mathbf{x}_{-s})$ be the first component of the minimizer $\hat{\beta}$ of the following weighted sum of squares:

$$\sum_{j=1}^{n} \left\{ Y_j - \sum_{u=1}^{r_s} \sum_{l=0}^{p} \beta_{sul}(X_{js} - x_s)^l T_{jsu} - \sum_{s' \neq s}^{d_0} \sum_{u'=1}^{r_{s'}} \beta_{s'u'} T_{js'u'} \right\}^2 K_h(X_{js} - x_s) L_g(X_{j,-s} - \mathbf{x}_{-s}).$$

The integration estimator of $f_{su}(x_s)$ is given by a weighted average of $\beta_{su0}(\mathbf{X}_{i,-s})$'s, i.e.

$$\hat{f}_{su}(x_s) = \sum_{i=1}^{n} w_{-s}(\mathbf{X}_{i,-s}) \hat{\beta}_{su0}(\mathbf{X}_{i,-s}) / \sum_{i=1}^{n} w_{-s}(\mathbf{X}_{i,-s}).$$

As in Model I, one may predict Y given any realization $(\mathbf{x}, \mathbf{t})$ of $(\mathbf{X}, \mathbf{T})$ by the predictor

$$\hat{m}(\mathbf{x}, \mathbf{t}) = \sum_{s=1}^{d_0} \sum_{u=1}^{r_s} \hat{f}_{su}(x_s) t_{su}. \tag{9}$$

We have the following theorem which is an analogue of Theorem 1.

Theorem 3 *Under the assumptions A1-A2, A3', A4', A5-A6 and A7 given in the appendix, we have, for any $u = 1, \ldots, r_s$ and $s = 1, \ldots, d_0$,*

$$\sqrt{nh} \left\{ \hat{f}_{su}(x_s) - f_{su}(x_s) - h^{p+1} b_{su}(x_s) \right\} \xrightarrow{\mathcal{L}} N \left\{ 0, \sigma_{su}^2(x_s) \right\} \tag{10}$$

as $n \to \infty$, where $b_{su}(x_s) = \kappa_{su}(x_s)/\eta_s$ and $\sigma_{su}^2(x_s) = \tau_{su}^2(x_s)/\eta_s^2$. The definition of κ_{su} and τ_{su}^2 are given at (A.19) and (A.20)

9

Each pair of the entries $\hat{f}_{su}$ and $\hat{f}_{s'u'}$ for $1 \le s, s' \le d_0$ and $1 \le u, u' \le r_s$ has a negligible asymptotic covariance when $s \ne s'$. However, it has the same magnitude as the variance of each entry when $s = s'$. The following theorem is an analogue of Theorem 2.

Theorem 4 *Under the assumptions of Theorem 3, we have as $n \to \infty$*

(i) when $s' \ne s$

$$\text{cov}\left[\sqrt{nh}\left\{\hat{f}_{su}(x_s) - f_{su}(x_s)\right\}, \sqrt{nh}\left\{\hat{f}_{s'u'}(x_{s'}) - f_{s'u'}(x_{s'})\right\}\right] \to 0;$$

(ii) when $s' = s$

$$\text{cov}\left[\sqrt{nh}\left\{\hat{f}_{su}(x_s) - f_{su}(x_s)\right\}, \sqrt{nh}\left\{\hat{f}_{su'}(x_s) - f_{su'}(x_s)\right\}\right] \to \tau_{suu'}(x_s)/\eta_s^2$$

where $\tau_{suu'}$ is defined at (A.23). Hence

$$\sqrt{nh}\left\{\hat{m}(\mathbf{x}, \mathbf{t}) - m(\mathbf{x}, \mathbf{t}) - h^{p+1}\tilde{b}_m(\mathbf{x}, \mathbf{t})\right\} \xrightarrow{\mathcal{L}} N\left\{0, \tilde{\sigma}_m^2(\mathbf{x}, \mathbf{t})\right\}$$

where $\tilde{b}_m(\mathbf{x}, \mathbf{t}) = \sum_{s=1}^{d_0} \sum_{u=1}^{r_s} b_{su}(x_s)t_{su}$, $\tilde{\sigma}_m^2(\mathbf{x}, \mathbf{t}) = \sum_{s=1}^{d_0} \sum_{u=1}^{r_s} \sum_{u'=1}^{r_s} \sigma_{suu'}(x_s)t_{su}t_{su'}$ and $\sigma_{suu'}(x_s) = \tau_{suu'}(x_s)/\eta_s^2$.

3   TESTING FOR VARYING COEFFICIENTS

Suppose we are interested in testing the hypothesis

$$H_0 : f_s(x_s) \equiv \text{constant} \tag{11}$$

for a specific s in Model I. Testing the hypothesis (11) is a very important first step in model building procedure. If this hypothesis is true, one would get $\min_\alpha E\{f_s(X_s) - \alpha\}^2 w_s(X_s) = 0$ where w_s is an arbitrary positive weight function with a compact support. This leads us to propose the following test statistic:

$$
\begin{aligned}
V_{ns} &= n^{-1}\min_\alpha \sum_{i=1}^{n}\{\hat{f}_s(X_{is}) - \alpha\}^2 w_s(X_{is}) \\
&= n^{-1}\sum_{i=1}^{n} \hat{f}_s(X_{is})^2 w_s(X_{is}) - n^{-1}\{\sum_{i=1}^{n} w_s(X_{is})\}^{-1}\{\sum_{i=1}^{n} \hat{f}_s(X_{is})w_s(X_{is})\}^2, \tag{12}
\end{aligned}
$$

where the obvious solution of the least squares problem is given by

$$\hat{\alpha}_s = \{\sum_{i=1}^{n} w_s(X_{is})\}^{-1}\{\sum_{i=1}^{n} \hat{f}_s(X_{is})w_s(X_{is})\}. \tag{13}$$

The next theorem describes the asymptotic distribution of the test statistic (12) under the null hypothesis (11).

Theorem 5 *Under the null hypothesis (11) and the assumptions A1-A7 given in the appendix, we have, for any* $s = 1, ..., d$,

$$nh^{1/2}(V_{ns} - n^{-1}h^{-1}v_s) \xrightarrow{\mathcal{L}} N\{0, \gamma_s^2\} \tag{14}$$

as $n \to \infty$, *where* v_s *and* γ_s *are as given in (A.17) and (A.16).*

For the practical implementation of the test, we suggest to use a bootstrap procedure instead of the asymptotic normal distribution theory in Theorem 5. The reason is that for a test statistic based on kernel type of smoothing, the normal approximation to the distribution of the test statistic is very poor, as shown in Härdle and Mammen (1993) and, more recently, confirmed by Sperlich, Tjøstheim and Yang (2002). Another reason is that the normal approximation given in Theorem 5 involves too complicated expressions, which makes the task of obtaining asymptotic critical values out of reach.

It is well-known that the ordinary method of resampling residuals fails to work when the error variances are allowed to be different. See Wu (1986), Liu (1988), and Mammen (1992). Härdle and Mammen (1993) also pointed out that it breaks down even for homoscedastic errors in the case of the goodness-of-fit test statistic for testing a parametric hypothesis against the nonparametric alternative. As an alternative, we suggest to use the wild bootstrap procedure which was first introduced by Wu (1986) and implemented in various settings by Liu (1988), Härdle and Mammen (1993), and Sperlich, Tjøstheim and Yang (2002) among others. Basically, this approach attempts to mimic the conditional distribution of each response given covariate using the corresponding *single* residual, in such a way that the first three moments of the bootstrap population equal to those of the single residual.

To describe the procedure in our setting, let $\tilde{m}(\mathbf{x}, \mathbf{t}) = \hat{\alpha}_s t_s + \sum_{k \neq s}^{d} \hat{f}_k(x_k) t_k$ be the regression estimator under the hypothesis (11), where $\hat{\alpha}_s$ is an estimate of the constant $\hat{f}_s(x_s)$ given by (13) while $\hat{f}_k(x_k)$ ($k \neq s$) is the marginally integrated estimate of $f_k(x_k)$ in (4). The wild bootstrap procedure to estimate the sampling distribution of V_{ns} under the null hypothesis then consists of the following steps:

(i) Find the residuals $\tilde{\varepsilon}_i = Y_i - \tilde{m}(\mathbf{X}_i, \mathbf{T}_i)$ for $i = 1, \ldots, n$.

(ii) Generate i.i.d. random variables Z_i^W such that $E(Z_i^W) = 0$, $E(Z_i^W)^2 = 1$ and $E(Z_i^W)^3 = 1$. Put $Y_i^* = \tilde{m}(\mathbf{X}_1, \mathbf{T}_1) + \tilde{\varepsilon}_i Z_i^W$.

(iii) Compute the bootstrap test statistic V_{ns}^* using the wild bootstrap sample $\{(Y_i^*, \mathbf{X}_i, \mathbf{T}_i)\}_{i=1}^{n}$

(iv) Repeat the steps (ii) and (iii) M times, obtaining $V^*_{ns,1}, \ldots, V^*_{ns,M}$. Estimate the null distribution of V_{ns} by the empirical distribution of $V^*_{ns,1}, \ldots, V^*_{ns,M}$.

For examples of Z^W_i satisfying the moment conditions, see Mammen (1992). For the empirical example in the next section, we used a two-point distribution : $Z^W_i = (1 - \sqrt{5})/2$ with probability $(5 + \sqrt{5})/10$, and $Z^W_i = (1 + \sqrt{5})/2$ with probability $(5 - \sqrt{5})/10$, with $M = 200$.

For Model II, we consider the following hypothesis:

$$f_{su}(x_s) \equiv \text{constant}. \tag{15}$$

The corresponding test statistic for the hypothesis (15) is given by

$$V_{nsu} = n^{-1} \sum_{i=1}^{n} \hat{f}_{su}(X_{is})^2 w_s(X_{is}) - n^{-1} \{\sum_{i=1}^{n} w_s(X_{is})\}^{-1} \{\sum_{i=1}^{n} \hat{f}_{su}(X_{is}) w_s(X_{is})\}^2.$$

The next theorem describes the asymptotic distribution of the test statistic V_{nsu} under the null hypothesis (15).

Theorem 6 *Under the null hypothesis (15) and the assumptions of Theorem 3, we have, for any $u = 1, \ldots, r_s$ and $s = 1, \ldots, d_0$,*

$$nh^{1/2}(V_{nsu} - n^{-1}h^{-1}v_{su}) \xrightarrow{\mathcal{L}} N\{0, \gamma^2_{su}\}$$

as $n \to \infty$, where v_{su} and γ_{su} are as given in (A.22) and (A.21).

For testing the hypothesis (15), let $\bar{m}(\mathbf{x}, \mathbf{t}) = \hat{\alpha}_{su} t_{su} + \sum_{s' \neq s}^{d_0} \sum_{u'=1}^{r_{s'}} \hat{f}_{s'u'}(x_{s'}) t_{s'u'}$ where

$$\hat{\alpha}_{su} = \{\sum_{i=1}^{n} w_s(X_{is})\}^{-1} \{\sum_{i=1}^{n} \hat{f}_{su}(X_{is}) w_s(X_{is})\}.$$

A wild bootstrap procedure may be obtained by simply replacing $\tilde{m}$, V_{ns} and V^*_{ns} by $\bar{m}$, V_{nsu} and V^*_{nsu}, respectively, in the four steps described above for testing (11).

Some related work on this testing problem includes Chen and Liu (2001), and Cai, Fan and Yao (2000). The former paper treated testing, in the FAR model, whether all the coefficient functions are constant, i.e., whether the underlying process is simply a linear AR model. The latter proposed a testing procedure for the hypothesis that all the coefficient functions have known parametric forms. We think testing for a parametric form in our models is also an interesting topic for future research.

12

4 SIMULATION STUDY

In this section we investigate the finite sample properties of the estimation and testing methods through two simulated examples. One is the case where $(\mathbf{X}_i, \mathbf{T}_i)$ are independent and identically distributed (i.i.d.), and the other is the case where they are endogenous and are lagged observations of the response Y. We employed local linear smoothing ($p = 1$) in all cases. Both of the kernels K and L were the quartic kernel $K(x) = L(x) = 0.9375 \left(1 - x^2\right)^2 I_{(-1,1)}(x)$, while the bandwidths were chosen as described Section 2.1.

4.1 The i.i.d. case

In this case we generated the data from the following varying coefficient model:

$$Y = f_1(X_1) + f_2(X_2)T_1 + f_3(X_3)T_2 + \delta(\mathbf{X}, \mathbf{T})\varepsilon,$$

where $f_1(X_1) = 1 + \exp\left(2X_1 - 1\right)$, $f_2(X_2) = \cos\left(2\pi X_2\right)$, $f_3(X_3) = 2$. The heteroscedastic conditional standard deviation was set to be

$$\delta(\mathbf{X}, \mathbf{T}) = 0.5 + \frac{T_1^2 + T_2^2}{1 + T_1^2 + T_2^2} \exp\left(-2 + (X_1 + X_2)/2\right).$$

The particular form of $\delta(\mathbf{X}, \mathbf{T})$ was considered to ensure the variance to be bounded. The vector $\mathbf{X} = (X_1, X_2, X_3)^T$ was generated from the uniform distribution over the unit cube $[0, 1]^3$, and $\mathbf{T} = (T_1, T_2)^T$ was generated from the bivariate normal with mean zero and covariance matrix $\begin{pmatrix} 1 & 0.5 \\ 0.5 & 1 \end{pmatrix}$. The vectors $\mathbf{X}$ and $\mathbf{T}$ were generated independently. Finally, the error term ε was generated from the standard normal distribution independently of $(\mathbf{X}, \mathbf{T})$.

A total of 100 independent data sets with sizes $n = 50, 100$ and 250 were generated. The estimated functions of f_s, $s = 1, 2, 3$ were evaluated on a grid of 91 equally-spaced points x_j, $j = 1, ..., 91$ with $x_1 = 0.05$, $x_{91} = 0.95$. To assess the performance of $\hat{f}_s$ for $s = 1, 2, 3$, we calculated the mean integrated squared error (MISE) of $\hat{f}_s$ which is defined by

$$\mathrm{MISE}(\hat{f}_s) = \frac{1}{R} \sum_{r=1}^{R} \mathrm{ISE}(\hat{f}_{r,s}) = \frac{1}{R} \sum_{r=1}^{R} \frac{1}{g} \sum_{j=1}^{g} \left\{ \hat{f}_{r,s}(x_j) - f_s(x_j) \right\}^2.$$

Here $\hat{f}_{r,s}(x_j)$ denotes the estimated value of f_s at x_j for the rth data set, $R = 100$ and $g = 91$ are the numbers of data sets and grid points, respectively. Table 1 summarizes the MISE values of the function estimators. This simulation study numerically supports our theoretical results for the estimation method as given in Section 2.

13

(Insert Table 1 about here)

To see how the marginal integration improves the three dimensional function estimators, we also computed the mean average squared errors for the case where $n = 50$. Consider $\hat{\beta}_{s0}$, as defined in Section 2.1, evaluated at the observed X_{i1}, X_{i2}, X_{i3}. Write them $\hat{\beta}_{s0}(X_{i1}, X_{i2}, X_{i3})$. These are the estimates before the marginal integration. We computed the mean average squared error

$$
\begin{aligned}
\text{MASE}_1 \ = \ & \frac{1}{R}\sum_{r=1}^{R}\frac{1}{n}\sum_{i=1}^{n}\Big\{\hat{\beta}_{r,10}(X_{i1}, X_{i2}, X_{i3}) + \hat{\beta}_{r,20}(X_{i1}, X_{i2}, X_{i3})T_{i1} \\
& +\hat{\beta}_{r,30}(X_{i1}, X_{i2}, X_{i3})T_{i2} - f_1(X_{i1}) - f_2(X_{i2})T_{i1} - f_3(X_{i3})T_{i2}\Big\}^2,
\end{aligned}
$$

and also for the marginal integration estimate

$$
\begin{aligned}
\text{MASE}_2 \ = \ & \frac{1}{R}\sum_{r=1}^{R}\frac{1}{n}\sum_{i=1}^{n}\Big\{\hat{f}_{r,1}(X_{i1}) + \hat{f}_{r,2}(X_{i2})T_{i1} + \hat{f}_{r,3}(X_{i3})T_{i2} \\
& -f_1(X_{i1}) - f_2(X_{i2})T_{i1} - f_3(X_{i3})T_{i2}\Big\}^2,
\end{aligned}
$$

where $\hat{\beta}_{r,s0}(X_{i1}, X_{i2}, X_{i3})$ and $\hat{f}_{r,s}(X_{is})$ are the estimates for the rth dataset. We found $\text{MASE}_1 = 0.3164$ and $\text{MASE}_2 = 0.2761$.

Next, we give some numerical results for the testing method. For each of the simulated data sets above, we applied the proposed wild bootstrap method with $M = 500$ to test the null hypothesis $H_{s0} : f_s = c_s$ for some constants c_s. Table 2 provides for each s the proportion of the cases where the null hypothesis H_{s0} was rejected at the significance level $\alpha = 0.05$ among the 100 replications.

(Insert Table 2 about here)

4.2 The time series case

In this simulation, $R = 200$ time series were generated. Each time, 1000 observations were generated from the following varying coefficient AR (VCAR) model, among which only the last 250 observations were used:

$$
Y_t = f_1(Y_{t-3})Y_{t-1} + f_2(Y_{t-4})Y_{t-2} + 0.2\varepsilon_t \tag{16}
$$

where $f_1(Y_{t-3}) = 0.4 + (0.1 + Y_{t-3})\exp(-3Y_{t-3}^2)$, $f_2(Y_{t-4}) = -0.2 - (0.6 + Y_{t-4})\exp(-3Y_{t-4}^2)$, and ε_t are i.i.d. standard normal random variates. Again, the performance of the estimators of f_1 and f_2 were assessed by MISE. We obtained $\text{MISE}(\hat{f}_1) = 0.0137$ and $\text{MISE}(\hat{f}_2) = 0.0151$.

14

We found that the Monte Carlo average over 200 time series of $\sqrt{\sum_{t=1}^{250}(Y_t - \bar{Y})^2/250}$ equals 0.6374 with the standard error 0.0026, where $\bar{Y} = \sum_{t=1}^{250} Y_t/250$. The obtained values of MISEs are much smaller than the variation of Y, which means that the fitted model with $\hat{f}_1$ and $\hat{f}_2$ is useful to explain the variation of Y.

Similarly as in the i.i.d. case, we report here a numerical result for testing $H_{10} : f_1 =$ constant and $H_{20} : f_2 =$ constant. For each of the simulated time series we applied the wild bootstrap method with $M = 500$ and used the significant level 0.05. We found that the proportion of the cases where the null hypothesis was rejected among the 200 replications was 0.57 for H_{10} and 0.943 for H_{20}.

It is also of interest to examine the effectiveness of the varying coefficient model (16) in comparison with some related FAR models, discussed in Cai, Fan and Li (2000) and Cai, Fan and Yao (2000), where all the coefficient functions depend on the same variable(s). For this purpose, we considered the following three FAR models:

$$Y_t = g_1(Y_{t-3})Y_{t-1} + g_2(Y_{t-3})Y_{t-2} + 0.2\varepsilon_t, \tag{17}$$

$$Y_t = g_1(Y_{t-4})Y_{t-1} + g_2(Y_{t-4})Y_{t-2} + 0.2\varepsilon_t, \tag{18}$$

$$Y_t = g_1(Y_{t-3}, Y_{t-4})Y_{t-1} + g_2(Y_{t-3}, Y_{t-4})Y_{t-2} + 0.2\varepsilon_t. \tag{19}$$

We fitted the three FAR models with the same series generated by (16). For comparison we computed the mean average squared residuals (MASR) defined by

$$\text{MASR} = \sum_{r=1}^{200}\sum_{t=1}^{250} (y_{r,t} - \hat{y}_{r,t})^2 /(200 \times 250),$$

where $y_{r,t}$ denotes the tth observation in the rth replication, and $\hat{y}_{r,t}$ is the corresponding fitted value based on the underlying model. We note that average squared residuals (ASR), as a statistic that can be computed from any data, real or simulated, is a very useful measure of goodness-of-fit. This is illustrated in the next section where ASR is used to select an optimal forecasting model. Thus, MASR is a sensible criterion to compare different models. Although it varies with the bandwidth, an incorrect model would have an MASR asymptotically greater than that of a correct model by a positive constant, which is of larger magnitude than any variation caused by bandwidth tuning. The three FAR models (17), (18) and (19) gave the MASR values 1.020, 0.343 and 0.081, respectively, whereas the VCAR model (16) gave a much smaller 0.075.

We also compared the mean average squared prediction errors (MASPE) of these models. For this, we generated additional 50 observations for each of the 200 times series of size 250

and computed

$$\text{MASPE} = \sum_{r=1}^{200} \sum_{t=251}^{300} (y_{r,t} - \widehat{y}_{r,t})^2 /(200 \times 50),$$

where $\widehat{y}_{r,t}$ is the predicted value of $y_{r,t}$ based on the estimated model from the first 250 observations. The three FAR models (17), (18) and (19) gave the MASPE values 0.075, 0.071 and 0.062, respectively, while the VCAR model (16) gave 0.059.

5 AN EMPIRICAL EXAMPLE

We illustrate our estimation and testing methods with an analysis of the quarterly (seasonally non-adjusted) West German real GNP data collected from 1960:1 to 1990:4. The data G_t, $1 \leq t \leq n = 124$, which was compiled by Wolters (1992, p. 424, note 4), is plotted in Figure 1(a). One sees clearly a linear trend and a seasonal pattern. Based on the seasonal unit root test of Franses (1996), we took the first differences of the logs, and obtained a time series data, D_t, $1 \leq t \leq n = 124$, which is plotted in Figure 1(b). This time series no longer reveals any linear or higher order trends, but is obviously seasonal. Following the de-seasonalization procedure of Yang and Tschernig (2002), the sample means of the four seasons -0.065116, 0.038595, 0.051829 and 0.008944, respectively, were calculated and subtracted from D_t so that the de-seasonalized Y_t, $1 \leq t \leq n = 124$, became the growth rates with respect to the spring season. As such, it is reasonable to assume that the Y_t's satisfy our strict stationarity and mixing conditions. In Figure 4, the data $Y_t, 1 \leq t \leq n = 124$, is plotted as the dotted curve.

(Insert Figure 1 and Table 3 about here)

According to the semiparametric lag selection performed in Yang and Tschernig (2002), the significant variables for the prediction of Y_t are Y_{t-4} and Y_{t-2}. Calculation of the autocorrelation functions indicated that Y_t is more correlated with Y_{t-1} and Y_{t-3} than other lagged values. Hence we fitted all the twelve VCAR models of Model I type, consisting of the lagged variables Y_{t-1}, Y_{t-2}, Y_{t-3} and Y_{t-4}. According to the definition (4) of the marginal integration estimator, we estimated all VCAR models using the first 114 observations and made out-of-sample predictions for the last 10 observations. Their average squared residuals (ASR) and average squared prediction errors (ASPE) are reported in Table 3. One may expect the ASRs should be smaller than the ASPEs. But we found in the residual plots that there were some very large residual terms that made all the ASRs larger than their

16

corresponding ASPEs. The model with the smallest ASR is

$$Y_t = f_1(Y_{t-1})Y_{t-2} + f_2(Y_{t-3})Y_{t-4} + \text{(noise)}, \tag{20}$$

while the model with the smallest ASPE is

$$Y_t = f_1(Y_{t-3})Y_{t-2} + f_2(Y_{t-1})Y_{t-4} + \text{(noise)}. \tag{21}$$

Both of the above models include as special case the following linear AR(2) model:

$$Y_t = c_1 Y_{t-2} + c_2 Y_{t-4} + \text{(noise)}. \tag{22}$$

In Table 3, the ASR and ASPE of the linear AR model (22) are also included. Both the optimal VCAR models (20) and (21) have much smaller ASR and ASPE than the linear AR model. These two VCAR models have similar values of ASR and ASPE. Figure 2 depicts the estimates of the functions f_1 and f_2 for each model. To test if these functions are significantly different from a constant, we carried out the wild bootstrap procedures. For the model (20), the p-values were 0.80 for f_1 and 0.01 for f_2, while for the model (21) they were 0.22 and 0.48, respectively. This means that for the model (20) the function f_1 is not significantly different from a constant but there is a strong evidence in the data for that f_2 is not a constant. Thus one may conclude that a parsimonious model is the partially linear model:

$$Y_t = f_1 Y_{t-2} + f_2(Y_{t-3})Y_{t-4} + \text{(noise)}.$$

We further computed the ASR and ASPE of this semiparametric partially linear model, which are 0.00032 and 0.00024 respectively, as seen in Table 3. In terms of these estimation and forecasting errors, the semiparametric model is much inferior to its nonparametric parent model (20). Thus the simpler semiparametric model is preferred only for its parsimony while the nonparametric model (20) should be used if optimal forecasting is the goal. The testing for coefficient functions, therefore, works in a similar fashion as BIC works for linear AR time series where ASR is similar to AIC. For linear AR time series, it is well known that AIC is optimal for forecasting while BIC is consistent in identifying a correct AR model. It should be noted also that ASR can be compared across models not necessarily nested within each other, while the testing procedure selects the most parsimonious model from a nested hierarchy of models.

To further verify the validity of the models (20) and (21), we examined the residuals $\hat{\varepsilon}_t$ to check the independence of the error terms as it is another way of assessing goodness-of-fit for the models. At a practical level, such independence can be checked using the autocorrelation

functions (ACF) of powers of $|\hat{\varepsilon}_t|$. Figure 3 shows the ACFs of both $|\hat{\varepsilon}_t|$ and $\hat{\varepsilon}_t^2$ for the models (20) and (21). As can be seen from the plots, within the confidence levels of $\pm 2 \times n^{-1/2}$ lie more than 95% of all the sample ACFs, and hence we can conclude that both $|\hat{\varepsilon}_t|$ and $\hat{\varepsilon}_t^2$ have no autocorrelation. The ACF plots for $|\hat{\varepsilon}_t|^3$, $\hat{\varepsilon}_t^4$, etc., led to the same conclusion. Thus, the models (20) and (21) fit well the structure of the data Y_t. As a further evidence, Figure 4 shows the overlay of Y_t together with the predicted series $\hat{Y}_t$ obtained from fitting the models (20) and (21). The predicted series follows the actual series very closely.

(Insert Figures 2, 3 and 4 about here)

APPENDIX: PROOFS

A longer version of the paper with proofs of greater detail may be found at *http: ace.snu.ac.kr/ theostat/papers/jasa-ypxh.pdf.*

A.1 Preliminaries

We shall need the following technical assumptions on the kernels.

A1: *The kernels K and L are symmetric, Lipschitz continuous with $\int K(u)\,du = \int L(u)\,du = 1$, and have compact supports with nonempty interiors. While K is nonnegative, the kernel L is of order q.*

When estimating the function f_s for a particular s, a multiplicative kernel is used consisting of K for the s-th variable and L for all other variables. To accommodate dependent data, such as those from varying-coefficient autoregression models, we assume that

A2: *The vector process $\{(\mathbf{X}_i, \mathbf{T}_i)\}_{i=1}^{n}$ is strictly stationary and β-mixing with mixing coefficients $\beta(k) \leq C_2 \rho^k, 0 < \rho < 1$. Here*

$$\beta(n) = \sup_k E \sup \left\{ \left| P(A|\mathcal{F}_{-\infty}^k) - P(A) \right| : A \in \mathcal{F}_{n+k}^{\infty} \right\}$$

where $\mathcal{F}_t^{t'}$ is the σ-algebra generated by $(\mathbf{X}_t, \mathbf{T}_t), (\mathbf{X}_{t+1}, \mathbf{T}_{t+1}), ..., (\mathbf{X}_{t'}, \mathbf{T}_{t'})$ for $t < t'$.

The following assumptions are on the smoothness of the functions involved in the estimation and testing, and on the moments of the process for the proofs of Theorems 1, 2 and 5.

18

A3: *The functions f_s's have bounded continuous $(p+1)$-th derivatives for all $1 \leq s \leq d$, and $p \geq q - 1$*

A4: *The distribution of $(\mathbf{X}, \mathbf{T})$ has a density ψ and $\mathbf{X}$ has a marginal density φ. On the supports of weight functions w_{-s} and w_s, the densities φ_{-s} of $\mathbf{X}_{-s}$ and φ_s of X_s, respectively, are uniformly bounded away from zero and infinity. The marginal density φ and $E(T_s T_{s'} | \mathbf{X} = \cdot)$ for $1 \leq s, s' \leq d$ are Lipschitz continuous. Also, $\sigma^2(\cdot, \mathbf{t})$ and $\psi(\cdot, \mathbf{t})$ are equicontinuous.*

A5: *The weight functions w_{-s} and w_s are nonnegative, have compact supports with nonempty interiors, and are continuous on their supports.*

A6: *The error term ε_t satisfies $E|\varepsilon_t|^{4+\delta} < \infty$ for some $\delta > 0$. For $j < k < l < m$ there exists a joint probability density function $\psi_{j,k,l,m}$ of $(X_j, T_j; X_k, T_k; X_l, T_l; X_m, T_m)$. Let $\mathcal{X} = \{x : x_s \in \text{supp}(w_s), \mathbf{x}_{-s} \in \text{supp}(w_{-s})\}$, and for $\epsilon > 0$ define $\mathcal{X}_\epsilon = \{\mathbf{x} : \text{there exists } \mathbf{z} \in \mathcal{X} \text{ such that } \|\mathbf{z} - \mathbf{x}\| \leq \epsilon\}$. There exist $\epsilon > 0$, $\bar{\sigma}(\mathbf{t})$ and $\tilde{\varphi}_{j,k,l,m}(\mathbf{t}_j, \mathbf{t}_k, \mathbf{t}_l, \mathbf{t}_m)$ such that $\sigma(\mathbf{x}, \mathbf{t}) \leq \bar{\sigma}(\mathbf{t})$ for all $\mathbf{x} \in \mathcal{X}_\epsilon$, $\psi_{j,k,l,m}(\mathbf{x}_j, \mathbf{t}_j; \mathbf{x}_k, \mathbf{t}_k; \mathbf{x}_l, \mathbf{t}_l; \mathbf{x}_m, \mathbf{t}_m) \leq \tilde{\varphi}_{j,k,l,m}(\mathbf{t}_j, \mathbf{t}_k, \mathbf{t}_l, \mathbf{t}_m)$ for all $\mathbf{x}_j, \mathbf{x}_k, \mathbf{x}_l, \mathbf{x}_m$ in $\mathcal{X}_\epsilon$, and $\int (\|\mathbf{t}_j\| \|\mathbf{t}_k\| \|\mathbf{t}_l\| \|\mathbf{t}_m\|)^{2+c} |\bar{\sigma}(\mathbf{t}_j) \bar{\sigma}(\mathbf{t}_k) \bar{\sigma}(\mathbf{t}_l) \bar{\sigma}(\mathbf{t}_m)|^{2+c} \tilde{\varphi}_{j,k,l,m} (\mathbf{t}_j, \mathbf{t}_k, \mathbf{t}_l, \mathbf{t}_m) \, d\mathbf{t}_j d\mathbf{t}_k d\mathbf{t}_l d\mathbf{t}_m \leq C < \infty$ for some $c > 0$ and $C > 0$.*

Also, we assume that the bandwidths, g for the kernel L and h for the kernel K, satisfy

A7: *$(\ln n)(nhg_{\text{prod}})^{-1/2} = O(n^{-a})$ for some $a > 0$ and $(nh \ln n)^{1/2} g_{\text{max}}^q \rightarrow 0$ as $n \rightarrow \infty$ where $g_{\text{prod}} = g_1 \cdots g_{s-1} g_{s+1} \cdots g_d$ and $g_{\text{max}} = \max(g_1, \ldots, g_{s-1}, g_{s+1}, \ldots, g_d)$, and h is asymptotic to $n^{-1/(2p+3)}$.*

For Theorems 3, 4 and 6, we need to modify the assumptions A3, A4 and A7 slightly as follows:

A3': *The functions f_{su}'s have bounded continuous $(p+1)$-th derivatives for all $1 \leq s \leq d_0$, $1 \leq u \leq r_s$ and $p \geq q - 1$.*

A4': *It is the same as A4 except that now we require $E(T_{su} T_{s'u'} | \mathbf{X} = \cdot)$ for $1 \leq s, s' \leq d_0$ and $1 \leq u, u' \leq r_s$ are Lipschitz continuous.*

A7': *It is also the same as A7 except that d is replaced by d_0.*

One should note here that for existence of the bandwidth vector **g** satisfying the assumption A7 and A7' it is necessary that q, the order of the kernel L, should be larger than $(d - .1)/2$ and $(d_0 - 1)/2$, respectively.

To prove many of our results, we make use of some inequalities about U-statistic and von Mises statistic of dependent variables derived from Yoshihara (1976). Let $\xi_i, 1 \leq i \leq n$ be a strictly stationary sequence of random variables with values in R^d and β-mixing coefficients $\beta(k), k = 1, 2, ...,$ and r a fixed positive integer. Let $\{\theta_n(F)\}$ denote the functionals of the distribution function F of ξ_i

$$\theta_n(F) = \int g_n(x_1, ..., x_m) \, dF(x_1) \cdots dF(x_m)$$

where $\{g_n\}$ are measurable functions symmetric in their m arguments such that

$$\int |g_n(x_1, ..., x_m)|^{2+\delta} \, dF(x_1) \cdots dF(x_m) \leq M_n < +\infty,$$

$$\sup_{(i_1, ..., i_m) \in S_c} \int |g_n(x_1, ..., x_m)|^{2+\delta} \, dF_{\xi_{i_1}, ..., \xi_{i_m}}(x_1, ..., x_m) \leq M_{n,c} < +\infty, \quad c = 0, ..., m-1$$

for some $\delta > 0$, where $S_c = \{(i_1, ..., i_m) | \#_r(i_1, ..., i_m) = c\}, c = 0, ..., m-1$ and for every $(i_1, ..., i_m), 1 \leq i_1 \leq \cdots \leq i_m \leq n$, $\#_r(i_1, ..., i_m) = $ the number of $j = 1, ..., m-1$ satisfying $i_{j+1} - i_j \leq r$. Clearly, the cardinality of each set S_c is less than n^{m-c}.

The von Mises' differentiable statistic and the U-statistic

$$
\begin{aligned}
\theta_n(F_n) &= \int g_n(x_1, ..., x_m) \, dF_n(x_1) \cdots dF_n(x_m) \\
&= \frac{1}{n^m} \sum_{i_1=1}^{n} \cdots \sum_{i_m=1}^{n} g_n(\xi_{i_1}, ..., \xi_{i_m}) \\
U_n &= \frac{1}{\binom{n}{m}} \sum_{1 \leq i_1 < \cdots < i_m \leq n} g_n(\xi_{i_1}, ..., \xi_{i_m}),
\end{aligned}
$$

allow decompositions as

$$
\begin{aligned}
\theta_n(F_n) &= \theta_n(F) + \sum_{c=1}^{m} \binom{m}{c} V_n^{(c)}, \\
U_n &= \theta_n(F) + \sum_{c=1}^{m} \binom{m}{c} U_n^{(c)}.
\end{aligned}
$$

Here, $g_{n,c}$ are the projections of g_n defined by

$$g_{n,c}(x_1, ..., x_c) = \int g_n(x_1, ..., x_m) \, dF(x_{c+1}) \cdots dF(x_m), c = 0, 1, ..., m$$

20

so that $g_{n,0} = \theta_n(F)$, $g_n = g_{n,m}$, and

$$V_n^{(c)} = \int g_{n,c}(x_1, ..., x_c) \prod_{j=1}^{c} [dF_n(x_j) - dF(x_j)],$$

$$U_n^{(c)} = \frac{(n-c)!}{n!} \sum_{1 \le i_1 < \cdots < i_c \le n} \int g_{n,c}(x_{i_1}, ..., x_{i_c}) \prod_{j=1}^{c} \left[dI_{R_+^d}(x_j - \xi_{i_j}) - dF(x_j) \right]$$

where $I_{R_+^d}$ is the indicator function of $R_+^d = \{(y_1, ..., y_d) \in R^d \mid y_j \ge 0, j = 1, ..., d\}$.

Lemma A.1 *If $\beta(k) \le C_1 k^{-(2+\delta')/\delta'}$ for $\delta > \delta' > 0$, then*

$$EV_n^{(c)2} + EU_n^{(c)2} \tag{A.1}$$

$$\le C(m, \delta, r) n^{-c} \left\{ M_n^{2/(2+\delta)} \sum_{k=r+1}^{n} k\beta^{\delta/(2+\delta)}(k) + \sum_{c'=0}^{m-1} n^{-c'} M_{n,c'}^{'2/(2+\delta)} \sum_{k=1}^{r} k\beta^{\delta/(2+\delta)}(k) \right\}$$

for some constant $C(m, \delta, r) > 0$. In particular, if one has $\beta(k) \le C_2 \rho^k$ for $0 < \rho < 1$, then

$$EV_n^{(c)2} + EU_n^{(c)2} \le C(m, \delta, r) C_2 C(\rho) n^{-c} \left\{ M_n^{2/(2+\delta)} + \sum_{c'=0}^{m-1} n^{-c'} M_{n,c'}^{'2/(2+\delta)} \right\}. \tag{A.2}$$

Proof. The proof essentially is the same as Lemma 2 in Yoshihara (1976), which dealt with the special case of $g_n \equiv g, r = 1, M_n = M_n'$ and yielded (A.1). The inequalities in the proof of this lemma do not require all g_n's to be the same for $n = 1, 2, ...$, and terms in $U_n^{(c)}$ where exactly c' pairs of neighboring indices differ by at most r form a subset of terms with cardinality of order $n^{c-c'}$. Elementary arguments then establish (A.2) under geometric mixing conditions. ∎

A.2   Proofs of Theorems 1, 2 and 5

Define the following square matrix of dimension $(p + d)$

$$S_s(\mathbf{x}) = \begin{bmatrix} \int \mathbf{p}(u)\mathbf{p}^T(u)K(u)du \, E(T_s^2|\mathbf{X} = \mathbf{x}) & \int \mathbf{p}(u)K(u)du \, E(T_s \mathbf{T}_{-s}^T|\mathbf{X} = \mathbf{x}) \\ E(T_s \mathbf{T}_{-s}|\mathbf{X} = \mathbf{x}) \int \mathbf{p}^T(u)K(u)du & E(\mathbf{T}_{-s}\mathbf{T}_{-s}^T|\mathbf{X} = \mathbf{x}) \end{bmatrix}.$$

The identifiability condition given at (3) is closely related to the invertibility of the matrix $S_s(\mathbf{X})$. To see this, we note that for vectors λ_1 and λ_2 of dimensions $p + 1$ and $d - 1$, respectively,

$$(\lambda_1^T, \lambda_2^T)S_s(\mathbf{x})(\lambda_1^T, \lambda_2^T)^T = \int E\left[\{\lambda_1^T \mathbf{p}(u)T_s + \lambda_2^T \mathbf{T}_{-s}\}^2 \mid \mathbf{X} = \mathbf{x} \right] K(u) \, du.$$

Thus, if $\left[\lambda_1(X_s)^T, \lambda_2(\mathbf{X}_{-s})^T\right] S_s(\mathbf{X}) \left[\lambda_1(X_s)^T, \lambda_2(\mathbf{X}_{-s})^T\right]^T = 0$, $a.s.$, then $\lambda_1(X_s)^T \mathbf{p}(u)T_s + \lambda_2(\mathbf{X}_{-s})^T T_{-s} = 0$, $a.s.$ $(\mathbf{X}, \mathbf{T})$ and $u \in \text{supp}(K)$. Since K has a nonempty interior, the identifiability condition (3) implies $\lambda_1 \equiv 0$ and $\lambda_2 \equiv 0$ by the uniqueness of polynomial expansion.

The next lemma shows that the matrix $S_s(\mathbf{x})$ is proportional to the limiting dispersion matrix

Lemma A.2 *As* $n \to \infty$

$$\sup_{x_s \in \text{supp}(w_s), \mathbf{x}_{-s} \in \text{supp}(w_{-s})} \left| \mathbf{Z}_s^T \mathbf{W}_s(\mathbf{x}_{-s})\mathbf{Z}_s - \varphi(x_s, \mathbf{x}_{-s})S(x_s, \mathbf{x}_{-s}) \right| = o(b) \ a.s.$$

where $b = \ln n \left(h + g_{\max}^q + 1/\sqrt{nhg_{\text{prod}}} \right)$.

Proof. The conclusion follows by directly using the covering technique and exponential inequalities for β-mixing processes, as in the proof of Theorem 2.2 of Bosq (1998). ∎

Now let c be an integer such that $b^{c+1} = o(h^{p+2})$, the next lemma decomposes the dispersion matrix.

Lemma A.3 *For any integer* k,

$$\left(\mathbf{Z}_s^T \mathbf{W}_s(\mathbf{x}_{-s})\mathbf{Z}_s\right)^{-1} - \frac{S_s^{-1}(x_s, \mathbf{x}_{-s})}{\varphi(x_s, \mathbf{x}_{-s})}$$

$$= \frac{S_s^{-1}(x_s, \mathbf{x}_{-s})}{\varphi(x_s, \mathbf{x}_{-s})} \sum_{\ell=1}^{c} \left\{ I_{p+d} - \frac{\mathbf{Z}_s^T \mathbf{W}_s(\mathbf{x}_{-s})\mathbf{Z}_s S_s^{-1}(x_s, \mathbf{x}_{-s})}{\varphi(x_s, \mathbf{x}_{-s})} \right\}^{\ell} + R_s(x_s, \mathbf{x}_{-s})$$

as $n \to \infty$, *where the matrix* $R_s(x_s, \mathbf{x}_{-s})$ *satisfies*

$$\sup_{x_s \in \text{supp}(w_s), \mathbf{x}_{-s} \in \text{supp}(w_{-s})} |R_s(x_s, \mathbf{x}_{-s})| = o\left(h^{p+2}\right) \ a.s.$$

Proof. By a Taylor expansion for the matrix inversion operation, Lemma A.2 immediately yields the result. ∎

Lemma A.4 *Define*

$$D_{s1}(x_s) = \frac{1}{n} \sum_{i=1}^{n} w_{-s}(\mathbf{X}_{i,-s}) R_s(x_s, \mathbf{X}_{i,-s}) \mathbf{Z}_s^T \mathbf{W}_{is} \mathbf{E},$$

$$D_{s2}(x_s) = \frac{1}{n} \sum_{i=1}^{n} w_{-s}(\mathbf{X}_{i,-s}) R_s(x_s, \mathbf{X}_{i,-s}) \mathbf{Z}_s^T \mathbf{W}_{is} \left[\{f_s(X_{js})\}_{j=1}^{n} - \sum_{\nu=0}^{p} \frac{f_s^{(\nu)}(x_s)h^{\nu}}{\nu!} \mathbf{Z}_s e_{\nu} \right],$$

$$D_{s3}(x_s) = \frac{1}{n} \sum_{i=1}^{n} w_{-s}(\mathbf{X}_{i,-s}) R_s(x_s, \mathbf{X}_{i,-s}) \mathbf{Z}_s^T \mathbf{W}_{is}$$

$$\times \left[\left\{ \sum_{s' \neq s} f_{s'}(X_{js'}) \right\}_{j=1}^{n} - \sum_{s' \neq s} f_{s'}(X_{is'}) \mathbf{Z}_s e_{p+s'} \right].$$

22

Then, as $n \to \infty$

$$\sup_{x_s \in \text{supp}(w_s)} \{|D_{s1}(x_s)| + |D_{s2}(x_s)| + |D_{s3}(x_s)|\} = o(h^{p+2}) \ a.s.$$

Proof. The lemma follows directly from Lemmas A.3. ∎

Lemma A.5 *Write* $\mathbf{W}_{is} = \mathbf{W}_s(\mathbf{X}_{i,-s})$ *and* $\mathbf{E} = \{\sigma(\mathbf{X}_1, \mathbf{T}_1)\varepsilon_1, ..., \sigma(\mathbf{X}_n, \mathbf{T}_n)\varepsilon_n\}^T$. *For* $\ell = 1, 2, ...,$ *define*

$$
\begin{aligned}
R_{\ell 1}(x_s) &= \frac{1}{n} \sum_{i=1}^{n} \frac{w_{-s}(\mathbf{X}_{i,-s})}{\varphi(x_s, \mathbf{X}_{i,-s})} S_s^{-1}(x_s, \mathbf{X}_{i,-s}) \left\{ I_{p+d} - \frac{\mathbf{Z}_s^T \mathbf{W}_{is} \mathbf{Z}_s S_s^{-1}(x_s, \mathbf{X}_{i,-s})}{\varphi(x_s, \mathbf{X}_{i,-s})} \right\}^\ell \\
&\quad \times \mathbf{Z}_s^T \mathbf{W}_{is} \mathbf{E}
\end{aligned}
\tag{A.3}
$$

$$
\begin{aligned}
R_{\ell 2}(x_s) &= \frac{1}{n} \sum_{i=1}^{n} \frac{w_{-s}(\mathbf{X}_{i,-s})}{\varphi(x_s, \mathbf{X}_{i,-s})} S_s^{-1}(x_s, \mathbf{X}_{i,-s}) \left\{ I_{p+d} - \frac{\mathbf{Z}_s^T \mathbf{W}_{is} \mathbf{Z}_s S_s^{-1}(x_s, \mathbf{X}_{i,-s})}{\varphi(x_s, \mathbf{X}_{i,-s})} \right\}^\ell \\
&\quad \times \mathbf{Z}_s^T \mathbf{W}_{is is} \left[\{f_s(X_{js})\}_{j=1}^{n} - \sum_{\nu=0}^{p} \frac{f_s^{(\nu)}(x_s) h^\nu}{\nu!} \mathbf{Z}_s e_\nu \right]
\end{aligned}
\tag{A.4}
$$

$$
\begin{aligned}
R_{\ell 3}(x_s) &= \frac{1}{n} \sum_{i=1}^{n} \frac{w_{-s}(\mathbf{X}_{i,-s})}{\varphi(x_s, \mathbf{X}_{i,-s})} S_s^{-1}(x_s, \mathbf{X}_{i,-s}) \left\{ I_{p+d} - \frac{\mathbf{Z}_s^T \mathbf{W}_{is} \mathbf{Z}_s S_s^{-1}(x_s, \mathbf{X}_{i,-s})}{\varphi(x_s, \mathbf{X}_{i,-s})} \right\}^\ell \\
&\quad \times \mathbf{Z}_s^T \mathbf{W}_{is} \left[\left\{ \sum_{s' \neq s} f_{s'}(X_{js'}) \right\}_{j=1}^{n} - \sum_{s' \neq s} f_{s'}(X_{is'}) \mathbf{Z}_s e_{p+s'} \right].
\end{aligned}
\tag{A.5}
$$

Then, as $n \to \infty$,

$$|R_{\ell 1}(x_s)| + |R_{\ell 2}(x_s)| + |R_{\ell 3}(x_s)| = o_p\left(b^\ell / \sqrt{nh}\right). \tag{A.6}$$

Proof. For simplicity of notations, consider the case of $R_{\ell 1}(x_s)$ and only $\ell = 1$. The term $R_{\ell 1}(x_s)$ equals $P_1 - P_2$ in which

$$
\begin{aligned}
P_1 &= \frac{1}{n} \sum_{i=1}^{n} w_{-s}(\mathbf{X}_{i,-s}) S_s^{-1}(x_s, \mathbf{X}_{i,-s}) \left\{ \frac{S(x_s, \mathbf{X}_{i,-s})}{\varphi(x_s, \mathbf{X}_{i,-s})} - \frac{E\left(\mathbf{Z}_s^T \mathbf{W}_{is} \mathbf{Z}_s | x_s, \mathbf{X}_{i,-s}\right)}{\varphi^2(x_s, \mathbf{X}_{i,-s})} \right\} \\
&\quad \times S_s^{-1}(x_s, \mathbf{X}_{i,-s}) \mathbf{Z}_s^T \mathbf{W}_{is} \mathbf{E},
\end{aligned}
$$

$$
\begin{aligned}
P_2 &= \frac{1}{n} \sum_{i=1}^{n} w_{-s}(\mathbf{X}_{i,-s}) S_s^{-1}(x_s, \mathbf{X}_{i,-s}) \left\{ \frac{\mathbf{Z}_s^T \mathbf{W}_{is} \mathbf{Z}_s}{\varphi(x_s, \mathbf{X}_{i,-s})} - \frac{E\left(\mathbf{Z}_s^T \mathbf{W}_{is} \mathbf{Z}_s | x_s, \mathbf{X}_{i,-s}\right)}{\varphi^2(x_s, \mathbf{X}_{i,-s})} \right\} \\
&\quad \times S_s^{-1}(x_s, \mathbf{X}_{i,-s}) \mathbf{Z}_s^T \mathbf{W}_{is} \mathbf{E}.
\end{aligned}
$$

Denote $\xi_i = (\mathbf{X}_i, \mathbf{T}_i, Y_i)$, The term P_1 can be written as the von Mises' differentiable statistic $V_n = (2n^2)^{-1} \sum_{i,j=1}^n g_n(\xi_i, \xi_j)$ where $g_n(\xi_i, \xi_j)$ equals

$$w_{-s}(\mathbf{X}_{i,-s}) S_s^{-1}(x_s, \mathbf{X}_{i,-s}) \left\{ \frac{S(x_s, \mathbf{X}_{i,-s})}{\varphi(x_s, \mathbf{X}_{i,-s})} - \frac{E\left(\mathbf{Z}_s^T \mathbf{W}_{is} \mathbf{Z}_s | x_s, \mathbf{X}_{i,-s}\right)}{\varphi^2(x_s, \mathbf{X}_{i,-s})} \right\}$$

$$\times S_s^{-1}(x_s, \mathbf{X}_{i,-s}) \begin{pmatrix} T_{js} K_h(X_{js} - x_s) L_{\mathbf{g}}(\mathbf{X}_{j,-s} - \mathbf{X}_{i,-s}) \sigma(\mathbf{X}_j, \mathbf{T}_j) \varepsilon_j \\ \mathbf{p}\{(X_{js} - x_s)/h\} T_{js} K_h(X_{js} - x_s) L_{\mathbf{g}}(\mathbf{X}_{j,-s} - \mathbf{X}_{i,-s}) \sigma(\mathbf{X}_j, \mathbf{T}_j) \varepsilon_j \\ \mathbf{T}_{j,-s} K_h(X_{js} - x_s) L_{\mathbf{g}}(\mathbf{X}_{j,-s} - \mathbf{X}_{i,-s}) \sigma(\mathbf{X}_j, \mathbf{T}_j) \varepsilon_j \end{pmatrix}$$

$$+ w_{-s}(\mathbf{X}_{j,-s}) S_s^{-1}(x_s, \mathbf{X}_{j,-s}) \left\{ \frac{S(x_s, \mathbf{X}_{j,-s})}{\varphi(x_s, \mathbf{X}_{j,-s})} - \frac{E\left(\mathbf{Z}_s^T \mathbf{W}_{js} \mathbf{Z}_s | x_s, \mathbf{X}_{j,-s}\right)}{\varphi^2(x_s, \mathbf{X}_{j,-s})} \right\}$$

$$\times S_s^{-1}(x_s, \mathbf{X}_{j,-s}) \begin{pmatrix} T_{is} K_h(X_{is} - x_s) L_{\mathbf{g}}(\mathbf{X}_{i,-s} - \mathbf{X}_{j,-s}) \sigma(\mathbf{X}_i, \mathbf{T}_i) \varepsilon_i \\ \mathbf{p}\{(X_{is} - x_s)/h\} T_{js} K_h(X_{is} - x_s) L_{\mathbf{g}}(\mathbf{X}_{i,-s} - \mathbf{X}_{j,-s}) \sigma(\mathbf{X}_i, \mathbf{T}_i) \varepsilon_i \\ \mathbf{T}_{i,-s} K_h(X_{is} - x_s) L_{\mathbf{g}}(\mathbf{X}_{i,-s} - \mathbf{X}_{j,-s}) \sigma(\mathbf{X}_i, \mathbf{T}_i) \varepsilon_i \end{pmatrix}.$$

First, one calculates that $g_{n,0} = 0$ and $g_{n,1}(\xi_j)$ equals

$$\int S_s^{-1}(x_s, \mathbf{z}_{-s}) w_{-s}(\mathbf{z}_{-s}) S_s^{-1}(x_s, \mathbf{z}_{-s}) \cdot \left\{ \frac{S(x_s, \mathbf{z}_{-s})}{\varphi(x_s, \mathbf{z}_{-s})} - \frac{E\left(\mathbf{Z}_s^T \mathbf{W}_{is} \mathbf{Z}_s | x_s, \mathbf{z}_{-s}\right)}{\varphi^2(x_s, \mathbf{z}_{-s})} \right\}$$

$$\times S_s^{-1}(x_s, \mathbf{z}_{-s}) \begin{pmatrix} T_{js} K_h(X_{js} - x_s) L_{\mathbf{g}}(\mathbf{X}_{j,-s} - \mathbf{z}_{-s}) \sigma(\mathbf{X}_j, \mathbf{T}_j) \varepsilon_j \\ \mathbf{p}\{(X_{js} - x_s)/h\} T_{js} K_h(X_{js} - x_s) L_{\mathbf{g}}(\mathbf{X}_{j,-s} - \mathbf{z}_{-s}) \sigma(\mathbf{X}_j, \mathbf{T}_j) \varepsilon_j \\ \mathbf{T}_{j,-s} K_h(X_{js} - x_s) L_{\mathbf{g}}(\mathbf{X}_{j,-s} - \mathbf{z}_{-s}) \sigma(\mathbf{X}_j, \mathbf{T}_j) \varepsilon_j \end{pmatrix}$$

$$\times \varphi_{-s}(\mathbf{z}_{-s}) d\mathbf{z}_{-s}$$

which has mean zero and variance of order b^2/nh. So $V_n^{(1)} = n^{-1} \sum_{j=1}^n g_{n,1}(\xi_j) = o_p\left(b/\sqrt{nh}\right)$. Next, take a small constant $\delta > 0$. Then, the $(2+\delta)$-th moment of $g_n(\xi_i, \xi_j)$, $i < j$, is not greater than

$$Cb^{2+\delta} C(\rho) \left(\frac{1}{g_{\text{prod}}^{1+2\delta}} E \left| \begin{array}{c} T_{js} K_h(X_{js} - x_s) \sigma(\mathbf{X}_j, \mathbf{T}_j) \varepsilon_j \\ \mathbf{p}\{(X_{js} - x_s)/h\} T_{js} K_h(X_{js} - x_s) \sigma(\mathbf{X}_j, \mathbf{T}_j) \varepsilon_j \\ \mathbf{T}_{j,-s} K_h(X_{js} - x_s) \sigma(\mathbf{X}_j, \mathbf{T}_j) \varepsilon_j \end{array} \right|^{2+2\delta} \right)^{(2+\delta)/(2+2\delta)}$$

$$\leq Cb^{2+\delta} C(\rho) \left(\frac{1}{h^{1+2\delta} g_{\text{prod}}^{1+2\delta}} \right)^{(2+\delta)/(2+2\delta)}$$

by Lemma 1 of Yoshihara (1976).

Hence, one can take $M_n = M_{n,0} = Cb^{2+\delta} \left(h^{1+2\delta} g_{\text{prod}}^{1+2\delta}\right)^{-(2+\delta)/(2+2\delta)}$ in the context of Lemma A.1 with $m = c = 2$ and $r = 1$. Similarly, one can show that $M_{n,1} = Cb^{2+\delta} h^{-(1+\delta)} g_{\text{prod}}^{-(2+\delta)}$.

Now applying Lemma A.1 with $m = c = 2$ and $r = 1$, (A.2) gives

$$
\begin{aligned}
EP_1^2 &\leq Cn^{-2}b^2 \left(hg_{\text{prod}}\right)^{-2(1+2\delta)/(2+2\delta)} + Cn^{-3}b^2 h^{-(1+\delta)2/(2+\delta)} g_{\text{prod}}^{-(2+\delta)2/(2+\delta)} + Cb^2/nh \\
&\leq Cn^{-1}h^{-1}b^2
\end{aligned}
$$

by making δ sufficiently small. Similar arguments establish that $EP_2^2 \leq Cn^{-1}h^{-1}b^2$. Hence $P_1 - P_2 = o_p(b/\sqrt{nh})$. We have thus concluded the proof of the lemma. ∎

Now write $\mathbf{q}_s(u; \mathbf{t})$ for the $(p+d)$-dimensional vector given by

$$
\mathbf{q}_s(u; \mathbf{t})^T = \left(\mathbf{p}(u)^T t_s, \mathbf{t}_{-s}^T\right) = (t_s, ut_s, \ldots, u^p t_s, \mathbf{t}_{-s}^T),
$$

and define an equivalent kernel

$$
K_s^*(u; \mathbf{t}, \mathbf{x}) = e_0^T S_s^{-1}(\mathbf{x}) \mathbf{q}_s(u; \mathbf{t}) K(u). \tag{A.7}
$$

Write $K_{s,h}^*(u; \mathbf{t}, \mathbf{x}) = (1/h) K_s^*(u/h; \mathbf{t}, \mathbf{x})$, i.e.

$$
K_{s,h}^*(u; \mathbf{t}, \mathbf{x}) = (1/h) e_0^T S_s^{-1}(\mathbf{x}) \mathbf{q}_s(u/h; \mathbf{t}) K(u/h). \tag{A.8}
$$

This kernel satisfies the moment conditions as are given in the following lemma, which follows directly from the definition of $S_s(\mathbf{x})$ and $S_s^{-1}(\mathbf{x})$.

Lemma A.6 *Let δ_{jk} equal 1 if $j = k$ and 0 otherwise. Then*

$$
\begin{aligned}
E\left\{\int u^q T_s K_s^*(u; \mathbf{T}, \mathbf{X}) du | \mathbf{X} = \mathbf{x}\right\} &= \delta_{0q}, \quad 0 \leq q \leq p; \\
E\left\{\int T_{s'} K_s^*(u; \mathbf{T}, \mathbf{X}) du | \mathbf{X} = \mathbf{x}\right\} &= 0, \qquad s' = 1, \ldots, d, s' \neq s.
\end{aligned} \tag{A.9}
$$

In order to prove Theorem 1, we begin by observing

$$
e_0^T \left(\mathbf{Z}_s^T \mathbf{W}_{is} \mathbf{Z}_s\right)^{-1} \mathbf{Z}_s^T \mathbf{W}_{is} \mathbf{Z}_s e_l = \delta_{0l}, \quad l = 0, \ldots, p + d - 1.
$$

Define $Q_{1n} = \sum_{i=1}^n w_{-s}(\mathbf{X}_{i,-s})/n$ and

$$
\begin{aligned}
Q_{2n}(x_s) &= n^{-1} \sum_{i=1}^n w_{-s}(\mathbf{X}_{i,-s}) e_0^T \left(\mathbf{Z}_s^T \mathbf{W}_{is} \mathbf{Z}_s\right)^{-1} \mathbf{Z}_s^T \mathbf{W}_{is} \left\{ \mathbf{Y} - \sum_{\nu=0}^p \frac{f_s^{(\nu)}(x_s) h^\nu}{\nu!} \mathbf{Z}_s e_\nu \right. \\
&\qquad \left. - \sum_{s' \neq s}^d f_{s'}(X_{is'}) \mathbf{Z}_s e_{p+s'} \right\}
\end{aligned}
$$

Then, we obtain $Q_{1n}\left\{\hat{f}_s(x_s) - f_s(x_s)\right\} = Q_{2n}(x_s)$. By Lemmas A.5, A.4 and A.3 and by the definition of $K_{s,h}^*(u, \mathbf{t}; \mathbf{x})$ in (A.8), we now write

$$
Q_{2n}(x_s) = \sum_{a=1}^3 \left\{ P_{an}(x_s) + \sum_{l=1}^c R_{la}(x_s) + D_{sa}(x_s) \right\} \tag{A.10}
$$

where for $a = 1, 2, 3$

$$P_{an}(x_s) = n^{-2} \sum_{i,j=1}^{n} \frac{w_{-s}(\mathbf{X}_{i,-s})}{\varphi(x_s, \mathbf{X}_{i,-s})} K_{s,h}^* \left(X_{js} - x_s; T_j, x_s, \mathbf{X}_{i,-s} \right) L_g(\mathbf{X}_{j,-s} - \mathbf{X}_{i,-s}) H_{js} \quad \text{(A.11)}$$

with H_{js} being $\sigma(\mathbf{X}_j, T_j)\varepsilon_j$ for $a = 1$, $\{f_s(X_{js}) - \sum_{\nu=0}^{p} f_s^{(\nu)}(x_s)(X_{js} - x_s)^{\nu}/\nu!\}T_{js}$ for $a = 2$ and $\sum_{s'=1, s' \neq s}^{d} \{f_{s'}(X_{js'}) - f_{s'}(X_{is'})\}T_{js'}$.

In the following three lemmas, we derive the asymptotics for P_{1n}, P_{2n} and P_{3n}.

Lemma A.7 *As $n \to \infty$,*

$$P_{1n}(x_s) = n^{-1} \sum_{j=1}^{n} p_{js}(x_s)\varepsilon_j + o_p\{(nh \log n)^{-1/2}\}$$

where $p_{js}(x_s) = w_{-s}(\mathbf{X}_{j,-s}) K_{s,h}^* \left(X_{js} - x_s; T_j, x_s, \mathbf{X}_{j,-s} \right) \varphi_{-s}(\mathbf{X}_{j,-s})\sigma(\mathbf{X}_j, T_j)/\varphi(x_s, \mathbf{X}_{j,-s})$.

Proof. By the definition (A.11) and using Lemma A.1 for geometrically β-mixing processes,

$$P_{1n}(x_s) = n^{-1} \sum_{j=1}^{n} \int \frac{w_{-s}(\mathbf{x}_{-s})}{\varphi(x_s, \mathbf{x}_{-s})} K_{s,h}^* \left(X_{js} - x_s; T_j, x_s, \mathbf{x}_{-s} \right)$$

$$\times L_g(\mathbf{X}_{j,-s} - \mathbf{x}_{-s})\varphi_{-s}(\mathbf{x}_{-s})d\mathbf{x}_{-s}\sigma(\mathbf{X}_j, T_j)\varepsilon_j + o_p\{(nh \log n)^{-1/2}\}.$$

By the change of variable $\mathbf{x}_{-s} = \mathbf{X}_{j,-s} - \mathbf{g}\mathbf{v}_{-s}$ and the fact that L is of order q, it equals

$$n^{-1} \sum_{j=1}^{n} \frac{w_{-s}(\mathbf{X}_{j,-s})}{\varphi(x_s, \mathbf{X}_{j,-s})} K_{s,h}^* \left(X_{js} - x_s; T_j, x_s, \mathbf{X}_{j,-s} \right) \varphi_{-s}(\mathbf{X}_{j,-s})\sigma(\mathbf{X}_j, T_j)\varepsilon_j$$

$$+ o_p\{(nh \log n)^{-1/2}\}.$$

This completes the proof of the lemma. ∎

Lemma A.8 *As $n \to \infty$, $P_{2n}(x_s) = \kappa_s(x_s)h^{p+1} + o_p(h^{p+1})$ where*

$$\kappa_s(x_s) = (p+1)!^{-1} f_s^{(p+1)}(x_s) \int u^{p+1} E\left\{ w_{-s}(\mathbf{X}_{-s})T_s K_s^*(u; \mathbf{T}, x_s, \mathbf{X}_{-s}) \right\} du.$$

Proof. By definition (A.11) and again using Lemma A.1, we derive

$$P_{2n}(x_s) = \int \frac{w_{-s}(\mathbf{x}_{-s})}{\varphi(x_s, \mathbf{x}_{-s})} K_{s,h}^* \left(z_s - x_s; \mathbf{t}, x_s, \mathbf{x}_{-s} \right) L_g(\mathbf{z}_{-s} - \mathbf{x}_{-s})$$

$$\times \left\{ f_s(z_s) - \sum_{\nu=0}^{p} f_s^{(\nu)}(x_s)(z_s - x_s)^{\nu}/\nu! \right\} t_s \psi(\mathbf{z}, \mathbf{t})\varphi_{-s}(\mathbf{x}_{-s}) d\mathbf{z}d\mathbf{t}d\mathbf{x}_{-s} \{1 + o_p(1)\}.$$

26

By the changes of variables $z_s = x_s + hu$ and $\mathbf{z}_{-s} = \mathbf{x}_{-s} + \mathbf{g}\mathbf{v}_{-s}$, we obtain

$$
\begin{aligned}
P_{2n}(x_s) &= h^{p+1}(p+1)!^{-1} \int \frac{w_{-s}(\mathbf{x}_{-s})}{\varphi(x_s, \mathbf{x}_{-s})} K_s^* (u; \mathbf{t}, x_s, \mathbf{x}_{-s}) f_s^{(p+1)}(x_s) u^{p+1} t_s \\
&\quad \times \varphi_{-s}(\mathbf{x}_{-s})\psi(x_s, \mathbf{x}_{-s}, \mathbf{t}) du d\mathbf{x}_{-s} d\mathbf{t} \{1 + o_p(1)\} \\
&= h^{p+1}(p+1)!^{-1} f_s^{(p+1)}(x_s) E\Big[w_{-s}(\mathbf{X}_{-s}) \int K_s^* (u; \mathbf{t}, x_s, \mathbf{X}_{-s}) u^{p+1} t_s \\
&\quad \psi(\mathbf{t}|x_s, \mathbf{X}_{-s}) du d\mathbf{t} \Big] + o_p(h^{p+1}).
\end{aligned}
$$

This completes the proof of the lemma. ∎

Lemma A.9 *As* $n \to \infty$, $P_{3n}(x_s) = O_p\left(g_{\max}^q\right)$.

Proof. By definition (A.11) and applying Lemma A.1, one has

$$
\begin{aligned}
P_{3n}(x_s) &= \int \frac{w_{-s}(\mathbf{x}_{-s})}{\varphi(x_s, \mathbf{x}_{-s})} K_{s,h}^* (z_s - x_s; \mathbf{t}, x_s, \mathbf{x}_{-s}) L_{\mathbf{g}}(\mathbf{z}_{-s} - \mathbf{x}_{-s}) \\
&\quad \times \Big[\sum_{s' \neq s} \{f_{s'}(z_{s'}) - f_{s'}(x_{s'})\} t_{s'} \Big] \psi(\mathbf{z}, \mathbf{t}) \varphi_{-s}(\mathbf{x}_{-s}) d\mathbf{z} d\mathbf{t} d\mathbf{x}_{-s} \{1 + o_p(1)\}.
\end{aligned}
$$

After the changes of variables $\mathbf{z}_{-s} = \mathbf{x}_{-s} + \mathbf{g}\mathbf{v}_{-s}$ and $z_s = x_s + hu$, we obtain

$$
\begin{aligned}
P_{3n}(x_s) &= \int \frac{w_{-s}(\mathbf{x}_{-s})}{\varphi(x_s, \mathbf{x}_{-s})} K_s^* (u; \mathbf{t}, x_s, \mathbf{x}_{-s}) L(\mathbf{v}_{-s}) \Big[\sum_{s' \neq s} \{f_{s'}(x_{s'} + g_{s'} v_{s'}) - f_{s'}(x_{s'})\} t_{s'} \Big] \\
&\quad \times \psi(x_s + hu, \mathbf{x}_{-s} + \mathbf{g}\mathbf{v}_{-s}, \mathbf{t}) \varphi_{-s}(\mathbf{x}_{-s}) du d\mathbf{v}_{-s} d\mathbf{t} d\mathbf{x}_{-s} \{1 + o_p(1)\} \\
&= O_p(g_{\max}^q)
\end{aligned}
$$

since L is of order q by the assumption A1. Thus, we have proved the lemma. ∎

Proof of Theorem 1. By Lemma A.7 and the martingale central limit theorem of Liptse and Shirjaev (1980), $\sqrt{nh} P_{1n}(x_s)$ for each $x_s \in \text{supp}(w_s)$ is asymptotically normal with mean 0 and variance

$$
h \int \frac{w_{-s}^2(\mathbf{z}_{-s})}{\varphi^2(x_s, \mathbf{z}_{-s})} K_{s,h}^{*2} (z_s - x_s; \mathbf{t}, x_s, \mathbf{z}_{-s}) \varphi_{-s}^2(\mathbf{z}_{-s}) \sigma^2(\mathbf{z}, \mathbf{t}) \psi(\mathbf{z}, \mathbf{t}) d\mathbf{z} d\mathbf{t} \{1 + o(1)\}.
$$

By the change of variable $z_s = x_s + hu$, the leading term of this equals

$$
\tau_s^2(x_s) = \int \frac{w_{-s}^2(\mathbf{z}_{-s})}{\varphi^2(x_s, \mathbf{z}_{-s})} K_s^{*2} (u; \mathbf{t}, x_s, \mathbf{z}_{-s}) \varphi_{-s}^2(\mathbf{z}_{-s}) \sigma^2(x_s, \mathbf{z}_{-s}, \mathbf{t}) \psi(x_s, \mathbf{z}_{-s}, \mathbf{t}) du d\mathbf{z}_{-s} d\mathbf{t}.
$$

The theorem now follows immediately from Lemmas A.7, A.8, the conditions on the bandwidths as given in A7, and the fact that $Q_{1n} = \int w_{-s}(\mathbf{z}_{-s}) \varphi_{-s}(\mathbf{z}_{-s}) d\mathbf{z}_{-s} + O_p(n^{-1/2})$. ∎

Proof of Theorem 2. One first notes that (8) follows directly from (7), so we will only show the latter. Now, from Lemmas A.7, A.8, A.9 and the conditions on the bandwidths, we obtain

$$\hat{f}_s(x_s) - f_s(x_s) = b_s(x_s)h^{p+1} + n^{-1}\eta_s^{-1}\sum_{j=1}^{n} p_{js}(x_s)\varepsilon_j + o_p(h^{p+1}). \qquad (A.12)$$

Applying (A.12), one only needs to show that the two stochastic terms $n^{-1}\sum_{j=1}^{n} p_{js}(x_s)\varepsilon_j$ and $n^{-1}\sum_{j=1}^{n} p_{js'}(x_{s'})\varepsilon_j$ for $s \neq s'$ have covariance of order $o(n^{-1}h^{-1})$. Noting that the ε_j's are i.i.d. white noise and each ε_i is independent of the vectors $(\mathbf{X}_j, \mathbf{T}_j), j = 1, ..., i$ for each $i = 1, ..., n$, we need only to show that

$$E\left\{p_{js}(x_s)p_{js'}(x_{s'})\right\} = o(h^{-1}). \qquad (A.13)$$

By change of variables technique for X_s and $X_{s'}$ which are contained in $p_{js}(x_s)$ and $p_{js'}(x_{s'})$ respectively, one may show that the left hand side of (A.13) is actually $O(1)$, which proves the theorem. ■

Proof of Theorem 5. For this proof, we use (A.10) again. Under the hypothesis (11), $P_{2n}(x_s) = R_{l2}(x_s) = D_{s2}(x_s) = 0$ and thus

$$Q_{1n}\left\{\hat{f}_s(x_s) - \alpha\right\} = P_{1n}(x_s) + \sum_{l=1}^{c} R_{l1}(x_s) + D_{s1}(x_s) + P_{3n}(x_s) + \sum_{l=1}^{c} R_{l3}(x_s) + D_{s3}(x_s).$$

Hence to study $\sum_{k=1}^{n} \hat{f}_s(X_{ks})^2 w_s(X_{ks})/n$, we derive the asymptotics of such as $\sum_{k=1}^{n} w_s(X_{ks}) P_{1n}^2(X_{ks})/n$. Let $\xi_i = (\mathbf{X}_i, \mathbf{T}_i, Y_i)$ and define

$$\begin{aligned}
\tilde{g}_n\left(\xi_i, \xi_j, \xi_k, \xi_l, \xi_m\right) &= w_s(X_{ks})\frac{w_{-s}(\mathbf{X}_{i,-s})}{\varphi(X_{ks}, \mathbf{X}_{i,-s})}K_{s,h}^*\left(X_{js} - X_{ks}; \mathbf{T}_j, X_{ks}, \mathbf{X}_{i,-s}\right) \\
&\quad \times L_g(\mathbf{X}_{j,-s} - \mathbf{X}_{i,-s})\sigma(\mathbf{X}_j, \mathbf{T}_j)\varepsilon_j \\
&\quad \times \frac{w_{-s}(\mathbf{X}_{l,-s})}{\varphi(X_{ks}, \mathbf{X}_{l,-s})}K_{s,h}^*\left(X_{ms} - X_{ks}; \mathbf{T}_m, X_{ks}, \mathbf{X}_{l,-s}\right) \\
&\quad \times L_g(\mathbf{X}_{m,-s} - \mathbf{X}_{l,-s})\sigma(\mathbf{X}_m, \mathbf{T}_m)\varepsilon_m.
\end{aligned}$$

Then, by the definition (A.11)

$$\sum_{k=1}^{n} w_s(X_{ks})P_{1n}^2(X_{ks})/n = n^{-5}\sum_{i,j,k,l,m=1}^{n} \tilde{g}_n\left(\xi_i, \xi_j, \xi_k, \xi_l, \xi_m\right).$$

Next, we define $g_n\left(\xi_i, \xi_j, \xi_k, \xi_l, \xi_m\right) = \sum \tilde{g}_n\left(\xi_{i'}, \xi_{j'}, \xi_{k'}, \xi_{l'}, \xi_{m'}\right)/5!$, where the sum is over all possible permutations (i', j', k', l', m') of (i, j, k, l, m). Then $\sum_{k=1}^{n} w_s(X_{ks})P_{1n}^2(X_{ks})/n$ is

28

expressed as a V statistic $n^{-5} \sum_{i,j,k,l,m=1}^{n} g_n (\xi_i, \xi_j, \xi_k, \xi_l, \xi_m)$. It is easy to see that $g_{n,0} = 0$, $g_{n,1} = 0$, and by changes of variables $g_{n,2} (\xi_j, \xi_m)$ equals

$$
\begin{aligned}
& \sigma(\mathbf{X}_j, \mathbf{T}_j)\sigma(\mathbf{X}_m, \mathbf{T}_m)\varepsilon_j\varepsilon_m \int \frac{w_s(X_{js} - hu_{ks})w_{-s}(\mathbf{X}_{j,-s} - \mathbf{g}\mathbf{u}_{i,-s})w_{-s}(\mathbf{X}_{m,-s} - \mathbf{g}\mathbf{u}_{l,-s})}{\varphi(X_{js} - hu_{ks}, \mathbf{X}_{j,-s} - \mathbf{g}\mathbf{u}_{i,-s})\varphi(X_{js} - hu_{ks}, \mathbf{X}_{m,-s} - \mathbf{g}\mathbf{u}_{l,-s})} \\
& \times K_s^* \left(u_{ks}; \mathbf{T}_j, X_{js} - hu_{ks}, \mathbf{X}_{j,-s} - \mathbf{g}\mathbf{u}_{i,-s}\right) L(\mathbf{u}_{i,-s})L(\mathbf{u}_{l,-s}) \\
& \times K_{s,h}^* \left(X_{ms} - X_{js} + hu_{ks}; \mathbf{T}_m, X_{js} - hu_{ks}, \mathbf{X}_{m,-s} - \mathbf{g}\mathbf{u}_{l,-s}\right) \\
& \times \psi(x_{is}, \mathbf{X}_{j,-s} - \mathbf{g}\mathbf{u}_{i,-s}, \mathbf{t}_i)\psi(x_{ls}, \mathbf{X}_{m,-s} - \mathbf{g}\mathbf{u}_{l,-s}, \mathbf{t}_l)\psi(X_{js} - hu_{ks}, \mathbf{x}_{k,-s}, \mathbf{t}_k) \\
& \times dx_{is}d\mathbf{u}_{i,-s}dx_{ls}d\mathbf{u}_{l,-s}du_{ks}d\mathbf{x}_{k,-s}d\mathbf{t}_i d\mathbf{t}_l d\mathbf{t}_k.
\end{aligned}
$$

To establish the asymptotic normality of the off-diagonal sum $2\,n^{-2} \sum_{1 \leq j < m \leq n} g_{n,2} (\xi_j, \xi_m)$, we use Lemma 3.2 of Hjellvik, Yao and Tjøstheim (1998). Let δ_n^2 be their σ_n^2, i.e., $\delta_n^2 = \sum_{1 \leq j < m \leq n} \text{var} \{2\,n^{-2} g_{n,2} (\xi_j, \xi_m)\}$. Define λ_{nk} in the same way as their M_{nk} for $k = 1, \ldots, 6$ with $2\,n^{-2} g_{n,2} (\xi_j, \xi_m)$ taking the role of their φ_{jm}. If we prove that for some $\zeta > 0$

$$
n^2\delta_n^{-2} \left(\lambda_{n1}^{1/(\zeta+1)} + \lambda_{n5}^{1/(\zeta+1)} + \lambda_{n6}^{1/2} \right) \longrightarrow 0 \tag{A.14}
$$

$$
n^{3/2}\delta_n^{-2} \left(\lambda_{n2}^{1/\{2(\zeta+1)\}} + \lambda_{n3}^{1/2} + \lambda_{n4}^{1/\{2(\zeta+1)\}} \right) \longrightarrow 0, \tag{A.15}
$$

then we establish that $2\,n^{-2} \sum_{1 \leq j < m \leq n} g_{n,2} (\xi_j, \xi_m)$ is asymptotically normal with mean 0 and variance δ_n^2.

We compute δ_n^2 first. Note that

$$
\begin{aligned}
\delta_n^2 &= \frac{2}{n^2} \int \left\{ \sigma(\mathbf{x}_j, \mathbf{t}_j)\sigma(\mathbf{x}_m, \mathbf{t}_m) \int \frac{w_s(x_{js})w_{-s}(\mathbf{x}_{j,-s})w_{-s}(\mathbf{x}_{m,-s})}{\varphi(x_{js}, \mathbf{x}_{j,-s})\varphi(x_{js}, \mathbf{x}_{m,-s})} \right. \\
& \times K_s^* \left(u_{ks}; \mathbf{t}_j, x_{js}, \mathbf{x}_{j,-s}\right) K_{s,h}^* \left(x_{ms} - x_{js} + hu_{ks}; \mathbf{t}_m, x_{js}, \mathbf{x}_{m,-s}\right) \\
& \times L(\mathbf{u}_{i,-s})L(\mathbf{u}_{l,-s})\psi(x_{is}, \mathbf{x}_{j,-s}, \mathbf{t}_i)\psi(x_{ls}, \mathbf{x}_{m,-s}, \mathbf{t}_l)\psi(x_{js}, \mathbf{x}_{k,-s}, \mathbf{t}_k) \\
& \left. \times dx_{is}d\mathbf{u}_{i,-s}dx_{ls}d\mathbf{u}_{l,-s}du_{ks}d\mathbf{x}_{k,-s}d\mathbf{t}_i d\mathbf{t}_l d\mathbf{t}_k \right\}^2 \psi(\mathbf{x}_j, \mathbf{t}_j)\psi(\mathbf{x}_m, \mathbf{t}_m) \\
& \times d\mathbf{x}_j d\mathbf{x}_m d\mathbf{t}_j d\mathbf{t}_m \left\{1 + O \left(h^{p+1} + g^q\right)\right\}.
\end{aligned}
$$

By the change of variable $x_{ms} = x_{js} + hv_s$ and further approximations of the functions, we obtain $\delta_n^2 = \{1 + O(h^{p+1} + g^q)\}n^{-2}h^{-1}\eta_s^4\gamma_s^2$, where

$$
\begin{aligned}
\gamma_s^2 &= \frac{2}{\eta_s^4} \int \frac{w_{-s}^2(\mathbf{x}_{-s})w_{-s}^2(\mathbf{z}_{-s})w_s^2(x_s)}{\varphi^2(x_s, \mathbf{x}_{-s})\varphi^2(x_s, \mathbf{z}_{-s})} \left\{ K_s^{*(c)} \left(u; \mathbf{t}_1, \mathbf{t}_2, x_s, \mathbf{x}_{-s}, \mathbf{z}_{-s}\right) \right\}^2 \\
& \times \varphi_{-s}^2(\mathbf{x}_{-s})\varphi_{-s}^2(\mathbf{z}_{-s})\sigma^2(x_s, \mathbf{x}_{-s}, \mathbf{t}_1)\sigma^2(x_s, \mathbf{z}_{-s}, \mathbf{t}_2)\varphi_s^2(x_s) \\
& \times \psi(x_s, \mathbf{x}_{-s}, \mathbf{t}_1)\psi(x_s, \mathbf{z}_{-s}, \mathbf{t}_2)dud x_s d\mathbf{x}_{-s}d\mathbf{z}_{-s}d\mathbf{t}_1 d\mathbf{t}_2
\end{aligned} \tag{A.16}
$$

and $K_s^{*(c)} (w; \mathbf{t}_1, \mathbf{t}_2, x_s, \mathbf{x}_{-s}, \mathbf{z}_{-s}) = \int K_s^* (u; \mathbf{t}_1, x_s, \mathbf{x}_{-s}) K_s^* (w + u; \mathbf{t}_2, x_s, \mathbf{z}_{-s}) du$.

Next, we approximate λ_{nj}. We only illustrate the calculation of λ_{n4}. For $j < k$ and $l < m$ with all j, k, l, m different, we obtain

$$E\left|g_{n,2}(\xi_j, \xi_k)g_{n,2}(\xi_l, \xi_m)\right|^{2(1+\zeta)}$$

$$\leq \text{const.}h^{-4(1+\zeta)+2}\left(E|\varepsilon_1|^{2(1+\zeta)}\right)^4 \int \left|\sigma(\mathbf{x}_j, \mathbf{t}_j)\sigma(x_{js} + hv, \mathbf{x}_{k,-s}, \mathbf{t}_k)\sigma(\mathbf{x}_l, \mathbf{t}_l)\right.$$

$$\times \sigma(x_{ls} + hv', \mathbf{x}_{m,-s}, \mathbf{t}_m)\Big|^{2(1+\zeta)}\left|K_s^{*(c)}(v; \mathbf{t}_j, \mathbf{t}_k, x_{js}, \mathbf{x}_{j,-s}, \mathbf{x}_{k,-s})\right.$$

$$\times K_s^{*(c)}(v'; \mathbf{t}_l, \mathbf{t}_m, x_{ls}, \mathbf{x}_{l,-s}, \mathbf{x}_{m,-s})\Big|^{2(1+\zeta)}\psi_{j,k,l,m}(\mathbf{x}_j, \mathbf{t}_j; x_{js} + hv, \mathbf{x}_{k,-s}, \mathbf{t}_k; \mathbf{x}_l, \mathbf{t}_l;$$

$$x_{ls} + hv', \mathbf{x}_{m,-s}, \mathbf{t}_m)\, d\mathbf{x}_j dt_j dv d\mathbf{x}_{k,-s} dt_k d\mathbf{x}_l dt_l dv' d\mathbf{x}_{m,-s} dt_m,$$

where the integrations with respect to $\mathbf{x}_j$, dv, $d\mathbf{x}_{k,-s}$, $\mathbf{x}_l$, dv', $d\mathbf{x}_{m,-s}$ are over compact sets. By the assumption A6 the right hand side of the above inequality is bounded by

$$\text{const.}h^{-4(1+\zeta)+2}\int (\|\mathbf{t}_j\|\|\mathbf{t}_k\|\|\mathbf{t}_l\|\|\mathbf{t}_m\|)^{2(1+\zeta)}|\tilde\sigma(\mathbf{t}_j)\tilde\sigma(\mathbf{t}_k)\tilde\sigma(\mathbf{t}_l)\tilde\sigma(\mathbf{t}_m)|^{2(1+\zeta)}$$

$$\times \tilde\varphi_{j,k,l,m}(\mathbf{t}_j, \mathbf{t}_k, \mathbf{t}_l, \mathbf{t}_m)\, dt_j dt_k dt_l dt_m$$

$$\leq \text{const.}h^{-4(1+\zeta)+2}.$$

This shows

$$n^{3/2}\delta_n^{-2}\lambda_{n4}^{1/\{2(\zeta+1)\}} \asymp n^{3/2} \times n^2 h \times n^{-4}h^{-(1+2\zeta)/(1+\zeta)} = n^{-(2p+2p\zeta+3+\zeta)/\{2(1+\zeta)(2p+3)\}}.$$

Similarly, we can establish

$$n^2\delta_n^{-2}\lambda_{n1}^{1/(\zeta+1)} \asymp n^2 \times n^2 h \times n^{-4}h^{-2\zeta/(\zeta+1)} = h^{(1-\zeta)/(1+\zeta)},$$

$$n^{3/2}\delta_n^{-2}\lambda_{n2}^{1/\{2(\zeta+1)\}} \asymp n^{3/2} \times n^2 h \times n^{-4}h^{-(1+2\zeta)/(1+\zeta)} = n^{-(2p+2p\zeta+3+\zeta)/\{2(1+\zeta)(2p+3)\}},$$

$$n^{3/2}\delta_n^{-2}\lambda_{n3}^{1/2} \asymp n^{3/2} \times n^2 h \times n^{-4}h^{-3/2} = (nh)^{-1/2},$$

$$n^2\delta_n^{-2}\lambda_{n5}^{1/\{2(\zeta+1)\}} \asymp n^2 \times n^2 h \times n^{-4}h^{-(1+2\zeta)/\{2(\zeta+1)\}} = h^{1/\{2(1+\zeta)\}},$$

$$n^2\delta_n^{-2}\lambda_{n6}^{1/2} \asymp n^2 \times n^2 h \times n^{-4}h^{-1/2} = h^{1/2}.$$

Thus, if we take ζ such that $0 < \zeta < 1$, the convergences (A.14) and (A.15) hold.

By the martingale central limit theorem again, the diagonal sum $n^{-2}\sum_{j=1}^n g_{n,2}(\xi_j, \xi_j)$ is also asymptotically normal with mean $\eta_s^2 v_s n^{-1} h^{-1}\{1 + O(h^{p+1})\}$, where v_s is given by

$$v_s = \int \frac{w_{-s}^2(\mathbf{x}_{-s})w_s(x_s)}{\eta_s^2 \varphi^2(\mathbf{x})} K_s^{*2}(u; t, \mathbf{x})\, \varphi_{-s}^2(\mathbf{x}_{-s})\sigma^2(\mathbf{x}, t)\psi(\mathbf{x}, t)\varphi_s(x_s)dudxdt. \qquad (\text{A.17})$$

The asymptotic variance of $n^{-2}\sum_{j=1}^n g_{n,2}(\xi_j, \xi_j)$ is likewise calculated, and may be shown to be of order $n^{-3}h^{-2}$.

Therefore, we establish

$$nh^{1/2}\left\{n^{-2}\sum_{j=1}^{n}\sum_{m=1}^{n}g_{n,2}\left(\xi_j,\xi_m\right)-\frac{\eta_s^2}{nh}v_s\right\}\xrightarrow{\mathcal{L}}N\left(0,\eta_s^4\gamma_s^2\right).\qquad(A.18)$$

Application of Lemma A.1 reveals that $n^{-c}\sum_{j_1,...,j_c=1}^{n}g_{n,c}\left(\xi_{j_1},...,\xi_{j_c}\right)=o\left(n^{-1}h^{-1/2}\right)$ for $c=3,4,5$. Using Lemma A.1 again, now to terms such as $\sum_{k=1}^{n}w_s(X_{ks})P_{3n}^2(X_{ks})/n$, $\sum_{k=1}^{n}w_s(X_{ks})R_{\ell1}^2(X_{ks})/n$ and $\sum_{k=1}^{n}w_s(X_{ks})R_{\ell3}^2(X_{ks})/n$, one may show that they are all of order $o\left(n^{-1}h^{-1/2}\right)$ as well. Using Lemma A.4 one may also prove $\sum_{k=1}^{n}w_s(X_{ks})D_{s1}^2(X_{ks})/n$ and $\sum_{k=1}^{n}w_s(X_{ks})D_{s3}^2(X_{ks})/n$ are both of order $o\left(h^{2p+4}\right)=o\left(n^{-1}h^{-1/2}\right)$. Similar arguments establish that $\{\sum_{i=1}^{n}\hat{f}_s(X_{is})w_s(X_{is})\}^2=o\left(n^{-1}h^{-1/2}\right)$. Hence,

$$V_{ns}=Q_{1n}^{-2}\sum_{k=1}^{n}P_{1n}^2(X_{ks})w_s(X_{ks})/n+o\left(n^{-1}h^{-1/2}\right).$$

This completes the proof of Theorem 5. ∎

A.3 Proofs of Theorems 3, 4 and 6

Define $\mathbf{q}_{su}(v;t)^T=\left(\mathbf{p}(v)^Tt_{su},\mathbf{t}_{-(su)}^T\right)$ and $S_{su}(\mathbf{x})$ in the same way as $S_s(\mathbf{x})$ with T_s and $\mathbf{T}_{-(s)}$ being replaced by T_{su} and $\mathbf{T}_{-(su)}$, respectively. Define an equivalent kernel $K_{(su)}^*(v;t,\mathbf{x})=e_0^T S_{su}^{-1}(\mathbf{x})\mathbf{q}_{su}(v;t)K(v)$. Let $K_{(su)}^{*(c)}$ denote the two-folded convolution of $K_{(su)}^*$. Theorems 3 and 6 can be proved in the same way as in the proofs of Theorems 1 and 5 with the following definitions of κ_{su}, τ_{su}^2, γ_{su} and v_{su}:

$$\kappa_{su}(x_s)=\frac{f_{su}^{(p+1)}(x_s)}{(p+1)!}\int u^{p+1}E\left\{w_{-s}(\mathbf{X}_{-s})T_{su}K_{(su)}^*(v;\mathbf{T},x_s,\mathbf{X}_{-s})\right\}dv,\qquad(A.19)$$

$$\tau_{su}^2(x_s)=\int\frac{w_{-s}^2(\mathbf{z}_{-s})}{\varphi^2(x_s,\mathbf{z}_{-s})}K_{(su)}^{*2}(v;t,x_s,\mathbf{z}_{-s})\varphi_{-s}^2(\mathbf{z}_{-s})\sigma^2(x_s,\mathbf{z}_{-s},t)$$
$$\times\psi(x_s,\mathbf{z}_{-s},t)dvd\mathbf{z}_{-s}dt,\qquad(A.20)$$

$$\gamma_{su}^2=\frac{2}{\eta_s^4}\int\frac{w_{-s}^2(\mathbf{x}_{-s})w_{-s}^2(\mathbf{z}_{-s})w_s^2(x_s)}{\varphi^2(x_s,\mathbf{x}_{-s})\varphi^2(x_s,\mathbf{z}_{-s})}\left\{K_{(su)}^{*(c)}(v;t_1,t_2,x_s,\mathbf{x}_{-s},\mathbf{z}_{-s})\right\}^2$$
$$\times\varphi_{-s}^2(\mathbf{x}_{-s})\varphi_{-s}^2(\mathbf{z}_{-s})\sigma^2(x_s,\mathbf{x}_{-s},t_1)\sigma^2(x_s,\mathbf{z}_{-s},t_2)\varphi_s^2(x_s)\qquad(A.21)$$
$$\times\psi(x_s,\mathbf{x}_{-s},t_1)\psi(x_s,\mathbf{z}_{-s},t_2)dvdx_sd\mathbf{x}_{-s}d\mathbf{z}_{-s}dt_1dt_2,$$

$$v_{su}=\int\frac{w_{-s}^2(\mathbf{x}_{-s})w_s(x_s)}{\eta_s^2\varphi^2(\mathbf{x})}K_{(su)}^{*2}(v;t,\mathbf{x})\varphi_{-s}^2(\mathbf{x}_{-s})\sigma^2(\mathbf{x},t)\psi(\mathbf{x},t)$$
$$\times\varphi_s(x_s)dvd\mathbf{x}dt.\qquad(A.22)$$

The proof of Theorem 4(i) is the same as that of the first part of Theorem 2. For the

proof of Theorem 4(ii), define

$$p_{jsu}(x_s) = \frac{w_{-s}(\mathbf{X}_{j,-s})}{\varphi(x_s, \mathbf{X}_{j,-s})} K^*_{(su),h}\left(X_{js} - x_s; \mathbf{T}_j, x_s, \mathbf{X}_{j,-s}\right) \varphi_{-s}(\mathbf{X}_{j,-s})\sigma(\mathbf{X}_j, \mathbf{T}_j)$$

where $K^*_{(su),h}(v; \mathbf{t}, \mathbf{x}) = (1/h)K^*_{(su)}(v/h; \mathbf{t}, \mathbf{x})$. We observe

$$\hat{f}_{su}(x_s) - f_{su}(x_s) = b_{su}(x_s)h^{p+1} + n^{-1}\eta_s^{-1}\sum_{j=1}^{n} p_{jsu}(x_s)\varepsilon_j + o_p(h^{p+1}).$$

Thus for the case $s = s'$, we obtain $\mathrm{cov}\left(\hat{f}_{su}(x_s), \hat{f}_{su'}(x_s)\right) = \eta_s^{-2}\tau_{suu'}(x_s)n^{-1}h^{-1}\{1 + o(1)\}$ where

$$\tau_{suu'}(x_s) = \int \frac{w_{-s}^2(\mathbf{z}_{-s})}{\varphi^2(x_s, \mathbf{z}_{-s})} K^*_{(su)}(v; \mathbf{t}, x_s, \mathbf{z}_{-s}) K^*_{(su')}(v; \mathbf{t}, x_s, \mathbf{z}_{-s}) \varphi_{-s}^2(\mathbf{z}_{-s})$$
$$\times \sigma^2(x_s, \mathbf{z}_{-s}, \mathbf{t})\psi(x_s, \mathbf{z}_{-s}, \mathbf{t})dvd\mathbf{z}_{-s}d\mathbf{t}, \tag{A.23}$$

We note that $\tau_{su}^2(x_s) = \tau_{suu}(x_s)$. This completes the proof of Theorem 4.

REFERENCES

Bosq, D. (1998), *Nonparametric Statistics for Stochastic Processes*, New York: Springer-Verlag.

Cai, Z., Fan, J. and Li, R. Z. (2000), "Efficient Estimation and Inferences for Varying-Coefficient Models", *Journal of the American Statistical Association*, 95, 888-902.

Cai, Z., Fan, J. and Yao, Q. (2000), "Functional-Coefficient Regression Models for Nonlinear Times Series", *Journal of the American Statistical Association*, 95, 941-956.

Chen, R. and Liu, L. M. (1993a), "Functional Coefficient Autoregressive Models: Estimation and Tests of Hypotheses", *Journal of Time Series Analysis*, 22, 151-173.

Chen, R. and Tsay, R. S. (1993a), "Nonlinear Additive ARX Models", *Journal of the American Statistical Association*, 88, 955-967.

Chen, R. and Tsay, R. S. (1993b), "Functional-Coefficient Autoregressive Models", *Journal of the American Statistical Association*, 88, 298-308.

Fan, J. and Gijbels, I. (1996), *Local Polynomial Modelling and Its Applications*, London: Chapman and Hall.

Fan, J. Härdle, W. and Mammen, E. (1998), "Direct Estimation of Low-Dimensional Components in Additive Models", *The Annals of Statistics*, 26, 943-971.

Franses, H. F. (1996), *Periodicity and Stochastic Trends in Economic Time Series*, Oxford: Oxford University Press.

Härdle, W., Hlavka, Z. and Klinke, S. (2000), *XploRe Application Guide*, Heidelberg: Springer Verlag.

Härdle, W., Liang, H. and Gao, J. T. (2000), *Partially Linear Models*, Heidelberg: Springer-Verlag.

Härdle, W. and Mammen, E. (1993), "Comparing Nonparametric versus Parametric Regression Fits", *The Annals of Statistics*, 21, 1926-1947.

Hastie, T. J. and Tibshirani, R. J. (1990), *Generalized Additive Models*, London: Chapman and Hall.

Hastie, T. J. and Tibshirani, R. J. (1993), "Varying-Coefficient Models", *Journal of the Royal Statistical Society Series B*, 55, 757-796.

Hjellvik, V., Yao, Q. and Tjøstheim, D. (1998), "Linearity testing using local polynomial approximation", *Journal of Statistical Planning and Inference*, 68, 295-321

Linton, O. B. (1997), "Efficient Estimation of Additive Nonparametric Regression Models", *Biometrika*, 84, 469-473.

Linton, O. B. and Härdle, W. (1996), "Estimation of Additive Regression Models with Known Links", *Biometrika*, 83, 529-540.

Linton, O. B. and Nielsen, J. P. (1995), "A Kernel Method of Estimating Structured Nonparametric Regression Based on Marginal Integration", *Biometrika*, 82, 93-100.

Liptser, R. Sh. and Shirjaev, A. N. (1980), "A Functional Central Limit Theorem for Martingales", *Theory of Probability and Applications*, 25, 667-688.

Liu, R. (1988), "Bootstrap Procedures Under Some Non I.I.D. Models", *The Annals of Statistics*, 16, 1696-1708.

Mammen, E. (1992), *When Does Bootstrap Work: Asymptotic Results and Simulations*, Lecture Notes in Statistics 77, Berlin: Springer-Verlag.

33

Mammen, E. Linton, O. and Nielsen, J. (1999), "The Existence and Asymptotic Properties of a Backfitting Projection Algorithm under Weak Conditions", *The Annals of Statistics*, 5, 1443-1490.

Masry, E. and Tjøstheim, D. (1995), "Non-parametric Estimation and Identification of ARCH Nonlinear Time Series: Strong Convergence and Asymptotic Normality", *Econometric Theory*, 11, 258-289.

Masry, E. and Tjøstheim, D. (1997), "Additive Nonlinear ARX Time Series and Projection Estimates", *Econometric Theory*, 13, 214-252.

Ruppert, D., Sheather, S. J. and Wand, M. P. (1995), "An Effective Bandwidth Selection for Local Least Squares Regression", *Journal of the American Statistical Association*, 90, 1257-1270.

Sperlich, S., Tjøstheim, D. and Yang, L. (2002), "Nonparametric Estimation and Testing of Interaction in Additive Models", *Econometric Theory*, 18, 197-251.

Stone, C. J. (1977), "Consistent Nonparametric Regression", *The Annals of Statistics*, 5, 595 - 645.

Tjøstheim, D. and Auestad, B. (1994), "Nonparametric Identification of Nonlinear Time Series: Projections", *Journal of the American Statistical Association*, 89, 1398-1409.

Wand, M. P. and Jones, M. C. (1995), *Kernel Smoothing*, London, Chapman and Hall.

Wolters, J. (1992), "Persistence and Seasonality in Output and Employment of the Federal Republic of Germany", *Rechereches Economiques de Louvain*, 58, 421–439.

Wu, C. F. J. (1986), "Jackknife, Bootstrap and Other Resampling Methods in Regression Analysis (with discussion)", *Annals of Statistics*, 14, 1261-1350.

Yang, L. and Tschernig, R. (2002), "Non- and Semiparametric Identification of Seasonal Nonlinear Autoregression Models", *Econometric Theory*, 18, 1408-1448.

Yoshihara (1976), "Limiting Behavior of U-statistics for Stationary, Absolutely Regular Processes", *Zeitschrift für Wahrscheinlichkeitstheorie und verwandte Gebiete*, 35, 237-252.

34

Table 1: MISEs of the estimators $\hat{f}_1$, $\hat{f}_2$ and $\hat{f}_3$ for the i.i.d. case.

	f_1	f_2	f_3
n=50	0.0559	0.1144	0.1336
n=100	0.0300	0.0515	0.0617
n=250	0.0108	0.0223	0.0225

Table 2: Proportions among the 100 replications of rejecting the null hypotheses H_{s0}, $s = 1, 2, 3$, at the significant level 0.05 for the i.i.d. case.

	H_{10}	H_{20}	H_{30}
n=50	0.94	0.85	0.02
n=100	1	1	0.04

Table 3: Average squared residuals (ASR) and average squared prediction errors (ASPE) obtained from fitting twelve VCAR models with the German real GNP data. Each model is identified by the four digits which indicate the order in which the lagged variables Y_{t-1}, Y_{t-2}, Y_{t-2} and Y_{t-4} enter the VCAR model. For example, the model '1234' means $Y_t = f_1(Y_{t-1})Y_{t-2} + f_2(Y_{t-3})Y_{t-4} + $ (noise). The partially linear VCAR model at the bottom is $Y_t = f_1 Y_{t-2} + f_2(Y_{t-3})Y_{t-4} + $ (noise).

Model	ASR	ASPE
1234	**0.00021**	0.00011
1243	0.00040	0.00019
1324	0.00025	0.00013
1342	0.00039	0.00016
1423	0.00026	0.00014
1432	0.00024	**0.00009**
2134	0.00023	0.00017
2143	0.00051	0.00037
2341	0.00049	0.00032
2431	0.00024	0.00015
3142	0.00041	0.00023
3241	0.00038	0.00017
Linear AR	0.00059	0.00041
Partially Linear VCAR	0.00032	0.00024

36
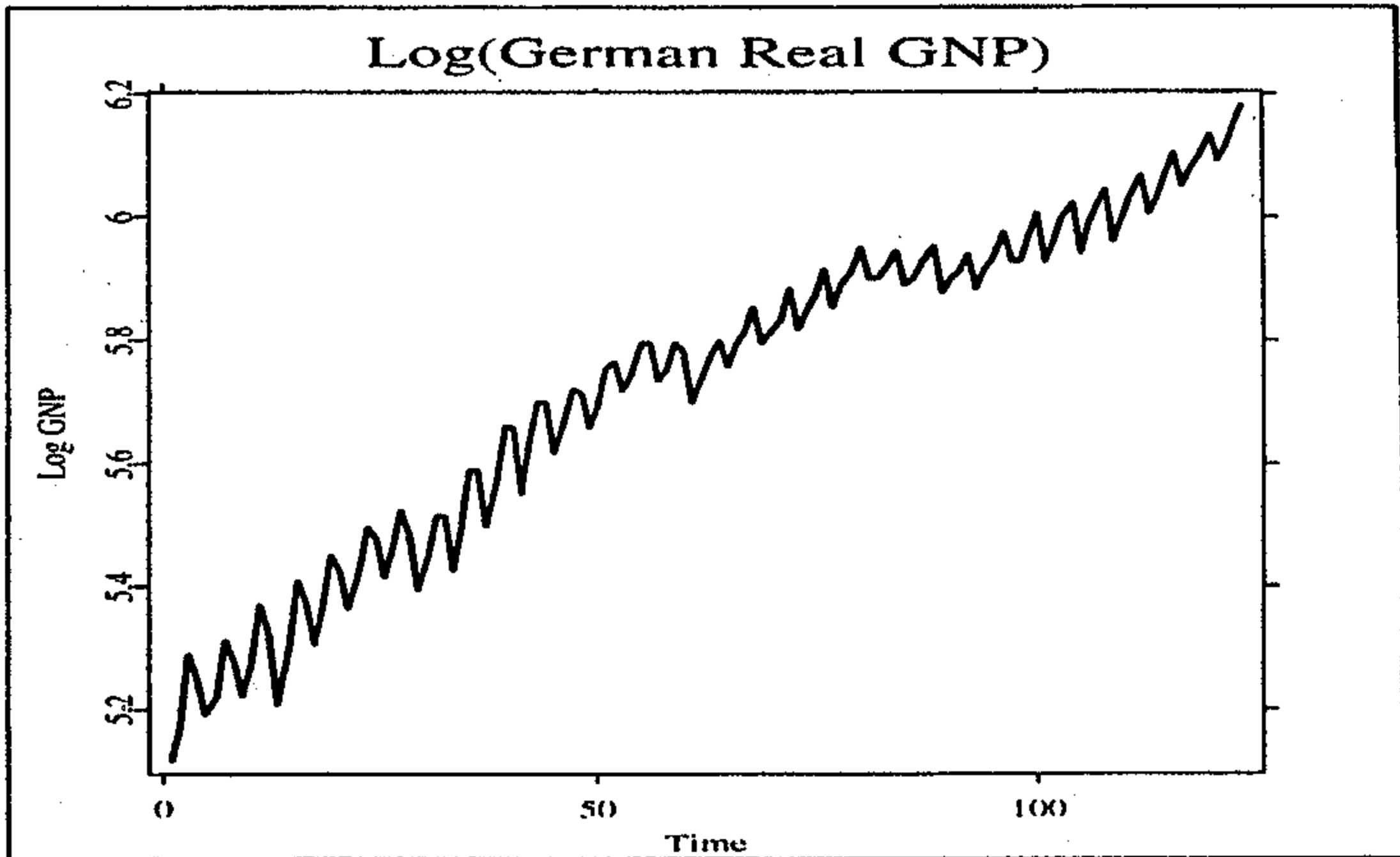
(a)
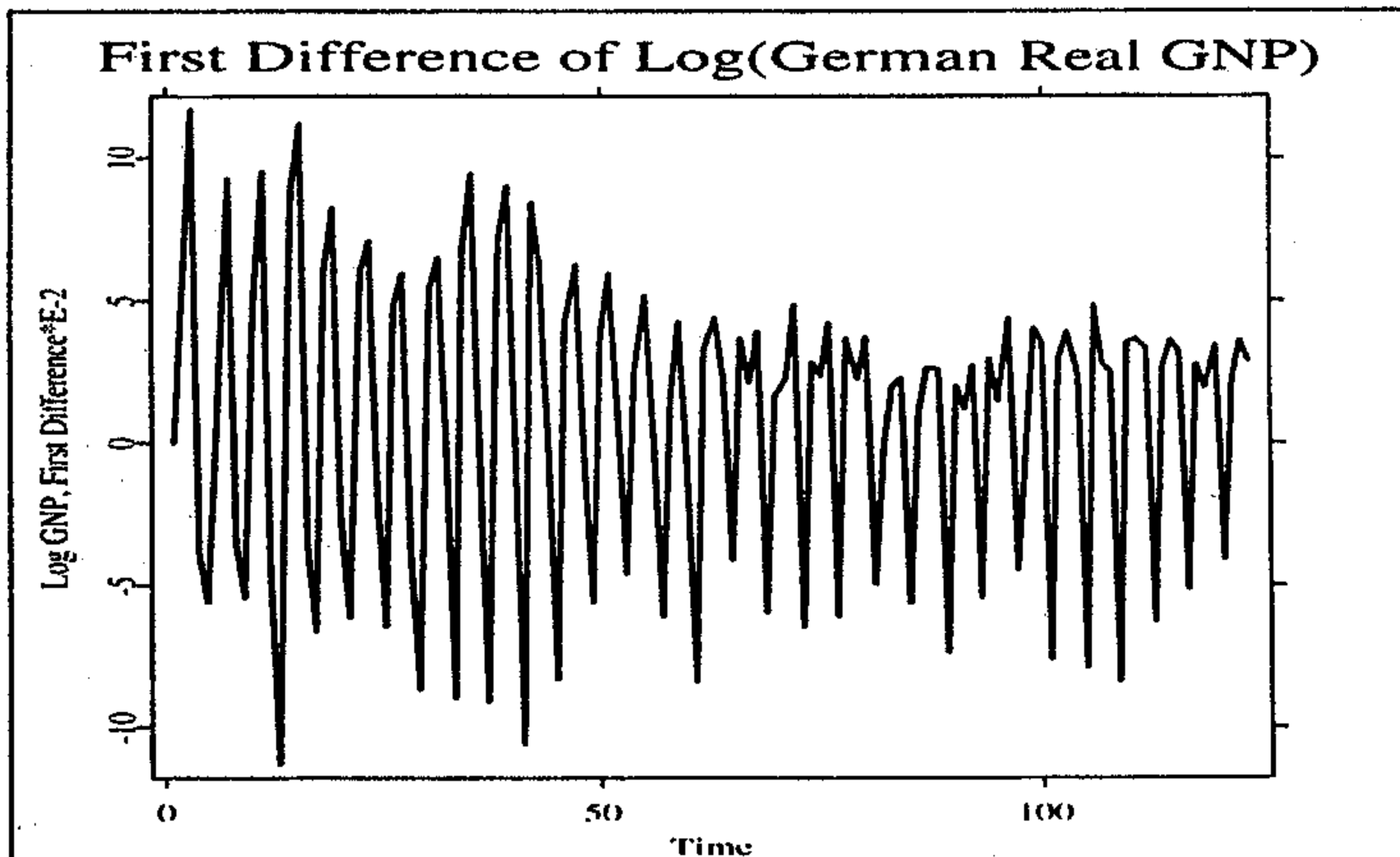
(b)

Figure 1: Plots of the West German real GNP quarterly data from 1960:1 to 1990:4. Panel (a) shows log(GNP) over time, and (b) depicts the first difference of log(GNP).

37
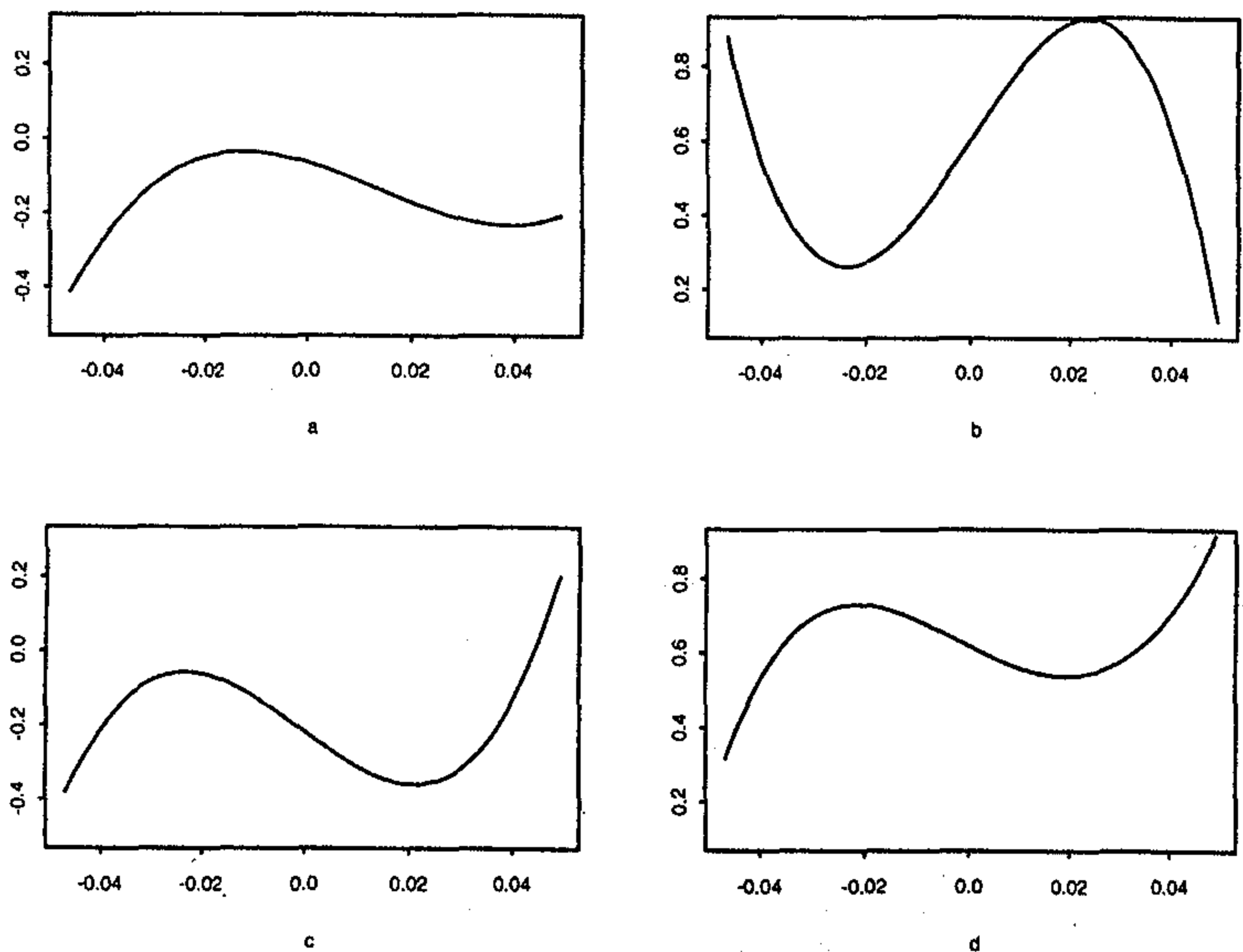
Figure 2: Estimated functions under the models (20) and (21). Panels (a) and (b) depict $\hat{f}_1$ and $\hat{f}_2$, respectively, for the model (20), while (c) and (d) are for the model (21).
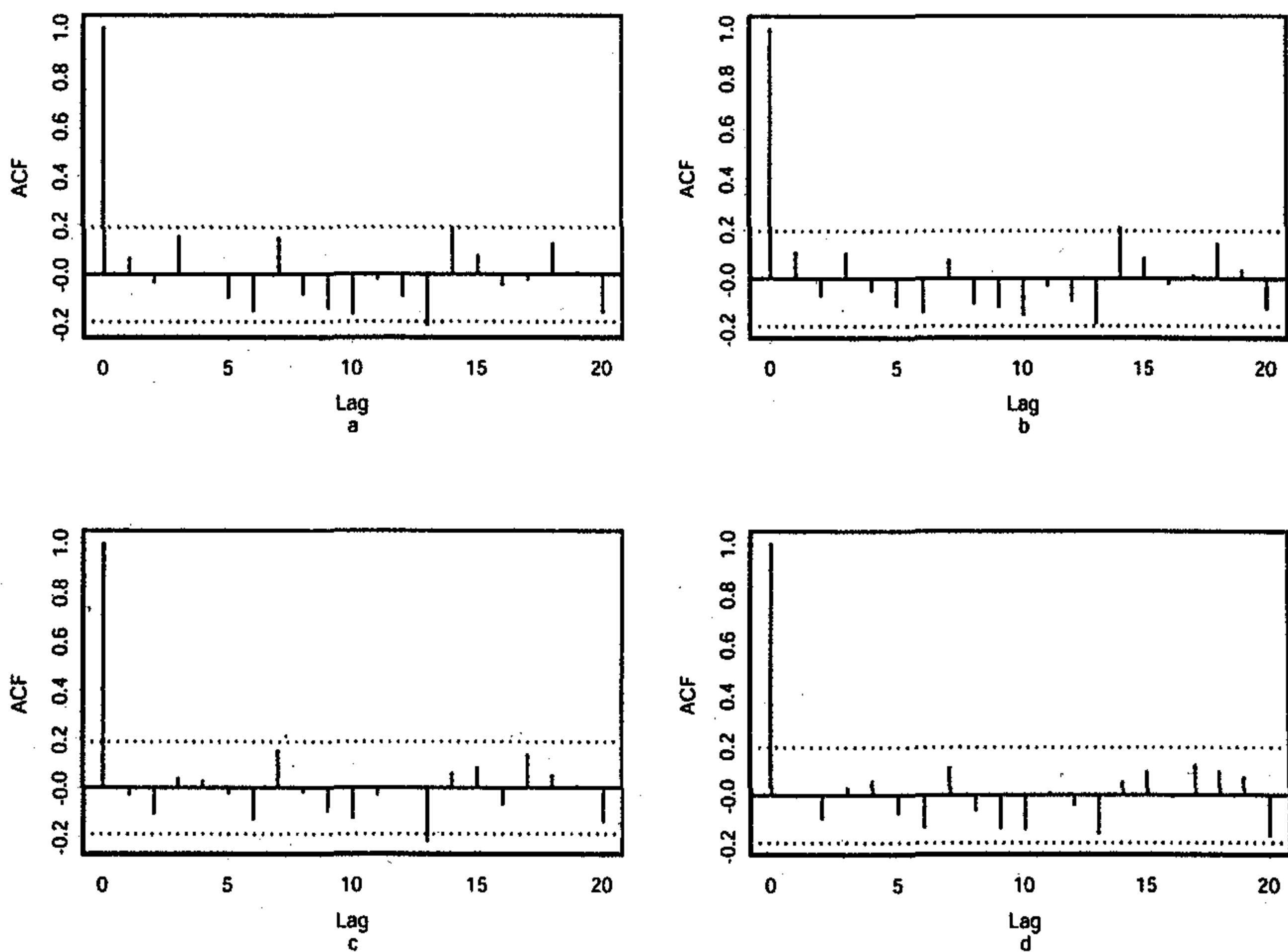
Figure 3: Autocorrelations of standardized residuals $\hat{\varepsilon}_t$. Panels (a) and (b) are for the model (20) and depict the autocorrelations of $|\hat{\varepsilon}_t|$ and $\hat{\varepsilon}_t^2$, respectively, while (c) and (d) are for the model (21). The dotted horizontal lines at levels $\pm 2 \times n^{-1/2}$ represent the 95.44% confidence bands of the autocorrelation functions.
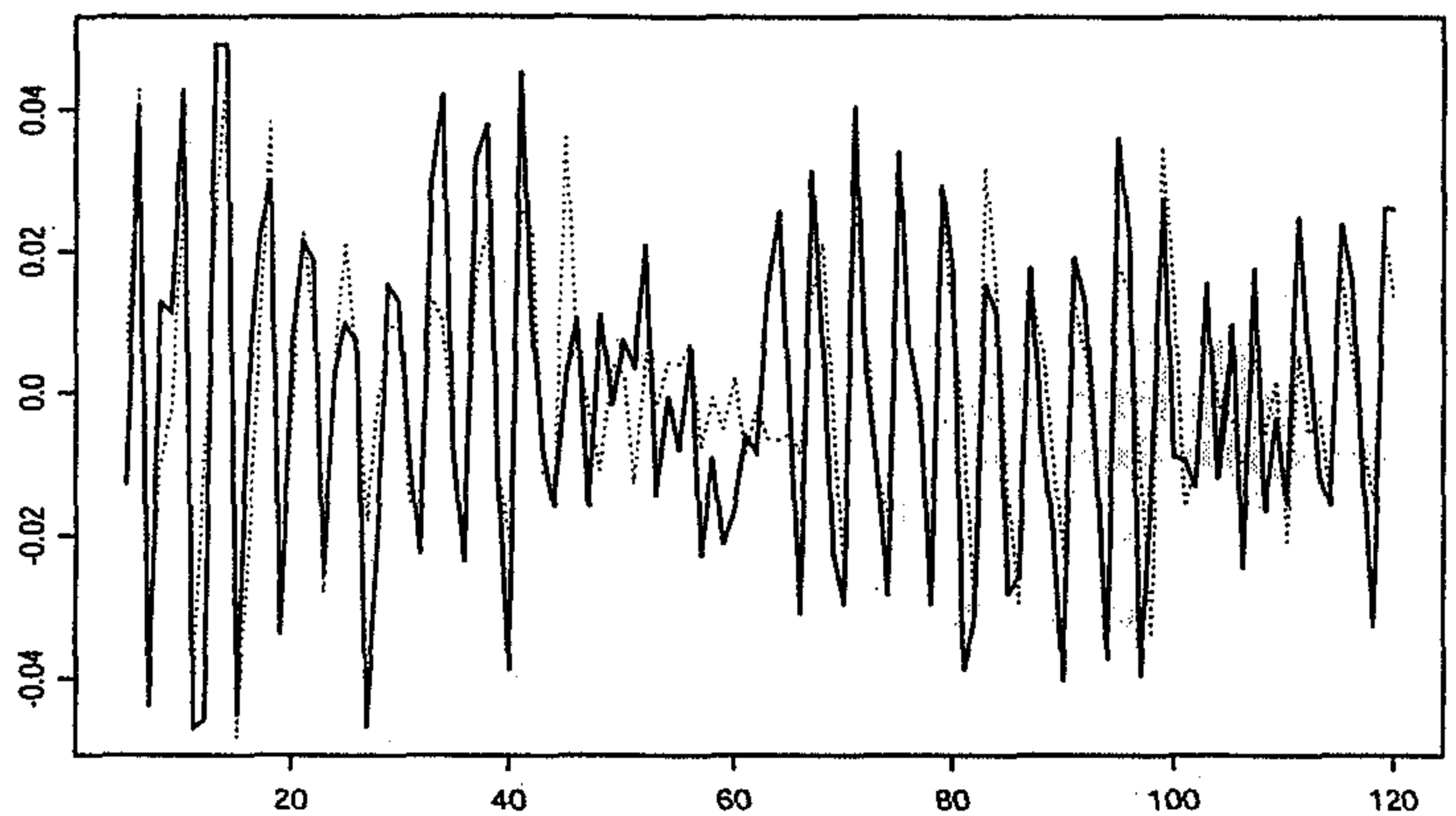
(a)
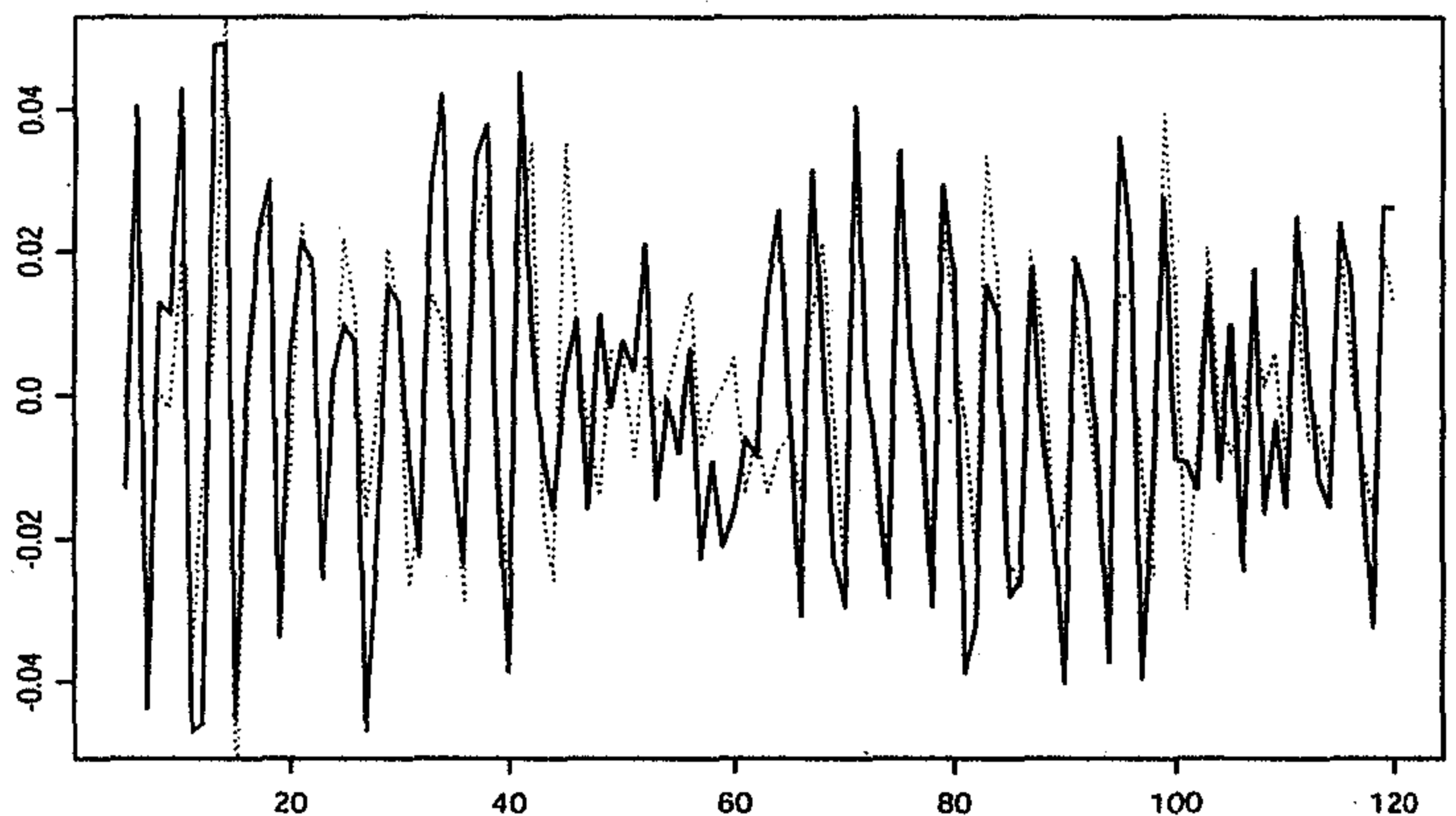
(b)

Figure 4: Prediction for the West German real GNP quarterly data based on the marginal integration fits of the varying coefficient models (20) and (21). Panel (a) is for the model (20), and (b) is for (21). Solid lines represent the predicted values $\hat{Y}_t$, while the dotted are for the observed values Y_t.

SFB 649 Discussion Paper Series

For a complete list of Discussion Papers published by the SFB 649, please visit http://sfb649.wiwi.hu-berlin.de.

Yang, L., Härdle W., Park, B., Xue, L., Estimation and Testing for Varying Coefficients in Additive Models with Marginal Integration

Using Wiki to Build an E-learning System in Statistics in Arabic Language

Taleb Ahmad*
Wolfgang Härdle*
Sigbert Klinke*

* Humboldt-Universität zu Berlin, Germany

BERLIN

ECONOMIC RISK

SFB 649

Using Wiki to build an e-learning system in statistics in Arabic language [*]

Taleb Ahmad , Wolfgang Härdle and Sigbert Klinke

CASE - Center for Applied Statistics and Economics

Humboldt-Universität zu Berlin

Wirtschaftswissenschaftliche Fakultät

Spandauerstrasse 1, 10178 Berlin, Germany

May 18, 2007

Abstract

E-learning plays an important role in education as it supports online education via computer networks and provides educational services by utilising information technologies. We present a case study describing the development of an Arabic language e-learning course in statistics. Discussed are issues concerning e-learning in Arab countries with special focus on problems of the application of e-learning in the Arab world and the difficulties concerning the design Arabic platforms such as language problems, cultural and technical problems, especially ArabTeX works difficulty with LaTeX format. Thus Wiki is offered as a solution to such problems. Wiki supports LaTeX and other statistical programs, for instance R, and Wiki offers the solution to language problems. Details of this technology are discussed and a solution as to how this system can serve in building an Arab platform is presented.

Keywords: E-learning, MM*Stat, Wiki, ArabTeX, Statistical software

JEL Classification I21, C19

1 Introduction

Due to the proliferation of the Internet, e-learning has become significant within education. Universities and educational institutions have created web sites and e-learning systems. In the future e-learning is expected to provide a complement to classical learning and the e-learning market is expected to become for software companies. In spite of this the e-learning market is still weak in Arab countries when compared to America and Europe. The reason why the software industry is weak in the Arab world has several reasons and will be discussed later. Our internet research shows that electronic platforms are few in Arab universities and an Arab e-learning platform in statistics could not be found. For this reason we find the creation a platform that would aid Arabic students in learning statistics highly necessary. From this perspective we are currently working on the development of an Arab e-learning platform in statistics (Arabic MM*Stat). The platform will contain the basic statistical topics, will be supported by multiple examples and the easy-of-use will be adapted for Arab students. It will be an important reference point in the study of statistics through the Internet. MM*Stat is an HTML based multimedia environment for supporting teaching and learning statistics via the internet or from a CD. The main aim is to discuss why LaTeX2HTML does not work with Arabic MM*Stat, therefore we use a new technology, namely the "Wiki technology". Wiki defines as software tool for collective operation through the internet and it uses easy rules to create web pages. Despite its young age e-learning has begun spreading worldwide and in particular in western countries. Statistics show that the size of the worldwide e-learning market is estimated to be 11 billion dollars yearly, with this ratio located between $60 - 70\%$ in the United States and Europe. Statistics indicate that 30% of the education was delivered electronically. There is a big gap between Europe and the United States and the Arab countries, where the size of e-learning market is around 15 million dollars yearly. Al-Reeise (2004). The reasons for this gap can be shortly summarized using the following facts:

- The diffusion of the internet services in the most Arabic countries is weak compared to other regions of the world. This is mainly due to the government monopolies over the telecommunications sector, resulting in higher prices. As a consequence only 0.5% of the internet users come from the Arab world, even though the Arab population is 5% of the world population.

Another example for this gap is the number of personal Computers. 20 per thousand in Arab countries compared with 200 per thousand in the developed countries. Al-Reeise (2004).

- As English is the most common language in the e-learning platforms and most Arab users have difficulties in understanding and speaking English, users do not have enough experience and ability to contact with others using English and that distances them from e-learning sources and educational courses.

- Social and cultural problems: Every language contains within the traditions and customs of the people speaking it. Restrictions are imposed by some Arab governments for internet users as it contains to opinions that violate the traditions and cultural values.

- The educational problems: A high level of illiteracy in the Arab world. $23, 21\%$. See: http://www.sajed.org/news.php?ID=48.

- A limited number of specialized cadres and the scientific expertise in the area of e-learning in the Arab countries.

Due to the mentioned problems the Arab countries need more time to acquire the advantage of e-learning. The dissemination of the culture of e-learning in the schools and universities needs a new generation of qualified people who can deal successfully with modern technology and the experiences of e-learning.

2 Difficulties to design Arabic platforms

There are many problems associated with the making of an Arabic platform, these relate to language, cultural and also technical. We summarize these issues :

Problem 1: The writing in Arabic is from right to left opposite to most other languages. Therefore all lists, paragraphs, statistical forms, tables and graphics run from right to left. In some cases however Arabic text may contain information that needs to run in the opposite direction (from left to right) like numbers and Latin texts. Any program that supports the Arabic language should provide the possibility of changing the direction when needed.

Problem 2: There are some items related with translation, some words and scientific terms are similar in Arabic and could create a problem when translated. For example, see table 1. Arabic makes no distinction between "administration" and "management". Another example, "calculate" and "calculator" are ambiguous as well as "compute" and "computer."

Arabic	English
ادارة	Administration
ادارة	Management
يحسب	Calculate
يحسب	Compute
حاسوب	Calculator
حاسوب	Computer

Table 1: Some similar words in Arabic

Problem 3: We must consider the issues relating specifically to Arabic culture. As traditions, cultural values play an important role in every society. "The cultural effects of using a different language. Each language has its own characteristics acting as a container in which morals and values are presented. It carries the culture of those people speaking it". Al-Khatib (2004).

Problem 4: Another problem is the LaTeX format. ArabTeX works difficulty with LaTeX. We aim to find a solution to this problem in our project. There was a project for Arabtex in LaTeX presented by Prof. Klaus Lagally - University Stuttgart (1993). This project was a positive step but it is not enough as the letters in Arab texts in this project were written in English and not Arabic and this project did not have the ability to produce statistics on an Arabic Web site. Figure 1 shows how to implemented ArabTeX into LaTeX in that project.

```
\documentstyle[12pt,arabtex]{article}
\begin{document}

\setarab    % choose the language conventions
\vocalize   % diacritics for short vowels on
\transtrue  % additionally switch on the transliteration
\arabtrue   % print arabic text ... is on anyway
\spreadtrue % spread out caption

\centerline {<^gu.hA wa-.himAruhu>}

\begin{arabtext}
'at_A .sadIquN 'il_A ^gu.hA ya.tlubu minhu .himArahu li-yarkabahu
fI safraTiN qa.sIraTiN wa-qAla lahu:

sawfa 'u`Iduhu 'ilayka fI al-masA'i, wa-'adfa`u laka 'u^graTaN. \\
fa-qAla ^gu.hA:

'anA 'AsifuN ^giddaN 'annI lA 'asta.tI`u 'an 'u.haqqiqa
laka ra.gbataka, fa-al.himAru laysa hunA al-yawma. \\
wa-qabla 'an yutimmu ^gu.hA kalAmahu
bada'a al-.himAru yanhaqu fI i.s.tablihi. \\
fa-qAla lahu .sadIquhu:

'innI 'asma`u .himAraka yA ^gu.hA yanhaqu. \\
fa-qAla lahu ^gu.hA:

.garIbuN 'amruka yA .sadIqI!
'a-tu.saddiqu al-.himAra wa-tuka_d_dibunI?
\end{arabtext}

\end{document}
```

جُحَا وَحِمَارُهُ *ǧuḥā wa-ḥimāruhu*

ʾatā ṣadīqun ʾilā ǧuḥā yaṭlubu minhu ḥimārahu li-yarkabahu fī safratin qaṣīratin wa-qāla lahu:

أَتَى صَدِيقٌ إِلَى جُحَا يَطْلُبُ مِنْهُ حِمَارَهُ لِيَرْكَبَهُ فِي سَفْرَةٍ قَصِيرَةٍ وَقَالَ لَهُ :

sawfa ʾuʿīduhu ʾilayka fī ʾl-masāʾi , wa-ʾadfaʿu laka ʾuǧratan.

سَوْفَ أُعِيدُهُ إِلَيْكَ فِي الْمَسَاءِ ، وَأَدْفَعُ لَكَ أُجْرَةً.

fa-qāla ǧuḥā:

فَقَالَ جُحَا :

ʾanā ʾāsifun ǧiddan ʾannī lā ʾastaṭīʿu ʾan ʾuḥaqqiqa laka raǧbataka, fa-ʾlḥimāru laysa hunā ʾl-yawma.

أَنَا آسِفٌ جِدًّا أَنِّي لَا أَسْتَطِيعُ أَنْ أُحَقِّقَ لَكَ رَغْبَتَكَ ، فَالْحِمَارُ لَيْسَ هُنَا الْيَوْمَ .

wa-qabla ʾan yutimmu ǧuḥā kalāmahu badaʾa ʾl-ḥimāru yanhaqu fī ʾsṭablihi.

وَقَبْلَ أَنْ يُتِمُّ جُحَا كَلَامَهُ بَدَأَ الْحِمَارُ يَنْهَقُ فِي اصْطَبْلِهِ .

fa-qāla lahu ṣadīquhu:

فَقَالَ لَهُ صَدِيقُهُ :

ʾinnī ʾasmaʿu ḥimāraka yā ǧuḥā yanhaqu.

إِنِّي أَسْمَعُ حِمَارَكَ يَا جُحَا يَنْهَقُ .

fa-qāla lahu ǧuḥā:

فَقَالَ لَهُ جُحَا :

ǧarībun ʾamruka yā ṣadīqī! ʾa-tuṣaddiqu ʾl-ḥimāra wa-tukaḏḏibunī?

غَرِيبٌ أَمْرُكَ يَا صَدِيقِي ! أَتُصَدِّقُ الْحِمَارَ وَتُكَذِّبُنِي ؟

Figure 1: *Left panel: Sample Arabtex input in LaTeX, right panel: Sample Arabtex output in LaTeX*

.

Problem 5: How to include the interactive examples in Arabic MM*Stat?

3   WiKi Technology

We propose the Wiki technology as a solution to these problems. Wiki supports the solutions for language problems, and on top of this it is an easy tool to create web page. Wiki software supports LaTeX format to writing the complex statistical forms, and we will try to integrate of LaTeX into Arabic MM*Stat. We will use R program to do the interactive examples and multiple choice question in Arabic MM*Stat, thus R will be integrated into Wiki.

3.1   What is Wiki ?

Wiki is a system that allows users to collaborate in forming the content of a web site, Jennings (2006). The first web site in wiki, "WikiWikiWeb" was designed by Ward Cunnigham and Bo Leuf in 1995. They describe the Wiki system as a simple database that can operate on the World Wide Web. The goal is to simplify the process of participation and cooperation in the development of web content with maximum flexibility. The main advantages of Wiki are that:

- Wiki simplifies the process of content editing. Each page contains a link for changing contents through an editing screen which can then be saved. The saved pages can be viewed on the web site.

- It uses simple markup to coordinate contents, and it is suitable for users with little experience with computers or web site development as no HTML language knowledge is required.

- Wiki sites keep record of the history pages and therefore makes the comparison of older and newer web pages an easy task. If a mistake is made, one can revert to the older version of the page.

- Most Wiki sites are publicly open and therefore content editing by users is open.

- Wiki simplifies the organization of a site: Wiki sites create hypertext databases and can regulate the content in any manner desired; many of the content management systems

require the planning of a classifications for of content before anything is written; in Wiki one can organize the contents by the sections or without the sections. Visitors to the web site can use the links between the pages. This flexibility is not available in the content management systems.

3.2   Application of Wiki

The flexibility of the Wiki concept makes it an ideal knowledge transfer tool, at universities, educational institutes, in companies and with specialized web sites." Wiki serves easily as an effective tool for communication between students and professors in the universities and for publishing articles, documents as well as lectures." (See: http://www.serdal.com/articles/6/). In many universities students are not required to visit lectures since lectures are written and published on the university web site. Each teacher can write his course using Wiki and offer it to his students as an useful material for the study. There are many examples of web sites dependent on Wiki as tool for the development of the contents, like Wikipedia, see the link: http://en.wikipedia.org. The Wikipedia project started 15 January 2001 and today there are more than 5.5 million articles in the encyclopedia in all languages, more than 1.5 million articles in the English encyclopedia alone. Millions of volunteers around the world modify and add to the contents daily and a new articles are created. The Arabic version of the free encyclopedia was launched in July 2003 and currently contains only 30,210 articles only as the Arab encyclopedia is in the content-building phase. There is a great potential as a tool for online collaboration in Humboldt university from Dr. Sigbert Klinke, this project was applied in the web site: http://teachwiki.wiwi.hu-berlin.de. And there is Arabic e-learning platform in Wiki. http://wiki.arabeyes.org. Arabeyes provides a good environment for discussion and exchange of experience and knowledge about the Arabic language and support it with the computer and free software, Arabeyes offers the translation for a free open-source programs. In addition there is in Arabeyes the technical dictionary that aims to translate and standardize the technical terms used in translating the software to the Arab user. Arabeyes is a solution for problem 2, see later figure 2. Wiki supports the solution for problem 1, see figures 4, 6 and 9. It also supports LaTeX to writing the complex statistical forms into Arabic MM*Stat, a solution for problem 4, see figure 4.
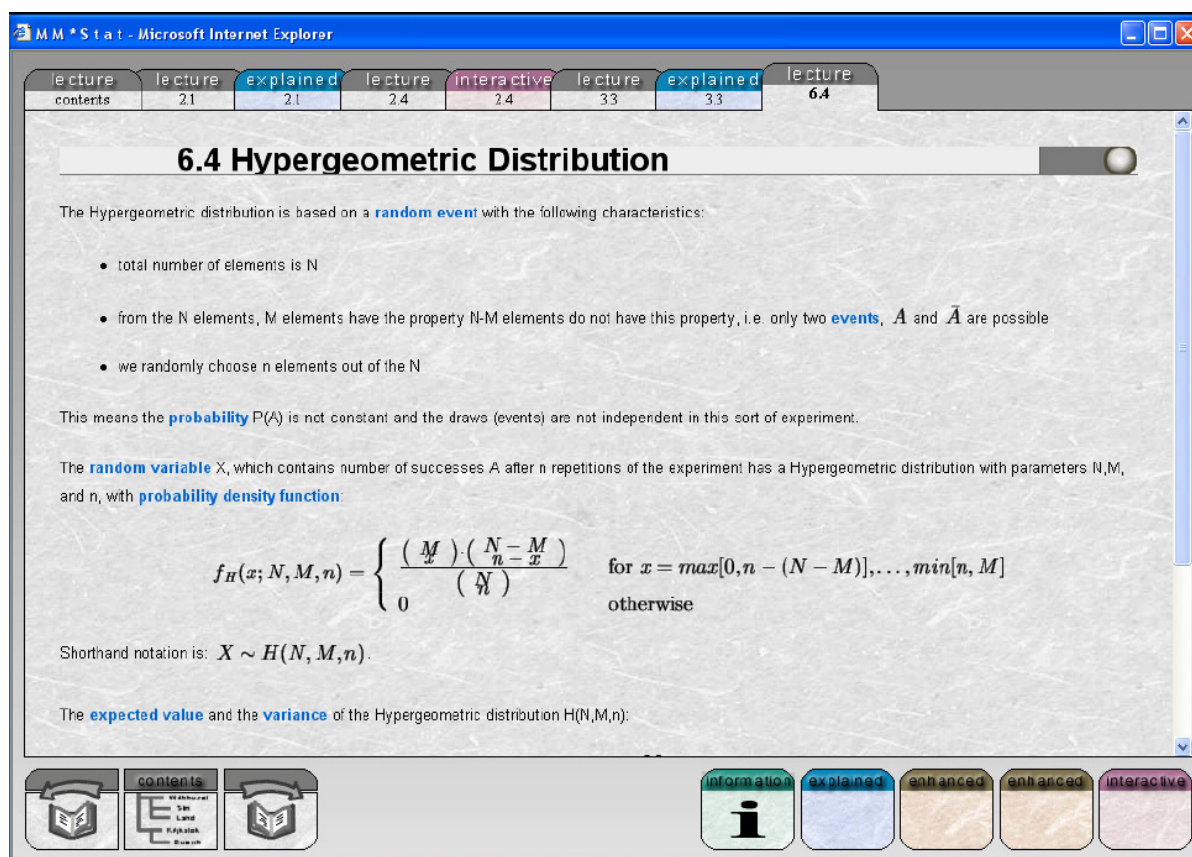
Figure 2: Graphical user interface (GUI) of Arabeyes Todo

4 An Arabic e-learning system

4.1 Description of Arabic MM*Stat

The basic frame for this platform, the system MM*Stat was developed at the School for Business and Economics of Humboldt - Universität zu Berlin. MM*Stat is a platform for e-learning statistics and it is an HTML based multimedia environment to support teaching and learning statistics via internet or CD, see: http://www.quantlet.com/mdstat/products.html. It provides e-learning of Statistics in many languages including English, German, French, Polish, Czech, Spanish and Italian.

Figure 3: Standardization is via an HTML filing card system

Arabic MM*Stat is directed to students and Arab users that serve the e-learning issues in the Arab region. As we saw earlier there is a lack of interest in the e-learning issues within
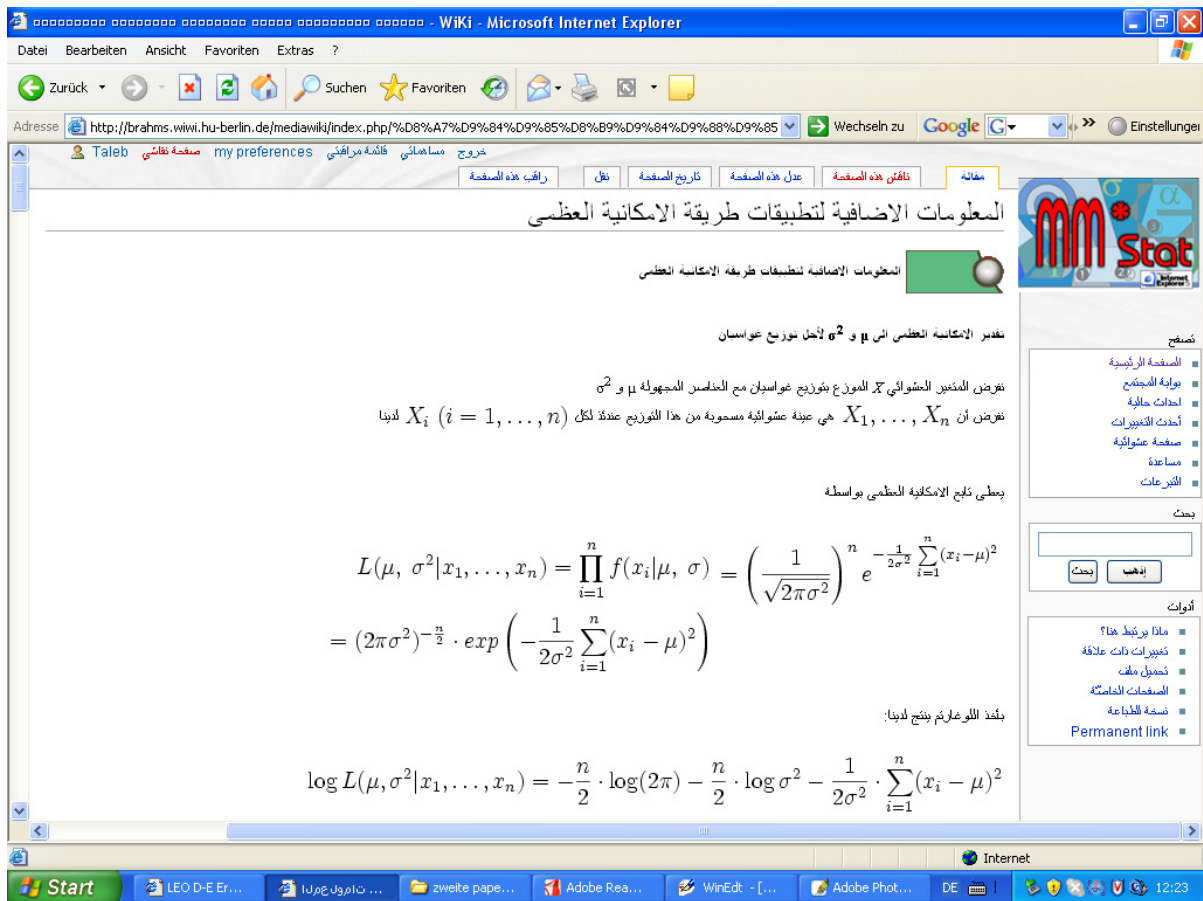
Figure 4: Graphical user interface (GUI) of Arabic MM*Stat in WiKi

universities and Arab institutes and there are not enough e-learning platforms. This platform will be the first developed platform for Arab users and it is applied in the following web page see: (http://pluto.wiwi.hu-berlin.de/mediawiki/index.php).

MM*Stat course has five components structures: The lecture, information, explained, enhanced and interactive. Each lecture gives the basic concepts of the general statistical theory, definitions, formulae as well as mathematical proofs. MM*Stat provides the students to compute distribution functions, graphics and derive results for statistical tests. The students or anyone interested in statistics can interactively learn about basic concepts of statistics at anytime and anywhere.

4.2 How to introduce new contents into Arabic MM*Stat?

These principles and rules of Wiki are very important for designing Arabic MM*Stat that operates using Wiki. These principles reduce the effort and cost associated with the creation of simple web pages that works with the HTML language.

- Adding of new pages: The establishment of the pages for links existed already and sometimes refers to the pages of the articles not yet established. Before pages are established it must be confirmed that no page has been previously been created and that it is separate from the other pages.

- Editing pages: Editing and modifying pages is an easy process. Click the link "edit this page" located at the top of the page, moved this link to another page "edit page", where the contents of the page are for the adding and preview and when we finish the writing. The page must be saved and we can see all the changes made by clicking the button "show preview".

- Adding pictures: All the loading pictures are from the file MM*Stat.

- Addition the internal and external links to Arabic MM*Stat: When we write a word or sentence between brackets, this sentence will work as an internal link in Arabic MM*Stat and pages already existing will appear in blue. If the page does not exist, the link color will appear in red. This relates only to internal links in Wiki. However external links can also be created with ease.

- Addition variables and statistical formulas: Writing in Arabic MM*Stat dependent basically on the mathematical and statistical formulas. All statistical formulas can be rewritten from Arabic Wikipedia. See: http://ar.wikipedia.org/wiki/.

- The creation of tables in Arabic MM*Stat: There are two ways to create tables. The first is using HTML. Formulas can be applied to the HTML environment, and the following example in figure 5 explains that. Since some users are unfamiliar with HTML or because some symbols of tables are difficult, we apply Wiki method as the figure 6.
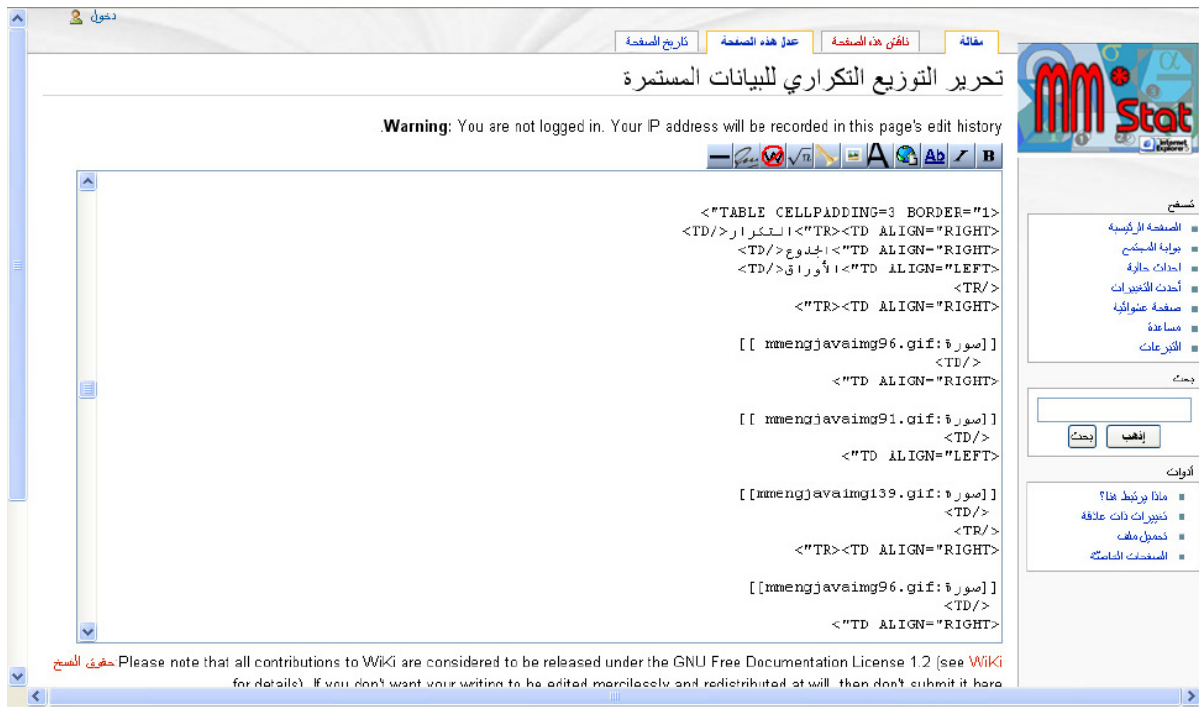
Figure 5: Edit table in Arabic MM*Stat using HTML
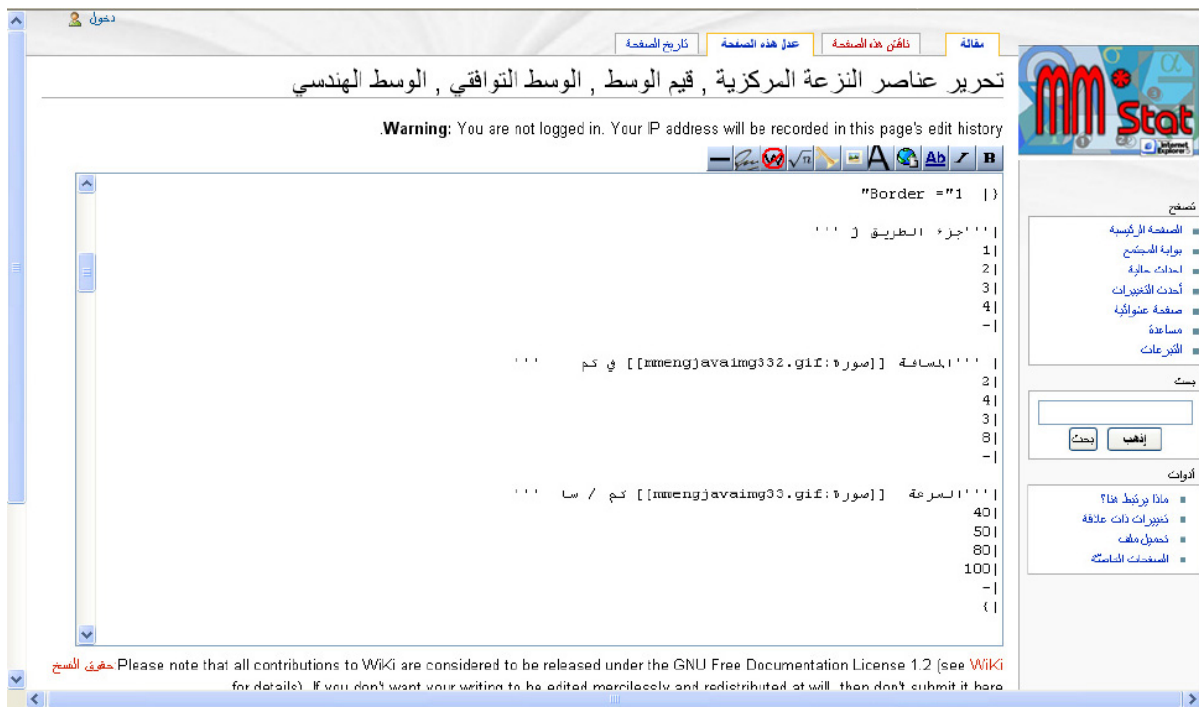
Figure 6: Edit table in Arabic MM*Stat using Wiki

11

We must consider another issues, when we translate MM*Stat from German or English to Arabic, we must consider the issues relating specifically to Arabic culture. For example: Interest rates or the prohibition of alcohol. There were in MM*Stat many examples relating to this. We must delete them from Arabic MM*Stat like figure 7 and 8 . The user can delete these examples directly during the working on Wiki. Here comes the importance of Wiki as a edit tool. It also a solution for problem 3.

Explained: A Deck of Cards

Assume you have shuffled a standard deck of 52 playing cards. You are interested in the **probability** of a randomly drawn card being a queen or a 'heart'. We are thus interested in probability of the event $(\{\text{Queen}\} \cup \{\text{Heart}\})$.

Following Laplace's notion of probability, we proceed as follows: There are 4 queens and 13 hearts in the deck. Hence,

- $P(\{\text{Queen}\}) = \frac{4}{52}$

- $P(\{\text{Heart}\}) = \frac{13}{52}$

But there is also one card which is both a queen *and* a heart. As this card is included in both counts, we would overstate the probability of either queen or heart appearing if we simply added both probabilities. In fact, the addition rule of probability requires one to deduct the probability of this joint event:

$$P(A \cup B) = P(A) + P(B) - P(A \cap B)$$

Figure 7: An example for deck of cards

Combinations

Millions of Germans try every Saturday their luck in the lottery called Lotto. They choose 6 numbers from 49 and hope that, thanks to these 6 numbers, they will get rich. They base the choice often on various almost ``mystical" numbers--numbers such as the date of somebody's birthday, the birthday of their dog, numbers hinted by a horoscope and so on. How many possibilities for a choice of 6 numbers out of 49 actually exists?

From 49 numbers (elements), exactly 6 is chosen. The order in which the numbers are chosen in unimportant--it does not matter whether one crosses first 4 and then 23 or vice versa. That means that ordering of elements is not taken into consideration. Therefore, permutations (simple reorderings of n elements) and variations as well (ordering of elements matters) are not the right choice. The right concept is a **combination**.

Nevertheless, there are still two possibilities--combinations with or without repetition. Since every number of the lottery ticket can be crossed just once, repetition of numbers (elements) is not possible and we use **combinations without repetition**.

$$n = 49 \qquad k = 6$$

$$K(n,k) = \binom{n}{k} = \frac{V(n,k)}{P(k)} = \frac{n!}{k! \cdot (n-k)!}$$

$$K(n,k) = \frac{49!}{6! \cdot (49-6)!} = 13\,983\,816$$

Figure 8: An example for lotto

12

5 Integrate R program into Wiki

R is a language and environment for statistical computing and graphics. It was origi-
nally created by Ross Ihaka and Robert Gentleman at the University of Auckland, New
Zealand, and is now developed by the R Development Core Team. R provides a wide
variety of statistical (linear and nonlinear modelling, classical statistical tests, time-
series analysis, classification, clustering, ...) and graphical techniques, and is highly
extensible (see: http://www.r-project.org/about.html). Arabic MM*Stat provides R
program which can be incorporated with the courses notes. Based on R program, the
student can interactively learn about basic concepts of statistics and that lead the stu-
dents to communicate with the lectures. The Mediawiki is the software behind the
Wikipedia. This extension allows to run R programs within the Mediawiki and to dis-
play graphics, HTML output and the interactive examples in the Arabic MM*Stat. As
a solution for problem 5. These interactive examples enable the students and learner to
apply the statistics distributions and the probability tables via the internet. (For exam-
ples, see figure 9). This example explains how to apply the hypergeometric distribution
for various values. Other example (see figure 10) shows the graphic for the binomial
distribution and you can get other graphic when you input other values. There are
in Arabic MM*Stat other examples for other distributions like normal, poisson and
exponential distribution.
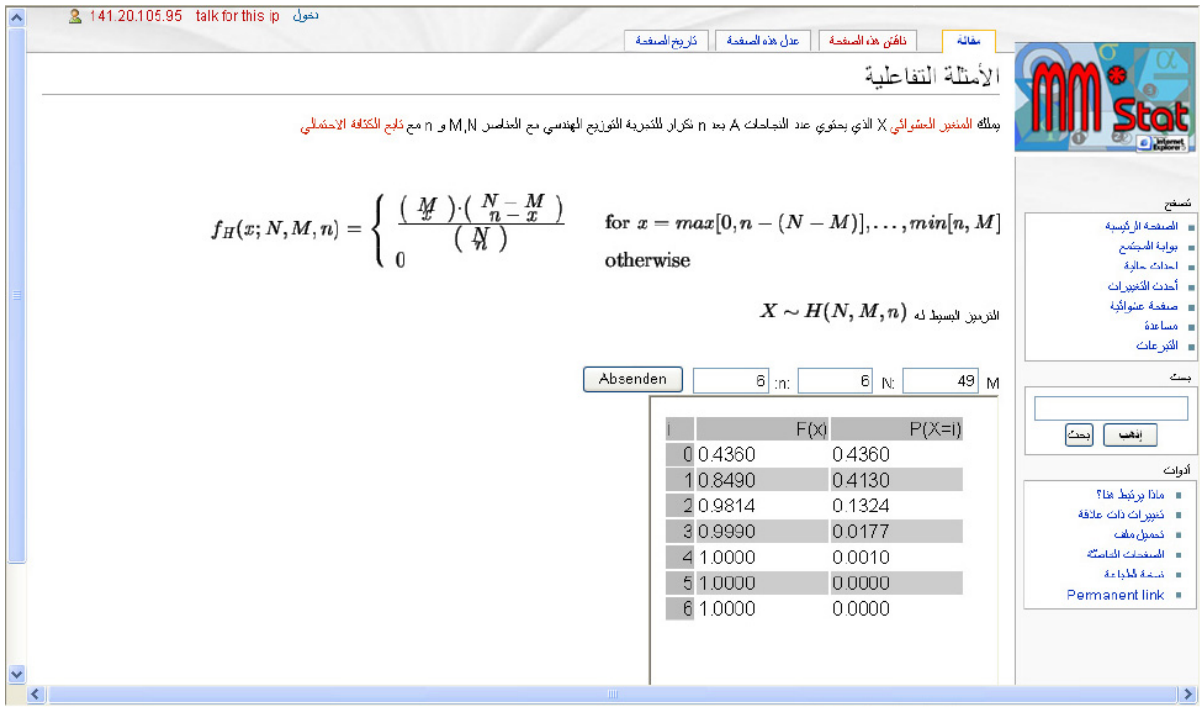
Figure 9: The interactive example for hypergeometric distribution.
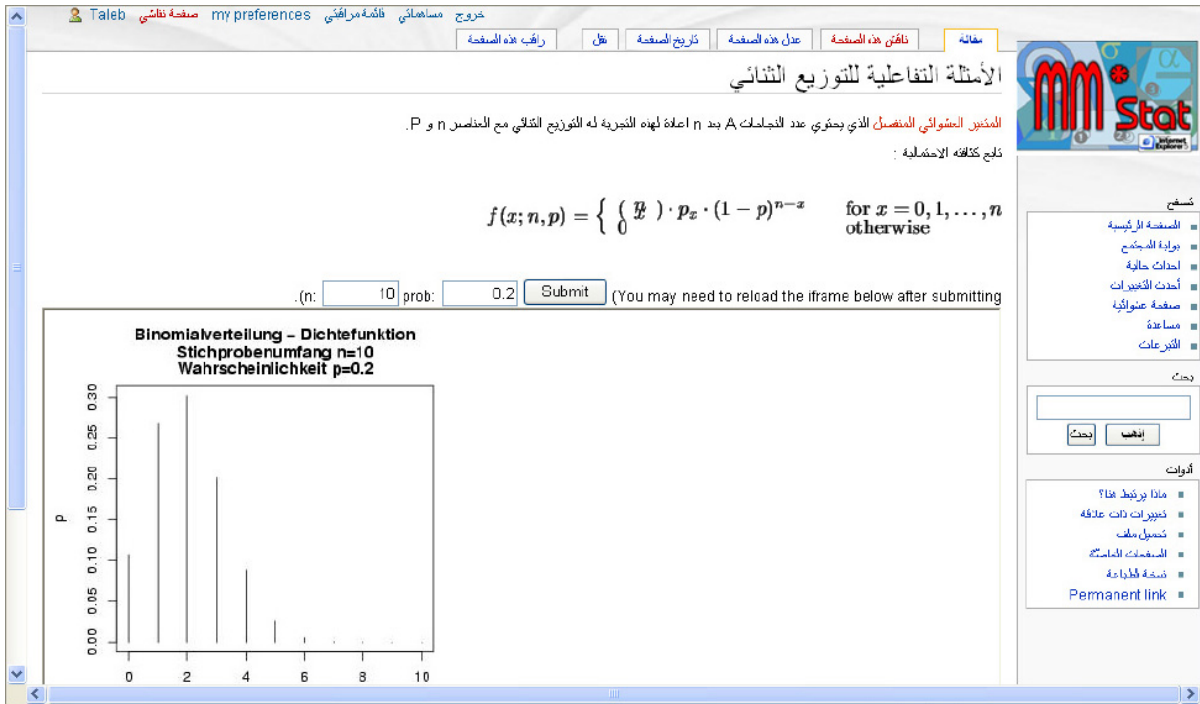
Figure 10: The interactive example for binomial distribution.

6 Conclusion

Using E-learning/e-teaching tools to offer an effective learning of statistics is a necessity for students. There is the possibility of creating an e-learning system with Arabic MM*Stat through the application of Wiki technology. Since some of the specific characteristics we have discussed earlier for developing an Arabic platform already exist in the Wiki. We see that Rwiki is the solution for the interactive examples in Arabic MM*Stat. We hope that Arabic MM*Stat platform for e-learning of statistics will be a significant contributions to the Arab user as it clearly overcomes weaknesses in developing such electronic platforms in Arabic.

References

Ahmad, T., Härdle, W., Mungo, J., *On the Difficulty to Design Arabic E-learning System in Statistics*, COMPSTAT 2006 Proceedings, Roma, Italy (2006).

Al-Khatib, W., *e-learning: Opportunities, Challenges and Future*. Information and Computer Science Department. King Fahd University of Petroleum and Minerals, Dhahran, Saudi Arabia (2004) .

Al-Reeise, A., *E-learning in Arab world*, Oman 2004,
http://www.elearning.edu.sa/forum/showthread.php?t=264

Dutta, S., El-Hage, C., Sabbagh, K., Tarazi, P., *Challenges for Information and Communication Technology Development in the Arab World* 2003.

Härdle, W., Klinke, S., Ziegenhagen, U., *eLearning-A Critical Review*, COMPSTAT 2006 Proceedings, Roma, Italy (2006).

Hoffmann, M., *Arabic and Bidirectional Challenges for Translation and Software Development*, Manager of Applications Engineering, ENLASO Corporation (2006).

Klinke, S., Schmerbach, S., Troitschanskaia, O., *Use of wikis in teaching*. Statistics under one Umbrella, Bielefeld University, Germany (2007).

Lagally, K., *ArabTeX : a System for Typesetting Arabic*. User Manual Version 3.00, University Stuttgart (1993).

Mungo, J., *e-learning/e-teaching An Implementation and Evaluation of a Finance Introductory Course*, Humboldt-Universität zu Berlin (2004).
http://edoc.hu-berlin.de/docviews/abstract.php?id=26937

Müller, M., Rönz, B., Ziegenhagen, U., *The Multimedia Project MM*Stat for Teaching Statistics*, In COMPSTAT. Proceedings in Computational Statistics, Physica Verlag (2000).

Maegaard, B., Choukri, K., Mokbel, C., Yaseen, M.,*Language Technology for Arabic*, NEMLAR, Center for Sprogteknolo gi. University of Copenhagen, July 2005.

http://www.linux.org.sy/node/51

http://www.serdal.com/articles/6/

http://www.sajed.org/news.php?ID=48

http://searchwebservices.techtarget.com/sDefinition/0,,sid26$_g$ci943070, 00.$html$

$http://mars.wiwi.hu-berlin.de/mediawiki/slides/index.php/Main_page$

$http://www.r-project.org/about.html$

$http://en.wikipedia.org/wiki/R_programming_language$

Klinke, S., Schmerbach, S., Troitschanskaia, O., *Use of wikis in teaching*. Statistics under one Umbrella, Bielefeld University, Germany (2007).

Lagally, K., *ArabTeX : a System for Typesetting Arabic*. User Manual Version 3.00, University Stuttgart (1993).

Mungo, J., *e-learning/e-teaching An Implementation and Evaluation of a Finance Introductory Course*, Humboldt-Universität zu Berlin (2004).
http://edoc.hu-berlin.de/docviews/abstract.php?id=26937

Müller, M., Rönz, B., Ziegenhagen, U., *The Multimedia Project MM*Stat for Teaching Statistics*, In COMPSTAT. Proceedings in Computational Statistics, Physica Verlag (2000).

Maegaard, B., Choukri, K., Mokbel, C., Yaseen, M.,*Language Technology for Arabic*, NEMLAR, Center for Sprogteknolo gi. University of Copenhagen, July 2005.

http://www.linux.org.sy/node/51

http://www.serdal.com/articles/6/

http://www.sajed.org/news.php?ID=48

http://searchwebservices.techtarget.com/sDefinition/0,,sid26$_g$ci943070, 00.$html$

$http://mars.wiwi.hu-berlin.de/mediawiki/slides/index.php/Main_page$

$http://www.r-project.org/about.html$

$http://en.wikipedia.org/wiki/R_programming_language$

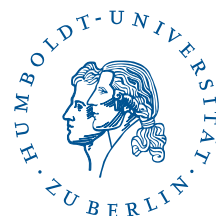# SFB 649 Discussion Paper Series 2007

For a complete list of Discussion Papers published by the SFB 649, please visit http://sfb649.wiwi.hu-berlin.de.

Taleb Ahmad
Wolfgang Härdle
Uwe Ziegenhagen

E-Learning interkulturell

_Arab_MM*Stat

Entwicklung einer Lernplattform für arabische Nutzer

Abb. 2
Die arabische MM*Stat-Version im MediaWiki

In vielen Bereichen der universitären Ausbildung haben Elemente des E-Learning ihren Einzug gehalten, sei es in Form von Video- oder Audioinhalten, Lern-und Lerninhaltsumgebungen (Moodle, Mneme) oder als web- und computerbasierte Trainingsanwendung. – Wir stellen eine Fallstudie vor, die die Entwicklung der ersten arabischen Trainingsanwendung für angewandte Statistik beschreibt. Neben den technischen Schwierigkeiten gehen wir ebenso auf die kulturellen und sprachlichen Schwierigkeiten ein, die bei der Entwicklung dieser Plattform eine Rolle spielten.

MM*STAT auf Arabisch

Die inhaltliche Grundlage dieses Projekts wurde vor einigen Jahren am Fachgebiet für Statistik und Okonometrie der Wirtschaftswissenschaftlichen Fakultät unter Leitung von Prof. W. Härdle und Prof. B. Rönz geschaffen. Die Software mit dem Titel MM*Stat wurde entwickelt, um Studenten des Grundstudiums einen leichteren Zugang zur Materie zu verschaffen und ihnen die Möglichkeit zu bieten, Vorlesungsinhalte am heimischen Rechner zu wiederholen.

Das System läuft komplett im Web-Browser und besteht aus HTML-Seiten, die in Form von Karteikarten angeordnet sind (siehe Abb. 1). Von den Vorlesungsseiten gelangt man über Hyperlinks zu den Beispielseiten, die detailliertere Erklärungen, Multiple-Choice-Aufgaben oder interaktive Beispiele auf Basis von Java-Applets bieten.

MM*Stat wurde in diverse Sprachen übersetzt, darunter englisch, französisch, italienisch, polnisch, spanisch und tschechisch. Der technische Aufwand der Übersetzung in Sprachen mit lateinischen Buchstaben blieb dadurch in überschaubaren Grenzen, dass die

Abb. 1
Beispiel einer MM*Stat-Webseite

Quelltexte, die in LaTeX erstellt worden waren, über die selbst entwickelte Software MD*Book (siehe Humboldt-Spektrum 01/2002) in das HTML-Format konvertiert wurden. _Arab_MM*Stat wird die erste E-Learning-Umgebung in Statistik für arabische Nutzer sein. Eine Beispielseite der arabischen MM*Stat-Version im MediaWiki zeigt Abb. 2 (http://pluto.wiwi.hu-berlin.de/mediawiki/index.php).

Bei der Übersetzung ins Arabische stießen wir jedoch auf Schwierigkeiten, die technische, spachliche und kulturelle Ursachen haben:

- Im Arabischen wird im Gegensatz zu den meisten anderen Sprachen von rechts nach links geschrieben, dies gilt auch für Formeln, Tabellen und Grafiken. Die bisher erstellten Inhalte mussten allesamt konvertiert werden.

- Der bei den anderen Sprachen genutzte Workflow (LaTeX → HTML → MD*Book → MM*Stat-Format) konnte nicht genutzt werden, da die zugrunde-liegenden Konverter in das HTML-Format allesamt nur auf lateinische Sprachen ausgelegt sind und mit der gängigen Eingabeform für arabische Inhalte (ArabTeX) nicht umgehen konnten.

- Die sozio-kulturellen Probleme betreffen in erster Linie die Unterschiede zwischen arabischen und westlichen Gesellschaften, wie die Verbote von Glücksspiel, Zinsen und Alkohol. In den Inhalten von MM*Stat fanden sich verschiedene Beispiele wie die Veranschaulichung von Wahrscheinlichkeitsberechnungen durch Spielkarten, die für arabische / muslimische Lerninhalte ungeeignet sind.

Die technischen Probleme ließen sich durch einen Wechsel der Plattform lösen. Anstelle der LaTeX-basierten Lösung nutzen wir für die arabische Version von MM*Stat ein Wiki, das neben der Sprachunterstützung für Arabisch noch eine ganze Reihe interessanter Funktionen bietet.

WIKI, WIKI

Ein Wiki (hawaiisch für »schnell, schnell«) ist eine im Intra- oder Internet verfügbare Sammlung von HTML-Seiten, die von Benutzern nicht nur gelesen, sondern auch online über entsprechende Eingabefenster verändert werden kann. Die Ur-Form aller Wikis wurde 1994 von Ward Cunningham entwickelt, das bekannteste Wiki-System ist MediaWiki (http://www.mediawiki.org), das auch die Grundlage für die Wikipedia bildet. Für die Formatierung von Inhalten nutzen Wikis eine einfache Syntax, so dienen beispielsweise =, == und === der Abgrenzung verschiedener Gliederungsebenen. Zwei Anführungszeichen dienen der Auszeichnung von kursivem Text, drei Anführungszeichen der Auszeichnung von fettem Text. Die Vorteile eines Wiki sind:

- Die Seiten lassen sich online ändern und stehen allen Nutzern sofort wieder zur Verfügung.
- Es sind keine HTML-Kenntnisse nötig, um Seiten ändern zu können.
- Das Wiki-System speichert die Historie des Dokuments. Falls notwendig, kann so eine ältere Version der Seite wiederhergestellt werden.
- Komfortablere Wiki-Systeme, wie z.B. MediaWiki, können beispielsweise LaTeX-Formeln automatisch in Bilder umwandeln und bieten sogar die Möglichkeit, eigene Plugins einzusetzen (Abb. 3).

Interaktive Beispiele mit *Rwiki*

Ein Beispiel für ein solches Plugin ist das von Dr. S. Klinke am Fachgebiet Statistik entwickelte *Rwiki*. Mit Hilfe dieses Plugins ist es möglich, innerhalb des Wikis Programmcodes der statistischen Programmiersprache *R* (http://www.r-project.org) auszuführen. In der entsprechenden Wiki-Seite eines jeden Beispiels finden sich spezielle <R> Tags, die den *R*-Quellcode enthalten. Über einen output-Parameter lässt sich steuern, ob eine Grafik, Text oder eine Tabelle generiert werden soll. Innerhalb von *Arab*MM*Stat wird dieses Plugin genutzt, um beispielsweise verschiedene Verteilungen zu visualisieren; Abb. 4 zeigt ein Beispiel für die hypergeometrische Verteilung.

Abb. 3

Nutzung von LaTeX für ArabMM*Stat
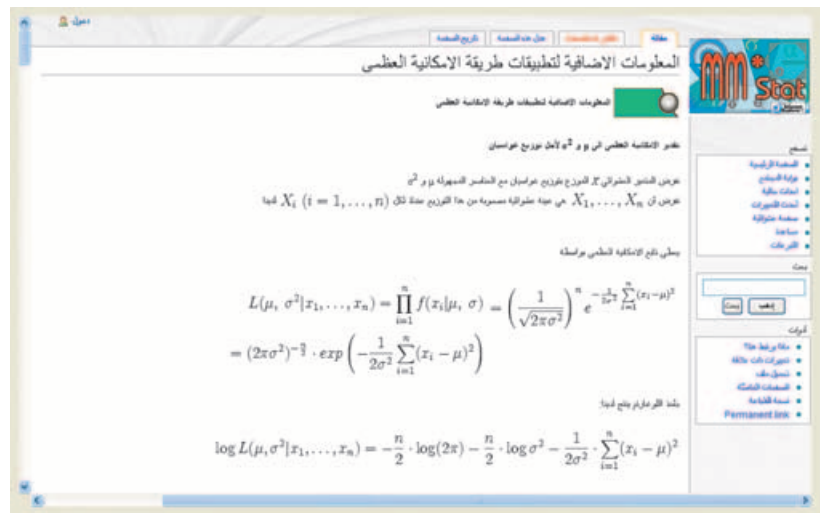
Abb. 4

Interaktives Beispiel für die Hypergeometrische Verteilung
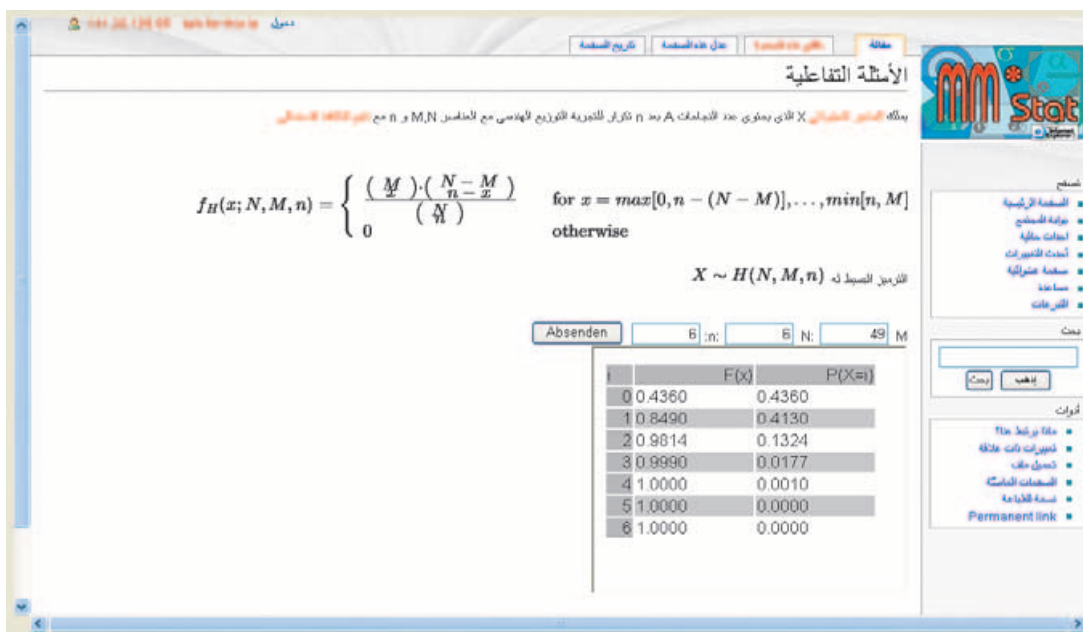
Taleb Ahmad
Jg. 1976. Studium der Wirt-
schaftswissenschaften an
der Tischrin-Universität Lat-
takia, Syrien, 1994–1999;
Diplom für Statistik in Latta-
kia 2000. Nach seinem
Masterstudiengang Statistik
an der Humboldt-Universität
zu Berlin promoviert er dort
seit 2005 an der Wirt-
schaftswissenschaftlichen
Fakultät. Von März 2006 bis
Juni 2007 war er Mitarbei-
ter am Multimedia-Projekt
*Arab*MM*Stat.

Kontakt
Humboldt-Universität
zu Berlin
Wirtschaftswissenschaft-
liche Fakultät
Fachgebiet Statistik
Ziegelstr. 13a
D–10117 Berlin
Tel.: +49 30 2093–5654
Fax: +49 30 2093–5649
E-Mail: tahmad@
wiwi.hu-berlin.de

Prof. Dr. Wolfgang Härdle
Jg. 1953. Nach Promotion
(Universität Heidelberg) und
Habilitation (Universität
Bonn) war W. Härdle Profes-
seur Ordinaire, CORE, an
der Université Catholique de
Louvain (1990–92). Seit
1992 C4-Professor an der
Wirtschaftswissenschaftli-
chen Fakultät der Humboldt-
Universität zu Berlin, Fach-
gebiet Statistik, Direktor des
Interdisziplinären Zentrums
CASE – Center for Applied
Statistics and Economics
seit 2001; Sprecher des SFB
373 »Quantifikation und
Simulation ökonomischer
Prozesse« (1/1994–2/2003);
jetzt Sprecher des SFB 649
»Ökonomisches Risiko«.

Kontakt
Humboldt-Universität
zu Berlin
Wirtschaftswissenschaft-
liche Fakultät
Spandauer Str. 1
D–10178 Berlin
Tel.: +49 30-2093–5630
Fax: +49 30-2093–5649
E-Mail: stat@wiwi.
hu-berlin.de
http://ise.wiwi.hu-berlin.de

Uwe Ziegenhagen
Jg. 1977. Studium der BWL
mit den Schwerpunkten
Wirtschaftsinformatik,
Entrepreneurship und
Statistik. Masterstudium
Statistik mit Schwerpunkt
Computational Statistics.
Forschungsthemen: E-Learn-
ing, Statistische Software.

Kontakt
Humboldt-Universität
zu Berlin
Wirtschaftswissenschaft-
liche Fakultät
Fachgebiet Statistik
Ziegelstr. 13a
D–10117 Berlin
Tel.: +49 30 2093–1469
Fax: +49 30 2093–5649
E-Mail: ziegenhagen@
wiwi.hu-berlin.de

Zusammenfassung
E-Learning-Elemente bilden in der Statistik eine wert-
volle Ergänzung zum regulären Lehr-Programm. Mit
*Arab*MM*Stat entwickeln wir ein auf gängigen Stan-
dards aufbauendes System, das langfristig von den
Nutzern selbst mit weiteren Inhalten gefüllt werden
kann und sogar als Basis für andere arabische Lern-
projekte dienen kann. Mit *Rwiki* und der Möglichkeit,
hochwertigen Formelsatz aus LaTeX zu erzeugen, kön-
nen auch komplexere mathematische und statistische
Inhalte sauber präsentiert und veranschaulicht wer-
den. Wir hoffen, dass *Arab*MM*Stat eine Bereicherung
des E-Learning-Angebots in der arabischen Welt dar-
stellt und eine Möglichkeit bietet, auf universitärer
Ebene zusammenzuarbeiten.

Literatur
Ahmad, T. / Härdle, W. / Mungo, J. (2006): On the Dif-
ficulty to Design Arabic E-learning System in Statis-
tics, COMPSTAT Proceedings in Computational Statis-
tics, Physica Verlag.
Ahmad, T. / Härdle, W. / Klinke, S. (2007): Using Wiki
to build an e-learning system in statistics in Arabic
language, ICMS, Vienna, Austria.
Dutta, S. / El-Hage, C. / Sabbagh, K. / Tarazi, P.
(2003): Challenges for Information and Communica-
tion Technology Development in the Arab World.
Härdle, W. / Klinke, S. / Ziegenhagen, U. (2006):
E-Learning – A Selective Review, COMPSTAT Proceed-
ings in Computational Statistics, Physica Verlag, S.
417–428.
Lagally, K. (1993): ArabTeX : a System for Typesetting
Arabic. User Manual Version 3.00, University Stutt-
gart, Germany.
Müller, M. / Rönz, B. / Ziegenhagen, U. (2000): The
Multimedia Project MM*Stat for Teaching Statistics.
In: COMPSTAT Proceedings in Computational Statis-
tics, Physica Verlag, S. 409–415.

Statistics of Risk Aversion

Enzo Giacomini*
Wolfgang Härdle*

* Humboldt-Universität zu Berlin, Germany

BERLIN

ECONOMIC RISK

SFB 649

Statistics of Risk Aversion

Giacomini, Enzo
Humboldt-Universität zu Berlin, CASE - Center for Applied Statistics and Economics
Spandauer Strasse 1
10178 Berlin, Germany
E-mail: giacomini@wiwi.hu-berlin.de

Härdle, Wolfgang
Humboldt-Universität zu Berlin, CASE - Center for Applied Statistics and Economics
Spandauer Strasse 1
10178 Berlin, Germany
E-mail: haerdle@wiwi.hu-berlin.de

1. Introduction

Option prices are a valuable source of information concerning risk assessments from investors about future financial payoffs. The information is summarized in the state price densities (SPD), the continuous counterpart (normalized by a constant) from Arrow-Debreu security prices. Under no arbitrage assumptions the state price densities - corresponding to a risk neutral measure Q - are derived from option prices as in Breeden and Litzenberger (1978). In contrast to the state price density, the historical density $p(x)$ describes the random variations of the underlying price.

In a utility based framework, investors facing financial risk have preference-indifference relations represented by a utility function. Standard economic theory enforces that the utility function is concave or equivalently, that investors are risk averse. Equilibrium and non-arbitrage arguments, as in Merton (1973), show that preferences from investors are related to the state price and historical densities, allowing to conclude the functional form of one given the functional form of the remaining two. Part of this relation is given by the pricing kernel (PK).

In this paper we investigate pricing kernels from DAX and ODAX data in a *time varying* approach and consider their *term structure*. In order to approximate and analyse the complex dynamic structure from pricing kernels and risk preferences across different maturities, we use *dynamic semiparametric factor models* (DSFMs). The obtained PKs exhibit risk proclivity in certain regions of returns and maturities. The dynamics of risk aversion and proclivity from investors preferences is investigated through sensitivity analysis with respect to the basis functions.

In the sequel pricing kernels are defined (section 2), their relation to utility functions and the DSFM estimation method are described (section 3). In the empirical part (section 4) pricing kernels are estimated from DAX and ODAX data sets and sensitivity analysis based on estimated loading coefficients and basis functions is performed.

2. Pricing Kernels

A flexible approach in a complete market is to assume that the price of a security follows a

diffusion process described by

$$
(1) \quad \frac{dS_t}{S_t} = \mu(S_t, t)dt + \sigma(S_t, t)dB_t
$$

where $t \in [0, T]$, B_t is a standard Brownian motion under measure P. Defining

$$
(2) \quad \lambda_t = \frac{\mu(S_t, t) - r}{\sigma(S_t, t)}
$$

as the market price of risk, the risk neutral measure Q is obtained by $\left. \frac{dQ}{dP} \right|_{\mathcal{F}_t} = \zeta_t$ where with (2)

$$
\zeta_t = \exp\left(-\int_0^t \lambda_u dB_u - \frac{1}{2} \int_0^t \lambda_u^2 du \right)
$$

The pricing kernel, or stochastic discount factor is defined for maturity $\tau = s - t$, $0 \le t \le s \le T$ as:

$$
(3) \quad M_{t,\tau} = e^{-r\tau} \frac{\zeta_s}{\zeta_t}
$$

Assuming the existence of a representative investor with utility function u, constant interest rates r and defining a wealth $\{W_s\}$ and a consumption process $\{C_s\}$, $C_s = 0$, the investor adjusts the amounts $\{\xi_s\}$ invested in the asset S_s at times s and consumes all his wealth at T, $C_T = W_T$. That means, he chooses ξ_s via the Merton optimization problem:

$$
\max_{\{\xi_s, t \le s \le T\}} E[u(W_T)]
$$

subjected to $W_s \ge 0$ and $dW_s = \{rW_s + \xi_s(\mu - r)\}ds + \xi_s \sigma dB_s$. In equilibrium all wealth is invested in the asset, at the end all wealth is consumed $C_T = W_T = S_T$. In the Merton model the PK is path independent and (3) is equal to the marginal rate of substitution:

$$
(4) \quad \frac{u'(S_T)}{u'(S_t)} = M_{t,\tau} = e^{-r\tau} \frac{q_t(S_T)}{p_t(S_T)}
$$

Here q_t, p_t denote the risk neutral and historical density at time t. The functional relation between the utility u and the densities q_t, p_t is given by

$$
(5) \quad u(S_T) = e^{-r\tau} u'(S_t) \int \frac{q_t(S_T)}{p_t(S_T)} dS_T
$$

3. PK Estimation with Dynamic Semiparametric Factor Models

Breeden and Litzenberger (1978) showed how the SPD $q_t(S_T)$ may be obtained from option prices. Ait-Sahalia and Lo (1998) used the estimate:

$$
(6) \quad \widehat{q}_t(S_T) = e^{r\tau} \left. \frac{\partial^2 C_{t,BS}\{S_t, K, \tau, r_t, \widehat{\sigma}_t(\kappa, \tau)\}}{\partial K^2} \right|_{K = S_T}
$$

where $C_{t,BS}$ is the Black-Scholes price at time t and $\widehat{\sigma}_t(\kappa, \tau)$ is a nonparametric estimator for the implied volatility.

Implied volatilities may be estimated from option prices. On each day $i = 1, \ldots, I$ there are J_i options traded. Each intra-day trade $j = 1, \ldots, J_i$ corresponds to an implied volatility $\sigma_{i,j}$, and a
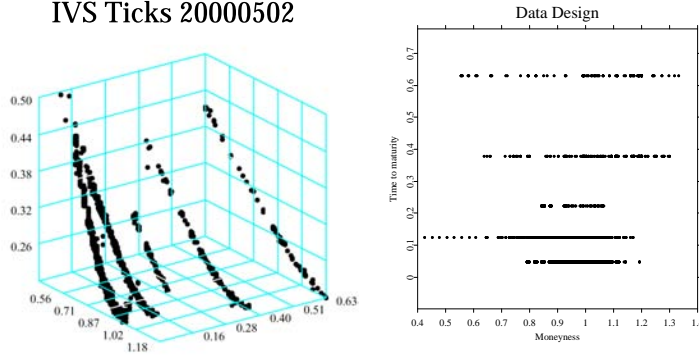
Figure 1: Call and put implied volatilities observed (left), data design (right), ODAX on 20000502

pair of moneyness and maturity $X_{i,j} = (\kappa_{i,j}, \tau_{i,j})^\top$. Moneyness is defined as $\kappa_{i,j} = \frac{K}{F(t_{i,j})}$ where K is strike and $F(t_{i,j}) = S_{t_{i,j}} \exp(r_{\tau_{i,j}} \tau_{i,j})$ are futures prices. Figure 1 depicts the implied volatilities corresponding to trades on ODAX in day 20000502 (dates are written as year, month, day).

Dynamic semi-parametric factor models (DSFM), Fengler et al. (2007), employ the time series structure of implied volatilities regressing log implied volatilities $Y_{i,j} = \log \sigma_{i,j}$ on $X_{i,j}$ using smooth basis functions m_l, $l = 0, \ldots, L$ weighted with factor loadings $z_{i,l}$:

$$(7) \quad Y_{i,j} = \sum_{l=0}^{L} z_{i,l} m_l(X_{i,j}) + \varepsilon_{i,j}$$

where $\varepsilon_{i,j}$ is noise and $z_{i,0} \equiv 1$. Following Borak et al. (2007), the basis functions are expanded using a series estimator for functions $\psi_k : \mathbb{R} \to \mathbb{R}$ and coefficients $\gamma_{l,k} \in \mathbb{R}$

$$m_l(X_{i,j}) = \sum_{k=1}^{K} \gamma_{l,k} \psi_k(X_{i,j})$$

Defining the matrices $Z = (z_{i,l})$, $\Gamma = (\gamma_{l,k})$ for $i = 1, \ldots, I$, $k = 1, \ldots, K$ and $l = 0, \ldots, L$ we obtain the least square estimators as

$$(\widehat{\Gamma}, \widehat{Z}) = \arg \min_{\Gamma \in \mathcal{G}, Z \in \mathcal{Z}} \sum_{i=1}^{I} \sum_{j=1}^{J} \left\{ Y_{i,j} - z_i^\top \Gamma \psi(X_{i,j}) \right\}^2$$

where $z_i = (z_{i,0}, \ldots, z_{i,L})^\top$, $\psi(x) = \{\psi_1(x), \ldots, \psi_K(x)\}^\top$, $\mathcal{G} = \mathcal{M}(L+1, K)$, $\mathcal{Z} = \{Z \in \mathcal{M}(I, L+1) : z_{i,0} \equiv 1\}$ and $\mathcal{M}(a, b)$ is the set of all $(a \times b)$ matrices. The estimators for the basis functions in (7) are $\widehat{m}_l(x) = \widehat{\gamma}_l^\top \psi(x)$ where $\gamma_l = (z_{l,1}, \ldots, z_{l,K})^\top$. Denoting $\widehat{m} = (\widehat{m}_0, \ldots, \widehat{m}_L)^\top$, the implied volatility at time i is estimated as $\widehat{\sigma}_i(\kappa, \tau) = \exp\left\{\widehat{z}_i^\top \widehat{m}(\kappa, \tau)\right\}$.

Using (6), the state price density may be approximated by

$$(8) \quad \widehat{q}_t(\kappa, \tau, \widehat{z}_t, \widehat{m}) = \varphi(d_2) \left\{ \frac{1}{K\widehat{\sigma}_t \sqrt{\tau}} + \frac{2d_1}{\widehat{\sigma}_t} \frac{\partial \widehat{\sigma}_t}{\partial K} + \frac{K\sqrt{\tau} d_1 d_2}{\widehat{\sigma}_t} \left(\frac{\partial \widehat{\sigma}_t}{\partial K} \right)^2 + K\sqrt{\tau} \frac{\partial^2 \widehat{\sigma}_t}{\partial K^2} \right\} \Bigg|_{K=S_T}$$

where $\varphi(x)$ is the standard normal pdf, $d_1 = \frac{\log\left(\frac{S_t}{K}\right) + (r + \frac{1}{2}\widehat{\sigma}_t^2)\tau}{\widehat{\sigma}_t \sqrt{\tau}}$ and $d_2 = d_1 - \widehat{\sigma}_t \sqrt{\tau}$. As in Ait-Sahalia and Lo (2000) we define an estimate $\widehat{M}_t(\kappa, \tau)$ of the PK as the ratio between the estimated SPD and
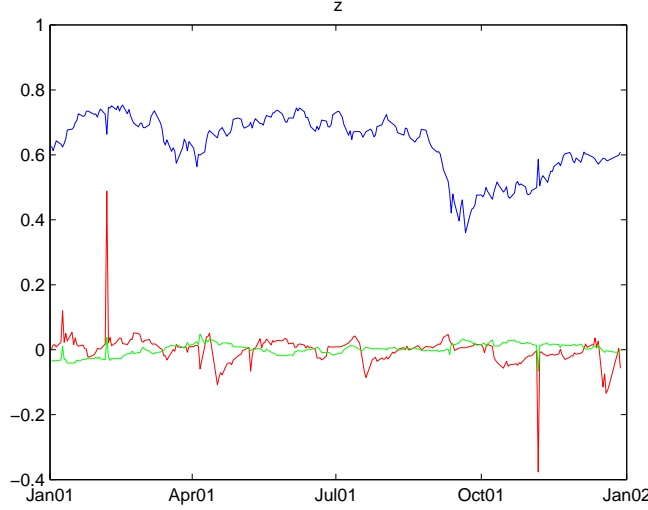
Figure 2: Loading factors $\widehat{z}_{t,l}$, $l = 1, 2, 3$ from the top

the estimated p:

$$\widehat{M}_t(\kappa, \tau, \widehat{z}_t, \widehat{m}) \;=\; e^{-r_t\tau}\frac{\widehat{q}_t(\kappa, \tau, \widehat{z}_t, \widehat{m})}{\widehat{p}_t(\kappa, \tau)}$$

Here $\widehat{p}_t$ is estimated by a GARCH(1,1) model. It is our interest in this paper to examine the dynamic structure of (9).

4. Empirical Results

Here SPDs and PKs are estimated from intraday DAX and ODAX data. The options data ranges from 20010101 to 20020101 corresponding to 253 trading days. The SPD $\widehat{q}_t$ (8) is estimated with 4 basis functions. The historical density $\widehat{p}_t$ is estimated with GARCH(1,1) from the last 240 observations. The time series of loading factors $\{\widehat{z}_{t,l}\}$ and basis functions $\widehat{m}_l$, $l = 0, \ldots, 3$, $t = 1, \ldots, 253$ are in figures 2 and 3. The summary statistics from loading factors series are in table 1. From (8) and (9) we obtain a sequence of 253 SPDs and PKs moving on time over a grid of moneyness and maturities, figure 4 shows one shot of this sequence, at day 20010710.

Risk averse utilities u are concave. Hence, (4) implies that under risk aversion pricing kernels are monotone decreasing in moneyness. Figure 4 shows that DAX PKs - similarly to S&P PKs, Ait-Sahalia and Lo (2000) - present risk proclivity at certain moneyness levels (PKs are increasing between 0.9 and 1) and these levels vary across maturities, hence we verify the *empirical pricing kernel paradox*.

In order to analyse the dynamics of risk aversion and proclivity implied in PKs we focus on the SPDs, as these contain information about preferences from investors. In figure 5 we can see that the mean of $\widehat{q}_t(S_T)$ and the factor loadings $\widehat{z}_{t,1}$ from the basis function $\widehat{m}_1$ are correlated. We investigate the influence of the basis functions $\widehat{m}_l$ on the mean, variance and skewness of SPDs by sensitivity analysis to changes in the loadings $\{\widehat{z}_{t,l}\}$: the SPDs are recalculated with new loadings, where one of the coefficients varies linearly and the remaining are hold fixed.

We define *scenario loadings* W^l corresponding to a linear increase on factor l in N steps from

Table 1: Loading factors, descriptive statistics

	min	max	median	mean	std.dev.
$\widehat{z}_{t,1}$	0.36	0.75	0.66	0.63	0.09
$\widehat{z}_{t,2}$	-0.37	0.49	0.01	0.00	0.05
$\widehat{z}_{t,3}$	-0.07	0.05	0.00	0.00	0.02
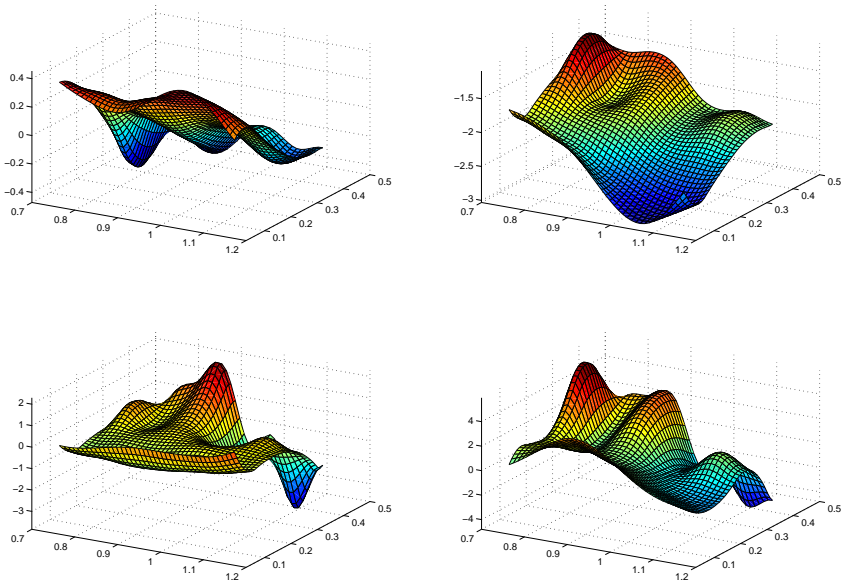
Figure 3: Basis functions $\widehat{m}_l$, $l = 0, \ldots, 3$ clockwise
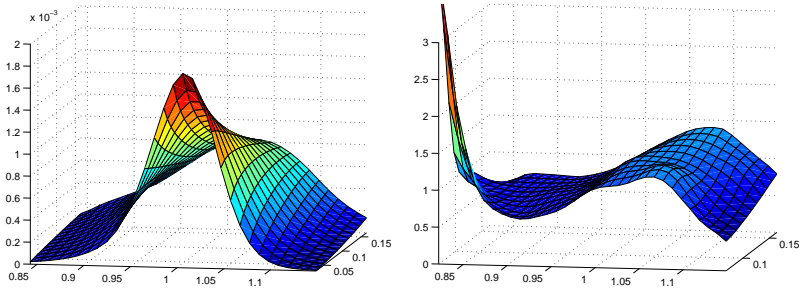
Figure 4: Estimated SPD (left) and PK (right) across κ and τ at $t = 20010710$
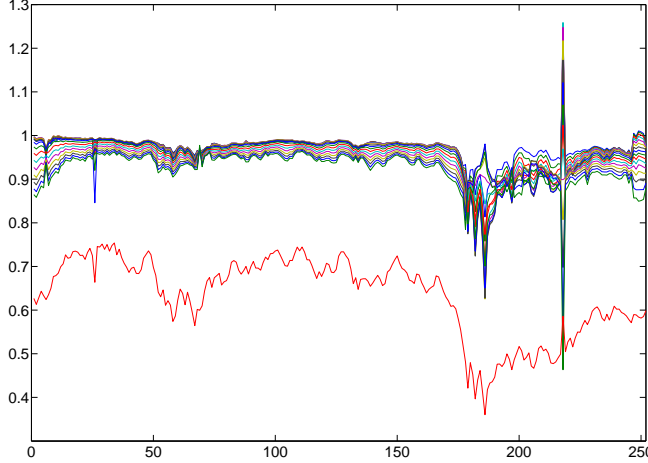
Figure 5: Mean from SPD for $\tau = 25, \ldots, 75$ days (above), $\widehat{z}_{t,1}$ (below)

levels d_l to u_l. The remaining coefficients are held constant at median of estimated values. The matrices $W^l = (w^l_{n,j})$, $l, j = 0, \ldots, 3$, and $n = 1, \ldots, N$ contain the scenario loadings to factor l as

$$w^l_{n,j} = \left\{ d_j + \frac{n-1}{N-1}(u_j - d_j) \right\} \mathbf{1}(j = l) + med(\widehat{z}_{t,j})\mathbf{1}(j \neq l)$$

with levels given by $d_l = \min \widehat{z}_{t,l} - 0.5|\min \widehat{z}_{t,l}|$ and $u_l = \max \widehat{z}_{t,l} + 0.5|\max \widehat{z}_{t,l}|$. SPDs estimated with loadings W^l are under influence of variations in factor l while the remaining factor are constant at a typical level (median). The observed changes in mean, variance and skewness may be considered as a typical effect of variation in factor l.

Figure 6 shows the mean, variance and skewness of SPD obtained with W^l, for $l = 1, 2, 3$, $N = 50$ and three different maturities (25, 40 and 75 days) plotted against n. From the middle column, we verify that increasing factor 2 decreases mean while increasing variance and skewness for short maturities (full lines) and increases mean while decreasing variance and skewness for long maturities (dashed lines). From the left column, we verify that increasing factor 1 increases mean while decreasing variance and skewness for all maturities.

Figure 7 plots PKs obtained with W^l, for $l = 1, 2, 3$, $N = 50$ and two maturities, short (20) and long (75 days), against moneyness. The pricing kernels are calculated with weak ($n = 1$) medium (25) and strong (50) loadings from each factor. The influence of factor 1 is represented by the left pair in figure 7: for both maturities weak loadings result in risk proclivity for $\kappa > 1$ while strong loadings in risk proclivity for $\kappa < 1$.

In the middle pair we see that factor 2 has the opposite influence at short maturities: weak loadings result in risk proclivity for $\kappa < 1$, strong loadings for $\kappa > 1$. This effect is not clear at long maturities. Increasing loadings of the factor 3 shifts the risk proclivity region to higher values of κ in both maturities, as observed in the right pair of figure 7.
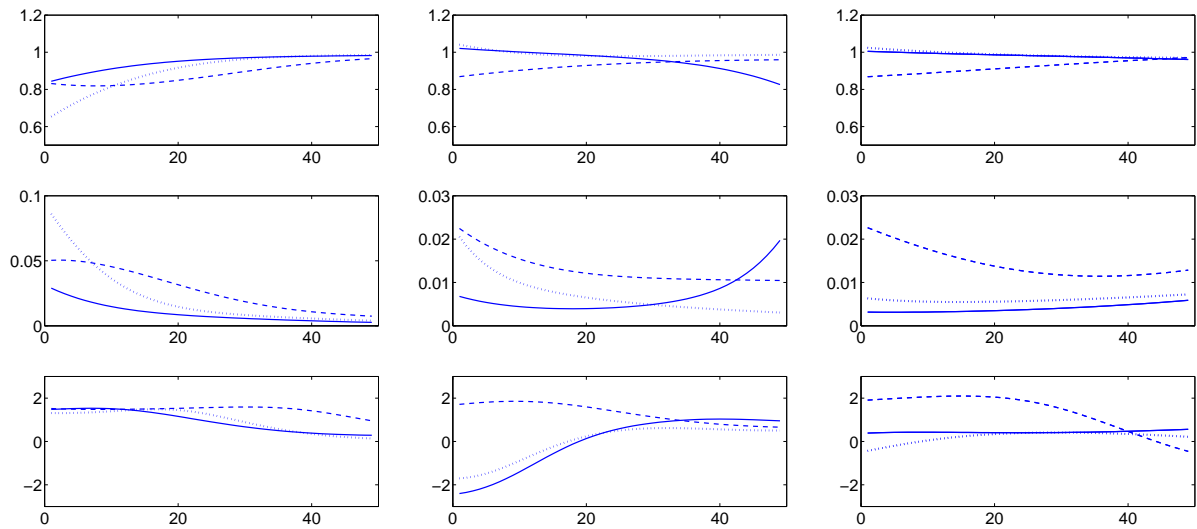
Figure 6: *Mean, variance and skewness from SPD (from the top) plotted against n. For W^1, W^2 and W^3 (from the left), $\tau = 25$ (full), 40 (dotted) and 75 (dashed) days, $N = 50$*
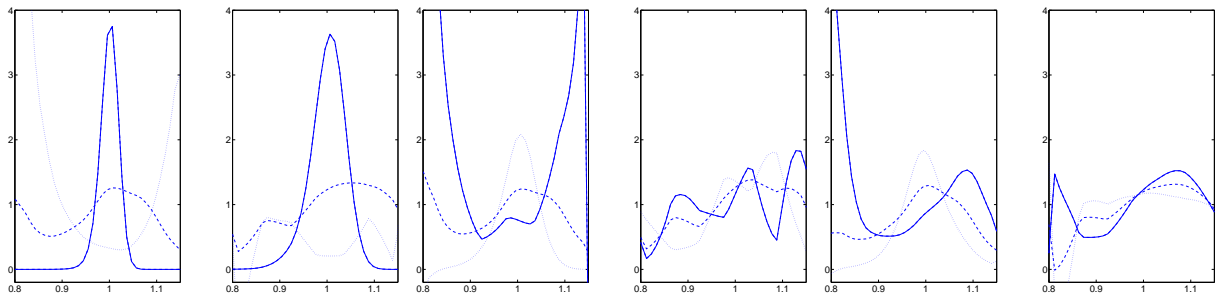
Figure 7: *PK obtained with loadings W^1 (left pair), W^2 (middle pair) and W^3 (right pair), $\tau = 25$ (first) and 75 (second in each pair) days, $n = 1$ (dotted), 25 (dashed) and 50 (full)*

REFERENCES

Ait-Sahalia, Y. and Lo, A. (1998). Nonparametric estimation of state-price densities implicit in financial asset prices. *Journal of Finance*, 53:499–547.

Ait-Sahalia, Y. and Lo, A. (2000). Nonparametric risk management and implied risk aversion. *Journal of Econometrics*, 94:9–51.

Borak, S., Härdle, W., Mammen, E., and Park, B. (2007). Time series modelling with semiparametric factor dynamics. *Discussion paper, SFB 649 - Humboldt-Universität zu Berlin*, 2007–23.

Breeden, D. and Litzenberger, R. (1978). Prices of state-contingent claims implicit in options prices. *Journal of Business*, 51:621–651.

Debreu, G. (1959). *Theory of Value. An Axiomatic Analysis of Economic Equilibrium.* John Wiley and Sons, Inc., New York.

Fengler, M., Härdle, W., and Mammen, E. (2007). A semiparametric factor model for implied volatility surface dynamics. *Journal of Financial Econometrics*, 5:189–218.

Merton, R. (1973). Theory of rational option pricing. *The Bell Journal of Economics and Management Science*, 4:141–183.

ABSTRACT

Information about risk preferences from investors is essential for modelling a wide range of quantitative finance applications. Valuable information related to preferences can be extracted from option prices through pricing kernels. In this paper, pricing kernels and their term structure are estimated in a time varying approach from DAX and ODAX data using dynamic semiparametric factor model (DSFM). DSFM smooths in time and space simultaneously, approximating complex dynamic structures by basis functions and a time series of loading coefficients. Contradicting standard risk aversion assumptions, the estimated pricing kernels indicate risk proclivity in certain levels of return. The analysis of the time series of loading coefficients allows a better understanding of the dynamic behaviour from investors preferences towards risk.

AKNOWLEDGEMENTS

SFB 649 Discussion Paper Series 2007

001 "Trade Liberalisation, Process and Product Innovation, and Relative Skill Demand" by Sebastian Braun, January 2007.

002 "Robust Risk Management. Accounting for Nonstationarity and Heavy Tails" by Ying Chen and Vladimir Spokoiny, January 2007.

003 "Explaining Asset Prices with External Habits and Wage Rigidities in a DSGE Model." by Harald Uhlig, January 2007.

004 "Volatility and Causality in Asia Pacific Financial Markets" by Enzo Weber, January 2007.

005 "Quantile Sieve Estimates For Time Series" by Jürgen Franke, Jean-Pierre Stockis and Joseph Tadjuidje, February 2007.

006 "Real Origins of the Great Depression: Monopolistic Competition, Union Power, and the American Business Cycle in the 1920s" by Monique Ebell and Albrecht Ritschl, February 2007.

007 "Rules, Discretion or Reputation? Monetary Policies and the Efficiency of Financial Markets in Germany, 14th to 16th Centuries" by Oliver Volckart, February 2007.

008 "Sectoral Transformation, Turbulence, and Labour Market Dynamics in Germany" by Ronald Bachmann and Michael C. Burda, February 2007.

009 "Union Wage Compression in a Right-to-Manage Model" by Thorsten Vogel, February 2007.

010 "On σ−additive robust representation of convex risk measures for unbounded financial positions in the presence of uncertainty about the market model" by Volker Krätschmer, March 2007.

011 "Media Coverage and Macroeconomic Information Processing" by Alexandra Niessen, March 2007.

012 "Are Correlations Constant Over Time? Application of the CC-TRIG$_t$-test to Return Series from Different Asset Classes." by Matthias Fischer, March 2007.

013 "Uncertain Paternity, Mating Market Failure, and the Institution of Marriage" by Dirk Bethmann and Michael Kvasnicka, March 2007.

014 "What Happened to the Transatlantic Capital Market Relations?" by Enzo Weber, March 2007.

015 "Who Leads Financial Markets?" by Enzo Weber, April 2007.

016 "Fiscal Policy Rules in Practice" by Andreas Thams, April 2007.

017 "Empirical Pricing Kernels and Investor Preferences" by Kai Detlefsen, Wolfgang Härdle and Rouslan Moro, April 2007.

018 "Simultaneous Causality in International Trade" by Enzo Weber, April 2007.

019 "Regional and Outward Economic Integration in South-East Asia" by Enzo Weber, April 2007.

020 "Computational Statistics and Data Visualization" by Antony Unwin, Chun-houh Chen and Wolfgang Härdle, April 2007.

021 "Ideology Without Ideologists" by Lydia Mechtenberg, April 2007.

022 "A Generalized ARFIMA Process with Markov-Switching Fractional Differencing Parameter" by Wen-Jen Tsay and Wolfgang Härdle, April 2007.

023 "Time Series Modelling with Semiparametric Factor Dynamics" by Szymon Borak, Wolfgang Härdle, Enno Mammen and Byeong U. Park, April 2007.

024 "From Animal Baits to Investors' Preference: Estimating and Demixing of the Weight Function in Semiparametric Models for Biased Samples" by Ya'acov Ritov and Wolfgang Härdle, May 2007.

025 "Statistics of Risk Aversion" by Enzo Giacomini and Wolfgang Härdle, May 2007.

Yxilon – A Client/Server Based Statistical Environment

Wolfgang Härdle*
Sigbert Klinke*
Uwe Ziegenhagen*

* Humboldt-Universität zu Berlin, Germany

BERLIN

ECONOMIC RISK

SFB 649

Yxilon – A Client/Server Based Statistical Environment*

May 30, 2007

Wolfgang Härdle, Sigbert Klinke, Uwe Ziegenhagen
Humboldt-Universität zu Berlin
School of Business and Economics
Unter den Linden 6
10099 Berlin, Germany
haerdle/sigbert/ziegenhagen@wiwi.hu-berlin.de

Along with many others, we agree that a modern education in statistics needs to incorporate the practical analysis of real datasets, which are usually more complex than the common examples found in standard textbooks.

The software used in the teaching of statistics includes standard spreadsheet environments such as OpenOffice and Excel and dedicated commercial and non-commercial packages such as R, Minitab or SPSS. With the freely available Yxilon environment we add another package and proliferate the statistical programming language XploRe, using a modern client/server based architecture. This architecture has the capabilities of serving

---

statistical results in a variety of flavors for different groups of users. In this paper we describe the general setup of the Yxilon environment and present selected technical details.

1 From XploRe to Yxilon

Following the requirements we gave in Härdle, Klinke & Ziegenhagen (2004), the value of a statistical software environment does not only depend on general features such as stability of the software and the set of available methods, but also on the way it allows an efficient interaction with the user. Nowadays the user operates statistical software either via character input in a terminal or mouse. Well-known packages for the former method are R and Matlab, for the latter Minitab and SPSS, which also allows the extension of its methods via VisualBasic.

XploRe, the predecessor of Yxilon and its scientific fundament, belonged to the category of language-operated statistical packages. The codes, the so-called *quantlets*, were written in a C-style programming language, which was then evaluated by XploRe. Details of the XploRe syntax can be found in Härdle, Klinke & Müller (2000). XploRe included more than 1500 quantlets, covering especially non- and semiparametric methods and the statistics of financial markets. The internal design of XploRe was monolithic, which means that a large proportion of components was tied closely together in one executable. This made the maintenance of the code complicated and led us to the decision to reimplement the XploRe language in the Yxilon project, using a simplified and modular client/server approach with two tiers. On the one hand this allows the easier adding of new modules, on the other hand existing modules can easily be maintained or exchanged.

2 The Client

The Yxilon environment is operated by the user through the client. In contrast to the server which was written in C++, it was implemented in Sun Microsystems' Java. For our purpose Java offers various advantages if compared with other programming languages. Compiled Java code can be run on any platform for which a Java Runtime Environment (JRE) is available, it offers a large set of methods for internet-related techniques (TCP/IP, XML, SOAP), it can be embedded into HTML pages and is supported by a large user community.

The client satisfies different needs: It hides the technical implementation of the communication and offers an test platform for the implementation of new functions. The current version does not only offer basic editing and the exchange of code and results

with the server but provides advanced editing features such as the highlighting and autocompletion of code and database connectivity.

When the user presses the connect button, the client reads the communication settings from a textfiles and exchanges information with the server, e.g. login and password. After this handshake process the user edits his sourcefiles and sends them to the server.
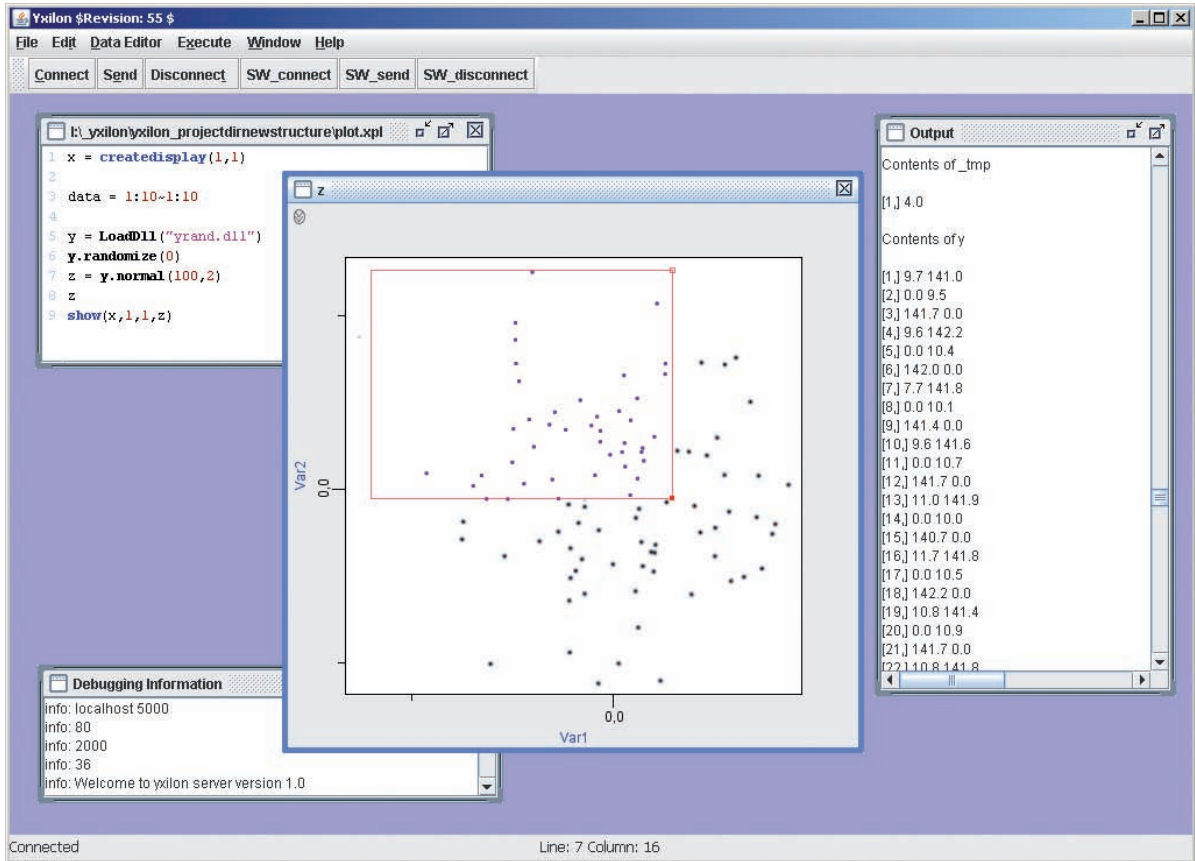
Figure 1: Yxilon Java client

For the display of graphics under Java there are different solutions. We have chosen the Jaspplot library (Yamamoto, Nakano, Fujiwara & Kobayashi 2004), a library explicitly designed for the display of statistical graphics. This library can easily be embedded into Yxilon and offers a wide range of statistical plots. It furthermore allows the linking of different plots, so selected (brushed) observations in one plot are automatically highlighted in another.

Figure 1 depicts a screenshot of the Yxilon client, showing relevant windows of the interface. In the topleft corner the editor window with the Yxilon code needed to produce the scatterplot, in the bottom left the 'Debug' window, which shows selected information

3

about the communication packages sent an received, and the output window used for the display of numerical results. Furthermore the client also contains a database wizard, which allows to connect to different databases such as MySQL, MS Access or Oracle via the Java Database Connector (JDBC).

3 The Server and the Communication Protocol

For the communication between a server and its client there are numerous ways to exchange information. We evaluated common techniques such as *Remote Procedure Calls* (RPC) and *Remote Method Interaction* (RMI) but finally decided to develop our own binary protocol. This approach had the disadvantage that a significant amount of time had to be spent on low-level network programming but provided the advantage that we kept the complete control of the communication, which resulted in a very small overhead, therefore in a very fast communication.

Table 1 gives an overview about different types of objects currently used for the communication. There are ten types of objects; different types of lists, numeric and character matrices as well as a different types of exceptions. These exceptions are extremely useful when the code submitted by the user is incomplete or erroneous since they contain detailed information about the location and type of the error. A runtime exception is usually caused by errors in the evaluation of syntactically correct programs, e.g. when a dll could not be loaded properly. User exceptions are defined in the Yxilon code. Their task is to give feedback in form of warnings, if for example conditions concerning the size and values of matrices are not met. The parser exception is the third type, capturing incomplete input. Here the user receives a message that the input, he or she submitted, had syntactical errors.

When the server executable is started, it reads the necessary settings from a textfile, e.g. the network port it binds to, the paths for the dlls and quantlets and the login/password needed for clients to connect. After the successful loading of these settings a TCP/IP server port is opened. From now on clients can connect.

ID	Object Type	Description
0	empty object	to denote `null` objects
1	numeric matrix	stores numeric matrices (up to 3 dimensions)
2	string matrix	stores string objects (up to 3 dimensions)
3	object list	may contain matrices or lists
4	named list	a named list object
5	parameter list	stores named key/value pairs
9	runtime exception	captures errors in the server
10	user exception	allows user to define own errors/warnings
13	parser exception	captures incomplete input

Table 1: Overview of Communication Objects

When Yxilon objects are sent from the server to the client as a serialized data stream, the first information transmitted is the object type. Based on this type identification the client reads the remaining bytes and generates the necessary objects as matrices and lists which are then used e.g. to generate a statistical graphics or character output.

The server component of Yxilon is currently available for Microsoft Windows only, however we will provide versions for Linux and Solaris in the future. Due to the concept of modularity and exchangable parts the server itself has only very limited functionality, such as the management of objects, basic calculations and the communication stack. All computational functionality above the level of simple algebra and basic matrix manipulations has been sourced out to dynamic link libraries (DLLs). Listing 1 illustrates the simple structure of one of these libraries.

Following the initial import statements we define in the *LibMain* function a set of variables such as the author and the version of the DLL. These information, accessible also from Yxilon code, provide valuable information in the development process, e.g. to ensure that a required dynamic link library has a certain version level. The current list of available libraries for Yxilon contains the handling of matrices, mathematical functions and the generation of various random numbers, further libraries are planned.

```
1  #include <LibInclude.hpp>
2  #define CExtObj CExpCompObject
3
4  EXPORT LibMain(CExtLibrary *pLib){
5         pLib->setAuthor(L"Sheldon Kelly");
6         pLib->setVersion(L"1.0");
7         pLib->setName(L"Mathematical Functions");
8         pLib->addFunction(L"abs", L"Yabs");
9         pLib->addFunction(L"ceil", L"Yceil");
10        pLib->addFunction(L"floor", L"Yfloor");
11        pLib->addFunction(L"cos", L"Ycos");
12 };
```

Listing 1: Example of C++ code for a dynamic link library

4 Conclusion

The Yxilon project is our attempt to proliferate the ideas and concepts of the XploRe language in a modern, modular environment. By using innovative user features together with a fast and reliable communication protocol we want to provide an easily learnable environment especially suitable for the education of students. The next steps on the agenda include the further integration of statistical graphics into the client and the conversion of existing XploRe code to Yxilon. We highly appreciate feedback from interested students and scientists, more information and downloads can be found online at `http://www.quantlet.org`.

REFERENCES

Härdle, W., Klinke, S. & Müller, M. (2000), XploRe Learning Guide, Springer-Verlag Heidelberg.

Härdle, W., Klinke, S. & Ziegenhagen, U. (2004), Yxilon – Designing The Next Generation, Vertically Integrable Statistical Software Environment.*in* Proceedings of the 36th Syposium on the Interface, Baltimore

Yamamoto, Y., Nakano, J., Fujiwara, T. & Kobayashi, I. (2004), A mixed User Interface for a Statistical System, *in* Computational Statistics, 17:379–393.

SFB 649 Discussion Paper Series 2007

For a complete list of Discussion Papers published by the SFB 649, please visit http://sfb649.wiwi.hu-berlin.de.

001 "Trade Liberalisation, Process and Product Innovation, and Relative Skill Demand" by Sebastian Braun, January 2007.

002 "Robust Risk Management. Accounting for Nonstationarity and Heavy Tails" by Ying Chen and Vladimir Spokoiny, January 2007.

003 "Explaining Asset Prices with External Habits and Wage Rigidities in a DSGE Model." by Harald Uhlig, January 2007.

004 "Volatility and Causality in Asia Pacific Financial Markets" by Enzo Weber, January 2007.

005 "Quantile Sieve Estimates For Time Series" by Jürgen Franke, Jean-Pierre Stockis and Joseph Tadjuidje, February 2007.

006 "Real Origins of the Great Depression: Monopolistic Competition, Union Power, and the American Business Cycle in the 1920s" by Monique Ebell and Albrecht Ritschl, February 2007.

007 "Rules, Discretion or Reputation? Monetary Policies and the Efficiency of Financial Markets in Germany, 14th to 16th Centuries" by Oliver Volckart, February 2007.

008 "Sectoral Transformation, Turbulence, and Labour Market Dynamics in Germany" by Ronald Bachmann and Michael C. Burda, February 2007.

009 "Union Wage Compression in a Right-to-Manage Model" by Thorsten Vogel, February 2007.

010 "On σ−additive robust representation of convex risk measures for unbounded financial positions in the presence of uncertainty about the market model" by Volker Krätschmer, March 2007.

011 "Media Coverage and Macroeconomic Information Processing" by Alexandra Niessen, March 2007.

012 "Are Correlations Constant Over Time? Application of the CC-TRIG$_t$-test to Return Series from Different Asset Classes." by Matthias Fischer, March 2007.

013 "Uncertain Paternity, Mating Market Failure, and the Institution of Marriage" by Dirk Bethmann and Michael Kvasnicka, March 2007.

014 "What Happened to the Transatlantic Capital Market Relations?" by Enzo Weber, March 2007.

015 "Who Leads Financial Markets?" by Enzo Weber, April 2007.

016 "Fiscal Policy Rules in Practice" by Andreas Thams, April 2007.

017 "Empirical Pricing Kernels and Investor Preferences" by Kai Detlefsen, Wolfgang Härdle and Rouslan Moro, April 2007.

018 "Simultaneous Causality in International Trade" by Enzo Weber, April 2007.

019 "Regional and Outward Economic Integration in South-East Asia" by Enzo Weber, April 2007.

020 "Computational Statistics and Data Visualization" by Antony Unwin, Chun-houh Chen and Wolfgang Härdle, April 2007.

021 "Ideology Without Ideologists" by Lydia Mechtenberg, April 2007.

022 "A Generalized ARFIMA Process with Markov-Switching Fractional Differencing Parameter" by Wen-Jen Tsay and Wolfgang Härdle, April 2007.

023   "Time Series Modelling with Semiparametric Factor Dynamics" by Szymon Borak, Wolfgang Härdle, Enno Mammen and Byeong U. Park, April 2007.

024   "From Animal Baits to Investors' Preference: Estimating and Demixing of the Weight Function in Semiparametric Models for Biased Samples" by Ya'acov Ritov and Wolfgang Härdle, May 2007.

025   "Statistics of Risk Aversion" by Enzo Giacomini and Wolfgang Härdle, May 2007.

026   "Robust Optimal Control for a Consumption-Investment Problem" by Alexander Schied, May 2007.

027   "Long Memory Persistence in the Factor of Implied Volatility Dynamics" by Wolfgang Härdle and Julius Mungo, May 2007.

028   "Macroeconomic Policy in a Heterogeneous Monetary Union" by Oliver Grimm and Stefan Ried, May 2007.

029   "Comparison of Panel Cointegration Tests" by Deniz Dilan Karaman Örsal, May 2007.

030   "Robust Maximization of Consumption with Logarithmic Utility" by Daniel Hernández-Hernández and Alexander Schied, May 2007.

031   "Using Wiki to Build an E-learning System in Statistics in Arabic Language" by Taleb Ahmad, Wolfgang Härdle and Sigbert Klinke, May 2007.

032   "Visualization of Competitive Market Structure by Means of Choice Data" by Werner Kunz, May 2007.

033   "Does International Outsourcing Depress Union Wages? by Sebastian Braun and Juliane Scheffel, May 2007.

034   "A Note on the Effect of Outsourcing on Union Wages" by Sebastian Braun and Juliane Scheffel, May 2007.

035   "Estimating Probabilities of Default With Support Vector Machines" by Wolfgang Härdle, Rouslan Moro and Dorothea Schäfer, June 2007.

036   "Yxilon – A Client/Server Based Statistical Environment" by Wolfgang Härdle, Sigbert Klinke and Uwe Ziegenhagen, June 2007.
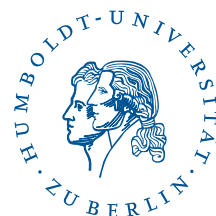
On extracting information implied in options

M. Benko · M. Fengler · W. Härdle · M. Kopa

Abstract   Options are financial instruments with a payoff depending on future states of the underlying asset. Therefore option markets contain information about expectations of the market participants about market conditions, e.g. current uncertainty on the market and corresponding risk. A standard measure of risk calculated from plain vanilla options is the implied volatility (IV). IV can be understood as an estimate of the volatility of returns in future period. Another concept based on the option markets is the state-price density (SPD) that is a density of the future states of the underlying asset. From raw data we can recover the IV function by nonparametric smoothing methods. Smoothed IV estimated by standard techniques may lead to a non-positive SPD which violates no arbitrage criteria. In this paper, we combine the IV smoothing with SPD estimation in order to correct these problems. We propose to use the local polynomial smoothing technique. The elegance of this approach is that it yields all quantities needed to calculate the corresponding SPD. Our approach operates only on the IVs—a major improvement comparing to the earlier multi-step approaches moving through the Black–Scholes formula from the prices to IVs and vice-versa.

Keywords   Implied volatility · Nonparametric regression

M. Benko (✉) · M. Fengler · W. Härdle
Humboldt-Universität zu Berlin, Berlin, Germany
e-mail: benko@wiwi.hu-berlin.de

M. Kopa
Charles University, Prague, Czech Republic

1 Introduction

Visualization techniques play an enormous role in practical finance. To the broad audience, the best-known examples include technical analysis or charting techniques from asset management which aim at exhibiting geometric patterns in historical price data (Edwards and Magee 1966; Murphy 1986). But also in option trading, visualization techniques are of high importance: market makers and option traders continuously monitor plots of the so-called implied volatility (IV) smile during a trading day. Despite violating the assumptions of the Black–Scholes (BS) option pricing model, from which it is derived, IV is a widely accepted state variable of market sentiments. In particular IV can be understood as an estimate of the volatility of future returns of the underlying asset, and is an indicator of the uncertainty of the market (Britten-Jones and Neuberger 2000). Moreover, IV and its correlate, the so-called state price density (SPD), serve as a decisive pricing factor for valuing plain vanilla and exotic options alike. Naturally, adequate visualization and modeling techniques are crucial tools in daily practice.

Implied volatility data are obtained from observed price data of European options, which are derivatives paying out $\max(S_T - K, 0)$, if it is a call option, and, in case of a put option, $\max(K - S_T, 0)$, where S_T is the terminal stock price at some future date T (the expiry date), and K is called the strike price. Assuming a geometric Brownian motion as driver for the underlying stock price dynamics analytic formulae for time-t prices of calls $C_t, t < T$, can be obtained (Black and Scholes 1973):

$$C_t(S_t, K, \tau, r, \sigma) = S_t \Phi(d_1) - K e^{-r\tau} \Phi(d_2) \tag{1}$$

where $d_1 = \frac{\ln(S_t/K) + (r + 0.5\sigma^2)\tau}{\sigma\sqrt{\tau}}, d_2 = d_1 - \sigma\sqrt{\tau}; r$ is a riskless interest rate, $\tau = T - t$ is referred to as time-to-maturity, and σ is an unknown and constant volatility parameter. The price for a put option P_t can be derived by no-arbitrage principles from the put-call parity: $P_t = C_t - S_t + e^{-r\tau}K$. The IV $\tilde{\sigma}$ is defined as the volatility σ, which matches the Black–Scholes price C_t with the price $\tilde{C}_t$ actually observed on the market. In antagonism to the actual assumptions of the BS model IV is *not* a constant, but a u-shaped or skewed function in strikes K, called the IV 'smile'. Also across time-to-maturity τ IV can exhibit pronounced curvatures, in which case one speaks of the IV 'surface', $\tilde{\sigma}_t(K, \tau)$. For visualization, and eventually pricing, parametric approaches (Brockhaus et al. 2000, Chap. 2) or non- and semiparametric methods are employed (Fengler 2005b), which are fitted to the observed option price data such as quote ticks or closing prices.

These approaches need to cope with a number of challenges. First the functional choice must have sufficient flexibility to generate the salient patterns and shapes of IV smiles and surfaces. And second the estimation or calibration process must be robust against noise present in price observations due to market micro-structures effects, such as bid-ask spreads, discrete ticks in prices or quotes, non-synchronous trading, effects due to the auction mechanism itself, or simply to misprints (for a detailed analysis of errors in IV data we refer to Hentschel 2003). While nonparametric estimation strategies are natural candidates in such a situation, great care must be taken in order

to avoid the estimates violating natural conditions given by financial theory, as this can have hazardous consequences for price computations. Recent advances in this field include the work by Kahale (2004) and Fengler (2005b), who introduced interpolation schemes in the space of call prices. After estimating the price functions under suitable shape constraints they derive the IV surface from these estimates.

The aim of this paper is to propose a direct method of estimating the IV function from observed noisy IVs without resorting to option price functions. The input data can be intra-day prices, or some subset of them—e.g. closing prices for a given period (day). Denoting IVs observed on K_i and τ_i by $\tilde{\sigma}_i$ and the corresponding 'true' IV function by $\sigma(K_i, \tau_i)$, $i = 1, \ldots, n$ we assume the following model:

$$\tilde{\sigma}_i = \sigma(K_i, \tau_i) + \varepsilon_i, \tag{2}$$

where ε_i models the noise, n denotes the number of observed IVs. Since we do not wish to assume a particular model for the underlying stock price process we cannot estimate model (2) parametrically. Standard approaches include local polynomial estimators as presented by Fan and Gijbels (1996) amongst others, see Shimko (1993), Fengler et al. (2003) and Cont and da Fonseca (2002) for such applications to estimate the IV surface. We follow these approaches using local polynomial estimators, but impose constraints that avoid violating the no-arbitrage conditions. To this end we introduce the concept of the state price density:

$$q_{t, S_T}(x, \tau) \stackrel{\text{def}}{=} e^{r\tau} \frac{\partial^2 C_t(K, T)}{\partial K^2} \bigg|_{K=x}. \tag{3}$$

As already mentioned, the SPD, which is a density function, can be interpreted as a correlate of the IV surface, as it can be expressed as a function of the IV surface and its derivatives, see Sect. 2 for the details. With noisy IV data, nonparametric techniques can lead to estimates where the SPD is negative in some regions of the estimation domain. A negative SPD, however, corresponds to arbitrage opportunities (possibility of a 'free-lunch') in the market. Since the no-arbitrage condition is an integral part of financial theory, Musiela and Rutkowski (1997), we constrain the SPD to be non-negative. This leads to a nonlinear constrained optimization. To obtain local polynomial estimators a system of nonlinear minimization problems is solved for a given grid of strikes and time-to-maturities. All computations are done in system GAMS 22.0 using the nonlinear solver MINOS.

Advantages of the proposed method are that it does not assume arbitrage-free input data as in Kahale (2004) and operates only on the IVs—a major improvement comparing to the earlier approaches moving through the BS formula from the prices to IVs and vice-versa.

The paper is organized as follows: in Sect. 2 the concept of the SPD and its relation to IV is explained. Estimating the IV function for a fixed and observed time-to-maturity is discussed in Sect. 3. In Sect. 4 the estimation of the IV surface on an arbitrary grid of time-to-maturities is discussed. The proposed methods are applied to the IVs of options on the German stock index (DAX).

2 State price density

In a dynamically complete market, the absence of arbitrage opportunities implies the existence of an equivalent martingale measure Q, Harrison and Kreps (1979) and Harrison and Pliska (1981), that is uniquely characterized by the state price density q_{t,S_T} of the underlying price process S_t. Therefore the price $\Pi_t(H)$ of a derivative with a payoff-function $H(S_T)$ depending on the asset with price S_T at the expiration date T, is given by the well-known arbitrage-free pricing formula:

$$\Pi_t(H) = e^{-r\tau} E_Q(H|\mathcal{F}_t)$$

$$= e^{-r\tau} \int_0^\infty H(s) q_{t,S_T}(s,\tau) ds \quad \text{for all } t \in [0, T].$$

The last formula is of vital practical importance, since given an estimate of q_{t,S_T} one can immediately price any path-independent derivative.

The connection between SPD and IV can be established by combining (3) and (1). After some algebra we obtain:

$$q_{t,S_T}(x,\tau) = e^{r\tau} S_t \sqrt{\tau} \varphi(d_1(x,\tau)) \left\{ \frac{1}{x^2 \sigma(K,\tau)\tau} + \frac{2d_1(x,\tau)}{x\sigma(x,\tau)\sqrt{\tau}} \frac{\partial \sigma(K,\tau)}{\partial K} \right|_{K=x}$$

$$+ \frac{d_1(x,\tau)d_2(x,\tau)}{\sigma(x,\tau)} \left(\left. \frac{\partial \sigma}{\partial K} \right|_{K=x} \right)^2 + \left. \frac{\partial^2 \sigma}{\partial K^2} \right|_{K=x} \right\}, \tag{4}$$

where $d_1(x,\tau) = \frac{\ln(S_t/x) + (r + 0.5\sigma^2(x,\tau))\tau}{\sigma(x,\tau)\sqrt{\tau}}$, $d_2 = d_1(x,\tau) - \sigma(x,\tau)\sqrt{\tau}$ and φ is the p.d.f. of a standard normal random variable, Brunner and Hafner (2003).

The IV function and the SPD are naturally connected. This motivates an estimation technique combining the concepts of SPD estimation and IV smoothing. In Sect. 3 we present such a method for estimating the IV function for a given time-to-maturity τ observed on a given day. For an exhaustive review on the literature of option-implied SPDs we refer to Jackwerth (2004).

3 Estimating the IV smile for fixed maturity

In this section, we focus on estimating the IV function from a set of observed option prices with fixed time-to-maturity τ. Since, in this section τ is fixed, we may simplify the notation in (2) by setting $\sigma(K_i) \stackrel{\text{def}}{=} \sigma(K_i, \tau_i)$ for $i = 1, \ldots, n_\tau$, where n_τ denotes the number of observed IVs with time-to-maturity τ. Then (2) simplifies to

$$\tilde{\sigma}_i = \sigma(K_i) + \varepsilon_i. \tag{5}$$

In Fig. 1 we present IVs calculated from the daily option prices on February 2, 2006 for the time-to-maturity of 15 days. As can be seen the IVs are quite rough and especially
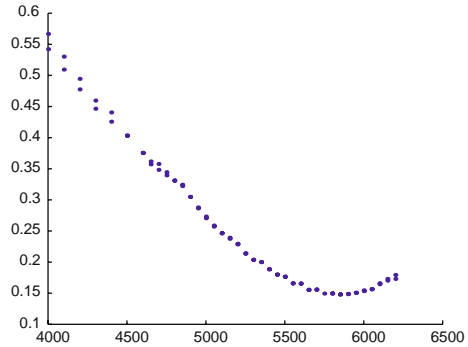
for the boundary regions (strike below 5,000 and above 6,000) two IV values are observed for the same strike (corresponding to puts and calls). By the put-call parity, these IVs must coincide. This noise, observed in IV in Fig. 1, is modeled via error term in (5).

As already outlined, our aim is to combine the regression model (5) with the SPD in (4). According to (4) the SPD is a function of the IV function and its first and second derivative. We therefore propose a local quadratic estimator that automatically provides an estimate of the IV function and its derivatives, see Fan and Gijbels (1996). The local quadratic estimator $\hat{\sigma}(K)$ of the regression function $\sigma(K)$ in the point K is defined by the solution of the following local least squares criterion:

$$\min_{\alpha_0,\alpha_1,\alpha_2} \sum_{i=1}^{n_\tau} \left\{ \tilde{\sigma}_i - \alpha_0 - \alpha_1(K_i - K) - \alpha_2(K_i - K)^2 \right\}^2 \mathcal{K}_h(K - K_i), \qquad (6)$$

where $\mathcal{K}_h(K - K_i) \stackrel{\text{def}}{=} \frac{1}{h}\mathcal{K}\left(\frac{K-K_i}{h}\right)$ and $\mathcal{K}$ is a so-called kernel function—typically a symmetric density function with compact support, e.g. $\mathcal{K}(u) = \frac{3}{4}(1 - u^2)I(|u| \leq 1)$ (Epanechnikov kernel) and h is called bandwidth. Since $\mathcal{K}_h$ is nonnegative within the (localization) window $[K - h, K + h]$, points outside of this interval have no influence on the estimator $\hat{\sigma}(K)$. The choice of h governs the trade-off between bias and variance of $\hat{\sigma}(K)$—large h yields a small variance but the large bias, small h vice versa, see Härdle (1990). Comparing (6) with the Taylor expansion of σ yields

$$\alpha_0 = \hat{\sigma}(K_i), \quad \alpha_1 = \hat{\sigma}'(K_i), \quad 2\alpha_2 = \hat{\sigma}''(K_i), \qquad (7)$$

which makes the estimation of the regression function and its first two derivatives ($\hat{\sigma}'$ and $\hat{\sigma}''$ respectively) possible. In order to take the non-negativity of the SPD into account, we need to perform (6) under the condition $q_{t,S_T} \geq 0$, on the entire support. Since we consider only one time point t we will ease the notation in this section to $S_t = S$, and $q_{t,S_T}(K, \tau)$ to $q(K, \tau)$. For the fixed point (K, τ) and by plugging (7) into (6) we can rewrite $d_1 = \frac{\ln(S/K)+(r+0.5(\alpha_0)^2)\tau}{\alpha_0\sqrt{\tau}}$, $d_2 = d_1 - \alpha_0\sqrt{\tau}$ and the SPD can

be estimated at the point (K, τ) by

$$\hat{q}(K, \tau) = F\sqrt{\tau}\varphi(d_1) \left\{ \frac{1}{K^2\alpha_0\tau} + \frac{2d_1}{K\alpha_0\sqrt{\tau}}\alpha_1 + \frac{d_1 d_2}{\alpha_0}(\alpha_1)^2 + 2\alpha_2 \right\}, \quad (8)$$

where $F = Se^{r\tau}$. Summarizing, the optimization problem can be written as:

$$\min_{\alpha_0, \alpha_1, \alpha_2} \sum_{i=1}^{n_\tau} \left\{ \tilde{\sigma}_i - \alpha_0 - \alpha_1(K_i - K) - \alpha_2(K_i - K)^2 \right\}^2 \mathcal{K}_h(K - K_i) \quad (9)$$

subject to $F\sqrt{\tau}\varphi(d_1) \left\{ \frac{1}{K^2\alpha_0\tau} + \frac{2d_1}{K\alpha_0\sqrt{\tau}}\alpha_1 + \frac{d_1 d_2}{\alpha_0}(\alpha_1)^2 + 2\alpha_2 \right\} \geq 0$ where $d_1 = \frac{\ln(S/K)+(r+0.5(\alpha_0)^2)\tau}{\alpha_0\sqrt{\tau}}$, $d_2 = d_1 - \alpha_0\sqrt{\tau}$. As already mentioned, this leads to a nonlinear optimization problem. All computations were done in GAMS 22.0 using the solver MINOS. For an overview on the nonlinear optimization, see Bertsekas (1999) among others.

The constrained estimate, and the corresponding SPD for the dataset using the Epanechnikov kernel and $h = 200$ are displayed on the Fig. 2. The smoothing parameter has been chosen by keeping the bias small on the one hand and guaranteeing enough data for each point K where the estimate (9) was constructed. More sophisticated choice of the parameter h seems to be possible, by using the standard cross-validation arguments. Adaptive methods proposed recently by Spokoiny (2006) could be employed under some further conditions. The confidence intervals for the estimated IV and the SPD can be constructed by using the classical idea of wild residual bootstrap, see Härdle (1990) among others.

Let us now consider the more complicated situation of intra-day data. The IVs are calculated for each realized trade on the exchange EUREX. The crucial difference
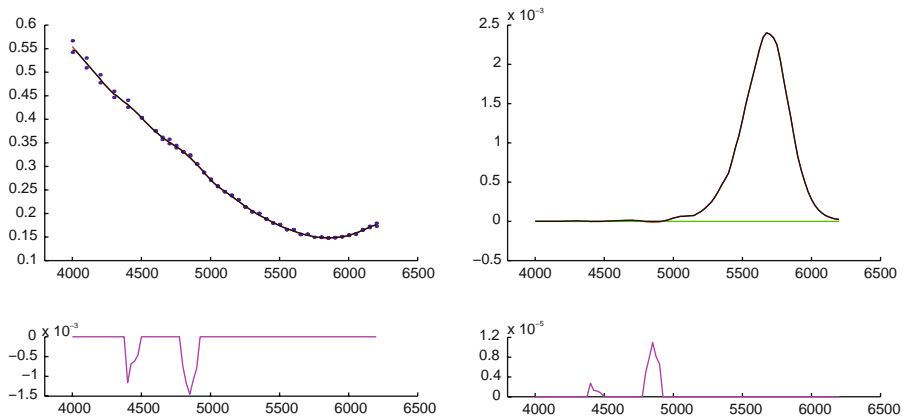
Fig. 2 *Left figure* Smoothed IV function, constrained (*black line*) and unconstrained (*red line*), and corresponding SPDs, constrained (*black line*) and unconstrained (*red line*). Daily data, February 2, 2006, horizontal axis is the strike level. The lower figures illustrate the difference between constrained and unconstrained smoothing

in comparison to the daily data is that the underlying DAX prices that are used in the calculation of the IV are not constant over time. In order to standardize the IV with respect to the underlying stock (in our case DAX index), we express the IV as a function of *futures moneyness* $\kappa \stackrel{\text{def}}{=} \frac{K}{F}$. The IV function is located at $\kappa = 1$—so called at-the-money (ATM). Using this standardization, the local quadratic estimate $\hat{\sigma}(\kappa)$ of $\sigma(\kappa)$ is given by the solution of:

$$\min_{\alpha_0,\alpha_1,\alpha_2} \sum_{i=1}^{n_\tau} \left\{ \tilde{\sigma}_i - \alpha_0 - \alpha_1(\kappa_i - \kappa) - \alpha_2(\kappa_i - \kappa)^2 \right\}^2 \mathcal{K}_h(\kappa - \kappa_i). \qquad (10)$$

Again, using same argument as in (7), we obtain: $\alpha_0 = \hat{\sigma}(\kappa_i)$, $\alpha_1 = \hat{\sigma}'(\kappa_i)$, $2\alpha_2 = \hat{\sigma}''(\kappa_i)$. Next, from the definition of futures moneyness we obtain: $K = F\kappa$, $\frac{\partial K}{\partial \kappa} = F$, $\frac{\partial \sigma}{\partial K} = \frac{\partial \sigma}{\partial \kappa} \frac{1}{F}$, $\frac{\partial^2 \sigma}{\partial^2 K} = \frac{1}{F^2} \frac{\partial^2 \sigma}{\partial^2 \kappa}$. After some straightforward calculations we obtain

$$d_1 = \frac{-\ln(\kappa e^{r\tau}) + (r + 0.5\sigma^2)\tau}{\sigma\sqrt{\tau}} = \frac{\sigma^2\tau/2 - \ln(\kappa)}{\sigma\sqrt{\tau}}, \ d_2 = d_1 - \sigma\sqrt{\tau}.$$

Finally we obtain the SPD expressed as a function of κ (note that after analytical calculations the SPD needs to be rescaled in order to have $\int q(\kappa)d\kappa = 1$):

$$q(\kappa, \tau) = \sqrt{\tau}\varphi(d_1) \left\{ \frac{1}{\kappa^2\sigma\tau} + \frac{2d_1}{\kappa\sigma\sqrt{\tau}} \frac{\partial \sigma}{\partial \kappa} + \frac{d_1 d_2}{\sigma} \left(\frac{\partial \sigma}{\partial \kappa} \right)^2 + \frac{\partial^2 \sigma}{\partial \kappa^2} \right\}. \qquad (11)$$

Hence the analogue of (9) can be obtained by constraining (10) with respect to the corresponding non-negative SPD (11).

The left plot of the Fig. 3 shows the intra-day data (blue points) on December 29, 2003, the red line is the constrained local quadratic smoother with Epanechnikov kernel and $h = 0.045$, the black line is the constrained local polynomial estimator with the same bandwidth.

Since S_t and F_t are not constant in intra-day data, the daily average S of (S_t) and $F = Se^{r\tau}$ is used in (11). As a nice side-effect, we can see that the constrained estimator is more robust against outlyers. The corresponding SPDs are plotted in the right plot.

It should be noted that we are using one functional optimization criterion (9) for estimating the function (IV) and its first and second derivative simultaneously. As argued in Fan and Gijbels (1996) if we were interested in these functions separately, it could be advantageous to consider a separate objective function for each of these functions with different bandwidths or different order of the polynomial used in (9). However, the elegance of our approach is that all quantities needed for determining the SPD are obtained from (9) immediately.

So far we focused on the positivity of the SPD function. The additional condition— the integral condition ($\int q_{t,S_T}(s,\tau)ds = 1$) could be considered similarly. However, since this condition is not a local but a global condition, it will be computationally more involved. It will lead to single large nonlinear minimization problem. Solving it is much
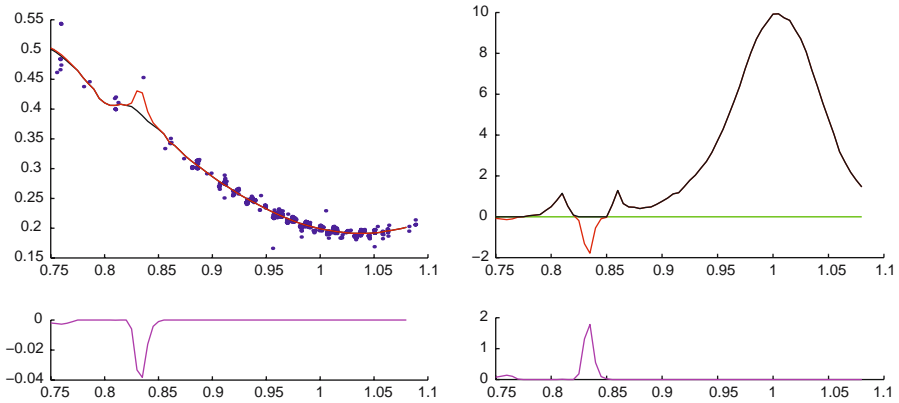
Fig. 3 *Left figure* Smoothed IV function, constrained (*black line*) and unconstrained (*red line*), and corresponding SPDs, constrained (*black line*) and unconstrained (*red line*). Intra-day data, December 29, 2003, horizontal axis is moneyness level. The lower figures illustrate the difference between constrained and unconstrained smoothing

more demanding than solving the local system (9). Moreover in this case the tails of the SPD that are outside the interval of the observed IV need to be considered. These tails can be estimated by considering the cumulative distribution function corresponding to the SPD that is a function of the IV and its first derivative, see also Brunner and Hafner (2003) for a discussion of this strategy.

In the next section we comment on the application of these ideas to the two dimensional smoothing—recovering of the whole IV surface—function of strike (or futures moneyness) and time-to-maturity.

4 Estimating the IV-surface

In the previous section we considered the estimation of the IV-function for a single time-to-maturity. The aim of this section is to develop a technique for the (arbitrage-free) estimation for any time-to-maturity, i.e. for the 2-dimensional IV-surface $\sigma(K, \tau)$.

The condition on the non-negative SPD can be taken from (9). The arbitrage in 'τ direction' is often referred to as calendar arbitrage. Kahale (2004) argues, assuming zero interest-rates, that so called total-variance $w(K, \tau) \stackrel{\text{def}}{=} \sigma^2(K, \tau)\tau$ is strictly increasing in τ. Fengler (2005a) generalizes this result to a deterministic time-varying interest rate, arguing that $w(\kappa, \tau) \stackrel{\text{def}}{=} \sigma^2(\kappa, \tau)\tau$ should be strictly increasing in τ under no-arbitrage. Assuming the model (2), our aim is to use the condition on the total variance in our smoothing estimate. Let us first introduce the two dimensional local polynomial estimator.

The idea of local polynomial estimation in higher dimensions is a straightforward generalization of the one-dimensional case. A standard—unconstrained

two-dimensional local quadratic estimator $\hat{\sigma}(\kappa, \tau)$ is given by the minimizer of:

$$
\min_{\alpha} \sum_{i=1}^{n} \mathcal{K}_H(\kappa - \kappa_i, \tau - \tau_i)\{\tilde{\sigma}_i - \alpha_0 - \alpha_1(\kappa_i - \kappa)
$$
$$
-\alpha_2(\tau_i - \tau) - \alpha_{1,1}(\kappa_i - \kappa)^2 - \alpha_{1,2}(\kappa_i - \kappa)(\tau_i - \tau) - \alpha_{2,2}(\tau_i - \tau)^2\}^2 \quad (12)
$$

where $\mathcal{K}_H(u) \overset{\text{def}}{=} \frac{1}{\det H}\mathcal{K}(H^{-1}u)$ is a (bivariate) kernel function with bandwidths (matrix) H. Comparing (12) with a truncated bi-variate Taylor expansion of $\sigma(\kappa, \tau)$ shows $\alpha_0 = \hat{\sigma}(\kappa, \tau)$, $\alpha_1 = \frac{\partial\hat{\sigma}}{\partial\kappa}(\kappa, \tau)$, $\alpha_2 = \frac{\partial\hat{\sigma}}{\partial\tau}(\kappa, \tau)$, $\alpha_{1,1} = \frac{\partial^2\hat{\sigma}}{2\partial\kappa^2}(\kappa, \tau)$, $\alpha_{2,2} = \frac{\partial^2\hat{\sigma}}{2\partial\tau^2}(\kappa, \tau)$, $\alpha_{1,2} = \frac{\partial^2\hat{\sigma}}{\partial\kappa\partial\tau}(\kappa, \tau)$. Since in our application it is typical to have small number of design points in the τ direction, we propose a parsimonious smoother $\hat{\sigma}(\kappa, \tau)$ given by the solution of:

$$
\min_{\alpha} \sum_{i=1}^{n} \mathcal{K}_H(\kappa - \kappa_i, \tau - \tau_i)\{\tilde{\sigma}_i - \alpha_0 - \alpha_1(\kappa_i - \kappa)
$$
$$
-\alpha_2(\tau_i - \tau) - \alpha_{1,1}(\kappa_i - \kappa)^2 - \alpha_{1,2}(\kappa_i - \kappa)(\tau_i - \tau)\}^2. \quad (13)
$$

The idea of (13) is to construct a local smoother quadratic in κ and linear in τ. Again the unconstrained estimate (13) can yield an estimate that contradicts the no-arbitrage assumptions. Our aim is to solve (13) with respect to non negative corresponding SPD and total variance strictly increasing in τ.

Consider first a problem of estimating the IV function for a fixed τ which is not observed in the data set. Define $\hat{w}(\kappa, \tau) = \hat{\sigma}^2(\kappa, \tau)\tau$. Since $\frac{\partial\hat{w}}{\partial\tau} > 0$ can be rewritten as $2\tau\alpha_0\alpha_2 + \alpha_0^2 > 0$ for a given (single) τ we solve the optimization problem (13) constrained by:

$$
\hat{q}(\kappa, \tau) = \sqrt{\tau}\varphi(d_1)\left\{\frac{1}{\kappa^2\alpha_0\tau} + \frac{2d_1}{\kappa\alpha_0\sqrt{\tau}}\alpha_1 + \frac{d_1 d_2}{\alpha_0}\alpha_1^2 + 2\alpha_{1,1}\right\} \geq 0
$$
$$
2\tau\alpha_0\alpha_2 + \alpha_0^2 > 0 \quad (14)
$$

where $d_1 = \frac{\alpha_0^2\tau/2 - \ln(\kappa)}{\alpha_0\sqrt{\tau}}$, $d_2 = d_1 - \alpha_0\sqrt{\tau}$ and $F_\tau = Se^{r\tau}$, for a given (but arbitrary) κ.

If we are interested in estimating the entire IV-surface $\hat{\sigma}(\kappa, \tau)$ for a set of maturities $\{\tau_1, \ldots, \tau_L\}$ and for a given value κ, we need to ensure $\hat{w}(\kappa, \tau_l,) \leq \hat{w}(\kappa, \tau_{l'})$, for all $\tau_l < \tau_{l'}'$. This leads to the following optimization problem:

$$
\min_{\alpha(l)} \sum_{l=1}^{L}\sum_{i=1}^{n} \mathcal{K}_H(\kappa - \kappa_i, \tau_l - \tau_i)\{\tilde{\sigma}_i - \alpha_0(l)
$$
$$
-\alpha_1(l)(\kappa_i - \kappa) - \alpha_2(l)(\tau_i - \tau) - \alpha_{1,1}(l)(\kappa_i - \kappa)^2
$$
$$
-\alpha_{1,2}(l)(\kappa_i - \kappa)(\tau_i - \tau)\}^2 \quad (15)
$$
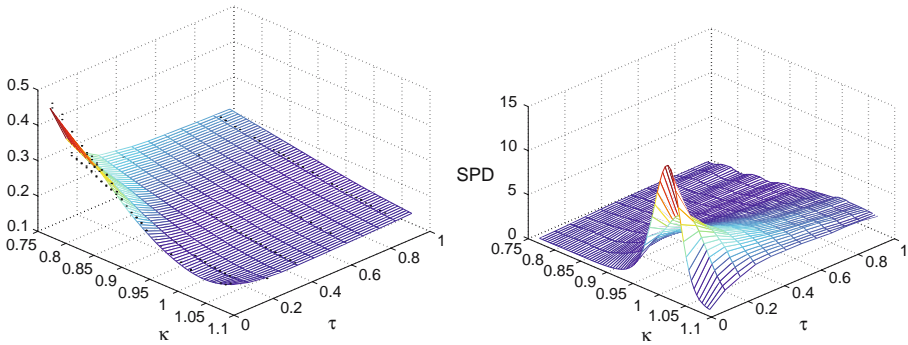
Fig. 4 *Left figure* Smoothed IV surface left figure and corresponding family of SPDs, daily Data, February 2, 2006, horizontal axis are the moneyness (κ) and time-to-maturity (τ)

subject to

$$\sqrt{\tau_l}\varphi(d_1(l))\left\{\frac{1}{\kappa^2\alpha_0(l)\tau_l} + \frac{2d_1(l)}{\kappa\alpha_0(l)\sqrt{\tau_l}}\alpha_1(l) + \frac{d_1(l)d_2(l)}{a_0(l)}\alpha_1^2(l) + 2\alpha_{1,1}(l)\right\} \geq 0,$$

$$d_1(l) = \frac{\alpha_0^2(l)\tau_l/2 - \ln(\kappa)}{\alpha_0(l)\sqrt{\tau_l}}, \quad d_2(l) = d_1(l) - a_0(l)\sqrt{\tau_l}, \quad l = 1, \dots, L$$

$$2\tau_l\alpha_0(l)\alpha_2(l) + \alpha_0^2(l) > 0 \quad l = 1, \dots, L$$

$$\alpha_0^2(l)\tau_l < \alpha_0^2(l')\tau_l', \quad \tau_l < \tau_l'. \tag{16}$$

Comparing (15)–(16) with the one-dimensional problem, (15)–(16) calculates for a given κ the estimates for all given τ_l in one step in order to guarantee increasing $\hat{w}$ in τ. The bivariate kernel function $\mathcal{K}_H(\kappa - \kappa_i, \tau_l - \tau_i)$ is given by the product of two univariate kernel functions: $\mathcal{K}_{h_\kappa}(\kappa - \kappa_i) = \frac{1}{h_\kappa}\mathcal{K}\left(\frac{\kappa - \kappa_i}{h_\kappa}\right)$, $\mathcal{K}_{h_\tau}(\tau - \tau_i)$ where $\mathcal{K}$ is the Epanechnikov kernel.

Figure 4 shows the results for the daily data on February 2, 2006. By analogy to the univariate case, we consider a global $h_\kappa = 0.05$. The h_τ is chosen as a function increasing in τ: $h_\tau = 0.2$ for $0 < \tau \leq \frac{1}{3}$, 0.3 for $\frac{1}{3} < \tau \leq \frac{2}{3}$ and 0.4 for $\frac{2}{3} < \tau \leq 1$. This choice was made again in such a way to cover a sufficient number of observations. Problem (15)–(16) was solved in system GAMS 22.0—solver MINOS, for each κ separately. Similar to the one-dimensional problem, a more sophisticated choice of the smoothing parameter H can be done by considering the cross-validation, however, in the general situation, we need to optimize the cross-validation criterion with respect to a 2×2 matrix. This choice of smoothing parameters is computationally much more involved.

As an alternative to estimating the IV surface by smoothing the IVs, we may smooth the total variance w directly. This can be done via the same route, since the SPD can be expressed in terms of the total variance w. From the definition of w follows:

$d_1 = \sqrt{w}/2 - \ln(\kappa)/\sqrt{w}, d_2 = -\ln(\kappa)/\sqrt{w} - \sqrt{w}/2$ and

$$q(\kappa, \tau) = \frac{\sqrt{\tau}\varphi(d_1)}{\sqrt{w\tau}} \left\{ \frac{1}{\kappa^2} + \frac{d_1}{k\sqrt{w}} \left(\frac{\partial w}{\partial k} \right) + \frac{d_1 d_2 - 1}{4w} \left(\frac{\partial w}{\partial k} \right)^2 + \frac{1}{2} \left(\frac{\partial^2 w}{\partial k^2} \right) \right\}$$

since $\partial w/\partial \kappa = 2\sigma\tau\partial\sigma/\partial\kappa$ and $\partial^2 w/\partial\kappa^2 = \frac{1}{2w}(\frac{\partial w}{\partial \kappa})^2 + 2\sqrt{w\tau}(\frac{\partial^2 \sigma}{\partial \kappa^2})$. Using these expressions we can design an estimator for the IV-surface. The comparative advantage to smoothing in IV is the simpler structure of the constraints. On the other hand, in this case, the IV surface must be calculated by setting $\hat{\sigma}(\kappa, \tau_l) \overset{\text{def}}{=} \sqrt{\hat{w}(\kappa, \tau_l)\tau_l^{-1}}$.

References

Bertsekas D (1999) Nonlinear programming. Athena Scientific, Belmont

Black F, Scholes M (1973) The pricing of options and corporate liabilities. J Polit Econ 81:637–654

Breeden D, Litzenberger R (1978) Price of state-contingent claims implicit in options prices. J Bus 51:621–651

Britten-Jones M, Neuberger A (2000) Option prices, implied price process and stochastic volatility. J Fin 55(2):839–866

Brockhaus O, Farkas M, Ferraris A, Long D, Overhaus M (2000) Equity derivatives and market risk models. Risk Books, London

Brunner B, Hafner R (2003) Arbitrage-free estimation of the risk-neutral density from the implied volatility smile. J Comput Fin 7:75–106

Cont R, da Fonseca J (2002) The dynamics of implied volatility surfaces. J Quant Fin 2(1):45–60

Edwards R, Magee J (1966) Technical analysis of stock trends, 5th edn. John Magee, Boston

Fan J, Gijbels I (1996) Local polynomial modelling and its applications. Chapman and Hall, London

Fengler M (2005a) Arbitrage-free smoothing of the implied volatility surface. Working paper 2005-019, SFB 649, Humboldt-Universität zu Berlin

Fengler M (2005b) Semiparametric modeling of implied volatility. Springer, Berlin

Fengler M, Härdle W, Villa P (2003) The dynamics of implied volatilities: a common principle components approach. Rev Deriv Res 6:179–202

Harrison J, Kreps D (1979) Martingales and stochastic integral in the theory of continuous trading. Stochast Process Appl 11:215–260

Harrison J, Pliska S (1981) Martingales and arbitrage in multiperiod securities markets. J Econ Theory 20:381–408

Härdle W (1990) Applied nonparametric regression. Cambridge University Press, Cambridge

Hentschel L (2003) Errors in implied volatility estimation. J Fin Quant Anal 38:779–810

Jackwerth JC (2004) Option-implied risk neutral distributions and risk aversion, Research Foundation of AIMR, Charlotteville, USA

Hull CJ, White A (1987) The pricing of options on assets with stochastic volatilities. J Fin 42:281–300

Kahale N (2004) An arbitrage-free interpolation of volatilities. RISK 17(5):102–106

Murphy J (1986) Technical analysis of the futures market. New York Institute of Finance, New York

Musiela M, Rutkowski M (1997) Martingale methods in financial modelling. Springer, Heidelberg

Rebonato R (1999) Volatility and correlation. Wiley series in financial in financial ingeniering. Wiley, New York

Spokoiny V (2006) Local parametric methods in nonparametric estimation. Springer, Heidelberg

Shimko D (1993) Bounds on probability. RISK 6(4):33–37

Realoptionen und Immobilienbewertung: Eine Umsetzungsstudie

Steffen Brenner, Rainer Schulz und Wolfgang Härdle*

Zusammenfassung

In der Studie werden entwicklungsfähige Grundstücke des Berliner Stromversorgers Bewag als Realoptionen bewertet. Im Gegensatz zu traditionellen Bewertungsverfahren werden dadurch die Opportunitätskosten einer sofortigen und irreversiblen Entwicklung explizit berücksichtigt. Wir zeigen, dass Realoptionswerte konzeptionell als auch praktisch besser sind als andere Bewertungen nach dem Ertragswertansatz. Rein auf Bodenwerten beruhende Vergleichswerte sind im Mittel ebenfalls schlechter. Lediglich durch Gutachter angepasste Vergleichswerte sind Realoptionswerten im Mittel immer überlegen. Aber auch in Bezug auf die Gutachterwerte tragen Realoptionswerte noch Informationen bei, die zu insgesamt besseren Bewertungen führen können. Um den Einfluss von Annahmen und Parameterwerten auf Realoptionswerte abzuschätzen, werden umfangreiche Sensitivitätsanalysen durchgeführt. Insgesamt zeigt die Studie, dass das Realoptionsverfahren als sinnvolle Ergänzung zum gutachterlichen Vergleichswertverfahren angesehen werden kann. Das Realoptionsverfahren sollte deshalb in der Zukunft verstärkt bei der Immobilienbewertung herangezogen werden.

Realoptionen und Immobilienbewertung: Eine Umsetzungsstudie

1 Einleitung

Der Marktwert einer entwicklungsfähigen Immobilie ist der Preis, der bei einer gewöhnlichen Transaktion zu erwarten ist. Eine Bewertung ist eine Schätzung des Marktwertes.[1] Potenzielle Käufer und Verkäufer von entwicklungsfähigen Immobilien benötigen solche Bewertungen für ihre Investitionsentscheidungen und sind insbesondere an den folgenden zwei Aspekten interessiert: einer genauen Schätzung des Marktwertes und einer Diskussion der ökonomischen Faktoren, die den Marktwert beeinflussen.

Nicht alle in der Praxis angewandten Bewertungsverfahren geben Auskünfte zu beiden Aspekten. Die Verfahren lassen sich dabei nach dem Vergleichswert- und dem Ertragswertansatz unterteilen. Bei Verfahren nach dem Vergleichswertansatz werden beobachtete Transaktionspreise vergleichbarer Grundstücke herangezogen und darauf basierend der Marktwert geschätzt. Diese Verfahren behandeln nur den ersten der oben genannten Aspekte und erklären nicht, wodurch der Marktwert beeinflusst wird. Bei Verfahren nach dem Ertragswertansatz werden Annahmen über zukünftige Reinerträge der entwickelten Immobilien und Entwicklungskosten getroffen und auf dieser Grundlage der Marktwert geschätzt. Die Berechnungsformeln basieren auf finanzwirtschaftlichen Modellen und können ökonomisch interpretiert werden. Diese Verfahren sind prinzipiell in der Lage, beide der oben genannten Aspekte zu behandeln.

Könnte man davon ausgehen, dass Verfahren nach dem Ertragswertansatz immer auch bessere Marktwertschätzungen liefern, dann wäre die Wahl für den Anwender einfach. Von einer solchen Eindeutigkeit ist a priori nicht auszugehen. In dieser Studie untersuchen wir deshalb die Güte von Marktwertschätzungen nach unterschiedlichen Verfahren des Vergleichswert- und des Ertragswertansatzes. Hierzu stehen uns detaillierte Informationen und Transaktionspreise für zwölf Grundstücke des Berliner Stromversorgers Bewag zur Verfügung. Zugleich untersuchen wir, inwieweit das in der theoretischen Literatur etablierte Realoptionsverfahren für konkrete Bewertungen angewendet werden kann.

---

[1] Die immobilienwirtschaftliche Terminologie verwendet den Begriff *Verkehrswert* statt *Marktwert* und dieser wird *ermittelt* statt *geschätzt*, siehe § 194 BauGB.

Ausgangspunkt für die Studie war die Erfahrung der Bewag, dass die in der Praxis üblichen Bewertungsverfahren nicht geeignet schienen, die inherente Flexibilität bei der Immobilienentwicklung hinreichend abzubilden. Diese Kritik trifft insbesondere auf das Residualwertverfahren zu, welches das dominierende Verfahren nach dem Ertragswertansatz ist. Nach diesem Verfahren wird der Marktwert der unentwickelten Immobilie mit dem Residualwert geschätzt, was dem Kapitalwert (ohne Kosten für das Grundstück) bei sofortiger Entwicklung entspricht.[2] Das Residualwertverfahren geht von der Prämisse einer sofortigen Entwicklung aus, was dazu führen kann, dass der Residualwert für ein Grundstück null wird (wenn eine sofortige Entwicklung nicht lohnend ist). Dies ist nicht plausibel, da beobachtete Grundstückspreise strikt positiv sind. Die ökonomische Erklärung dafür ist schnell gefunden: auch wenn sich eine sofortige Entwicklung nicht lohnt, so kann eine Entwicklung in der Zukunft profitabel sein. Eigentum am Grundstück gibt das Recht auf diese zukünftigen Entwicklungsmöglichkeiten, und das Grundstück muss deshalb einen strikt positiven Wert haben.

Der Optionscharakter von unentwickelten Grundstücken und der Wert des Wartens wurde erstmals von *Arrow und Fisher* (1974) aufgezeigt, allerdings ohne eine Formel herzuleiten, die zur Bewertung genutzt werden könnte. *Titman* (1985) nutzt ein einfaches Binomialmodel, um eine solche Bewertungsformel für entwicklungsfähige Grundstücke herzuleiten. In der Arbeit von *McDonald*/*Siegel* (1986), die sich generell mit dem Optionscharakter von Investitionen befasst, werden Bewertungsformeln unter der Annahme konkreter stetiger stochastischer Prozesse für Reinerträge und Entwicklungskosten hergeleitet. Diese Bewertungsformeln werden ausführlich im Überblicksartikel von *Pindyck* (1991) und den Monographien von *Dixit/Pindyck* (1994) und *Trigeorgis* (1996) diskutiert.[3] Neben der Option des Wartens sind bei der Immobilienentwicklung oftmals weitere Optionen vorhanden: die Möglichkeit, die Größe des Gebäudes an die Marktlage anzupassen,[4] die Möglichkeit, ein bestehendes Gebäude zu erweitern oder umzunutzen oder die Entwicklung frühzeitig abzubrechen[5]. *Lucius* (2000) gibt einen Überblick zu Optionen bei der Immobilienentwicklung.

---

[2] Der Residualwert ist der Barwert der entwickelten Immobilie minus aller Kosten der Entwicklung (insbesondere Baukosten) ohne Kosten für das Grundstück. Es handelt sich damit um den Kapitalwert des Grundstücks.

[3] *Hommel/Pritsch* (1999a,b) bieten einen Überblick zu Optionseigenschaften von Investitionen in deutscher Sprache.

[4] Titman (1985); *Capozza/Li* (2002)

[5] Williams (1991)

Der umfangreichen Literatur zur Theorie der Realoptionen stehen nur wenige Arbeiten gegenüber, die sich mit der Tauglichkeit des Verfahrens für die Immobilienbewertung befassen. *Paddock* et al. (1988) bewerten Bohrrechte und stellen fest, dass Optionswerte näher an Gebotspreisen liegen als Residualwerte. *Quigg* (1993) zeigt, dass Preise für unentwickelte Grundstücke durch Bewertungen als Realoptionen erklärt werden können und Optionsprämien aufgrund der zeitlichen Flexibilität den Wert von Grundstücken positiv beeinflussen.[6] *Patel* und *Paxson* (2001) zeigen für Londons *Canary Wharf,* dass Optionswerte vergleichbar mit gutachterlichen Bewertungen sind. *Leung* und *Hui* (2002) stellen für das Hongkong Disneyland Projekt fest, dass das Realoptionsverfahren anderen Verfahren überlegen ist. Während die beiden erstgenannten Arbeiten keine gutachterlichen Bewertungen in den Vergleich einbeziehen, werden in den beiden letztgenannten Arbeiten die Bewertungen nicht mit Transaktions- oder Gebotspreisen verglichen. Insgesamt vergleicht somit keine der Arbeiten gleichzeitig Bewertungen nach dem Vergleichswert- und dem Ertragswertansatz mit Transaktionspreisen.

Der Beitrag unserer Studie besteht in einem umfassenden Gütevergleich der üblicherweise in der Immobilienbewertung verwendeten Verfahren. Der Gütevergleich zeigt, dass die gutachterlichen Vergleichswerte im Mittel die besten Preisvorhersagen liefern. Die Realoptionswerte folgen auf dem zweiten Rang und damit vor den Residual- und einfachen Vergleichswerten. Das gute Abschneiden der gutachterlichen Vergleichswerte sollte nicht überraschen, da die Gutachter über Expertenwissen verfügen und einen weit besseren Einblick in das Marktgeschehen haben, als dies durch das von uns "mechanisch" angewendete Realoptionsverfahren erreicht werden könnte. Wir zeigen jedoch auch, dass die Bewertung insgesamt noch verbessert werden können, wenn ein gleichgewichtetes Mittel aus gutachterlichem Vergleichswert und Realoptionswert verwendet wird. Realoptionswerte enthalten damit Informationen, die bei den angepassten Vergleichswerten nicht (oder zumindest nicht hinreichend) berücksichtigt wurden.

Wir diskutieren weiterhin, ob und wie Abweichungen der Realoptionswerte von den Transaktionspreisen erklärt werden können. So könnten Realoptionswerte etwa die Flexibilität bei der Entwicklung einer Immobilie über- beziehungsweise unterschätzen.

---

[6] Bezogen auf den Optionswert betragen die Prämien zwischen 1% bis 30%. Einschränkend muss angemerkt werden, dass *Quigg* die für die Bewertungen benötigten (durchschnittlichen) implizierten Volatilitäten mit Transaktionspreisen schätzt, um diese dann wiederum vorherzusagen.

Dies wird für einzelne Objekte im Detail diskutiert. Für eine allgemeine Unterschätzung der Flexibilität spricht, dass sich die Realoptionswerte im Mittel verbessern, wenn eine höhere einheitliche Volatilität der Reinerträge angesetzt wird. Der Einfluss des Liegenschaftszinses auf Realoptionswerte ist ebenfalls nicht zu vernachlässigen. In Sensitivitätsanalysen wird deshalb geklärt, wie stark der Einfluss von Parametervariationen auf Realoptionswerte ist. Insgesamt zeigen die Sensitivitätsanalysen, dass die Preise mit Realoptionswerten bei plausiblen Parameterkombinationen prinzipiell repliziert werden können. Zudem berechnen wir implizierte Volatilitäten, wie sie oft bei der Bewertung von Finanzoptionen angewendet werden.[7] Der implizierte Volatilitätsparameter berücksichtigt weitere, noch unberücksichtigte Flexibilität bei der Entwicklung. Da uns nur wenige Beobachtungen zur Verfügung stehen, können wir keine umfassende Untersuchung durchführen. Wir glauben jedoch, dass die Nutzung von implizierten Volatilitäten einen viel versprechenden Ansatz für die Immobilienbewertung darstellt.

Unsere Studie zeigt, dass das Realoptionsverfahren nicht nur ökonomisch plausibel ist, sondern in der Praxis auch angewendet werden kann. Zwar schneiden gutachterliche Vergleichswerte im Mittel besser ab, der Unterschied in der Güte der Verfahren ist aber gering genug, um sich intensiv mit dem Realoptionsverfahren zu befassen. Dies gilt umso mehr, da Realoptionswerte offenbar Informationen enthalten, die von Gutachtern nicht gebührend berücksichtigt wurden. Das Realoptionsverfahren ist deshalb als sinnvolle Ergänzung zum gutacherlichen Vergleichswertverfahren zu sehen, welches Gutachtern das Instrumentatrium an die Hand gibt, explizit über die ökonomischen Werttreiber sprechen zu können.

Die Arbeit ist wie folgt strukturiert: *Abschnitt 2* stellt die Bewertungsverfahren und Berechnungsformeln vor und diskutiert den Zusammenhang zwischen Residual und Realoptionswerten. *Abschnitt 3* präsentiert die Ergebnisse der empirischen Untersuchung und der Sensitivitätsanalysen. *Abschnitt 4* fasst die Ergebnisse zusammen und gibt einen Ausblick.

---

[7] *Dumas et al.* (1998)

2 Bewertungsverfahren

2.1 Aufgaben

Bewertungsverfahren für entwicklungsfähige Immobilien werden für zwei Aufgaben eingesetzt:

1. Schätzung des Marktwertes des Grundstücks

2. Bestimmung der Entwicklungsmöglichkeiten und des optimalen Entwicklungszeitpunktes.

Verfahren nach dem Vergleichswertansatz erfüllen nur die erste Aufgabe, während Verfahren nach dem Ertragswertansatz prinzipiell beide Aufgaben erfüllen können.

2.2 Vergleichswertansatz

Bewertungsverfahren nach dem Vergleichswertansatz schätzen den Marktwert der entwicklungsfähigen Immobilie mit Hilfe beobachteter Transaktionspreise vergleichbarer entwicklungsfähiger Immobilien. Diese Verfahren basieren auf dem Kalkül, dass kein Investor mehr für ein Grundstück zahlen wird, als für vergleichbare Grundstücke in der Vergangenheit gezahlt wurde. Wir konzentrieren uns in dieser Studie auf zwei in der Praxis häufige angewandte Verfahren.

Einfaches Vergleichswertverfahren: Der einfache Vergleichswert L_E entspricht der Fläche des zu bewertenden Grundstücks multipliziert mit dem Bodenwert. Bodenwerte können auf Bodenrichtwerten basieren, welche von örtlichen Gutachterausschüssen für Grundstückswerte anhand vergangener Grundstückstransaktionen berechnet und jährlich veröffentlicht werden (§ 196 BauGB). Auf Grund der allgemeinen Marktentwicklung seit Veröffentlichung der Bodenrichtwerte müssen eventuell marktweite Anpassungen vorgenommen werden, um zum Bodenwert zu gelangen.

Angepasstes Vergleichswertverfahren: Der angepasste Vergleichswert L_A entspricht dem einfachen Vergleichswert mit Zu- beziehungsweise Abschlägen auf Grund spezieller Eigenheiten des zu bewertenden Grundstücks, etwa einer vorhandenen Altbebauung.

Vergleichswerte nach der Wertermittlungsverordnung werden mit einem solchen mehrstufigen Verfahren geschätzt (§ 14 WertV).

2.3 Ertragswertansatz

Bewertungsverfahren nach dem Ertragswertansatz schätzen den Marktwert der entwicklungsfähigen Immobilie anhand der zukünftig zu erwartenden Reinerträge und berücksichtigen die Kosten der Entwicklung. Zudem werden die Reinerträge bei sofortiger Entwicklung in Bezug gesetzt zu kritischen Reinerträgen. Die kritischen Reinerträge sind als Wert definiert, bei welchem der Investor im aktuellen Zeitpunkt indifferent zwischen Abwarten und Entwickeln ist. Liegen die tatsächlichen Erträge darüber, sollte mit der Entwicklung unverzüglich begonnen werden. Anderenfalls wäre Abwarten die optimale Alternative.

Modellannahmen: Die von uns angewendeten Verfahren nach dem Ertragswertansatz folgen dem Modellrahmen von *Dixit/Pindyck* (1994, Kapitel 6). Es wird angenommen, dass es keine Flexibilität hinsichtlich der Art der zu entwickelnden Immobilie gibt.[8] Die Reinerträge *D(t)* der entwickelten Immobilie in Zeitpunkt *t* sind Mieteinnahmen minus Bewirtschaftungskosten und folgen der geometrischen Brownschen Bewegung

$$\frac{dD(t)}{D(t)} = g\,dt + \sigma\,dW\,. \tag{1}$$

g ist die erwartete Wachstumsrate der Reinerträge, σ die Volatilität der Wachstumsrate und *dW* das Inkrement eines Standard-Wiener-Prozesses. (1) lässt sich wie folgt plausibilisieren: In gewerblichen Mietverhältnissen wird oftmals vereinbart, dass die Miethöhe angepasst werden kann, wenn sich der Lebenshaltungskostenindex ändert (Wertsicherungsklausel). Die Mieten wachsen somit im Mittel mit der allgemeinen Inflationsrate. Es wird weiterhin angenommen, dass dies auch für die Bewirtschaftungskosten gilt. Dann wachsen auch die Reinerträge im Mittel mit dieser Rate.

---

[8] Dies ist mit unseren Daten weitgehend vereinbar und wird in *Abschnitt 3* ausführlicher diskutiert.

Mit den stochastischen Schocks des Wiener-Prozesses wird berücksichtigt, dass nicht jede Mieterhöhung durchgesetzt werden kann, dass einzelne Mieter ausfallen können und Neuvermietungen nicht immer garantiert sind.

Zusätzlich zu den Reinerträgen kann der Eigentümer der entwickelten Immobilie von Wertsteigerungen profitieren. Diese Wertsteigerungen hängen von den zukünftigen Reinerträgen ab, die unsicher sind. Mit $V(t)$ als dem Wert der entwickelten Immobilie und dem bedingten Erwartungswertoperator $\mathcal{E}_t[\cdot]$ ist die erwartete Wertsteigerung $\mathcal{E}_t[dV(t)]/dt.$ Im Marktgleichgewicht entspricht die geforderte Rendite r der Immobilieninvestition (diese wird weiter unten ausführlicher diskutiert) der erwarteten Rendite

$$r = \frac{D(t) + \mathcal{E}_t\left[dV(t)\right]/dt}{V(t)} \; . \tag{2}$$

Aufgrund der Annahmen über den Reinertragsprozess entspricht die erwartete Wertsteigerung $gV(t)$ und es folgt aus (2) mit $\theta \stackrel{\text{def}}{=} r - g$ [9]

$$V(t) \;=\; \frac{D(t)}{\theta}$$

Dies entspricht der Summe der diskontierten erwarteten zukünftigen Reinerträge

$$\int_t^\infty \mathcal{E}_t[D(\tau)]e^{-r(\tau-t)}d\tau$$

und wir nennen V (t) im Folgenden Ertragswert.[10]

Schließlich bezeichnen I die Kosten, die bei der Entwicklung des unentwickelten Grundstücks entstehen. Die Entwicklungskosten setzen sich aus Kosten für Planung, Grundstücksaufbereitung und Bau zusammen. Die Kosten für das Grundstück sind darin nicht enthalten.

---

[9] *Dixit/Pindyck*, (1994), Kap. 6.1, S. 181

[10] Für die Brownsche Bewegung (1) gilt $\mathcal{E}_t[D(\tau)] = D(t)\exp\{g(\tau - t)\},\ \tau \geq t.$ In immobilienwirtschaftlicher Terminologie ist θ der Liegenschaftszins und es wird $\theta > 0$ angenommen.

Residualwertverfahren: Das unentwickelte Grundstück wird unter der Prämisse bewertet, dass es nur in der aktuellen Periode und nicht später entwickelt werden kann. Der Residualwert

$$L_R(t) = \max\left[\frac{D(t)}{\theta} - I, 0\right] \tag{3}$$

ist der maximale Betrag, den ein potentieller Investor unter dieser Prämisse für das unentwickelte Grundstück zu zahlen bereit wäre. Der Residualwert ist nur dann strikt größer als 0, wenn der aktuelle Ertragswert größer ist als die Entwicklungskosten. Nur in diesem Fall ist eine sofortige Entwicklung lohnend. Ist der Ertragswert kleiner als die Entwicklungskosten, so unterbleibt eine Entwicklung. In diesem Fall ist der Wert des Grundstücks 0.

Das Residualwertverfahren erfüllt beide der in Unterabschnitt 2.1 genannten Aufgaben. Es ermittelt den Marktwert des Grundstücks und gibt an, dass nur dann *sofort* entwickelt werden soll, wenn die gegenwärtigen Reinerträge *D(t)* mindestens so groß sind wie die kritischen Reinerträge *θ I*. In diesem Fall ist der Ertragswert mindestens so groß wie die Entwicklungskosten.[11] *Abbildung 1* zeigt den Residualwert L_R als Funktion des Reinertrages. Das Grundstück hat einen strikt positiven Residualwert, wenn die Reinerträge größer als *D´* sind, ansonsten ist eine sofortige Entwicklung nicht lohnend und der Residualwert ist 0.

[Abbildung 1 ungefähr hier.]

Realoptionsverfahren: Der Realoptionswert berücksichtigt eine möglicherweise bestehende Flexibilität bei der Wahl des Entwicklungszeitpunktes. Wir gehen dabei davon aus, dass grundsätzlich eine Baugenehmigung in der Zukunft erteilt werden wird.[12]

---

[11] Wird die dynamische interne Rendite

$$r_d(t) \stackrel{\text{def}}{=} \frac{D(t)}{I} + g$$

zur Diskontierung der Reinerträge verwendet, so entspricht der Ertragswert in *t* gerade den Entwicklungskosten *I*. $L_R(t) > 0$ ist damit äquivalent zum Renditekriterium $r_d(t) > r$.

[12] Es bestehen zwei Unsicherheiten: (i) die Genehmigung wird nicht rechtzeitig erteilt, (ii) sie wird gar nicht erteilt. Da Genehmigungen je nach Bundesland 3 bis 4 Jahre gültig sind (und oftmals verlängert werden können), lässt sich (i) kontrollieren. Solange die Wahrscheinlichkeit für eine Genehmigung zeitpunktunabhängig ist, lässt (ii) den optimalen Entwicklungszeitpunkt unbeeinflusst. Dies ist für unseren Anwendungsfall plausibel. Eine Abhängigkeit könnte auftreten, wenn die Größe der zu entwickelnden Immobilie von *t* abhängt, wie in *Titman* (1985) oder *Capozza/Li* (2002) und Genehmigungen etwa für größere Projekte schwerer zu erhalten sind.

Der Realoptionswert des entwicklungsfähigen Grundstücks ist [13]

$$
L_O(t) = \begin{cases} \left(\dfrac{D^*}{\theta} - I\right)\left\{\dfrac{D(t)}{D^*}\right\}^{\alpha} & , \text{ wenn } D(t) < D^*, \\[2em] \dfrac{D(t)}{\theta} - I & , \text{ wenn } D(t) \geqslant D^* \end{cases} \tag{4}
$$

mit

$$
\begin{aligned}
D^* &= \frac{\alpha}{\alpha - 1}\theta I \\
\alpha &= \frac{1}{\sigma^2}\left[0.5\sigma^2 - (i - \theta) + \sqrt{\{0.5\sigma^2 - (i - \theta)\}^2 + 2i\sigma^2}\right],
\end{aligned}
$$

wobei *i* der Zinssatz für risikolose Anlagen ist. *Abbildung 1* zeigt den Realoptionswert L_O in Abhängigkeit vom Reinertrag. Bei $D \leq D´$ ist der Residualwert 0. Aufgrund der Chance zukünftig steigender Reinerträge ist der Realoptionswert strikt positiv. Bleiben die Reinerträge auch in Zukunft unterhalb $D´$, bleibt das Grundstück unentwickelt und sein Wert unverändert 0, steigen sie jedoch über $D´$, lohnt sich eine Entwicklung. Zugleich bedeutet dies, dass ein positiver Residualwert nicht hinreichend für die Optimalität einer sofortigen Entwicklung ist. Durch die Entwicklung verliert der Eigentümer die Möglichkeit, noch höhere Reinerträge abzupassen. Es sollte erst dann entwickelt werden, wenn die Vorteile weiteren Wartens mindestens aufgehoben werden. Dies ist bei $D \geq D^*$ der Fall. Zwar besteht hier ebenfalls die Chance, dass die Reinerträge in Zukunft noch weiter steigen. Während dieser zusätzlichen Wartezeit würde die entwickelte Immobilie allerdings bereits Reinerträge generieren, während dies für die unentwickelte nicht gilt. Diese entgangenen Reinerträge sind die Opportunitätskosten weiteren Wartens.

2.4 Diskussion der Annahmen

Realoptionswerte werden in der Praxis noch selten berechnet. Nach *Lander* und *Pinches* (1998) sind hierfür die folgenden zwei Aspekte verantwortlich:

1. Realoptionsmodelle sind sehr komplex, und es fehlt an Verständnis und Erfahrung, um damit umzugehen;

---

[13] *Dixit/Pindyck* (1994) S. 183f

2. die den Modellen zu Grunde liegenden Annahmen werden bei praktischen Anwendungen oftmals als zu einschränkend empfunden

Zum ersten Aspekt ist zu nennen, dass das Realoptionsverfahren sicherlich komplexer ist als das Residualwertverfahren. Allerdings berücksichtigt der Realoptionswert im Gegensatz zum Residualwert den Wert von Flexibilität bei der Entwicklung. Die Komplexität des Realoptionsverfahrens ist damit Folge der Komplexität in wirklichen Entscheidungssituationen und nicht Selbstzweck.

Zum zweiten Aspekt sind zuerst die grundsätzlichen Annahmen zu nennen, die zur Herleitung der Bewertungsformeln (3) und (4) nach dem Ertragswertansatz getroffen wurden. Es wird erstens angenommen, dass Investoren friktionslos ein Portfolio aus gehandelten Wertpapieren zusammenstellen können, dessen Zahlungsströme proportional zu den Reinerträgen der entwickelten Immobilie sind. Der Kapitalmarkt ist damit vollständig genug, um die Zahlungsströme aus der entwickelten Immobilie zu replizieren.[14] Zweitens wird angenommen, dass potentielle Käufer der entwickelten Immobilie wohldiversifiziert sind und deshalb die gleiche Rendite r für die entwickelte Immobilie fordern.[15] Diese grundsätzlichen Annahmen bedeuten, dass die zu bewertende Immobilie Zahlungsströme generiert, die sich nicht grundsätzlich von denen anderer Anlageobjekte unterscheiden. Diese Zahlungsströme werden von unterschiedlichen potenziellen Investoren gleich bewertet. Sowohl die Annahme der Marginalität der zu bewertenden Immobilie als auch die Annahme homogener Investoren sind den Verfahren nach dem Vergleichwert- und dem Ertragswertansatz gemeinsam. Wären Grundstücke und die Zahlungsströme, die mit ihnen generiert werden können, verschieden, dann könnten beobachtete Grundstückspreise nicht zum Vergleich herangezogen werden. Wären Immobilieninvestoren nicht homogen, dann würden Preise, die andere Immobilieninvestoren in der Vergangenheit gezahlt haben, nicht bei der Einschätzung helfen, wieviel Investoren für das zu bewertende Grundstück zu zahlen bereit sind.

---

[14] *Williams* (1991), S. 191, betont, dass dies eine Annahme über den Kapitalmarkt ist und nicht über den Immobilienmarkt.

[15] Die Bewertungsgleichungen lassen sich grundsätzlich auch mit einer investorenspezifischen geforderten Rendite herleiten (*Dixit/Pindyck*, 1994, Kap. 5.2). Es kann aber nur dann sinnvoll vom Marktwert der entwicklungsfähigen Immobilie gesprochen werden, wenn die geforderte Rendite nicht vom potenziellen Käufer abhängt.

Ferner sind im Zusammenhang mit dem zweiten Aspekt die konkreten Annahmen zu nennen, die bei der Herleitung der Bewertungsformeln getroffen wurden. Man kann es als einschränkend empfinden, dass für die Reinerträge eine Brownsche Bewegung mit konstanter Wachstumsrate und Volatilität angenommen wird. Dies kann unrealistisch sein, wenn langfristige Mietverträge mit fixem Mietzins oder mit festen Erhöhungsklauseln vorliegen. Sollten solche Angaben für eine Bewertung vorliegen, so können sie grundsätzlich berücksichtigt werden. Die Modellierung mit einer Brownschen Bewegungen scheint jedoch gerechtfertigt, wenn nicht sicher ist, wie die Mietverträge der entwickelten Immobilie ausgestalten sein werden. Ferner könnte man gegen die Optionsformel (4) einwenden, dass konstante Entwicklungskosten angenommen werden. Dies ist für unsere Daten allerdings plausibel, da die Entwicklungskosten in Berlin kaum gewachsen sind und ihre Wachstumsrate eine im Vergleich zur Wachstumsrate der Reinerträge vernachlässigbar geringe Volatilität aufweist.[16] Grundsätzlich können stochastische Entwicklungskosten in Realoptionsverfahren berücksichtigt werden[17]. Die resultierende Bewertungsformel hat große Ähnlichkeit mit (4), hängt jedoch von zusätzlichen Eingangsparametern ab. Diese Parameter müssen geschätzt werden und sind mit Unsicherheit behaftet. Aufgrund dieser zusätzlichen Parameterunsicherheit muss ein realistischeres Modell nicht unbedingt zu besseren Bewertungen führen.

Zusammenfassend ist festzustellen, dass die Annahmen des Realoptionsverfahrens nur dann als einschränkend anzusehen sind, wenn sie zu unrealistischen Preisvorhersagen führen. Dies kann nur durch den Vergleich von Bewertungen nach verschiedenen Bewertungsverfahren mit den tatsächlich erzielten Transaktionspreisen geklärt werden. Dies geschieht im folgenden empirischen Teil dieser Studie. Dabei wird ausführlich die Sensitivität der Realoptionswerte im Hinblick auf die angesetzten Werte für den Volatilitäts- und den Liegenschaftszins diskutiert.

---

[16] Die durchschnittliche jährliche Wachstumsrate der Baukostenindexe für Gewerbe- und Bürogebäude des Statistischen Landesamtes Berlin seit 1992 ist 0,4% und die Volatilität 2,0%.

[17] *Williams* (1991); *Quigg,* (1993)

3 Empirische Untersuchung

3.1 Objektdaten

Unser Datensatz umfasst Angaben zu zwölf Berliner Immobilien aus dem ehemaligen Anlagevermögen der Bewag Aktiengesellschaft & Co. KG, die zwischen Dezember 1998 und Juni 2002 unentwickelt verkauft wurden.[18] Die Immobilien waren allesamt nicht mehr für die Stromversorgung oder für die Verwaltung der Bewag betriebsnotwendig. *Tabelle 1* gibt einen Überblick zur Lage und die ursprüngliche Nutzung der zwölf Immobilien.

[Tabelle 1 ungefähr hier.]

Bei den Immobilien handelt es sich um architektonisch sehr attraktive Umspannwerke aus den zwanziger Jahren, eines der ersten Berliner Hochhäuser aus den dreißiger Jahren, ein Gästehaus an der Spree sowie mehrere Objekte ohne wertvolle Bausubstanz.[19] Die Grundstücke bilden einen großen Teil der Variation innerhalb des Berliner Grundstücks-marktes ab, da die Standorte von Zentrums- und Innenstadtlagen bis zur Stadtrandlage reichen. Alle Objekte können erst nach einer Entwicklung genutzt werden, wobei viele der Objekte unter Denkmalschutz stehen und für sie nur eine Nutzung für Gewerbe- und Wohnzwecke in Frage kommt.

Für die Verhandlungen mit potenziellen Käufern wurde von der Bewag für jedes der zwölf Grundstücke eine Entwicklungsperspektive ermittelt und darauf basierend Reinerträge und Entwicklungskosten geschätzt. Zudem wurden für die Grundstücke Sachverständigengutachten in Auftrag gegeben, die Auskunft über Grundstücksgröße, Lage, vorhandene Bebauung und deren Zustand sowie geschätzte Boden- und Marktwerte erteilen. Die Daten stehen einheitlich zum Stichtag Dezember 1997 zur Verfügung. *Tabelle 2* gibt beschreibende Statistiken zu den Immobiliendaten.

[Tabelle 2 ungefähr hier.]

---

[18] Im Januar 2002 wurde aus den Unternehmen Bewag, HEW, VEAG und Laubag der Konzernverbund Vattenfall Europe gebildet.
[19] Zu detaillierten Informationen zu einigen der Immobilien siehe Bewag Aktiengesellschaft (2000).

Teil A enthält Angaben zu den unentwickelten Immobilien und deren möglicher Entwicklung. Die Kosten entstehen bei der baulichen Entwicklung der Immobilien. Nach einer Entwicklung bleibt die Bruttogrundfläche (BGF) bei drei Immobilien unverändert und Erhöht sich bei den restlichen neun. Die objektspezifischen Liegenschaftszinsen variieren, womit aus $\theta = r - g$ und einheitlichem r folgt, dass die Wachstumsraten der Reinerträge g für die einzelnen Objekte von den Gutachtern unterschiedlich eingeschätzt wurden. Setzen wir für die geforderte Rendite 7,6% an, so ist die durchschnittlich erwartete implizite Wachstumsrate 1,2%.[20] Die beiden Objekte mit den höchsten objektspezifischen Liegenschaftszinsen haben den größten Neubaubedarf bei einer Entwicklung (100% bzw. 63,9%). Die objektspezifischen Liegenschaftszinsen für die restlichen Objekte (Neubau höchstens 50%) sind dagegen niedriger und deren Reintragswachstum wird folglich höher eingeschätzt. Vernachlässigen wir die beiden Neubauobjekte, so ist der durchschnittliche Liegenschaftszins 6,1% und die durchschnittlich erwartete implizite Wachstumsrate 1,5%. Dies ist plausibel und entspricht der durchschnittlichen Inflationsrate des Verbraucherpreisindex für Berlin seit 1992. Die Bodenwerte basieren auf Bodenrichtwerten aus dem Jahr 1996 und wurden teilweise für Wertänderungen seit diesem Zeitpunkt angepasst. Vergleicht man die Statistiken für Bodenricht- und Bodenwerte, so erkennt man, dass die Anpassungen relativ gering sind. Teil B von Tabelle 2 enthält Angaben zu den Transaktionen der zwölf Immobilien. Zu Vergleichszwecken werden auch Statistiken für diskontierte Verkaufspreise gegeben. Diese wurden mit der geforderten Rendite r auf den Stichtag Dezember 1997 diskontiert.

3.2 Schätzung der Marktwerte

Der einfache Vergleichswert L_E wird als das Produkt von Bodenwert und Grundstücks-fläche berechnet. Der angepasste Vergleichswert L_A wurde von Gutachtern berechnet. L_A basiert auf L_E, es wurden aber bei zehn Objekten Anpassungen für besondere Bedingungen der Grundstücke – insbesondere Abbruchkosten für bestehende Bebauung – vorgenommen. Bei den beiden Objekten ohne Anpassung handelt es sich um ein unbebautes Grundstück und ein Grundstück, bei welchem die BGF vor und nach Entwicklung unverändert bleibt.

---

[20] Der Schätzwert für *r* entspricht der durchschnittlichen DIMAX-Rendite. Der DIMAX ist ein Performanceindex für ein Portfolio von Aktiengesellschaften, die ihren Schwerpunkt (mindestens 75%) im Immobiliengeschäft haben. Der Index wird vom *Bankhaus Ellwanger & Geiger* berechnet.

Der Residualwert L_R wird von uns mit *Formel (3)* mit Hilfe der Reinerträge, Nutzfläche, Liegenschaftszins und Entwicklungskosten berechnet.

Um die Realoptionswerte nach (4) zu berechnen, müssen die Parameter *i* und *σ* festgelegt werden. Als Zinssatz *i* für eine risikolose Anlage nutzen wir den Zins einer hypothetischen Null-Kupon-Anleihe mit Restlaufzeit von 2 Jahren (Deutsche Bundesbank, 1997, Reihe WZ9810). Die durchschnittliche Zeit der Entwicklung wird für die meisten Objekte mit rund zwei Jahren veranschlagt. Der hypothetische Zins betrug im Dezember 1997 4,2%. Die Volatilität *σ* der Wachstumsrate der Reinerträge wird getrennt für Büro- und Wohnimmobilien mit der Standardabweichung von Mietwachstumsraten geschätzt. Wir nehmen damit implizit an, dass Reinerträge und Mieten proportional sind. Für Büroimmobilien werden Mietangaben genutzt, die für unterschiedliche Berliner Lagen von *Aengevelt* für die Jahre 1991 bis 2001 erhoben wurden. Es handelt sich um Neuvertragsmieten, wobei spezielle Anreize wie Staffelmieten oder Indexklauseln berücksichtigt wurden. Bei Mieten für Wohnimmobilien wurde auf den RDM-Mietspiegel zurückgegriffen und die Angaben für fertiggestellte Objekte im Berichtsjahr mit mittlerem Wohnwert genutzt. Die geschätzten Volatilitäten für Büromieten liegen zwischen 10,3% und 16,5%, wobei die Volatilität für Zentrumslagen am höchsten ist. Die Volatilität für Wohnmieten ist mit 8,1% geringer.[21] Da die Wahl einer homogenen Volatilität die Bewertung vereinfacht, setzen wir für die Volatilität *σ* zunächst den höchsten Wert von 16,5% an, der sich für Objekte in Zentrumslage ergibt. Wir diskutieren später, in welcher Weise diese Annahme die Realoptionswerte beeinflusst.

Das Realoptionsverfahren schätzt nicht nur den Marktwert der unentwickelten Immobilie, sondern gibt auch an, ob sofort mit der Entwicklung begonnen werden soll. Es sollte immer dann sofort entwickelt werden, wenn

$$d \stackrel{\text{def}}{=} \frac{D(t)}{D^*} \geqslant 1$$

gilt, dass heißt, wenn die aktuellen Reinerträge *D(t)* nicht kleiner als die kritischen Reinerträge *D** sind. *Tabelle 3* zeigt dieses Kriterium für die zwölf Objekte.

---

[21]Die durchschnittlichen Mietwachstumsraten über den betrachteten Zeitraum sind allesamt negativ. Dies ist sicherlich ein temporäres Phänomen, da für die lange Frist eine strikt positive durchschnittliche Wachstumsrate zu erwarten ist. Wie in Abschnitt 3.1 diskutiert, impliziert der durchschnittliche Liegenschaftszins mit g = r − θ > 0 genau dies.

In der immobilienwirtschaftlichen Praxis sind zudem die Kriterien der statischen und der dynamischen Rendite

$$r_s(t) = \frac{D(t)}{I} \quad \text{und} \quad r_d(t) = \frac{D(t)}{I} + g$$

gebräuchlich. Übersteigt die jeweilige Rendite die vorgegebene kritische Rendite r^*_s beziehungsweise r^*_d, so soll sofort mit der Entwicklung begonnen werden. *Tabelle 3* gibt zusätzlich die aktuellen sowie die kritischen statischen und dynamischen Renditen an. Für die dynamischen Renditen wird die jeweilige erwartete Wachstumsrate g als Differenz aus der geforderten Rendite *r* und dem objektspezifischen Liegenschaftszins θ berechnet. *Tabelle 3* zeigt, dass die Entwicklungsentscheidung selbstverständlich unabhängig von der Wahl des verwendeten Kriteriums ist.[22]

Am Stichtag Dezember 1997 sollten von den insgesamt zwölf Objekten nur die sechs Objekte 6 bis 11 sofort entwickelt werden. Für diese Objekte entspricht der Realoptionswert dem Residualwert und die Prämie weiteren Wartens mit der Entwicklung beträgt Null. Bis zum Jahr 2004 waren fünf dieser Objekte entwickelt und beim sechsten wurde mit der Entwicklung begonnen. Für die restlichen sechs Objekte war es am Stichtag lohnend, mit der Entwicklung zu warten. Für die beiden Objekte 2 und 4 war *d* am Stichtag nahe 1 und die Wahrscheinlichkeit hoch, dass deren Entwicklung in naher Zukunft profitabel sein würde.17 Beide Objekte wurden bis zum Jahr 2004 entwickelt. Bei Objekt 5 hat der Investor nach dem Kauf im Jahre 1999 mit der Entwicklung begonnen, ging bald danach jedoch in Insolvenz. Mittlerweile wurde das Objekt mit einer anderen Nutzung entwickelt. Bei den restlichen drei Objekten wurde noch nicht mit der Entwicklung begonnen.

---

[22] Man sieht sofort, dass die statische und die dynamische Rendite immer zur gleichen Entscheidung führen. Da weiterhin

$$\frac{r_s}{r_s^*} = d$$

gilt, folgt dies auch für das erste Kriterium, welches die aktuellen mit den kritischen Reinerträgen ins Verhältnis setzt.

17 Die erwartete diskontierte Zeit am Stichtag, die es dauert, bis *d = 1* eintritt, ist im Durchschnitt für die Objekte 2 und 4 um rund 90% kleiner als für die anderen vier Objekte. Zur Berechnung der erwarteten diskontierten Zeit siehe *Shackleton und Wojakowski* (2002), S. 552.

3.3 Vergleich der Bewertungen

Die Güte der einzelnen Bewertung wird durch das Bewertungsverhältnis

$$Q_j \stackrel{\text{def}}{=} \frac{L_j}{P} \quad \text{mit} \quad j \in \{E, A, R, O, G\}$$

gemessen. Es folgt, dass $Q - 1$ den relativen Vorhersagefehler in Bezug auf den Preis misst. Um Vergleichbarkeit herzustellen, wurden die Preise mit der geforderten Rendite *r* auf den Bewertungsstichtag Dezember 1997 diskontiert. Der Index *j* wurde bereits in *Abschnitt 2* eingeführt: *E* steht für den einfachen Vergleichswert, *A* für den durch Gutachter angepassten Vergleichswert, *R* für den Residualwert und *O* für den Realoptionswert. Im Gegensatz zu den vier reinen Verfahren benennt G das arithmetische Mittel aus angepasstem Vergleichs- und Realoptionswert, welches später noch ausführlicher diskutiert wird. Es können nur elf Bewertungsverhältnisse berechnet werden, da die Objekte 11 und 12 gemeinsam mit einem weiteren Objekt verkauft wurden. Für das dritte Objekt, dessen Wert von untergeordneter Bedeutung ist, liegt eine gutachterliche Marktwertschätzung vor. Das Bewertungsverhältnis für Objekte 11 und 12 wurde als Verhältnis der Summe der drei Bewertungen zum Gesamtpreis berechnet. Objekt 10 wurde zusammen mit einem vermieteten Wohnhaus verkauft. Für das Wohnhaus lag eine gutachterliche Marktwerteinschätzung vor und weiterhin alle Angaben, die zur Berechnung des Ertragswertes (3) benötigt werden. Auch für Objekt 10 wurde das Bewertungsverhältnis als Summe der Bewertungen zum Gesamtpreis berechnet. Dabei wurde für die beiden Vergleichswertverfahren die gutachterliche Bewertung des Wohnhauses angesetzt und für die beiden Ertragswertverfahren dessen Ertragswert.

Tabelle 4 zeigt in Teil A die Bewertungsverhältnisse für die verschiedenen Verfahren. Ein Bewertungsverfahren ist umso besser, je näher dessen Bewertungsverhältnisse bei Eins liegen und desto geringer die Verhältnisse um diesen Wert streuen. Ein Verhältnis von Eins bedeutet, dass der Preis exakt vorhergesagt wurde.

[Tabelle 4 ungefähr hier.]

Man erkennt an den Bewertungsverhältnissen, dass keines der vier reinen Bewertungsverfahren die anderen dominiert und immer die beste Preisvorhersage gibt. Teil B von Tabelle 4 gibt auf der Diagonalen an, wie oft ein Verfahren den Preis überschätzt. Die vier reinen Verfahren überschätzen die Preise in etwa der Hälfte aller Transaktionen.

Allerdings werden nicht immer die Preise der gleichen Objekte von den vier reinen Verfahren über- beziehungsweise unterschätzt. Die Werte oberhalb der Diagonalen geben an, in wie vielen Fällen zwei Verfahren den Preis übereinstimmend über- oder unterschätzen. Man sieht, dass die Realoptions- und Residualwerte vollständig übereinstimmen. Die Übereinstimmung ist auch für die beiden Vergleichswertverfahren hoch, während sie zwischen den Verfahren nach den beiden Ansätzen deutlich geringer ist.

Güte der reinen Verfahren: *Tabelle 5* fasst die einzelnen Bewertungsverhältnisse in Kennzahlen zusammen. Die mittleren Bewertungsverhältnisse zeigen, dass der einfache Vergleichswert im Mittel die Preise um 12,5% überschätzt, während die anderen Verfahren die Preise im Durchschnitt unterschätzen. Die Realoptionswerte haben mit -5,5% den geringsten durchschnittlichen Vorhersagefehler. Der Fehler für die angepassten Vergleichswerte beträgt -13,2% und für die Residualwerte sogar -16,0%. Die größte Variation der Vorhersagefehler haben die einfachen Vergleichswerte, gefolgt von den Residual- und den Realoptionswerten. Die geringste Variation haben die angepassten Vergleichwerte.

[Tabelle 5 ungefähr hier.]

Ein gutes Bewertungsverfahren sollte im Mittel keinen Vorhersagefehler machen und die Variation der Fehler sollte gering sein. Zur Gesamtbewertung der Güte eines Verfahrens verwenden wir den mittleren quadratischen Fehler

$$\mathrm{MQF}_j = \{\mathcal{E}[Q_j] - 1\}^2 + \sigma_j^2 \quad \mathrm{mit} \quad j \in \{E, A, R, O, G\} \,. \tag{5}$$

Hierbei steht σ_j für die Standardabweichung der Zufallsvariablen Q_j und $\mathcal{E}[Q_j]$ für deren Erwartungswert. Der MQF berücksichtigt somit den mittleren Vorhersagefehler und die Variation der Fehler. Der MQF nimmt den minimalen Wert Null an, wenn ein Verfahren die Preise exakt vorhersagt. Je höher der MQF wird, desto schlechter ist ein Verfahren. Die letzte Spalte in *Tabelle 5* zeigt, dass von den reinen Bewertungsverfahren die angepassten Vergleichswerte am besten abschneiden, gefolgt von den Realoptions- und den Residualwerten. Die einfachen Vergleichswerte schneiden am schlechtesten ab.

Um die Robustheit dieser Rangordnung zu überprüfen, wurden die Realoptionswerte zum Vergleich auch für einen einheitlichen Liegenschaftszins und objektspezifischen

Volatilitäten der Mietwachstumsrate berechnet. Der MQF beträgt 46,6%, wenn objektspezifische Volatilitäten und Liegenschaftszinsen angesetzt werden. Er beträgt 25,6%, wenn objektspezifische Volatilitäten und ein einheitlicher Liegenschaftszins angesetzt werden und beträgt 26,7% bei einer einheitlichen Volatilität und einem einheitlichen Liegenschaftszins.[23] Weiterhin wurde die einheitliche Volatilität σ bestimmt, welche den MQF der Realoptionswerte minimiert. Die MQF-minimierende Volatilität mit objektspezifisch angesetzten Liegenschaftszinsen beträgt 20,7% und der resultierende MQF ist 30,0%. Die MQF-minimierende Volatilität mit einheitlich angesetztem Liegenschaftszins beträgt 13,2% und der resultierende MQF ist 22,6%. Beide Volatilitäten sind mit den von uns mit Hilfe der Mietdaten berechneten Volatilitäten vereinbar. Die Wahl von σ und θ beeinflusst damit erwartungsgemäß den MQF der Residual- und Realoptionswerte. Die Rangordnung der Verfahren bleibt davon jedoch unberührt. Die Robustheitsanalyse bestätigt damit die Rangordnung der Bewertungsverfahren. Zugleich unterstreicht die Analyse unsere Vermutung, dass Detailtreue nicht unbedingt zu besseren Bewertungen führen muss. Obwohl ein einheitlicher Liegenschaftszins und ein einheitliche Volatilität aus theoretischer Sicht eine Vereinfachung sind, müssen sie Bewertungen nicht verschlechtern, sondern können sie sogar verbessern.[24]

Zusammenfassend können wir feststellen, dass Vergleichswerte die besten Bewertungen liefern, dies aber nur, wenn die Vergleichswerte von Gutachtern ermittelt wurden. Einfache (unangepasste) Vergleichswerte liefern eindeutig die schlechtesten Bewertungen. Die zweitbesten Bewertungen erzielt das Realoptionsverfahren, welches das Residualwertverfahren größtenteils dominiert. Insgesamt sollte die gefundene Rangordnung nicht überraschen. Sie zeigt, dass die richtige Einschätzung des Wertes von Flexibilität entscheidend sein kann. Gutachter passen die zugrunde liegenden einfachen Vergleichswerte mit Hilfe ihres Expertenwissens für Entwicklungsmöglichkeiten an.

---

[23] Als einheitlicher Liegenschaftszins wurde der Durchschnittswert von 6,4% angesetzt, als einheitliche Volatilität wurde 16,5% angesetzt. Für Immobilien, die als Büro- oder Gewerbeobjekt entwickelt werden sollen, wurde die relevante lagespezifische Volatilität für die Wachstumsrate der Büromieten als objektspezifisch angesetzt. Für Immobilien, die als Wohnobjekt entwickelt werden sollen, wurde die entsprechende Volatilität von Wohnungsmieten angesetzt. Der MQF für Residualwerte verbessert sich bei einen einheitlich angesetzten Liegenschaftszins ebenfalls auf 43,6%, ist damit aber immer noch weit höher als der jeweils relevante MQF für die Realoptionswerte.

[24] Die Parameter σ und θ gehen in nicht-linearer Weise in die Bewertungsformeln (3) und insbesondere (4) ein. Der Effekt von ungenau geschätzten Parametern auf die Bewertungen ist deshalb komplex. Ein genau geschätzter Parameterdurchschnittswert muss deshalb nicht unbedingt schlechter sein als ungenau geschätzte objektspezifische Parameter.

Allerdings heißt dies nicht, dass die angepassten Vergleichswerte selbst nicht verbesserungswürdig sind. Die Bewertungsverhältnisse der Vergleichswerte in *Tabelle 4* zeigen, dass gutachterliche Anpassungen den Vorhersagefehler zwar für fünf Objekte reduzieren, ihn zugleich aber bei vier Objekten leicht erhöhen.

Güte der gemittelten Bewertungen: In der Praxis wird oftmals nicht nach dem einzigen und besten reinen Bewertungsverfahren verlangt. Auch Verfahren, die in einem Einzelvergleich unterlegen sind, können für Bewertungen hilfreiche Informationen beitragen. Dies ist vereinbar mit dem Ergebnis der statistischen Vorhersagetheorie, dass gemittelte Vorhersagen unter Umständen besser sein können als die Vorhersagen nach einzelnen Verfahren. Da keines der vier reinen Verfahren die anderen eindeutig dominiert, haben wir deshalb auch paarweise gleichgewichtete gemittelte Bewertungen berechnet.[25] Es bestehen vier Kombinationsmöglichkeiten für die jeweils zwei Verfahren nach dem Vergleichswert- und dem Ertragswertansatz. Kombinationen von Bewertungen nach dem gleichen Ansatz werden nicht berücksichtigt. Die gewichteten Bewertungen von einfachen Vergleichswerten und Residualwerten beziehungsweise von Realoptionswerten führen jeweils zu einem geringeren MQF als dies für die reinen Verfahren der Fall ist. Allerdings ist keiner der MQF geringer als 10,1%, was dem MQF der angepassten Vergleichswerte entspricht. Dies gilt ebenfalls für die gemittelten Bewertungen aus Residual- und angepasstem Vergleichswert. Die Kombination aus angepasstem Vergleichswert und Realoptionswert führt jedoch zu einer deutlichen Verbesserung des MQF, der sich um mehr als ein Viertel auf 7,5% reduziert. *Tabelle 4* gibt in Teil A die Bewertungsverhältnisse für die Kombination aus Realoptions- und angepassten Vergleichswerten und *Tabelle 5* gibt die Kennziffern für die Bewertungsgüte. Die Realoptionswerte enthalten demnach Informationen, die von den Gutachtern nicht hinreichend in ihren Vergleichswerten berücksichtigt wurden.

---

[25] §7 Wert V sagt zur Gewichtung: "Sind mehrere Verfahren herangezogen worden, ist der Verkehrswert aus den Ergebnissen der angewandten Verfahren unter Würdigung ihrer Aussagefähigkeit zu bemessen". Dies bedeutet nicht, dass die einzelnen Bewertungen mit gleichem Gewicht eingehen müssen. Auch die statistische Theorie zur optimalen Kombination von Vorhersagen schreibt dies nicht vor. Vielmehr sollten die Gewichte derart gewählt werden, dass etwa der MQF minimiert wird. Die Gewichte werden dabei für einen Teil von Vorhersagen und Preisen bestimmt und dann für weitere Vorhersagen benutzt. Aufgrund der geringen Anzahl an Beobachtungen können wir ein solches zweistufiges Verfahren nicht anwenden und beschränken uns deshalb auf das arithmetische Mittel.

Alternative Bewertungsverhältnisse: Statt der von uns verwendeten Bewertungsverhältnisse *Q* hätte man auch deren Reziprokwert *1/Q = P/L* zur Beurteilung der Bewertungsgüte verwenden können. Da es sich um eine nicht-lineare Transformation von *Q* handelt, kann dies zu einer anderen Rangfolge der Verfahren führen. Wir haben uns aus zwei Gründen auf die Bewertungsverhältnisse *Q* konzentriert: Einerseits können die Reziprokwerte für zwei Residualwerte nicht berechnet werden, und wir verlieren zwei Beobachtungen zur Beurteilung dieses Verfahrens. Andererseits misst *1/Q − 1* die relative Abweichung des Preises von der Bewertung und macht damit die Bewertung zum Bezugspunkt. Demgegenüber misst *Q − 1* den Vorhersagefehler in Bezug auf den Preis, was wir für ein weit sinnvolleres Maß halten. Nutzt man die Reziprokwerte zur Beurteilung der Verfahren, so sind die gutachterlichen Vergleichswerte auch hier im Mittel am besten, die Realoptionswerte sind jedoch schlechter als die einfachen Vergleichswerte. Diese Rangordnung ist wiederum robust. Die zwei wichtigsten Ergebnisse unserer Analyse der Bewertungsverhältnisse erhalten wir aber auch in diesem Fall: Das Realoptionsverfahren ist das bessere Verfahren nach dem Ertragswertansatz und der MQF gemittelter angepasster Vergleichs- und Realoptionswerte ist um mehr als ein Viertel kleiner als der MQF der angepassten Vergleichswerte.

3.4 Detailbetrachtung der Realoptionswerte

Die von uns berechneten Realoptionswerte beruhen auf den durch die Bewag angenommen Entwicklungsmöglichkeiten der einzelnen Grundstücke. Für einige der Objekte haben wir weitere Informationen, die über die in den *Tabellen 1 und 2* gegebenen Informationen hinausgehen. Hierzu zählen insbesondere Informationen über den aktuellen Entwicklungsstand der Immobilien. In diesem Abschnitt wird deshalb untersucht, inwieweit diese zusätzlichen Informationen Vorhersagefehler der Realoptionswerte erklären können. Besondere Aufmerksamkeit gilt Objekteigenschaften, die eine höhere oder geringere Flexibilität bei der Entwicklung vermuten lassen. Weiterhin diskutieren wir den Einfluss, den die Parameter der Volatilität und des Liegenschaftszinses auf die Realoptionswerte haben. Wir haben bereits weiter oben gesehen, dass Änderungen dieser Parameter großen Einfluss auf die Realoptionswerte haben können. Solche Sensitivitätsanalysen dürften für Bewertungen in der Praxis von großem Interesse sein, da der Einfluss der Wertreiber auf Realoptionsbewertung explizit quantifiziert werden kann.

Dies ist ein eindeutiger Vorteil der Verfahren nach dem Ertragswertansatz. Sie basieren auf ökonomischen Kalkülen, so dass der Effekt von Parametervariationen qualitativ eindeutig ist, und sie besitzen explizite Bewertungsformeln, so dass die quantitative Wirkung berechnet werden kann.

Diskussion von Objekteigenschaften: Sechs der zwölf entwicklungsfähigen Immobilien unterliegen Restriktionen des Denkmalschutzes, was die Flexibilität bei der Entwicklung einschränken könnte. Im Extremfall besteht nur die Möglichkeit, das Gebäude den historischen Vorgaben getreu zu einem frei gewählten Zeitpunkt zu sanieren. Demgegenüber besteht bei Objekten, die nicht dem Denkmalschutz unterliegen, weit mehr Flexibilität für den Investor, was Größe, Form, Funktionalität und Zweck des zu erstellenden Gebäudes betrifft. Die Bewag hat lediglich bei vier der sechs denkmalgeschützten Objekte geringe Änderungen der BGF nach Entwicklung für möglich erachtet. Bei diesen vier Objekten liegt der Realoptionswert jeweils über dem Preis, im Durchschnitt um 16,2%. Hier wird von uns möglicherweise mehr Flexibilität angenommen, als tatsächlich vorlag. Bei den anderen beiden Objekten liegt der Realoptionswert unterhalb des Preises. Dabei ist insbesondere *Objekt 9* zu nennen, bei welchem die Bewag von einer Sanierung des bestehenden Gebäudes ausging. Die tatsächliche Entwicklung dieses Objektes erfolgte dann aber nicht durch die Sanierung des vorhandenen Gebäudes, sondern durch den Neubau der inneren Struktur unter Erhalt der Fassade. Die BGF des Objektes konnte damit um 54,5% erhöht werden. Dies kann plausibilisieren, warum der Preis um 31,0% über dem Realoptionswert liegt. In diesem speziellen Fall lag demnach mehr Flexibilität vor, als dies bei der Bewertung angenommen wurde.[26] Insgesamt liegt der durchschnittliche Vorhersagefehler für die denkmalgeschützten Objekte bei -6,0%, und es lässt sich nicht pauschal sagen, dass die Realoptionswerte die Entwicklungsflexibilität bei den denkmalgeschützten Objekten überschätzen.

Für die verbleibenden Objekte beträgt der durchschnittliche Vorhersagefehler der Realoptionswerte -5,1%. Die Variation der Fehler ist insgesamt größer als bei den Objekten mit denkmalgeschützten Gebäuden. Bei vier Objekten unterschätzt der Realoptionswert den Preis, bei zweien wird der Preis überschätzt. Dies deutet daraufhin, dass die Flexibilität bei

---

[26] Erhöht man die Reinerträge und die Entwicklungskosten für dieses Objekt pauschal um 54,5%, so entspricht der Realoptionswert weiterhin dem Residualwert, der Bewertungsfehler wechselt jedoch sein Vorzeichen und reduziert sich zu 17,9%.

diesen Objekten nicht hinreichend modelliert wurde. Dies gilt insbesondere für *Objekt 5*, bei welchem der Realoptionswert weit unterhalb des tatsächlich erzielten Preises liegt. Bei diesem Objekt weicht die von der Bewag angesetzte Entwicklungsidee von der tatsächlich durchgeführten Entwicklung ab. Es handelt es sich um ein größeres Grundstück in grüner Lage mit sehr guter Anbindung an die Innenstadt. Die Bewag hatte die Entwicklung als Seniorenheim angesetzt, es wurde letztendlich jedoch eine Bebauung mit hochwertigen Reihenhäusern durchgeführt. Zudem ging, wie bereits erwähnt, der ursprüngliche Investor in Insolvenz. Beides könnte den relativ hohen Vorhersagefehler des Realoptionswertes erklären, der aber immer noch weit kleiner als der Fehler des Residualwertes ist. Einerseits war die von der Bewag projektierte Entwicklung möglicherweise nicht optimal und der damit berechnete Realoptionswert zu niedrig. Andererseits könnte der ursprüngliche Investor dass Potenzial seiner Entwicklungsidee überschätzt und einen zu hohen Preis gezahlt haben.

Schließlich haben wir untersucht, ob der Vorhersagefehler mit der Lage der Objekte zusammenhängt. Fünf der Objekte befinden sich im Osten der Stadt und sieben im Westen. Man könnte spekulieren, dass die Flexibilität bei der Entwicklung im Osten zum Stichtag weit größer war als im Westen, da die komplette Infrastruktur modernisiert werden musste. Die von uns berechneten Realoptionswerte berücksichtigen dies nicht, und der Vorhersagefehler für Objekte im östlichen Teil sollte dann deutlich negativ sein. Wir können solche lagespezifischen Einflüsse jedoch nicht entdecken, da der durchschnittliche Vorhersagefehler für die fünf Objekte im östlichen Teil -6,5% beträgt und für die Objekte im westlichen Teil -4,7%.

Sensitivität in Bezug auf den Liegenschaftszins: Um zu untersuchen, wie stark die Realoptionswerte auf Änderungen des Liegenschaftszinses reagieren, berechnen wir die Elastizitäten

$$\epsilon_\theta \stackrel{\text{def}}{=} \frac{\frac{\Delta L_O}{L_O}}{\frac{\Delta \theta}{\theta}} \, . \tag{6}$$

Der Wert ϵ_θ gibt an, um wieviel Prozent sich der Realoptionswert ändert, wenn der jeweilige Liegenschaftszins um ein Prozent erhöht wird. Da die objektspezifischen Liegenschaftszinsen unterschiedlich sind, ist dies nicht gleichbedeutend mit einer gleichen

absoluten Veränderung der Liegenschaftszinsen. Deshalb wird zusätzlich die prozentuale Änderung der Realoptionswerte für gleiche absolute Änderungen des Liegenschaftszinses berechnet. *Tabelle 6* gibt die Liegenschaftszinsen und die relativen Änderungen der Realoptionswerte an.

[Tabelle 6 ungefähr hier.]

Es sei darauf hingewiesen, dass die in den Wertermittlungsrichtlinien empfohlenen Liegenschaftschaftszinsen in der Bandbreite von 5,0% bis 6,5% liegen (3.5.4 WertR 2002). Diese Liegenschaftszinsen sollen nach WertR 2002 angesetzt werden, wenn keine andere Marktevidenz zur Verfügung steht. Für die meisten Objekte wird diese Bandbreite durch die in der Tabelle angegebenen Änderungen $\Delta\theta$ gut abgedeckt. Die in *Abschnitt 3.3* durchgeführte Robustheitsanalyse hat gezeigt, dass selbst geringe Veränderungen der Liegenschaftszinsen den MQF deutlich verbessern können. Der dort einheitlich angesetzte Liegenschaftszins entsprach einer durchschnittlichen betragsmäßigen Änderung der objektspezifischen Liegenschaftszinsen von 0,5%. Bereits geringe Abweichungen von den objektspezifischen Liegenschaftszinsen können somit zu einer starken Reduktion der Vorhersagefehler führen. Dies bedeutet gleichzeitig, dass kleine Fehler bei der Wahl des Liegenschaftszinses in großen Vorhersagefehlern resultieren können.

Sensitivität in Bezug auf die Volatilität: Wir untersuchen zunächst, wie sensitiv die Realoptionswerte auf eine kleine Änderung der Volatilität reagieren. Hierzu berechnen wir die Elastizität der Realoptionswerte

$$\epsilon_\sigma \stackrel{\text{def}}{=} \frac{\dfrac{\Delta L_O}{L_O}}{\dfrac{\Delta \sigma}{\sigma}} \ . \tag{7}$$

Teil A in *Tabelle 7* gibt die Elastizitäten an.

[Tabelle 7 ungefähr hier.]

Objekte, bei denen der Realoptionswert dem Residualwert entspricht, sind nicht aufgeführt, da für diese die Elastizität Null ist. Die Elastizität ist umso höher, je höher die Optionsprämie ist, dass heißt, desto mehr der Realoptionswert vom Residualwert abweicht. Die Elastizitäten ϵ_σ sind zwar im Durchschnitt kleiner als die Elastizitäten ϵ_θ, aber es zeigt

23

sich auch hier, dass bereits kleine Änderungen zu großen Änderungen der Vorhersagefehler führen können.

Um den Einfluss der Volatilität eingehender zu untersuchen, haben wir für Objekte, bei welchen der Preis über dem Residualwert liegt, die implizierte Volatilität σ_{imp} berechnet. Mit dieser Volatilität entspricht der Realoptionswert dem Preis

$$P = \left\{ \frac{D^*(\sigma_{imp})}{\theta} - I \right\} \left\{ \frac{D}{D^*(\sigma_{imp})} \right\}^{\alpha(\sigma_{imp})} . \tag{8}$$

Bei P handelt es sich um den auf den Bewertungsstichtag diskontierten Transaktionspreis. Die rechte Seite von (8) entspricht der *Bewertungsformel (4)* für den Realoptionswert. Liegenschaftszinssätze werden von Gutachterausschüssen üblicherweise, auf vergleichbare Weise bestimmt, in dem Preise und Ertragswerte (nach WertV) gleichgesetzt werden und für *θ* gelöst wird. Solcherart berechnete Liegenschaftszinsen werden dann für Bewertungen genutzt. Implizierte Volatilitäten könnten in äquivalenter Weise genutzt werden.

Die implizierten Volatilitäten sind in Teil B in *Tabelle 7* für verschiedene Liegenschaftszinsen aufgeführt.[27] Die implizierten Volatilitäten für die *Objekte 1, 3 und 10* unterscheiden sich mit Werten zwischen 9,2% und 22,5% nicht stark von den geschätzten Volatilitäten der Mietwachstumsraten, die eine Bandbreite von 8,1% bis 16,5% haben. Dies gilt umso mehr, da der objektspezifische Liegenschaftszins für die *Objekte 1 und 3* mit 7,7% sehr hoch ist und die implizierten Volatilitäten bei niedrigeren Liegenschaftszinsen von 6,0% und 6,5% vollständig innerhalb dieser Bandbreite liegen. *Objekt 4* ist eine Wohnimmobilie, dessen implizierte Volatilität recht hoch sein kann. Auch hier gilt, dass der objektspezifische Liegenschaftszins mit 6,7% hoch ist. Wäre ein niedrigerer Liegenschaftszins angesetzt worden, so wäre die implizierte Volatilität weitaus niedriger. Würde gar der Liegenschaftszins von 5,0% angesetzt werden, wie in der WertR 2002 für Wohngrundstücke empfohlen, so sollte die Immobilie sofort entwickelt werden, und der Vorhersagefehler würde nur 7,5% betragen. Die implizierten Volatilitäten für *Objekt 5* (in geringerem Maße auch für *Objekt 6*) sind allesamt sehr hoch. Wir haben die Besonderheiten von *Objekt 5* allerdings bereits ausführlich diskutiert

---

[27] *Objekt 9* wurde nicht berücksichtigt, da bei diesem Objekt die projektierte von der tatsächlichen Entwicklung stark abweicht. Wird die tatsächliche BGF zur Bewertung angesetzt, dann sollte dieses Objekt immer sofort entwickelt werden.

Anhand der beiden *Objekte 1 und 3* kann man zeigen, wie implizierte Volatilitäten für die Bewertung genutzt werden können. Beide Objekte liegen im Ostteil der Stadt, für beide ist eine gewerbliche Nutzung vorgesehen und der Neubau maßgeblich, und für beide ist der von den Gutachtern angesetzte Liegenschaftszins identisch. *Objekt 3* wurde rund 2 Jahre früher als *Objekt 1* verkauft. Wäre die implizierte Volatilität von *Objekt 3* für den Realoptionswert von *Objekt 1* verwendet worden, so hätte der Vorhersagefehler nur -8,9% betragen.

Zusammenfassend ist festzuhalten, dass eine höhere als die von uns einheitlich angesetzte Volatilität von 16,5% nicht unplausibel sein muss. *Quigg* (1993) etwa erhält für den Immobiliemarkt in Seattle durchschnittliche implizierte Volatilitäten in einer Bandbreite von 18,6% bis 28,1%. Die von uns in Abschnitt 3.3 berechnete Volatilität, welche den MQF der Realoptionswerte minimiert, liegt mit 20,7% innerhalb dieser Bandbreite. Es gibt mindestens zwei sich ergänzende Erklärungen, warum die Volatilität höher sein könnte: Einerseits berücksichtigen wir nur die Option des Wartens und vernachlässigen damit andere Optionen, die es erlauben, die Entwicklung an die jeweilige Marktsituation anzupassen. Dies wurde bei der Diskussion der *Objekte 5 und 9* deutlich. Andererseits wurde Unsicherheit in Bezug auf die Entwicklungskosten nicht berücksichtigt. Sind die Wachstumsraten von Reinerträgen und Baukosten nicht stark positiv korreliert, so kann dies die relevante Volatilität erhöhen.[28]

Fazit zu den Sensitivitätsanalysen: Der Einfluss der Liegenschaftszinsen auf die Realoptionswerte ist erwartungsgemäß stark und geringe Variationen können zu großen Änderungen des Realoptionswertes führen. Die dabei durch die Liegenschaftszinsen implizierten Volatilitäten müssen nicht immer plausibel sein. Der Einfluss der Volatilität auf den Realoptionswert ist im Mittel geringer, aber dennoch nicht zu vernachlässigen. Die hohe Sensitivität der Realoptionswerte bedeutet zugleich jedoch auch, dass die beobachteten Preise prinzipiell mit Realoptionswerten repliziert werden können. Die Analyse hat gezeigt, dass dies mit plausibler Kombination der Parameterwerte möglich scheint.

---

28 Die Volatilität ist in diesem Fall $\sigma = \sqrt{\sigma_D^2 + \sigma_I^2 - 2\sigma_{DI}}$, wobei σ_{DI} für die Kovarianz der Wachstums-raten von Reinerträgen und Entwicklungskosten steht (*Williams*, 1991; *Quigg*, 1993). Mit Mieten für das Berliner Zentrum und Baukosten für Gewerbegebäude erhalten wir $\sigma = 17,9\%$. Um die Realoptionswerte unter der Annahme stochastischer Entwicklungskosten zu berechnen, werden weitere Parameterschätzer benötigt, über die wir nicht verfügen. Der Gesamteffekt auf die Realoptionswerte kann deshalb nicht abgeschätzt werden.

4 Zusammenfassung und Ausblick

Aus ökonomischer Sicht ist das Realoptionsverfahren von großer Relevanz, da es im Gegensatz zu anderen Bewertungsverfahren die Flexibilität bei der Immobilien-entwicklungen explizit berücksichtigt. Ziel der Studie war es, das Realoptionsverfahren auf seine Praxistauglichkeit hin zu überprüfen und einen Vergleich mit traditionell angewandeten Bewertungsverfahren durchzuführen. Die Analyse basiert auf einem Datensatz von zwölf Berliner Grundstücken, welche erst nach erheblichen Bauinvestitionen zu Erträgen führen. Für jedes der Grundstücke liegen weitreichende Angaben, gutachterliche Bewertungen und Transaktionspreise vor, die einen Vergleich der Bewertungsverfahren in Hinsicht auf deren Vorhersagegüte erlauben. Zudem wurden Angaben zur Mietentwicklung in Berlin herangezogen, um alle notwendigen Eingangsparameter schätzen zu können. Diese Angaben waren ausreichend, um für jedes der Objekte den Realoptionswert zu berechnen. Die Robustheitsanalyse hat zudem gezeigt, das Vereinfachungen, die bei der Berechnung der Realoptionswerte genutzt wurden, nicht unbedingt zu schlechteren Bewertungen führen müssen.

Im Mittel schneiden die gutachterlichen Vergleichswerte am besten ab, Realoptionswerte sind jedoch eindeutig besser als Residualwerte. Das Realoptionsverfahren ist damit das bessere Verfahren nach dem Ertragswertansatz. Sensitivitätsanalysen haben zudem gezeigt, dass Realoptionswerte prinzipiell die Preise replizieren können. Zugleich haben die Sensitivitätsanalysen aber auch gezeigt, dass die Realoptionswerte immer nur so gut sein können, wie die Annahmen, die über die Entwicklung des Grundstücks und die Eingangsparameter getroffen werden. Dieses Problem betrifft unseres Erachtens jedoch alle Bewertungsverfahren gleichermaßen. Aufgrund der Bewertungsformel und des ökonomischen Modellrahmens erlaubt es das Realoptionsverfahren immerhin explizit, den Einfluss von Parametervariationen auf die einzelne Bewertung zu berechnen. Gutachtern wird damit ein Instrument an die Hand gegeben, welches in Kombination mit deren Expertenwissen insgesamt zu guten Bewertungen führen sollte.

Das Ergebnis unserer Umsetzungsstudie sollte keinesfalls als abschließendes Urteil zur Güte des Realoptionsverfahrens in der Immobilienbewertung verstanden werden. Um zu einem abschließenden Urteil zu gelangen, müssten ähnliche Studien für eine größere Zahl von Objekten folgen. Besonderes Augenmerk sollte dabei auf der Schätzung der Eingangsparameter liegen, denn je verlässlicher die Schätzwerte sind, desto eindeutiger

kann danach die Güte der Realoptionswerte beurteilt werden. Statt Zeitreihen der Mietentwicklung für die Schätzung der historischen Volatilität zu verwenden, erscheinen uns implizierte Volatilitäten ein geeigneter Weg, um zuverlässige Schätzwerte für den Volatilitätsparameter abzuleiten. Dabei wäre es wünschenswert, ähnlich wie bei Finanzmarktstudien, statistische Methoden für die Modellierung implizierter Volatilitäten einzusetzen. Modelle wie GARCH, die für Finanzmarktdaten erfolgreich angewandt werden[29], sind jedoch auf eine größere Datenmenge angewiesen. Es ist zu hoffen, dass für solche Studien in Zukunft ausreichend Daten verfügbar werden.

Schließlich sei auf eine weitere Erweiterung des Realoptionsverfahrens hingewiesen, in welcher Wettbewerbseffekte explizit berücksichtigt werden[30]. Der Wettbewerb um die Entwicklung bestimmter Objekte kann, wie etwa empirische Evidenz in den Arbeiten von *Bulan et al.* (2004) und *Schwartz* und *Torous* (2006) zeigt, die Optionsprämie vermindern. Für unsere Daten scheint dies allerdings nur von untergeordneter Bedeutung zu sein, da positive Optionsprämien meist mit den Transaktionspreisen vereinbar sind. Dies zeigt jedoch, dass die Theorie der Realoptionen in ständiger Weiterentwicklung begriffen ist, um die Interaktion und Preisbildung auf Immobilienmärkten so realistisch wie möglich zu modellieren. Ob realistischere Modelle dann allerdings zu besseren Bewertungen führen, können wiederum nur empirische Studien zeigen.

Literatur

Arrow, Kenneth. J. und *Fisher, Adrian C.* (1974), Environmental preservation, uncertainty, and irreversibility, in: Quarterly Journal of Economics, Vol. 88, S.312–319.

Bewag Aktiengesellschaft (2000), Elektropolis Berlin. Historische Bauten der Stromverteilung, Bewag Aktiengesellschaft, Berlin. Ohne Jahr.

Bulan, Laarni, Mayer, Christopher J. und *Somerville, Tsuriel C.* (2004), Irreversible investment, real options, and competition: Evidence from real estate development, Technical report, Columbia University, Graduate School of Business New York.

Capozza, Dennis. R. und *Li, Y.* (2002), Optimal land development decisions, in: Journal of Urban Economics, Vol. 51, S. 123–142.

---

[29] *Heston/Nandi* (2000)
[30] Grenadier (2005)

Deutsche Bundesbank (1997), Schätzung von Zinsstrukturkurven, in: Deutsche Bundesbank Monatsbericht (Oktober), S. 61–66.

Dixit, Avinash K. und *Pindyck, Robert S.* (1994), Investment under Uncertainty, in: Princeton University Press, Princeton, New Jersey.

Dumas, Bernard; Fleming, Jeff und *Whaley, Robert E.* (1998), Implied volatility functions: Empirical tests, in: Journal of Finance, Vol. 53, S. 2059–2106.

Grenadier, Steven R. (2005), An equilibrium analysis of real estate leases, in: Journal of Business, Vol. 78, S. 1173–1213.

Heston, Steven L. und *Nandi, Saikat* (2000), A closed-form GARCH option valuation model, in: Review of Financial Studies, Vol.13, S. 585–625.

Hommel, Ulrich und *Pritsch, Gunnar* (1999a), Investitionsbewertung und Unternehmensführung mit dem Realoptionsansatz, in A.-K. Achleitner und G. F. Thoma (Hrsg.), Handbuch Corporate Finance, Ergänzungslieferung 4/1999, Verlag Deutscher Wirtschaftsdienst, Köln, S. 1–68.

Hommel, Ulrich und *Pritsch, Gunnar* (1999b), Marktorientierte Investitionsbewertung mit dem Realoptionsansatz: Ein Implementierungsleitfaden für die Praxis, Finanzmarkt und Portfolio Management, Vol. 13, S. 121–144.

Lander, D. M. und *Pinches, G. E.* (1998), Challenges to the practical implementation of modeling and valuing real options, Quarterly Review of Economics and Finance, Vol. 38, S. 537–567.

Leung, B. Y. P. und *Hui, Chi Man Eddie* (2002), Option pricing for real estate development: Hong Kong Disneyland, Journal of Property Investment & Finance, Vol. 20, S. 473–495.

Lucius, Dominik I. (2000), Real options in real estate development, Journal of Property Investment & Finance, Vol. 19, S. 73–78.

McDonald, R. und *Siegel, Daniel* (1986), The value of waiting to invest, Quarterly Journal of Economics, Vol. 101, S. 707–728.

Paddock, James L., Siegel, Daniel R. und *Smith, James L.* (1988), Option valuation of claims on real assets: The case of offshore petroleum leases, Quarterly Journal of Economics, Vol. 103, S. 479–508.

Patel, Kanak und *Paxson, Dean* (2001), Real urban development options at Canary Wharf, in S. Howell, A. Stark, D. Newton, D. Paxson, M. Cavus, J. Pereira und K. Patel (Hrsg.),

Real Options. Evaluating Corporate Investment Opportunities in a Dynamic World, Pearson Education, Harlow, S. 163–176.

Pindyck, Robert S. (1991), Irreversibility, uncertainty, and investment, Journal of Economic Literature, Vol. 29, S. 1110–1148.

Quigg, Laura (1993), Empirical testing of real option-pricing models, Journal of Finance, Vol. 48, S. 621–640.

Schwartz, Eduardo S. und *Torous, Walter N*. (2006), Commercial office space: Tests of a real options model with competitive interactions. Erscheint demnächst in Real Estate Economics.

Shackleton, Mark und *Wojakowski, Ralf* (2002), The expected return and exercise time of Merton-style real options, Journal of Business Finance & Accounting, Vol. 29, S. 541–555.

Titman, Sheridan (1985), Urban land prices under uncertainty, American Economic Review, Vol. 75, S. 505–514.

Trigeorgis, Lenos (1996), Real Options, Managerial Flexibility and Strategy in Resource Allocation, MIT Press, Cambridge MA.

Williams, Joseph T. (1991), Real estate development as an option, Journal of Real Estate Finance and Economics, Vol. 4, S. 191–208.
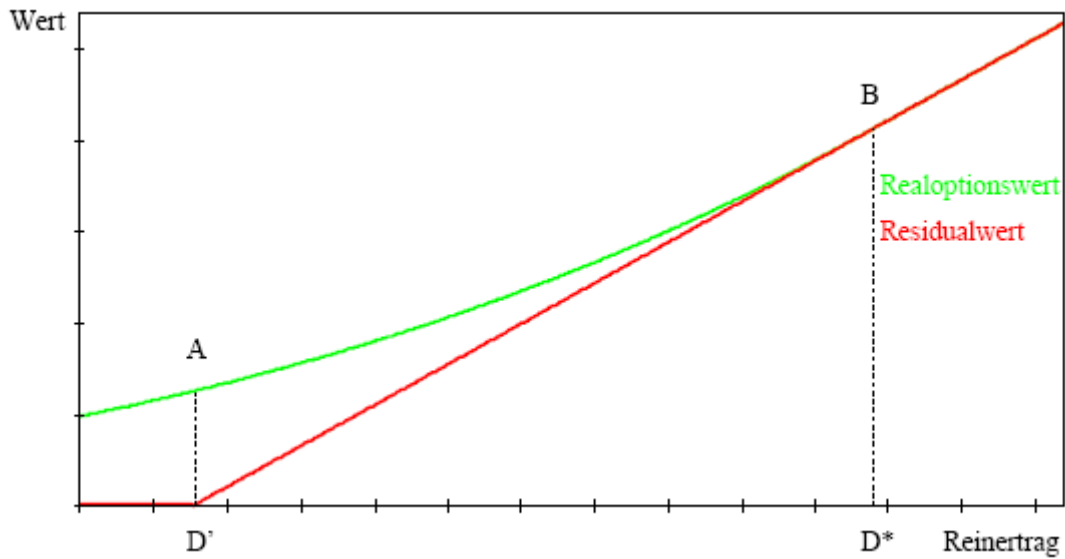
Abbildung 1: Realoptionswert L_O (obere Linie) und Residualwert L_R (untere Linie) als Funktion des Reinertrages D, gegeben Entwicklungskosten I. Bei $D < D´$ ist der Ertragswert bei sofortiger Entwicklung kleiner als die Entwicklungskosten und der Residualwert ist 0. Der Realoptionswert ist aber strikt positiv (links von A), da zukünftige Reinerträge über $D´$ liegen können. Bei $D´ < D < D*$ ist eine sofortige Entwicklung lohnend und der Residualwert ist strikt positiv. Warten ist jedoch optimal und der Realoptionswert ist größer (rechts von A). Bei $D \geq D*$ sind Residual- und Realoptionswert gleich (rechts von B), und eine sofortige Entwicklung ist optimal.

Tabelle 1: Bewag Objekte

Objekt	Ursprüngliche Nutzung	Lage
1	Schaltstation	Äußerer Stadtraum
2	Umspannwerk	Äußerer Stadtraum
3	Heizwerk	Innenstadt
4	Umspannwerk	Äußerer Stadtraum
5	Umspannwerk	Innenstadt
6	Umspannwerk	Äußerer Stadtraum
7	Gästehaus an der Spree	Äußerer Stadtraum
8	Geschäftshaus	Zentrum West
9	Wohn- und Geschäftshaus	Zentrum Ost
10	Umspannwerk	Zentrum Ost
11	Umspannwerk	Innenstadt
12	Umspannwerk	Innenstadt

Tabelle 2: Statistiken für Bewertungsausgangsdaten und Transaktionsdaten für alle zwölf Bewag Objekte zum Stichtag Dezember 1997.

	Mittelwert	Median	Standab.	Min	Max	Einheiten
Teil A: Bewertungsausgangsdaten						
Grundstücksfläche	6189	1587	13101	560	46657	qm
BGF vor Entwicklung	6786	1390	14562	0	51653	qm
BGF nach Entwicklung	8699	2158	18659	700	66653	qm
Nutzfläche	6189	1587	13101	560	46657	qm
Reinertrag	141	126	49	89	245	EUR pro qm
Entwicklungskosten	8615	2095	17032	476	58788	Tsd. EUR
Liegenschaftszins	6,4	6,3	0,7	5,4	7,7	Prozent p.a.
Bodenrichtwert	1349	614	1842	143	6136	EUR pro qm
Bodenwert	1346	614	1845	128	6136	EUR pro qm
Teil B: Transaktionsdaten						
Preis	11799	677	27330	189	93055	Tsd. EUR
Preis Stichtag	9927	590	22759	150	77545	Tsd. EUR
Zeit bis zum Verkauf	26	23	11	12	54	Monate

Erläuterungen: Die meisten Angaben in Teil A beruhen auf gutachterlichen Einschätzungen. BGF ist die Bruttogrundfläche. Nutzfläche, jährlicher Reinertrag pro Quadratmeter Nutzfläche, Kosten und Liegenschaftszins sind für die entwickelte Immobilie. Bodenrichtwerte beruhen auf dem Berliner Bodenrichtwertatlas, Stand Dezember 1996. Teil B ist für Transaktionspreise, diskontierte Preise und die Zeit seit Dezember 1997 bis zum Verkauf der unentwickelten Immobilie, Diskontierungsrate ist die geforderte Rendite von 7,6%

Tabelle 3: Kriterien für eine sofortige Entwicklung

Objekt	d	d^*	r_s	r_s^*	r_d	r_d^*	Sofortige Entwicklung
1	0,632	1	6,3	10,0	6,2	9,9	Nein
2	0,972	1	8,3	8,6	10,0	10,2	Nein
3	0,589	1	5,9	10,0	5,8	9,9	Nein
4	0,948	1	8,7	9,2	9,6	10,1	Nein
5	0,753	1	6,7	8,8	8,0	10,2	Nein
6	1,037	1	8,9	8,6	10,6	10,2	Ja
7	2,327	1	20,6	8,8	21,9	10,2	Ja
8	1,821	1	15,6	8,6	17,3	10,2	Ja
9	2,276	1	18,7	8,2	20,8	10,4	Ja
10	1,007	1	8,9	8,8	10,2	10,2	Ja
11	2,078	1	19,1	9,2	20,0	10,1	Ja
12	0,888	1	7,9	8,8	9,2	10,2	Nein

Erläuterungen: Angaben für Renditen in Prozent. Das Verhältnis des aktuellen zum kritischen Reinertrag ist $d = D/D^*$. Die statischen Renditen sind $r_s = D/I$ und $r_s^* = D^*/I$. Die dynamischen Renditen sind $r_d = r_s + g$ und $r_d^* = r_s^* + g$, wobei für g die geforderte Rendite von 7,6% minus den jeweiligen Liegenschaftszins θ angesetzt wurde.

Tabelle 4: Verhältnis von Bewertung und Transaktionspreis nach verschiedenen Verfahren.

Teil A: Bewertungsverhältnisse					
Objekt	Q_E	Q_A	Q_R	Q_O	Q_G
1	122,2	122,2	0,0	49,6	85,9
2	80,5	77,0	198,0	198,6	137,8
3	355,9	131,2	0,0	50,3	90,7
4	85,7	116,8	44,3	45,0	80,9
5	132,5	80,3	9,4	22,7	51,5
6	113,5	109,6	63,4	63,4	86,5
7	41,7	86,7	184,8	184,8	135,7
8	51,8	54,1	125,0	125,0	89,5
9	58,5	58,5	76,3	76,3	67,4
10	119,6	50,3	106,4	106,4	78,4
11 und 12	75,3	68,4	116,8	117,3	92,9
Teil B: Übereinstimmung der Bewertungen					
	Q_E	Q_A	Q_R	Q_O	Q_G
Q_E	5	8	3	3	4
Q_A	.	4	2	2	5
Q_R	.	.	5	11	8
Q_O	.	.	.	5	8
Q_G	.	.	.	.	2

Erläuterung: Angaben in Teil A in Prozent. Teil B gibt auf der Diagonalen die Anzahl von Fällen, in welchen der Preis überschätzt wird und oberhalb der Diagonalen die Anzahl von Fällen, in welchen zwei Verfahren bei der Über- bzw. Unterschätzung übereinstimmen.

Tabelle 5: Statistiken für die Bewertungsverhältnisse nach den einzelnen Verfahren

	Mittelwert	Median	Standardabweichung	Min	Max	MQF
Q_E	112,5	85,7	86,3	41,7	355,9	76,1
Q_A	86,8	80,3	28,8	50,3	131,2	10,1
Q_R	84,0	76,3	69,4	0,0	198,0	50,7
Q_O	94,5	76,3	57,8	22,7	198,6	33,8
Q_G	90,7	86,5	25,7	51,5	137,8	7,5

Erläuterungen: Angaben in Prozent. Der Mittlere Quadratische Fehler (MQF) wurde nach (5) berechnet.

Tabelle 6: Sensitivitätsanalyse bezüglich der Liegenschaftszinsen: Prozentuale Änderungen der Realoptionswerte

Objekt	θ	ϵ_θ	Änderung des Liegenschaftszinses $\Delta\theta$					
			-0,75	-0,5	-0,25	0,25	0,5	0,75
1	7,7	-6,3	84,6	51,1	23,2	-19,2	-35,1	-48,1
2	5,9	-3,3	49,5	31,3	14,8	-13,2	-25,0	-35,4
3	7,7	-6,6	90,6	54,5	24,6	-20,1	-36,5	-49,7
4	6,7	-3,8	51,8	32,6	15,4	-13,7	-25,8	-36,4
5	6,3	-4,4	67,1	41,4	19,2	-16,4	-30,5	-42,4
6	5,9	-2,9	43,3	27,5	13,1	-12,1	-23,2	-33,2
7	6,3	-1,4	19,6	12,5	6,0	-5,5	-10,6	-15,4
8	5,9	-1,6	23,5	14,9	7,1	-6,5	-12,6	-18,2
9	5,4	-1,4	22,7	14,3	6,8	-6,2	-11,9	-17,2
10	6,3	-3,3	45,7	29,1	14,0	-12,8	-24,3	-34,5
11	6,7	-1,5	19,5	12,5	6,0	-5,6	-10,7	-15,5
12	6,3	-3,8	55,9	34,9	16,4	-14,4	-27,0	-38,0

Erläuterungen: Angaben in Prozent. θ ist der objektspezifische Liegenschaftszins. Die Elastizität ε_θ wurde nach (6) berechnet. $\Delta\theta$ ist die absolute Veränderung des jeweiligen Liegenschaftszinses.

Tabelle 7: Sensitivitätsanalyse bezüglich der Volatilität: Elastizitäten und implizierte Volatilitäten

Teil A: Elastizitäten ϵ_σ					
Objekt					
1	2	3	4	5	12
2,6	0,1	3,0	0,2	1,1	0,5

Teil B: Implizierte Volatilitäten σ_{imp}					
Objekt					
θ	1	3	4	5	6	10
objektspezifisch	22,5	21,5	60,0	88,9	52,1	.
5,0	.	.	.	49,1	26,3	.
5,5	.	.	32,7	63,5	41,9	.
6,0	9,2	10,2	44,6	79,7	55,3	.
6,5	14,0	13,9	56,1	98,8	69,3	19,7

Erläuterungen: Angaben in Prozent. Werte in Teil A wurden nach (7) berechnet und Werte in Teil B nach (8). θ ist der Liegenschaftszins.

Portfolio value at risk based on independent component analysis

Ying Chen[a,b,*], Wolfgang Härdle[a], Vladimir Spokoiny[a,b]

[a]*CASE—Center for Applied Statistics and Economics, Humboldt-Universität zu Berlin, Wirtschaftswissenschaftliche Fakultät, Spandauerstrasse 1, 10178 Berlin, Germany*
[b]*Weierstraß—Institute für Angewandte Analysis und Stochastik, Mohrenstrasse 39, 10117 Berlin, Germany*

Abstract

Risk management technology applied to high-dimensional portfolios needs simple and fast methods for calculation of value at risk (VaR). The multivariate normal framework provides a simple off-the-shelf methodology but lacks the heavy-tailed distributional properties that are observed in data. A principle component-based method (tied closely to the elliptical structure of the distribution) is therefore expected to be unsatisfactory. Here, we propose and analyze a technology that is based on independent component analysis (ICA). We study the proposed ICVaR methodology in an extensive simulation study and apply it to a high-dimensional portfolio situation. Our analysis yields very accurate VaRs.
© 2006 Elsevier B.V. All rights reserved.

1. Introduction

The value at risk (VaR) calculation of large portfolios is a challenging task—both numerically and statistically. VaR indicates the possible loss over a given time horizon at a risk level *a*. From a statistical point of view, it is the *a*-quantile of the joint distribution of the portfolio's risk factors which are modeled heteroscedastically:

$$x_t = \Sigma_t^{1/2}\varepsilon_t, \tag{1}$$

where $x_t \in \mathbb{R}^d$ is the risk factor vector, e.g., the (log) returns of d individual financial instruments. The matrix Σ_t denotes the corresponding time-dependent covariance and ε_t is the d-dimensional standardized residual vector. The portfolio VaR calculation becomes technically difficult for high dimensionality of the portfolio.

In order to solve this and other numerical problems, portfolio variations are typically mapped into a conditional multivariate normal framework such as RiskMetrics launched by J.P. Morgan. Recall that, Gaussian distributed residuals or devolatilized returns $\varepsilon_t = \Sigma_t^{-1/2}x_t$ are independent after devolatilization; the joint density of the residuals is the product of d marginals. In this sense, the portfolio VaR calculation is simplified and only covariance based.

---

* Corresponding author. CASE—Center for Applied Statistics and Economics, Humboldt-Universität zu Berlin, Wirtschaftswissenschaftliche Fakultät, Spandauerstrasse 1, 10178 Berlin, Germany.
E-mail address: ychen@wiwi.hu-berlin.de (Y. Chen).
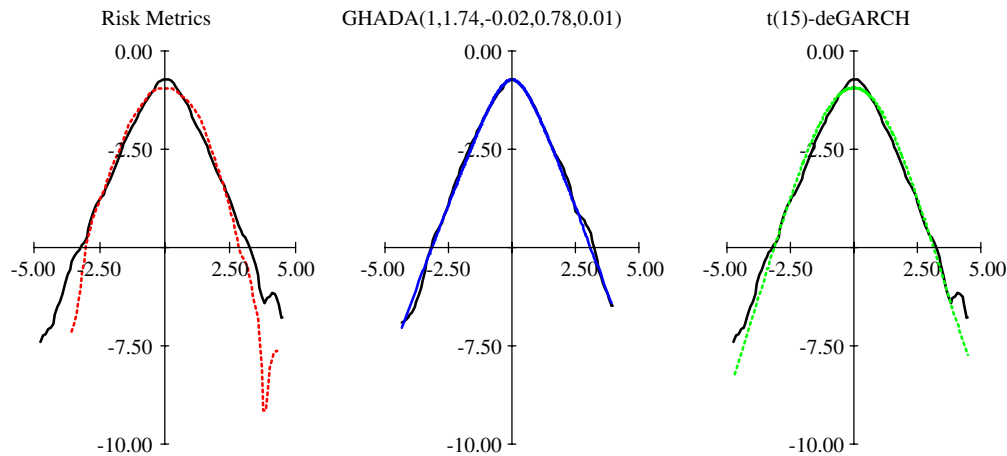
Fig. 1. Graphical comparison of density estimations based on the devolatilized DEM/USD returns from 1979-12-01 to 1994-04-01 (3719 observations). The non-parametrical kernel density estimations are considered as benchmarks. The RiskMetrics-based density estimation is the dotted curve in the left panel whereas the $t(15)$-deGARCH-based estimation is displayed in the right. The corresponding GARCH$(1, 1)$ process is: $\hat{\sigma}_t^2 = 1.65 * 10^{-6} + 0.07 x_{t-1}^2 + 0.89 \hat{\sigma}_{t-1}^2$. The GHADA technique with the local constant volatility process and the GH$(1, 1.74, -0.02, 0.78, 0.01)$ density estimation is displayed in the middle panel. Data source: FEDC (http://sfb649.wiwi.hu-berlin.de/).

Many well-developed covariance estimation methodologies in a high-dimensional space have been easily applied in practice, for example, the constant conditional correlation (CCC) model proposed by Bollerslev [3] and the subsequent dynamic conditional correlation (DCC) model proposed by Engle [7] and Engle and Sheppard [8]. The simplicity of this kind of covariance-based methodology nevertheless bears a risk of modeling bias since, among other things, the assumed conditional Gaussian marginals are unable to mimic the heavy-tailedness of financial time series observed in markets. This issue has been addressed in a variety of papers. For example, Jaschke and Jiang [14] have studied the conditional Gaussian distribution fits to VaR that deliver satisfactory estimates at a moderate (e.g., 95%) confidence level but underestimate VaRs at more extreme levels such as 99%.

The tail problems are evident from Fig. 1, where we compare the marginal density estimations of devolatilized returns of foreign exchange rates, the German Mark to the US Dollar (DEM/USD), from 1979-12-01 to 1994-04-01. In order to mimic the empirical distributional behavior of the real data, we assume three different distributional types. The RiskMetrics (left panel) and $t(15)$-deGARCH (right panel) methods fit the devolatilized returns, based on a GARCH(1,1) volatility process: $\hat{\sigma}_t^2 = 1.65 * 10^{-6} + 0.07 x_{t-1}^2 + 0.89 \hat{\sigma}_{t-1}^2$, by the normal and student-t with degrees of freedom (df) 15 distributions. The GHADA (middle panel) technique assumes that the devolatilized returns follow a time stationary generalized hyperbolic (GH) distributional mechanism based on locally constant volatilities, see [4,17] for details. The non-parametric kernel density estimations (solid lines) corresponding to the devolatilized return processes are considered as benchmarks. According to the graphical comparison, the GHADA technique is superior to the other two techniques since the empirical GH density coincides to the benchmark, especially in the tails. The t-deGARCH technique although shows a better tail fit compared to the RiskMetrics, it still deviates from the benchmark. This small comparison provides evidence that the normally distributional assumption is unreliable and will lead to low accuracy of univariate and portfolio VaR calculations.

The weak performance of the normal assumption motivates us to search for a different approach solving the technical problems of portfolio VaR calculations. An "ideal" situation is, as mentioned before, that the residuals ε_t are independent. Since based on the independence, the estimation of the joint distribution can be converted to marginals' estimations.

In the context of sound engineering, signal detection from unknown filters and sources is treated by a method called independent component analysis (ICA). This engineering method is designed for detection of blind folded signals and retrieves out of a high-dimensional time series stochastically independent source components. A tutorial on ICA can be found in [12] and a variety of numerical techniques to uncover independent components (ICs) are discussed in [11]. Promoted by the success in engineering, ICA has been applied in different areas such as brain imaging [6] and telecommunication study [19]. An early implementation of ICA in financial time series is given in [1], drawing

comparisons of ICs and principal components (PCs) applied to 28 Japanese stocks from 1986 to 1989. Few contributions however exist for the application of ICA in risk management.

The first aim of this paper is to bring together the lines of thought of the engineering signal processing literature and newer statistical insights on the high-dimensional VaR calculations. We coin the name ICVaR owing to the ICA technology. Given a trading strategy $b_t \in \mathbb{R}^d$, the portfolio return $R_t \in \mathbb{R}$ is

$$R_t = b_t^\top x_t. \tag{2}$$

Note that the ICVaR technique does not rely on a direct joint density estimation of the high-dimensional returns $x_t \in \mathbb{R}^d$ such as $x_t = \Sigma_t^{1/2} \varepsilon_t$. Instead the ICVaR procedure in a first step applies a linear transformation to x_t, i.e., a non-singular matrix W yields (approximately) ICs $y_t \in \mathbb{R}^d$:

$$x_t = W^{-1} y_t. \tag{3}$$

The matrix W is different from the Mahalanobis transformation $\text{cov}(x)^{-1/2}$—that creates ICs in the Gaussian regime—only in the case of non-Gaussian marginals as we will see later. The second ICVaR step concerns the fit of each IC univariately:

$$y_t = \text{diag}(\sigma_{1t}, \ldots, \sigma_{dt})\varepsilon_t = D_t^{1/2} \varepsilon_t,$$
$$\text{or equivalently: } y_{jt} = \sigma_{jt}\varepsilon_{jt}, \quad j = 1, \ldots, d, \tag{4}$$

where the covariance D_t of the ICs is diagonal due to independence and ε_t are cross-independent shocks. Based on the ICA, the high-dimensional VaR problem is now converted to simpler univariate VaR calculations. Given the discussion above on fitting VaR this opens a wide avenue of alternative VaR determination.

The second purpose of this paper is to compare the proposed ICVaR technique with the industry standard RiskMetrics and the most-often used t-deGARCH method. As discussed before, the RiskMetrics often gives underestimated VaRs, an inevitable cost of the Gaussian assumption. On the contrary, the students-t distribution can better mimic the heavy-tailed distributional behavior of financial risk factor than the RiskMetrics, but it is hard to reflect the leptokurtic scenario exactly, as illustrated in Fig. 1. In this paper, we show how high accuracy can be reached by the ICVaR compared to the RiskMetrics and t-deGARCH methods in real data analysis.

The methodological contribution of the study unfolds in Section 2 where the ICA method is discussed. The simulation study is presented in Section 3. Further, we apply the ICVaR to exchange rate portfolios with different artificial trading strategies. The ICVaR predicts risk levels precisely and outperforms the RiskMetrics and t-deGARCH methods. Finally, we conclude our study in Section 5. All algorithms and figures can be recalculated by following the net linked indicator below the figures.

2. ICVaR methodology

2.1. Basic model

The proposed ICVaR methodology consists of two main steps: searching for ICs based on a linear transformation and modeling the univariate volatility and stochastic distribution.

$$R_t = b_t^\top x_t \tag{5}$$
$$= b_t^\top W^{-1} y_t \tag{6}$$
$$= b_t^\top W^{-1} D_t^{1/2} \varepsilon_t. \tag{7}$$

The idea of the ICA is that risk factors $x_t \in \mathbb{R}^d$ can be represented by a linear combination of d-dimensional ICs. The linear transformation matrix W is assumed to be non-singular. Due to the independence property of ICs, the covariance D_t must be a diagonal matrix and the elements of the stochastic vector ε_t are cross-independent. Furthermore, it fulfills that $\mathbb{E}[\varepsilon_t | F_{t-1}] = 0$ and $\text{Var}[\varepsilon_t | F_{t-1}] = I_d$.

From a statistical viewpoint, this projection technique is desirable since the d-dimensional portfolio is decomposed to univariate and independent risk factors through a simple linear transformation. Recall that, the joint density (f) and the covariance of any linear transformed ICs such as $x_t = W^{-1} y_t$ are analytically computable:

$$f_y = \prod_{j=1}^{d} f_{y_j}, \quad \Sigma_{t,y} = D_t,$$

$$f_x = \mathrm{abs}(|W|) f_y(Wx), \quad \Sigma_{t,x} = W^{-1} D_t W^{-1\top}.$$

In the second step of the proposed ICVaR, the diagonal elements of D_t and each component of ε_t are estimated univariately since the matrix manipulation is equivalent to

$$y_{jt} = \sigma_{jt} \varepsilon_{jt}, \quad j = 1, \ldots, d, \tag{8}$$

where σ_{jt} is the square root of the jth diagonal element of D_t and ε_{jt} is the univariate stochastic term with $\mathrm{E}[\varepsilon_{jt}|F_{t-1}]=0$ and $\mathrm{Var}[\varepsilon_{jt}|F_{t-1}]=1$. There are various univariate models that estimate the volatility and approximate the distribution of the stochastic term. Here we focus on the IC identification step. For the fit of the marginal IC factors we refer to the GHADA technique as in [4], where one estimates the local constant volatility adaptively in a homogeneous interval and fits the devolatilized returns in the GH distributional framework. An alternative approach is given by the t-deGARCH setup. It could be used to estimate the heteroscedastic volatility process and the student-t could be applied to pick up the heavy-tailedness of ε_j.

After these two steps, we compute the quantile of portfolio risk by Monte Carlo simulation or analytical methods e.g., saddle point approximation, see [13]. For the ease of presentation, we concentrate in this paper on the simulation methodology. In particular, we generate d-dimensional samples of the fitted distributions with sample size M, from which we calculate the daily empirical a-quantile of the portfolio variations. The simulation will repeat N times and the average value of the empirical quantiles is considered as the portfolio VaR at level a:

$$\mathrm{VaR}_{a,t} = \frac{1}{N} \sum_{n=1}^{N} \hat{F}_{a,t}^{-1}(R_t^{(n)}) = \frac{1}{N} \sum_{n=1}^{N} \hat{F}_{a,t}^{-1} \{ b_t^\top \hat{W}^{-1} \hat{D}_t^{1/2} \hat{\varepsilon}_t^{(n)} \},$$

where $\hat{F}_{a,t}^{-1}$ denotes the empirical quantile function of R_t.

2.2. ICA: properties and estimation

Since ICA is a relatively new technique in this context, we present a small pedagogical illustration of its usage.

Example. Generate three independent GH random variables GH($y; 1, 2, 0, 1, 0$), GH($y; 1, 1.7, 0, 0.5, 0$) and GH($y; 1, 1.5, 0, 1, 0$) as sources. The first distributional parameter specifies the subclass of the random variables. With the value of 1, they are hyperbolic (HYP) distributed. The other four parameters control the location, scale, asymmetry and likeliness of extreme events, see [18]. These source components have mean $(-0.02, 0.05, -0.00)^\top$ and standard deviation (SD) $(0.83, 0.92, 0.99)^\top$, respectively. The linear transformation matrix is the estimate based on three real German stocks' returns: ALLIANZ, BASF and BAYER from 1974-01-02 to 1996-12-30 (Data source: FEDC at http://sfb649.wiwi.hu-berlin.de).

$$W^{-1} = \begin{pmatrix} 1.31 & 0.14 & 0.18 \\ -0.42 & -1.26 & -1.25 \\ -0.03 & 0.41 & -0.49 \end{pmatrix} 10^{-2}. \tag{9}$$

The time series $x_t = W^{-1} y_t$ are analyzed by the ICA. The covariance of x_t is

$$\mathrm{cov}(x) = \begin{pmatrix} 13.66 & 6.04 & 6.49 \\ 6.04 & 15.54 & 11.64 \\ 6.49 & 11.64 & 16.07 \end{pmatrix} 10^{-5} \tag{10}$$
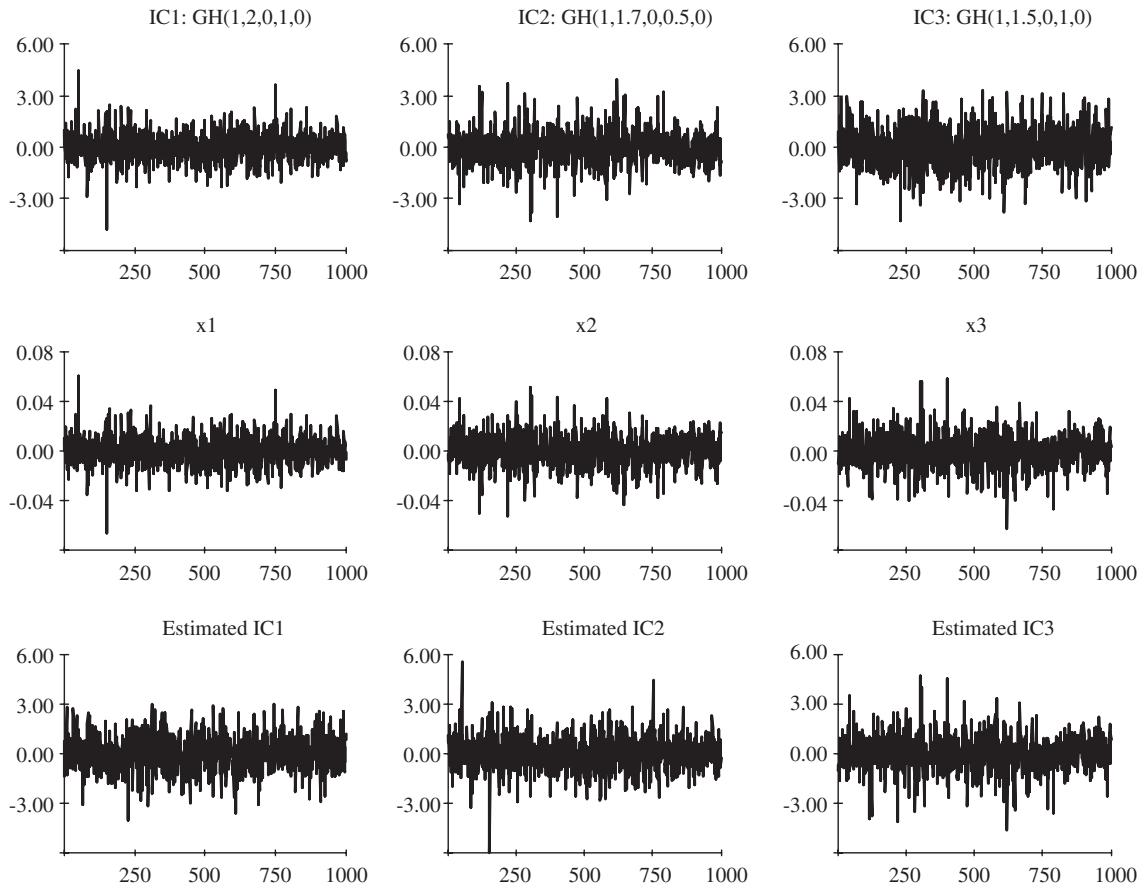
Fig. 2. Simulation: on the top are time plots of three independent GH variables. The mixed variables $x_t = W^{-1} y_t$ are displayed on the middle row where the linear transformation matrix W is estimated based on three German stocks' returns: ALLIANZ, BASF and BAYER from 1974-01-02 to 1996-12-30. The time series of the estimated ICs are displayed on the bottom.

which is comparable to $W^{-1} D_t W^{-1\top}$. Recall that in the Gaussian framework, the Mahalanobis transformation delivers independent variables:

$$\widehat{\text{cov}}(x)^{-1/2} = \begin{pmatrix} 0.91 & -0.09 & -0.12 \\ -0.09 & 1.03 & -0.41 \\ -0.12 & -0.41 & 1.04 \end{pmatrix} 10^2 \tag{11}$$

which is clearly distinct from

$$W = \begin{pmatrix} 0.79 & 0.10 & 0.03 \\ -0.11 & -0.44 & 1.08 \\ -0.15 & -0.38 & -1.10 \end{pmatrix} 10^2 \tag{12}$$

indicated by (9). Fig. 2 provides an illustration of this procedure. The top row contains the three independent source signals y_t. The middle row displays the time series $x_t = W^{-1} y_t$. One sees the scale changes and different random patterns display. The last row of Fig. 2 shows the ICs estimated by the ICA method. The time series on the top and the bottom look familiar but the sign and the ordering may change as displayed in the figure; the first estimated IC displays the similar movement as the third generated IC. Furthermore, the third estimated IC has a mirror pattern of the second true IC.

2.2.1. Scale identification

In fact, the scales of y_t and W are not identifiable. Given a matrix $C = \text{diag}(c_1, \ldots, c_d)$ with $c_j \neq 0$, the new ICs (Cy_t) with the transformation ($W^{-1}C^{-1}$) also fulfill (3). In order to avoid the identification problem, it is suggested to prewhiten x_t and assume y_t to be standardized. The Mahalanobis transformation $\widehat{\text{cov}}(x)^{-1/2}$ does the prewhitening job. It is not hard to see that W becomes then an orthogonal matrix. Denote by $\tilde{x}_t$—the prewhitened x_t and $y_t = \tilde{W}\tilde{x}_t$ the corresponding ICs. $W = \tilde{W}\widehat{\text{cov}}(x)^{-1/2}$ is then the linear transformation for the original observations. For notational convenience, we assume from now on that x_t has been prewhitened.

2.2.2. Order identification

Furthermore, the order of the ICs is ambiguous. Given a permutation matrix P, the ICs (Py_t) fulfill (3) with a new transformation ($W^{-1}P^{-1}$) as well.

2.2.3. IC is necessarily non-Gaussian

Given a d-dimensional standardized Gaussian vector x_t and an orthogonal matrix W, the joint pdfs of x_t and y_t are in fact identical and unrelated with W:

$$
\begin{aligned}
f_x &= |2\pi I_d|^{-1/2} \exp\left(-\frac{x^\top x}{2}\right) \\
&= f_x(W^{-1}y)|\det W^{-1}| \\
&= |2\pi I_d|^{-1/2} \exp\left(-\frac{y^\top y}{2}\right) = f_y.
\end{aligned}
$$

The third condition is although strict, it is naturally fulfilled in financial applications, where financial time series, even after devolatilization, display heavy-tailed behavior.

How can we estimate the linear transformation matrix W after we have prewhitened the non-Gaussian variables? Independence of the components of a random vector $y \in \mathbb{R}^d$ can be measured by the mutual information:

$$
\begin{aligned}
I(W, y) = I(W, f_y) &= \sum_{j=1}^{d} H(y_j) - H(y) \\
&= \sum_{j=1}^{d} H(y_j) - H(x) - \log|\det(W)|,
\end{aligned}
\tag{13}
$$

where $H(y) = H(f_y) = -\int f_y(u) \log f_y(u)\, du$ is the entropy of the vector y with a joint pdf f_y. If the components of y are independent, then the mutual information will reach its minimum with a value of 0. Therefore, the IC searching is identical to minimizing (13) w.r.t. W. Since $H(x)$ is fixed given the data, and the matrix W is orthogonal after prewhitening, this problem is further equivalent to minimizing the term $\sum_{j=1}^{d} H(y_j)$. Now we replace the objective function of the optimization problem by $\min H(y_j)$ w.r.t. w_j, the jth row of W. Note that

$$
\sum_{j=1}^{d} \min H(y_j) \leqslant \min \sum_{j=1}^{d} H(y_j) = \min \sum_{j=1}^{d} H(w_j x).
\tag{14}
$$

This replacement leads to some loss in the W estimation but extensively speeds up the estimation procedure. Moreover, the entropy and negentropy $J(w_j, y_j) = H\{N(0,1)\} - H(y_j)$ are in one-to-one correspondence, we can also formulate the optimization problem as

$$
\hat{w}_j = \arg\min H(y_j) = \arg\max J(w_j, y_j).
$$

The negentropy is always non-negative, since with a fixed variance, the Gaussian random variable has the largest entropy among all distributional types, [5]. Therefore, the negentropy is considered as a non-Gaussian measure and widely used in projection pursuit (PP), see [15]. Furthermore, the PP methods of searching non-Gaussian direction can be applied in the IC identification as well. Nevertheless, compared to the cumulant-based PP method, the entropy or negentropy is less sensitive to outliers and therefore preferable.

A question however raises at this stage, i.e., the marginal pdfs of the ICs in the entropy or the negentropy are unknown. A distributional free approximation of the univariate negentropy has been proposed by [9]

$$J(w_j, y_j) \approx C\{\mathsf{E}\{G(y_j)\} - \mathsf{E}[G\{N(0, 1)\}]\}^2,$$

where C is a constant and G is an even function, e.g., $G(y_j) = (1/\kappa) \log \cosh(\kappa y_j)$, $1 \leqslant \kappa \leqslant 2$.

The estimation of the linear transformation matrix in the ICA is estimated by

$$\hat{w}_j = \operatorname{argmax} \{\mathsf{E}\{G(w_j X)\} - \mathsf{E}[G\{N(0, 1)\}]\}^2 \tag{15}$$

$$\text{s.t.} \quad W^\top W = I_d. \tag{16}$$

Based on the Kuhn–Tucker condition and the Newton's method, Hyvärinen and Oja [12] have proposed the FastICA algorithm. Hyvärinen [10] has shown that the estimates based on the FastICA algorithm are consistent.

FastICA algorithm. *Set $j = 1$.*

1. Choose an initial vector w_j of unit norm, $W = (w_1, \ldots, w_d)^\top$.
2. Let $w_j^{(n)} = \mathsf{E}\{g(w_j^{(n-1)}x)x\} - \mathsf{E}\{g'(w_j^{(n-1)}x)\}w_j^{(n-1)}$, where g denotes the first derivative of $G(y_j)$ and g' the second derivative. In practice, the sample mean is applied for $\mathsf{E}[\cdot]$.
3. Orthogonalization: $w_j^{(n)} = w_j^{(n)} - \sum_{k \neq j}(w_j^{(n)\top} w_k)w_k$.
4. Normalization: $w_j^{(n)} = w_j^{(n)}/\|w_j^{(n)}\|$, $\|\cdot\|$ denotes the norm.
5. If not converged, i.e., $\|w_j^{(n)} - w_j^{(n-1)}\| \neq 0$, go back to 2.
6. Set $j = j + 1$. For $j \leqslant d$, go back to step 1.

3. Simulation study

The reliability of the proposed ICVaR depends on the linear transformation matrix estimation and the univariate modeling on the ICs estimated. In order to fit the local distributions of ICs, we apply the GHADA technique due to its good performance in the simulation and empirical studies in [4]. The target of the simulation study here is to search for ICs and compare the marginal densities of the estimated and generated ICs.

In particular, we pursue an experiment to check the validation of the FastICA approach with normal inverse gaussian (NIG) distributed ICs, where the NIG distribution is a subclass of the GH distribution with the fixed parameter $\lambda = -0.5$, see [2]. We generate $d = 50$ NIG samples with $T = 1000$ observations, i.e., $y_j \sim \text{NIG}(\alpha_j, \beta_j, \delta_j, \mu_j)$ for $j = 1, \ldots, 50$. Without loss of generality, we set $\mu_j = 0$ and $\delta_j = 1$. The parameter α_j is uniformly distributed in $[1, 2]$ and β fulfills the condition:

$$\text{Var}[y_j] = \frac{1}{\sqrt{\alpha_j^2 - \beta_j^2}} \frac{\alpha_j^2}{\alpha_j^2 - \beta_j^2} = 1$$

such that the generated ICs have unit variances. Furthermore, the sign of β is chosen arbitrarily. Table 1 shows the distributional parameters of the generated ICs. The linear transformation matrix W^{-1} is obtained via the Jordan decomposition of a square matrix, whose elements are standard normally distributed. The mixed time series $x_t = W^{-1}y_t$ are analyzed by the ICA.

We apply the FastICA algorithm to the transformed time series x_t and estimate the NIG parameters of each estimated IC. We order the 50 estimated independent series to minimize the mean absolute error (MAE) of the marginal pdfs between the estimated and generated ICs. Overlapping is avoided in the ordering:

$$\text{MAE}_j = \frac{1}{T} \sum_{t=1}^{T=1000} |f(\hat{y}_{jt}) - f(y_{jt})|.$$

The largest two MAEs are 0.09 (ICest 20) and 0.04 (ICest 17), indicating the worst cases of IC searching. In this case, it is expected to get accurate VaR estimations based on these fits.

Table 1
ML estimators $\hat{\alpha}_j$ and $\hat{\beta}_j$ of the estimated ICs, the parameters of the true ICs and the MAE (unit: 10^{-2})

ICest	$\hat{\alpha}$	$\hat{\beta}$	IC	α	β	MAE	ICest	$\hat{\alpha}$	$\hat{\beta}$	IC	α	β	MAE
1	1.89	0.63	23	1.39	0.61	0.63	26	1.84	0.60	2	1.92	1.14	2.56
2	1.76	0.60	25	1.97	1.19	1.57	27	1.82	−0.18	16	1.57	−0.80	2.56
3	1.45	−0.51	30	1.06	−0.21	0.32	28	1.30	−0.37	8	1.17	−0.37	0.32
4	1.64	0.42	36	1.22	0.43	0.52	29	1.34	0.13	17	1.06	0.21	0.68
5	2.11	−0.67	13	1.31	−0.53	0.70	30	2.02	−0.39	44	1.54	0.77	2.73
6	2.05	−0.62	28	1.39	−0.61	0.93	31	3.05	0.80	35	1.78	1.00	3.27
7	1.73	−0.41	50	1.60	−0.83	0.76	32	1.79	−0.45	32	1.34	−0.57	0.77
8	1.62	0.45	7	1.15	0.35	0.45	33	1.58	0.08	4	1.24	0.46	0.88
9	1.27	0.16	6	1.96	1.17	1.58	34	1.59	0.35	47	1.45	0.67	0.74
10	1.51	0.43	27	1.51	−0.74	2.33	35	2.00	−0.33	34	1.43	−0.66	1.18
11	2.44	1.33	38	1.27	0.49	0.98	36	1.63	−0.06	31	1.85	1.07	1.37
12	1.65	−0.25	26	1.35	−0.57	0.79	37	1.81	0.29	29	1.93	1.15	1.29
13	2.58	−1.35	10	1.89	−1.11	1.69	38	1.49	−0.01	22	1.66	−0.89	2.97
14	1.60	0.15	24	1.54	0.77	0.98	39	2.39	−0.30	42	1.53	−0.76	1.27
15	1.52	−0.03	20	1.04	0.18	0.67	40	1.52	−0.22	21	1.39	0.62	1.48
16	2.14	−1.00	1	1.72	−0.94	3.09	41	2.32	−0.001	41	1.70	0.93	3.37
17	2.20	0.64	3	1.75	0.97	4.07	42	2.87	0.30	9	1.50	0.73	1.37
18	1.44	0.48	15	1.40	0.63	2.28	43	2.27	0.24	14	1.28	0.50	1.13
19	1.44	−0.40	33	1.70	0.93	3.22	44	2.44	0.46	43	1.87	1.09	1.39
20	1.90	−0.54	39	1.72	0.95	9.73	45	1.88	−0.07	18	1.44	−0.67	1.23
21	1.57	0.39	11	1.63	0.86	0.72	46	2.22	0.13	37	1.39	0.62	1.24
22	1.80	−0.56	12	1.78	−1.00	0.84	47	1.79	0.26	49	1.45	−0.68	1.41
23	1.69	0.20	46	1.69	0.92	1.09	48	3.03	0.75	40	1.95	1.17	2.37
24	1.76	−0.27	5	1.39	0.62	1.53	49	3.31	−0.18	48	1.50	−0.73	1.49
25	1.50	0.26	19	1.69	−0.92	1.38	50	3.77	0.29	45	1.53	0.76	1.52

4. Empirical study

In this section, we analyze foreign exchange rate portfolios with artificial and time-invariant trading strategies. Since the position of these individual risk factors are constant in time, one can simply use the historical simulation approach where the portfolio returns are considered as the single risk factor. It speeds up VaR computations without losing the portfolio's dependence information. However, in practice, it may slow down the calculation as the trading strategies change, see [16]. Three VaR models are evaluated: the proposed ICVaR approach based on the multivariate risk factors, the RiskMetrics and t-deGARCH methodologies based on the univariate portfolio returns. In the RiskMetrics and t-deGARCH frameworks, we apply the GARCH(1,1) setup to estimate the dependence structure of real data and assume that the devolatilized returns are Gaussian or student-t distributed. The df of the student-t distribution are selected by the maximum likelihood method.

The foreign exchange market, or "Forex" market, is by far the largest financial market in the world with trading volumes surpassing USD 1.5 trillion on some days. The very active buying and selling of traders make it further the most liquid financial market. Here, we consider portfolios including two exchange rates: DEM/USD and the British Pound to the US Dollar (GBP/USD) from 1979-12-01 to 1994-04-01. We forecast VaRs one day ahead w.r.t. four artificial trading strategies, i.e., $b_1 = (1, 1)^\top$, $b_2 = (1, 2)^\top$, $b_3 = (-1, 2)^\top$ and $b_4 = (-2, 1)^\top$. For example, b_1 means holding one unit DEM/USD and one unit GBP/USD forward contracts.

The data is available at FEDC (http://sfb649.wiwi.hu-berlin.de/). Each time series consists of 3721 observations. Table 2 summarizes the statistical properties of the original data and the estimated ICs based on the linear transformation. These four time series are all centered around 0, approximately symmetric but distinctly non-Gaussian indicated by their large kurtoses. The temporal correlations are very small; the ACF plots show that the serial correlations decay at the very beginning lags, indicating a weak stationarity of the four series, see Fig. 3. On the other hand, the cross-correlation of the two exchange returns is over 0.77, referring a strongly linear dependence between them.

Table 2
Descriptive statistics of the log returns and the two estimated independent processes of the DEM/USD and GBP/USD rates

Time series	Mean	SD	Skewness	Kurtosis	ρ_1	ρ_2
DEM/USD	0.00	0.71×10^{-2}	-0.13	4.94	0.02	0.01
GBP/USD	0.00	0.69×10^{-2}	-0.01	5.64	0.08	0.01
IC1	-0.02	1.00	-0.62	8.71	0.07	0.02
IC2	0.01	1.00	-0.08	5.19	0.05	0.01
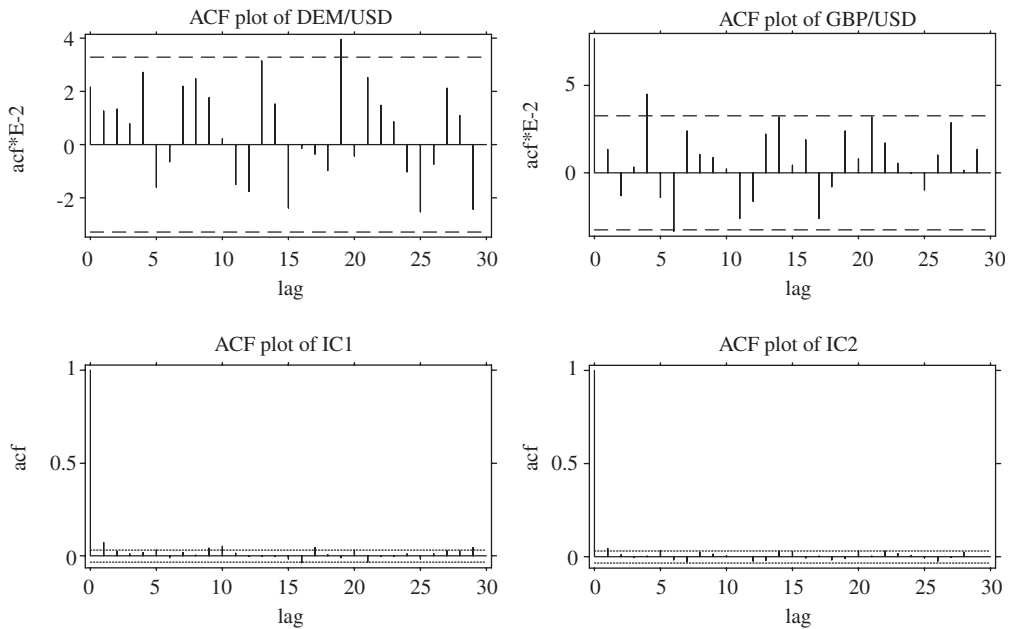
Fig. 3. ACF plots of the log returns of the DEM/USD (left) and the GBP/USD (right) are displayed on the top. Below are the ACF plots of the estimated IC series: IC1 (left) and IC2 (right).

Applying the FastICA algorithm, we estimate the linear transformation matrix and its inverse:

$$\hat{W} = \begin{pmatrix} 207.93 & -213.63 \\ 77.72 & 73.29 \end{pmatrix}, \quad \hat{W}^{-1} = \begin{pmatrix} 2.30 & 6.71 \\ -2.44 & 6.53 \end{pmatrix} 10^{-3}. \tag{17}$$

We then implement the GHADA approach to fit the distributional feature of each estimated IC. Fig. 4 shows the adaptive volatility series based on the two ICs. The well-known volatility clustering is reflected and volatility jumps appear. These jumps happen in most cases at different time and have individual influences on the return processes.

The estimated HYP and NIG parameters of the devolatilized ICs are reported in Table 3. In the Monte Carlo simulation, to find the empirical quantiles of the portfolios, we generate $d = 2$ samples with $M = 10,000$ observations. The daily empirical quantiles at three risk levels $a = 5\%$, 1% and 0.5% are the average values of $N = 100$ repetitions. We redo the simulation for the last $T = 1000$ days. The daily means and SDs of the three empirical quantiles given different trading strategies and two GH subclasses (HYP and NIG) are reported in Tables 4–7. The largest SD of daily empirical quantiles is underlined in each category. The values are small due to the large sample size, indicating efficient estimation of the daily quantiles.

The backtesting results based on the ICVaR, the RiskMetrics and t-deGARCH methodologies are reported in Table 8. The dfs of the student-t fits are 16, 16, 13 and 14 w.r.t. different trading strategies. Based on two
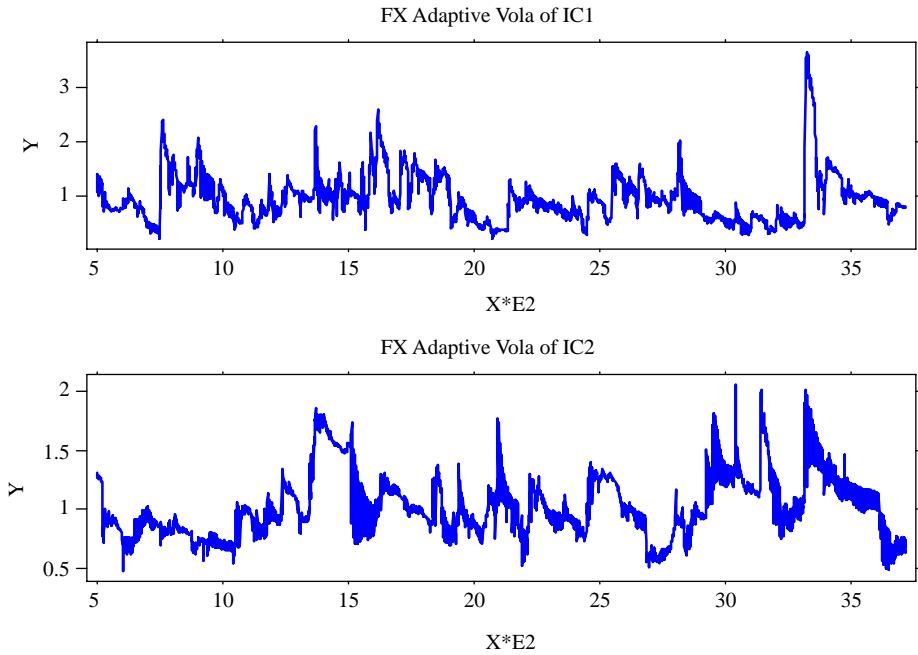
Fig. 4. Adaptive volatility processes of the FX ICs.

Table 3
GH parameters of the ICs

GH type	Time series	$\hat{\alpha}$	$\hat{\beta}$	$\hat{\delta}$	$\hat{\mu}$
HYP	IC1	1.71	−0.17	0.55	0.12
HYP	IC2	1.77	0.02	0.71	−0.01
NIG	IC1	1.22	−0.18	1.10	0.13
NIG	IC2	1.37	0.03	1.28	−0.02

Table 4
Descriptive statistics of daily empirical quantile estimates: $b = (1, 1)^{\top}$, MC simulation with $M = 10,000$, $N = 100$, $T = 1000$

HYP	a 5%		a 1%		a 0.5%		NIG	a 5%		a 1%		a 0.5%	
Day	Mean	SD%	Mean	SD%	Mean	SD%	Day	Mean	SD%	Mean	SD%	Mean	SD%
1	−0.01	0.01	−0.02	0.04	−0.02	0.06	1	−0.01	0.02	−0.02	0.05	−0.02	0.06
2	−0.01	0.01	−0.01	0.04	−0.02	0.05	2	−0.01	0.01	−0.02	0.04	−0.02	0.06
424	−0.03	0.07	−0.06	0.15	−0.07	0.21	423	−0.04	0.06	−0.07	0.15	−0.07	0.23
425	−0.03	0.06	−0.06	0.14	−0.07	0.22	425	−0.04	0.06	−0.07	0.14	−0.07	0.24
426	−0.03	0.04	−0.05	0.11	−0.06	0.18	426	−0.03	0.05	−0.06	0.12	−0.06	0.18
597	−0.03	0.06	−0.06	0.16	−0.06	0.22	607	−0.04	0.06	−0.06	0.15	−0.07	0.21
999	−0.01	0.02	−0.02	0.04	−0.02	0.06	622	−0.03	0.07	−0.06	0.12	−0.07	0.18
1000	−0.01	0.02	−0.02	0.04	−0.02	0.07	1000	−0.01	0.02	−0.02	0.05	−0.02	0.07

Table 5
Descriptive statistics of daily empirical quantile estimates: $b = (1, 2)^\top$, MC simulation with $M = 10,000$, $N = 100$, $T = 1000$

HYP	a 5%		a 1%		a 0.5%		NIG	a 5%		a 1%		a 0.5%	
Day	Mean	SD%	Mean	SD%	Mean	SD%	Day	Mean	SD%	Mean	SD%	Mean	SD%
1	−0.01	0.02	−0.03	0.06	−0.03	0.09	1	−0.01	0.02	−0.03	0.06	−0.03	0.09
2	−0.01	0.02	−0.02	0.05	−0.03	0.08	2	−0.01	0.02	−0.03	0.06	−0.03	0.09
424	−0.05	0.10	−0.09	0.23	−0.10	0.31	325	−0.06	0.11	−0.10	0.22	−0.11	0.30
425	−0.05	0.10	−0.10	0.21	−0.10	0.33	423	−0.06	0.11	−0.10	0.25	−0.11	0.32
597	−0.05	0.09	−0.09	0.24	−0.10	0.33	424	−0.06	0.10	−0.10	0.22	−0.11	0.32
999	−0.01	0.03	−0.03	0.07	−0.03	0.08	597	−0.05	0.09	−0.10	0.22	−0.10	0.34
1000	−0.02	0.03	−0.03	0.06	−0.03	0.10	1000	−0.02	0.03	−0.03	0.08	−0.04	0.11

Table 6
Descriptive statistics of daily empirical quantile estimates: $b = (-1, 2)^\top$, MC simulation with $M = 10,000$, $N = 100$, $T = 1000$

HYP	a 5%		a 1%		a 0.5%		NIG	a 5%		a 1%		a 0.5%	
Day	Mean	SD%	Mean	SD%	Mean	SD%	Day	Mean	SD%	Mean	SD%	Mean	SD%
1	−0.01	0.01	−0.01	0.03	−0.01	0.04	1	−0.01	0.01	−0.01	0.03	−0.02	0.05
2	−0.01	0.01	−0.01	0.03	−0.01	0.05	2	−0.01	0.01	−0.01	0.03	−0.02	0.05
601	−0.03	0.05	−0.06	0.15	−0.06	0.22	601	−0.03	0.05	−0.06	0.15	−0.07	0.19
607	−0.04	0.06	−0.06	0.13	−0.07	0.18	606	−0.03	0.06	−0.06	0.15	−0.07	0.23
611	−0.03	0.06	−0.06	0.13	−0.06	0.20	607	−0.04	0.07	−0.07	0.13	−0.07	0.19
613	−0.03	0.05	−0.06	0.15	−0.07	0.20	613	−0.04	0.06	−0.06	0.12	−0.07	0.18
999	−0.01	0.01	−0.01	0.03	−0.01	0.04	999	−0.01	0.01	−0.01	0.03	−0.01	0.04
1000	−0.01	0.01	−0.01	0.03	−0.01	0.04	1000	−0.01	0.01	−0.01	0.03	−0.01	0.05

Table 7
Descriptive statistics of daily empirical quantile estimates: $b = (-2, 1)^\top$, MC simulation with $M = 10,000$, $N = 100$, $T = 1000$

HYP	a 5%		a 1%		a 0.5%		NIG	a 5%		a 1%		a 0.5%	
Day	Mean	SD%	Mean	SD%	Mean	SD%	Day	Mean	SD%	Mean	SD%	Mean	SD%
1	−0.01	0.01	−0.01	0.03	−0.01	0.04	1	−0.01	0.01	−0.01	0.03	−0.02	0.04
2	−0.01	0.01	−0.01	0.03	−0.01	0.05	2	−0.01	0.01	−0.01	0.03	−0.02	0.05
607	−0.04	0.06	−0.06	0.13	−0.07	0.18	601	−0.03	0.06	−0.06	0.15	−0.07	0.22
608	−0.03	0.05	−0.06	0.13	−0.06	0.19	605	−0.04	0.06	−0.06	0.14	−0.07	0.19
612	−0.03	0.05	−0.06	0.14	−0.07	0.21	612	−0.04	0.07	−0.07	0.14	−0.07	0.19
999	−0.01	0.01	−0.01	0.03	−0.01	0.04	999	−0.01	0.01	−0.01	0.02	−0.01	0.04
1000	−0.01	0.01	−0.01	0.03	−0.01	0.04	1000	−0.01	0.01	−0.01	0.03	−0.02	0.04

likelihood ratio tests: LR1 for risk level and LR2 for exceedances clustering [16], the proposed ICVaR model is superior to the other two candidates. In the risk level test (LR1), the NIG fit performs even better than the HYP fit. In the RiskMetrics framework, the exceedances happen minimal 2.6 times and maximal 23 times more than the expected risk level, e.g., for the trading strategy $b = (1, 1)^\top$. In some cases, the underestimation is even over 25 times. Compared to the RiskMetrics, the t-deGARCH method improves the VaR forecasting as the extreme risk levels such as 0.5% are considered. However, both models are rejected everywhere in the two tests at 99% level, meaning unreliable predictions. Exemplary graphical illustrations of the VaR time plot are displayed in Figs. 5 and 6.

Table 8
Backtesting of the VaR forecast of the exchange portfolios

$b^\top$	a (10^{-2})	ICVaR (HYP fit)			ICVaR (NIG fit)			Risk metrics method			t-deGARCH method		
		$n/T\%$	LR1	LR2	$n/T\%$	LR1	LR2	$n/T\%$	LR1	LR2	$n/T\%$	LR1	LR2
(1, 1)	5	6.7	5.52	[a]27.60	5.8	1.28	[a]28.82	10.8	[a]53.95	[a]19.84	11.0	[a]57.33	[a]26.05
(1, 1)	1	1.9	6.47	[a]16.29	1.4	1.43	[a]11.14	5.5	[a]99.59	[a]21.70	5.3	[a]92.67	[a]22.37
(1, 1)	0.5	0.9	2.59	5.58	0.7	0.71	0.00	5.0	[a]142.33	[a]23.44	4.2	[a]106.16	[a]18.96
(1, 2)	5	6.8	6.16	[a]27.23	5.4	0.32	[a]30.50	10.9	[a]55.63	[a]24.78	11.0	[a]57.33	[a]24.48
(1, 2)	1	1.6	3.07	[a]10.51	1.4	1.43	[a]11.14	5.3	[a]92.67	[a]22.37	4.9	[a]79.30	[a]23.81
(1, 2)	0.5	0.9	2.59	0.00	0.8	1.52	0.00	4.7	[a]128.43	[a]21.77	4.4	[a]114.93	[a]22.83
(−1, 2)	5	6.4	3.80	[a]38.56	5.5	0.51	[a]28.17	12.0	[a]75.40	[a]35.66	12.0	[a]75.40	[a]35.66
(−1, 2)	1	1.4	1.43	[a]11.14	1.2	0.37	[a]11.86	6.0	[a]117.58	[a]25.86	5.5	[a]99.59	[a]23.40
(−1, 2)	0.5	1.2	[a]7.06	[a]11.86	1.1	5.38	[a]15.84	4.4	[a]114.93	[a]18.33	3.2	[a]65.54	[a]20.90
(−2, 1)	5	6.5	4.34	[a]22.35	5.0	0.00	[a]20.34	12.7	[a]89.18	[a]18.41	12.7	[a]89.18	[a]18.41
(−2, 1)	1	1.6	3.07	0.00	1.1	0.09	0.00	5.6	[a]103.12	[a]14.79	4.7	[a]72.87	4.49
(−2, 1)	0.5	0.9	2.59	0.00	0.8	1.52	0.00	4.2	[a]106.16	2.40	3.2	[a]65.54	5.54

MC simulation with $M = 10,000$, $N = 100$, $T = 1000$.

[a] Indicates the model is rejected at 99% level.

Fig. 5. VaR time plots of the exchange rate portfolio with trading strategy $b_4 = (-2, 1)^\top$ and risk level $a = 0.5\%$. The one-day-ahead predicted VaRs are based on the HYP distribution. The dots are the real portfolio returns, the RiskMetrics (red) and the t-deGARCH (green) forecasts are displayed as solid line and circles while the ICVaR is displayed as a straight line. The exceedances w.r.t. the risk management models are displayed as cross on the bottom (from the top are the exceedances w.r.t. the ICVaR, the t-deGARCH and the RiskMetrics).

5. Conclusion

In this paper, we proposed an easy and fast multivariate risk management model. The study is mainly based on the ICA. Instead of estimating the joint density and covariance of high-dimensional returns, the searching of ICs transfers the calculation to unidimensional studies. In the empirical study, the proposed ICVaR is superior to the RiskMetrics and t-deGARCH methods, above all in the risk level controlling. In addition, in the ICVaR methodology, the joint distribution of portfolio does not rely on trading strategy and therefore can be further applied to calculate VaRs as the
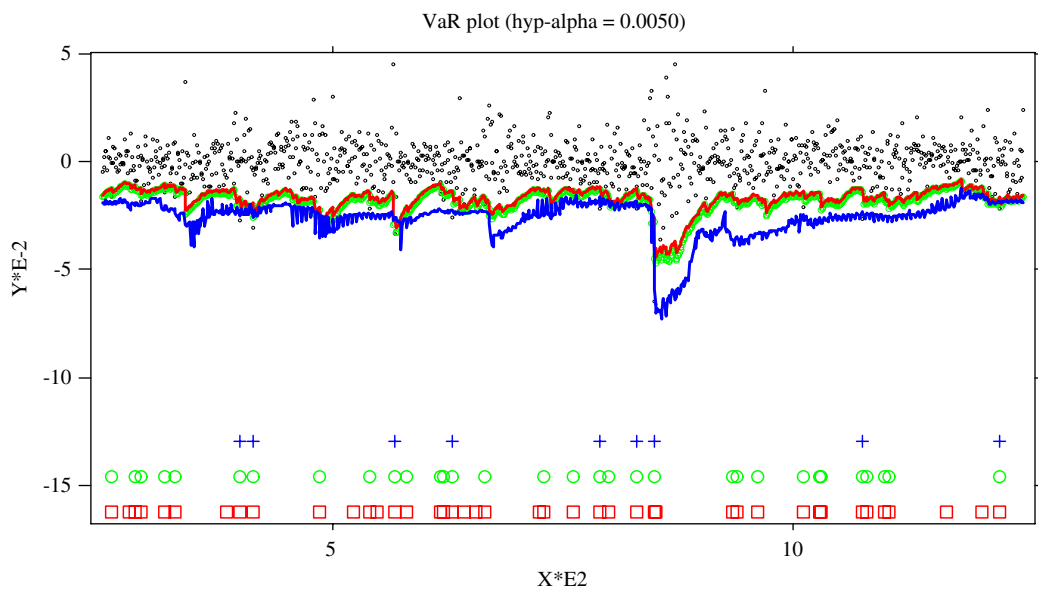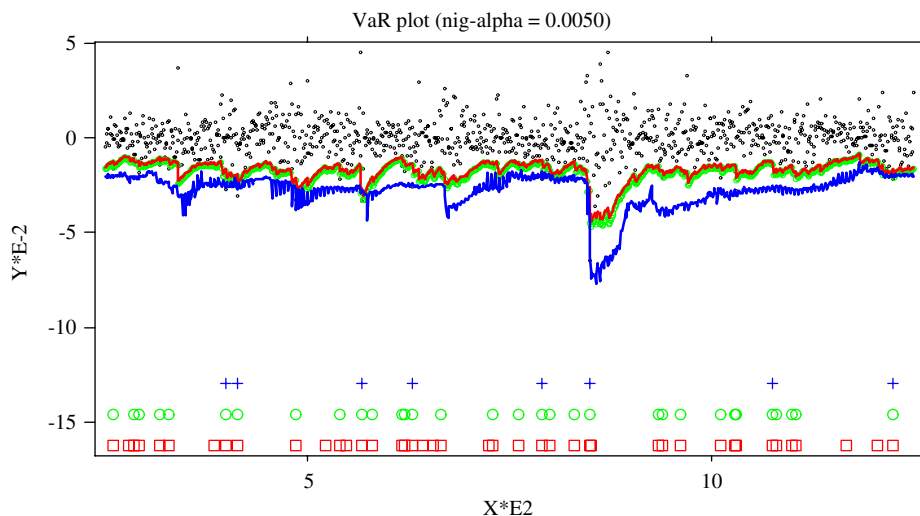
Fig. 6. VaR time plots of the exchange rate portfolio with trading strategy $b_4 = (-2, 1)^\top$ and risk level $a = 0.5\%$. The one-day-ahead predicted VaRs are based on the NIG distribution. The dots are the real portfolio returns, the RiskMetrics (red) and the t-deGARCH (green) forecasts are displayed as solid line and circles while the ICVaR is displayed as a straight line. The exceedances w.r.t. the risk management models are displayed as cross on the bottom (from the top are the exceedances w.r.t. the ICVaR, the t-deGARCH and the RiskMetrics).

investing positions change. Moreover, the ICVaR approach can be easily applied to calculate and forecast other risk measures such as expected shortfall.

Acknowledgement

References

[1] A.D. Back, A.S. Weigend, A first application of independent component analysis to extracting structure from stock returns, Internat. J. Neural Systems 8 (1998) 473–484.

[2] O.E. Barndorff-Nielsen, P. Blæsild, Hyperbolic distribution and ramifications: Contributions to theory and applications, in: C. Taillie, P.G. Patil, A.B. Baldessari (Eds.), Statistical Distributions in Scientific Work, vol. 4, D. Reidel, Dordrecht, 1981, pp. 19–44.

[3] T.P. Bollerslev, Modelling the coherence in short-run nominal exchange rates: a multivariate generalized ARCH model, Rev. Econom. Statist. 72 (1990) 498–505.

[4] Y. Chen, W. Härdle, S.O. Jeong, Nonparametric risk management with generalized hyperbolic distributions. SFB 649, discussion paper 2005-001.

[5] T.M. Cover, J.A. Thomas, Elements of Information Theory, Wiley, New York, 1991.

[6] J.R. Duann, T.P. Jung, W.J. Kuo, T.C. Yeh, S. Makeig, J.C. Hsieh, T.J. Sejnowski, Single-trial variability in event-related bold signals, NeuroImage 15 (2002) 823–835.

[7] R. Engle, Dynamic conditional correlation—a simple class of multivariate GARCH models, J. Business Econom. Statist. 20 (3) (2002) 339–350.

[8] R.F. Engle, K.K. Sheppard, Theoretical and empirical properties of dynamic conditional correlation multivariate GARCH, NBER Working Paper 8554, 2001.

[9] A. Hyvärinen, New Approximations of Differential Entropy for Independent Component Analysis and Projection Pursuit, MIT Press, Cambridge, MA, 1998 pp. 273–279.

[10] A. Hyvärinen, Fast and robust fixed-point algorithms for independent component analysis, Neural Networks 10 (1999) 626–634.

[11] A. Hyvärinen, J. Karhunen, E. Oja, Independent Component Analysis, Wiley, New York, 2001.

[12] A. Hyvärinen, E. Oja, Independent component analysis: algorithms and applications, Neural Networks 13 (1999) 411–430.

[13] S. Iyengary, M. Mazumdar, A saddle point approximation for certain multivariate tail probabilities, SIAM J. Sci. Comput. 19 (4) (1998) 1234–1244.

[14] S. Jaschke, Y. Jiang, Approximating value at risk in conditional Gaussian models, in: W. Härdle, T. Kleinow, G. Stahl (Eds.), Applied Quantitative Finance, Springer, Berlin, 2002.

[15] M.C. Jones, R. Sibson, What is projection pursuit?, J. Roy. Statist. Soc. A 150 (1) (1987) 1–36.

[16] P. Jorion, Value at Risk, McGraw-Hill, New York, 2001.

[17] D. Mercurio, V. Spokoiny, Statistical inference for time-inhomogeneous volatility models, Ann. Statist. 12 (2004) 577–602.

[18] K. Prause, The generalized hyperbolic model: estimation, financial derivatives and risk measures, Dissertation, 1999.

[19] T. Ristaniemi, K. Raju, J. Karhunen, Jammer mitigation in DS-CDMA using independent component analysis, in: Proceedings of the IEEE International Conference on Communications, 2002.

Calibration Risk for Exotic Options

K. Detlefsen and Wolfgang K. Härdle

K. Detlefsen
is a PhD candidate in finance at the Center for Applied Statistics and Economics, (CASE), Humboldt-Universität zu Berlin in Berlin, Germany.
detlefsen@wiwi.hu-berlin.de

Wolfgang K. Härdle
is professor of statistics and director of CASE at Humboldt-Universität zu Berlin in Berlin, Germany.
haerdle@wiwi.hu-berlin.de

Equity derivative pricing models are calibrated to market data of plain vanilla options by minimization of an error functional. From the economic viewpoint, there are several possibilities to measure the error between the market and the model. These different specifications of the error give rise to different sets of calibrated model parameters and the resulting prices of exotic options vary significantly.

We provide evidence for this calibration risk in a time series of DAX implied volatility surfaces from April 2003 to March 2004. We analyze factors influencing these price differences for exotic options in the Heston and in the Bates models and recommend an error functional. Moreover, we determine the model risk of these two stochastic volatility models for the time series and compare it to calibration risk.

Recently, there has been considerable interest, both from a practical and a theoretical point of view, in the risks involved in option pricing. Schoutens et al., [2004] have analyzed model risk in an empirical study and Cont [2004] has put this risk into a theoretical framework. Another source of risk is hidden in the calibration of models to market data. This calibration risk is also fundamental for the banking industry because it significantly influences the prices of exotic options. Moreover, calibration risk exists even if an appropriate model has been chosen and model risk does not exist anymore.

Calibration risk arises from the different possibilities to measure the error between the observations on the market and the corresponding quantities in the model world. A natural approach to specify this error is to consider the absolute price (AP) differences. (See e.g., Schoutens et al., [2004]). But the importance of absolute price differences depends on the magnitude of these prices. Hence, another useful way for measuring the error is relative price (RP) differences. (See e.g., Mikhailov et al., [2003]). As models are often judged by their capability to reproduce implied volatility surfaces, other measures can be defined in terms of implied volatilities. There are again the two possibilities of absolute implied volatilities (AI) and relative implied volatilities (RI). We consider these four ways to measure the difference between model and market data and explore the implications for the pricing of exotic options.

To this end, we focus on the stochastic volatility model of Heston. In order to analyze the influence of the goodness of fit on calibration risk, we consider in addition the Bates model, which is an extension of the Heston model with similar qualitative features. These two models are calibrated to the prices of plain vanilla options on the DAX. We use a time series of implied volatility surfaces from April 2003 to March 2004. As exotic options we consider down and out puts, up and out calls, and cliquet options for one, two, or three years to maturity.

In this framework we determine the size of calibration risk and analyze factors influencing it.

Besides calibration risk, there is also model risk, which represents wrong prices because a wrong parametric model has been chosen. We consider the model risk between the Heston and the Bates models and analyze the relation between the two forms of risk in pricing exotic options.

The next section introduces the models and describes their risk neutral dynamics that we use for option pricing. Moreover, this section contains information about the data used for the calibration. The following section describes the calibration method and defines the error functionals analyzed in this work. The goodness of fit is shown by representative surfaces and statistics on the errors. We then present the exotic options that we consider for calibration risk and price these products by simulation. In our next section we analyze the model risk for the two stochastic volatility models under the four error functionals. In the last section, we summarize the results and draw our conclusions.

MODELS AND DATA

In this section, we describe briefly the Heston model and the Bates model for which we are going to analyze calibration risk. Moreover, we provide some descriptive statistics of the implied volatility surfaces that we use as input data for the calibration.

Heston Model

We consider the popular stochastic volatility model of Heston [1993]:

$$\frac{dS_t}{S_t} = \mu dt + \sqrt{V_t} dW_t^1 \tag{1}$$

where the volatility process is modelled by a square-root process:

$$dV_t = \xi(\eta - V_t)dt + \theta\sqrt{V_t} dW_t^2 \tag{2}$$

and W^1 and W^2 are Wiener processes with correlation ρ.

The variance process (V_t) remains positive if its volatility θ is small enough with respect to the product of the mean reversion speed ξ and the average variance level η:

$$\xi\eta > \frac{\theta^2}{2} \tag{3}$$

The dynamics of the price process are analyzed under a martingale measure under which the characteristic function of (S_t) is given by:

$$\phi_t^H(z) = \exp\left\{\frac{-(z^2 + iz)V_0}{\gamma(z)\coth\frac{\gamma(z)t}{2} + \xi - i\rho\theta z}\right\}$$
$$\times \frac{\exp\{\frac{\xi\eta t(\xi - i\rho\theta z)}{\theta^2} + iztr + iz\log(S_0)\}}{(\cosh\frac{\gamma(z)t}{2} + \frac{\xi - i\rho\theta z}{\gamma(z)}\sinh\frac{\gamma(z)t}{2})^{\frac{2\xi\eta}{\theta^2}}} \tag{4}$$

where $\gamma(z) \stackrel{\text{def}}{=} \sqrt{\theta^2(z^2 + iz) + (\xi - i\rho\theta z)^2}$, see e.g., cont et al. [2004].

Bates Model

Bates [1996] extended the Heston model by considering jumps in the stock price process:

$$\frac{dS_t}{S_t} = \mu dt + \sqrt{V_t} dW_t^1 + dZ_t$$
$$dV_t = \xi(\eta - V_t)dt + \theta\sqrt{V_t} dW_t^2 \tag{5}$$

where Z is a compound Poisson process with intensity λ and jumps k that have a lognormal distribution:

$$\log(1 + k) \sim N\left(\log(1 + \bar{k}) - \frac{\delta^2}{2}, \delta^2\right) \tag{6}$$

We analyze the dynamics of this model under a martingale measure under which the characteristic function of (S_t) is given by:

$$\phi_t^B(z) = \exp\{t\lambda(e^{-\delta^2 z^2/2 + i\{\log(1+\bar{k}) - \frac{1}{2}\delta^2\}z} - 1)\}$$
$$\times \exp\left\{\frac{-(z^2 + iz)V_0}{\gamma(z)\coth\frac{\gamma(z)t}{2} + \xi - i\rho\theta z}\right\}$$
$$\times \frac{\exp\{\frac{\xi\eta t(\xi - i\rho\theta z)}{\theta^2} + izt(r - \bar{k}) + iz\log(S_0)\}}{(\cosh\frac{\gamma(z)t}{2} + \frac{\xi - i\rho\theta z}{\gamma(z)}\sinh\frac{\gamma(z)t}{2})^{\frac{2\xi\eta}{\theta^2}}} \tag{7}$$

where $\gamma(z) \stackrel{\text{def}}{=} \sqrt{\theta^2(z^2 + \mathbf{i}z) + (\xi - \mathbf{i}\rho\theta z)^2}$. (See e.g., Cont et al. [2004].)

The Bates model has eight parameters while the Heston model has only five parameters. Because of these three additional parameters, the Bates model can better fit an observed surface but parameter stability is more difficult to achieve.

Data

Our data consists of EUREX-settlement implied volatilities on the DAX. Hence, these volatilities are given by inversion of the Black-Scholes formula, i.e., if you plug these volatilities into the Black–Scholes formula together with the other corresponding parameters then you get the settlement prices of European options on the DAX. In this context, we approximate the risk-free interest rates by the EURIBOR. On each trading day we use the yields corresponding to the maturities of the implied volatility surface. As the DAX is a performance index, it is adjusted to dividend payments. Thus, we do not consider dividend payments explicitly.

We analyze the time period from April 2003 to March 2004. Since March 2003 the EUREX trades plain vanilla options with maturities up to five years. Until March 2004, it has not changed its range of products. Hence, the data is homogeneous in the sense that the implied volatility surfaces are derived from similar products.

From this time period, we analyze the surfaces from all the Wednesdays when trading has taken place. We restrict ourselves to these days because of the computationally intense Monte Carlo simulations for the pricing. Thus, we consider 51 implied volatility surfaces. We exclude observations that are deep out of the money because of illiquidity of these products. More precisely, we consider only options with moneyness $K/S_0 \in [0.75, 1.35]$ for small times to maturity $T \leq 1$. As we analyze exotic options that expire in one, two, or three years we exclude also plain vanillas with time to maturity less than three months.

Some information about the resulting implied volatility surfaces is summarized in Exhibit 1. The surfaces contain on average 140 prices and nine times to maturity with a mean moneyness range of 65%.

The values of the underlying in the sample period are shown in Exhibit 2. This exhibit also shows the (interpolated) at the money implied volatilities for one year to maturity. The market value of the DAX went up in this period and accordingly the implied volatilities went down as Exhibit 2 shows.

CALIBRATION

In this section, we specify the calibration routine and describe the four error functionals. The calibration results illustrate how well the plain vanilla prices can be replicated by the Heston and the Bates models.

Calibration Method

Carr and Madan [1999] found a representation of the price of a European call option by one integral for a whole class of option pricing models. Their method, that

EXHIBIT 1

	mean number of maturities	mean number of observations	mean money-ness range
short maturities (0.25 ≤ T < 1.0)	3.06	64	0.553
long maturities (1.0 ≤ T)	5.98	76	0.699
total	9.04	140	0.649

Note: Description of the implied volatility surfaces.

EXHIBIT 2
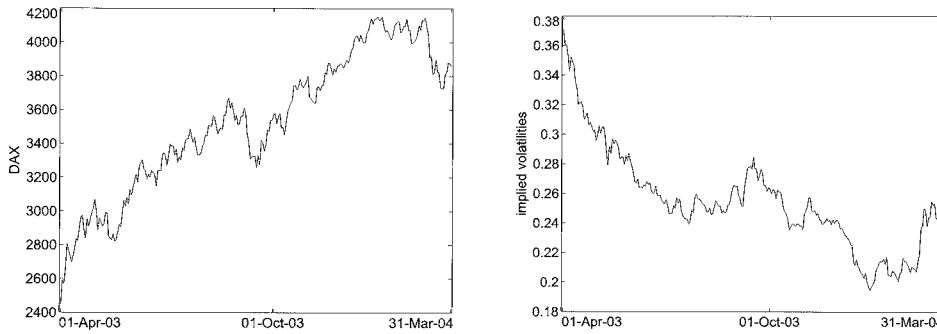
Note: DAX and ATM implied volatility with 1 year to maturity on the trading days from 01 April 2003 to 31 March 2004.

is applicable to the Heston and the Bates model, is based on the characteristic function of the log stock price under the risk neutral measure.

Carr and Madan showed that the price $C(K,T)$ of a European call option with strike K and maturity T is given by

$$C(K,T) = \frac{\exp\{-\alpha \ln(K)\}}{\pi}$$
$$\int_0^{+\infty} \exp\{-i\nu \ln(K)\}\psi_T(\nu)d\nu \tag{8}$$

for a (suitable) damping factor $\alpha > 0$. The function ψ_T is given by

$$\psi_T(\nu) = \frac{\exp(-rT)\phi_T\{\nu - (\alpha+1)i\}}{\alpha^2 + \alpha - \nu^2 + i(2\alpha+1)\nu} \tag{9}$$

where ϕ_T is the characteristic function of $\log(S_T)$; see the prior section on Models and Data.

For the difference between market and model, we consider the following four objective functions based on the root weighted square error:

$$AP \overset{def}{=} \sqrt{\sum_{i=1}^{n} w_i \left(P_i^{mod} - P_i^{mar} \right)^2}$$

$$RP \overset{def}{=} \sqrt{\sum_{i=1}^{n} w_i \left(\frac{P_i^{mod} - P_i^{mar}}{P_i^{mar}} \right)^2}$$

$$AI \overset{def}{=} \sqrt{\sum_{i=1}^{n} w_i \left(IV_i^{mod} - IV_i^{mar} \right)^2}$$

$$RI \overset{def}{=} \sqrt{\sum_{i=1}^{n} w_i \left(\frac{IV_i^{mod} - IV_i^{mar}}{IV_i^{mar}} \right)^2} \tag{10}$$

where *mod* refers to a model quantity and *mar* to a quantity observed on the market, P to a price, and IV to an implied volatility. The index i runs over all n_t observations of the surface on day t. The weights w_i are non negative with $\Sigma_i w_i = 1$. Hence, the objective functions can be interpreted as mean average errors. While the error functionals AP and RP measure the differences between option prices, the other two error measures focus on Black-Scholes implied volatilities because these quantities are normally used in reality for price quotations. The model implied volatilities IV^{mod} are computed from the prices of European options in the Heston and in the Bates models by numerical inversion of the Black-Scholes formula.

We choose the weights in such a way that on each day all maturities have the same influence on the objective function. In order to make different surfaces comparable, each maturity gets the weight $1/n_{mat}$ where n_{mat} denotes the number of maturities in this surface. Moreover, we assign the same weight to all points of the same maturity. This leads to the weights

$$w_i \overset{def}{=} \frac{1}{n_{mat} n_{str}^i} \tag{11}$$

where n_{str}^i denotes the number of strikes with the same maturity as observation i. This weighting leads asymptotically to a uniform density on each maturity.

Given these weights, the average time to maturity of an implied volatility surface can be measured by a modified duration:

$$\sum_{i=1}^{n} \frac{\tau_i w_i}{\sum_{i=1}^{n} w_i} \tag{12}$$

where τ_i is the time to maturity of the option i. The mean duration of the 51 surfaces is 2.02 and the minimal (maximal) is 1.70 (2.30). Thus, the point of balance for the maturities lies around 2 for our time series of surfaces. As we analyze exotic options with one, two or three years to maturity, this point of balance confirms a correct weighting for our purposes.

We consider only out-of-the-money prices. Thus, we use call prices for strikes higher than the spot and put prices for strikes below the spot. This approach ensures that we compare only prices of similar magnitude. It has no impact on the errors based on implied volatilities. Because of put-call parity, the use of OTM options has no impact on the absolute price error (AP). But the relative prices are weighted in such a way that the observations around the spot receive less weight. Hence, only the relative price error (RP) is influenced by this choice of prices.

In order to estimate the model parameters, we apply a stochastic global optimization routine and minimize the objective functions with respect to the model parameters. In addition to some natural constraints on the range of the parameters, we have taken into account inequality (3) that ensures the positivity of the variance process.

Calibration Results

We consider 51 implied volatility surfaces between April 2003 and March 2004. Each of these is calibrated with respect to the four error functions described in our prior sub-section. These calibrations are done for the Heston and the Bates models.

The resulting errors of these 408 calibrations have been summarized in Exhibit 3 for the Heston model and in Exhibit 4 for the Bates model. Descriptive statistics on the calibrated parameters are given in Exhibit 5 for the Heston model and in Exhibit 6 for the Bates model. Exhibit 7 shows the fit of the implied volatility surface in the Heston model on a day that is representative for the AI error.

Exhibit 3 reports in each line the means of the four errors when the objective function given in the left column is minimized. In the Heston model, we get a mean absolute price error of 7.3 and a mean relative price error of 9.7% when we calibrate with respect to AP. Using the RP error functional we get the opposite result with a mean absolute price error of 11 and a mean relative price error of 6.1%. The errors based on implied volatilities are smaller for the RP objective function than for

the AP objective function. The results for the AI and RI objective functionals differ only slightly: the mean absolute implied volatility error is about 0.68% and the mean relative implied volatility error is about 2.5%. Moreover, the price errors for these objective functions lie between the price errors of the other two objective functions. The calibration with respect to RI gives the best overall fit because it has the smallest RI error and the second best errors for the rest. The meaning of these numbers is illustrated by Exhibit 7, which shows an implied volatility fit that is representative for an AI error of 0.68%. In order to make the AP errors comparable for different days (with different values of the spot) we have computed the mean of AP/DAX as 0.21%, 0.34%, 0.27%, 0.25% for the four error functionals.

The calibrated parameters which are described by Exhibit 5 form two groups because the parameters for the RP, AI, and RI calibration are quite similar. The initial variance V_0 and the average variance level η are both about 0.07 for all objective functionals. For the AP calibration we get a reversion speed $\xi = 0.9$, a volatility of variance of $\theta = 0.34$, and a correlation $\rho = -0.82$. The other calibrations lead to similar parameters with a reversion speed $\xi = 1.3$, a volatility of volatility of $\theta = 0.44$, and a correlation $\rho = -0.75$. As the correlations are sig-

EXHIBIT 3

objective fct.	AP	RP	AI	RI
AP	7.3	0.097	0.0081	0.031
	(2.2)	(0.048)	(0.0025)	(0.012)
RP	11.	0.061	0.0074	0.029
	(4.6)	(0.025)	(0.0022)	(0.010)
AI	9.4	0.073	0.0068	0.026
	(3.2)	(0.032)	(0.0020)	(0.009)
RI	8.8	0.070	0.0070	0.025
	(2.9)	(0.030)	(0.0021)	(0.009)

Note: Mean calibration errors in the Heston model for 51 days.
(AP = absolute price differences, RP = relative price differences, AI = absolute implied volatility differences, RI = relative implied volatility differences)

EXHIBIT 4

objective fct.	AP	RP	AI	RI
AP	7.0	0.130	0.0076	0.028
	(2.2)	(0.100)	(0.0023)	(0.009)
RP	12	0.051	0.0067	0.026
	(4.8)	(0.019)	(0.0020)	(0.009)
AI	8.9	0.064	0.0060	0.023
	(3.2)	(0.026)	(0.0018)	(0.008)
RI	8.7	0.062	0.0062	0.022
	(3.2)	(0.028)	(0.0020)	(0.008)

Note: Mean calibration errors in the Bates model for 51 days. (AP = absolute price differences, RP = relative price differences, AI = absolute implied volatility differences, RI = relative implied volatility differences)

EXHIBIT 5

	ξ	η	θ	ρ	V_0
AP	0.87	0.07	0.34	-0.82	0.07
	(0.48)	(0.02)	(0.08)	(0.08)	(0.02)
RP	1.38	0.07	0.44	-0.74	0.08
	(0.35)	(0.02)	(0.06)	(0.03)	(0.02)
AI	1.32	0.07	0.43	-0.77	0.08
	(0.40)	(0.02)	(0.06)	(0.04)	(0.02)
RI	1.20	0.07	0.41	-0.75	0.08
	(0.35)	(0.02)	(0.06)	(0.05)	(0.02)

Note: Mean parameters (std. dev.) in the Heston model for 51 days.

nificantly greater than −1, the calibrated Heston models have really two stochastic factors.

The Bates model exhibits similar qualitative results as the Heston model: the AP and the RP calibrations differ clearly while the AI and the RI calibrations lead to similar results. The Bates model can be regarded as an extension of the Heston model. The additional three parameters for the jumps in the spot process lead to better calibration results for all error functionals: the AP error is reduced (on average) by 4%, the RP error by 16%, the AI and the RI error both by 12%.

The calibrated parameters of the Bates model are given in Exhibit 6. As in the Heston model, they form two groups with the AP calibration on the one hand and the RP, AI, and RI calibrations on the other hand. The parameters ξ, η, θ, and V_0 are similar to the calibrations for the Heston model. Only the correlation ρ rises to a level of −0.93 for all objective functions. Hence, this criterion for distinguishing between the two groups disappears. It is replaced by the expected number of jumps per year: for the AP calibration we expect (on average) a jump every three years while we expect a jump every two years for the other calibrations. It is interesting that all calibrations lead to a mean jump up of about +8% for the returns. The expected jumps upwards correspond to the market going up as shown in Exhibit 2.

Schoutens et al. [2004] found that the Heston and the Bates option models can both be calibrated well to the EuroStoxx50. In summarizing the results of this section, we can say that DAX implied volatility surfaces can be replicated by these models for different error functionals as well as the EuroStoxx50 data in Schoutens et al. [2004]. As in that work, we find that the Bates model gives only slightly better fits for the AP calibration. In addition we have shown that it leads to a considerable improvement in the fit for the other objective functions.

EXOTIC OPTIONS

In this section, we analyze the price differences of exotic options for calibrations with respect to different error measures. We consider barrier and cliquet options. The prices of these products are calculated by Monte Carlo simulations using Euler discretizations.

Simulation

We price all exotic options by Monte Carlo simulations. To this end, we use for each derivate one million

EXHIBIT 6

	ξ	η	θ	ρ	V_0	λ	$\bar{k}$	δ
AP	0.92	0.07	0.33	-0.94	0.07	0.33	0.07	0.08
	(0.50)	(0.02)	(0.08)	(0.07)	(0.02)	(0.21)	(0.03)	(0.06)
RP	1.56	0.07	0.45	-0.89	0.08	0.54	0.05	0.08
	(0.47)	(0.02)	(0.07)	(0.07)	(0.02)	(0.23)	(0.03)	(0.06)
AI	1.43	0.07	0.43	-0.95	0.07	0.50	0.06	0.09
	(0.44)	(0.02)	(0.06)	(0.06)	(0.02)	(0.22)	(0.03)	(0.04)
RI	1.36	0.07	0.41	-0.93	0.07	0.52	0.05	0.08
	(0.44)	(0.02)	(0.07)	(0.09)	(0.02)	(0.26)	(0.04)	(0.08)

Note: Mean parameters (std. dev.) in the Bates model for 51 days.

paths generated by Euler discretization. (See e.g., Glasserman [2004]). As we take into account the positivity constraint (3) the square root process for the variance can be simulated by truncation at zero

$$V_{t_{i+1}} = \left(V_{t_i} + \xi(\eta - V_{t_i})\Delta_t + \theta\sqrt{V_{t_i}}\sqrt{\Delta_t}Z_i\right)^+$$

where Δt is the time step and Z_i are independent standard normal variables. This simple scheme leads to an acceptable small bias when the positivity constraint is fulfilled. For each exotic option we consider three maturities: 1 year, 2 years, and 3 years. We analyze three exotic options: up and out calls, down and out puts, and cliquet options. These products are described in the following sections where remaining parameters are also specified.

The payoffs of barrier options depend only on whether the underlying price process exceeded the barrier in some time interval. Hence, the value of barrier options depends on the minimum or maximum of the underlying price process. We approximate such continuous extrema by discrete extrema using one observation for each trading day. We use 252 time steps to simulate a process for a year assuming 252 trading days a year.

The calibration results are presented in the following sections together with a discussion of the options. The accuracy of the Monte Carlo results is given by the relative standard error in Exhibit 8. This exhibit confirms that the estimators have sufficiently small variance after one million paths compared to the price differences that we observe in Exhibits 10, 13, and 16.

Barrier Options

We consider two types of barrier options: up and out calls and down and out puts. These options are quite popular on the market. Down and out puts are, for example, sold together with zero-strike calls as "bonus certificates." These structured products are actively traded in Germany and also in many other markets.

Up and out call options The prices of up and out calls with strike K, barrier B, and maturity T on an underlying (S_t) are given by

$$\exp(-rT)\, E[1_{\{M_T < B\}}(S_T - K)^+] \tag{13}$$

where

Kai Detlefsen and Wolfgang K. Härdle: Calibration Risk for Exotic Option

Exhibit 7
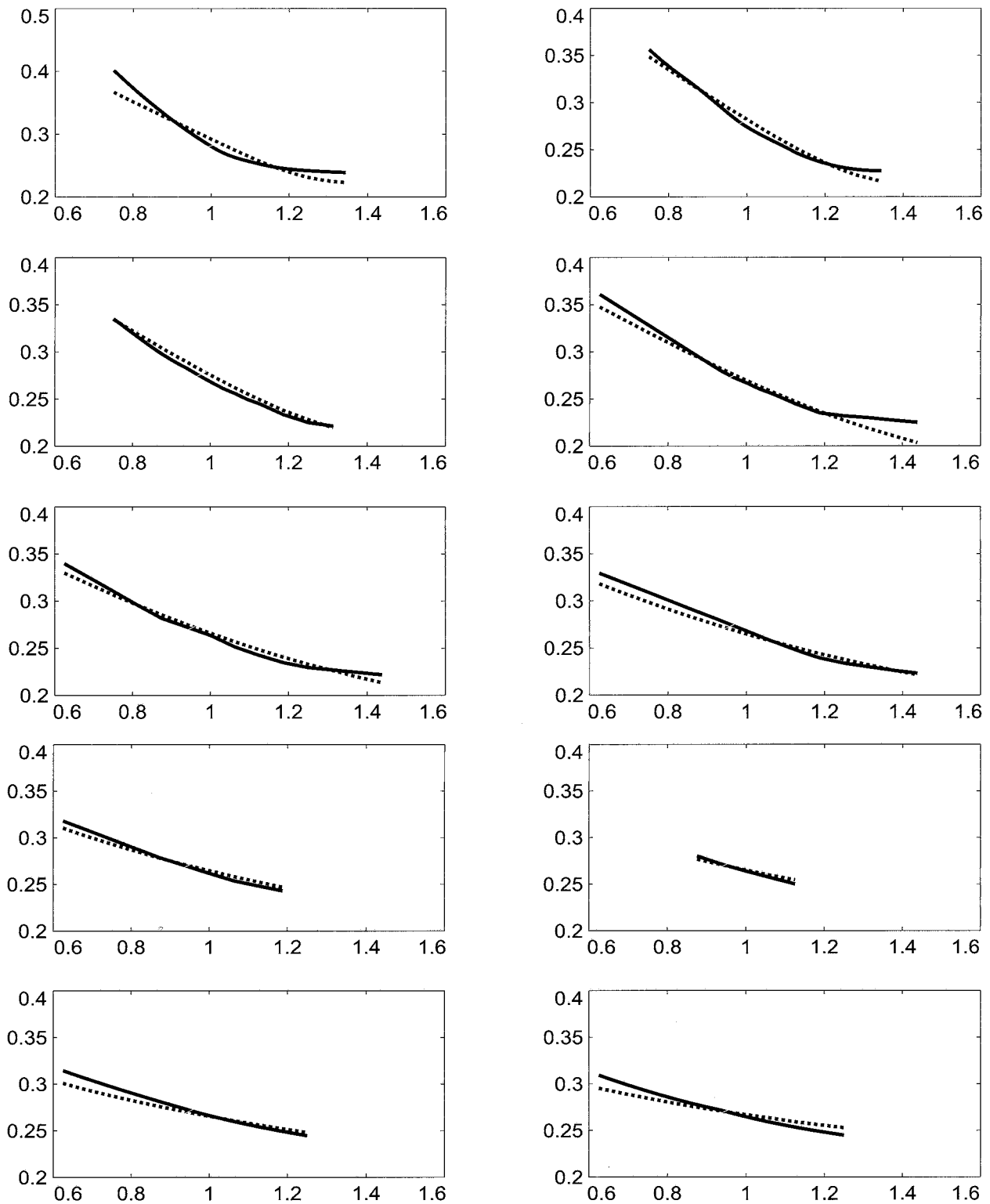
Note: Implied volatilities in the market and in the Heston model for the maturities 0.26, 0.52, 0.78, 1.04, 1.56, 2.08, 2.60, 3.12, 3.64, 4.70 (left to right, top to bottom) for AI parameters on 25/06/2003. Solid: model, dotted: market. X-axis: moneyness.

$$M_T \overset{\text{def}}{=} \max_{0 \le t \le T} S_t \tag{14}$$

We choose as strike K and barrier B

$$K = (1 - 0.1T)S_0 \tag{15}$$

$$B = (1 + 0.2T)S_0 \tag{16}$$

where T denotes time to maturity. Up and out calls with such strikes and barriers are widely traded on the German market.

Up and out call options have the payoff profile of European call options if the underlying has not exceeded the barrier. Otherwise their payoff is zero. Thus, up and out calls are path dependent exotic options.

We want to analyze the difference between the prices of the exotic options when the underlying model has been calibrated with respect to different error measures. To this end, we have calibrated the Heston and the Bates models to implied volatility or price data on each day with respect to the four error functionals introduced in the prior sub-section on Calibration Method. Hence, we have four time series of calibrated models parameters that are described in the sub-section on Calibration Results. By Monte Carlo simulations we calculate on each day the prices of up and out calls for the four sets of model para-

meters. In this way, we get four time series of up and out call prices corresponding to the four error measures for the Heston model and four corresponding time series of prices for the Bates model. We are interested in how the prices of exotic options differ when the four different error functionals are used. If the prices of up and out calls that are computed from the AP parameters are denoted by P_{AP}^{UOC} and the corresponding prices from the RP parameters by P_{AP}^{UOC} then we measure the difference between these prices by the ratio $P_{AP}^{UOC}/P_{AP}^{UOC}$. The other five price differences are measured by corresponding price ratios. Hence, we observe on each day six price ratios that describe the differences of the prices of exotic options resulting from different error measures.

The six time series of price ratios are summarized in Exhibit 9 for up and out calls with three years to maturity in the Heston model. In the boxplots the central line gives the median and the box contains 50% of the observations. Hence, the AP prices lie on average about 6% over the other prices and the AP prices are on 75% of the 51 days at least 4% higher than the other prices. The RP prices are about 2% below the AI or RI prices, which are very similar to each other.

We analyze the influence of time to maturity on these price ratios by also considering one year and two years to maturity (and by adjusting the barrier and the strike appropriately). The medians of the price ratios are presented in Exhibit 10 for all three times to maturity. This exhibit

EXHIBIT 8

	Heston			Bates		
	$T = 1$	$T = 2$	$T = 3$	$T = 1$	$T = 2$	$T = 3$
up and out calls	0.002	0.001	0.001	0.002	0.001	0.001
down and out puts	0.002	0.001	0.001	0.002	0.001	0.001
cliquet options	0.001	0.001	0.001	0.001	0.001	0.001

Note: Maximal ratio of standard error and price in Monte Carlo simulations. (Maximum over all time points and all objective functions.) Up and out calls with strike $1 - 0.1T$ and barrier $1 + 0.2T$ relative to initial spot, down and out puts with strike $1 + 0.1T$ and barrier $1 - 0.2T$ relative to initial spot, cliquets with reset times $t_i = Ti/3 (i = 0, \ldots, 3)$ and local caps of 8% and local floors of -8%. (T denotes time to maturity).

shows that the price differences become smaller for shorter times to maturity for the AP prices. The other price ratios remain almost constant. For one year to maturity, the price differences are about 2% – 3% and the AP prices are lower than the other prices. For two years to maturity, the AP prices are again higher than the other prices.

In order to analyze the influence of the goodness of fit on the price ratios, we consider also the Bates model. The boxplots of the price ratios in this model are given in Exhibit 11 for three years to maturity. Compared to the Heston boxplots, the boxes are longer in the Bates model. Thus there is more variation between the prices for different error functionals. Moreover, the median ratios between the AP prices and the other prices are bigger than in the Heston model—especially for AP/AI and AP/RI. The ratios between RP, AI, and RI are similar to those in the Heston model. The corresponding results for one year and two years to maturity are presented in Exhibit 10. Qualitatively the situation is similar to the Heston model: For shorter times to maturity the price differences decrease—especially for AP prices.

Thus, the price ratios in the Heston and in the Bates model are similar among the RP, AI, and RI prices while the AP price differences are bigger in the Bates model. Moreover, the variation of the price ratios is higher in the Bates model.

Down and out p put options The prices of the down and out puts with strike K, barrier B, and maturity T on an underlying (S_t) are given by

EXHIBIT 9

Note: Relative prices of the up and out calls in the Heston model for 3 years to maturity.

$$\exp(-rT) \, E[1_{\{m_T > B\}}(K - S_T)^+] \qquad (17)$$

where

$$m_T \stackrel{def}{=} \min_{0 \le t \le T} S_t \qquad (18)$$

For our analysis, we use the strike K and the barrier B

$$K = (1 + 0.1\,T)S_0 \qquad (19)$$

$$B = (1 - 0.2\,T)S_0 \qquad (20)$$

where T denotes time to maturity. The strikes and barriers are set analogously to those for the up and out calls. Such down and out puts are often part of bonus certificates.[1]

Down and out put options have the payoff profile of European put options if the underlying has been above the barrier during the lifetime of the option. Otherwise their payoff is zero.

As described above, we calculate on each day the prices of the down and out puts for the four parameter sets. The resulting six time series of price ratios are shown in the Exhibit 12 for three years to maturity in the Heston model. The AP prices are (on average) about 3.5% smaller than the other prices and 75% of the AP prices are at least 2% smaller than the other prices. The RP prices lie above the prices from the calibrations to implied volatilities. These AI and RI prices are quite similar so that we can identify again the two groups that we have already observed for the up and out calls.

Compared to the up and out calls, the price differences are smaller for the down and out puts. This can be seen also from Exhibit 13 that reports the median of the price ratios for one, two, and three years to maturity. This exhibit shows that the price ratios change for increasing time to maturity: For one year to maturity, the AP prices lie above the other prices but with increasing time to maturity, the AP prices become relatively smaller. The RP and AI prices remain on a similar level for all times to maturity and the RI prices tend to this level for longer times to maturity.

The situation in the Bates model that gives a better fit to the plain vanilla data is described in Exhibit 13 and Exhibit 14. The AP prices lie about 7% below the other prices. Thus this difference is bigger than in the Heston model. The other price ratios still lie on average on the same level but their variance has grown compared to the Heston model.
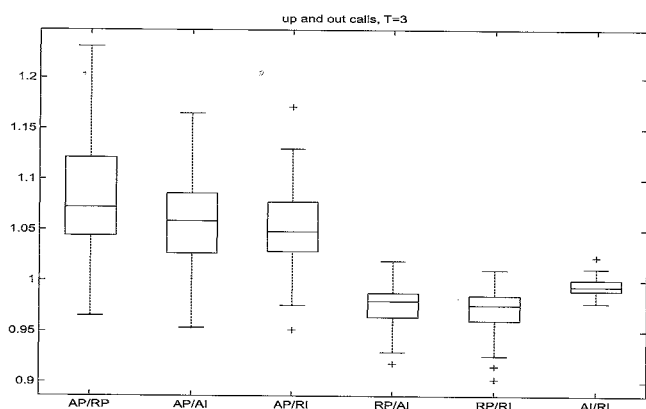
EXHIBIT 10

		AP/RP	AP/AI	AP/RI	RP/AI	RP/RI	AI/RI
Heston	T = 1	0.986	0.968	0.967	0.984	0.984	0.999
	T = 2	1.051	1.024	1.022	0.979	0.978	0.998
	T = 3	1.072	1.059	1.048	0.980	0.976	0.994
Bates	T = 1	0.988	0.985	1.002	1.002	1.006	1.012
	T = 2	1.070	1.083	1.104	0.970	0.986	1.018
	T = 3	1.106	1.123	1.129	0.972	0.975	1.013

Note: Median of price ratios of up and out calls.

The situation for the barrier options can be summarized as follows: the AP prices differ significantly from the other prices for both types of barrier options. While the AP prices are higher for up and out calls, they are lower for down and out puts relative to the other prices. In this sense the situation is symmetrical. The differences become bigger for longer times to maturity and the better fit of the Bates model does not lead to smaller price differences.

Cliquet Options

We consider cliquet options with prices

$$\exp(-rT)E[H] \tag{21}$$

EXHIBIT 11
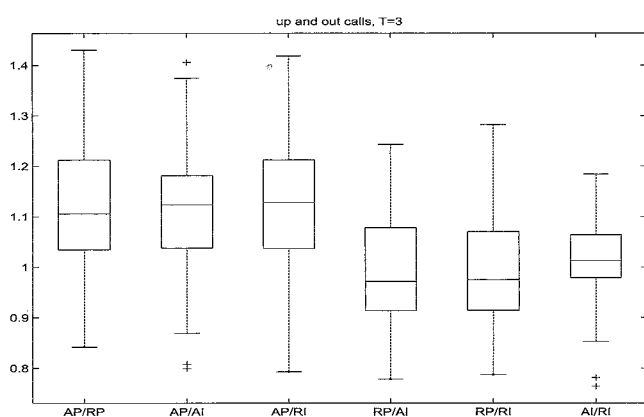
Note: Relative prices of the up and out calls in the Bates model for 3 years to maturity.

EXHIBIT 12
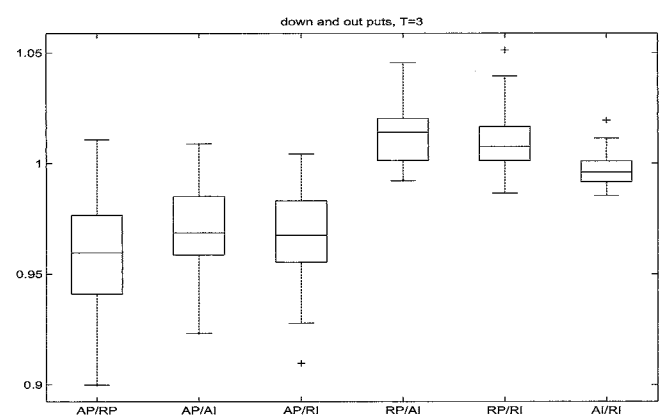
Note: Relative prices of the down and out puts in the Heston model for 3 years to maturity.

Kai Detlefsen and Wolfgang K. Härdle: Calibration Risk for Exotic Option

EXHIBIT 13

		AP/RP	AP/AI	AP/RI	RP/AI	RP/RI	AI/RI
Heston	$T = 1$	1.025	1.031	1.005	1.007	0.980	0.977
	$T = 2$	0.983	0.994	0.984	1.011	0.997	0.986
	$T = 3$	0.960	0.969	0.968	1.014	1.008	0.996
Bates	$T = 1$	1.021	1.012	1.019	1.004	1.006	0.998
	$T = 2$	0.968	0.975	0.966	1.031	1.022	0.990
	$T = 3$	0.922	0.935	0.931	1.026	1.022	0.995

Note: Median of price ratios of down and out puts.

where the payoff H is given by

$$H \stackrel{def}{=} \min\left(c_g, \max\left[f_g, \sum_{i=1}^{N} \min \left\{ c_l^i, \max\left(f_l^i, \frac{S_{t_i} - S_{t_{i-1}}}{S_{t_{i-1}}} \right) \right\} \right] \right) \quad (22)$$

EXHIBIT 14

down and out puts, T=3

Note: Relative prices of the down and out puts in the Bates model for 3 years to maturity.

Here c_g (f_g) is a global cap (floor) and c_g^i (f_g^i) is a local cap (floor) for the period $[t_{i-1}, t_i]$.

We consider three periods with $t_i = \frac{T}{3}i (i = 0,...,3)$ and the caps and floors are given by

$$
\begin{aligned}
c_g &= \infty \\
f_g &= 0 \\
c_l^i &= 0.08, \ i = 1, 2, 3 \\
f_l^i &= -0.08, \ i = 1, 2, 3
\end{aligned}
\quad (23)
$$

While barrier options are simple exotic options, cliquet options are more difficult because of their forward structure. Moreover, there exist many different types corresponding to the parameters. Hence, our specification cannot give a representative picture of all the traded cliquets. But these caps and floors are typical because the option holder cannot lose money and the returns are bounded above only by the local return bounds.

Cliquet options pay out basically the sum of the returns $R_i \stackrel{def}{=} (S_{t_i} - S_{t_i})/S_{t_{i-1}}$. In order to reduce risk, local and global floors f are introduced for the returns R. In the same way the returns are bounded from above by local and global caps c.

EXHIBIT 15
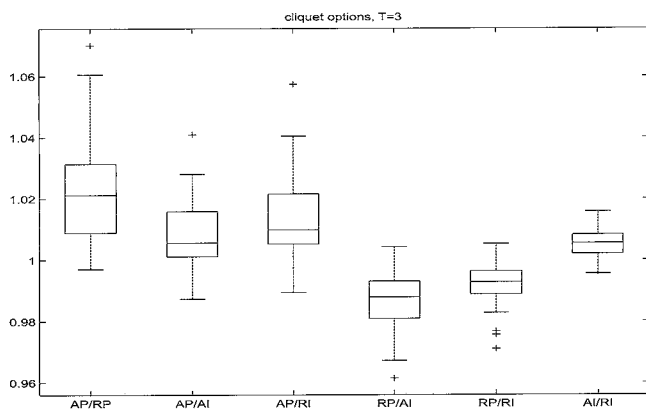
Note: Relative prices of the cliquet options in the Heston model for 3 years to maturity.

The distributions of the six time series of price ratios for cliquet options are described in Exhibit 15 for three years to maturity in the Heston model. The ratios are closer to 1 than in the case of the barrier options. The AP prices lie above the other prices but the difference is significant only for the AP and RP prices. The differences between the other prices is also small. Thus, we cannot recognize directly from this exhibit the two groups that we identified for the barrier options.

Exhibit 16 that reports the median price ratios for one, two, and three years to maturity gives some insight into this situation: the AP prices are about 2% smaller than the other prices for one year to maturity. With increasing time to maturity, the AP prices grow relatively and are about 1.5% higher than the other prices for three years to maturity. As Exhibit 16 confirms, the other price ratios remain relatively constant for different times to maturity. Thus there are again the two groups that we have identified for the barrier options: The changing relative AP prices on the one hand and the constant other price ratios on the other hand.

The relative prices of the cliquet options in the Bates model are presented in Exhibit 17 for three years to maturity. Here we see that the AP prices are about 7% smaller than the other prices. The RP prices lie about 2% under the AI prices that are 3% higher than the RI prices. The RP and RI prices are similar. Thus, there are quite big differences for the cliquet options in the Bates model. Moreover, the variance is larger relative to the Heston model. Exhibit 16 describes the situation of

different times to maturity and shows that the AP prices grow relatively with increasing time to maturity while the other price ratios remain rather constant for different times to maturity.

Comparing the results for the two types of barrier options and the cliquet options we see in all cases two groups, the AP prices and the other prices. The AP prices differ a lot from the other prices and in addition this difference changes for different times to maturity. Moreover, the variance of the price ratios with AP prices is bigger in general than for the other price ratios. The other group of RP, AI, and RI prices shows similar prices and small variances. The Bates model that gives better fits to implied volatility surfaces has higher price differences (with higher variances).

MODEL RISK

In the last section, we have described price differences that result from the calibration with respect to the four error functionals. In this section, we consider model risk and its relation to calibration risk and compare our results with the findings of Schoutens et al. [2004]. Model risk is generally understood as the risk of "wrong" prices because an inappropriate parametric model has been chosen for modelling the price process of the underlying.

In order to analyze this model risk for the two stochastic volatility models, we consider the ratios of the prices of the exotic options in the Bates model and the corresponding prices in the Heston model. The distribution of these ratios for up and out calls with three years to maturity is described by Exhibit 18. The prices in the Bates model lie below the prices in the Heston model for all four error functionals: The difference varies between 2% for the AP prices and 6% for the RI prices. Thus model risk is not independent of the calibration method, i.e., calibration risk. The results for smaller times to maturity are given in Exhibit 19. The exhibit suggests that model risk does not change significantly for different times to maturity.

The model risk of down and out puts is shown in Exhibit 20 for three years to maturity. The prices in the Bates model lie below the prices in the Heston model for all error functionals. Compared to the up and out calls, the model risk is bigger for the down and out puts: it varies between 9% for AP prices and 14% for RI prices. But again we observe the highest difference for RI prices

EXHIBIT 16

		AP/RP	AP/AI	AP/RI	RP/AI	RP/RI	AI/RI
Heston	T = 1	0.983	0.976	0.989	0.993	1.006	1.013
	T = 2	1.002	0.991	1.000	0.989	0.998	1.010
	T = 3	1.022	1.008	1.014	0.987	0.992	1.005
Bates	T = 1	0.917	0.899	0.917	0.987	1.005	1.024
	T = 2	0.931	0.903	0.923	0.980	0.999	1.029
	T = 3	0.946	0.912	0.933	0.976	0.995	1.029

Note: Median of price ratios of cliquet options.

and the smallest for AP prices. Moreover, the variance is bigger than for the up and out calls. Exhibit 19 that gives the results for smaller times to maturity can be interpreted in such a way that the model risk becomes smaller for shorter times to maturity.

Finally, we consider the model risk of cliquet options in Exhibit 21. For these options the Bates prices lie above

the corresponding Heston prices for all calibration methods. The smallest price difference that appears for the AP prices is about 8% while the biggest difference of 16% is for RI prices. Exhibit 19 shows again smaller price differences for shorter times to maturity.

The model risk between the Heston and the Bates model can be described for barrier and cliquet options as

EXHIBIT 17
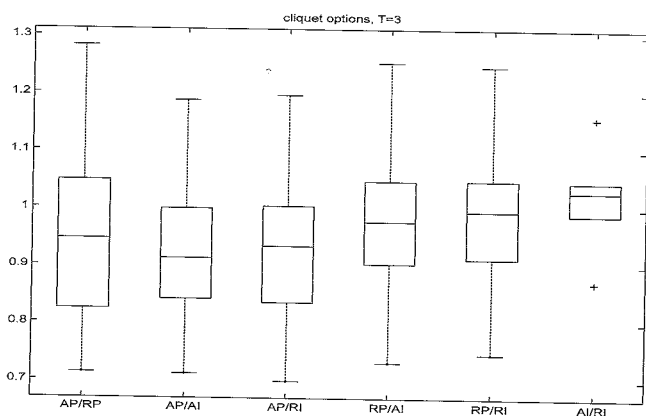
Note: Relative prices of the cliquet options in the Bates model for 3 years to maturity.

EXHIBIT 18
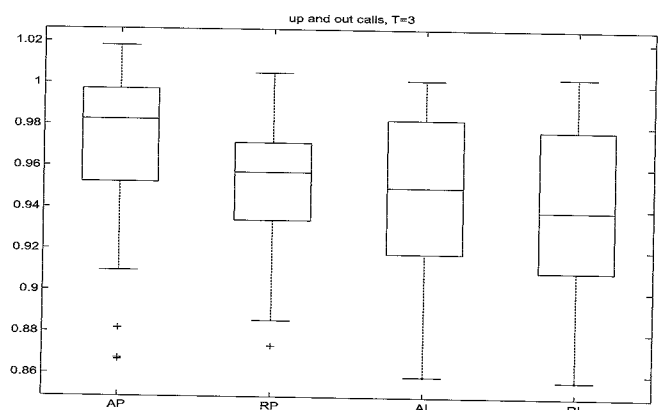
Note: Ratio of Bates to Heston prices for up and out calls with 3 years to maturity on 51 days.

Exhibit 19

		AP	RP	AI	RI
up and out calls	$T = 1$	0.973	0.953	0.944	0.941
	$T = 2$	0.980	0.954	0.953	0.940
	$T = 3$	0.983	0.957	0.950	0.939
down and out puts	$T = 1$	0.933	0.892	0.877	0.878
	$T = 2$	0.918	0.883	0.872	0.860
	$T = 3$	0.916	0.881	0.873	0.860
cliquets	$T = 1$	1.057	1.100	1.109	1.119
	$T = 2$	1.076	1.128	1.130	1.144
	$T = 3$	1.086	1.138	1.140	1.162

Note: Median of ratios of Bates to Heston prices.

follows: model risk measured by the price ratios in the two models increases for longer times to maturity. Moreover, it is ordered with respect to the calibration method. The calibration with respect to implied volatilities leads to bigger price differences than the calibration with respect to prices. The model risk is smallest for the AP calibration and bigger for the RP calibration. It is even bigger for the AI calibration and the price differences are the biggest for the RI calibrations. This emphasizes the importance of the implied volatility surfaces and their calibration. Moreover, model risk differs across option types. The more exotic cliquet options have a higher model risk than the barrier options.

Schoutens et al. [2004] consider up and out calls (with strike equal to spot) and cliquet options with three years to maturity. For a barrier 50% above the spot, they find a model risk for the up and out calls of about 14%. For the cliquet options, they do not find a significant model risk. These results do not correspond completely to our AP results. There may be several reasons for these different results: while

we look at a time series of 51 implied volatility surfaces, they focus only on one day. Moreover, they have analyzed EuroStoxx50 prices and we use DAX data.

CONCLUSION

We have looked at the stochastic volatility model of Heston and analyzed different calibration methods and their impact on the pricing of exotic options. Our analysis has been carried out for a time series of DAX implied volatility surfaces from April 2003 to March 2004.

We have shown that different ways to measure the error between the model and the market in the calibration routine lead to significant price differences for exotic options. We have considered the four error measures that are defined by the root mean squared error of absolute or relative differences between prices or implied volatilities in the market and in the model. Among these measures we have identified two groups. Calibrations with respect to relative prices, absolute implied volatilities, or relative

EXHIBIT 20
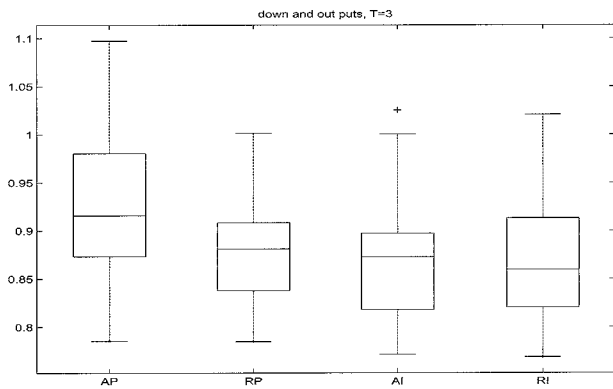
Note: Ratio of Bates to Heston prices for down and out puts with 3 years to maturity on 51 days.

EXHIBIT 21

Note: Ratio of Bates to Heston prices for cliquet options with 3 years to maturity on 51 days.

implied volatilities lead to similar prices of exotic options. Calibrations with respect to absolute prices imply exotics prices that are quite different from the prices of the first group. As implied volatilities are of comparable size for different strikes and different times to maturity, calibrations with respect to absolute or relative implied volatility differences lead to similar prices of exotic options. As these prices are also rather similar to those computed from calibrations with respect to relative price differences, we conclude that all three of these error measures have a similar weighting of the errors. The price differences between the described two groups of error measures increase for longer times to maturity. Moreover, the differences do not decrease in the Bates model although it is an extension of the Heston model with similar qualitative features and a better fit to plain vanilla data. The price differences of exotic options differ also across option types and are bigger for barrier options than for cliquets.

The observed price differences of exotic options between the two groups of error measures can be explained in the following way: as we use out of the money options for the calibration, measuring the error by absolute prices puts most weight at the money. The characteristic feature of implied volatility surfaces around the money is the skew because the smile can be identified only from deep out-of-the-money calls. Hence, errors in terms of absolute prices focus on the skew. In the Heston model, the skew is mainly controlled by the correlation between the Brownian motions of the stock and the variance process. Our calibrated model parameters confirm

this because the calibrations with respect to absolute prices lead to higher (absolute) correlations between these two processes. At the same t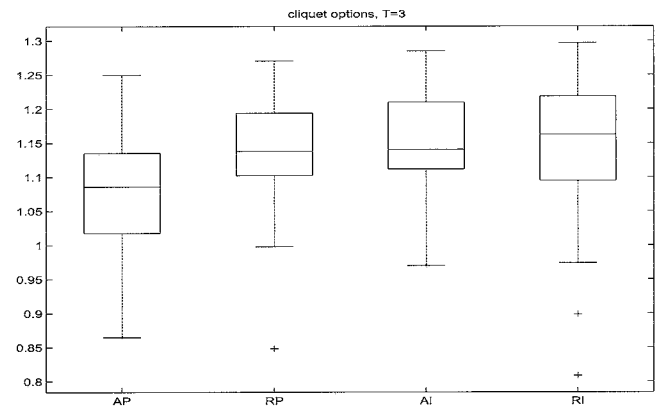ime such calibrations imply smaller volatilities of variance which control the smile. These two effects make the distribution of the stock more right skewed so that up and out calls (down and out puts) are more (less) expensive in models that are calibrated to absolute price differences.

Moreover, we have looked at the model risk between the Heston and the Bates model. Model risk and calibration risk are not independent because model risk is lowest for calibrations with respect to absolute prices and highest for calibrations with respect to relative implied volatilities. As this holds for all considered options, model risk seems to be ordered with respect to the error measure used in the calibration. In the analyzed time period the DAX experienced a strong upward trend. This is also reflected in the positive expected jumps in the Bates model. These jumps can explain why the barrier (cliquet) options are more (less) expensive in the Heston model than in the Bates model. Because of the jumps, the considered down and out puts often finish out of the money. The up and out calls are knocked out more often because of the upward jumps. The higher prices of the cliquet options in the Bates model may be due to the higher returns that result from the upward jumps. If the market has a downward trend that is also reflected in negative expected jumps in the Bates model, then model risk should have the same size in general but could have a different sign, e.g., for barrier options.

As model risk is bigger than calibration risk, calibrations should be carried out with respect to absolute prices if the choice of an appropriate model is unclear. But if a model has already been chosen we suggest measuring the error between the model and the market in terms of (relative) implied volatilities, because this error measure best reflects the characteristics of the model that are essential for pricing exotic options.

ENDNOTE

[1]A bonus certificate is a structured product incorporating a zero-strike call and a down and out put option. The payoff structure at expiry can be described as follows: if the value of the underlying at expiry is above the strike then the investor receives the underlying. If the price of the underlying is below the barrier then the investor bears the full loss of the underlying because she gets only the underlying. Otherwise the investor gets the underlying and in addition a bonus if the barrier has not been exceeded before expiration. Because of its payoff profile, this product is constructed mainly for sideways markets.

REFERENCES

Bates, D. "Jump and Stochastic Volatility: Exchange Rate Processes Implicit in Deutsche Mark Options." *Review of Financial Studies*, 9 (1996), pp. 69–107.

Carr, P., and D. Madan. "Option Valuation Using the Fast Fourier Transform." *Journal of Computational Finance*, 2 (1999), pp. 61–73.

Cont, R. "Model Uncertainty and its Impact on the Pricing of Derivative Instruments." *Mathematical Finance*, 16 (2005), pp. 519–542.

Cont, R., and P. Tankov. *Financial Modelling With Jump Processes*. Chapman and Hall/CRC, 2004.

Glasserman, P. *Monte Carlo Methods in Financial Engineering*. New York: Springer, 2004.

Heston, S. "A Closed-Form Solution for Options with Stochastic Volatility with Applications to Bond and Currency Options." *Review of Financial Studies*, 6 (1993), pp. 327–343.

Mikhailov, S., and U. Nögel. "Heston's Stochastic Volatility Model. Implementation, Calibration and Some Extensions." *Wilmott magazine* (July 2003), pp. 74–94.

Schoutens, W., E. Simons, and J. Tistaert. "A Perfect Calibration! Now What?" *Wilmott magazine* (March 2004), pp. 66–78.

To order reprints of this article, please contact Dewey Palmieri at dpalmieri@iijournals.com or 212-224-3675

Special Issue on Econometrics of Financial and Insurance Risks

A Semiparametric Factor Model for Implied Volatility Surface Dynamics

Matthias R. Fengler
Sal. Oppenheim jr. & Cie.

Wolfgang K. Härdle
Center for Applied Statistics and Economics

Enno Mammen
University of Mannheim

ABSTRACT

We propose a semiparametric factor model, which approximates the implied volatility surface (IVS) in a finite dimensional function space. Unlike standard principal component approaches typically used to reduce complexity, our approach is tailored to the degenerated design of IVS data. In particular, we only fit in the local neighborhood of the design points by exploiting the expiry effect present in option data. Using DAX index option data, we estimate the nonparametric components and a low-dimensional time series of latent factors. The modeling approach is completed by studying vector autoregressive models fitted to the latent factors.

KEYWORDS: functional principal component analysis, implied volatility surface, semiparametric factor models

The implied volatility surface (IVS), which is derived by applying the Black-Scholes (BS) formula to a set of traded plain vanilla options across different strikes and expiries, is a key financial variable for trading, hedging, and the risk management of option portfolios. In equity markets, the IVS is usually observed as a nonflat surface with a strong skew for out-of-the-money puts, as shown in the left panel of Figure 1. Despite this observation, which invalidates the constant volatility

Figure 1 Left panel: DAX index call and put implied volatilities observed on May 2, 2000, plotted across futures moneyness and time-to-maturity. Right panel: same data as seen from top.

assumption in the underlying model framework by Black and Scholes (1973), the IVS is a widely accepted state variable, which—unlike estimates based on historical data—is believed to reflect the current perception of market risk [Bakshi et al. (2000); Britten-Jones and Neuberger (2000)]. Consequently, market makers at plain vanilla desks continuously monitor and update the IVS they trade on, and risk managers study the impact of large market moves on entire portfolios in terms of joint spot and implied volatility shifts or other deformations of the IVS. To be implemented in a realistic way, all these applications require a suitable statistical model of the IVS.

Modeling the IVS poses two main challenges. First and foremost, the implied volatility data have a degenerated design: by "degenerated design" we refer to the institutional convention that implied volatility data exist only for a small number of maturities such as 1, 2, 3, 6, 9, 12, 18, and 24 months to expiry on the date of issue. Hence, implied volatility observations appear in "strings"; see the right panel of Figure 1, which shows the observations in the moneyness and time-to-maturity space only. Options belonging to the same string have the same time-to-maturity. As time passes, these strings move along the maturity axis toward expiry while changing their levels and shapes in a random fashion, see Figure 2. A second challenge is that the observations do not always cover the desired estimation grid completely and observations can be missing in certain sub-regions in the moneyness dimension. For instance, the options belonging to the third and fifth time-to-maturity (from top) are observed only near-the-money, while options belonging to the other time-to-maturities appear shifted either to the out-of-the-money put region or to the out-of-the-money call region. However, despite that appearance, implied volatilities are thought as being the observed structure of a smooth surface. This is because in practice one needs to price and hedge over-the-counter options whose expiry dates do not coincide with the expiry dates of options that are traded at the futures exchange.
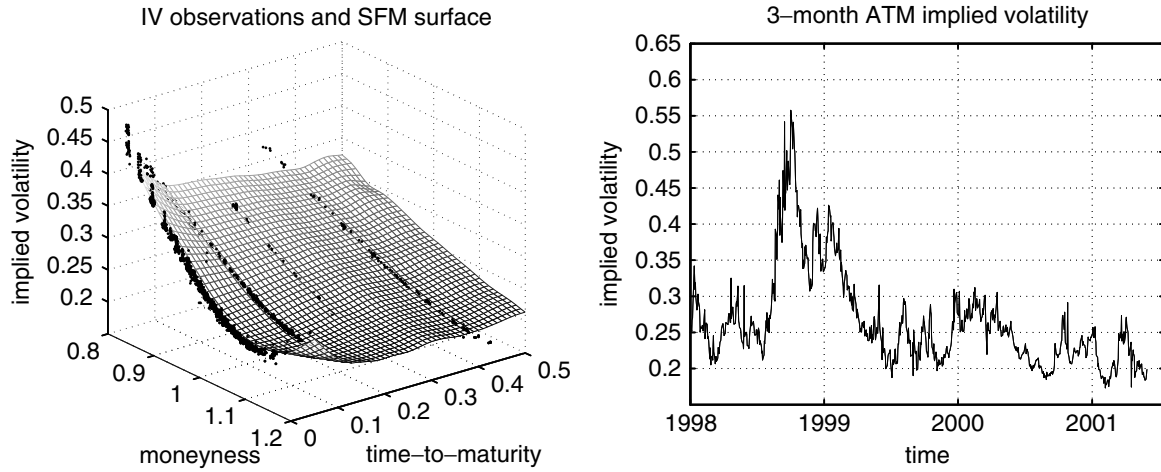
Figure 2 Left panel: DAX index IV observations and SFM fit for May 2, 2000. Bandwidths $h_1 = 0.04$ for the moneyness and $h_2 = 0.06$ for the time-to-maturity dimension. Right panel: Time series of three-months at-the-money implied volatility data obtained from linear interpolation of the raw data.

For the semi- or nonparametric approximations to the IVS used in Aït-Sahalia and Lo (1998), Aït-Sahalia et al. (2001), Cont and da Fonseca (2002), Fengler et al. (2003), and Benko et al. (2007) this design can pose difficulties. The reason is that for standard nonparametric estimators to achieve a reasonable fit to the IVS, the bandwidth in the time-to-maturity dimension must be chosen very large in order to bridge the gaps between the different time-to-maturity strings. This induces an estimation bias that becomes the more severe the more irregular the observations are distributed across moneyness.

In this article we propose a semiparametric factor model (SFM) for modeling the IVS. The dynamic structure of the IVS is approximated by unknown basis functions moving in a finite dimensional function space. To introduce our model, denote by $Y_{i,j}$ the log-implied volatility, where $i = 1, \ldots, I$ is an index of time, in our case the number of the day, and $j = 1, \ldots, J_i$ is an intra-day numbering of the option traded on day i. The observations $Y_{i,j}$ are regressed on two-dimensional covariables $X_{i,j}$ that contain moneyness $\kappa_{i,j}$ and time-to-maturity $\tau_{i,j}$. Moneyness is defined as $\kappa_{i,j} \stackrel{\text{def}}{=} K_{i,j}/F_{i,j}$, that is, strike $K_{i,j}$ divided by the futures price $F_{i,j}$ belonging to the option trade (i, j). The SFM approximates the IVS by

$$Y_{i,j} \approx m_0(X_{i,j}) + \sum_{l=1}^{L} \beta_{i,l}\, m_l(X_{i,j}), \tag{1}$$

where m_l are smooth basis functions ($l = 0, \ldots, L$) and $\beta_{i,l}$ are weights depending on time i. The factor loading $\beta_i \stackrel{\text{def}}{=} (\beta_{i,1}, \ldots \beta_{i,L})^\top$ forms an unobserved multivariate time series. By fitting model (1) to the implied volatility observations, we obtain approximations $\widehat{\beta}_i$. Their dynamic structure can be understood by applying classical vector autoregression (VAR) techniques. We argue that the VAR estimation based on $\widehat{\beta}_i$ is asymptotically equivalent to an estimation based on

the unobserved β_i. A justification for this is given in Borak et al. (2006), where the functions m_l are estimated by orthogonal series estimators. Our conjecture is that their arguments carry over to kernel smoothing. In the left panel of Figure 2, we display a fit coming from the SFM.

Factor models that represent processes whose dynamics are driven by latent variables have a long tradition in the asset pricing literature. For equity derivatives, the main approach is to model stock returns with stochastic volatility as latent factor, such as in Hull and White (1987), Wiggins (1987), Heston (1993), and Schöbel and Zhu (1999) among others. Empirical research suggests to add a jump component, [see Bates (1996, 2000) and Duffie et al. (2000)]. A further refinement includes stochastic volatility with Lévy innovations, [see inter alia Barndorff-Nielsen (1997), Barndorff-Nielsen and Shepard (2001), and Carr et al. (2003)]. For parameter identification in the presence of latent variables a number of different estimation techniques were proposed[1]: analytically tractable filtration approaches as in Harvey et al. (1994) and Fridman and Harris (1998), or the approximate maximum likelihood methodology due to Bates (2006); generalized method of moments (GMM) approaches, either based on analytical moment restrictions, such as Melino and Turnbull (1990), Andersen and Sørensen (1996), Chacko and Viceira (2003), or based on simulated moments, such as Duffie and Singleton (1993); efficient method of moments techniques as in Gallant and Tauchen (1998), Andersen et al. (1999), Chernov and Ghysels (2000), and the indirect inference method due to Gouriéroux et al. (1993); implied-state methods, for example, as an extension of the expectation-maximization methodology as in Renault and Touzi (1996) and Pastorello et al. (2003), or in a GMM sense as in Pan (2002); and Bayesian Monte Carlo Markov Chain approaches as suggested by Jacquier et al. (1994) and Eraker et al. (2003).

To better capture unknown nonlinearities, a more recent strand of literature employs non- and semiparametric techniques, for example, to estimate the option pricing function [see e.g. Broadie et al. (2000a,b) and Aït-Sahalia and Duarte (2003)], or to recover the yield curve [Linton et al. (2001)]. Semiparametric factor models were proposed by Ghysels and Ng (1989) and Jeffrey et al. (2004) for modeling the term structure of interest rates. Connor and Linton (2006) use a semiparametric model for stock returns based on observable characteristics, and in Connor et al. (2006) a heteroskedastic latent factor model is designed which has nonparametric components in volatility. We complement this literature, as low-dimensional factor approximations to the IVS are traditionally based on principal components analysis (PCA), such as in Zhu and Avellaneda (1997), Skiadopoulos et al. (1999), Cont and da Fonseca (2002), and Fengler et al. (2003). The SFM differs in the following respect: for example, in Cont and da Fonseca (2002) the IVS is fitted on a grid for each day and a functional norm is applied to the surfaces as described in Ramsay and Silverman (1997). In our approach, we use only finite dimensional fits which are obtained in the local neighborhood of strikes and maturities for

---

[1] This overview is by no means complete; for a comprehensive review see Garcia et al. (2007) and Gallant and Tauchen (2007).
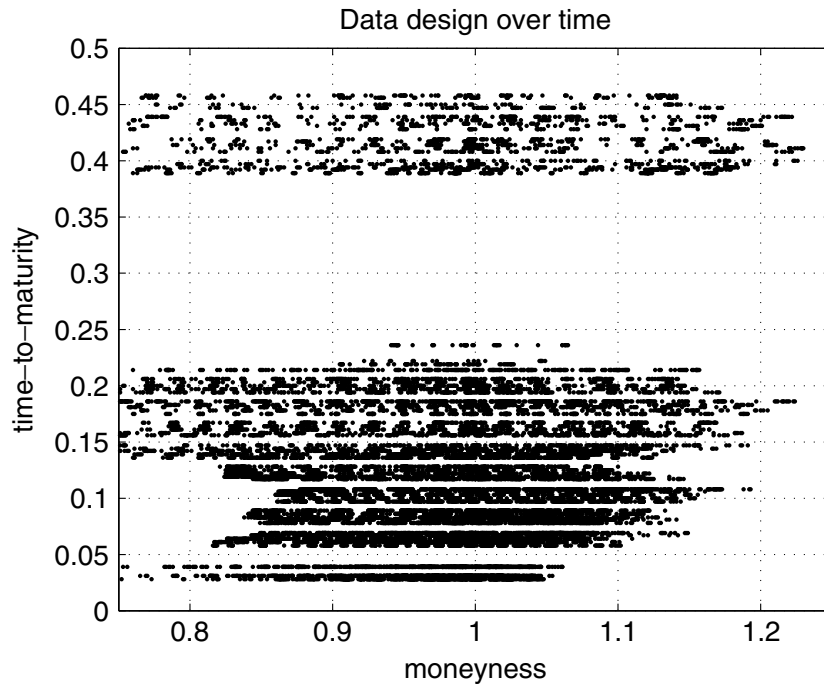
Figure 3 Evolution of data design as time passes. Data taken from January 01 to 31,1998; weekends appear as the white lines, since the scale is calendar days. Each dot corresponds to an actual trade.

which implied volatilities are recorded at the specific observation days. This leads to a minimization with respect to functional norms that depend on time i. The method can be understood as smoothing through time, which is made possible by the expiry effect present in IVS data, but which also exists in other economic variables, for example, in fixed income securities. This is illustrated in Figure 3 which shows the evolution of the data design for the period of one month. To our knowledge there do not yet exist estimation techniques that explicitly take advantage of the expiry effect.

The SFM is close to Connor and Linton (2006), but differs in that the explanatory variables $X_{i,j}$ do not depend on l. Under this assumption, the "factor coefficients" $m_l(X_{i,j,l})$ are identified and model (1) reduces to a nonparametric additive model where the additive components m_l can be estimated by smoothing techniques.[2] In our case $X_{i,j}$ does not depend on l and therefore, as in classical factor analysis, factor loadings and factor functions are not uniquely defined. Furthermore, our research has obvious relations to the literature on functional or varying-coefficient models, such as Hastie and Tibshirani (1993), Cai et al. (2000), Fan et al. (2003), or Kauermann et al. (2005).

Using transactions-based DAX index implied volatility data from January 1998 to May 2001, we recover three basis functions and three latent factors which we model with a VAR(2) model. The factor functions have natural interpretations as level, moneyness, and term structure slope effects. In comparison to a Nadaraya–Watson estimator, we demonstrate that the SFM mitigates bias

---

[2] For simplicity, Connor and Linton (2006) assume that $X_{i,j,l}$ does not depend on time i.

effects caused by the global daily fits, which are usually used in functional PCA.

The article is organized as follows. In the following section, implied volatilities are described and the SFM is presented. In Section 2 we estimate the factor model on DAX index option data. Section 3 concludes. The appendix gives details on the issues of model choice and on the preparation of the data.

1 IVS MODELING USING A SEMIPARAMETRIC FACTOR MODEL

1.1 Model Characterization

Implied volatilities are derived from the BS option pricing formula for European calls and puts [Black and Scholes (1973)]. European style calls and puts are contingent claims on a stock S_t (paying no dividends for simplicity, here), which yield the pay-off $\max(S_T - K, 0)$ and $\max(K - S_T, 0)$, respectively, for a strike price K at a given expiry day T. The asset price process S_t in the BS model is assumed to be a geometric Brownian motion. The BS option pricing formula for calls is given by

$$C_t^{BS}(S_t, K, \tau, r, \sigma) = S_t \Phi(d_1) - e^{-r\tau} K \Phi(d_2), \tag{2}$$

where $d_1 = \frac{\log(S_t/K) + (r + 0.5\sigma^2)\tau}{\sigma\sqrt{\tau}}$ and $d_2 = d_1 - \sigma\sqrt{\tau}$. Here, $\Phi(\cdot)$ denotes the cumulative distribution function of the standard normal distribution, $\tau = T - t$ time-to-maturity of the option, r the riskless interest rate over the option's life time, and σ the diffusion coefficient of the Brownian motion. Put prices P_t are obtained via the put–call parity $C_t - P_t = S_t - e^{-\tau r} K$. Given observed market prices $\tilde{C}_t$, one defines implied volatility $\hat{\sigma}$ by:

$$C_t^{BS}(S_t, K, \tau, r, \hat{\sigma}) - \tilde{C}_t = 0 . \tag{3}$$

Because of monotonicity of the BS price in σ, there exists a unique solution $\hat{\sigma} > 0$ which can be found using numerical methods [see the discussion by Manaster and Koehler (1982)]. Practically the computation of Equation (3) can fail, in particular, when the data violate standard arbitrage arguments or when the options are very far in or very far out-of-the-money.

In the SFM, we regress log-implied volatility $Y_{i,j} = \log\{\hat{\sigma}_{i,j}(\kappa, \tau)\}$ on the covariates $X_{i,j} = (\kappa_{i,j}, \tau_{i,j})$, where $\kappa_{i,j} = K_{i,j}/F_{i,j}$ is a moneyness metric with $F_{i,j}$ as the futures price at time (i, j) and where $\tau_{i,j}$ denotes time-to-maturity. In borrowing ideas from fitting additive models, such as as in Stone (1986), Hastie and Tibshirani (1990) and Mammen et al. (1999), we will estimate the nonparametric components m_l and the state variables $\beta_{i,l}$ with least squares methods. This is motivated by rewriting Equation (1) as

$$Y_{i,j} = m_0(X_{i,j}) + \sum_{l=1}^{L} \beta_{i,l} \, m_l(X_{i,j}) + \varepsilon_{i,j} , \tag{4}$$

where $\varepsilon_{i,j}$ is mean zero, conditional on $X_{i,j}$ and independent of the stationary process β_i. From an empirical point of view, one might expect that the errors exhibit heteroskedasticity and correlation patterns that depend on time i and the spatial index j.[3] This suggests the use of weighted least squares as proposed in Renault (1997, Section 2.3). In our setting, weights that correct for spatial heteroskedasticity lead to second-order corrections of the estimate only, because of the localizing smoothing approach. Furthermore, correcting for heteroskedasticity in time would require to specify an additional model for these time effects. Therefore, we will use unweighted least squares.

In an option pricing framework such as (4) a statistical singularity emerges, when the number of observed asset prices exceeds the number of state variables. This is because the state variables can be identified exactly, when the number of observed option prices matches the number of state variables [see e.g., Pan (2002) for approaches of this kind]. Therefore, rather than interpreting the error as mispricings which could be exploited by arbitrage strategies, we follow the notion of Renault (1997) and perceive the error term as an option pricing error, that is, we interpret the observed option price as the price given by a pricing formula plus an error term which might be due to a neglected heterogeneity factor. Hence, the econometric specification is with respect to another martingale measure equivalent to the actual pricing measure. This notion allows us to stay within the Harrison and Kreps (1979) framework.

In specifying the factor functions m_l as functions of moneyness, the model is homogeneous of degree zero in strikes and spot. This allows us—although it is not explicitly stated—to more precisely characterize our underlying asset pricing model [Renault (1997)]. First, an implied volatility function homogeneous of degree zero in strikes and spot is equivalent to a (European) option pricing function that is homogeneous of degree one in strikes and spot.[4] Second, by this homogeneity property, the delta hedge ratio of the option pricing function can be decomposed into the BS delta minus a vega correction, which depends on the strike slope of the implied volatility function. This relationship allows splitting of the hedging error into a component due to the misspecification of the BS model and into another component due to the smile effect.[5] Finally, a necessary and sufficient condition for this homogeneity property is that future returns and the concurrent underlying asset price be conditionally independent under the risk-neutral pricing measure, given the currently available information other than the current underlying asset price. Therefore, it is possible to characterize our setting as a multifactor, stochastic volatility framework,

---

[3] The index j was introduced as an intra-day numbering of the observed options; but comprising the data for a given day i, as we do, one can interpret it in a spatial sense across the surface.

[4] Strictly speaking, this homogeneity property holds only in the absence of cash dividends; since the DAX index is a performance index, this assumption can be justified, see Section 2.1 for details.

[5] Detailed discussions of the implications for delta hedging in homogeneous models are given by Garcia and Renault (1998a) and Alexander and Nogueira (2007).

where the underlying asset price process does not cause the latent variables β_i in a Granger sense; for proofs, an illustration, and the precise formulation of the notion of conditional independence we refer to Garcia and Renault (1998b).

1.2 Estimation Algorithm

We define the estimates $\widehat{m}_l, (l = 0, \ldots, L)$ and $\widehat{\beta}_{i,l} \, (i = 1, \ldots, I; \; l = 1, \ldots, L)$ of model (4) as minimizers of the following least squares criterion $(\widehat{\beta}_{i,0} \overset{\mathrm{def}}{=} 1)$:

$$\sum_{i=1}^{I} \sum_{j=1}^{J_i} \int \left\{ Y_{i,j} - \sum_{l=0}^{L} \widehat{\beta}_{i,l} \widehat{m}_l(u) \right\}^2 K_h(u - X_{i,j}) \, du \, . \tag{5}$$

Here, K_h denotes a two-dimensional product kernel, $K_h(u) = k_{h_1}(u_1) \times k_{h_2}(u_2)$, $h = (h_1, h_2)^\top$, based on the one-dimensional kernel $k_h(v) = h^{-1}k(h^{-1}v)$. In our applications we use the quartic kernel which is defined as $k(v) = \frac{15}{16}(1 - v^2)^2 \, \mathbf{1}(|v| \le 1)$, where $\mathbf{1}(\cdot)$ denotes the indicator function.

In (5) the minimization runs over all functions $\widehat{m}_l : \mathbb{R}^2 \to \mathbb{R}$ and all values $\widehat{\beta}_{i,l} \in \mathbb{R}$. Consider the case $L = 0$: the implied volatilities $Y_{i,j}$ are approximated by a surface $\widehat{m}_0$ that does not depend on time i. In this degenerated case, $\widehat{m}_0(u) = \sum_{i,j} K_h(u - X_{i,j}) Y_{i,j} / \sum_{i,j} K_h(u - X_{i,j})$, which is the Nadaraya–Watson estimate based on the pooled sample of all days. In the case $L > 0$, the implied volatility surfaces are approximated by surfaces moving in an L-dimensional affine function space $\{\widehat{m}_0 + \sum_{l=1}^{L} \beta_l \widehat{m}_l : \; \beta_1, \ldots, \beta_L \in \mathbb{R}\}$. The estimates $\widehat{m}_l$ are not uniquely defined. They can be replaced by functions that span the same affine space. In order to respond to this problem, we select $\widehat{m}_l$ such that they are orthonormal. As shall be seen in Section 2, the orthonormalization also facilitates the interpretation of the functions.

In order to derive an estimator of the SFM we proceed in the following way. For an l' with $1 \le l' \le L$, for $\delta > 0$ and a function g, we replace in (5) $\widehat{m}_{l'}$ by $\widehat{m}_{l'} + \delta g$ and take the first derivative with respect to δ at the point $\delta = 0$. This yields

$$\sum_{i=1}^{I} \sum_{j=1}^{J_i} \int \left\{ Y_{i,j} - \sum_{l=0}^{L} \widehat{\beta}_{i,l} \widehat{m}_l(u) \right\} \widehat{\beta}_{i,l'} K_h(u - X_{i,j}) g(u) du = 0 \, . \tag{6}$$

Since this equation holds for all functions g, it implies that for $1 \le l' \le L$ and for all u

$$\sum_{i=1}^{I} \sum_{j=1}^{J_i} \left\{ Y_{i,j} - \sum_{l=0}^{L} \widehat{\beta}_{i,l} \widehat{m}_l(u) \right\} \widehat{\beta}_{i,l'} K_h(u - X_{i,j}) = 0 \, . \tag{7}$$

Second, by replacing $\widehat{\beta}_{i',l'}$ by $\widehat{\beta}_{i',l'} + \delta$ in (4) and by taking derivatives with respect to δ we get, for $1 \le l' \le L$ and $1 \le i' \le I$:

$$\sum_{j=1}^{J_{i'}} \int \left\{ Y_{i',j} - \sum_{l=0}^{L} \widehat{\beta}_{i',l}\widehat{m}_l(u) \right\} \widehat{m}_{l'}(u)K_h(u - X_{i',j})du = 0. \tag{8}$$

For convenience, we introduce the notation

$$\widehat{p}_i(u) = \frac{1}{J_i} \sum_{j=1}^{J_i} K_h(u - X_{i,j}) , \tag{9}$$

$$\widehat{q}_i(u) = \frac{1}{J_i} \sum_{j=1}^{J_i} K_h(u - X_{i,j})Y_{i,j} , \tag{10}$$

for $1 \le i \le I$. The quantity $\widehat{p}_i(u)$ is recognized as the kernel density estimate of the data of day i, while $\widehat{q}_i(u)$ is a weighted sum of the dependent variable known from the numerator of a Nadaraya–Watson estimator. Then, we obtain from Equations (7)–(8), for $1 \le l' \le L, 1 \le i \le I$:

$$\sum_{i=1}^{I} J_i\widehat{\beta}_{i,l'}\widehat{q}_i(u) = \sum_{i=1}^{I} J_i \sum_{l=0}^{L} \widehat{\beta}_{i,l'}\widehat{\beta}_{i,l}\widehat{p}_i(u)\widehat{m}_l(u) , \tag{11}$$

$$\int \widehat{q}_i(u)\widehat{m}_{l'}(u) \, du = \sum_{l=0}^{L} \widehat{\beta}_{i,l} \int \widehat{p}_i(u)\widehat{m}_{l'}(u)\widehat{m}_l(u) \, du . \tag{12}$$

We calculate the estimates by iterative use of Equations (11) and (12). We start with initial values $\widehat{\beta}_{i,l}^{(0)}$ for $\widehat{\beta}_{i,l}$. A possible choice of the initial $\widehat{\beta}$ could correspond to fits of surfaces that are piecewise constant on time intervals $I_1, \ldots, I_L$. This means, for $l = 1, \ldots, L$, put $\widehat{\beta}_{i,l}^{(0)} = 1$ (for $i \in I_l$), and $\widehat{\beta}_{i,l}^{(0)} = 0$ (for $i \notin I_l$). Here $I_1, \ldots, I_L$ are pairwise disjoint subsets of $\{1, \ldots, I\}$ and $\bigcup_{l=1}^{L} I_l$ is a strict subset of $\{1, \ldots, I\}$. For $r \ge 0$, we put $\widehat{\beta}_{i,0}^{(r)} = 1$. Now, we define the matrix $B^{(r)}(u)$ by its elements

$$\left(B^{(r)}(u)\right)_{l,l'} \stackrel{\text{def}}{=} \sum_{i=1}^{I} J_i\widehat{\beta}_{i,l'}^{(r-1)}\widehat{\beta}_{i,l}^{(r-1)}\widehat{p}_i(u) , \qquad 0 \le l, \, l' \le L , \tag{13}$$

and introduce a vector $Q^{(r)}(u)$ with elements

$$Q^{(r)}(u)_l \stackrel{\text{def}}{=} \sum_{i=1}^{I} J_i\widehat{\beta}_{i,l}^{(r-1)}\widehat{q}_i(u) , \qquad 0 \le l \le L. \tag{14}$$

In the r-th iteration, the estimate $\widehat{m} = (\widehat{m}_0, \ldots, \widehat{m}_L)^\top$ is given by:

$$\widehat{m}^{(r)}(u) = B^{(r)}(u)^{-1}Q^{(r)}(u) . \tag{15}$$

This update step is motivated by Equation (11). The values of $\widehat{\beta}$ are updated in the r-th cycle as follows: for the matrix $M^{(r)}(i)$ with elements

$$\left(M^{(r)}(i)\right)_{l,l'} \stackrel{\text{def}}{=} \int \widehat{p}_i(u)\widehat{m}_{l'}^{(r)}(u)\widehat{m}_l^{(r)}(u)\, du\,, \qquad 1 \le l, l' \le L, \tag{16}$$

and for a vector $S^{(r)}(i)$ with elements

$$S^{(r)}(i)_l \stackrel{\text{def}}{=} \int \widehat{q}_i(u)\widehat{m}_l(u)\, du - \int \widehat{p}_i(u)\widehat{m}_0^{(r)}(u)\widehat{m}_l^{(r)}(u)\, du, \qquad 1 \le l \le L, \tag{17}$$

we compute

$$\left(\widehat{\beta}_{i,1}^{(r)}, ..., \widehat{\beta}_{i,L}^{(r)}\right)^{\top} = M^{(r)}(i)^{-1}S^{(r)}(i)\,, \tag{18}$$

which is motivated by Equation (12). The algorithm is iterated, until only minor changes in an L^2-fitting criterion occur (see Appendix A for the details). In the implementation, we choose a regularly spaced grid in the moneyness and time-to-maturity space and calculate $\widehat{m}_l$ for these points. In the calculation of $M^{(r)}(i)$ and $S^{(r)}(i)$, we replace the integrals by Riemann sums.

As discussed above, $\widehat{m}_l$ and $\widehat{\beta}_{i,l}$ are not uniquely defined. Therefore, we orthonormalize $\widehat{m}_0, \ldots, \widehat{m}_L$ in $L^2(\widehat{p})$, where $\widehat{p}(u) = I^{-1}\sum_{i=1}^I \widehat{p}_i(u)$ is the average design density in our sample, and order them according to the variance of the corresponding factor loadings $\widehat{\beta}_l$. This can be achieved by the following two steps. First, replace

$$\widehat{m}_0 \qquad \text{by} \qquad \widehat{m}_0^* \;=\; \widehat{m}_0 - \gamma^{\top}\Gamma^{-1}\widehat{m}\,,$$

$$\widehat{m} \qquad \text{by} \qquad \widehat{m}^* \;=\; \Gamma^{-1/2}\widehat{m}\,,$$

$$\begin{pmatrix} \widehat{\beta}_{i,1} \\ \vdots \\ \widehat{\beta}_{i,L} \end{pmatrix} \quad \text{by} \quad \begin{pmatrix} \widehat{\beta}_{i,1}^* \\ \vdots \\ \widehat{\beta}_{i,L}^* \end{pmatrix} \;=\; \Gamma^{1/2}\left\{ \begin{pmatrix} \widehat{\beta}_{i,1} \\ \vdots \\ \widehat{\beta}_{i,L} \end{pmatrix} + \Gamma^{-1}\gamma \right\}, \tag{19}$$

where $\widehat{m} = (\widehat{m}_1, \ldots, \widehat{m}_L)^{\top}$ and the $(L \times L)$ matrix $\Gamma = \int \widehat{m}(u)\widehat{m}(u)^{\top}\widehat{p}(u)\, du$. The vector γ has elements $\gamma_l = \int \widehat{m}_0(u)\widehat{m}_l(u)\widehat{p}(u)du$. By applying steps (19), $\widehat{m}_0$ is replaced by a function that minimizes $\int \widehat{m}_0^2(u)\widehat{p}(u)du$. This is evident because $\widehat{m}_0^*$ is orthogonal to the linear space spanned by $\widehat{m}_1, \ldots \widehat{m}_L$. By the second equation of steps (19), $\widehat{m}_1, \ldots, \widehat{m}_L$ are replaced by orthonormal functions in $L^2(\widehat{p})$.

In a second step, we proceed as in PCA and define a matrix B^* with $B_{l,l'}^* = \sum_{i=1}^I \widehat{\beta}_{i,l}^*\widehat{\beta}_{i,l'}^*$. We calculate the eigenvalues of B^*, $\lambda_1 > \cdots > \lambda_L$, and the corresponding eigenvectors $z_1, \ldots z_L$. Put $Z = (z_1, \ldots, z_L)$. Replace

$$\widehat{m}^* \quad \text{by} \quad \widehat{m}^{**} \;=\; Z^{\top}\widehat{m}^*\,, \tag{20}$$

and

$$
\begin{pmatrix} \widehat{\beta}^*_{i,1} \\ \vdots \\ \widehat{\beta}^*_{i,L} \end{pmatrix} \quad \text{by} \quad \begin{pmatrix} \widehat{\beta}^{**}_{i,1} \\ \vdots \\ \widehat{\beta}^{**}_{i,L} \end{pmatrix} = Z^\top \begin{pmatrix} \widehat{\beta}^*_{i,1} \\ \vdots \\ \widehat{\beta}^*_{i,L} \end{pmatrix}.
\tag{21}
$$

After application of (20) and (21), the orthonormal basis $\widehat{m}^{**}_1, \ldots, \widehat{m}^{**}_L$ is chosen so that $\widehat{\beta}^{**}_{i,1}$ has maximum variance, that is, $\widehat{m}^{**}_1$ is chosen so that as much as possible is explained by $\widehat{\beta}^{**}_{i,1} \widehat{m}^{**}_1$. Next $\widehat{m}^{**}_2$ is chosen to achieve maximum explanation by $\widehat{\beta}^{**}_{i,1} \widehat{m}^{**}_1 + \widehat{\beta}^{**}_{i,2} \widehat{m}^{**}_2$, and so forth, up to L. We point out that in all figures and tables and in the time series analysis to be presented in Section 2 we shall work with the quantities obtained after applying these orthonormalization steps.

The functions $\widehat{m}_l$ are not eigenfunctions of an operator as in the usual functional PCA [Ramsay and Silverman (1997)]. This is because we use a different norm, namely $\int f^2(u)\widehat{p}_i(u)du$, for each day. Through the norming procedure, the functions are chosen as eigenfunctions in an L-dimensional approximating linear space. The L-dimensional approximating spaces are not necessarily nested for increasing L. For this reason the estimates cannot be calculated by an iterative procedure that starts by fitting a model with L components, and that uses these components in the iteration step from L to $L + 1$ components. The calculation of $\widehat{m}_0, \ldots, \widehat{m}_L$ has to be redone for different choices of L.

2 THE FACTORS OF THE DAX INDEX IVS

2.1 Data Description and Preparation

The data set contains tick statistics on DAX futures contracts and DAX index options traded at the German-Swiss futures exchange EUREX in Frankfurt (Main) over the period from January 1998 to May 2001. The DAX option (ODAX) is of European type, is settled in cash, expires on the third Friday of the contract month, and has a minimum price movement of 0.1 index points. Expiries are the three nearest calendar months, the three following months of the cycle March, June, September, and December and the two following months of the cycle from June to December. For our purposes, we disregard data beyond half a year to expiry, as the liquidity is thin, in particular in the early years of our data set. The average daily trading volume over all strikes and expiries was about 140 000 contracts as of May, 2001. The futures contract is settled in cash as well, has a minimum price movement of half of an index point and the expiry months available are the three nearest months within the cycle March, June, September, December. The contract with the shortest time-to-maturity is by far the most liquid one, with an average daily trading volume of around 50 000 traded contracts as of May, 2001. In the data set, futures price and option price data are recorded per transaction, that is, each single trade is registered together with its price, contract size, and time of settlement (time is measured in seconds). The interest rate data we employ are one-, three-, and six-months FIBOR rates for the years 1998–1999

and EURIBOR rates[6] for the period 2000–2001 in daily frequency. The source is Thomson Financial Datastream. The interest rate data are linearly interpolated to approximate the riskless interest rate for the option's expiry. This is common practice [Dumas et al. (1998)].

Given the price data, implied volatilities have to be computed. To this end we cannot apply the Black (1976) formula, as it requires a liquid futures contract for each option expiry. This condition is not met by our data. Therefore, we work with the standard BS formula and derive implied DAX index values from the recorded futures prices. Since a futures price differs from the spot by the basis [Hull (2003, p. 75)], it may seem as a limitation that the analysis is based on futures prices rather than on spot prices. The DAX, however, is not a traded security: it is an index comprising 30 German blue chips. Thus, for option traders and market makers, trading the index at low transaction costs is possible only through futures contracts.[7] Backing out index implied volatilities from futures price data therefore is a natural process and corresponds to daily practice.

The implied DAX index values are recovered following Hafner and Wallmeier (2001), who use the same data: we pair each option price observation with the futures price F_t of the nearest available futures contract that was traded within a one-minute interval. The futures price observation is taken from the most heavily traded futures contract on that particular day. The no-arbitrage price of the underlying index in a frictionless market without dividends can be computed from the fair forward price formula via $S_t = F_t e^{-r_{T_F,t}(T_F-t)}$, where S_t and F_t denote the index and the futures price, respectively, T_F the expiry date of the futures contract, and $r_{T,t}$ the interest rate for the period $T-t$. It needs to be remarked that owing to the daily settlement of futures contracts, futures prices and forward prices are not equal, when interest rates are stochastic. However, for the time-to-maturities we consider in this work (up to half a year), we believe this difference to be negligible [see Hull (2003, p. 51--52) for a more detailed discussion and further references to this topic].

The DAX index is a capital-weighted performance index, that is, dividends less corporate tax are reinvested into the index [Deutsche Börse (2006)]. Therefore, dividend payments should have no impact on index options. However, when one uses only the discounted futures price to recover implied volatilities, implied volatilities of calls and puts can differ significantly. To accommodate for this fact, we apply a correction algorithm due to Hafner and Wallmeier (2001), which is described in Appendix B. The data are stored in the financial data base MD*base at the Center for Applied Statistics and Economics (CASE), Berlin.

Since trade-by-trade data may contain misprints or outliers, a filter is applied before estimating the model: observations with an implied volatility outside the

---

[6] FIBOR is the Frankfurt Interbank Offered Rate, which was a daily reference rate based on the interest rates at which banks offer to lend unsecured funds to other banks in the Deutsche Mark money market. With introducing the Euro in 1999, it was replaced by the EURIBOR, the European Interbank Offered Rate.

[7] To be very exact, there are now exchange-traded funds on the DAX that track the index remarkably close; they serve, however, as investment vehicles and are not used for hedging.

Table 1 Summary statistics of implied volatility data.

		Min.	Max.	Mean	Median	Stdd.	Skewn.	Kurt.
All	Time-to-maturity	0.028	2.014	0.131	0.083	0.148	3.723	23.373
	Moneyness	0.325	1.856	0.985	0.993	0.098	−0.256	5.884
	Implied volatility	0.041	0.799	0.279	0.256	0.090	1.542	6.000
1998	Time-to-maturity	0.028	2.014	0.134	0.081	0.148	3.548	22.957
	Moneyness	0.386	1.856	0.984	0.992	0.108	−0.030	5.344
	Implied volatility	0.041	0.799	0.335	0.306	0.114	0.970	3.471
1999	Time-to-maturity	0.028	1.994	0.126	0.083	0.139	4.331	32.578
	Moneyness	0.371	1.516	0.979	0.992	0.099	−0.595	5.563
	Implied volatility	0.047	0.798	0.273	0.259	0.076	0.942	4.075
2000	Time-to-maturity	0.028	1.994	0.130	0.083	0.151	3.858	23.393
	Moneyness	0.325	1.611	0.985	0.992	0.092	−0.337	6.197
	Implied volatility	0.041	0.798	0.254	0.242	0.060	1.463	7.313
2001	Time-to-maturity	0.028	0.978	0.142	0.083	0.159	2.699	10.443
	Moneyness	0.583	1.811	1.001	1.001	0.085	0.519	6.762
	Implied volatility	0.043	0.789	0.230	0.221	0.049	1.558	7.733

Summary statistics of the implied volatility data computed from DAX index option data for the period from January 1998 to May 2001, entire data set and annual subsamples. The figures are computed after filtering and data preparation.

range [4%, 80%] are dropped. Furthermore, we disregard all observations with a maturity of less than ten days [for similar approaches see Dumas et al. (1998) and Skiadopoulos et al. (1999)]. The filtering removed 635 000 observations from file, leaving us with a total number of about 2.24 million contracts, that is, around 2 600 observations per daily IVS. Clearly, this filter does not remove all potential outliers, but since our methodology relies on data of different observation days for the estimation of the factor functions, it is robust against spurious data points.

Table 1 gives a short summary of our IVS data. Most heavy trading occurs in short-term contracts, as is seen from the skewness of the term structure distribution of the observations. Median time-to-maturity is 30 days (0.083 years). Across moneyness, the distribution is slightly negatively skewed. This is because out-of-the-money puts are more heavily traded than out-of-the-money calls. Mean implied volatility over the sample period is 27.9%.

2.2 Model Setup

The implementation of the SFM requires decisions on a number of issues, most importantly the model size and the bandwidths. In this section, we outline the general procedure. Technicalities are transferred to Appendix A.

Because of skewness and positivity constraints, we regress log-implied volatility $Y_{i,j}$ on $X_{i,j} = (\kappa_{i,j}, \tau_{i,j})^\top$. The grid is regularly spaced and covers in

moneyness $\kappa \in [0.80, 1.20]$ and in time-to-maturity $\tau \in [0.05, 0.5]$ measured in years. To determine the model size L, we compute the portion of variance explained $1 - RV(L)$, where $RV(L)$ is the residual sum of squares per total variance [Equation (A1) in Appendix A]. We find that up to 89% of the variation are explained by $L = 1$ basis function, 94% by $L = 2$, and 97% by $L = 3$ basis functions. These findings are in line with the textbook literature of the standard PCA on the IVS [see e.g., Alexander (2001)]. Therefore, we decide for $L = 3$ basis functions as a sufficiently good approximation to the IVS. For measuring the goodness of fit and the progress of convergence we compute the L^2-measure Equation (A5) in the Appendix. Convergence is typically achieved after 25 cycles.

The choice of the smoothing parameter is essential in the practice of non- and semiparametric modeling. The literature offers a number of approaches, such as cross-validation procedures, penalization and bootstrap techniques, or plug-in methods for determining the bandwidths [see Härdle (1990, Section 5) for detailed discussions]. Since in this study we are interested in fitting the IVS, a natural approach for bandwidth choice can be based on the Akaike (1970) Information Criterion (AIC) and on the Schwarz (1978) Criterion (SC). More precisely, we work with the two computationally slightly differing versions of a weighted AIC and SC given in Equations (A3) and (A4). The reason for using weighted criteria is the uneven distribution of the data, which can lead to nonconvexities and which might result in unacceptably small bandwidths [see Fengler et al. (2003) for a first description of this problem]. We choose as weight function $p^{-1}(u)$ where $p(u)$ is the average design density. It can be shown that this choice ensures an equal weight everywhere.

Table 2 presents the criteria for a grid of bandwidths. It is seen—from bottom to top—that for a given bandwidth in moneyness (h_1) all criteria drop as the bandwidth in time-to-maturity (h_2) decreases, but that the criteria increase as the latter approaches the length of two weeks, i.e. when $h_2 \approx 0.03$ years. Criterion Ξ_{AIC_1} assumes its global minimum at $h = (h_1, h_2) = (0.04, 0.05)^\top$, Ξ_{Sc_1} at $h = (0.04, 0.06)^\top$. The second type of criteria proposes much smaller bandwidths: Ξ_{AIC_2} suggests $h = (0.02, 0.03)^\top$ and Ξ_{SC_2} proposes $h = (0.03, 0.04)^\top$, respectively. Since for the second criteria the functions still look quite volatile, we prefer the suggestion of the SC among the first group of criteria, that is, we take Ξ_{Sc_1} at $h^* = (0.04, 0.06)^\top$ as our optimal bandwidth vector. For further analysis, we check how the factor loadings and the basis functions change relative to this optimal bandwidth h^*, columns seven and eight in Table 2. As is seen, the bandwidth suggestions of Ξ_{AIC_1} and Ξ_{SC_1} are hardly distinguishable and the estimates are stable.

The choices for h and L are not independent. From this point of view, one may think about minimizing the information criteria over both parameters. Practical experience shows that for a given L, changes in the criteria from variations in h are small, compared to variations in L for a given h. Therefore, separating the two selection procedures can be justified. As an alternative to our constant bandwidth, one could consider bandwidths that are functions of the grid $h = h(u)$. This is discussed in Borak et al. (2005).

Table 2 Bandwidth selection for the factor model.

h_1	h_2	$\Xi_{AIC_1} \times 10^4$	$\Xi_{AIC_2} \times 10^3$	$\Xi_{SC_1} \times 10^4$	$\Xi_{SC_2} \times 10^3$	$V_{\widehat{\beta}}$	$V_{\widehat{m}}$
0.01	0.02	8.127	1.7562	8.3540	2.0461	0.042	0.096
0.01	0.03	5.386	1.3717	5.4825	1.5142	0.039	0.069
0.01	0.04	5.381	1.3732	5.4518	1.4762	0.038	0.043
0.01	0.05	5.373	1.3816	5.4282	1.4625	0.038	0.028
0.01	0.06	5.371	1.3952	5.4162	1.4620	0.038	0.021
0.01	0.07	5.376	1.4134	5.4144	1.4707	0.038	0.019
0.01	0.08	5.390	1.4369	5.4235	1.4874	0.038	0.019
0.02	0.02	8.037	1.7338	8.1459	1.8689	0.042	0.090
0.02	0.03	5.347	1.3560*	5.3946	1.4239	0.039	0.035
0.02	0.04	5.346	1.3636	5.3803	1.4134	0.038	0.020
0.02	0.05	5.345	1.3746	5.3718	1.4140	0.002	0.014
0.02	0.06	5.346	1.3893	5.3687	1.4220	0.001	0.010
0.02	0.07	5.353	1.4081	5.3718	1.4362	0.001	0.011
0.02	0.08	5.367	1.4319	5.3841	1.4568	0.002	0.014
0.03	0.02	8.009	1.7331	8.0811	1.8213	0.042	0.090
0.03	0.03	5.329	1.3585	5.3598	1.4031	0.038	0.025
0.03	0.04	5.327	1.3667	5.3494	1.3996*	0.002	0.015
0.03	0.05	5.326	1.3777	5.3435	1.4038	0.001	0.008
0.03	0.06	5.327	1.3921	5.3417	1.4138	0.000	0.005
0.03	0.07	5.333	1.4107	5.3457	1.4293	0.001	0.007
0.03	0.08	5.348	1.4342	5.3586	1.4507	0.002	0.012
0.04	0.02	8.018	1.7476	8.0720	1.8135	0.042	0.090
0.04	0.03	5.321	1.3743	5.3446	1.4079	0.038	0.022
0.04	0.04	5.318	1.3821	5.3354	1.4069	0.002	0.012
0.04	0.05	5.316*	1.3925	5.3294	1.4121	0.001	0.006
0.04	**0.06**	5.316	1.4062	**5.3273***	1.4226	n.a.	n.a.
0.04	0.07	5.322	1.4242	5.3312	1.4382	0.001	0.006
0.04	0.08	5.336	1.4472	5.3440	1.4596	0.002	0.011
0.05	0.02	8.086	1.7833	8.1292	1.8366	0.041	0.092
0.05	0.03	5.333	1.4083	5.3515	1.4357	0.038	0.020
0.05	0.04	5.328	1.4154	5.3420	1.4355	0.026	0.012
0.05	0.05	5.324	1.4249	5.3350	1.4409	0.026	0.007
0.05	0.06	5.323	1.4378	5.3320	1.4511	0.001	0.005
0.05	0.07	5.328	1.4550	5.3355	1.4664	0.001	0.008
0.05	0.08	5.341	1.4772	5.3479	1.4873	0.002	0.013
0.06	0.02	8.183	1.8424	8.2190	1.8880	0.041	0.094
0.06	0.03	5.373	1.4667	5.3887	1.4903	0.038	0.020
0.06	0.04	5.367	1.4728	5.3781	1.4902	0.026	0.014
0.06	0.05	5.361	1.4815	5.3697	1.4952	0.026	0.010
0.06	0.06	5.358	1.4934	5.3656	1.5048	0.026	0.009
0.06	0.07	5.362	1.5096	5.3685	1.5195	0.002	0.011
0.06	0.08	5.375	1.5309	5.3806	1.5396	0.003	0.015

Bandwidth selection for different choices of h_1 (moneyness) and h_2 (time-to-maturity in years); model size $L = 3$. For each selection criterion the optimal value is labeled with an asterisk, the optimal bandwidth eventually chosen according to SC$_1$ (column 5) is printed in bold. $V_{\widehat{\beta}}$ and $V_{\widehat{m}}$ measure the standard deviation of $\widehat{\beta}_h$ and $\widehat{m}_h$ relative to $\widehat{\beta}_{h^*}$ and $\widehat{m}_{h^*}$, i.e. $V_{\widehat{\beta}}^2(h_k) = (LI)^{-1} \sum_{l,i} (\widehat{\beta}_{i,l}(h_k) - \widehat{\beta}_{i,l}(h^*))^2$ and $V_{\widehat{m}}^2(h_k) = (LN)^{-1} \sum_{l,n} (\widehat{m}_l(u_n, h_k) - \widehat{m}_l(u_n, h^*) du)^2$, where $n = 1, \ldots, N$ runs over all points in the estimation grid and h_k over the values displayed in column one and two.

2.3 Estimation Results and VAR Modeling

Unlike the traditional literature on smoothing the IVS, our approach exploits the expiry effect of the implied volatility data and thus avoids global daily fits. Indeed, our bandwidth in the time-to-maturity dimension is so small that in a fit of a particular day, data belonging to contracts with two adjacent expiries do not enter together $\widehat{p}_i(u)$ in Equation (9) and $\widehat{q}_i(u)$ in Equation (10). To visualize the mechanics behind the algorithm consider Figure 4. Here, we illustrate the situation for a small number of trading days and for four highlighted grid points, which are depicted by the big bullets. The neighborhood that is realized using the kernel functions is sketched by the ellipses. It is seen for a given u', that is, for a given bullet point, that the quantities $\widehat{p}_i(u')$ and $\widehat{q}_i(u')$ must be zero most of the time, and only assume positive values for dates i, when observations are in the local neighborhood of u'. This happens because of the simple fact that the contracts move towards expiry.

Figures 5 and 6 display the estimated factor functions $\widehat{m}_1$ to $\widehat{m}_3$ after orthonormalization in $L^2(\hat{p})$. We do not display the invariant function $\widehat{m}_0$, as it turns out to be the zero function of the affine space fitted by the data. The remaining three functions exhibit interesting patterns: $\widehat{m}_1$ in Figure 5 is negative throughout, U-shaped and roughly constant across the term structure. Since this function belongs to the weights with highest variance, it can be interpreted as the time-dependent mean of the (log)-IVS, that is, as a shift function. Function $\widehat{m}_2$, depicted in Figure 6, changes sign around the at-the-money region, which implies that it drives the smile deformations of the IVS. Hence, it is the moneyness
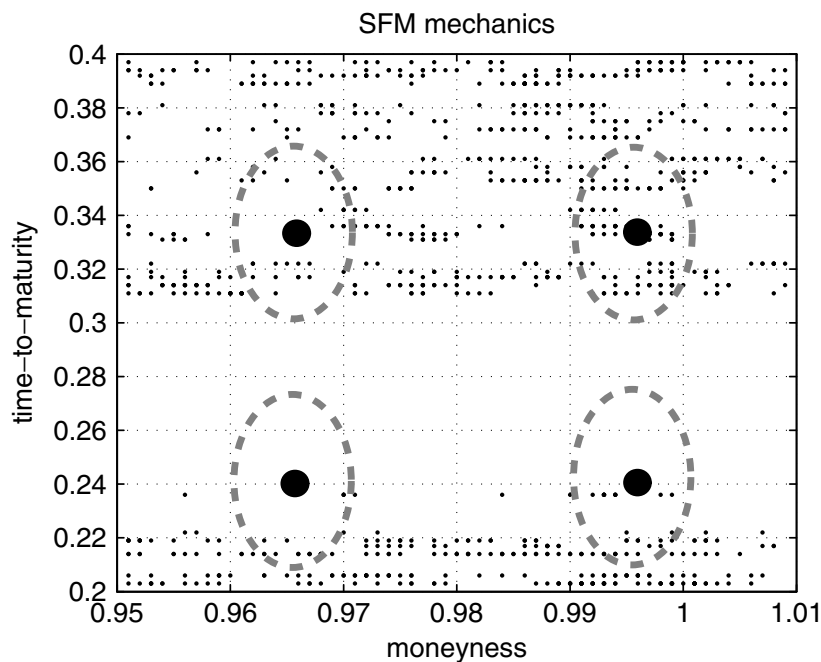
Figure 4 Sketch of the computational mechanics behind the SFM algorithm. The big bullets denote points from the estimation grid. The neighborhood is realized via the kernel functions and illustrated by the dotted ellipses. The small dots denote implied volatility data as observed for a sequence days, February to March 1998.
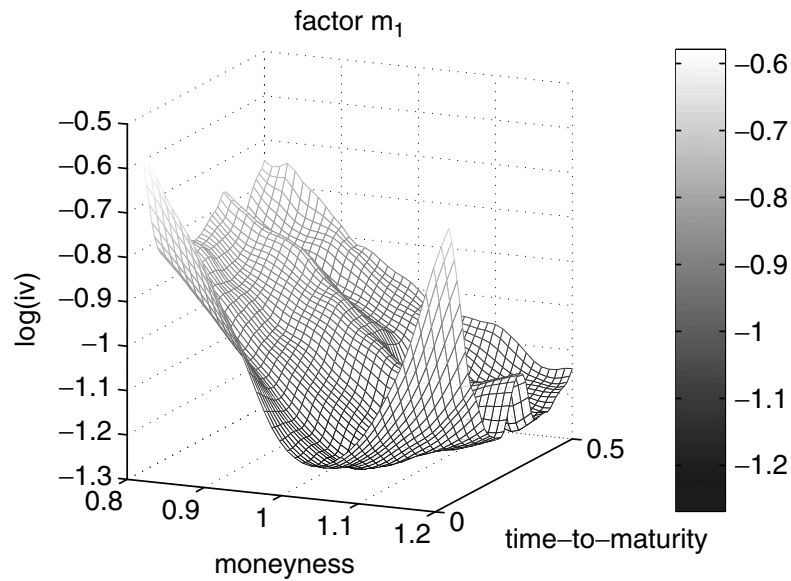
Figure 5 Factor $\widehat{m}_1$ of the SFM after orthonormalization in $L^2(\hat{p})$, where $\hat{p}(u) = I^{-1} \sum_{i=1}^{I} \hat{p}_i(u)$ is the average design density, that is, after applying steps (19) to (21). Bandwidths $h = (0.04, 0.06)$ chosen optimally according to SC_1, cf. Table 2. Estimated from ODAX data from January 1998 to May 2001.
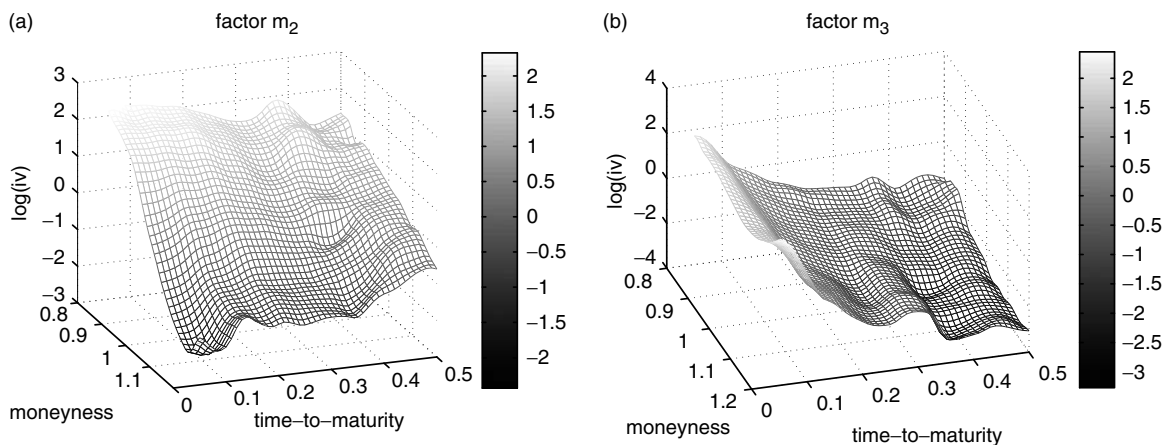
Figure 6 Factor $\widehat{m}_2$ (left) and $\widehat{m}_3$ (right) of the SFM after orthonormalization in $L^2(\hat{p})$, i.e. after applying steps (19) to (21). Bandwidths $h = (0.04, 0.06)$ chosen optimally according to SC_1, cf. Table 2. Estimated from ODAX data from January 1998 to May 2001.

slope function of the IVS. Finally, $\widehat{m}_3$ (also Figure 6) is positive for the very short term contracts and negative for contracts with a time-to-maturity of more than 0.1 years. Consequently, $\widehat{m}_3$ generates the term structure dynamics of the IVS [for similar interpretations see Skiadopoulos et al. (1999), Cont and da Fonseca (2002), and Fengler et al. (2003)].

To give an impression of the properties of the SFM, we compare the fit of the SFM with that of a Nadaraya–Watson estimator for a randomly chosen date from the sample, June 18, 1998. For the bandwidth choice in case of the Nadaraya–Watson estimator we compute a classical AIC and SC which

Figure 7 Fit comparisons. Left panel: Nadaraya–Watson estimator with bandwidths $h^*_{NW:AIC} = (h_1, h_2)^\top = (0.04, 0.12)^\top$ (solid line) and $h^*_{NW:SC} = (0.05, 0.19)^\top$ (dotted line), where h_1 is the bandwidth in moneyness and h_2 for time-to-maturity. Right panel: SFM using bandwidths $h^* = (0.04, 0.06)^\top$. Observed implied volatility data shown as gray dots, time-to-maturity is 29 days, data from June 18, 1998.
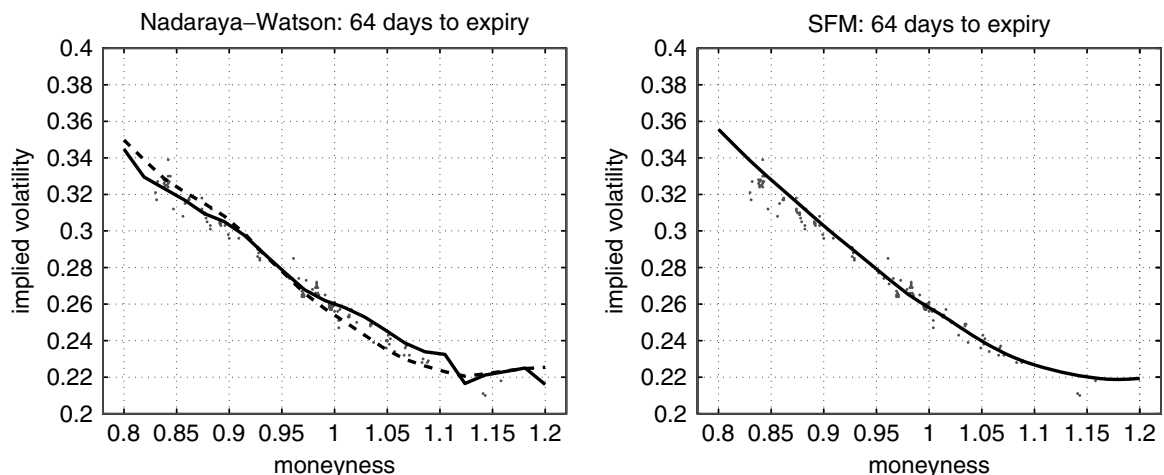
Figure 8 Fit comparisons. Left panel: Nadaraya–Watson estimator with bandwidths $h^*_{NW:AIC} = (h_1, h_2)^\top = (0.04, 0.12)^\top$ (solid line) and $h^*_{NW:SC} = (0.05, 0.19)^\top$ (dotted line), where h_1 is the bandwidth in moneyness and h_2 for time-to-maturity. Right panel: SFM using bandwidths $h^* = (0.04, 0.06)^\top$. Observed implied volatility data shown as gray dots, time-to-maturity is 64 days, data from June 18, 1998.

yield $h^*_{NW:AIC} = (h_1, h_2)^\top = (0.04, 0.12)^\top$ and $h^*_{NW:SC} = (0.05, 0.19)^\top$, respectively, as optimal bandwidths for moneyness h_1 and time-to-maturity h_2. Figures 7 to 10 show the slices through the surfaces at the observed expiries. Before studying the differences, we recall that the Nadaraya–Watson estimator is based on data of one single day, whereas the SFM takes advantage of the entire sample. Therefore, any reasonable comparison is necessarily unbalanced. Nevertheless, we believe that such a comparison can reveal the principal advantages the SFM offers over using a standard nonparametric estimator for fitting the IVS.
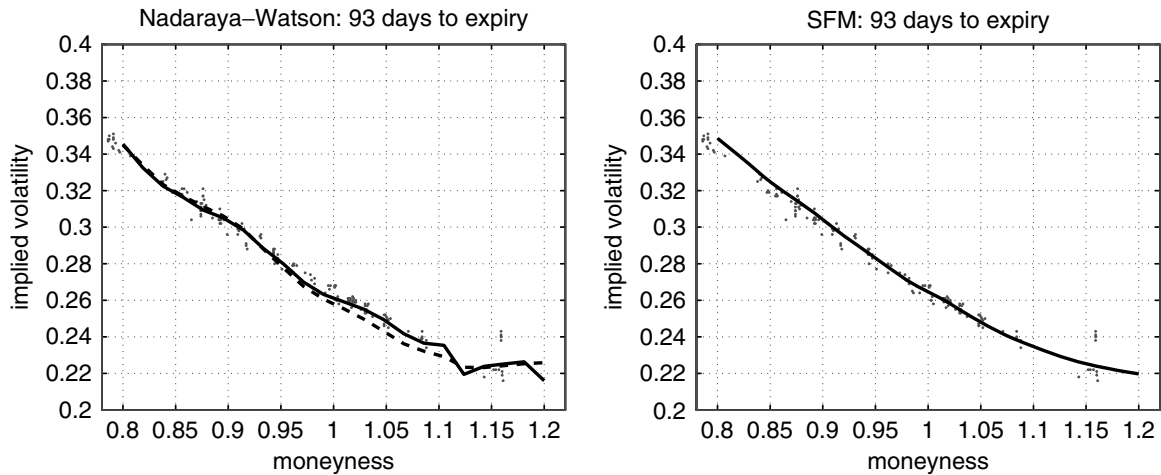
Figure 9 Fit comparisons. Left panel: Nadaraya–Watson estimator with bandwidths $h^*_{NW:AIC} = (h_1, h_2)^\top = (0.04, 0.12)^\top$ (solid line) and $h^*_{NW:SC} = (0.05, 0.19)^\top$ (dotted line), where h_1 is the bandwidth in moneyness and h_2 for time-to-maturity. Right panel: SFM using bandwidths $h^* = (0.04, 0.06)^\top$. Observed implied volatility data shown as gray dots, time-to-maturity is 93 days, data from June 18, 1998.
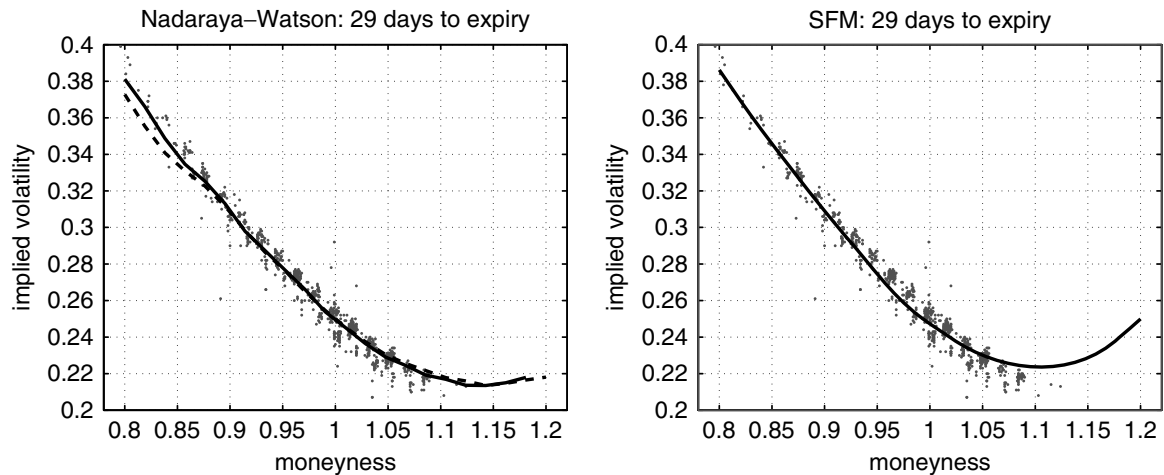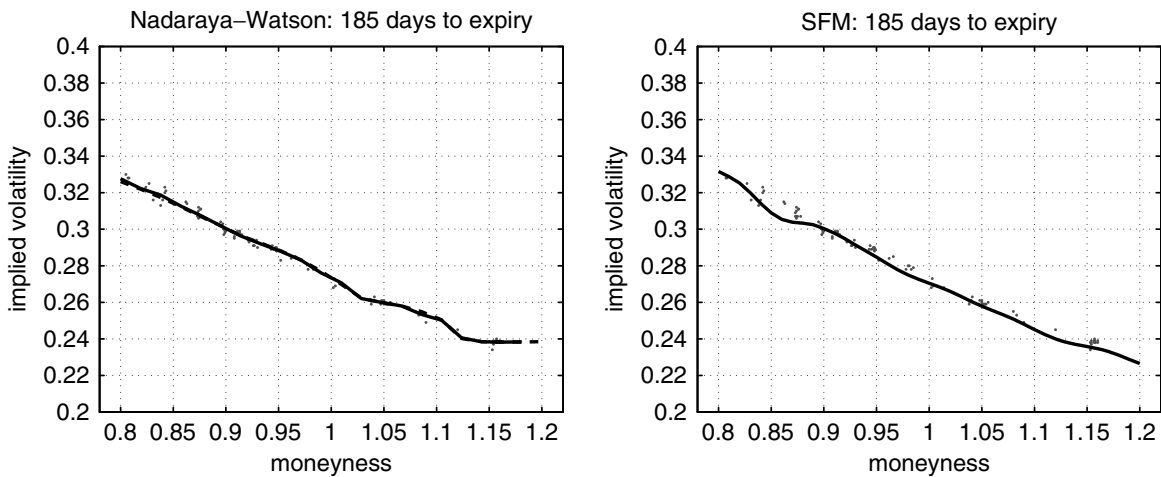

Figure 10 Fit comparisons. Left panel: Nadaraya–Watson estimator with bandwidths $h^*_{NW:AIC} = (h_1, h_2)^\top = (0.04, 0.12)^\top$ (solid line) and $h^*_{NW:SC} = (0.05, 0.19)^\top$ (dotted line), where h_1 is the bandwidth in moneyness and h_2 for time-to-maturity. Right panel: SFM using bandwidths $h^* = (0.04, 0.06)^\top$. Observed implied volatility data shown as gray dots, time-to-maturity is 185 days, data from June 18, 1998.

Figures 7 to 10 show that despite the smaller bandwidths, the SFM fit is smoother than the Nadaraya–Watson fit, in particular in comparison with the Nadaraya–Watson estimator based on the bandwidths as suggested by the AIC (solid lines). Necessarily, the Nadaraya–Watson estimate must be rough when the data are sparse, such as for the moneyness region greater than 1.1 in Figures 8 and 9. However, increasing the bandwidths to the size as proposed by the SC results in directional biases: for moneyness greater than 1.1 in Figures 8 and 9 the Nadaraya–Watson estimator (dotted line) tends to be below the actual data cloud. The SFM is not entirely unbiased either, as is seen in a small bump around

Table 3 Summary statistics of the factor series.

	Min.	Max.	Mean	Median	Stdd.	Skewn.	Kurt.
$\widehat{\beta}_1$	−1.504	−0.454	−1.191	−1.229	0.200	1.099	4.079
$\widehat{\beta}_2$	−0.109	0.074	−0.002	−0.003	0.033	0.009	2.710
$\widehat{\beta}_3$	−0.085	0.073	−0.002	0.002	0.023	−0.494	3.319

Descriptive summary statistics on the time series of the factor loadings $(\widehat{\beta}_1, \widehat{\beta}_2, \widehat{\beta}_3)^\top$ after application the orthonormalization of the factor functions, i.e. after applying steps (19) to (21). Data from January 1998 to May 2001.

moneyness 0.85 in Figure 10, but overall, the fit is superior. Another noteworthy feature can be observed in Figure 7 for moneyness bigger than 1.1. Note that on this day and for this time-to-maturity almost no observations are found beyond that point and the estimates of the Nadaraya–Watson estimator and the SFM are an extrapolation. The Nadaraya–Watson yields a flat extrapolation, which is due to the flat structure in the second time-to-maturity (presented in Figure 8), while the SFM produces the typical smile pattern. Of course, one cannot tell which extrapolation is the better one. But observing that the extrapolation of the SFM is based on a large sample and on data observations from the local neighborhood (though from other dates), the smiley SFM extrapolation may be considered as the more likely functional structure. This interpretation corresponds to the general knowledge about smile patterns in equity option data [see e.g. Rubinstein (1985) and Tompkins (2001)].

Figure 11 shows the time series of $\widehat{\beta}_1$ to $\widehat{\beta}_3$ and their correlograms.[8] By comparing the plot of $\widehat{\beta}_1$ with the time series of the mean ATM-implied volatility in Figure 2, the first factor loading is recognized as a shifted and reflected version of mean implied volatility levels. This observation neatly matches the aforementioned interpretation of the first factor function as a mean function of the IVS. Summary statistics of the loading time series are given in Table 3. An augmented Dickey–Fuller (ADF) test indicates a unit root for $\widehat{\beta}_1$ and $\widehat{\beta}_2$ at the 5% level (left part of Table 4). Thus, one can model first differences of the first two loading series. Alternatively, since the test statistics are not strongly significant, a model in levels is an equally good choice. Moreover, one avoids possible over-differencing. Therefore we prefer a VAR model of order p in the levels of the series[9], that is,

$$\beta_t = c + A_1 \beta_{t-1} + \cdots + A_p \beta_{t-p} + u_t ,\qquad(22)$$

---

[8] For all figures and in all subsequent steps in the time series modeling the factor loadings obtained after the orthonormalization of the factor functions are used, see Section 1.

[9] In slight abuse of notation, we remove the ''hat'' notation of the β from now on. We point out that the VAR estimation is carried out on the estimated series $\widehat{\beta}_t$.
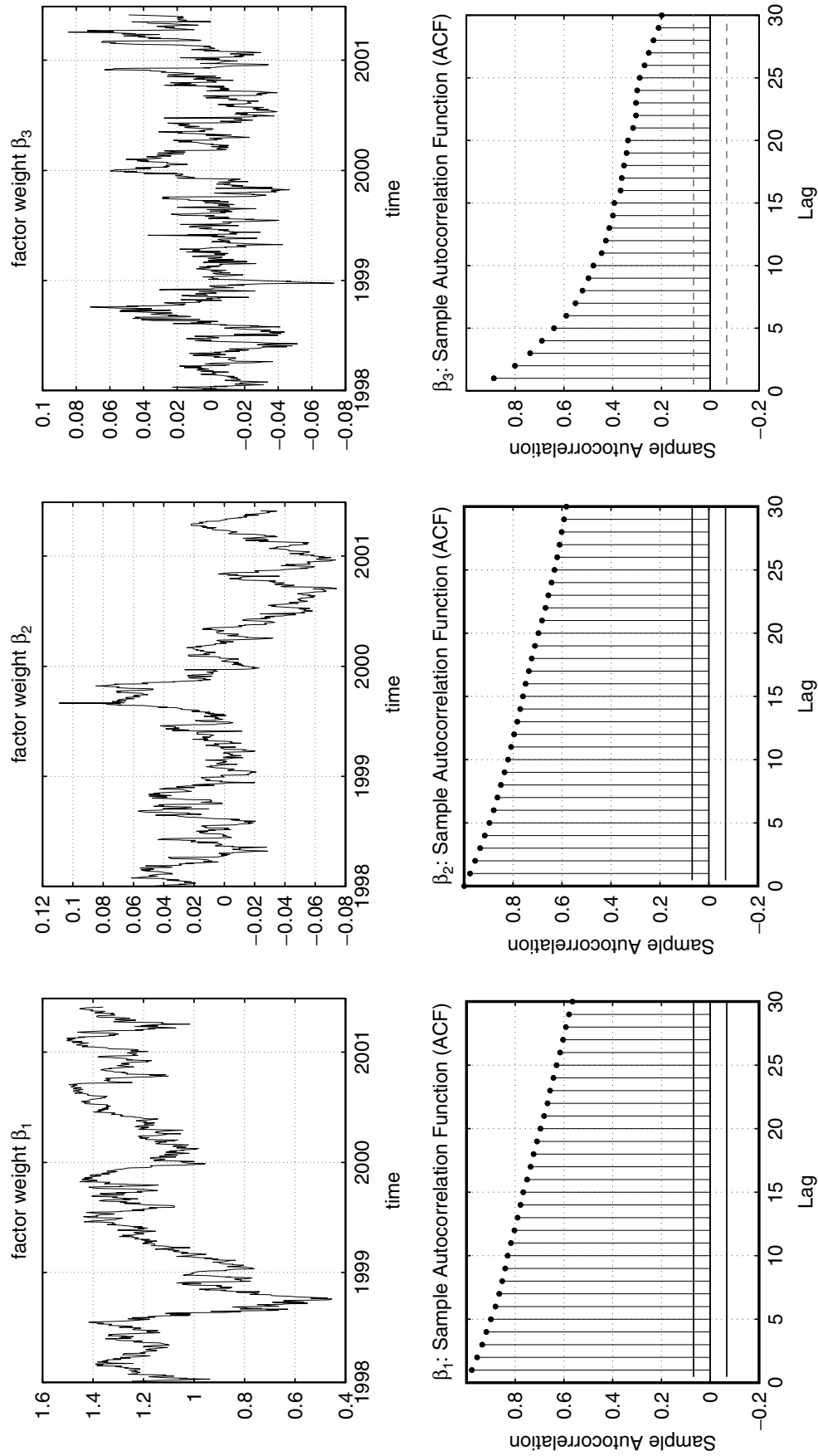
Figure 11 Upper panel: Time series of weights $(\widehat{\beta}_1, \widehat{\beta}_2, \widehat{\beta}_3)^\top$ after application of the steps (19) to (21), i.e. after orthonormalization of the factor functions. Bandwidths $h = (0.04, 0.06)$ chosen according to SC_1, cf. Table 2. Lower panel: autocorrelation functions; dotted horizontal lines indicate 95% confidence interval under the null hypothesis that the data are white noise.

Table 4 ADF diagnostics and correlation.

Coefficient	ADF test		Contemp. correlation		
	Test statistic	# of lags	$\widehat{\beta}_1$	$\widehat{\beta}_2$	$\widehat{\beta}_3$
$\widehat{\beta}_1$	-2.67	3	1.00	0.27***	0.39***
$\widehat{\beta}_2$	-3.01**	1		1.00	0.00
$\widehat{\beta}_3$	-5.63***	2			1.00

Left part: Augmented Dickey–Fuller (ADF) tests on $\widehat{\beta}_1$ to $\widehat{\beta}_3$ for the full IVS model, intercept included in each case. Third column gives the number of lags included in the ADF regression. For the choice of lag length, we start with four lags, and subsequently delete lag terms, until the last lag term is significant at least at the 5% level. MacKinnon critical values for rejecting the hypothesis of a unit root are -2.87 at the 5% significance level, and -3.44 at the 1% significance level. Right part: Contemporaneous correlation matrix; three asterisks denote an estimate significantly different from zero at the 1% level.

Table 5 Size of VAR model.

Suggested lag order	LR	AIC	HQ	SC
p	2	2	2	1

Choice of lag order p in VAR as suggested by the sequential modified likelihood ratio (LR) test, the Akaike (AIC), Hannan-Quinn (HQ), and Schwarz (SC) information criterion with maximum lag order $p_{max} = 12$; for details on these criteria in the context of VAR see Lütkepohl (1991).

where c is an $L \times 1$ vector of intercepts, $A_i, i = 1, ..., p$ are $L \times L$ parameter matrices and $u_t = (u_{1t}, u_{3t}, u_{3t})^\top$ an unobservable error term with mean zero and nonsingular covariance matrix Σ_u.

For the choice of the lag length p in VAR models, we employ standard information criteria as described, for example, in Lütkepohl (1991, Chap. 4.3). The sequential modified likelihood ratio test, the Akaike and Hannan-Quinn information criteria suggest a lag length of order $p = 2$, except for the SC, which prefers $p = 1$, see Table 5. Since it is known that the SC tends to propose lag orders which are too low to capture the salient dynamics in VAR models, we choose a VAR(2) model.

The residual diagnostics are given in Table 6. The adjusted Box–Pierce–Ljung Q-statistic and the LM test on serial correlation do not indicate the presence of serial correlation up to order 4 and 8, respectively. White's test for heteroskedasticity [White (1980)] rejects the null hypothesis of homoskedasticity in the residuals. Therefore we report in Table 7, together with the estimated coefficient matrix, heteroskedasticity-consistent standard errors which are obtained by applying a correction in the spirit of White (1980) [see Hafner and Herwartz (2002) for the consistent estimation of covariance matrices in heteroskedastic VAR models]. Besides the expected autoregressive patterns in own lagged observations, there are a number of remarkable cross dynamics. For instance, a positive shock in

Table 6 Portmanteau statistics.

	Q(4)	Q(8)	LM(4)	LM (8)
Test statistic	26.45	57.90	6.15	3.46
p-value	0.090	0.333	0.725	0.943
df of χ^2	18	54	9	9

We report the adjusted multivariate Box–Pierce–Ljung Q-statistics for residual serial correlation and the multivariate LM test statistic described in Greene (2003, Chap. 12.7). The number in parentheses gives the order up to which there is no serial correlation under the null hypothesis; all tests are chi-squared distributed with degrees of freedom (df) as given.

Table 7 VAR model.

	Equation		
Dependent variable	$\beta_{1,i}$	$\beta_{2,i}$	$\beta_{3,i}$
$\beta_{1,i-1}$	0.958***	0.002	−0.023*
	(0.049)	(0.010)	(0.014)
$\beta_{1,i-2}$	0.024	−0.005	0.024*
	(0.049)	(0.011)	(0.013)
$\beta_{2,i-1}$	−0.150	0.902***	0.114*
	(0.235)	(0.066)	(0.067)
$\beta_{2,i-2}$	0.104	0.071*	−0.109*
	(0.235)	(0.064)	(0.066)
$\beta_{3,i-1}$	−0.223	−0.055*	0.795***
	(0.157)	(0.046)	(0.062)
$\beta_{3,i-2}$	0.329*	0.053*	0.107*
	(0.155)	(0.043)	(0.058)
c	−0.022***	−0.004**	0.001
	(0.011)	(0.002)	(0.003)
$\overline{R}^2$	0.958	0.956	0.797

Estimation results of an VAR(2) of the factor loadings β_i and an intercept denoted by c. Heteroskedasticity-consistent standard errors are given in parentheses. $\overline{R}^2$ denotes the adjusted coefficient of determination. Three asterisks denote significance at a 1% level, two asterisks significance at a 5% level, and one significance at a 10% level.

β_1, that is, in the level component of the IVS, has a negative impact on the term structure component β_3. Given the shape of $\widehat{m}_3$, this shock implies a flattening term structure. This interpretation is well understood by noting that the IVS is typically upward sloping and that an increase in the overall risk perception of the market causes short-term implied volatilities to rise much stronger than long-term implied volatilities. This pattern necessarily implies a flatter term structure; furthermore, this case has to be distinguished from a positive shock in downside risk, that is, β_2, which implies a more pronounced term structure (positive impact on β_3).

There are also significant relationships from lagged β_3 on β_1 and β_2. However, an impulse–response analysis (not shown for want of space) suggests that only the shocks on β_3 have a permanent impact, while shocks from β_3 on β_1 and β_2 have none. Contrary to shocks on the term structure, it is to be observed that the interactions between the lags of level and moneyness shocks, that is, β_1 and β_2, are not significant, while their contemporaneous observations are significantly positively correlated (see the right part of Table 4). Thus, positive shocks tend to increase both the levels, β_1, and the downside risk, β_2, always instantly. This is common market experience. Overall, the VAR(2) model appears to capture the dynamics of the factor loadings, which generate the propagation of the IVS through time, in a suitable and well-interpretable way.

3   SUMMARY AND OUTLOOK

In this study we present a modeling approach to the IVS called SFM. Unlike other methods, our approach is tailored to the degenerated design of implied volatility data by fitting basis functions in the local neighborhood of the design points only. In comparison to a Nadaraya–Watson estimator, we demonstrate that the SFM can mitigate bias effects. Using transactions-based DAX index implied volatility data from January 1998 to May 2001, we recover three basis functions and three time series of factor loadings that generate the dynamics of the IVS. The factor functions allow for an interpretation as level, moneyness, and term structure slope effects. We study the time series properties of the parameter weights and complete the modeling approach by fitting vector autoregressive models. Standard lag-length criteria suggest that the VAR(2) model captures best the dynamics of the three-dimensional factor series.

There are important topics for further research. First, given the VAR(2) time series model of the factor loadings, the SFM can be integrated into scenario and worst case simulations of portfolios, for example, along the lines of Jamshidian and Zhu (1997). The SFM can be used to generate plausible deformations of the IVS, which should allow for better vega hedges of option portfolios. Second, from a theoretical point of view, the SFM has a natural relation with Kalman filtering. Writing our model more compactly as

$$\Theta_p(B)\beta_i = u_i \tag{23}$$

$$Y_{i,j} = m_0(X_{i,j}) + \sum_{l=1}^{L} \beta_{i,l} m_l(X_{i,j}) + \varepsilon_{i,j} , \tag{24}$$

where u_i and $\varepsilon_{i,j}$ are noise, and where $\Theta_p(B) \overset{\text{def}}{=} 1 - \theta_1 B - \theta_2 B^2 \ldots - \theta_p B^p$ denotes a polynomial of order p in the backshift operator B, we get the state-space representation. The state equation (23) depending on a parameter vector θ relates the (unobservable) state i of the system to the previous step $i-1$, while the measurement equation (24) relates the state to the measurement, the IVS in our

case. In difference to our work the time series modeling of β_i could be achieved jointly with the estimation of the factor functions.

We think that this modeling approach has potential in other fields of financial research, most naturally in interest rate modeling. Since interest rates are inferred from bond prices, there too exists an expiry effect. In this sense, the SFM might complement the modeling strategies proposed by Ghysels and Ng (1989), Linton et al. (2001), and Diebold and Li (2006).

APPENDIX A: MODEL SELECTION

Here, we outline the selection procedures for the SFM. For the model size we use the residual sum of squares for different L:

$$RV(L) \overset{\text{def}}{=} \frac{\sum_i^I \sum_j^{J_i} \left\{ Y_{i,j} - \sum_{l=0}^L \widehat{\beta}_{i,l} \, \widehat{m}_l(X_{i,j}) \right\}^2}{\sum_i^I \sum_j^{J_i} (Y_{i,j} - \overline{Y})^2} \,, \tag{A1}$$

where $\overline{Y}$ denotes the overall mean of the observations.

For a data-driven choice of bandwidths, we propose a weighted AIC, since the distribution of the observations is very uneven. For a given weight function w, consider:

$$\triangle(m_0, \ldots, m_L) \overset{\text{def}}{=} \mathsf{E} \frac{1}{N} \sum_{i,j} \{ Y_{i,j} - \sum_{l=0}^L \beta_{i,l} m_l(X_{i,j}) \}^2 \, w(X_{i,j}) \,, \tag{A2}$$

for functions $m_0, \ldots, m_L$. The expectation operator is denoted by E. We choose bandwidths such that $\triangle(\widehat{m}_0, \ldots, \widehat{m}_L)$ is minimal. According to the AIC this is asymptotically equivalent to minimizing:

$$\Xi_{AIC_1} \overset{\text{def}}{=} \frac{1}{N} \sum_{i,j} \{ Y_{i,j} - \sum_{l=0}^L \widehat{\beta}_{i,l} \widehat{m}_l(X_{i,j}) \}^2 \, w(X_{i,j}) \exp \left\{ 2 \frac{L}{N} K_h(0) \int w(u) du \right\} . \tag{A3}$$

Alternatively, one might consider the computationally more convenient criterion:

$$\Xi_{AIC_2} \overset{\text{def}}{=} \frac{1}{N} \sum_{i,j} \{ Y_{i,j} - \sum_{l=0}^L \widehat{\beta}_{i,l} \widehat{m}_l(X_{i,j}) \}^2 \, \exp \left\{ 2 \frac{L}{N} K_h(0) \frac{\int w(u) \, du}{\int w(u) p(u) \, du} \right\} . \tag{A4}$$

Putting $w(u) \overset{\text{def}}{=} 1$, delivers the common AIC. This, however, does not take into account the quality of the estimation at the boundary regions or in regions where data are sparse, since in these regions $p(u)$ is small. We choose $w(u) \overset{\text{def}}{=} p^{-1}(u)$, which can be shown to give equal weight everywhere. We do not punish for the number parameters $\widehat{\beta}_{i,l}$ that are employed to model the time series. This can be neglected because we use a finite dimensional model for the dynamics of $\beta_{i,l}$. The

corresponding penalty term is negligible compared to the smoothing penalty term. The SC are obtained by replacing the "2" in Equations (A3) and (A4) by $\log(N)$.

Convergence of the iterations is measured by

$$Q^2(r) \overset{\text{def}}{=} \sum_{i=1}^{I} \int \left(\sum_{l=0}^{L} \widehat{\beta}_i^{(r)} \widehat{m}_l^{(r)}(u) - \widehat{\beta}_i^{(r-1)} \widehat{m}_l^{(r-1)}(u) \right)^2 du \,. \tag{A5}$$

The rth cycle of the estimation is denoted by (r). Iterations are stopped when $Q^2(r) \le 10^{-5}$.

B DIVIDEND CORRECTION SCHEME

When the DAX spot price is approximated by simply discounting the DAX futures price for calculating implied volatilities, it can turn out that implied volatilities of calls and puts at same strikes diverge. As an explanation, Hafner and Wallmeier (2001) argue that due to German tax laws which discriminated foreign investors against domestic investors and which were effective till end of 2002, the marginal investor can be different from the one actually assumed to compute the DAX index. To foreign investors this tax scheme has the same impact as a dividend payment on an unprotected option, that is, it drives a wedge into the option prices and hence into implied volatilities. In order to obtain this unknown residual dividend $\Delta\hat{D}$, Hafner and Wallmeier (2001) solve the dividend-adjusted formula of a fair forward price for the spot and insert into the dividend-adjusted put–call parity.

For an estimate of $\Delta\hat{D}$, one identifies pairs of puts and calls of the same strikes and the same expiry date traded within a five minutes interval. The residual dividend for each pair is computed. $\Delta\hat{D}$ is estimated by the median of these residual dividends of all pairs for a given maturity at day t. Implied volatilities are recovered using as index value $\tilde{S}_t = F_t e^{-r_{T_F,t}(T_F-t)} + \Delta\hat{D}$. This procedure is applied on a daily basis throughout the entire data set from January 1998 to May 2001. Observations, for which the put–call pairs could not be identified, are eliminated.

REFERENCES

Aït-Sahalia, Yacine, and Jefferson Duarte. (2003). "Nonparametric option pricing under shape restrictions." *Journal of Econometrics 116*, 9–47.

Aït-Sahalia, Yacine, and Andrew Lo. (1998). "Nonparametric estimation of state-price densities implicit in financial asset prices." *Journal of Finance 53*, 499–548.

Aït-Sahalia, Yacine, Yubo Wang, and Francis Yared. (2001). "Do options markets correctly price the probabilities of movement of the underlying asset?" *Journal of Econometrics 102*, 67–110.

Akaike, Hirotugo. (1970). "Statistical predictor identification." *Annals of the Institute for Statistical Mathematics 22*, 203–217.

Alexander, Carol. (2001). *Market Models*, John Wiley & Sons, New York.

Alexander, Carol, and Leonardo M. Nogueira. (2007). ''Model-free hedge ratios and scale-invariant models.'' *Journal of Banking and Finance*. Forthcoming.

Andersen, Torben G., and Bent E. Sørensen. (1996). ''GMM estimation of a stochastic volatility model: A Monte Carlo study.'' *Journal of Business and Economic Statistics 14*, 328–352.

Andersen, Torben G., Hyung-Jin Chung, and Bent E. Sørensen. (1999). ''Efficient method of moments estimation of a stochastic volatility model: A Monte Carlo study.'' *Journal of Econometrics 91*, 61–87.

Bakshi, Gurdip, Charles Cao, and Zhiwu Chen. (2000). ''Do call and underlying prices always move in the same direction?'' *Review of Financial Studies 13*(3), 549–584.

Barndorff-Nielsen, Ole E. (1997). ''Normal inverse Gaussian distributions and stochastic volatility modelling.'' *Scandinavian Journal of Statistics 24*, 1–13.

Barndorff-Nielsen, Ole E. and Neil Shepard. (2001). ''Non-Gaussian Ornstein-Uhlenbeck-models and some of their uses in financial economics.'' *Journal of the Royal Statistical Society B 63*, 167–241.

Bates, David. S. (1996). ''Jumps and stochastic volatility: Exchange rate processes implicit in deutsche mark options.'' *Review of Financial Studies 9*, 69–107.

Bates, David. S. (2000). ''Post-'87 crash fears in the S&P 500 futures option market.'' *Journal of Econometrics 94*, 181–238.

Bates, David. S. (2006). ''Maximum likelihood estimation of latent affine processes.'' *Review of Financial Studies 19*(3), 909–965.

Benko, Michal, Matthias R. Fengler, Wolfgang Härdle, and Milos Kopa. (2007). ''On extracting information implied in options.'' *Computational Statistics*. Forthcoming.

Black, Fischer. (1976). ''Studies of stock price volatility changes.'' *Proceedings of the 1976 Meetings of the American Statistical Association* 177–181.

Black, Fischer, and Myron Scholes. (1973). ''The pricing of options and corporate liabilities.'' *Journal of Political Economy 81*, 637–654.

Borak, Szymon, Matthias R. Fengler, and Wolfgang Härdle. (2005). ''DSFM fitting of implied volatility surfaces.'' *Conference proceedings of the Fifth International Conference on Intelligent Systems Design and Applications*.

Borak, Szymon, Wolfgang Härdle, Enno Mammen, and Beyong Park. (2006). ''Time series modelling with semiparametric factor dynamics.'' *CASE Discussion Paper*, Humboldt-Universität zu Berlin.

Britten-Jones, Mark, and Anthony J. Neuberger. (2000). ''Option prices, implied price processes, and stochastic volatility, *Journal of Finance 55*(2), 839–866.

Broadie, Mark, Jérôme Detemple, Eric Ghysels, and Olivier Torrès. (2000a). ''American options with stochastics dividends and volatility: A nonparametric investigation.'' *Journal of Econometrics 94*, 53–92.

Broadie, Mark, Jérôme Detemple, Eric Ghysels, and Olivier Torrès. (2000b). ''Nonparametric estimation of American options exercise boundaries and call prices.'' *Journal of Economic Dynamics and Control 24*, 1829–1857.

Cai, Zongwu, Jianqing Fan, and Qiwei Yao. (2000). ''Functional-coefficient regression models for nonlinear time series, *Journal of the American Statistical Association 95*, 941–956.

Carr, Peter, Héliette Geman, Dilip B. Madan, and Marc Yor. (2003). ''Stochastic volatility for Lévy processes.'' *Mathematical Finance 13*(3), 345–382.

Chacko, George, and Luis M. Viceira. (2003). ''Spectral GMM estimation of continuous-time processes.'' *Journal of Econometrics 116*, 259–292.

Chernov, Mikhail, and Eric Ghysels. (2000). ''A study towards a unified approach to the joint estimation of objective and risk neutral measures for the purpose of options valuation, *Journal of Financial Economics 56*(3), 407−458.

Connor, Gregory, and Oliver Linton. (2006). ''Semiparametric estimation of a characteristic-based factor model of stock returns.'' *Empirical Finance*. Forthcoming.

Connor, Gregory, Robert A. Korajczyk, and Oliver Linton. (2006). ''The common and specific components of dynamic volatility.'' *Journal of Econometrics 132*, 231−255.

Cont, Rama, and José da Fonseca. (2002). ''The dynamics of implied volatility surfaces.'' *Quantitative Finance 2*(1), 45−60.

Deutsche Börse (2006). *''Guide to the Equity Indices of Deutsche Börse*, 5.12 edn, Deutsche Börse AG, 60485 Frankfurt am Main, Germany.

Diebold, Francis X., and Canlin Li. (2006). ''Modeling and forecasting the term structure of government bond yields.'' *Journal of Econometrics 130*, 337−364.

Duffie, Darrell, and Kenneth Singleton. (1993). ''Simulated moments estimation of markov models of asset prices.'' *Econometrica 61*, 929−952.

Duffie, Darrell, Jun Pan, and Kenneth Singleton. (2000). ''Transform analysis and asset pricing for affine jump-diffusions.'' *Econometrica 68*, 1343−1376.

Dumas, Bernard, Jeff Fleming, and Robert E. Whaley. (1998). ''Implied volatility functions: Empirical tests.'' *Journal of Finance 53*(6), 2059−2106.

Eraker, Bjørn, Michael Johannes, and Nicholas Polson. (2003). ''The impact of jumps in volatility and returns.'' *Journal of Finance 58*, 1269−1300.

Fan, Jianqin, Qiwei Yao, and Zongwu Cai. (2003). ''Adaptive varying-coefficient linear models.'' *Journal of the Royal Statistical Society B 65*, 57−80.

Fengler, Matthias R., Wolfgang Härdle, and Christophe Villa. (2003). ''The dynamics of implied volatilities: A common principle components approach.'' *Review of Derivatives Research 6*, 179−202.

Fridman, Moshe, and Lawrence Harris. (1998). ''A maximum likelihood approach for non-Gaussian stochastic volatility models.'' *Journal of Business and Economic Statistics 16*, 284−291.

Gallant, A. Ronald, and George Tauchen. (1998). ''Reprojecting partially observed systems with application to interest rate diffusions.'' *Journal of the American Statistical Association 93*, 10−24.

Gallant, A. Ronald, and George Tauchen. (2007). ''Simulated score methods and indirect inference for continuous-time models.'' *in* Yacine Aït-Sahalia and Lars Hansen (eds).'' *Handbook of Financial Econometrics*, North Holland, Amsterdam. Forthcoming.

Garcia, René, and Eric Renault. (1998a). ''A note on hedging in ARCH and stochastic volatility option pricing models.'' *Mathematical Finance 8*(2), 153−161.

Garcia, René, and Eric Renault. (1998b). ''Risk aversion, intertemporal substitution and option pricing.'' *Working Paper 98s-02*, CIRANO.

Garcia, René, Eric Ghysels, and Eric Renault. (2007). ''The econometrics of option pricing models.'' *in* Yacine Aït-Sahalia and Lars Hansen (eds).'' *Handbook of Financial Econometrics*, North Holland, Amsterdam. Forthcoming.

Ghysels, Eric, and Serena Ng. (1989). ''A semiparametric factor model of interest rates and tests of the affine term structure.'' *Review of Economics and Statistics 80*, 535−548.

Gouriéroux, Christian, Alain Monfort, and Eric Renault. (1993). ''Indirect inference.'' *Journal of Applied Econometrics 8*, S85−S118.

Greene, William. (2003). *Econometric Analysis*, 5th edn, Prentice Hall, New Jersey.

Hafner, Christian M., and Helmut Herwartz. (2002). ''Testing for vector autoregressive dynamics under heteroskedasticity, Econometric Institute Report EI 2002-36, Erasmus University Rotterdam, Netherlands.

Hafner, Reinhold, and Martin Wallmeier. (2001). ''The dynamics of DAX implied volatilities.'' *International Quarterly Journal of Finance 1*(1), 1–27.

Härdle, Wolfgang. (1990). *''Applied Nonparametric Regression*, Cambridge University Press, Cambridge, UK.

Harrison, J. Michael, and David Kreps. (1979). ''Martingales and arbitrage in multiperiod securities markets.'' *Journal of Economic Theory 20*, 381–408.

Harvey, Andrew, Esther Ruiz, and Neil Shephard. (1994). ''Multivariate stochastic variance models.'' *Review of Economic Studies 61*, 247–264.

Hastie, Trevor, and Robert Tibshirani. (1990). *Generalized additive models*, Chapman and Hall, London.

Hastie, Trevor, and Robert Tibshirani. (1993). ''Varying-coeffcient models.'' *Journal of the Royal Statistical Society B 55*(4), 757–796.

Heston, Steven. (1993). ''A closed-form solution for options with stochastic volatility with applications to bond and currency options.'' *Review of Financial Studies 6*, 327–343.

Hull, John. (2003). *''Options, Futures, and Other Derivatives*, 5th edn, Prentice Hall, New Jersey, USA.

Hull, John, and Alan White. (1987). ''The pricing of options on assets with stochastic volatilities.'' *Journal of Finance 42*, 281–300.

Jacquier, Eric, Nicholas Polson, and Peter Rossi. (1994). ''Bayesian analysis of stochastic volatility models.'' *Journal of Business and Economic Statistics 12*, 1–19.

Jamshidian, Farshid, and Yu Zhu. (1997). ''Scenario simulation: Theory and methodology.'' *Finance and Stochastics 1*, 43–67.

Jeffrey, Andrew, Dennis Kristensen, Oliver Linton, Thong Nguyen, and Peter Phillips. (2004). ''Nonparametric estimation of a multifactor Heath-Jarrow-Morton model: An integrated approach.'' *Journal of Financial Econometrics 2*(2), 251–289.

Kauermann, Göran, Gerhard Tutz, and Josel Brüderl. (2005). ''The survival of newly founded firms: a case-study into varying-coefficient models.'' *Journal of the Royal Statistical Society A 168*(1), 145–158.

Linton, Oliver, Enno Mammen, Jens Perch Nielsen, Carsten Tanggaard. (2001). ''Yield curve estimation by kernel smoothing.'' *Journal of Econometrics 105*(1), 185–223.

Lütkepohl, Helmut. (1991). *Introduction to Multiple Time Series Analysis*, Springer-Verlag, Berlin, Heidelberg.

Mammen, Enno, Oliver Linton, and Jens Perch Nielsen. (1999). ''The existence and asymptotic properties of a backfitting projection algorithm under weak conditions.'' *Annals of Statistics 27*(5), 1443–1490.

Manaster, Steven, and Gary Koehler. (1982). ''The calculation of implied variances from the Black-and-Scholes model: A note.'' *Journal of Finance 37*, 227–230.

Melino, Angelo, and Stuart Turnbull. (1990). ''Pricing foreign currency options with stochastic volatility.'' *Journal of Econometrics 45*, 239–265.

Pan, Jun. (2002). ''The jump-risk premia implicit in options: evidence from an integrated time-series study.'' *Journal of Financial Economics 63*, 3–50.

Pastorello, Sergio, Valentin Patilea, and Eric Renault. (2003). ''Iterative and recursive estimation of structural non-adaptive models.'' *Journal of Business and Economic Statistics 21*(4), 449–509.

Ramsay, Jim O. and Bernard Silverman. (1997). *Functional Data Analysis*, Springer-Verlag, Berlin, Heidelberg.

Renault, Eric. (1997). "Econometric models of option pricing errors." *in* David M. Kreps and Kenneth F. Wallis (eds)." *Advances in Economics and Econometrics, Seventh World Congress, Econometric Society Monographs*, Cambridge University Press 223–278.

Renault, Eric, and Nizar Touzi. (1996). "Option hedging and implied volatilities in a stochastic volatility model." *Mathematical Finance 6*(3), 279–302.

Rubinstein, Mark (1985). "Nonparametric tests of alternative option-pricing models using all reported trades and quotes on the 30 most active CBOE option classes from August 23, 1976 through August 31, 1978." *Journal of Finance 40*, 455–480.

Schöbel, Rainer, and Jianwei Zhu. (1999). "Stochastic volatility with an Ornstein-Uhlenbeck process: An extension." *European Finance Review 3*, 23–46.

Schwarz, Gideon. (1978). "Estimating the dimension of a model." *Annals of Statistics 6*, 461–464.

Skiadopoulos, George, Stuart Hodges, and Les Clewlow. (1999). "The dynamics of the S&P 500 implied volatility surface." *Review of Derivatives Research 3*, 263–282.

Stone, Charles. (1986). "The dimensionality reduction principle for generalized additive models." *The Annals of Statistics 14*, 592–606.

Tompkins, Robert. (2001). "Implied volatility surfaces: Uncovering regularities for options on financial futures." *European Journal of Finance 7*(3), 198–230.

White, Halbert. (1980). "A heteroskedasticity-consistent covariance matrix and a direct test for heteroskedasticity." *Econometrica 48*, 817–838.

Wiggins, James. (1987). "Option values under stochastic volatility." *Journal of Financial Economics 19*, 351–372.

Zhu, Yingzi, and Marco Avellaneda. (1997). "An E-ARCH model for the term-structure of implied volatility of FX options." *Applied Mathematical Finance 4*, 81–100.

Routledge
Taylor & Francis Group

Semiparametric diffusion estimation and application to a stock market index

WOLFGANG HÄRDLE†, TORSTEN KLEINOW*‡, ALEXANDER KOROSTELEV§,
CAMILLE LOGEAY¶ and ECKHARD PLATEN‖

†CASE—Center for Applied Statistics and Economics, Institut für Statistik und Ökonometrie,
Humboldt-Universität zu Berlin, Berlin, Germany
‡Department of Actuarial Mathematics and Statistics, Heriot-Watt University,
Edinburgh, Scotland, UK
§Wayne State University, Detroit, MI, USA
¶Deutsches Institut für Wirtschaftsforschung, Berlin, Germany
‖School of Finance and Economics, University of Technology Sydney,
PO Box 123, Broadway, Sydney NSW 2007, Australia

1

The analysis of diffusion processes in financial models is crucially dependent on the form of the drift and diffusion coefficient functions. A new model for a stock market index process is proposed in which the index is decomposed into an average growth process and an ergodic diffusion. The ergodic diffusion part of the model is not directly observable. A methodology is developed for estimating and testing the coefficient functions of this unobserved diffusion process. The estimation is based on the observations of the index process and uses semiparametric and non-parametric techniques. The testing is performed via the wild bootstrap resampling technique. The method is illustrated on S&P 500 index data.

Keywords: Diffusion; Identification; Continuous-time financial models; Semiparametric methods; Kernel smoothing; Bootstrap

1. Introduction

The analysis of time series and diffusion models with stochastic variance and covariance has been very intense in the last decade. In particular, the research on stochastic volatility models in finance has been driven in several directions; see Ghysels *et al.* (1996) and Frey (1997) for a survey. One direction is related to modern continuous-time finance where it is necessary to relate empirically observed volatility or implied volatility data to theoretically derived stochastic volatility models (Genon-Catalot *et al.* 2002). Another direction focuses on time series models for asset price and volatility, which provide a well-established framework for testing and modelling nonlinear variance structures in finance (Engle and Bollerslev 1986).

As shown by Nelson (1990) and Duan (1997), standard time series models, such as ARCH, GARCH, etc.,

converge for vanishing time step size between observations towards corresponding diffusion models. These diffusion models are fully characterized by their drift and diffusion coefficient functions. The particular choice of the time discretization is not essential for a discrete time approximation of a diffusion process as long as the maximum time step size is small enough. Since real data are very frequently observed it is natural to interpret these as values of a discretely observed diffusion process or those of a corresponding discrete time approximation of a diffusion. The main statistical task in such a diffusion approach is the identification of the underlying drift and diffusion coefficient functions.

To handle this task in standard situations, a well-developed statistical theory is now available. Parameter estimation methods for discretely observed stationary diffusion processes have been derived, for instance, by Bibby and Sørensen (1995), Hansen and Scheinkman (1995), Aït-Sahalia (1996), Gallant and Tauchen (1996) and Kessler and Soerensen (1999). Hansen *et al.* (1998) consider a non-parametric method based on the spectral

decomposition of the conditional expectation operator to identify the drift and diffusion coefficients. Jiang and Knoght (1997) propose non-parametric estimators for the drift and diffusion coefficient that are based on the approximation of the local time process and the estimation of the marginal density of the diffusion. A crucial assumption for most of the developed estimation techniques is that a stationary diffusion is directly observed. Unfortunately, in reality this assumption is often not fulfilled. For instance, Genon-Catalot *et al.* (2000) consider a stochastic volatility model where the observed stock price depends on a hidden volatility process, which is itself a stationary diffusion.

We study here a case where a non-stationary diffusion process, an index, is observed. To be able to apply methods that rely on ergodicity we express the observed process as the product of an ergodic process and a smooth function of time. This smooth function is interpreted as average growth of the index. Due to the unknown impact of the average growth on the observed data, the ergodic part of our model is not directly observable. The proposed methodology combines recently developed non-parametric and parametric methods in order to estimate and probe the drift and diffusion coefficients of the ergodic process.

One important issue in an empirical analysis of financial data is the identification and estimation of the drift function. It is well known that the estimates for the drift are very unreliable whenever the observed data are modeled as a general diffusion process; see, for example, Merton (1980). Since we express the observed index as a product of a smooth function of time and an ergodic process we add a structure to the model that allows us to identify the drift of the index itself; compare the discussion of (6). As financial practitioners often assume that the development of a stock market index or other financial time series is related to the long-term-development of macroeconomic data, our assumption seems reasonable. To our knowledge this approach to model financial data is new.

To illustrate our methodology we concentrate here on the empirical analysis of a particular stock market index, the S&P 500. The statistical analysis of stock prices, exchange rates, etc. is similar but not in the focus of this paper. We concentrate here on the case where an index is modeled by a scalar diffusion process.

The framework of Platen (2002) fully characterizes a financial market by the specification of the different denominations of the so-called best benchmark portfolio. The stock market index can be interpreted as domestic currency denomination of the best benchmark portfolio. As a consequence, exchange prices are ratios of corresponding denominations of the best benchmark portfolio. Furthermore, this portfolio represents the growth optimal portfolio (Karatzas and Shreve 1998). A well-diversified market index, like the S&P 500, comes close to the growth optimal portfolio (Platen 2004). For this reason, the inference for the index is also the first step in the statistical analysis of an exchange rate or stock price.

We assume that an appropriately normalized index process $X = \{X(t), t \geq 0\}$ can be interpreted as an ergodic process, as described by Platen (2004). Based on this assumption we focus on the inference of this normalized process X instead of the index $S = \{S(t), t \geq 0\}$ itself. This allows us to direct our attention towards the identification of the drift and diffusion coefficient functions of an ergodic diffusion.

In figure 1 we plot the S&P 500 index S with daily data from January 1950 to March 2006 (14 151 observations) together with an average index $\bar{S} = \{\bar{S}(t), t \geq 0\}$. Such an average index $\bar{S}$ can be obtained in different ways. For instance, it could be exogenously given by a function of economic and financial quantities, i.e. inflation rate, growth rate of the domestic product, interest rate, etc. It could also be derived by a kernel smoothing procedure, with an appropriate bandwidth or filter length h. This is the choice that we will study in this paper.

We construct the normalized index X by dividing the original index S by the above-described average index $\bar{S}$, that is

$$X(t) = \frac{S(t)}{\bar{S}(t)}, \tag{1}$$

for $t \geq 0$. The resulting normalized index X, derived via a kernel smoother, is shown in figure 1. Its path resembles that of a stationary diffusion process.

To make (1) precise we assume for the value of the index $S(t)$ at time t a representation of the form

$$S(t) = S(0)Z(t) \exp\left\{ \int_0^t \eta(s)\, \mathrm{d}s \right\}, \tag{2}$$

for $t \geq 0$. Here, $\eta(t)$ is interpreted as the deterministic, time-dependent growth rate of the index at time t. Furthermore, $Z(t)$ denotes the value at time t of a positive
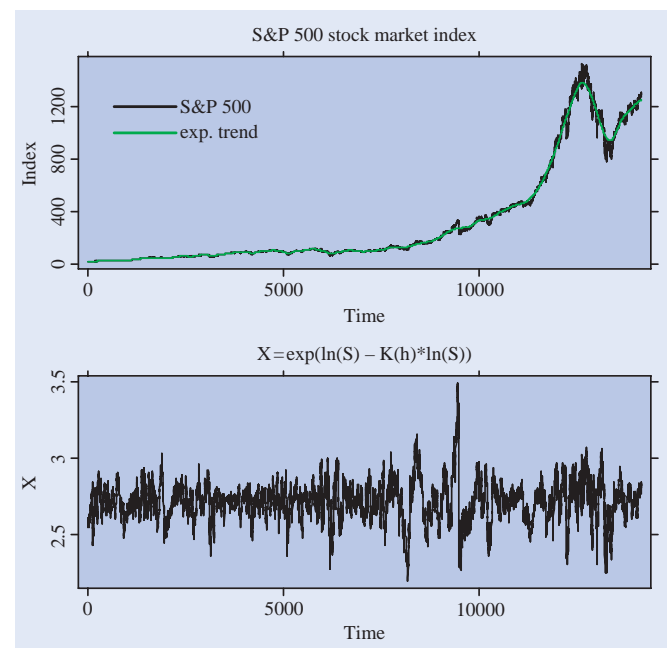

Figure 1. S&P 500 index S, average index $\bar{S}$ and normalized index X.

ergodic diffusion process Z, which solves the Itô stochastic differential equation (SDE)

$$dZ(t) = m\{Z(t)\}\,dt + \sigma\{Z(t)\}\,dW(t), \qquad (3)$$

for $t \geq 0$. Here, $W = \{W(t), t \geq 0\}$ denotes a standard Wiener process and $m\{.\}$ and $\sigma\{.\}$ are the drift and diffusion coefficient functions. Due to the factor $S(0)$ we assume that Z is stable about 1, which models a mean reverting behaviour. On the other hand, $Z(t)$ has to be positive for all $t \geq 0$.

To make our parametric model specific we may choose for Z a square root process that is positive and stable about an equilibrium reference level, where $m\{z\} = \psi - \varphi z$ and $\sigma\{z\} = \nu\sqrt{z}$. The square root process is also known as the Cox–Ingersoll–Ross (CIR) process (Cox *et al.* 1985). The functional form (2) that models the index is a special case of the minimal market model (MMM) proposed by Platen (2002).

Another parametric model arises if we choose

$$Z(t) = \exp\{U(t)\}, \qquad (4)$$

with an Ornstein–Uhlenbeck process $U = \{U(t), t \geq 0\}$, with

$$dU(t) = -\beta U(t)\,dt + \gamma\,dW(t). \qquad (5)$$

This leads us to the exponential of an Ornstein–Uhlenbeck process as index model. Such a model has been used, for instance, by Föllmer and Schweizer (1993), Platen and Rebolledo (1996) and Fleming and Sheu (1999).

To compute the average index $\bar{S}$ in (1), we apply a kernel smoother to the logarithm of S and then calculate $\bar{S}$ as the exponential of this smoothed process. This removes the average deterministic growth in (2). For the analysis of the resulting normalized index we have to take into account that the residuals $\ln S - \ln \bar{S}$ are corrupted by the smoother. This will be shown in detail later on. This means that the normalized index X is not a diffusion and, in particular, it does not equal the diffusion Z. For this reason we cannot directly apply estimation methods for discretely observed diffusions. In fact, the influence of the kernel smoother on X strongly depends on the chosen bandwidth. We will therefore analyse this influence in detail and provide a bandwidth selection method in that situation.

From the statistical point of view, we are faced with a non-parametric regression model with error terms that are not independent and identically distributed but are the discrete observations of a diffusion process. The analysis of these error terms and the clarification of their relationship to Z is a main task of this paper and a new contribution to the literature concerning non-parametric regression.

We remark that the index process S is itself a diffusion. When Z is specified according to (4) with an Ornstein–Uhlenbeck process U as given in (5), Itô's formula yields

the representation

$$\frac{dS(t)}{S(t)} = \left[\eta(t) - \beta\left(\ln S(t) - \int_0^t \eta(s)\,ds\right) + \frac{1}{2}\gamma^2\right]dt + \gamma\,dW(t), \qquad (6)$$

for $t \geq 0$. The parameters β, γ and η cannot be easily separated in this representation. For this reason, we develop a statistical methodology for models that are based on the representation (2).

In section 2.1 we introduce the parametric model for Z. The kernel smoothing and the computation of X is described in section 2.2. The choice of the kernel and bandwidth and its influence on the average index is discussed in section 2.3 together with the corresponding parameter estimation methods. Section 2.4 introduces discrete time approximations of diffusion processes and section 2.5 describes the non-parametric estimation of drift and diffusion coefficient functions. In section 2.6, a parametric model is tested versus a purely non-parametric alternative. This test is carried out by the bootstrap technique described in appendix A. In sections 3 and 4 we apply the introduced methodology to S&P 500 data as well as to simulated data.

We remark that the proposed methodology applies directly to situations where normalized data can be modeled by an ergodic diffusion process. Emphasis is given here to the case of an Ornstein–Uhlenbeck process, and results on the influence of the kernel smoother are included for this case.

2. Statistical methodology for a normalized diffusion

2.1. Parametric models

As discussed in the Introduction, one can, in principle, use various parametric ergodic diffusion models. Let us mention two examples. Both of them have mean reverting drift coefficients. In the case where the squared diffusion coefficient has the form

$$\sigma^2(z) = \nu^2 z, \quad z > 0, \qquad (7)$$

with a positive constant ν, we obtain in (2) a square root process Z. Here we assume that Z satisfies the SDE

$$dZ(t) = \{\psi - \varphi Z(t)\}\,dt + \nu\sqrt{Z(t)}\,dW(t), \qquad (8)$$

for $t \geq 0$ and with $\psi > \nu^2/2$, $\varphi > 0$. Note that a stationary and ergodic solution of (8) exists with the expected value $\mu_\infty = E[Z] = \psi/\varphi > 0$. Since the ratio $Z(t)/\mu_\infty$ is again a square root process and any constant term can be absorbed by $S(0)$ in (2), one can for simplicity assume that $\mu_\infty = E[Z] = 1$. This choice leads us to the SDE

$$dZ(t) = \varphi\{1 - Z(t)\}\,dt + \nu\sqrt{Z(t)}\,dW(t), \qquad (9)$$

for $t \geq 0$.

We obtain a second example for an ergodic diffusion by defining Z as in (4), where U denotes the well-known

Ornstein–Uhlenbeck process satisfying (5). Since U fluctuates about its reference level 0 and is ergodic, Z as given in (4) is an ergodic, positive diffusion process fluctuating about 1. In the following the statistical methodology is developed for the exponential of an Ornstein–Uhlenbeck process.

2.2. Kernel smoothing

Denote by K_h a smoother with a kernel K and a bandwidth h. The smoothing of any process is denoted by a convolution operator ($*$). As mentioned before, the normalized index $X(t)$ in (1) can be defined by the exponential of the difference of

$$L(t) = \ln\{Z(t)\}, \qquad (10)$$

and its smoother $(K_h * L)(t)$, that is

$$
\begin{aligned}
X(t) &= \exp\{\ln S(t) - (K_h * \ln S)(t)\} \\
&= \exp\{L(t) - (K_h * L)(t)\}.
\end{aligned}
\qquad (11)
$$

Equation (11) holds if we neglect the difference between the accumulated deterministic growth rate $\int_0^t \eta(s)\,ds$ in (2) and its smoother, which means

$$\int_0^t \eta(s)\,ds - \left(K_h * \int_0^{\cdot} \eta(s)\,ds\right)(t) \approx 0.$$

Here we arrive at a delicate point of our study. If we want to remove efficiently the deterministic growth rate in (2), then the value h should be chosen relatively small. Indeed, smaller values for h reduce the bias. On the other hand, the smaller the value of h chosen, the more X is corrupted by $K_h * L$ in (11).

Therefore, the choice of the bandwidth h is an important issue. We choose to investigate the properties of the proposed procedure for a fixed bandwidth and for $h \to \infty$ rather than $h \to 0$ since, in that case, $L(t) - (K_h * L)(t)$ converges to 0 and the estimated average growth index $\bar{S}(t)$ converges to the index $S(t)$ itself.

The smoother $K_h * L$ is differentiable for differentiable kernels K and thus of bounded variation. Due to the smoothing procedure $K_h * L$ involves also future information about L. Thus X is not a diffusion process. For this reason, we cannot treat $\ln X(t)$ in (11) as the logarithm of a square root process or as an Ornstein–Uhlenbeck process. A more detailed analysis of X has therefore to be performed. This is the objective of the next section. However, note that, in the case when $\bar{S}$ is obtained exogenously and not by a smoothing procedure, X might still be a diffusion.

2.3. Estimation of parameters and bandwidth selection

In this section we assume that the only observations available are those of

$$\ln X(t) = L(t) - (K_h * L)(t) \qquad (12)$$

Table 1. Estimated values for γ and the estimated variance of X for different bandwidths h.

h	100	200	300	400
Var (X)	0.00094158	0.0018558	0.0028745	0.0040556
$\hat{\gamma}$	0.0088847	0.0089054	0.0089125	0.0089171

in (11) and that L is the Ornstein–Uhlenbeck process U given in (5). The estimation problem that we now consider is that for the parameters β and γ in (5). In principle, the value of γ can be restored from the quadratic variation of either $Z(t)$ or $L(t)$. For differentiable kernels K_h in (11), the process $(K_h * L)(t)$ is also differentiable for every fixed bandwidth $h > 0$. For this reason it holds that

$$\lim_{\Delta t \to 0} \sum_{i=1}^{n} \{\ln X(i\Delta t) - \ln X(i\Delta t - \Delta t)\}^2 \overset{L^2}{=} \int_0^T d\langle L\rangle_t, \quad (13)$$

for $n = T/\Delta t$ if h is fixed or $h \to \infty$.

Here, $d\langle L\rangle_t$ denotes the differential of the quadratic variation of the process L at time t. Empirical results confirm that the quadratic variation is not sensitive to the choice of h; see table 1 for more details. The following formula provides a stable estimate of γ^2 in the form

$$
\begin{aligned}
\hat{\gamma}^2 &= T^{-1} \sum_{i=1}^{n} \{\ln X(i\Delta t) - \ln X(i\Delta t - \Delta t)\}^2 \\
&\approx T^{-1} \int_0^T d\langle L\rangle_t.
\end{aligned}
\qquad (14)
$$

To estimate the speed of adjustment parameter β in (5) we could use the well-known form of the stationary variance of the Ornstein–Uhlenbeck process L. Along with (14) this would result in a first estimator of β with

$$\hat{\beta}_1 = \hat{\gamma}^2/(2\mathrm{Var}[L]). \qquad (15)$$

Unfortunately, the substitution of $\mathrm{Var}[L]$ by $\mathrm{Var}[\ln X]$ makes $\hat{\beta}_1$ strongly dependent on h. Indeed, the variance

$$\mathrm{Var}[\ln X] = \mathrm{Var}[L - K_h * L] \qquad (16)$$

increases as h grows, and only for very large values of h can we expect that $\mathrm{Var}[\ln X] \approx \mathrm{Var}[L]$.

It is not just the variance of the random process $\ln X$ that changes with h. Its autocorrelation function also depends on the bandwidth h. The correlation between the values of the process $\ln X$, distant by a constant time length $\tau > 0$, diminishes as h decreases. For this reason we propose a selection method for h based on the simultaneous estimation of β from the variance and from the autocorrelation function of the process $\ln X$. The idea is simple, if for each value of h there are two different estimates of the same parameter β, then the best choice of h is considered to be the one that brings these estimates as close as possible to each other.

The autocorrelation function $\rho^{(L)}(\tau)$ of the Ornstein–Uhlenbeck process L equals

$$\rho^{(L)}(\tau) = e^{-\beta\tau}, \tag{17}$$

for $\tau > 0$. Thus, β represents the absolute value of the slope of this function at zero. Hence, another estimate of β from the observations of L would be

$$\hat{\beta}_2 = \left| \frac{\partial^+}{\partial \tau} \rho^{(L)}(\tau) \right|_{\tau=0}, \tag{18}$$

where

$$\frac{\partial^+}{\partial \tau} \rho(\tau) = \lim_{s \to 0,\, s>0} \frac{\rho(\tau+s) - \rho(\tau)}{s},$$

for $\tau \geq 0$ denotes the right-hand derivative of ρ with respect to τ.

Unfortunately, the estimator in (18) is not feasible since L is not observed. In appendix A.1 we show for the process $\ln X = L - K_h * L$ that its stationary variance is asymptotically

$$\mathrm{Var}\,[\ln X] = \frac{\gamma^2}{2\beta}\left(1 - \frac{c_K}{\beta h} + O(h^{-2})\right) \quad \text{as } h \to \infty, \tag{19}$$

where the constant c_K depends on the kernel K. Furthermore, we prove in appendix A.1 for the autocorrelation function $\rho_h^{(\ln X)}(\tau)$ of $\ln X$ that

$$\rho_h^{(\ln X)}(\tau) = [\ln X(\tau);\, \ln X(0)]$$

$$= \frac{e^{-\beta\tau} - (c_K/\beta h) + O(\tau h^{-2})}{1 - (c_K/\beta h) + O(h^{-2})}, \quad \tau \geq 0, \tag{20}$$

as $h \to \infty$ with the same constant c_K as in (19). In appendix A.1 this constant is calculated for the rectangle and the Epanechnikov kernels.

It follows from equation (19) that the first-order approximation of the stationary variance of $\ln X$ is

$$\mathrm{Var}\,[\ln X] \approx \left\{1 - \frac{c_K}{\beta h}\right\} \mathrm{Var}\,[L]. \tag{21}$$

By (20), the slope of the autocorrelation function $\rho_h^{(\ln X)}(\tau)$ of $\ln X$ at zero is asymptotically

$$\left| \frac{\partial^+}{\partial \tau} \rho_h^{(\ln X)}(\tau) \right|_{\tau=0} \approx \frac{\beta}{1 - (c_K/\beta h)}, \tag{22}$$

as $h \to \infty$. Thus this slope is steeper than that of ρ^L at $\tau = 0$.

The immediate consequence of (19) and (20) is that formulae (15) and (18) for $\hat{\beta}_1$ and $\hat{\beta}_2$, respectively, have to be modified if the process $\ln X$ rather than L is observed. In appendix A.1 we show that the correct modification is provided by the expressions

$$\hat{\beta}_1(h) = \frac{\hat{\gamma}^2}{2\mathrm{Var}\,[\ln X]} - \frac{c_K}{h} \tag{23}$$

and

$$\hat{\beta}_2(h) = \left| \frac{\partial^+}{\partial \tau} \rho_h^{(\ln X)}(\tau) \right|_{\tau=0} - \frac{c_K}{h}, \tag{24}$$

respectively. Finally, our method for the selection of h is based on balancing the two estimates $\hat{\beta}_1(h)$ and $\hat{\beta}_2(h)$. Therefore, the optimal value h_{opt} of h is

$$h_{\mathrm{opt}} = \arg\min_{h>0} |\hat{\beta}_1(h) - \hat{\beta}_2(h)|. \tag{25}$$

After h is chosen, we need to restore the process L, which is needed in the remaining non-parametric and parametric analysis. From (11), proceeding formally, one arrives at the following iterative formula:

$$L = \ln X + K_h * L = \ln X + K_h * (\ln X + K_h * L)$$
$$= \cdots$$
$$= \ln X + K_h * \ln X$$
$$\quad + K_h * K_h * \ln X + \cdots. \tag{31}$$

The justification for the restoration formula (26) comes from the fact that if one neglects the boundary effects, the smoothing operator K_h is a contracting operator in L_2, as shown in appendix A.1. In the practical application of (26), we rely on the fact that the smoother of the original process L is close to the smoother of $L - K_h * L$. In practice, only one or two convolutions are meaningful. After the restoration process is completed, the parameter β can be estimated directly from L by (15).

We were able to establish in this paper the above correction terms for the Ornstein–Uhlenbeck process. One could, in principle, estimate parameters also under the assumption that X itself is a square root process or another ergodic diffusion. However, if the average index $\bar{S}$ is calculated via a smoothing procedure, a similar bandwidth selection method has to be developed. At this stage this is left for future research.

2.4. Discrete-time approximation of a diffusion

We interpret the process $Z(t)$, which appears in (2), as a positive ergodic scalar diffusion process that is the solution of the SDE (3). The drift $m: [0, \infty) \mapsto (-\infty, \infty)$ and the diffusion coefficient $\sigma: [0, \infty) \mapsto [0, \infty)$ in (3) are assumed to be sufficiently regular, such that a unique solution of (3) exists.

For the existence of an ergodic solution of (3) the drift and the diffusion coefficient must satisfy some ergodicity conditions; see, for instance, Bibby and Sørensen (1995). The most important condition is that the stationary Kolmogorov forward equation,

$$\frac{1}{2}\frac{\partial}{\partial z}\{\sigma^2\{z\}p_0(z)\} - m\{z\}p_0(z) = 0,$$

must have a solution $p_0(z)$ which is then up to a constant the stationary probability density.

In particular, the above conditions hold for the exponential of an Ornstein–Uhlenbeck process with a log-normal stationary distribution and for the square root process which has a Gamma distribution as stationary distribution.

Let us assume that the diffusion process Z is observed at discrete times $t_i = i\Delta$, $i = 1, 2, \ldots$, with a time step size $\Delta > 0$. Here we assume that Δ is small or, more precisely, will tend to zero asymptotically. Under rather weak assumptions (Kloeden and Platen 1999) on the functions m and σ^2, it can be shown that the Euler approximation

$$Z^\Delta(t) = Z^\Delta(0) + \int_0^t m\{Z^\Delta(t_{i_s})\}\,\mathrm{d}s + \int_0^t \sigma\{Z^\Delta(t_{i_s})\}\,\mathrm{d}W(s),$$
$$(27)$$

with $Z^\Delta(0) = Z(0)$ and $t_{i_s} = \max\{t_i,\ t_i \leq s\}$, converges in a mean square sense to Z as $\Delta \to 0$, i.e.

$$\lim_{\Delta \to 0} E\left[\sup_{0 \leq t \leq T} |Z^\Delta(t) - Z(t)|^2\right] = 0, \quad T > 0. \qquad (28)$$

From now on, we assume that a discrete time approximation Z^Δ exists in the form of (27), and that the convergence property (28) holds. For the purposes of this paper, Δ will always be considered to be small enough so that Z can be substituted by Z^Δ without any major error in the interpretation of the observed data. The concrete choice of Δ does not matter since all the relevant information about the model is contained in the drift m and diffusion coefficient σ. As the estimates of moments of higher-order terms in the case of the daily observed S&P 500 show, the step size Δ corresponding to the given daily observations is small enough so that the difference between $Z^\Delta(t_i)$ and $Z(t_i)$ is indeed negligible. Thus we interpret the increments of the observed data as those of the Euler approximation (27), that is

$$Z^\Delta(t_{i+1}) - Z^\Delta(t_i) = m\{Z^\Delta(t_i)\}\Delta$$
$$+ \sigma\{Z^\Delta(t_i)\}\{W(t_{i+1}) - W(t_i)\}, \qquad (29)$$

for $i = 0, 1, \ldots$. Note that the observations $(Z^\Delta(t_i))$, $i = 0, 1, \ldots$, form a state-dependent time series.

2.5. Non-parametric estimation

For two consecutive observations $Z^\Delta(t_i)$ and $Z^\Delta(t_{i+1})$ of the above time series we define the increment

$$Y_i = Z^\Delta(t_{i+1}) - Z^\Delta(t_i) = m\{Z^\Delta(t_i)\}\Delta + \sigma\{Z^\Delta(t_i)\}\sqrt{\Delta}\varepsilon_i,$$
$$(30)$$

with independent standard Gaussian random variables

$$\varepsilon_i = \frac{W(t_{i+1}) - W(t_i)}{\sqrt{\Delta}} \sim \mathcal{N}(0, 1).$$

It is now possible to use the increments Y_i to estimate the functions $m\{.\}$ and $\sigma^2\{.\}$. In particular, we apply the local linear non-parametric method developed by Fan

and Gijbels (1986) and Härdle and Tsybakov (1997). To justify the use of this method, the time series $Z^\Delta(t_i)$ has to be ergodic and meet some technical conditions (Härdle and Tsybakov 1997). The square root process and the Ornstein–Uhlenbeck process are examples for which the discrete time approximation (30) satisfies these conditions.

In this non-parametric framework we obtain the drift function estimator $\hat{m}_{h_1}(z)$ from

$$\hat{m}_{h_1}(z) = \frac{1}{\Delta}\hat{b}_0(z), \qquad (31)$$

with

$$\hat{b}(z) = \begin{pmatrix} \hat{b}_0(z) \\ \hat{b}_1(z) \end{pmatrix}$$
$$= \arg\min_{b_0, b_1}\left(\sum_{i=1}^n \{Y_i - b_0 - b_1(Z^\Delta(t_i) - z)\}^2 K_{h_1}(z - Z^\Delta(t_i))\right),$$

and $K_{h_1}(x) = (1/h_1)K(x/h_1)$. The bandwidth $h_1 > 0$ is chosen with respect to Silverman's rule of thumb (Härdle 1990), and we choose the kernel K to be the density of the standard normal distribution.

We apply a two-step estimation to obtain an estimate of the squared diffusion coefficient $\sigma^2(z)$. In the first step we compute the values $\hat{m}_{h_1}\{Z^\Delta(t_i)\}$ using the drift function estimator (31). In the second step we use the squared diffusion function estimator

$$\hat{\sigma}_{h_1}^2(z) = \frac{1}{\Delta}\hat{\delta}_0(z), \qquad (32)$$

with

$$\hat{\delta}(z) = \begin{pmatrix} \hat{\delta}_0(z) \\ \hat{\delta}_1(z) \end{pmatrix}$$
$$= \arg\min\left(\sum_{i=1}^n \{(Y_i - \Delta\hat{m}_{h_1}\{Z^\Delta(t_i)\})^2$$
$$- \delta_0 - \delta_1(Z^\Delta(t_i) - z)\}^2 K_{h_1}(z - Z^\Delta(t_i))\right),$$

and bandwidth $h_1 > 0$. For details and the asymptotic properties of these estimators and the construction of corresponding confidence bands, we refer to appendices A.2 and A.3.

The application of the above methodology to the data Z^Δ yields corresponding non-parametric estimates with an estimated squared diffusion coefficient function as well as a drift coefficient function along with corresponding confidence bands as we demonstrate below.

2.6. Testing the parametric model

We now construct tests to compare the non-parametric estimates of the drift function $m\{.\}$ and the squared diffusion coefficient $\sigma^2\{.\}$ obtained in section 2.5 with their parametric counterparts. Although the proposed method

is applicable to a variety of diffusion models we will concentrate on the exponential of an Ornstein–Uhlenbeck process. We therefore consider the process $Z(t) = \exp(U(t))$ where U satisfies (5).

We apply Itô's formula to $Z(t)$ to derive the null hypotheses in this case and obtain

$$dZ(t) = d(\exp\{U(t)\})$$
$$= Z(t)\left\{-\beta \ln Z(t) + \frac{1}{2}\gamma^2\right\} dt + \gamma Z(t) dW(t), \quad (33)$$

for $t \geq 0$. The null hypotheses of the tests are therefore

$$H_0(m): m\{z\} = z\left\{-\beta \ln z + \frac{1}{2}\gamma^2\right\}$$

and

$$H_0(\sigma^2): \sigma^2\{z\} = \gamma^2 z^2,$$

while the alternatives are non-parametric.

We use a bootstrap method to construct confidence bands. The idea is to bootstrap the original discrete time series and estimate each time the drift and squared diffusion coefficients non-parametrically as described in section 2.5. Using these estimates we can construct confidence bands for the two functions.

The bootstrap method leads to better coverage probabilities than, for instance, a Gaussian approximation. Neumann and Kriss (1998) have shown for a time series similar to (30) that the coverage probability is of order $O(n^{-q})$ for some $q > 0$, where n is the number of observations. In contrast, a Gaussian approximation (Hall 1985) leads to a coverage probability of order $O(1/\ln(n))$.

In appendix A.3 we describe in detail how the confidence bands $KB(m)$ for the non-parametric drift estimate and $KB(\sigma^2)$ for the diffusion coefficient estimate are constructed. Franke *et al.* (2002) show that

$$P\{m\{z\} \in KB(m)\} \to 1 - \alpha$$

and

$$P\{\sigma^2\{z\} \in KB(\sigma^2)\} \to 1 - \alpha,$$

if the number of observations of the time series Z^Δ tends to infinity. The constant α is the confidence level.

The null hypothesis $H_0(m)$ is rejected if the parametric estimate of $m(z)$ is outside the confidence band $KB(m)$ for some values of z. Similarly, we reject $H_0(\sigma^2)$ if the parametric estimate of $\sigma^2(z)$ takes values outside $KB(\sigma^2)$ for some values of z.

3. Empirical analysis of the S&P 500

We apply the methods introduced in section 2 to daily observations of the S&P 500 index from 3 January 1950 to 30 March 2006 (14 151 observations).

For the kernel smoothing of S we choose the Epanechnikov kernel. The constant c_K that appears in

the correction terms in (23) and (24) is known for this particular kernel (see appendix A.1).

As already mentioned in section 2.3 the estimates for the parameter γ calculated from formula (14) are small relative to 1 and do not change significantly with h. Table 1 shows the estimated values for different values of h. The variance of the process X is also shown in the table. The small variance and the fact that X is stable about 1 justifies concentrating on the case of a geometric Ornstein–Uhlenbeck process defined by (4) and (5).

The next step in our analysis is the choice of h. Due to the long range of observations we apply a flexible bandwidth to the data. This flexible bandwidth was calculated by splitting the data into overlapping subintervals and calculating an optimal fixed bandwidth for every subinterval.

More precisely, we use the observations $S(t_i)$, $i = k - l, \ldots, k + l$, to calculate the optimal bandwidth at time t_k. We have chosen $l = 2500$ and $k = 2500, 3000, 3500, \ldots, 13\,500$. This means that every subinterval contains 5000 observations (corresponding to approximately 20 years) and that an optimal bandwidth is calculated every two years.

Within each subinterval the bandwidth is chosen to be optimal according to equation (25). To obtain a continuous optimal bandwidth function $h_{\text{opt}}(t)$ we interpolate the resulting values. $h_{\text{opt}}(0)$ and $h_{\text{opt}}(T)$ are chosen to be the mean of all the values of $h_{\text{opt}}(t_k)$. The function $t \mapsto h_{\text{opt}}(t)$ is shown in figure 2. The final values for β are $\hat{\beta}_1(h_{\text{opt}}) = 0.018981$ and $\hat{\beta}_2(h_{\text{opt}}) = 0.018148$ and the ratio is $\hat{\beta}_1(h_{\text{opt}})/\hat{\beta}_2(h_{\text{opt}}) = 1.0459$. For fixed bandwidths in the range of h_{opt} this ratio is given in table 2. The values for $h = 300$ and $h = 400$ are also very close to 1 and therefore we could use these fixed bandwidths instead of the flexible bandwidth. However, we have chosen a
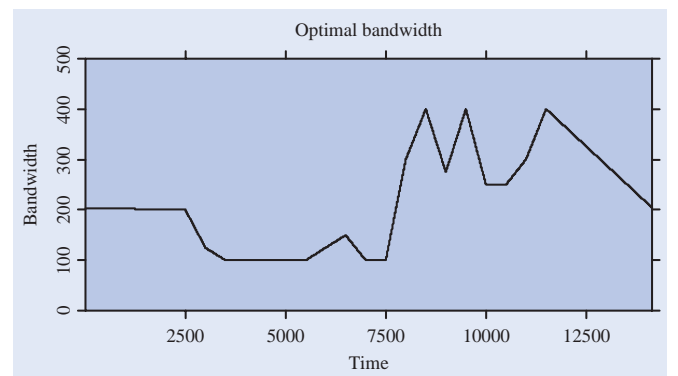
Figure 2. The optimal flexible bandwidth $h_{\text{opt}}(t)$.

Table 2. Estimated values for β and γ for different fixed bandwidths h together with the variance of X.

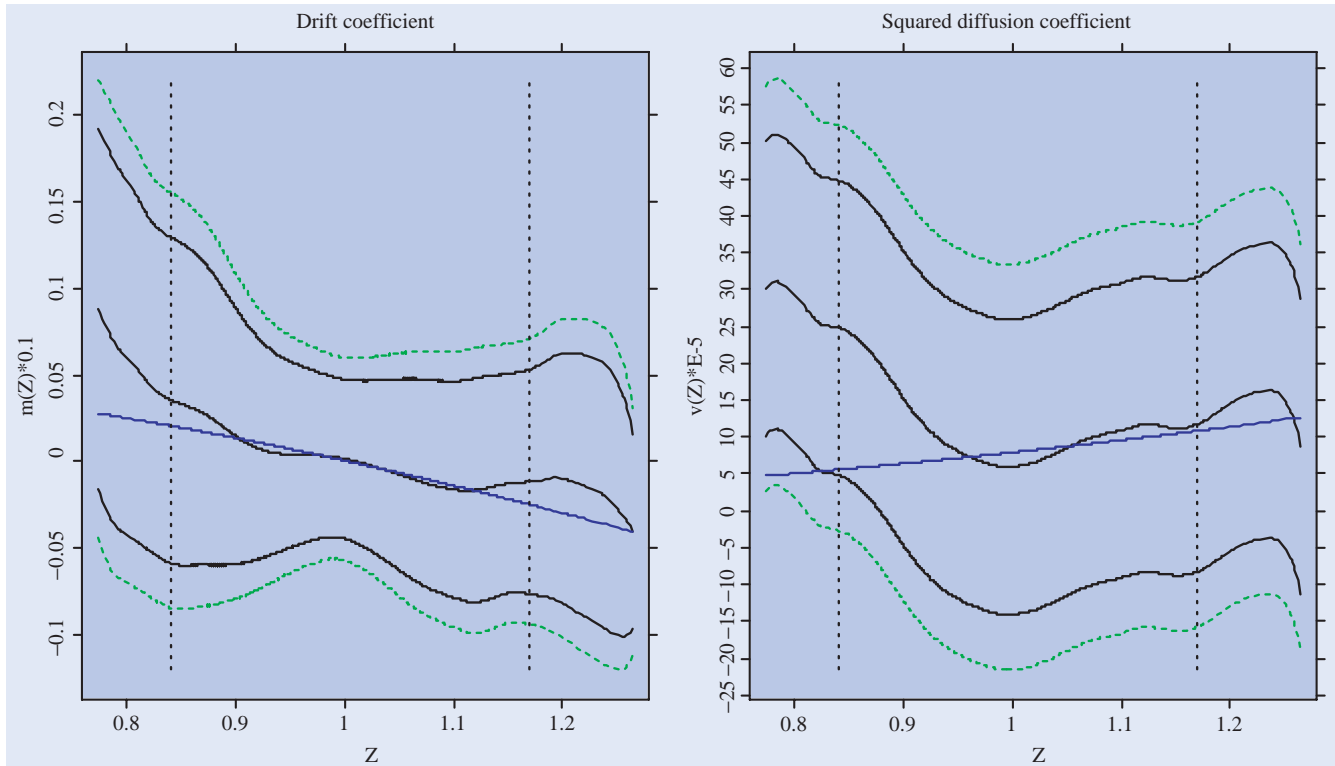| h	100	200	300	400
$\hat{\beta}_1(h)$	0.041199	0.020797	0.013338	0.0094164
$\hat{\beta}_2(h)$	0.033442	0.019147	0.013109	0.009632
$\hat{\beta}_1(h)/\hat{\beta}_2(h)$	1.2319	1.0861	1.0175	0.97762

Figure 3. Non-parametric and parametric estimates of the drift $m\{.\}$ and squared diffusion coefficient $\sigma^2\{.\}$ with 90% and 99% confidence bands.

flexible bandwidth due to the long range of observations, 55 years. The estimated value for γ is $\hat{\gamma}(h_{\text{opt}}) = 0.0089005$.

Now we are in the position to restore the path of the process L and to estimate its parameters. To restore L we apply equation (26). We use a first-order approximation and obtain $L = \ln X + K_h * \ln X$. We obtain the following estimates from the restored path:

$$\hat{\beta}_1 = 0.013881, \qquad \hat{\beta}_2 = 0.013718,$$

$$\hat{\beta}_1/\hat{\beta}_2 = 1.0119, \qquad \hat{\gamma} = 0.008905.$$

To finish the empirical analysis we apply the test procedure described in section 2.6. Figure 3 shows the non-parametric estimates of the drift and squared diffusion coefficient of the restored process Z together with the 90% and 99% confidence bands. The almost straight lines show the parametric estimates with respect to the estimated values of the restored process Z where we have chosen $\beta = 0.0138$, which is between the two estimates $\hat{\beta}_1$ and $\hat{\beta}_2$. The vertical lines enclose the interval where 99% of the observed data reside.

Both parametric functions are clearly inside the 99% confidence bands. Therefore, we do not reject the null hypothesis of the geometric Ornstein–Uhlenbeck process at the 99% significance level.

The parametric estimate for the squared diffusion coefficient takes values outside the 90% confidence band. However, these values are taken for observations of Z below the 0.5%-quantile. We can therefore conclude that the null hypothesis of the geometric

Ornstein–Uhlenbeck process does not hold for these very low values, but is fulfilled if we ignore the lower 0.5% of the observations.

4. Simulation study

We now perform a simulation study by applying the estimation methods introduced in section 2.3 to simulated trajectories of the Ornstein–Uhlenbeck process U. The drift and diffusion parameters β and γ in (5) are estimated directly from the observations of U as well as from the residual of a kernel smoothing procedure.

It is well known that the transition probability of an Ornstein–Uhlenbeck process is normal with conditional mean

$$E[U_{t+\Delta} \mid U_t = u] = ue^{-\beta\Delta},$$

and conditional variance

$$\text{Var}(U_{t+\Delta} \mid U_t = u) = \frac{\gamma^2}{-2\beta}(e^{-2\beta\Delta} - 1).$$

Using this Gaussian transition probability we simulate 100 paths of the process U with time step size $\Delta = 1$. The true parameters are set to $\beta = 0.01$ and $\gamma = 0.01$, which correspond approximately to the empirical estimates for the S&P 500 index in section 3.

For the analysis of the directly observed process U we apply three estimators for the speed of adjustment

630 parameter β. Besides $\hat{\beta}_1$ and $\hat{\beta}_2$ introduced in (15) and (18), we also use the estimator

$$\hat{\beta}_3 = -\frac{1}{\Delta} \ln \frac{\sum_{i=1}^{n} U_{i-1} U_i}{\sum_{i=1}^{n} U_{i-1}^2}, \qquad (34)$$

which is based on martingale estimating functions and was proposed by Bibby and Sørensen (1995). It is easy to see that $\hat{\beta}_3$ is related to the autocorrelation function of U. For details concerning this estimator and the theory of martingale estimating functions, we refer to Bibby and Sørensen (1995) and references therein. The diffusion coefficient is estimated via the slope of the quadratic-variation, similarly as in (14).

635 The first row of table 3 shows the average of the corresponding estimated values. In the second row the variance of the estimates are shown. We emphasize that the results are based on a directly observed simulated diffusion.

645 Furthermore, we simulate the logarithm of the index in (2) $\ln S$ as the sum of a linear function and U. In a second step we calculate $\ln X$ as in (11) with the Epanechnikov kernel (see appendix A.1). We then estimate from the simulated data the parameters β and γ of 650 U by the methods in (23), (24) and (14). This gives us

an idea concerning the fixed sample behaviour of these estimation methods when the residuals of a kernel smoothing procedure are observed instead of those of an Ornstein–Uhlenbeck process itself.

655 The estimated values calculated from the simulated trajectories of $\ln X$ are shown in the third and fourth rows of table 3. The results clearly demonstrate that the correction terms in (23) and (24) are necessary to obtain reasonable estimated values. In the situation considered here, 660 the correction terms equal each other and have approximately the value $c_k/h \approx 0.0061$ (see appendix A.1). Since the correction terms for $\hat{\beta}_3$ are not considered, we do not report them in table 3.

The table also shows the mean and the variance of the ratio $\hat{\beta}_1/\hat{\beta}_2$ used to select the bandwidth h (see equa- 665 tion (25)). The mean of the selected bandwidth h, which brings this ratio as close as possible to one, is given in the last column.

The second part of the simulation study treats the bootstrap procedure. We apply the bootstrap methodology as 670 introduced in appendix A.3 to a simulated path of an Ornstein–Uhlenbeck process U following the dynamics in (5) with parameters $\beta = 0.01$, $\gamma = 0.01$ and $\Delta = 1$. The values of the parameters are reasonable with respect to the empirical results for the S&P 500. The number of 675 observations is 5000 and the number of bootstrapped series for the confidence bands is 160. The two plots in figure 4 show the non-parametric estimators for the drift and squared diffusion coefficient together with their 90% confidence bands constructed by the bootstrap procedure. 680 The plots also show the true parametric functions for the drift and diffusion coefficient. The dotted vertical lines are

Table 3. Estimated parameters.

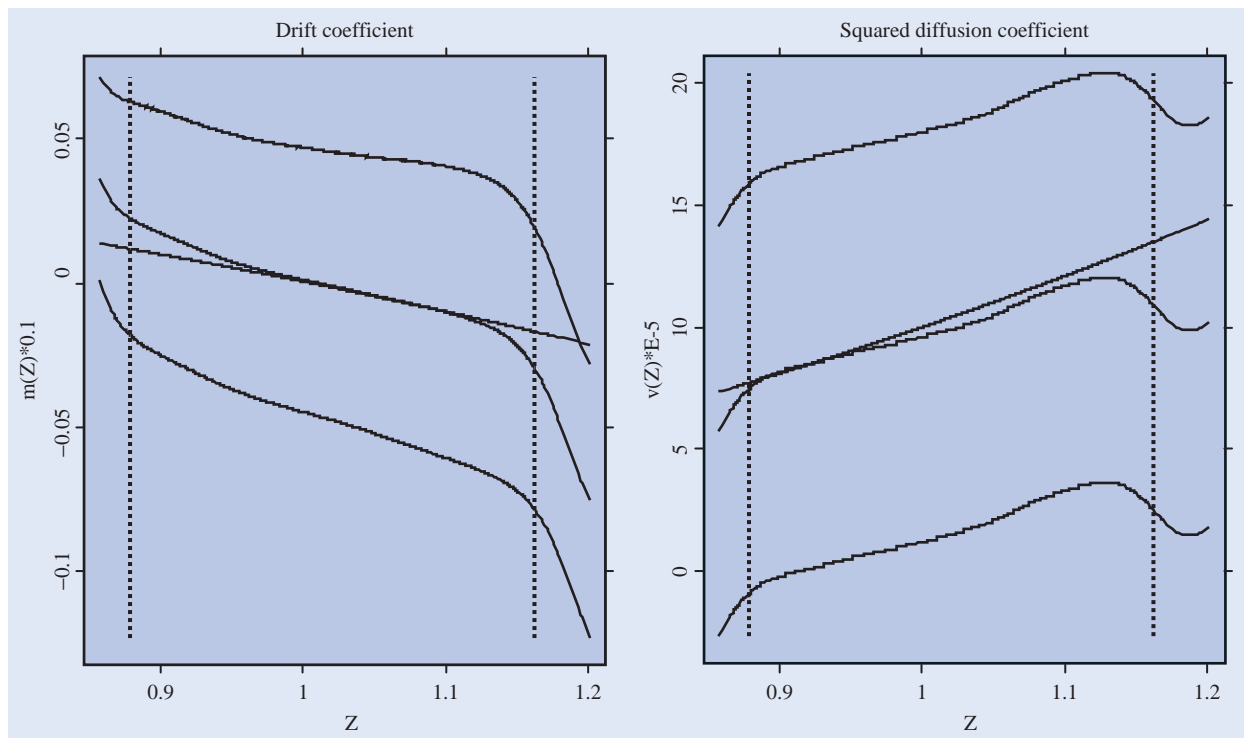| $\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_3$	γ	$\hat{\beta}_1/\hat{\beta}_2$	Opt h
0.01070	0.01018	0.01028	0.00996		
4.589e−06	3.695e−06	4.541e−06	1.059e−08		
0.00964	0.00967		0.00996	0.99435	295.0
7.161e−06	6.258e−06		1.067e−08	0.00248	

Figure 4. Non-parametric estimates for the drift and squared diffusion coefficients of a simulated geometric Ornstein–Uhlenbeck process, confidence bands and true functions.

the empirical 0.005 and 0.995 quantiles of the stationary distribution of $\exp(U)$. If we only consider the range between these quantities, i.e. the range where 99% of the data reside, then both of the parametric functions remain inside the confidence bands. This means that the null hypotheses $H_0(m)$ and $H_0(\sigma^2)$ as in section 2.6 cannot be rejected for data in this range.

5. Conclusion

In this paper we have modeled a stock market index as the product of an ergodic diffusion and a deterministic growth process. In the first part we proposed a methodology that allows us to separate the estimation of the average growth of the index and that of the parameters of the ergodic diffusion. The general methodology was carried out for the particular case of an Ornstein–Uhlenbeck process. A challenge for future research is to establish similar estimation methods for parameters of a square root process and other ergodic diffusions. The derivation of the corresponding correction terms will be the key problem in such an approach.

In the second part of the paper we developed a semi-parametric testing procedure for the drift and diffusion coefficient functions of an ergodic diffusion. The test is based on a comparison of the parametric forms of these functions with their non-parametric estimators. Finally, an empirical analysis of the S&P 500 stock market index and a simulation study completed the paper.

Acknowledgements

Appendix A

A.1. Parameter estimation

Let $U(t)$ be the Ornstein–Uhlenbeck process satisfying (5). Introduce the autocovariance of the process $U - K_h * U$ as

$$\text{Cov}(\tau) = \text{Cov}[(U - K_h * U)(\tau); \quad (U - K_h * U)(0)],$$
$$0 < \tau \ll h.$$

Let

$$\rho_h(\tau) = \frac{\text{Cov}(\tau)}{\text{Cov}(0)}$$

be the autocorrelation function of $U - K_h * U$.

Proposition A.1

(i) *If* $K(u)$ *is the rectangle kernel, i.e.* $K(u) = (1/2)\,\mathbb{I}(|u| \leq 1)$, *then as* $h \to \infty$ *we have*

$$\text{Cov}(\tau) = \frac{\gamma^2}{2\beta}\left(e^{-\beta\tau} - \frac{1}{\beta h} - \frac{1}{2\beta h}(\tau/h) + O(h^{-2})\right). \quad \text{(A1)}$$

(ii) *If* $K(u) = (3/4)(1 - u^2)\,\mathbb{I}(|u| \leq 1)$ *is the Epanechnikov kernel, then*

$$\text{Cov}(\tau) = \frac{\gamma^2}{2\beta}\left(e^{-\beta\tau} - \frac{1.8}{\beta h} - \frac{1.5}{\beta h}(\tau/h)^2 + O(h^{-2})\right). \quad \text{(A2)}$$

Proposition A.2 *Under the assumptions of proposition A.1, up to the terms of magnitude* $O(h^{-2})$, *the following equation holds for* β:

$$\left.\frac{\partial^+}{\partial\tau}\rho_h(\tau)\right|_{\tau=0} = -\frac{\beta}{1 - c_K/(\beta h)}.$$

The solution of this equation is approximately

$$\hat{\beta}_2 = \left|\frac{\partial^+}{\partial\tau}\rho_h(\tau)\right|\bigg|_{\tau=0} - \frac{c_K}{h} + O(h^{-2}),$$

where $c_K = 1$ *if* K *is the rectangle kernel, and* $c_K = 1.8$ *if* K *is the Epanechnikov kernel.*

Proof (proof of proposition A.1)

(i) Integrate both sides of (5) with the rectangle kernel. The integration results in

$$-\frac{1}{2\beta h}\{U(t+h) - U(t-h)\}$$
$$= (K_h * U)(t) - \frac{\gamma}{2\beta h}\{W(t+h) - W(t-h)\},$$

or

$$(K_h * U)(t) = \frac{\gamma}{2\beta h}(W(t+h) - W(t-h))$$
$$- \frac{1}{2\beta h}(U(t+h) - U(t-h)) \quad \text{(A3)}$$

The autocovariance function equals

$$\text{Cov}(\tau) = E[U(\tau)U(0)] - E[U(\tau)(K_h * U)(0)]$$
$$- E[U(0)(K_h * U)(\tau)]$$
$$+ E[(K_h * U)(\tau)(K_h * U)(0)]. \quad \text{(A4)}$$

The first term in the latter formula is $E[U(\tau)U(0)] = \{\gamma^2/(2\beta)\}e^{-\beta\tau}$. With the help of (A3) and the explicit representation

$$U(t) = \int_{-\infty}^{t} \exp\{-\beta(t-s)\}\gamma\,dW(s),$$

one finds by direct calculation that each of the negative terms on the right-hand side of (A4) contributes

$$-\frac{\gamma^2}{(2\beta^2 h)} + O(h^{-2}),$$

while the covariance of $K_h * U$ adds up to

$$\frac{\gamma^2(2h-\tau)}{(2\beta h)^2} + O(h^{-2}).$$

Combining these results, we arrive at (A1).

(ii) Integrating (5), we find as in (A3) that

$$(K_h * U)(t) = -\frac{3}{2\beta h^3} \int_{t-h}^{t+h} (s-t)U(s)\,\mathrm{d}s + \frac{\gamma}{\beta}(K_h * W)(t). \quad (A5)$$

It is straightforward to verify that the variance of the first term on the right-hand side of (A5) has magnitude $O(h^{-2})$ for h large. This term is negligible compared with the second one. As in part (i), we obtain

$$\begin{aligned} \mathrm{Cov}(\tau) &= \frac{\gamma^2}{2\beta}e^{-\beta\tau} - 2\left(\frac{3\gamma^2}{4\beta^2 h} + O(h^{-2})\right) \\ &= \quad + \frac{\gamma^2}{\beta^2}E[(K_h * W)(\tau)(K_h * W)(0)] \\ &= \frac{\gamma^2}{2\beta}e^{-\beta\tau} - \frac{3\gamma^2}{2\beta^2 h} + \frac{\gamma^2}{\beta^2}\left(\frac{3}{5} - \frac{3}{4}\left(\frac{\tau}{h}\right)^2\right) + O(h^{-2}). \end{aligned}$$
$$(A6)$$

This proves (A2). $\qquad\square$

Proof: (proof of proposition A.2) The slope of the autocorrelation at zero follows from (A1) and (A2). Let

$$A = \left\|\frac{\partial^+}{\partial\tau}\rho_h(\tau)\right\|_{\tau=0},$$

then for $\hat{\beta}_2$ the quadratic equation $\hat{\beta}_2^2 h - \hat{\beta}_2 hA + c_K A = 0$ holds with root $\hat{\beta}_2 = A - c_K/h + O(h^{-2})$. $\qquad\square$

Proposition A.3 *Let C_0 be a space of a continuous function with finite support. Define K as the rectangle or Epanechnikov kernel. Then the operator K_h is a contracting operator on the space $L_2 \bigcap C_5$ with the L_2-norm.*

Proof: The Fourier transformation for the rectangle kernel is $\tilde{K}(z) = (\sin z)/z$, and for the Epanechnikov kernel $\tilde{K}(z) = 3(\sin z - z\cos z)/z^3$ with unique maximum value 1 at $z = 0$. Thus, for the nth iterative convolution, $\|K^n\|_2 \to 0$ as $n \to \infty$. This confirms the result. $\qquad\square$

A.2. Asymptotic properties of LLP

The asymptotic properties of local polynomial estimates were studied by Fan and Gijbels (1996) and Härdle et al. (1999). Under some smoothness conditions with bandwidth $h_1 = k_0/n^{1/5}$ for a constant $k_0 > 0$, the results applied to our case provide the following formulae for the asymptotic normality:

$$n^{2/5}\{\hat{m}_{h_1}(z) - m(z)\} \xrightarrow{\mathcal{D}} N\left(\frac{k_0^2}{2}\mu_2(K)\Delta m''(z), \frac{\|K\|_2^2 \Delta\sigma^2(z)}{k_0 p_0(z)}\right), \quad (A7)$$

$$n^{2/5}\{\hat{\sigma}_{h_1}^2(z) - \sigma^2(z)\} \xrightarrow{\mathcal{D}} N\left(\frac{k_0^2}{2}\mu_2(K)(\Delta(\sigma^2(z))'' + 2(\Delta m'(z))^2), \frac{2\Delta^2\sigma^4(z)\|K\|_2^2}{k_0 p_0(z)}\right). \quad (A8)$$

Here, $\mu_2(K)$ is the second moment of the kernel K and $\|K\|_2$ is its L_2 norm. Furthermore, $p_0(z)$ denotes the stationary density of Z as given in section 2.4.

A.3. The bootstrap procedure

The confidence bands in section 2.6 for the non-parametric estimator are constructed by the following bootstrap method.

1. Choose a bandwidth g that is larger than the optimal h_1 in order to have oversmoothing. Then estimate non-parametrically $m(.)$ and $\sigma^2(.)$ and obtain the residual estimated errors:

$$\hat{\varepsilon}_i = \frac{Y_i - \Delta\hat{m}_g\{Z^\Delta(t_i)\}}{\sqrt{\Delta}\hat{\sigma}_g\{Z^\Delta(t_i)\}}.$$

Since we make the assumption that ε_i has zero mean, we subtract the sample mean of $\hat{\varepsilon}_i$.

2. Replicate N times the series of $(\hat{\varepsilon}_i)$ with wild bootstrap, obtaining $(\varepsilon_i^{*,n})$ for $n = 1, \ldots, N$ and build N new bootstrapped series $(Z^{*,n})$:

$$\begin{aligned} Z_1^{*,n} &= Z^\Delta(t_1), \\ Z_{i+1}^{*,n} - Z_i^{*,n} &= \Delta\hat{m}_g(Z_i^{*,n}) + \sqrt{\Delta}\hat{\sigma}_g(Z_i^{*,n})\varepsilon_i^{*,n}. \end{aligned}$$

Estimate again $m(z)$ and $\sigma^2(z)$ for each of the N bootstrapped series with bandwidth h_1.

3. Build the statistics

$$T_m^* = \sup_z \frac{|\hat{m}_{h_1}^{*,n}(z) - \hat{m}_{h_1}(z)|}{\hat{\sigma}_{h_1}^{*,n}(z)}$$

and

$$T_\sigma^* = \sup_z |(\hat{\sigma}^2)_{h_1}^{*,n}(z) - \hat{\sigma}_{h_1}^2(z)|.$$

4. Form the $(1 - \alpha)$ confidence bands KB

$$KB(m(.)) = [\hat{m}_{h_1}(z) - \hat{\sigma}_{h_1}(z)t_{m,1-\alpha}, \hat{m}_{h_1}(z) + \hat{\sigma}_{h_1}(z)t_{m,1-\alpha}]$$

and

$$KB(\sigma^2(.)) = [\hat{\sigma}^2_{h_1}(z) - t_{\sigma,1-\alpha}, \hat{\sigma}^2_{h_1}(z) + t_{\sigma,1-\alpha}],$$

where $t_{m,1-\alpha}$ and $t_{\sigma,1-\alpha}$ denote the empirical $1 - \alpha$-quantile of T^*_m and T^*_σ, respectively.

References

Aït-Sahalia, Y., Nonparametric pricing of interest rate derivative securities. *Econometrica*, 1996, **64**, 527–560.

Bibby, B.M. and Sørensen, M., Martingale estimation functions for discretely observed diffusion processes. *Bernoulli*, 1995, **1**, 17–40.

Cox, J., Ingersoll, J. and Ross, S., A theory of the term structure of interest rates. *Econometrica*, 1985, **2**, 385–407.

Duan, J.-C., Augmented GARCH(p, q) process and its diffusion limit. *J. Economet.*, 1997, **79**, 97–127.

Engle, R.F. and Bollerslev, T., Modelling the persistence of conditional variances. *Economet. Rev.*, 1986, **5**, 1–87.

Fan, J. and Gijbels, I., *Local Polynomial Modelling and its Applications—Theory and Methodologies*, 1996 (Chapman and Hall: New York).

Fleming, W.H. and Sheu, S.-J., Optimal long term growth rate of expected utility of wealth. *Ann. Appl. Probab.*, 1999, **9**, 871–903.

Föllmer, H. and Schweizer, M., A microeconomic approach to diffusion models for stock prices. *Math. Finan.*, 1993, **3**, 1–23.

Franke, J., Kreiss, J.-P., Mammen, E. and Neumann, M., Properties of the nonparametric autoregressive bootstrap. *Time Ser. Anal.*, 2002, **33**, 555–586.

Frey, R., Derivative asset analysis in models with level-dependent and stochastic volatility. *CWI Q*, 1997, **10**, 1–34.

Gallant, A. and Tauchen, G., Which moments to match? *Economet. Theory*, 1996, **12**, ██–██.

Genon-Catalot, V., Jeantheau, T. and Larédo, C., Stochastic volatility models as hidden Markov models and statistical applications. *Bernoulli*, 2000, **6**, ██–██.

Ghysels, E., Harvey, A. and Renault, E., *Statistical Methods in Finance.*, vol. 14 of Handbook of Statistics, pp. 119–191, 1996 (North-Holland: Amsterdam).

Hall, P., Resampling a coverage pattern. *Stoch. Proc. Appl.*, 1985, **20**, 231–246.

Hansen, L.P. and Scheinkman, J.A., Back to the future: generating moment implications for continuous-time Markov processes. *Econometrica*, 1995, **63**, 767–804.

Hansen, L.P., Scheinkman, J.A. and Touzi, N., Spectral methods for identifying scalar diffusions. *J. Economet.*, 1998, **86**, 1–32.

Härdle, W., *Applied Nonparametric Regression*, Econometric Society Monographs number 19, 1990 (Cambridge University Press: Cambridge).

Härdle, W., Klinke, S. and Müller, M., *XploRe—The Statistical Computing Environment*, 1999 (Springer: New York).

Härdle, W. and Tsybakov, A., Local polynomial estimators of the volatility function in nonparametric autoregression. *J. Economet.*, 1997, **81**, 223–242.

Jiang, G. and Knight, J., A nonparametric approach to the estimation of diffusion processes, with an application to a short-term interest rate model. *Economet. Theory*, 1997, **13**, 615–645.

Karatzas, I. and Shreve, S.E., *Methods of Mathematical Finance*, vol. 39 of Applications of Mathematics, Stochastic Modelling and Applied Probability, 1998 (Springer: New York).

Kessler, M. and Soerensen, M., Estimating equations based on eigenfunctions for a discretely observed diffusion process. *Bernoulli*, 1999, **5**, 299–314.

Kloeden, P.E. and Platen, E., *Numerical Solution of Stochastic Differential Equations*, vol. 23 of Applications of Mathematics, 1999 (Springer: Berlin).

Merton, R.C., On estimating the expected return on the market. *J. Finan. Econ.*, 1980, **8**, 141–183.

Nelson, D.B., ARCH models as diffusion approximations. *J. Econ.*, 1990, **45**, 7–38.

Neumann, M. and Kreiss, J.-P., Regression-type inference in nonparametric autoregression. *Ann. Statist.*, 1998, **26**, 1570–1613.

Platen, E., Arbitrage in continuous complete markets. *Adv. appl. Probab.*, 2002, **34**, 540–558.

Platen, E., Modeling the volatility and expected value of a diversified world index. *Int. J. Theor. Appl. Finan.*, 2004, **7**, 511–529.

Platen, E. and Rebolledo, R., Principles for modelling financial markets. *J. Appl. Probab.*, 1996, **33**, 601–613.

On the Utility of E-Learning in Statistics

Wolfgang Härdle[1,2], Sigbert Klinke[1] and Uwe Ziegenhagen[1,2]

[1]*Institute of Statistics and Econometrics (ISE), Berlin, Germany*
[2]*Center for Applied Statistics and Economics, School of Business and Economics,*
Humboldt-Universität zu, Berlin, Germany
E-mails: haerdle@wiwi.hu-berlin.de, sigbert@wiwi.hu-berlin.de, ziegenhagen@wiwi.hu-berlin.de

Summary

Students of introductory courses consider statistics as particularly difficult, as the understanding of the underlying concepts may require more time and energy than for other disciplines. For decades statisticians have tried to enhance understanding with the help of technical solutions such as animation, video or interactive tools. However, it is not clear if the added value generated by these e-learning tools justifies the work invested. In this paper the experience with various e-learning solutions in terms of utility and the impact on teaching is discussed.

Key words: E-learning; spreadsheets; PISA study; online learning tools.

1 Introduction

Before technical details are discussed it is necessary to clarify the vocabulary. E-learning describes all kinds of learning, where digital media are used for the presentation and transmission of learning materials and/or to support inter-human communication. In general e-learning tools can be grouped into:

- video conferencing and tele-teaching systems;
- simulations with or without user interaction;
- learning (content) management tools;
- content catalogues and podcasting;
- web and computerbased training applications.

In the context of this paper the term e-learning refers mainly to the last category of this list, to web and computerbased training applications.

The projects proposed by Nolan & Temple Lang (2007) and Darius *et al.* (2007) also fall into this category as they offer two different technical solutions to improve the understanding of statistics by students. Nolan & Temple Lang (2007) propose a system for interactive and dynamic documents which allows students to follow the complex decision making processes of a statistical analyst. As an example they take the classification of emails into spam and ham using different dialogues for, e.g. data loading, derivation of classification variables and statistical classification & prediction.

Darius *et al.* (2007) present a set of Java applets to assist student understanding of experimental design. For two common situations in experimental design, the optimization of an industrial

process and a greenhouse experiment to find the optimal dose of fertilizer, they have developed interactive applets to provide their students with a test-bed for experimental design studies.

In the last few decades a plethora of interactive tools have been developed that aimed at helping students to receive a better understanding of statistics and to develop 'statistical thinking'. Some prominent commercial examples are Fathom (Yates *et al.*, 2007) or Visual Statistics (McGraw-Hill, 2001). Common internet search engines deliver more than one million hits for the search term 'statistics applet'.

The availability of web based content and pro-active statistics teaching tools however prompts, at least, the following two important questions:

(1) Have students improved their statistics knowledge by using e-learning tools?
(2) What impact do e-learning tools have on our teaching?

As we will show in the next subsection, the results attained by students of statistics have not improved significantly, despite massive increases in the use of e-learning tools (see Section 1.2), so we are entitled to question the effectiveness of e-learning in statistics.

1.1 Students Versus Statistics

The reason that students encounter difficulties in statistics are complex and can, to a large part, be assigned to the students themselves and their lecturers. The PISA study (Baumert *et al.*, 2002) revealed that the mathematical ability of German school students is significantly lower than the average of all OECD countries; in mathematics and natural sciences Germany was ranked 20-th out of 32, in reading it was ranked even lower at place 21.

These results, which reflect the high school performance, consequently also affect the ability of university students. An even more dramatic scenario has recently been discovered at the Johannes Kepler University Linz, Austria: Some of the original PISA exercises were given to students on the statistics course to check their basic mathematical knowledge, with questions such as 'How much is 30% of 70% percent?'. As Duller (2004, 2007a, b) reported, the results were highly unsatisfactory and have even deteriorated since. The authors have the same impression. In fact exam grades have not improved although the standard of exams has been lowered over the last 10 years.

The comparatively low competence in mathematics may also provide hints for the evaluation of the statistics courses at the School of Business and Economics at Humboldt-Universität, Berlin. For more than 10 years now students have evaluated courses and lecturers. Figure 1 shows (at the bottom) the difference between the average answer to the questions 'General impression of course/course expectation fulfilled' for the compulsory course in statistics (comparable to STAT101 courses in the United States), the voluntary exercises (middle) and all other voluntary courses from the authors institute for the Winter Term 1999/2000 until the Summer Term 2006. As the average of all courses from the faculty was centred to 0, the graphs indicate that even a specifically good teacher (evaluated above average from the advanced students) does not fulfil all expectations for the introductory courses.

The overall average of the statistics courses is around −0.12, verbal comments from the students name three problems in the compulsory basic courses as well as in the voluntary advanced courses in statistics, these are:

(1) too formal, too much mathematics;
(2) too fast, too much content in the curriculum;
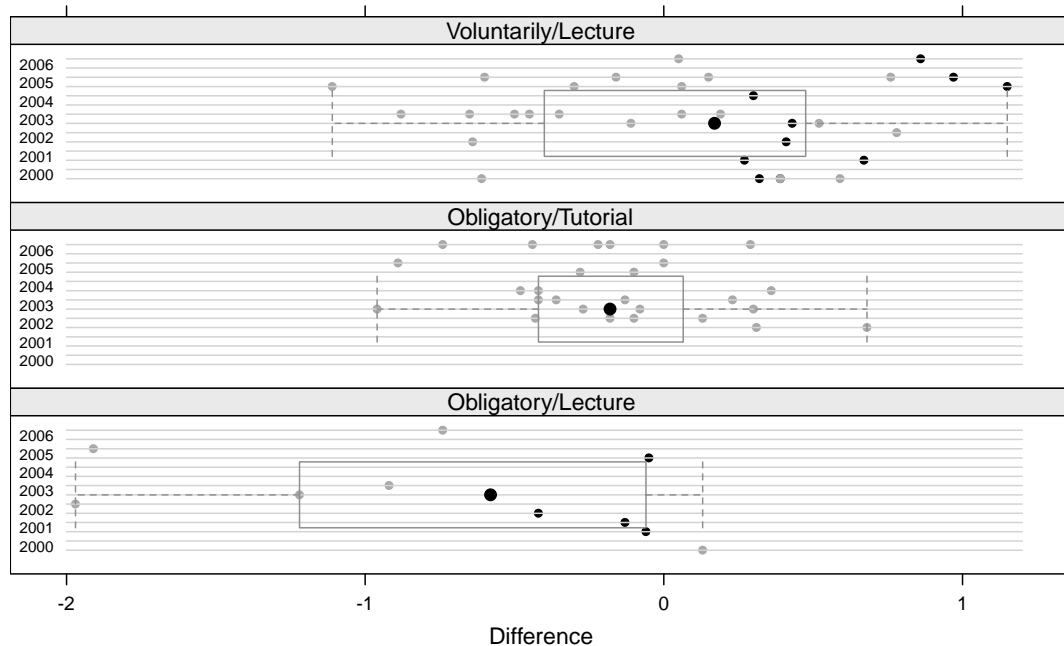(3) not enough examples.

Figure 1. *Deviation of the mean response for statistics courses from the mean for all courses for the question 'General impression of the course/course expectation fulfilled' (1 = fully to 5 = not at all) from winter term 1999/2000 until summer term 2006 by course attributes (obligatory/voluntarily visit and lecture/tutorial). The small black dots represent one particular teacher.*

The first point of criticism adresses the difficult problem of balancing between mathematical formalism and looseness. In the authors' opinion there will always be students who complain about the level of formalism. A significant part of statistics requires solid and basic mathematics, therefore there must be a minimum level of formalism and mathematical rigour. We agree with Nolan & Temple Lang (2007), that visualization, simulation and computing provide alternative lenses through which to view statistical problems; however we think, that these lenses may provide an excellent view when combined with lense of mathematics.

The topics covered in the curriculum are associated with the way courses are taught. In most universities there is little need to write on a board as there is a prepared set of digital slides either in Powerpoint or Adobe PDF format. The slide set for the two-term introductory course in statistics at the School of Business and Economics contains, for example, nearly 700 slides covering ten topics. The advantage of slides is obvious. One can easily add links to data sources or include low level dynamic content.

Experience and feedback from students together with the aim of proliferating existing content drove the development of electronic teaching material. By distributing course material in a convenient (electronic) form students have a chance to concentrate on lectures rather than copying the content of slides. By providing examples that would be too time-consuming or complex for the lecture we have tried to support the students' understanding. And finally various web based learning environments have been developed that will be described in the next section.

The last point of criticism by students is the lack of examples, a point we partly agree with as it can be very difficult to develop excellent examples especially for introductory courses. Our experience is that if one global example is used throughout the lecture to explain various concepts some students complain about the 'single boring example', if on the other hand, we use

several examples the students may fail to see the associations between the different topics in the course.

1.2 Online Learning Tools

Like Nolan & Temple Lang (2007) and Darius *et al.* (2007) we have endeavoured to supplement our class notes and enrich the educational experience of our students by providing a wide variety of e-learning tools. We will first list these tools and then describe each in more detail:

- MM*Stat, an interactive browser-based tool for statistics;
- Electronic books, resulting from a collaboration with Springer;
- Questions & Answers, a web based exam preparation tool;
- Microsoft Excel, to provide spread-sheet based examples;
- TeachWiki, a collaborative environment on a Wiki basis.

MM*Stat (Müller *et al.*, 2000) is aimed at supplementing the traditional lectures by creating a framework to repeat and practise the content of lectures. Besides versions in various latin-based languages such as English, German, French and Spanish, there is an Arabic version as well; for details see Ahmad *et al.* (2006, 2007).

Using a HTML-based tab structure, which was generated from LaTeX using the MD*Book system (Klinke & Lehmann, 2003) with 'lecture' tabs providing definitions, formulas and graphics and various example tabs, the student can follow the lecture step by step. Since the different tabs are connected with links the student may jump between the different lectures and the glossary, which provides brief definitions for each term used. The system is available online under http://www.quantlet.com/mdstat/mmstat.html.

To provide interactivity two different approaches have been used. There are not only multiple choice questions at the end of each chapter which are evaluated via Javascript, but also interactive examples based on a client/server solution featuring the XploRe Quantlet Server (XQS) and the XploRe Quantlet Client (Cizek *et al.*, 2005) which visualize, for example, parameter changes for statistical distributions and tests. The system is available for various languages, Figure 2 depicts a lecture tab, with links to the examples in the bottom right-hand corner.

In order to generate additional value for printed media, in terms of allowing the reader to interactively modify and run the examples from a textbook, the XQS technology used for MM*Stat has been applied to books, which have been designed for second year and above students. Each book (see http://www.xplore-stat.de/ebooks/ebooks.html), its HTML version and the corresponding set of slides contain references to web pages on a webserver. These HTML pages show the XploRe source code and allow the examples to be run in the Java Quantlet Client applet; experienced users may also modify the source code. Figure 3 shows an example taken from the book *Applied Multivariate Analysis* Härdle & Simar (2007).

The Questions & Answers (Q&A) system, Klinke (2004), was designed specifically for students who want to practice multiple choice questions that had played a significant role in the undergraduate exams. Despite the different books the students can use for exam preparation certain, very common errors can be seen. Within the Q&A project, available online under http://mars.wiwi.hu-berlin.de/qa the errors are specified on the basis of rated exams, for each exercise the different results are computed if a student makes a particular error. When the student then clicks one of the wrong answers the lecturer can see exactly where the source of the error is and explain the topic again in the next lecture. Figure 4 shows an example questions from Q&A: Given some information about the situation the student has to click one or more boxes; these

Figure 2. *Screenshot of MM*Stat, English version.*

are evaluated immediately. However, we have found that instead of thinking carefully about a solution, a significant proportion of students just click until they receive a correct answer.

The TeachWiki project, available under http://teachwiki.wiwi.hu-berlin.de/index.php/Main_Page, is another web based project but with a strong focus on collaborative work. Usually seminar theses remain in the basement, with TeachWiki we allow the students' theses to be published to a broader audience. The system is based on the MediaWiki technology of Wikipedia and enhanced to integrate LaTeX syntax for formulas and a plug-in to integrate R code. The idea is also to provide the students with a platform where they can easily work jointly on a seminar thesis.

In addition to these tools, standard spreadsheet-software (Microsoft Excel) is used in our courses. Excel is available on nearly every computer and the majority of students knows at least the basic commands to work with the software. The built-in functionality in term of graphical capabilities and interactive interface components allow easy and quick creation of, e.g. visualizations of the Central Limit Theorem or the parameter changes of distribution functions.

Figure 5 displays a screenshot of a spreadsheet to visualize eight different probability density functions where the parameters can be modified via sliders, Figure 6 shows a visualization of the basic algorithm for kernel density estimation.

1.3 Lessons Learned

As mentioned above the style and methods of teaching have changed from a board centred approach to presentations of electronic contents (Powerpoint, PDF). This solution may have
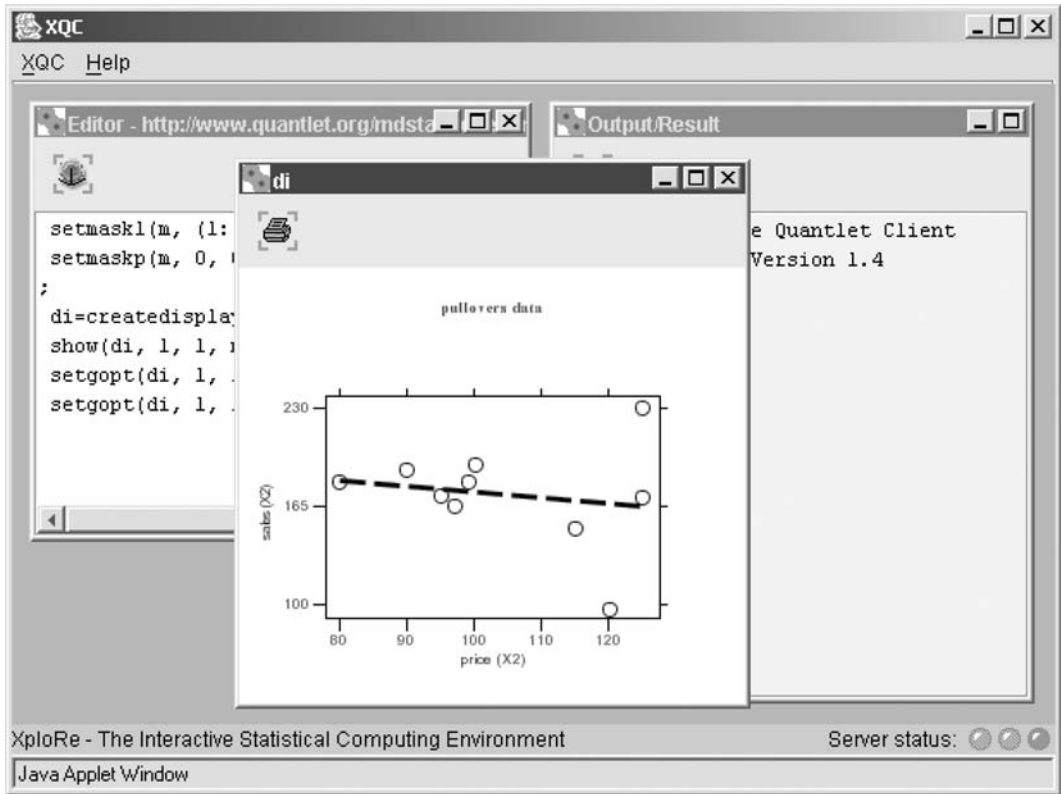
Figure 3. *Screenshot of an interactive example from the 'Applied Multivariate Statistics' book.*

disadvantages not only for the student (the curriculum is much larger), but also for the lecturer. In the same way as that it is generally not possible to do parallel digital-slide presentations and use a board—due to for example light conditions or the available projection space, the lecturer cannot leave the media and, for example, quickly add a drawing.

From over ten years of experience in the collaborative development and use of e-Learning tools we have learned the following:

Technical problems: E-Learning projects usually have to commit themselves to a certain platform. For many—especially smaller—opensource projects there is no guarantee that the software platform will be supported in a few years or that future versions will be compatible. Even commercial software products frequently disappear from the market, so the decision to use a certain platform must be well thought through. As the funding for a bigger project is often limited to a certain time window, one has to ask about the project maintenance after the funding has ceased.

Administrative overhead: The truth is simple: Excellent projects require excellent ideas and funding and usually provide more work than a single person can handle. The administrative work (security, access control) which has nothing to do with the original goal of teaching statistics often exceeds the usefulness of a supporting tool and rather becomes a burden.

Specialization: E-learning tools are often developed by small teams for a special concept only. Since the documentation and publication of the source codes usually receives a
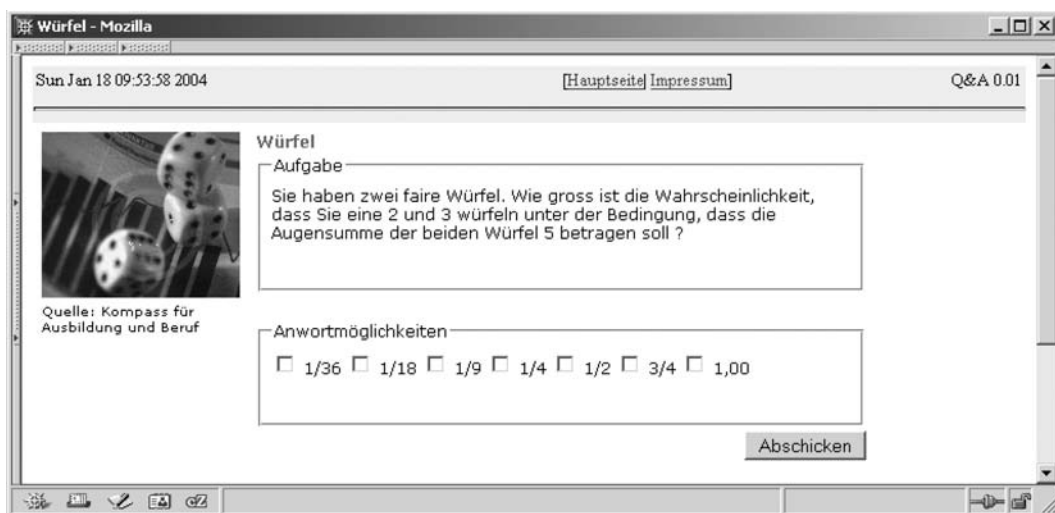
Figure 4. *An exercise from Q&A: 'Given you throw two fair dices what is the probability of throwing a 2 and a 3 given the sum of both throws was 5?'.*

comparatively low weight, it may be easier for other projects to implement their own solutions. A Google search for 'Java Applet Histogram' returns nearly 200,000 hits. Even if this number overestimates the true number of histogram applets there are still thousands of applets with more or less the same functionality.

Availability: The setup and handling of e-learning tools can be difficult especially if they require the installation of server software such as Apache, PHP or Java. Microsoft Excel on the contrary can be considered as one of the software packages that every student has access to; and although it may not provide enough precision or speed to do deep data analysis, its features are sufficient for interactive visualization of, e.g. parameter changes of distributions. This platform suits students that belong to the increasing number of 'button-pushers'. Only a very small fraction prefers to use command-driven interfaces. In our experience 'just clicking a button' also lowers the students willingness to question the appropriateness of an algorithm or method in certain situations. In this sense we must complement the recommendation of Nolan & Temple Lang (2007) of 'simplifying how we do what we do, not what we do' by 'and explain why we do'.

Evaluation: Like Mills (2002), we are not aware of any resonable study in the area of statistics where it is proven that the use of interactive tools improves student understanding. Projects which claim to be successful often required great effort in terms of human and technical resources. Traditional teaching methods with the same resources might be equally or even more successful. Most projects, for example, report the number of accesses on project web pages, etc., but no hard proof in terms of increased student grades in an experiment is given. We agree with Mills (2002) that an evaluation may be difficult, as ethical issues may arise in the selection, which students are taught according to the different methods.

Darius *et al.* (2007) present a questionnaire they had asked their students to complete but cannot now generate a numerical analysis since the experimental design software has changed over time.

In the introductory statistics course at the School of Business and Economics students receive a MM*Stat CD-ROM and, although the students' comments are very positive, an evaluation
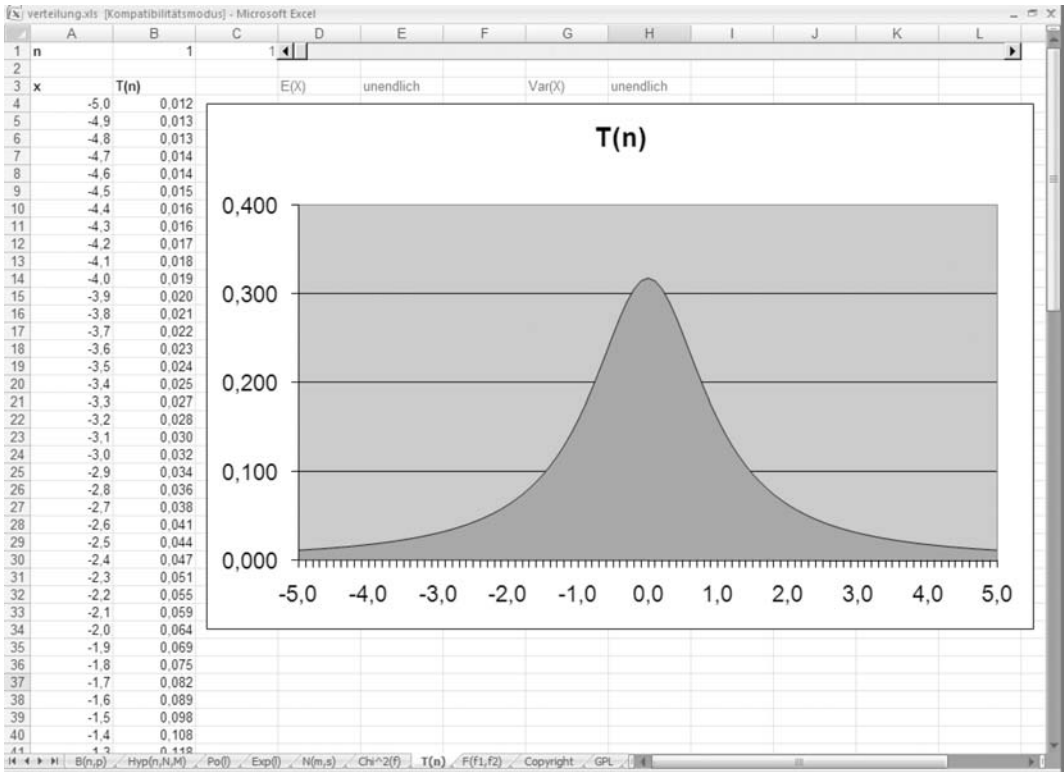
Figure 5. *Visualization of parameter changes with Excel.*

(Brandes, 2004) showed no significant improvement in the grades. Personal feedback from students has shown that the CD-ROM was used more as a replacement for textbooks than as an interactive e-learning tool.

Didactic concept: Didactic concepts are often not discussed or questions concerning technical implementations receive a much heavier importance in the development process. Most developers of e-learning projects are not educated in didactics but are mathematicians, statisticians or programmers.

All the above points boil down to one question: *Is the possible positive effect an e-learning tool may have on students understanding worth the effort the teacher needs to invest in it?*

We have considered the utility of e-teaching statistics for the particular field of web and computerbased-training applications and reviewed ten years worth of large-scale integrated electronic media elements in different areas of statistics education. The electronic media ranged from relatively simple Excel spreadsheets to integrated learning environments on the basis of HTML and Java. The aim of these implementations was to achieve improvements in the understanding of statistical concepts with the ultimate goal of better exam grades. We have found that the utility is marginal in comparison with the effort invested.

We believe that e-learning is required in modern statistics education but we do not share a too optimistic view that it will also deepen the knowledge of students in statistics. We rather believe that e-learning can be valuable for a certain proportion of students.
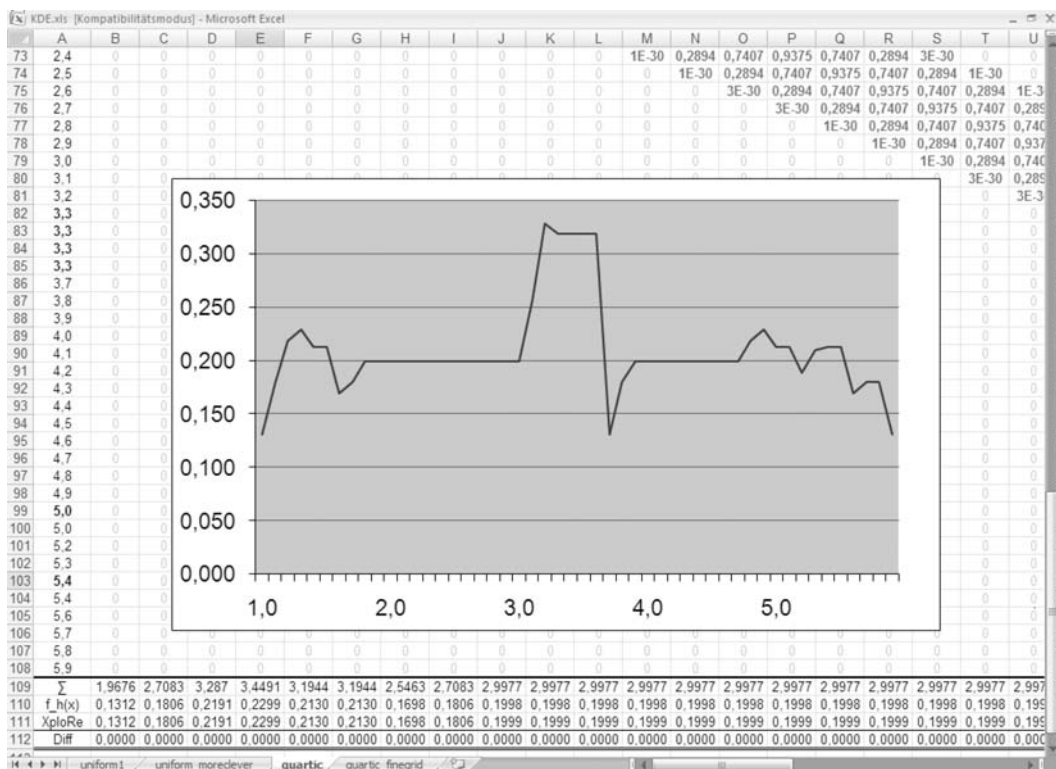
Figure 6. *Visualization of Kernel Density Estimation with Excel.*

Acknowledgement

This research was supported by the Deutsche Forschungsgemeinschaft through the SFB 649 'Economic Risk'.

References

Ahmad, T., Härdle, W. & Klinke, S. (2007). Using wiki to build an e-learning system in statistics in arabic language. In *CESSE 2007 Proceedings of the XXI. International Conference on Computer, Electrical, and Systems Science, and Engineering, Vienna Volume 21*, pp. 338–342.

Ahmad, T., Härdle, W. & Mungo, J. (2006). On the difficulty to design arabic e-learning systems in statistics. In *Proceedings in Computational Statistics—15th Symposium held in Rome*. Physika Verlag, Heidelberg, pp. 1619–1623.

Baumert, J., Artelt, C., Klieme, E., Neubrand, M., Prenzel, M., Schiefele, U., Schneider, W., Tillmann, K. & Weiß, M., Eds. (2002). PISA 2000—Die Länder der Bundesrepublik Deutschland im Vergleich, Vs Verlag.

Brandes, U. (2004). *Statistische Bewertung und Analyse der Klausurergebnisse Statistik (Grundstudium)*. Diploma thesis. URL: http://edoc.hu-berlin.de/docviews/abstract.php?lang=ger&id=26854

Cizek, P., Härdle, W. & Weron, R., Eds. (2005). *Statistical Tools for Finance and Insurance*. Chapter Working with the XQC. Heidelberg: Springer.

Darius, P. L., Portier, K. M. & Schrevens, E. (2007). Data collection skills: Challenges for teaching and training. *Int. Statist. Rev.*, **75**, 281–294.

Duller, C. (2004). A kind of PISA-survey at university, In *Proceedings in Computational Statistics—16th Symposium held in Prague*, Ed. J. Antoch, pp. 975–980. Heidelberg: Physika Verlag.

Duller, C. (2007a). Doing statistics versus understanding statistics. In *DAGStat—Statistics under one Umbrella*, p. 88.

Duller, C. (2007b). Do you speak statistics? In *Proceedings of the 56th Session of the ISI held in Lisbon*.

Härdle, W. & Simar, L. (2007). *Applied Multivariate Statistics*, 2nd ed. Heidelberg: Springer.

Klinke, S. (2004). Q&A—Variable multiple choice exercises with commented answers. In *Proceedings in Computational Statistics—16th Symposium held in Prague*, Ed. J. Antoch, pp. 1299–1304. Heidelberg: Physika Verlag.

Klinke, S. & Lehmann, H. (2003). *MD*Book and XQC/XQS—an architecture for reproducible research*, SFB 373 Discussion Paper.

McGraw-Hill (2001). *Visual Statistics*. Software. URL: http://www.mhhe.com/business/opsci/doane/home.htm

Mills, J. D. (2002). Using computer simulation methods to teach statistics: A review of the literature. *J. Stat. Educ.*, http://www.amstat-org/publications/jse/v30n1/mills.html.

Müller, M., Rönz, B. & Ziegenhagen, U. (2000). The multimedia project MM*STAT for teaching statistics. In *Proceedings in Computational Statistics—14th Symposium held in Utrecht*, Eds. J. Bethlehem & P. van der Heijden, pp. 409–415. URL: http://www.quantlet.com/mdstat/mmstat.html

Nolan, D. & Temple Lang, D. (2007). Dynamic, interactive documents for teaching statistical practise. *Int. Statist. Rev.*, **75**, 295–321.

Yates, D., Starnes, D.S. & Moore, D.S. (2007). *The Practice of Statistics & Fathom 2.0*. New York, USA: W. H. Freeman.

VAR Modeling for Dynamic Loadings Driving Volatility Strings

Ralf Brüggemann
University of Konstanz

Wolfgang Härdle
Humboldt-Universität zu Berlin

Julius Mungo
Humboldt-Universität zu Berlin

Carsten Trenkler
University of Mannheim

ABSTRACT

The implied volatility of an option as a function of strike price and time to maturity forms a volatility surface. Traders price according to the dynamics of this high dimensional surface. Recent developments that employ semiparametric models approximate the implied volatility surface (IVS) in a finite dimensional function space, allowing for a low dimensional factor representation of these dynamics. This paper presents an investigation into the stochastic properties of the factor loading time series using the vector autoregressive (VAR) framework and analyzes the dynamic relationship of these factors with economic indicators. (*JEL C14, C32*)

KEYWORDS: implied volatility surface, dynamic semiparametric factor model, vector autoregression, impulse responses

As undoubtedly the most important variable in finance, volatility appears consistently across a wide spectrum of theories and applications such as in asset pricing,

portfolio theory, risk management, derivatives, corporate finance, investment valuation and financial econometrics. For the fact that volatility is unknown, one studies the implied volatility (IV) which is derived from the Black and Scholes (1973) formula for a cross section of options with different strike and maturities traded at the same point in time. The implied volatility is seen as a mapping from time (t), strike price (K) and expiry date (T) to $\mathbb{R}^+$:

$$\hat{\sigma} : (t, K, T) \rightarrow \hat{\sigma}_t(K, T). \tag{1}$$

Calculating implied volatilities at different strikes and maturities yields an implied volatility surface (IVS). Due to the IVS dynamics, one faces the problem of high complexity in the sense that for each day simultaneously several maturities and levels of strikes are observed, forming a string structure that randomly changes shape and location as the options move towards expiry. The focus of several research work has been on finding efficient ways of modeling the IVS and to explain the driving forces behind its dynamics. An approach in the empirical finance literature for IV analysis is the utilization of factor type models, which can allow for low dimensional representation of the data. The estimation strategy involves estimating a two-dimensional IVS and applying principal component analysis on functional values. This modeling approach has been considered by Skiadopoulos, Hodges, and Clewlow (1999). Alexander (2001), Cont and Fonseca (2002). The models employed in the above studies are seen to disregard the specific string structure and thus tend not to capture the important features of the IVS dynamics. To overcome this problem and to ensure that the estimate is defined for the whole analyzed support, Fengler, Härdle, and Mammen (2007) have proposed a dynamic semiparametric factor model (DSFM), that offers flexible modeling for IVS fitting, dimension reduction and explanation of its dynamic behavior.

For a general presentation of the DSFM, consider the log-implied volatility Y_{tj}, where $t = 1, \ldots, I$ is the number of the day and $j = 1, \ldots, J_t$ is the number of IV observations on day t. Let $X_{t,j} = (\kappa_{tj}, \tau_{tj})$ be a two-dimensional covariate where κ_{tj} is a moneyness matrix and τ_{tj} denotes time-to-maturity. Here, future moneyness (a monotonic function of the strike price K_{tj}) defined as $\kappa_{tj} = \frac{K_{tj}}{S_t e^{(r_{tj}, \tau_{tj})}}$ is applied, where S_t is the spot price at time t and r_{tj} is the riskless interest rate for the option specific time to maturity obtained by linear interpolation of interest rate data, compare Dummas, Fleming, and Whaley (1998).

The DSFM approximates the IVS by regressing $Y_{tj} = \log\{\hat{\sigma}_{t,j}(\kappa, \tau)\}$ on $X_{t,j} = (\kappa_{tj}, \tau_{tj})$, expressed as

$$Y_{tj} = \sum_{k=0}^{K} z_{t,k} m_k(X_{t,j}) + \varepsilon_{t,j}, \tag{2}$$

where $z_{t,0} = 1$, m_k are smooth basis functions ($k = 0, \ldots, K$) and $z_{t,k}$ are weights or factor loadings depending on time. This model can be seen as a regression model with inherent time evolution. The estimation involves a nonparametric kernel

regression method with no assumption on the functional form of the factors except smooth basis function, earning the label "dynamic semiparametric factor model".

The estimates $\widehat{z}_{t,k}$ and $\widehat{m}_k$ of $z_{t,k}$ and m_k are obtained as minimizers of the following least squares criterion ($z_{t,0} = 1$):

$$\sum_{t=1}^{I} \sum_{j=1}^{J_t} \int \left\{ Y_{t,j} - \sum_{k=0}^{K} \widehat{z}_{t,k} \widehat{m}_k \right\}^2 K_h(u - X_{t,j})\, du, \tag{3}$$

where K_h denotes a two-dimensional kernel function, chosen as a product of one-dimensional kernels $K_h(u) = k_{h_1}(u_1) \times k_{h_2}(u_2)$, where $h = (h_1, h_2)^\top$ are bandwidths and $k_h(v) = k(h^{-1}v)/h$ is a one-dimensional kernel function. The optimization problem (3) is solved by an iterative procedure, see Borak, Härdle, and Fengler (2005). Hence, the time-dependent coefficients, $z_{t,k}$, of the smooth basis functions, which we refer to as factor loadings, are unobserved but can be estimated. An alternative estimation method for the DSFM has been proposed by Borak et al. (2007). Their approach uses nonparametric series estimation based on B-splines.

The choice for the number of dynamic basis function (K) is based on calculating the residual sum of squares per total variance,

$$RV(K) = \frac{\sum_t \sum_j \left\{ Y_{t,j} - \sum_{k=0}^{K} \widehat{z}_{t,k} \widehat{m}_k(X_{t,j}) \right\}^2}{\sum_t \sum_j (Y_{t,j} - \overline{Y})^2}, \tag{4}$$

where $\overline{Y}$ is the overall mean of the observation. K is chosen such that $1 - RV(K)$, the variance explained, is sufficiently high to give a good approximation to the IVS. For the technical details we refer to Borak et al. (2005), Fengler et al. (2007), and Borak et al. (2007).

It is worth mentioning that the DSFM differs from varying-coefficient models such as in Fan, Yao, and Cai (2003) and Yang et al. (2006) since the z_t are observables in these approaches. However, the DSFM methodology has some similarities to the model considered in Connor and Linton (2008) and Connor, Hagmann, and Linton (2007), which generalize the study of Fama and French (1992) on common movements of stock price returns. Here, the covariates are represented by $X_{l,j}$ and are time-invariant and differ for different m_l. The setting of these models allow for a direct application of backfitting procedures whereas the DSFM procedure can be seen as nonparametric curve estimation and backfitting for additive models. In addition, the DSFM nests linear models which allow for time varying coefficients as a special case, such as in Brumback and Rice (1998), and the one factor linear model for individual stocks implemented in Bakshi, Kapadia, and Madan (2003).

Using transaction-based DAX index implied volatility data from January 4, 1999 to February 25, 2003, Borak et al. (2005) obtained three factor loading series $z_t = (z_{t1}, \ldots, z_{tK})^\top$ with $K = 3$ after fitting the DSFM. Hence, three basis functions are used to model the implied volatility. They capture around 96% of the variation of the IVS, what seems to be high enough for a reliable approximation. Note that the DSFM operates on daily option data (closing prices). Hence, daily EURIBOR
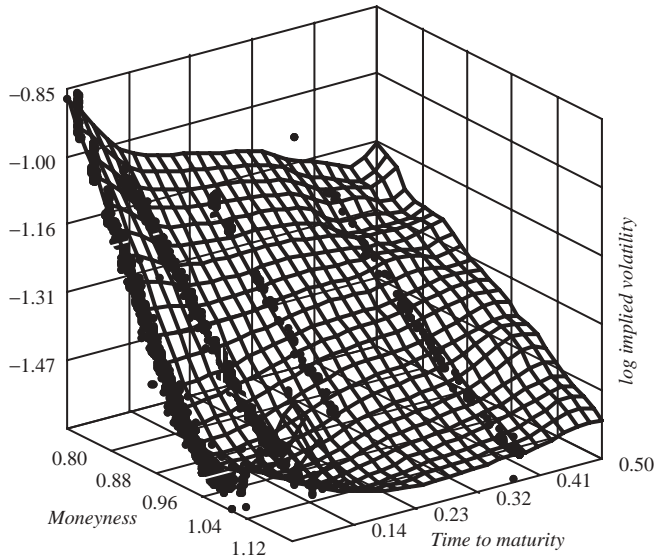
Figure 1 Implied volatility surface on the German DAX option on May 2, 2000 obtained through the DSFM estimator. Left axis denotes moneyness ($\kappa \in [0.8, 1.12]$) defined as $K/(S_t e^{r\tau})$, where K is the strike price, S_t is the spot price, r is the interest rate and τ is the time to maturity. Right axis denotes time to maturity ($\tau \in [0, 0.5]$) measured in years. For a specific maturity, the black bullets display the implied volatilities associated with different strike prices (moneyness values). This results in string structures for each of the maturities traded at the same time to which the implied volatility surface is fitted.

interest rate data have been used to define future moneyness. Figure 1 shows the IVS for the DAX trade on May 2, 2000 using a DSFM fit. The factor loadings describe the movements of the implied volatility surface such that z_1 may be interpreted as representing an overall volatility shift, z_2 represents changes in the maturity slope while z_3 represents changes in the smile curvature (moneyness slope) of the IVS.

Our analysis is guided by two aims. First, we want to provide an understanding of the dynamics and interaction of the factor loadings what may deliver insights in the behavior of the IVS. This, in turn, can help to provide a better assessment of market risks. Second, we want to study the relationship of the loadings with economic variables in order to identify possible spillover effects with respect to different markets.

To this end, we first examine the stochastic properties of the corresponding factor loading time series from Borak et al. (2005) by using VAR modeling techniques. The VAR framework is fairly general in describing the dynamics and interrelations between the variables of interest. We assume that the true loading series have been generated by a VAR process. Borak et al. (2007) show that the covariance structure of the estimated loading series converges in probability to the one of the unobservable series z_t if the latter is generated by a VAR process with certain properties. In a second step, we extend our VAR investigation by studying the associated

movements of the factor loadings with economic indicators representing returns and volatilities in money, bond, foreign exchange, commodity and global, i.e., U.S. stock markets.

Our VAR modeling results do reveal quite a bit of interaction between z_1 and z_2, i.e., between the volatility level and the maturity slope. This information may be used to assess market risks in a simplified way. One could develop a portfolio of stocks and (z_1, z_2) only, $Port(S_t, z_{t1}, z_{t2})$, with the intention to forecast its distribution in time $t + \Delta t$. Hence, z_{t1}, z_{t2} and their interaction, which represent the relevant systematic risk factors, can be used to obtain $Port(S_{t+\Delta t}, z_{1,t+\Delta t}, z_{2,t+\Delta t})$. Such an approach would avoid to simulate the whole IVS. Moreover, the consideration of economic indicators has uncovered significant mutual links between the loadings and the U.S. stock market. This applies both to the return dimension, represented by the log-returns of the S&P 500 index, as well as to the uncertainty dimension, represented by implied volatilities of call options on the S&P 500 index future. By contrast, the interaction with other markets is relatively weak (foreign exchange, money, and bond markets) or even insignificant (commodity markets).

Methodologically, we contribute to the market-spillover literature by analyzing spillover effects with respect to several dimensions of the IVS within a multivariate time series framework. This provides a more detailed picture of interactions. Volatility spillover effects have been analyzed e.g., by Billio and Pelizzon (2003), Veiga and McAleer (2004), and Andersen et al. (2007) using multivariate switching regime, VARMA-GARCH type and least squares regression model frameworks respectively. Furthermore, Andersen et al. (2003, 2007) have found that macroeconomic news releases have an simultaneous impact on stock, bond and foreign exchange markets. Related contributions in the macro-finance literature have focused on models of the term structure in which some connection between macroeconomic variables and yield curve components, *level*, *slope* and *curvature*, are investigated, see for example, Diebold and Li (2006), Diebold, Rudebusch, and Aruoba (2006), Hördahl, Tristani, and Vestin (2006) and Wu (2006).

The paper is organized as follows. We briefly describe the data used in our study and analyze the order of integration of the factor loading series in Section 1. Section 2 presents results from the VAR modeling and describes the dynamic interaction between the factor loading series. In Section 3 we relate these risk factors to economic indicators and conclude in Section 4.

1 THE DATA AND UNIT ROOT TESTS

We analyze time series data on factor loading series that have been obtained from a DSFM for DAX option data as specified by Borak et al. (2005). Their analysis is based on daily data on the DAX futures contract and DAX index options traded at the Futures exchange, EUREX, for the period from January 4, 1999 to February 25, 2003. The data set contains 1039 days of observations, excluding days with no option trades, and approximately 2.5 million contracts, i.e., about 2500 per day. Both futures and option price are contract based, i.e., each single contract is registered
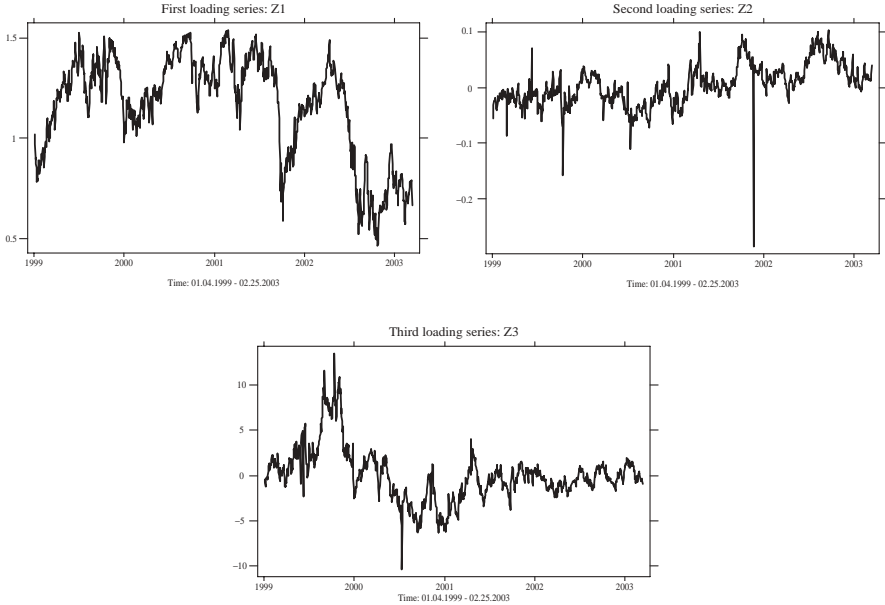
Figure 2 Factor loading time series from dynamic semiparametric factor model for implied volatility string dynamics.

together with its price, contract size, and time of settlement up to a hundredth second.

To reduce the noise and to diminish the effects of outlier a filter has been applied before estimating the model. Observations with implied volatility less than 4% and larger than 80% (extreme far out-of-the money options) are discarded. Time to maturity less than 10 days are discarded because of their sensitivity to small errors (expiry effect). The bandwidth in the time to maturity dimension is chosen relatively small to avoid that contracts with two adjacent time to maturities do not enter together in DSFM optimization. Thus, an effect of overlapping maturities on the estimated loading time series properties should be very weak.

The DSFM fitted by Borak et al. (2005) summarizes the IVS dynamics by $K = 3$ basis functions. Accordingly, we analyze three loading series for a period from January 4, 1999 to February 25, 2003 with $T = 1039$ observations in our sample. Corresponding time series plots are given in Figure 2.

Borak et al. (2007) have shown that under certain conditions the covariance structure of the estimated loading series converges in probability to the covariance structure of the unobservable loading series. Hence, with $\widehat{z}_t = (\widehat{z}_{t1}, \ldots, \widehat{z}_{t_K})^\top$ being the vector of the estimated loading series we have

$$\frac{1}{T} \sum_{t=h+1}^{T} (\widehat{z}_t - \bar{\widehat{z}}_t)(\widehat{z}_t - \bar{\widehat{z}}_t)^\top - \frac{1}{T} \sum_{t=h+1}^{T} (z_t - \bar{z}_t)(z_t - \bar{z}_t)^\top = o_p(T^{-1/2})$$

$$\bar{\widehat{z}}_t - \bar{z}_t = o_p(T^{-1/2}),$$

where $\bar{\hat{z}}_t = 1/T \sum_t^T \hat{z}_t$ and $\bar{z}_t = 1/T \sum_t^T z_t$. Besides a number of regularity conditions related to the estimation of the DSFM the following assumptions have to be imposed in order to obtain the convergence results.

(1) z_t is bounded with probability tending to one.
(2) z_t is strictly stationary with $\mathsf{E} \parallel z_t \parallel^\gamma < \infty$ for some $\gamma > 2$.
(3) z_t is strongly mixing with $\sum_{i=1}^\infty \alpha(i)^{(\gamma-2)/\gamma} < \infty$.
(4) The matrix $\mathsf{E}(z_t z_t^\top)$ has full rank.
(5) The process z_t is independent of X_{tj} and ε_{tj} $(t = 1, \ldots, T, j = 1, \ldots, J_t)$.

As pointed out by Borak et al. (2007), the asymptotic equivalence of the autocovariance structures carries over to classical estimation and testing procedures within the VAR model framework. Therefore, in order to simplify notation, we will use $z_t = (z_{t1}, z_{t2}, z_{t3})^\top$ in the following to describe the three estimated loading series obtained from Borak et al. (2005), i.e., z_t is understood to represent estimated series.

Assumption (2) requires the process z_t to be strictly stationary. A graphical inspection of the estimated loading series may doubt that they have been generated by a stationary process. The loading series do not only show volatility characteristics typical for financial market data but there are also indications of structural breaks in the series. To be precise, we observe a sudden downward movement in z_{t1} in September 2001, z_{t2} exhibits a clear outlier in November 2001 and the third loading series z_{t3} shows much stronger volatility in the first part of the sample than in the second one. In addition, the graph of z_{t3} also points towards a number of possible outliers. In order to account for possible structural breaks we have performed a unit root and VAR analysis for the two subsamples January 4, 1999–July 31, 2001 and August 1, 2001- February 25, 2003. However, the subsample results are not qualitatively different from the full sample outcomes. Therefore, we focus on the presentation of the latter ones and solely deal with stochastic nonstationarity in the following.

We have not found evidence for stochastic nonstationarity using the point-optimal unit root test of Elliott, Rothemberg, and Stock (1996) (ERS test). The small sample simulation results of Elliott et al. (1996) indicate that this test is superior to the augmented Dickey-Fuller procedure (cf. Fuller, 1976; Dickey and Fuller, 1979) in case of processes affected by conditional heteroscedasticity. The latter might be a relevant feature of our data. The ERS test is based on quasi-differences of z_{tk} which are defined by

$$d(z_{tk}|a) = \begin{cases} 1 & \text{if } t = 1 \\ z_{tk} - a\,z_{t-1,k} & \text{if } t > 1, \end{cases}$$

where a represents the point alternative against which the null of a unit root is tested. We follow the suggestion of Elliott et al. (1996) for a model with a constant only and use $a = \bar{a} = 1 - 7/T$.

Let $\hat{e}_t$, $t = 1, \ldots, T$, be the residuals from a regression of the time series on a quasi-differenced constant and let $S(\bar{a})$ and $S(1)$ be the sums of squared residuals for the cases $a = \bar{a}$ and $a = 1$ respectively. Then, the statistic testing the null hypothesis

Table 1 Unit root test statistics

Series	ERS-AIC	$\hat{b}$	ERS-HQ	$\hat{b}$
z_{t1}	3.787*	6	2.953**	6
Δz_{t1}	0.007***	5	0.075***	2
z_{t2}	5.295	8	3.338*	4
Δz_{t2}	0.663***	7	0.663***	7
z_{t3}	1.446***	7	1.446***	7
Δz_{t3}	0.005***	6	0.005***	6

ERS-AIC and ERS-HQ refer to the ERS tests using the AIC and HQ criteria to estimate the lag length b chosen for the estimation regression of the autoregressive spectral density estimator. Elliott et al. (1996) state 4.48 (10%), 3.26 (5%), and 1.99 (1%) as critical values for the ERS test. ***, **, and * denote significance at the 1%, 5%, and 10% level, respectively.

of a unit root is defined by $ERS = \{S(\bar{a}) - a\,S(1)\}/\hat{\omega}_b$, where $\hat{\omega}_b$ is the spectral density estimator of residuals $\hat{e}_t$, $t = 1, \ldots, T$, at frequency zero. We apply the autoregressive spectral density estimator as proposed by Elliott et al. (1996). In order to determine the lag length b of the corresponding estimation regression the Akaike information criterion (AIC) and the Hannan-Quinn (HQ) criterion are used.[1] The limiting distribution of the test statistic is nonstandard and critical values are stated in Elliott et al. (1996).

The ERS test results given in Table 1 clearly indicate that z_{t1} and z_{t3} are stationary. Regarding z_{t2} we reject the unit root null hypothesis at the 10% level when using the HQ criterion. The nonrejection in case of the AIC criterion could be due to the high lag length chosen for the estimation regression on which the autoregressive spectral density estimator is based. Given the unit root test results we proceed to work with the assumption that no unit root is present in the loading series.

2 VAR MODELS FOR FACTOR-LOADING DYNAMICS

VAR models are often used to investigate the dynamic relationship between the variables of interests. In our case, we use the VAR modeling framework to investigate the relationship between the factor loading times series described earlier. As before $z_t = (z_{t1}, z_{t2}, z_{t3})^\top$ denotes the vector containing the observations of the $K = 3$ estimated factor loadings z_{t1}, z_{t2} and z_{t3} at time t. Hence, we model the dynamics underlying the factor loadings by a VAR process of order p, VAR(p),

$$z_t = \nu + A_1 z_{t-1} + \cdots + A_p z_{t-p} + u_t, \tag{5}$$

where ν is a $K \times 1$ vector of intercept parameters, $A_i, i = 1, \ldots, p$, are fixed $K \times K$ parameter matrices and, $u_t = (u_{t1}, \ldots, u_{tK})^\top$, is an unobservable error term vector with mean zero and time-invariant, non-singular covariance matrix Σ_u, i.e., $\mathsf{E}[u_t u_t^\top] = \Sigma_u$.

---

[1] The criteria are described in Lütkepohl (2005, Chapter 4).

Table 2 Lag length suggested by information criteria

	01/04/1999–02/25/2003			02/16/2000–02/25/2003		
Model	AIC	HQ	SC	AIC	HQ	SC
z_{t1}, z_{t2}, z_{t3}	7	3	2	—	—	—
z_{t1}, z_{t2}, z_{t3}, EX_t, COM$_t$, R12M$_t$, SP500$_t$	3	2	1	—	—	—
z_{t1}, z_{t2}, z_{t3}, EXV_t, GOLV$_t$, BUNDV$_t$, SP500V$_t$	—	—	—	3	2	1

Table entries give the optimal number of lags determined using standard information criteria with a maximum lag order of $p_{max} = 12$.

Table 3 p-values of diagnostic tests

Model	Sample	p	Q(20)	LMF(4)	LMF(8)	LBJ	ARCH(1)
z_{t1}, z_{t2}, z_{t3}	Full	7	0.22	0.09	0.38	0.00	0.00
z_{t1}, z_{t2}, z_{t3}, EX_t, COM$_t$, R12M$_t$, SP500$_t$	Full	5	0.06	0.06	0.38	0.00	0.00
z_{t1}, z_{t2}, z_{t3}, EX_t, COM$_t$, R12M$_t$, SP500$_t$	Sub	8	0.24	0.31	0.67	0.00	0.00

Diagnostic tests: full sample 01/04/1999–02/25/2003 and sub-sample 02/16/2000–02/25/2003. Q(20) denotes an adjusted portmanteau test involving 20 autocorrelation matrices, LMF(4) and LMF(8) are LM tests for autocorrelation of order 4 and 8. LJB represents the multivariate Lomnicki–Jarque–Bera tests for nonnormality and ARCH(1) is the multivariate first-order ARCH test. All the tests are described in more detail in Lütkepohl (2004). Computations are performed with JMulTi, (Lütkepohl and Krätzig 2004).

To select the lag length p, we have applied standard information criteria as in Lütkepohl (2005, Chapter 4) to VAR models with a maximum lag order of $p_{max} = 12$. These criteria balance the trade-off between model fit and the number of parameters to be estimated and allow a fairly parsimonious model specification. The results are given in Table 2. The three considered information criteria suggest different lag lengths for the sample under consideration. As we want to model the dynamic relations between the factor loading series by impulse response functions we start out by a more general model with $p = 7$ lags as suggested by the AIC.

Choosing the lag length by AIC may be advantageous for our purpose given the results in Brüggemann (2004), as a fairly large model allows to capture the underlying dynamics in a more flexible way. We check the adequacy of the model by applying a number of standard diagnostic tests whose results are reported in Table 3. While the estimated residuals do not show signs of autocorrelation, normality and conditional homoscedasticity are clearly rejected. Both, non-normality and ARCH effects are often observed in empirical models for financial variables. In fact, these results are not surprising given the time series plots in Figure 2.

However, non-normality does not affect the asymptotic properties of the methods applied later on. The presence of conditional heteroscedasticity may distort estimators referring to second moments as e.g., confidence intervals. Hence, corresponding results should be interpreted with some care. For instance, the empirical coverage of the reported confidence bands might be slightly lower than the nominal coverage.

The dynamic relations between the loadings is captured by the impulse response function, Lütkepohl (2005, Chapter 3). If the VAR(p) model is stationary it has a moving average (MA) representation

$$z_t = \mu + \sum_{i=0}^{\infty} \Phi_i u_{t-i}, \tag{6}$$

where $\mu := (I_K - A_1 - \cdots - A_p)^{-1} v$ is the mean of the process. The MA matrices Φ_i contain the responses to forecast errors u_t that occurred i periods ago. These responses are difficult to interpret if the contemporaneous residual correlation is high. In fact, the estimated residual correlation matrix for our benchmark model is

$$\hat{P}_u = \begin{pmatrix} 1 & -0.49 & -0.23 \\ -0.49 & 1 & -0.10 \\ -0.23 & -0.10 & 1 \end{pmatrix}. \tag{7}$$

Given the sample size of $T = 1039$ observations, the off-diagonal elements of $\hat{P}_u$ are fairly large. A common approach in this situation is to consider orthogonalized shocks obtained by a Cholesky decomposition of Σ_u. However, the results of the impulse response analysis may then depend on the ordering of the variables in the system. To remove the dependence on the variable ordering we use generalized impulse responses introduced by Pesaran and Shin (1998). In their approach shocks are orthogonalized by looking at a shock in variable k and integrating out the effects of other shocks using the distribution of the errors, i.e., the correlation among the components of u_t is taken into account. If u_t has a multivariate normal distribution, then it can be shown that

$$E(u_t | u_{kt} = \delta_k) = (\sigma_{1k}, \sigma_{2k}, \ldots, \sigma_{Kk})^\top \sigma_{kk}^{-1} \delta_k = \Sigma_u e_k \sigma_{kk}^{-1} \delta_k,$$

where $\{\sigma_{jk}, j, k = 1, \ldots, K\}$ denotes the elements of Σ_u and e_k is a $K \times 1$ selection vector with a 1 in position k and zeros elsewhere. Hence, the response vector to a shock in variable k that occurred i periods ago is

$$\frac{\Phi_i \Sigma_u e_k}{\sqrt{\sigma_{kk}}} \frac{\delta_k}{\sqrt{\sigma_{kk}}}. \tag{8}$$
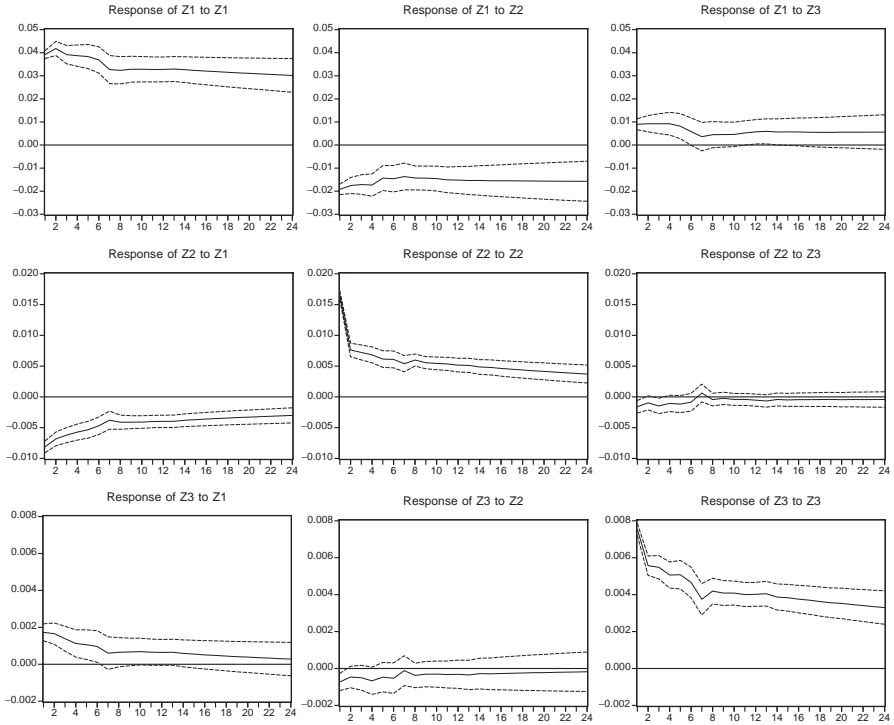
Figure 3 Generalized impulse responses of factor loadings obtained from a VAR(7) for $z_t = (z_{t1}, z_{t2}, z_{t3})^{\top}$. Dashed lines indicate approximate 95% confidence intervals. Sample: 01/04/1999–02/25/2003.

We scale δ_k such that the shocks have a size of one standard deviation, i.e., we set $\delta_k = \sqrt{\sigma_{kk}}$ and thus report the response vectors

$$\theta_k(i) = \sigma_{kk}^{-\frac{1}{2}} \Phi_i \Sigma_u e_k, \quad i = 0, \ldots, h, \tag{9}$$

which gives the generalized impulse responses of the variables in z_t to a shock in variable k that occurred i periods ago.

In the first column of Figure 3, we give the responses to a positive innovation in the first loading series. This innovation has a permanent negative effect on the second factor loading z_{t2} and a small but positive effect on z_{t3}, which becomes insignificant after about 6 periods. Similarly, we see that an innovation in the second loading factor (column 2 of Figure 3) has a permanent negative effect on z_{t1}, while the effect on z_{t3} is not significantly different from zero. In contrast, a shock in the third loading series has only a short-lived positive effect on z_{t1} but no significant response in z_{t2} is observed.

To further investigate the relationship between the variables we have also conducted Granger causality tests, see e.g., Granger (1969) and Lütkepohl (2005). The results of the tests are given in Table 4. Granger non-causality of z_{t1} for z_{t2} and

Table 4 Causality tests

H_0	Test result
$z_{t1} \nrightarrow z_{t2}, z_{t3}$	$F(14, 3030) = 4.69[0.00]$
$z_{t2} \nrightarrow z_{t1}, z_{t3}$	$F(14, 3030) = 1.87[0.02]$
$z_{t3} \nrightarrow z_{t1}, z_{t2}$	$F(14, 3030) = 0.76[0.71]$
$z_{t1}, z_{t2} \nrightarrow z_{t3}$	$F(14, 3030) = 0.74[0.74]$
$z_{t3} \nrightarrow z_{t1}$	$\chi^2(7) = 6.41[0.49]$
$z_{t3} \nrightarrow z_{t2}$	$\chi^2(7) = 3.19[0.87]$
$z_{t1} \nrightarrow z_{t3}$	$\chi^2(7) = 8.22[0.31]$
$z_{t2} \nrightarrow z_{t3}$	$\chi^2(7) = 4.80[0.68]$

Results are based on a model for z_t using $p = 7$ and data for the full sample period 01/04/1999–02/25/2003. $\nrightarrow$ denotes 'does not Granger cause'. p-values are given in square brackets. All the tests are described in more detail in Lütkepohl (2004). Computations are performed with JMulTi, (Lütkepohl and Krätzig 2004).

z_{t3} and of z_{t2} for z_{t1} and z_{t3} is rejected at the 5% significance level. Investigating the causal relationships more closely reveals that z_{t3} is neither Granger-caused by z_{t1} nor z_{t2}. Moreover, Granger non-causality from z_{t1} to z_{t3} and from z_{t2} to z_{t3} cannot be rejected. In other words, including z_{t3} does not help predicting z_{t1} and z_{t2} and predictions of z_{t3} are not improved by including z_{t1} and z_{t2}. This result is in line with the impulse response pattern discussed earlier. A possible conclusion is that z_{t3} may as well be excluded from the VAR system. In fact, using any of the sequential model reduction algorithms discussed in Brüggemann (2004, Chapter 2) leads to a model where lags of z_{t3} are excluded from the equations for z_{t1} and z_{t2} and moreover, z_{t3} only depends on its own lags.[2]

To sum up, the VAR model analysis has revealed quite a bit of interaction between the first and second loading. A positive shock in the first loading has a permanent negative impact on the second and vice versa. Note that movements in z_{t1} may be interpreted as overall shifts (up or down) of the IVS whereas z_{t2} represents changes in the maturity slope of the IVS. Thus, an overall increase in financial market risk is associated with an upward tilt of the maturity slope. In other words, the risk of longer maturities increases relative to shorter maturities. Accordingly, a decrease in the relative risk of long-term options induces a general risk reduction (lower overall implied volatility). The third loading series is not importantly related to z_{t1} and z_{t2}. Hence, changes in the moneyness slope of the IVS are not directly linked to up- or downward shifts of the IVS or to adjustments in the relative risk of long- and short-term options.

3 LOADINGS AND ECONOMIC INDICATORS

As a first step to analyze the relationship between economic variables and the loading series we extend our benchmark specification VAR model by including

---

[2]We have repeated the impulse response analysis for a system including only z_{t1} and z_{t2} and found impulse-responses virtually identical to the upper left 2×2 block of Figure 4.
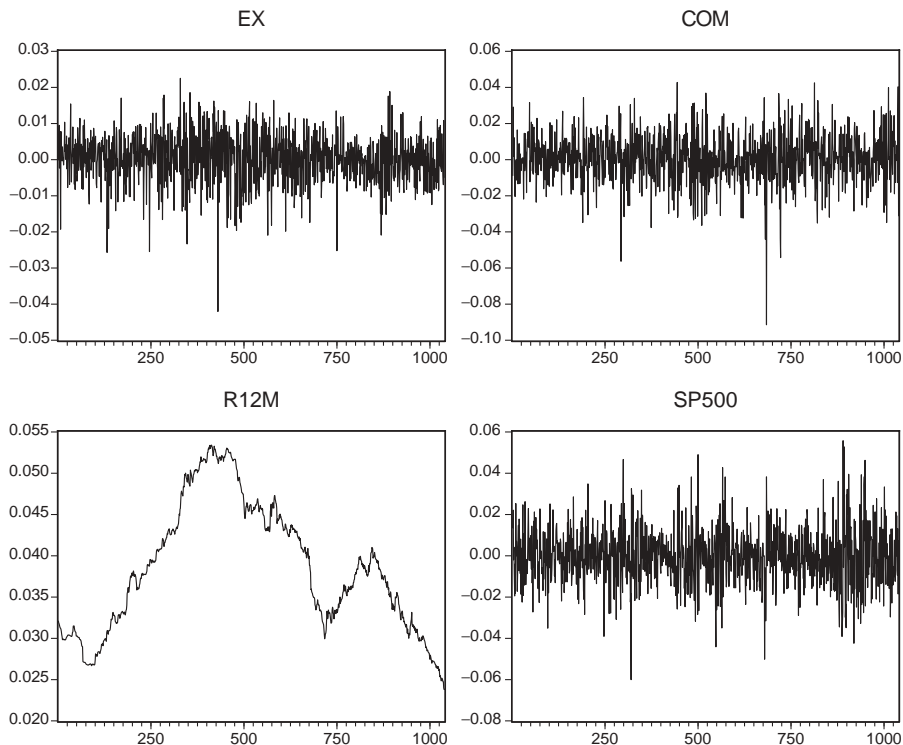
Figure 4 Time series plot of economic indicators for sample period: 01/04/1999–02/25/2003.

potentially important economic time series. To be precise, the log-return of the US dollar per Euro (EX_t), the log-return of the Goldman Sachs commodity price index (COM_t), the 12-months German money market interest rate ($R12M_t$), and the log-return of S&P 500 index ($SP500_t$) from 1.04.1999–2.25.2003 are considered (see Figure 4).[3] The first three variables are selected in order to analyze possible links between the loadings, i.e., risk in the German stock market, and the foreign exchange, money as well as the commodity markets. The choice of the 12-month interest rate ($R12M_t$) represents a compromise between short- and long-run rates and the return of the commodity index is chosen because the index captures potential effects of several commodity markets. $SP500_t$ represents global stock market factors that may effect the German stock market. Of course, global effects are also contained in the other economic indicators. However, we account for resulting contemporaneous error term correlation by using generalized impulse responses. For an identification of a global (risk) factor via exchange rate option prices see Bakshi, Carr, and Wu (2008). We have chosen log-returns instead of price series in order to avoid statistical and interpretational difficulties caused by the high persistence of price data. In a second step, we also use volatility data as specific

---

[3] All data have been retrieved from the Thomson Datastream database.
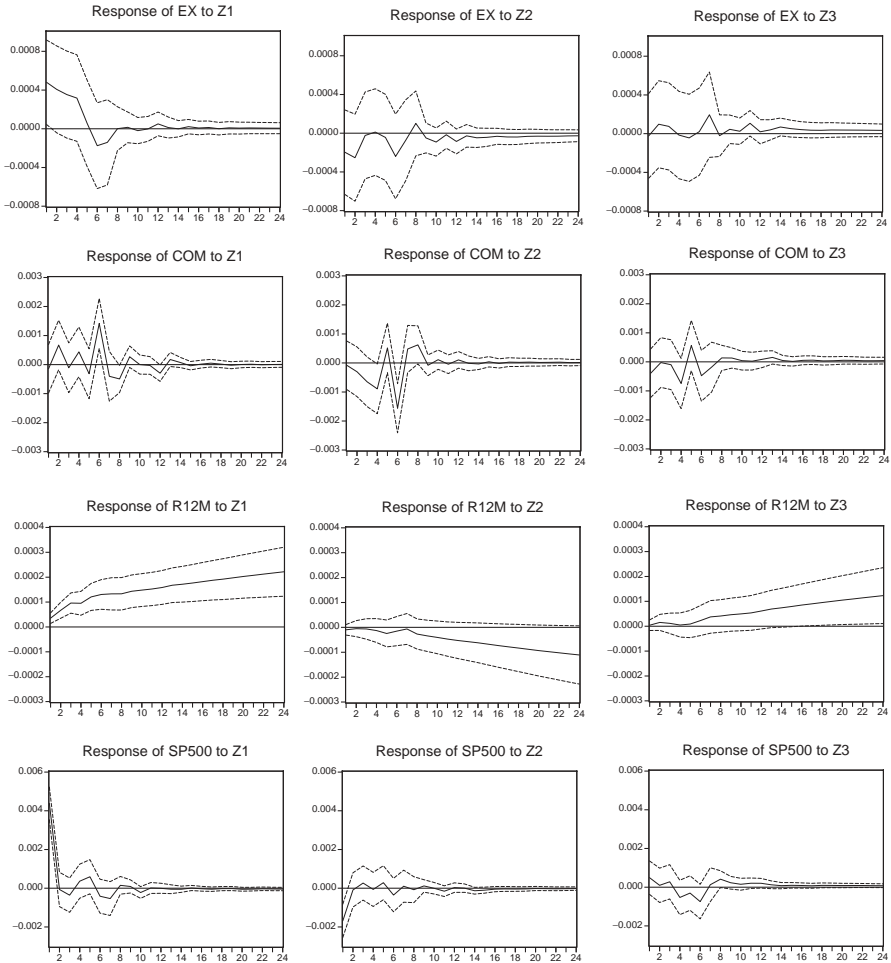
Figure 5 Generalized impulse responses of log-return of the US dollar per Euro exchange rate (EX_t), log-return of commodity index (COM_t), the 12-month money market interest rate ($R12M_t$) and the log-return of the S&P 500 index ($SP500_t$) to shocks in the factor loadings series obtained from a VAR(5) for $z_t = (z_{t1}, z_{t2}, z_{t3}, EX_t, COM_t, R12M_t, SP500_t)^\top$. Dashed lines indicate approximate 95% confidence intervals. Sample: 01/04/1999–02/25/2003.

measures of risk factors associated with the markets mentioned above. However, we have not obtained closer links with the loadings in general but a somewhat different structure of the relationships. We discuss this in more detail at the end of this section.

Given the results of information criteria and diagnostic tests (compare Tables 2 and 3), the best choice to analyze the system $z_t = (z_{t1}, z_{t2}, z_{t3}, EX_t, COM_t, R12M_t, SP500_t)^\top$ is a VAR(5). The dynamic characteristics of the system is exploited through the generalized impulse response technique and the results are presented in Figures 5 and 6. Most relevant are the relations between the return of the S&P 500
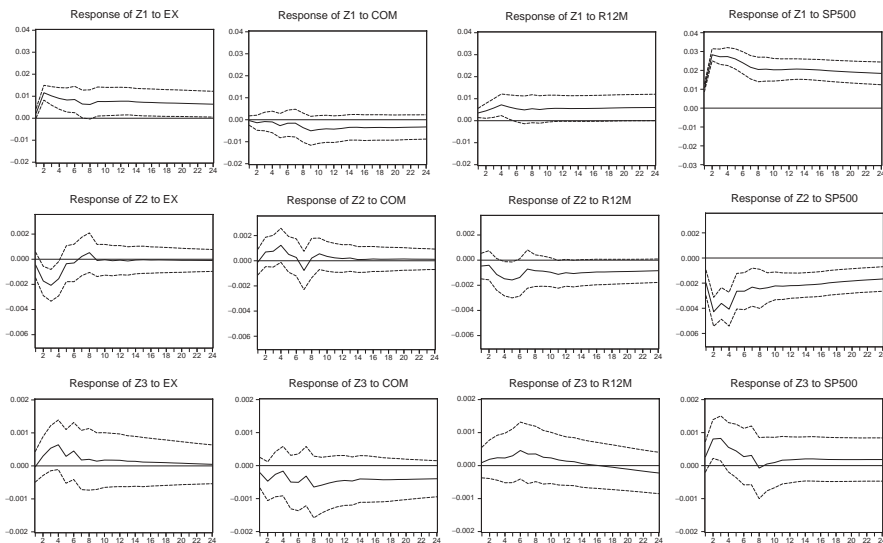
Figure 6 Generalized impulse responses in the factor-loadings series to shocks in log-return of the US dollar per Euro exchange rate (EX_t), log-return of commodity index (COM_t), the 12-month money market interest rate ($R12M_t$) and the log-return of the S&P 500 index ($SP500_t$) obtained from a VAR(5) for $z_t = (z_{t1}, z_{t2}, z_{t3}, EX_t, COM_t, R12M_t, SP500_t)^\top$. Dashed lines indicate approximate 95% confidence intervals. Sample: 01/04/1999–02/25/2003.

index and the loadings but we also observe a number of significant impulse responses involving the 12-month interest rate and the exchange rate return. However, there is no significant link with COM_t.

Clearly, all loadings respond significantly to a shock in $SP500_t$ (compare last column of Figure 6): z_{t1} is permanently positively affected, z_{t2} permanently negatively, and z_{t3} positively responds 1–3 days after the shock. Thus, given the interpretation of the first and second loading, a positive return shock in the U.S. stock market induces higher overall risk in the German market and increases the risk of longer maturities relative to shorter maturities due to the downward tilt of the IVS. Note that changes in the third loading affect the moneyness slope and smile curvature. Accordingly, the return shock temporarily enhances the smile curvature and raises the relative risk of options with low moneyness values. All these effects may be due to shifts in investment from the German to the U.S. stock market as a response to the positive U.S. return shock.

Not surprisingly, $SP500_t$ is much less affected by shocks in the loadings, confirming the conventional wisdom that German shocks do not greatly effect the U.S. market. Nevertheless, a positive overall risk shock in the German stock market and an increase in the risk of longer relative to shorter maturities lead to significant positive responses in $SP500_t$ on impact. We interpret these responses as reactions of the U.S. stock market to changes in outside risk factors. These risk factors may be captured by the loadings since our model does not include a general measure of outside forces from the point of view of the U.S. market.

Furthermore, the 12-month interest rate responds significantly positive to a shock in z_{t1} in a permanent way. This response may be regarded as a change in risk compensation such that higher overall risk induces an interest rate rise. Interestingly, we also observe in Figure 6 that z_{t1} significantly responds to a shock in R12M$_t$ although the permanent effect is only marginally significant. If we regard rising interest rates as an indicator for higher inflation then it is likely that higher interest rates are accompanied by increasing uncertainty in the financial markets. Positive shocks in EX$_t$ are relevant for the IVS dynamics. The first loading positively responds to a positive shock in EX$_t$ in a permanent way. Thus, a rise in EX$_t$, i.e., an appreciation of the Euro, increases the volatility of the DAX options. At a first glance, this may seem surprising. However, a stronger Euro increases uncertainty about future profits of German DAX companies, which may, in turn, lead to a volatility increase in the stock market. Some further responses involving R12M$_t$ and EX$_t$ are weakly significant or close to significance but given the uncertainty due to the ignored conditional heteroscedasticity we do not attempt to interpret them.

We have also analyzed alternative interest rates and specific commodities including returns of corresponding future series and briefly comment on the respective results. Replacing the return of the commodity index by the return of a specific commodity like oil or gold or by the return of some more specific index like energy and metals spot price index does not alter the results in any important way. Since volatility is backed-out by interest rates referring to the option specific time to maturity, the dynamic relationships between the loadings and R12M$_t$ may simply result from correlation of the level, controlled for by R12M$_t$, and the slope of the term structure. Therefore, we have also analyzed VAR models with the spread of the 12-month and 3-month money market interest rates instead of R12M$_t$. However, the significance of responses involving the spread is clearly lower compared to R12M$_t$. Therefore, we conclude that the obtained links between R12M$_t$ and the loadings are of direct nature. We have also found weaker links when replacing R12M$_t$ by the 3-month interest rate or an average bond rate referring to longer maturities, e.g., 5–6 years. The latter might be surprising since long-term interest rates contain information on the term premia, compare Cochrane and Piazzesi (2006). Neither did we obtain responses of higher significance when using returns of future series of the commodity index, of gold and of the exchange rate.

Subset VAR models with zero restrictions lead to qualitatively similar results. Nevertheless, a few more significant responses can be observed and some responses are now associated with a slightly higher degree of significance. Both slightly larger responses and tighter confidence intervals are the reasons for the changes. Yet, we believe that the unrestricted VAR results provide a more realistic impression of the significance of the responses given the ignored conditional heteroscedasticity.

Finally, we turn to the analysis of volatility data. To capture risk factors in foreign exchange rate, commodity, bond (money), and the U.S. markets we use the implied volatility of continuous call options on the CME Eurodollar future (EXV$_t$), Anglogold Ltd. (GOLDV$_t$), the Euro-BUND future (BUNDV$_t$), and the CME S&P 500 index future (SP500V$_t$) for the sample 2.16.2000–2.25.2003. We have referred to
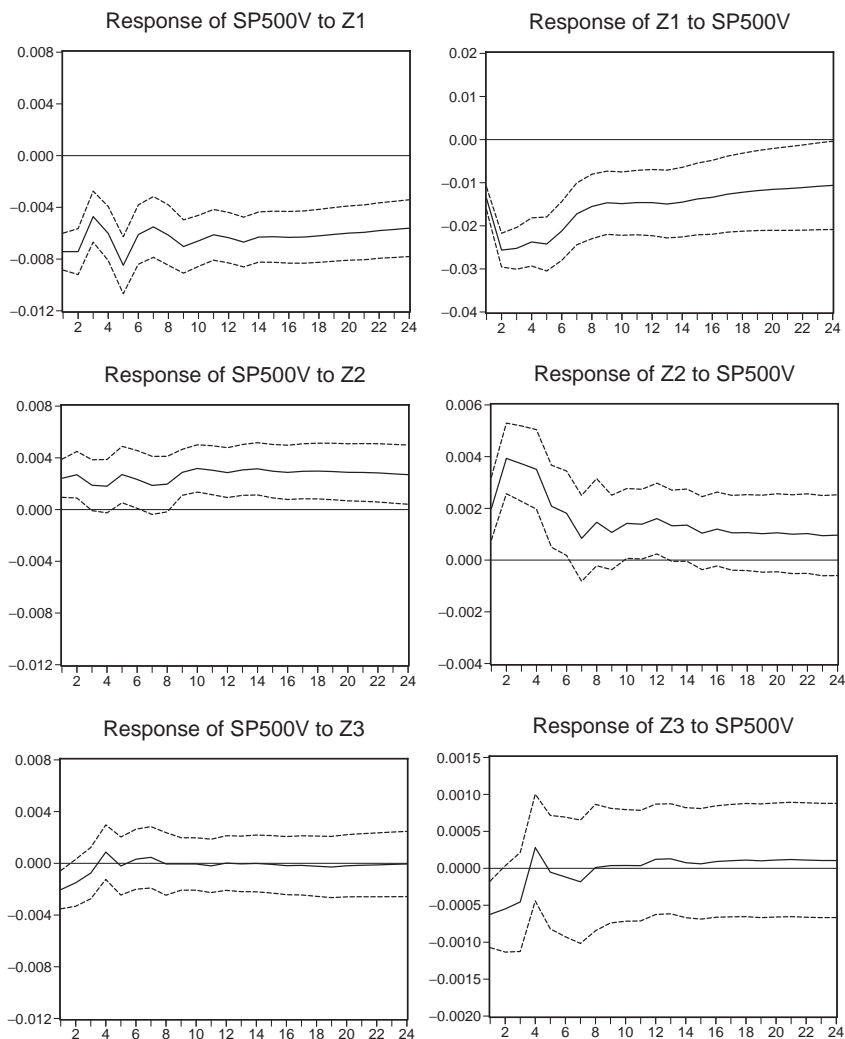
Figure 7 Generalized impulse responses of implied volatility of continuous call options of the CME S&P 500 index future (SP500V$_t$) to shocks in the factor-loadings series and responses in the factor-loadings series to shocks in implied volatility of continuous call options of the CME S&P 500 index future (SP500V$_t$) obtained from a VAR(8) for $z_t = (z_{t1}, z_{t2}, z_{t3}, \text{EXV}_t, \text{GOLDV}_t, \text{BUNDV}_t, \text{SP500V}_t)^\top$. Dashed lines indicate approximate 95% confidence intervals. Sample: 02/16/2000– 02/25/2003.

future options to have longer time series for the analysis. Using corresponding put option data leads to very similar results.

We need to fit a VAR(8) model in order to obtain residuals free of autocorrelation, i.e., we have to exceed the orders suggested by the information criteria, compare Table 2. We make three main observations compared to the responses

involving the return data. First, most of the relevant interactions involve again the U.S. market measure. There is practically no significant relation between BUNDV$_t$ and the loadings. Furthermore, z_{1t} does not significantly respond to a shock in EXV$_t$ compared to EX$_t$. Second, responses in SP500V$_t$ and EXV$_t$ to shocks in the loadings are clearly more pronounced than the corresponding reactions of the returns. In other words, changes in the risk level and structure of the German market as measured by the loadings are more important for the uncertainty dimension (volatility) than the return dimension of the U.S. and the foreign exchange markets. By contrast, volatility shocks to the foreign exchange or U.S. market do not feed stronger into the loading dynamics than shocks to the corresponding return levels do. However, the sign of the relationships between the U.S. market and the loadings has changed and this comprises the third observation.

As can be seen in Figure 7 we now have a negative link between the overall risk level in the German market and the implied volatility of the S&P 500 index future. Moreover, the relative risk of longer and shorter maturities is oppositely related to SP500V$_t$ than to SP500$_t$. The same signs for the responses have been obtained when replacing EXV$_t$, GOLDV$_t$, BUNDV$_t$, and SP500V$_t$ by standard deviations derived from GARCH (1,1) models fitted to the return series.[4] Since the shocks are orthogonalized, the negative link between z_{1t} and SP500V$_t$ may just reflect investment strategies of international investors. An isolated negative volatility shock in one market makes the other market more attractive inducing investments into the latter one. It is not obvious how larger investments affect volatility but higher market liquidity due to increased investments may lower risk.

4 CONCLUSIONS

In this paper, we have analyzed and modeled the stochastic properties of factors of volatility strings derived from a dynamic semiparametric model for implied volatility. The VAR modeling framework applied provides a fairly good description of the dynamics and interrelations between the factor loadings that determine the movements of the IVS. The uncovered interaction between the first two loadings, which represent changes in the volatility level and maturity slope respectively, is a very useful information for risk management. Not only the individual effects of the loadings but also their dynamic interaction is likely to be a relevant factor for assessing the effect of market risks e.g., on the performance of stock portfolios.

The relationship of the loadings to movements in economic variables was also investigated. Our results reveal that the loading series are mutually linked to the U.S. stock market return and volatility. Hence, global return and risk factors influence German stock market risk, but the German market also has an effect on the U.S. market. This study may be seen as a first step in associating movements in some risk factors with movements in other markets' conditions, an ingredient

---

[4]Note, however, that the number of significant responses is clearly lower when standard deviations derived from the GARCH models are used.

necessary to give an accurate assessment of market risks. Therefore, an important outlook is then to develop useful strategies for hedging against these risk factors.

A possible extension in the direction of market spillover effects would be to analyze the interaction of two sets of loadings that drive the IVS of different stock markets respectively. Having a set of loadings for an additional stock market would also allow to scrutinize the findings obtained for the German stock market. Additionally, one may classify options into different groups like at-the-money, out-of-the money and deep-out-of the money and compute the implied volatilities for these groups at a preferred frequency. In that, the IVS would reduce to a set of sampled series describing the development of implied volatilities of certain option groups. Such an approach could be a fruitful extension of our study.

Regarding the application of our method to individual stocks we should point out that the DSFM approach needs not to be adjusted if the definition of moneyness controls for stock volatility by using $K/Se^{r\tau}\sigma$, where σ is at-the-money volatility. Such an adjustment just reflects a shift in moneyness by some percentages. Since the first loading represents the level of implied volatility, it should be possible to exploit the empirical VAR framework used in the paper for volatility forecasting. Improved forecast quality could be possible if expected future volatility is an important component of implied volatility. Thereby, one may be able to complement recent advances in the volatility forecasting literature, see e.g., Andersen, Bollerslev, and Diebold (2007).

REFERENCES

Alexander, C. (2001). "Principles of the Skew." *Risk 14*, 29–32.

Andersen, T. G., T. Bollerslev, and F. X. Diebold (2007). "Roughing it Up: Including Jump Components in the Measurement, Modeling and Forecasting of Return Volatility." *The Review of Economics and Statistics 89*, 701–720.

Andersen, T. G., T. Bollerslev, F. X. Diebold, and C. Vega (2003). "Micro Effects of Macro Announcements: Real Time Discovery in Foreign Exchange." *American Economic Review 93*, 38–62.

Andersen, T. G., T. Bollerslev, F. X. Diebold, and C. Vega (2007). "Real-Time Price Discovery in Global Stock, Bond and Foreign Exchange Markets." *Journal of International Economics 73*, 251–277.

Bakshi, G., N. Kapadia, and D. Madan (2003). "Stock Return Characteristics, Skew Laws, and the Differential Pricing of Individual Equity Options." *Review of Financial Studies 16(1)*, 101–143.

Bakshi, G., P. Carr, and L. Wu (2008). "Stochastic Risk Premiums, Stochastic Skewness in Currency Options, and Stochastic Discount Factors in International Economies." *Journal of Financial Economics 87*, 132–156.

Billio, M., and L. Pelizzon (2003). "Volatility and Shocks Spillover before and after EMU in European Stock Markets." *Journal of Multinational Financial Management 13*, 323–340.

Black, F., and M. Scholes (1973). "The Pricing of Options and Corporate Liabilities." *Journal of Political Economy 81*, 637–659.

Borak, S., W. Härdle, E. Mammen, and B. U. Park (2007). "Time Series Modelling with Semiparametric Factor Dynamics." SFB 649 Discussion Paper No. 023. Humboldt-Universität zu Berlin.

Borak, S., W. Härdle, and M. Fengler (2005). "DSFM Fitting of Implied Volatility Surfaces." *Proceedings 5th International Conference on Intelligent, System Design and Applications*. IEEE Computer Society Number P2286, Library of Congress Number 2005930524.

Brüggemann, R. (2004). *Model Reduction Methods for Vector Autoregressive Processes*. Lecture Notes in Economics and Mathematical Systems 536. Berlin: Springer-Verlag.

Brumback, B., and J. A. Rice (1998). "Smoothing Splines Models for the Analysis of Nested and Cross Samples Curves." *Journal of the American Statistical Association 93*, 961–994.

Cochrane, J. H., and M. Piazzesi (2006). "Decomposing the Yield Curve." Working Paper. University of Chicago.

Connor, G., M. Hagmann, and O. Linton (2007). "Efficient Estimation of a Semiparametric Characteristic-Based Factor Model of Security Returns." Research Paper No. 07-26. Swiss Finance Institute.

Connor, G., and O. Linton (2008). "Semiparametric Estimation of a Characteristic-Based Factor Model of Common Stock Returns." *Journal of Empirical Finance, 14*, 694–717.

Cont, R., and J. Fonseca (2002). "Dynamics of Implied Volatility Surfaces." *Quantitative Finance 2*, 45–60.

Dickey, D., and W. Fuller (1979). "Estimators for Autoregressive Time Series with a Unit Root." *Journal of the American Statistical Association 74*, 427–431.

Diebold, F. X., and C. Li (2006). "Forecasting the Term Structure of Government Bond Yields." *Journal of Econometrics 130*, 337–364.

Diebold, F. X., G. D. Rudebusch, and S. B. Aruoba (2006). "The Macroeconomy and the Yield Curve: A Dynamic Latent Factor Approach." *Journal of Econometrics 131*, 309–338.

Dummas, B., J. Fleming, and R. E. Whaley (1998). "Implied Volatility Functions: Empirical Tests." *Journal of Finance 53*, 2095–2106.

Elliott, G., T. J. Rothemberg, and J. H. Stock (1996). "Efficient Tests for an Autoregressive Unit Root." *Econometrica 64*, 813–836.

Fama, E. F., and K. R. French (1992). "The Cross-Section of Expected Stock Returns." *Journal of Finance 47*, 427–465.

Fan, J., Q. Yao, and Z. Cai (2003). "Adaptive Varying-Coefficient Linear Models." *Journal of the Royal Statistical Society B 65*, 57–88.

Fengler, M., W. Härdle, and E. Mammen (2007). "A Semiparametric Factor Model for Implied Volatility Surface Dynamics." *Journal of Financial Econometrics 5*, 189–218.

Fuller, W. (1976). *Introduction to Statistical Time Series*. New York: John Wiley & Sons.

Granger, C. W. J. (1969). "Investigating Causal Relations by Econometric Models and Cross-Spectral Methods." *Econometrica 37*, 424–438.

Hördahl, P., O. Tristani, and D. Vestin (2006). "A Joint Econometric Model of Macroeconomic and Term Structure Dynamics." *Journal of Econometrics 131*, 405–444.

Lütkepohl, H. (2004). "Vector Autoregressive And Vector Error Correction Models." In H. Lütkepohl and M. Krätzig (eds.), *Applied Time Series Econometrics*. Cambridge, UK: Cambridge University Press.

Lütkepohl, H. (2005). *New Introduction to Multiple Time Series Analysis*. Berlin: Springer-Verlag.

Lütkepohl, H., and Krätzig, M. (eds.) (2004). *Applied Time Series Econometrics*. Cambridge, UK: Cambridge University Press.

Pesaran, M. H., and Y. Shin (1998). "Generalized Impulse Response Analysis in Linear Multivariate Models." *Economics Letters 58*, 17–29.

Skiadopoulos, G., S. Hodges, and L. Clewlow (1999). "The Dynamics of S&P 500 Implied Volatility Surface." *Review of Derivatives Research 3*, 263–282.

Veiga, B., and M. McAleer (2004). "Multivariate volatility and spillover effects in financial markets." In C. Pahl-Wostl, S. Schmidt, A. E. Rizzoli and A. J. Jakeman (eds.), *Complexity and Integrated Resources Management, Transactions of the 2nd Biennial Meeting of the International Environmental Modelling and Software Society*. Volume 3, 1442–1448. iEMSs: Manno, Switzerland.

Wu, T. (2006). "Macro Factors and the Affine Term Structure of Interest Rates." *Journal of Money, Credit, and Banking 38*, 1847–1875.

Yang, L., B. U. Park, L. Xue, and W. Härdle (2006). "Estimation and testing for varying coefficient in additive models with marginal integration." *Journal of the American Statistical Association 101*, 1212–1227.

Nonparametric Risk Management With Generalized Hyperbolic Distributions

Ying CHEN, Wolfgang HÄRDLE, and Seok-Oh JEONG

In this article we propose the generalized hyperbolic adaptive volatility (GHADA) risk management model based on the generalized hyperbolic (GH) distribution and on a nonparametric adaptive methodology. Compared with the normal distribution, the GH distribution has semiheavy tails and represents the financial risk factors more appropriately. Nonparametric adaptive methodology has the desirable property of being able to estimate homogeneous volatility over a short time interval and reflects a sudden change in the volatility process. For the German mark/U.S. dollar exchange rate and German bank portfolio data, the proposed GHADA model provides more accurate Value at risk calculations than the models with assumptions of the normal and t distributions.

KEY WORDS: Adaptive volatility estimation; Generalized hyperbolic distribution; Risk management; Value at risk.

1. INTRODUCTION

After the breakdown of the fixed exchange rate system of the Bretton Woods Agreement in 1971, a sudden increase in volatility in financial markets was observed. The subsequent boom in financial derivatives accelerated the turbulence of the markets. The subsequent scale of losses astonished the world and pushed the development of sound risk management systems.

One of the most challenging tasks in analyzing financial markets is measuring and managing risks properly. Financial risks have many sources and are typically mapped into a stochastic framework,

$$R_t = \sigma_t \varepsilon_t, \tag{1}$$

where R_t denotes the loss or the (log) negative return of financial instrument, that is, $\log(S_{t-1}) - \log(S_t)$, where S_t is the price of the financial instrument, σ_t is the time-dependent volatility, and ε_t is the white noise. Based on the distribution of the risk factor R_t, various kinds of risk measures, such as value at risk (VaR), expected shortfall, and lower partial moments, can be calculated. Among these, VaR has become the standard measure of market risk since J. P. Morgan launched RiskMetrics with the distributional assumption of normality in 1994, making the analysis of VaR simple and standard (Jorion 2001).

The importance of VaR was reinforced after it was used by central banks to govern and supervise the capital adequacy of banks in the Group of Ten (G10) countries in 1995. For a given financial instrument, VaR indicates the possible loss at a certain risk level over a certain time horizon. Andersen, Bollerslev, Christoffersen, and Diebold (2006) have pointed out that losses will converge to normality under temporal aggregation. This observation suggests that the principle of the RiskMetrics method is valid when a longer time horizon, such as 2 weeks or 1 month, is considered. On the other hand, financial institutions are also expected to report and control their daily VaRs that deviate from the normal assumption.

In this article we concentrate on proposing a risk management model to improve the calculation of daily VaR. Based on (1), daily VaR at a risk level p is defined mathematically as

$$VaR_{p,t} = F_t^{-1}(p) = \sigma_t q_{\varepsilon_t}(p), \tag{2}$$

where F_t^{-1} is the quantile function of R_t at time t, which is equal to the product of the volatility and the pth quantile of the stochastic term ε_t (Franke, Härdle, and Hafner 2004). It is clear that the accuracy of VaR and other risk measures depends heavily on the distributional assumption of the stochastic term and the volatility estimation. This observation motivates us to choose a heavy-tailed distribution family to mimic the empirical distribution of risk factors and an adaptive methodology to estimate and forecast volatility locally.

In the literature, for reasons of stochastic and numerical simplicity, it is often assumed that the involved risk factors are normally distributed, for example, in the RiskMetrics framework. However, this assumption contradicts an empirical fact observed in the market: Daily financial time series are heavytailed distributed. As more extreme risks appeared in the market, VaRs with higher risk quantiles, such as 99%, drew the attention of risk analysts, and discrepancies between these VaRs with normal distributional assumption and the actual VaRs became evident, although with the normality assumption, the VaR at 95% confidence level is almost identical to that with a more realistic leptokurtic distribution (Jaschke and Jiang 2002).

Figure 1 illustrates this empirical fact on the basis of the daily foreign exchange (FX) rates of the German mark to the U.S. dollar (DEM/USD) from December 1, 1979 to April 1, 1994. Here we use the daily devolatilized returns (i.e., $\varepsilon_t = R_t/\hat{\sigma}_t$) to fit the previously assumed stochastic distribution. Compared with R_t, the devolatilized returns eliminate the influence of volatility clustering and are more stationary. We discuss the technique used to estimate volatility, σ_t, later. The nonparametrically estimated kernel density and its log scale are considered benchmarks, with the quartic kernel function and Silverman's rule of thumb applied to determine the bandwidth h (Härdle, Müller, Sperlich, and Werwatz 2004). In the figure the estimated normal densities with the empirical mean and variance of the devolatilized returns obviously deviate from these benchmarks, which will lead to inaccurate VaR calculations.

Ying Chen is Assistant Professor, Department of Statistics and Applied Probability, National University of Singapore, Singapore, 117543 (E-mail: *stacheny@nus.edu.sg*). Wolfgang Härdle is Professor, Center for Applied Statistics and Economics, Humboldt University of Berlin, 10178 Berlin, Germany (E-mail: *haerdle@wiwi.hu-berlin.de*). Seok-Oh Jeong is Assistant Professor, Department of Information Statistics, Hankuk University of Foreign Studies, Mohyun, Yongin 449-791, Korea (E-mail: *seokohj@hufs.ac.kr*). This research was supported by the Deutsche Forschungsgemeinschaft through projects SFB 649, "Economic Risk," and SFB 373, "Simulation and Quantification of Economic Processes," at Humboldt-Universität zu Berlin. Special thanks are due to Dr. Ernst Eberlein for his kind contribution to the proof of Lemma 1. The authors also thank a referee for detailed comments.
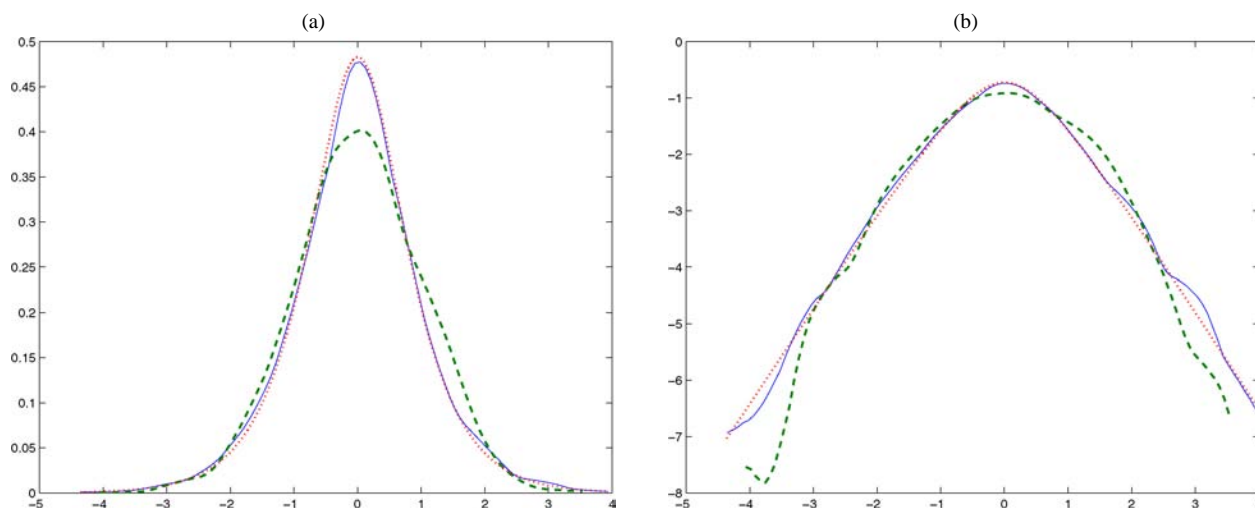
Figure 1. Density estimation of the daily DEM/USD standardized returns from December 1, 1979 to April 1, 1994 (3,719 observations) (a), along with the log scale of the estimation (b). [— nonparametric kernel; - - Gaussian; ····· GH(1, 1.74, −.02, .78, .01).] The nonparametric kernel density estimate is considered a benchmark. The bandwidth is $h' \approx .54$. Data source: FEDC (*http://sfb649.wiwi.hu-berlin.de*).

Due to the weak ability of normal distributions to capture this empirical distributional feature of financial risk factors, various heavy-tailed distribution families, such as the hyperbolic and Student t distributions and the Lévy process, have been introduced in finance by Eberlein and Keller (1995), Embrechts, McNeil, and Straumann (1999), and Barndorff-Nielsen and Shephard (2001). Among these heavy-tailed distributions, the generalized hyperbolic (GH) distribution family has attracted much attention from researchers. With five parameters, GH can match the distributional behavior of real data in a flexible way. Eberlein, Kallsen, and Kristen (2003) have proposed a model using the GH distribution that gives more accurate VaR values than the model with the normal distribution.

In addition to the stochastic distributional assumption, the role of volatility models also is of great significance in the VaR calculation. According to the definition of VaR (2), low accuracy of the volatility estimate will lead to low accuracy of the VaR estimate. The most frequently used estimations are the autoregressive conditional heteroscedasticity (ARCH) model (Engle 1982), generalized ARCH (GARCH) model (Bollerslev 1995), and stochastic volatility model (Harvey, Ruiz, and Shephard 1995). Although these models reflect the volatility clustering of financial time series, they are not flexible enough to react to possible structure shifts of volatility process. Table 1 presents an example in which the log-returns of the exchange rates Euro to the Japanese yen (EUR/JPY) before and after January 1, 2002 are modeled in the GARCH(1, 1) setting. Remember that on this date the Euro banknotes and coins were put into circulation. The maximum likelihood estimator (MLE) of one

involved parameter β is, as expected, distinct for the two subsets, changing from .75 before the currency transition to .95 after the transition. The standard deviation of the estimates are given in parentheses. (Refer to Bollerslev and Woolridge 1992 for maximum likelihood (ML) estimation of the parameters.) Therefore, using a time-constant closed form in estimation is questionable, especially in long time periods. Flexible estimation methods by providing a data-driven "local" model can help avoid this potential misspecification problem. We use the local constant model proposed by Mercurio and Spokoiny (2004) because of its good performance. The philosophy of this local constant model, in which volatility changes little over a short interval, is numerically tractable and economically meaningful. Moreover, compared with most adaptive estimation methods, such as the rectangular moving average, the local constant model can react very quickly once a sudden jump occurs. But an inconvenient assumption of Mercurio and Spokoiny (2004) is that the stochastic term is normally distributed.

Motivated by the aforementioned two lines of research, we estimate the local volatility adaptively and model the risk factors with heavy tails using the GH distribution. We call this new VaR technique the *generalized hyperbolic adaptive volatility* (GHADA) technique. The devolatilized return density plot in Figure 1 is in fact calculated using the GHADA technique. In addition, we check the validation of the GHADA technique by comparing it with the following risk management models:

- Volatility estimation with the GARCH(1, 1) model and distributional fit with the normal (NGARCH), Student t (tGARCH), or GH (GHGARCH) distribution
- Local constant volatility estimation and distributional fit with the normal (NADA) or Student t (tADA) distribution.

The article is organized as follows. In Section 2 we introduce the details of the GHADA technique. We illustrate the validation of the GHADA model is through Monte Carlo simulation in Section 3. In Section 4 we present VaR calculations based on the DEM/USD and German bank portfolio data. According to backtesting results, the GHADA technique provides

Table 1. ML estimates of the GARCH(1, 1) model based on the EUR/JPY rates

Time period	ω	α	β
1997/01/02–2001/12/31	9.07e−06(2.91e−06)	.13(.03)	.75(.05)
2002/01/02–2006/01/05	5.05e−07(3.70e−07)	.02(.01)	.95(.02)

NOTE: Standard deviations of the estimates are reported in parentheses.

more accurate forecasts than the models with assumptions of the normal and t distributions. Furthermore, in extreme events, the GHADA technique performs better than the models with GARCH(1, 1) volatility processes. Finally, we briefly conclude our study in Section 5.

2. THE GENERALIZED HYPERBOLIC ADAPTIVE VOLATILITY TECHNIQUE AND OTHER VARIATIONS

In risk management modeling, a major task is to estimate the future loss distribution accurately. As illustrated in (2), a realistic distributional assumption of the stochastic term ε_t and an accurate volatility estimation play important roles in modeling. In this section we describe two pillars of the proposed GHADA technique: the GH distribution and local constant volatility estimation.

2.1 The Generalized Hyperbolic Adaptive Volatility Technique

2.1.1 Generalized Hyperbolic Distribution. The GH distribution introduced by Barndorff-Nielsen (1977) is a heavy-tailed distribution that can closely replicate the empirical distribution of financial risk factors. The density of the GH distribution for $x \in \mathbb{R}$ is

$$f_{GH}(x; \lambda, \alpha, \beta, \delta, \mu)$$
$$= \frac{(\iota/\delta)^\lambda}{\sqrt{2\pi} K_\lambda(\delta\iota)} \frac{K_{\lambda-1/2}\{\alpha\sqrt{\delta^2 + (x-\mu)^2}\}}{\{\sqrt{\delta^2 + (x-\mu)^2}/\alpha\}^{1/2-\lambda}} \cdot e^{\beta(x-\mu)} \quad (3)$$

under the conditions

$$\delta \geq 0, \qquad |\beta| < \alpha \quad \text{if } \lambda > 0,$$
$$\delta > 0, \qquad |\beta| < \alpha \quad \text{if } \lambda = 0,$$

and

$$\delta > 0, \qquad |\beta| \leq \alpha \quad \text{if } \lambda < 0,$$

where λ, α, β, δ, and $\mu \in \mathbb{R}$ are the GH parameters with $\iota^2 = \alpha^2 - \beta^2$. The density's location and scale are mainly controlled by μ and δ,

$$\mathsf{E}[X] = \mu + \frac{\delta^2\beta}{\delta\iota} \frac{K_{\lambda+1}(\delta\iota)}{K_\lambda(\delta\iota)}$$

and

$$\text{var}[X]$$
$$= \delta^2 \left\{ \frac{K_{\lambda+1}(\delta\iota)}{\delta\iota K_\lambda(\delta\iota)} + \left(\frac{\beta}{\iota}\right)^2 \left[\frac{K_{\lambda+2}(\delta\iota)}{K_\lambda(\delta\iota)} - \left\{ \frac{K_{\lambda+1}(\delta\iota)}{K_\lambda(\delta\iota)} \right\}^2 \right] \right\},$$

whereas β and α play roles in the skewness and kurtosis of the distribution. (For more details of the parameters' domains, see Bibby and Sørensen 2001.) The $K_\lambda(\cdot)$ is the modified Bessel function of the third kind with index λ (Barndorff-Nielsen and Blæsild 1981),

$$K_\lambda(x) = \frac{1}{2} \int_0^\infty y^{\lambda-1} \exp\left\{ -\frac{x}{2}(y + y^{-1}) \right\} dy.$$

Furthermore, the GH distribution has the following tail behavior:

$$f_{GH}(x; \lambda, \alpha, \beta, \delta, \mu = 0) \sim x^{\lambda-1} e^{(\mp\alpha+\beta)x} \quad \text{as } x \to \pm\infty,$$
$$(4)$$

where $a(x) \sim b(x)$ as $x \to \infty$, meaning that both $a(x)/b(x)$ and $b(x)/a(x)$ are bounded as $x \to \infty$. Recall that the negative return, $R_t = \log(S_{t-1}) - \log(S_t)$, is considered a risk factor in risk management, indicating that the right tail of losses with large values is most important. On the other hand, the left tail concerns the values of profits that are less interesting in risk management. Compared with three popular heavy-tailed distributions—the Student t, Laplace, and Cauchy distributions—the GH distribution has a faster speed of exponential decay, which better matches the empirical tail behavior of real data. The four heavy-tailed and normal distributions, as well as their tail behaviors, are compared in Figure 2.

To retain the comparability of these distributions, we normalize the variables to have means equal to 0 and variances equal to 1. Figure 2(a) shows the complete shapes of these distributions. Among these, the Cauchy distribution has the lowest peak and the fattest tails; in brief, it has the flattest distribution. The GH distribution decays the second fastest in the tails but has the highest peak, which is similar to the distributional shape of real data (see, e.g., Fig. 1).

The moment-generating function of the GH distribution is

$$m_f(z) = e^{\mu z} \cdot \frac{\iota^\lambda}{\iota_z^\lambda} \cdot \frac{K_\lambda(\delta\iota_z)}{K_\lambda(\delta\iota)},$$
$$|\beta + z| < \alpha, \iota_z^2 = \alpha^2 - (\beta + z)^2, \quad (5)$$

indicating that m_f is differentiable infinitely many times near 0. As a result, every moment of a GH variable exists. In Section 2.1.2 this feature as well as the tail behavior seen in (4) of the GH distribution help extend the local constant volatility methodology from the normal distribution to the GH distribution (see the App.).

Given the distributional form of the GH distribution in (3), ML estimation can be applied straightforwardly. But such a direct estimation based on the GH distribution is rarely applied because it is computationally cumbersome and numerically unstable due to the estimation of λ in the modified Bessel function. Instead, subclasses of the GH distribution, such as the hyperbolic (HYP) and normal-inverse Gaussian (NIG) distributions, are frequently used. These subclasses fix the value of λ to avoid the numerical problem. Eberlein and Keller (1995) and Barndorff-Nielsen (1997) have shown that these subclasses are rich enough to model financial time series in an efficient way. In addition, the popularity of the subclasses is also motivated by the observation that the four parameters $(\mu, \delta, \beta, \alpha)^\top$ can simultaneously control the four moment functions of the distribution, that is, the trend, scale, asymmetry, and likeliness of extreme events. In our study we concentrate on the two subclasses of the GH distribution: HYP with $\lambda = 1$ and NIG distribution with $\lambda = -1/2$. In the simulation and empirical studies that follow, one subclass performs better than the other depending on the situation. The corresponding density functions are as follows:

- HYP distribution: $\lambda = 1$,

$$f_{HYP}(x; \alpha, \beta, \delta, \mu)$$
$$= \frac{\iota}{2\alpha\delta K_1(\delta\iota)} e^{\{-\alpha\sqrt{\delta^2+(x-\mu)^2}+\beta(x-\mu)\}}, \quad (6)$$

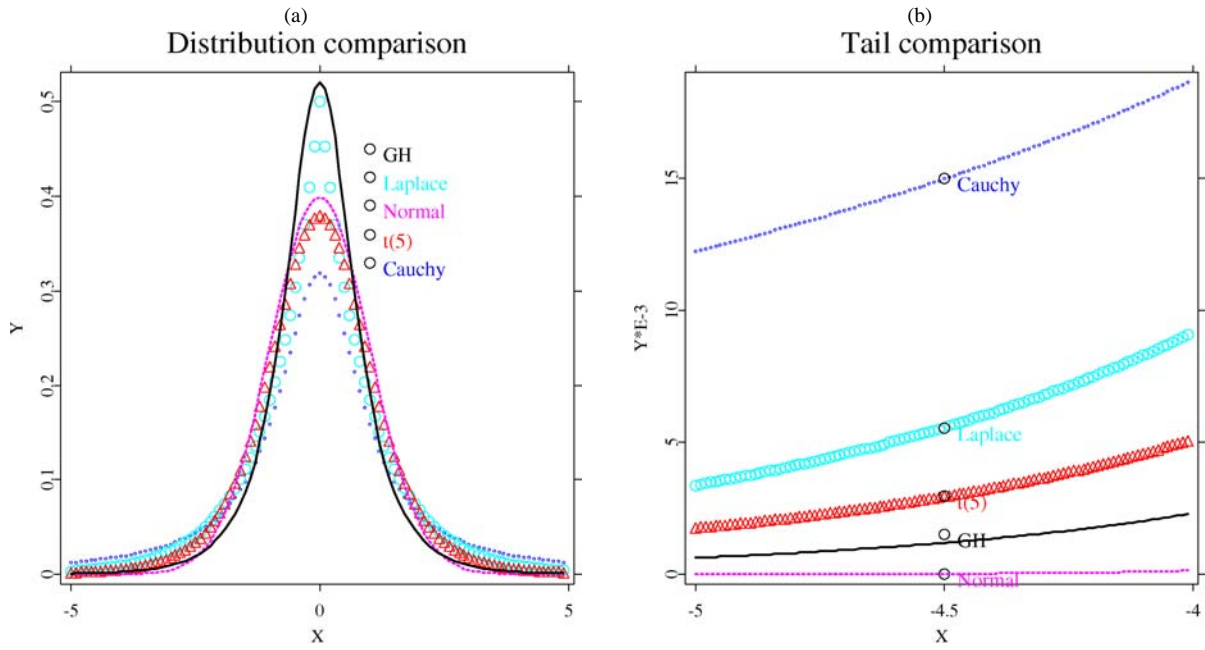where $x, \mu \in \mathbb{R}$, $0 \leq \delta$, and $|\beta| < \alpha$,

Figure 2. Graphical comparison of the tail behavior of five standardized distributions: GH distribution (——), standard normal distribution (- - - - -), Student t distribution with 5 degrees of freedom ($\triangle\triangle\triangle$), Laplace distribution ($\circ \circ \circ$), and Cauchy distribution ($\cdots\cdots$).

- NIG distribution: $\lambda = -1/2$,

$$f_{NIG}(x; \alpha, \beta, \delta, \mu)$$

$$= \frac{\alpha\delta}{\pi} \frac{K_1\{\alpha\sqrt{\delta^2 + (x-\mu)^2}\}}{\sqrt{\delta^2 + (x-\mu)^2}} e^{\{\delta\iota + \beta(x-\mu)\}}, \quad (7)$$

where $x, \mu \in \mathbb{R}$, $\delta > 0$, and $|\beta| \leq \alpha$. Note that the estimated GH density in Figure 2 is actually based on the NIG distributional assumption.

To estimate the unknown parameters $(\alpha, \beta, \delta, \mu)^\top$, ML and numerical optimization methods, such as the Powell method (Press, Teukolsky, Vetterling, and Flannery 1992), are used. For an independently and identically distributed (iid) HYP and NIG variable X, the respective log-likelihood functions are

$$L_{HYP} = T\log\iota - T\log 2 - T\log\alpha - T\log\delta - T\log K_1(\delta\iota)$$

$$+ \sum_{t=1}^{T}\left\{-\alpha\sqrt{\delta^2 + (x_t - \mu)^2} + \beta(x_t - \mu)\right\} \quad (8)$$

and

$$L_{NIG} = T\log\alpha + T\log\delta - T\log\pi + T\delta\iota$$

$$+ \sum_{t=1}^{T}\left[\log K_1\{\alpha\sqrt{\delta^2 + (x_t - \mu)^2}\}\right.$$

$$\left. - \frac{1}{2}\log\{\delta^2 + (x_t - \mu)^2\} + \beta(x_t - \mu)\right]. \quad (9)$$

Figure 1 displays the estimated HYP density with the MLEs $\hat{\alpha} = 1.744$, $\hat{\beta} = -.017$, $\hat{\delta} = .782$, and $\hat{\mu} = .012$ of the de-volatilized DEM/USD returns. The estimated density graphically coincides with the empirical density, namely the nonparametric kernel density estimation. The estimated NIG density has similar behavior to the empirical density, which is eliminated here.

2.1.2 Adaptive Volatility Estimation. Here we describe the adaptive estimation procedure for the volatility coefficients when risk factors follow the HYP and NIG subclasses of GH distributions. Originally, the concept of adaptive volatility estimation was proposed by Mercurio and Spokoiny (2004), which assumes that there exists an interval of local homogeneity of the volatility process σ_t, which means that for a fixed time point τ there exists a time interval $I = [\tau - m, \tau)$ such that σ_t varies little over I. Once an interval of homogeneity I is specified, the volatility at time τ is simply estimated by averaging the squared returns over the time interval I,

$$\hat{\sigma}_\tau^2 = \frac{1}{|I|}\sum_{t \in I} R_t^2, \quad (10)$$

where $|I|$ denotes the cardinality of I.

The squared returns R_t^2 are always nonnegative and have a skewed distribution with the stochastic errors ε_t. Therefore, the problem of estimating σ_t is transformed into an additive regression problem by a power transformation,

$$|R_t|^\gamma = C_\gamma \sigma_t^\gamma + D_\gamma \sigma_t^\gamma \zeta_{\gamma, t}$$

$$= \theta_t + s_\gamma \theta_t \zeta_{\gamma, t}, \quad (11)$$

with γ as the power transformation parameter. Theoretically, γ is a constant contained in [0, 1] (see the App.). Furthermore, $\zeta_{\gamma, t} = (|\varepsilon_t|^\gamma - C_\gamma)/D_\gamma$, $C_\gamma = \mathsf{E}(|\varepsilon_t|^\gamma|\mathcal{F}_{t-1})$, $D_\gamma^2 = \mathsf{E}[(|\varepsilon_t|^\gamma - C_\gamma)^2|\mathcal{F}_{t-1}]$, and $s_\gamma = D_\gamma/C_\gamma$. Equation (11) can be considered a regression model to estimate θ_t with heteroscedastic additive errors $s_\gamma \theta_t \zeta_{\gamma, t}$. Because ε_t is assumed to be iid, C_γ, D_γ, and s_γ are merely nonstochastic constants. For example, when ε_t is N(0, 1)-distributed, we can easily compute the exact values of C_γ, D_γ, and hence s_γ. However, the calculations of these constants do not necessarily specify an interval of homogeneity, as illustrated in what follows.

Note that σ_t has a one-to-one correspondence with θ_t. Therefore, if I is an interval of homogeneity, then θ_t is nearly constant for $t \in I$ and can be estimated by

$$\hat{\theta}_I = \frac{1}{|I|} \sum_{t \in I} |R_t|^\gamma. \qquad (12)$$

By (11), we have that

$$\hat{\theta}_I = \frac{1}{|I|} \sum_{t \in I} \theta_t + \frac{s_\gamma}{|I|} \sum_{t \in I} \theta_t \zeta_{\gamma,t}.$$

The conditional expectation and variance of $\hat{\theta}_I$ are

$$\mathsf{E}[\hat{\theta}_I | \mathcal{F}_{\tau-1}] = \mathsf{E} \frac{1}{|I|} \sum_{t \in I} \theta_t,$$

and

$$v_I^2 = \mathrm{var}[\hat{\theta}_I | \mathcal{F}_{\tau-1}] = \frac{s_\gamma^2}{|I|^2} \mathsf{E}\left(\sum_{t \in I} \theta_t \zeta_{\gamma,t}\right)^2 = \frac{s_\gamma^2}{|I|^2} \mathsf{E} \sum_{t \in I} \theta_t^2.$$

When I is an interval of homogeneity, v_I can be estimated by

$$\hat{v}_I = s_\gamma \hat{\theta}_I |I|^{-1/2}.$$

For testing whether an interval I is homogeneous, Mercurio and Spokoiny (2004) suggested investigating the homogeneity of θ_t in I instead of σ_t in I. They proposed a homogeneity test on θ_t based on a Martingale deviation probability bound (Mercurio and Spokoiny 2004, sec. 3), with $\varepsilon_t \sim N(0, 1)$. Thus we make a theoretical justification on ε_t from the GH distribution before we adopt their procedure for determining the interval of homogeneity. Details are given in the Appendix. We address this issue precisely in the following Theorem 1, the proof of which was given by Mercurio and Spokoiny (2004).

Theorem 1. If the volatility coefficient σ_t satisfies the condition $b \le \sigma_t^\gamma \le bB$ with some positive constants b and B, then there exists $a_\gamma > 0$ such that for every $\eta > 1$,

$$P\{|\hat{\theta}_I - \theta_\tau| > \Delta_I + \eta \hat{v}_I\} \le 4\sqrt{e} a_\gamma^{-1} \eta (1 + \log B) \exp\left\{-\frac{\eta^2}{2a_\gamma}\right\},$$

where Δ_I is the squared bias defined as $\Delta_I^2 = |I|^{-1} \sum_{t \in I} (\theta_t - \theta_\tau)^2$.

This theorem indicates that if I is a time-homogeneous interval, then the squared bias Δ_I is negligible and the estimation error $|\hat{\theta}_I - \theta_\tau|$ is small relative to $\eta \hat{v}_I$ for $\tau \in I$ with high probability. Moreover, it holds that for any subinterval J of a homogeneity interval I,

$$|\hat{\theta}_{I \setminus J} - \hat{\theta}_J| \le \eta (\hat{v}_J + \hat{v}_{I \setminus J})$$
$$= \eta'(\hat{\theta}_J |J|^{-1/2} + \hat{\theta}_{I \setminus J} |I \setminus J|^{-1/2}) \qquad (13)$$

with high probability when $\eta' = \eta s_\gamma$ is large enough. Therefore, if there exists a subinterval $J \subset I$ that makes $|\hat{\theta}_{I \setminus J} - \hat{\theta}_J|$ have significantly large positive value, then the homogeneity of the interval I should be denied.

Now we are ready to describe the precise procedure for estimating the interval of homogeneity with a constant η'. We start with a small interval, $I = [\tau - m_0, \tau)$, that satisfies the homogeneity and then suitably choose a step-increasing parameter m_0 and an integer $k (\ge 1)$. The choice of the step-increasing

parameter will influence the sensitivity of the estimation to a change point. A smaller value increases the sensitivity but slows the estimation speed, as illustrated in the simulation study in Section 3. For convenience, m_0 is suggested to be a multiple of 5. The procedure is as follows:

Step 1. Increase k to $k + 1$, and enlarge the interval I to $[\tau - m, \tau)$ with $m = k \times m_0$.

Step 2. Reject I if there exists $J(\ell) = [\tau - \frac{2m}{3} + \ell, \tau)$, $\ell = 1, 2, \ldots, \frac{m}{3}$ such that

$$|\hat{\theta}_{I \setminus J(\ell)} - \hat{\theta}_{J(\ell)}|$$
$$> \eta'(\hat{\theta}_{J(\ell)} |J(\ell)|^{-1/2} + \hat{\theta}_{I \setminus J(\ell)} |I \setminus J(\ell)|^{-1/2}). \quad (14)$$

Step 3. If I is rejected, then set $I = [\tau - m_0, \tau)$ as the interval of homogeneity, and the procedure is complete. If I is not rejected, then set $m_0 = m$ and go to Step 1.

There are still two parameters left to be specified: γ in the power transformation and the thresholding parameter η' in (13). According to Lemma A.1 in the Appendix, the parameter γ is contained in [0, 1]. In our study, we choose $\gamma = .5$, as did Mercurio and Spokoiny (2004). In fact, the choice of γ has little effect on the procedure for estimating the interval of homogeneity. In contrast, the choice of the threshold parameter η' is critical for the homogeneity test. We pursue a selection procedure similar to the choice of a smoothing parameter in nonparametric estimations. Take t_0 such that there are sufficient past observations to estimate $\hat{\theta}_{(t_0, \eta')}$ properly. Then select η' that gives minimal forecast error,

$$\eta' = \mathrm{argmin} \sum_{t=t_0}^{\tau-1} \{|R_t|^\gamma - \hat{\theta}_{(t, \eta')}\}^2, \qquad (15)$$

where $\hat{\theta}_{(t, \eta')}$ is the estimate (12) of θ_t from observations $R_1, \ldots, R_{t-1}$ with the threshold parameter η'. Note that such a selection is distribution-free and avoids the calculation of these constants based on the GH distribution in (11).

2.2 Some Risk Management Models

The research on VaR models has been ignited and prompted by the introduction of the rule of Basel Committee on Banking Supervision in 1995. Financial institutions may use their internal VaR models to calculate VaRs. Recall that VaR is formulated as $VaR_{p,t} = F_t^{-1}(p) = \sigma_t q_{\varepsilon_t}(p)$, with the notations as defined before. Table 2 lists eight models according to two factors: the estimation of volatility and the distributional assumption of stochastic term. The GARCH(1, 1) and local constant models are implemented with four distributional assumptions: the HYP, NIG, standard normal, and $t(df)$, where (df) represents the degrees of freedom. Among these models, the tGARCH, which combines the GARCH(1, 1) technique and the Student t distribution, is, to the best of our knowledge, the most frequently used model in practice. Compared with all of the other models mentioned here, the GHADA (HYPADA and NIGADA) models are expected to perform superiorly due to their desirable statistical characteristics, as discussed earlier. In the simulation and empirical studies, we compare the accuracy of these various combinations and check the expectations.

Table 2. Risk management models based on the heteroscedastic model, $R_t = \sigma_t \varepsilon_t$

Model	Volatility estimation	Distributional assumption
HYPGARCH	GARCH(1, 1)	HYP($\alpha, \beta, \delta, \mu$)
NIGGARCH	GARCH(1, 1)	NIG($\alpha, \beta, \delta, \mu$)
NGARCH	GARCH(1, 1)	N(0, 1)
tGARCH	GARCH(1, 1)	$t(df)$
HYPADA	Local constant	HYP($\alpha, \beta, \delta, \mu$)
NIGADA	Local constant	NIG($\alpha, \beta, \delta, \mu$)
NADA	Local constant	N(0, 1)
tADA	Local constant	$t(df)$

3. MONTE CARLO SIMULATION

As discussed earlier, good risk management modeling relies on two factors: estimating the volatility and fitting the distribution of risk factors. The previous distributional estimation based on the DEM/USD data, together with the comparison to the nonparametric density estimation in Section 2.1.1, provide evidence that with four parameters, the HYP and NIG distributions can represent the empirical distribution of the stochastic term very well. In this section we illustrate the reliability of the local constant volatility model with four different distributional assumptions. Two simple volatility processes with jumps and a GARCH(1, 1) process are considered here:

$$\sigma_{1,t} = \begin{cases} |.02t - 5|/100, & 1 \le t \le 300 \\ |.02t - 20|/100, & 300 < t \le 600 \\ |.12t - 30|/100, & 600 < t \le 1,000; \end{cases} \quad (16)$$

$$\sigma_{2,t} = \begin{cases} .01, & 1 \le t \le 400 \\ .03, & 400 < t \le 750 \\ .015, & 750 < t \le 1,000; \end{cases} \quad (17)$$

and

$$\sigma_{3,t} = 1.65e - 06 + .07\varepsilon_{t-1}^2 + .89\sigma_{t-1}^2, \quad (18)$$

where the parameters of the GARCH(1, 1) process (18) are the estimates of the DEM/USD negative returns from December 1, 1979 to April 1, 1994.

In each scenario we generate $n = 1,000$ observations with the HYP(2, 0, 1, 0), NIG(2, 0, 1, 0), N(0, 1), and $t(6)$ distributions. We generate the risk factors are based on the heteroscedastic model

$$R_{ij,t} = \sigma_{it}\varepsilon_{jt}, \qquad i = 1, 2, 3, \text{ and } j = \text{HYP, NIG, N}, t.$$

In the local constant (LC) model, the first 200 observations of $R_{ij,t}$ are considered a training set. The transformation parameter γ is fixed at .5, and a global η' that minimizes the mean of forecast error for $t \in [201, 1,000]$ is selected to perform the homogeneity test. Last but not least, two values of the step-increasing parameter m_0 in the homogeneous interval are used, because the value of m_0 will influence the detection speed of the LC model when a jump appears: the recommended value, $m_0 = 5$, and a more sensitive value, $m_0 = 2$. Note that in general, a smaller value of m_0 can find jumps faster than a larger one. All scenarios are repeated $M = 200$ times.

Figure 3 displays three examples of the estimated volatility series of σ_{1t} [Fig. 3(a)], σ_{2t} [Fig. 3(b)], and σ_{3t} [Fig. 3(c)] with the HYP or NIG distributional assumption, where the LC (solid
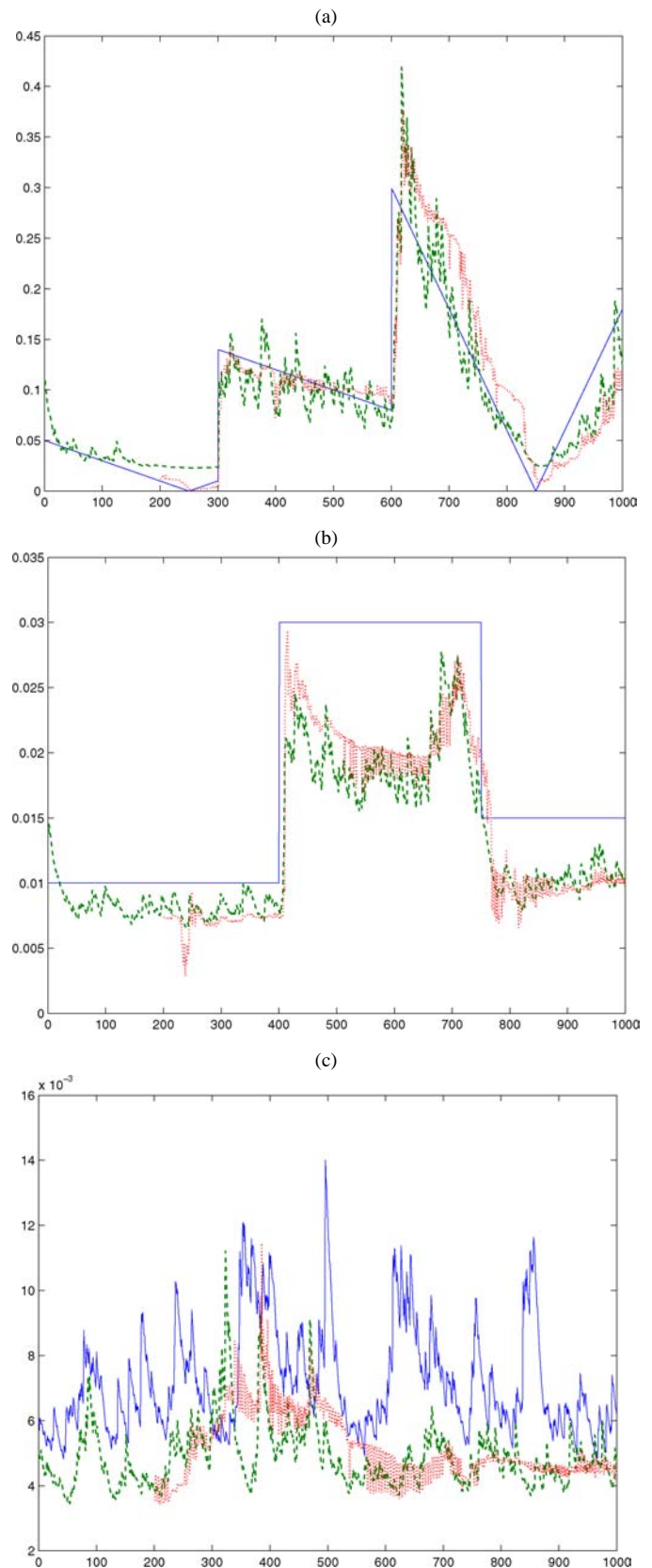

Figure 3. One realized estimation (—) based on the simulated data: (a) the HYP variables for σ_{1t}, (b) the NIG for σ_{2t}, and (c) the NIG for σ_{3t}. The involved parameters are $\gamma = .5$, $m_0 = 5$, and the starting point $t_0 = 201$. The volatility processes are estimated using the GARCH(1, 1) model (- -) and the local constant (LC) model (·····).

Table 3. Descriptive statistics of the RMAE and RMSE with respect to volatility estimations

	RMAE				RMSE			
	Mean	SD	Maximum	Minimum	Mean	SD	Maximum	Minimum
$\sigma_{1t}\varepsilon_t$								
HYP	.77	1.05	1.04	.06	.78	1.23	1.08	0
NIG	.87	1.00	1.04	.06	.83	1.15	1.07	0
N	.76	1.18	1.04	.05	.87	1.22	1.09	0
t	.93	1.28	1.06	.05	.95	1.14	1.11	0
$\sigma_{2t}\varepsilon_t$								
HYP	1.31	.76	1.27	1.42	1.47	.89	1.39	1.73
NIG	1.25	.70	1.24	1.29	1.44	.85	1.43	1.55
N	.61	.57	.61	.62	.50	0	.33	.50
t	.69	.78	.70	.69	.51	.88	.60	.50
$\sigma_{3t}\varepsilon_t$								
HYP	1.21	1.07	1.16	1.51	1.31	1.20	1.29	1.89
NIG	1.07	1.03	1.12	1.53	1.11	1.15	1.24	2.25
N	1.32	1.19	1.21	1.25	1.58	1.31	1.40	1.40
t	1.22	1.39	1.23	1.50	1.49	1.67	1.38	2.12

NOTE: Two volatility models, local constant (LC) ($\gamma = .5$ and $m_0 = 5$) and GARCH(1, 1) models, are applied to estimate three volatility processes based on four variables: HYP(2, 0, 1, 0), NIG(2, 0, 1, 0), N(0, 1), and t(6).

line) and GARCH(1, 1) estimations (dotted line) are compared with the generated volatilities (circles). In the first two estimations, the LC and GARCH(1, 1) models display comparable results, whereas the GARCH(1, 1) setup performs better in the estimation of σ_{3t}.

The quality of these two volatility estimation techniques is further measured by two ratios, the ratio of mean absolute error (RMAE) and the ratio of mean squared error (RMSE):

$$\text{RMAE} = \frac{\sum_{t=201}^{1,000} |\hat{\sigma}_{it}^{LC} - \sigma_{it}|}{\sum_{t=201}^{1,000} |\hat{\sigma}_{it}^{GARCH} - \sigma_{it}|}$$

and

$$\text{RMSE} = \frac{\sum_{t=201}^{1,000} (\hat{\sigma}_{it}^{LC} - \sigma_{it})^2}{\sum_{t=201}^{1,000} (\hat{\sigma}_{it}^{GARCH} - \sigma_{it})^2}, \qquad i = 1, 2, 3.$$

If the value of RMAE or RMSE is smaller than 1, this means that the LC model has a smaller estimation error on average than the GARCH and vice versa. Based on 200 repetitions, the mean, standard deviation (SD), maximum, and minimum of the two criteria are reported in Table 3. In the simulation of σ_{1t}, the LC model with $m_0 = 5$ gives more accurate volatility estimations than the GARCH(1, 1) technique on average. Concerning σ_{2t}, the GARCH model performs better than the LC when the stochastic terms are HYP and NIG distributed, but has lower accuracy when the ε_t's are normal and t distributed. As assumed, the GARCH model can better match the generated GARCH process σ_{3t}. Based on the simulation results, the LC model is comparable to the GARCH(1, 1) technique. However, the assumption of the GARCH technique, (i.e., the estimation form is time constant) is disputed as the nonstationary volatility processes σ_{1t} and σ_{2t} are given. In this sense, the LC model is considered better because it not only represents the empirical characteristics of volatility movements, but also provides a reasonable theory for the estimation.

Next, we compare the sensitivity of the LC and GARCH(1, 1) models to jumps in volatility. We introduce a percentage rule to study the sensitivity of the two volatility estimation techniques. The detection speed of the estimated volatility to a sudden jump is measured at a 40%, 50%, or 60% level of the jump size. The 40% rule, for example, refers to the number of time steps to reach 40% of the jump size. Table 4 provides examples of the detection steps. The GARCH(1, 1) process has a naturally fast reaction to jumps in a short interval, because it is actually an exponential smoothing process. In general, the LC model needs more time to detect a jump than GARCH, but the difference is very small. Sometimes the LC model reacts even faster than GARCH(1, 1) when adjusting the value of m_0. For example, concerning the jump of σ_{1t} at $t = 300$, the HYPADA needs 4.66 steps on average to detect 50% jump sizes and 6.10 steps to detect 60% jump sizes, whereas the GARCH(1, 1) requires 5.17 and 7.27 steps. In addition, the deviations of these two detections based on the LC method with values of 2.9 and 1.7 are smaller than those of the GARCH technique. Meanwhile, we find that the detection speed is slow for a deceased jump. For σ_{2t}, a downward jump from 5% to 1% occurs at $t = 750$. The LC model with $m_0 = 2$ needs 19.22 steps on average to detect 60% jump sizes. This number is three times greater than that of detection steps for an increased jump with 40% sizes at $t = 400$. This phenomenon results from a low test power in the homogeneity test (13), where the squared conditional variance v_I depends on θ_t and a larger value of θ_t will lead to a low test power.

4. EMPIRICAL STUDY

4.1 Data Set

We used two data sets, the DEM/USD exchange rate and a German bank portfolio, in the empirical analysis. These are available from the FEDC (*sfb649.wiwi.hu-berlin.de/fedc*). The loss of the exchange rate is calculated daily from December 1, 1979 to April 1, 1994. There are 3,719 observations. The bank portfolio data set reports the market value of the portfolio held

Table 4. Mean of the detection step for several sudden jumps based on the LC model with $m_0 = 2$ and $m_0 = 5$ and the GARCH(1, 1) model

Model	m_0	40% rule	50% rule	60% rule	40% rule	50% rule	60% rule
		$\sigma_{1,t=300}$			$\sigma_{1,t=600}$		
HYPADA	2	$3.72_{(1.7)}$	$4.66_{(2.9)}$	$6.10_{(4.7)}$	$5.01_{(3.2)}$	$6.81_{(4.7)}$	$8.82_{(6.5)}$
HYPADA	5	$4.47_{(2.2)}$	$5.85_{(3.3)}$	$7.87_{(4.8)}$	$5.52_{(3.0)}$	$7.28_{(3.9)}$	$9.66_{(5.6)}$
HYPGARCH		$3.63_{(2.3)}$	$5.17_{(3.5)}$	$7.27_{(5.3)}$	$3.42_{(2.6)}$	$5.28_{(3.8)}$	$7.69_{(5.5)}$
NIGADA	2	$4.35_{(2.5)}$	$5.64_{(4.3)}$	$8.88_{(8.3)}$	$5.81_{(3.6)}$	$8.09_{(6.2)}$	$13.22_{(13.4)}$
NIGADA	5	$5.92_{(2.9)}$	$7.94_{(4.5)}$	$14.69_{(20.5)}$	$9.25_{(4.2)}$	$11.98_{(7.2)}$	$17.89_{(13.1)}$
NIGGARCH		$4.73_{(3.3)}$	$6.74_{(4.9)}$	$10.90_{(8.1)}$	$4.19_{(2.9)}$	$6.95_{(5.2)}$	$10.76_{(9.7)}$
NADA	2	$3.05_{(1.6)}$	$3.62_{(1.9)}$	$4.23_{(2.4)}$	$3.77_{(2.1)}$	$5.00_{(2.7)}$	$6.14_{(3.5)}$
NADA	5	$4.29_{(1.8)}$	$5.17_{(2.0)}$	$6.19_{(2.4)}$	$7.24_{(3.6)}$	$9.49_{(4.1)}$	$11.19_{(5.4)}$
NGARCH		$2.56_{(1.7)}$	$3.27_{(2.2)}$	$4.27_{(2.9)}$	$2.27_{(1.5)}$	$3.11_{(2.0)}$	$4.00_{(2.5)}$
tADA	2	$2.86_{(1.4)}$	$3.43_{(1.8)}$	$3.82_{(2.0)}$	$1.97_{(1.9)}$	$3.76_{(2.5)}$	$4.83_{(2.9)}$
tADA	5	$3.78_{(1.7)}$	$4.50_{(1.9)}$	$5.14_{(2.2)}$	$1.39_{(2.1)}$	$6.69_{(3.8)}$	$8.61_{(3.9)}$
tGARCH		$2.42_{(1.6)}$	$3.03_{(2.1)}$	$3.76_{(2.5)}$	$1.64_{(1.4)}$	$2.47_{(2.0)}$	$3.25_{(2.4)}$
		$\sigma_{2,t=400}$			$\sigma_{2,t=750}$		
HYPADA	2	$5.24_{(3.7)}$	$7.58_{(4.9)}$	$10.79_{(7.5)}$	$48.51_{(33.5)}$	$28.81_{(19.2)}$	$19.22_{(12.6)}$
HYPADA	5	$6.90_{(3.9)}$	$9.44_{(5.2)}$	$12.74_{(9.5)}$	$60.67_{(45.5)}$	$30.60_{(21.8)}$	$20.20_{(13.8)}$
HYPGARCH		$4.09_{(3.0)}$	$7.65_{(5.0)}$	$12.27_{(9.2)}$	$73.28_{(35.4)}$	$30.96_{(11.4)}$	$17.10_{(7.7)}$
NIGADA	2	$6.84_{(4.3)}$	$10.09_{(6.9)}$	$15.70_{(12.9)}$	$29.77_{(21.7)}$	$18.84_{(13.6)}$	$9.92_{(8.4)}$
NIGADA	5	$8.93_{(4.6)}$	$12.04_{(6.7)}$	$18.03_{(13.5)}$	$39.27_{(30.1)}$	$22.56_{(13.8)}$	$13.38_{(9.0)}$
NIGGARCH		$6.63_{(4.4)}$	$12.71_{(7.8)}$	$20.49_{(12.4)}$	$39.49_{(18.1)}$	$18.17_{(8.9)}$	$7.94_{(6.1)}$

NOTE: The standard deviations of the detection steps are in parentheses. Two jumps with respect to σ_{1t} at $t = 300$ and $t = 600$ and two jumps with respect to σ_{2t} at $t = 400$ and $t = 750$ are considered.

by a German bank (anonymous due to the privacy protection law in Germany). There are 5,603 daily observations.

The mean, standard deviation, skewness, kurtosis, and the first two autocorrelations ρ_1 and ρ_2 of these two data sets are listed in Table 5. All time series are centered around 0 and have leptokurtic distributions, as indicated by their kurtoses. The first two autocorrelations of the returns are, as expected, close to 0, validating the first-order process (1). Two processes of devolatilized returns, $\varepsilon_t^{LC} = R_t / \hat{\sigma}_t^{LC}$ and $\varepsilon_t^{GARCH} = R_t / \hat{\sigma}_t^{GARCH}$, are analyzed according to the two volatility estimation techniques. As discussed earlier, the devolatilized returns ε_t^{LC} and ε_t^{GARCH}, compared with the loss series R_t, are expected to be more stationary. Note that these devolatilized returns still have the heavy-tailed distributional property even after eliminating the influence of the time-varying volatility.

The local constant and GARCH(1, 1) volatilities of the time plots of the daily losses are displayed in Figure 4. The plots of volatility estimations show volatility clustering, which graphi-

Table 5. Descriptive statistics for the daily devolatilized residuals of the exchange rate data and bank portfolio data

Data	Mean	SD	Skewness	Kurtosis	ρ_1	ρ_2
Exchange rate: $t \in [501, 3{,}719]$						
R_t	8.30e–05	7.00e–03	.07	4.94	.02	.01
ε_t^{LC}	5.24e–03	.99	.01	4.03	.03	.02
ε_t^{GARCH}	7.13e–03	.99	.04	4.38	.03	.02
Bank portfolio: $t \in [501, 5{,}602]$						
R_t	−9.51e–05	1.59e–02	.28	8.08	−.04	−.03
ε_t^{LC}	1.13e–02	.96	−.08	5.18	−.04	−.02
ε_t^{GARCH}	−1.31e–02	.99	.08	7.38	−.03	−.02

cally reflects the movement pattern of risks. For example, the local constant volatility displays a jump at $t = 3{,}044$ when a large loss is observed in the time plot of the exchange rate series. In the German bank portfolio data, the simultaneous movements of loss and volatility are more evident. High risk (i.e., large value of volatility) is observed in the turbulent periods as $t \in (3{,}000, 4{,}000)$, and small volatility appears in the quiet periods. The GARCH(1, 1) technique gives comparable estimations. Compared with the German bank portfolio, the loss series of the DEM/USD displays a more regular fluctuation, because the exchange rate market is liquid. In the German bank data, long quiet periods are observed with two extremely turbulent periods, suggesting that on average, large homogeneous intervals will be specified in the German bank data. The boxplots in Figure 5 provide evidence of this suggestion. The means of homogeneous intervals with respect to the two data sets are 51.37 (DEM/USD) and 76.42 (German bank). Furthermore, many outliers with large value of interval length (circles or stars with 1.5 or 3 times the box length distance from the upper level) are observed in the German bank data.

Given the estimated volatilities, the devolatilized returns are calculated and used to estimate the distributional parameters of four different assumptions: HYP, NIG, standard normal, and $t(6)$. The HYP and NIG distributional parameters with respect to the local constant and GARCH(1, 1) volatilities are listed in Table 6. The density estimations under these four assumptions are compared graphically. Once again, the nonparametric kernel density estimations implemented in Section 1 are used as benchmarks. Considering the influence of volatility estimation techniques on the devolatilized returns, the probability densities with respect to the local constant (top) and GARCH(1, 1) (bottom) techniques are graphed separately in Figure 6. Based

Figure 4. The time series of the daily losses, the local constant volatility (—) and the GARCH(1, 1) volatility (⋯⋯) with respect to the DEM/USD rate (left) and the German bank portfolio (right). The parameters in the local constant models are $t_0 = 501$, $m_0 = 5$, and $\eta' = 1.06$ in the DEM/USD data and $\eta' = 1.23$ for the German bank portfolio data.

on the DEM/USD data, the HYP (solid line) and NIG (triangle) models can better describe the distributions of the devolatilized

returns, ε_t^{LC} and ε_t^{GARCH} than the normal (dashed line) and $t(6)$ (dotted line). The normal density underestimates the right tail

Figure 5. Boxplots of the homogeneous interval length with respect to the DEM/USD exchange rates (a) and the German bank portfolio data (b).
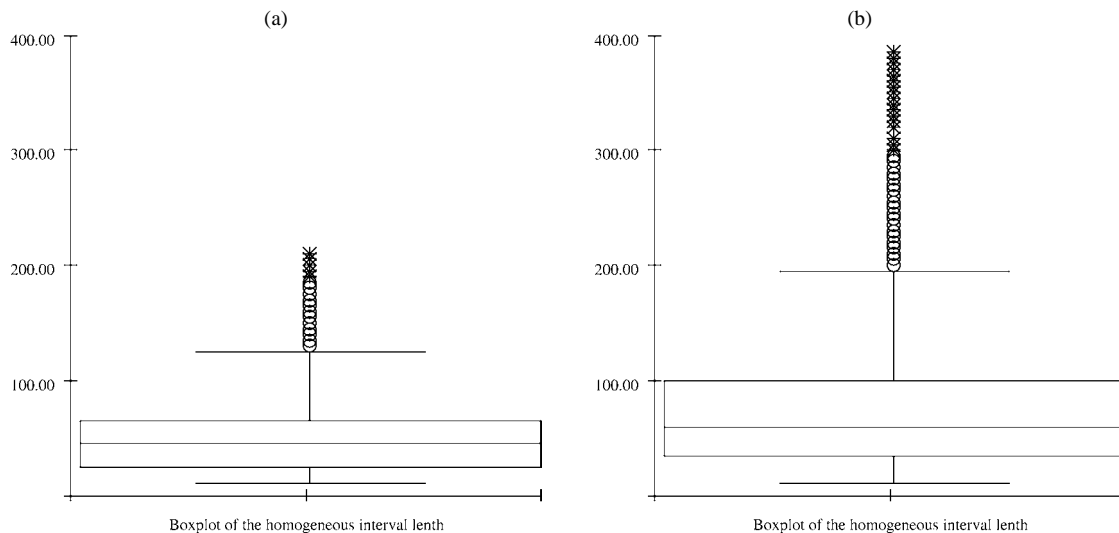
Table 6. Distributional parameters of the devolatilized residuals with respect to the local constant (LC) volatility and the GARCH(1, 1) volatility of the DEM/USD data and the German bank portfolio data

Model	Exchange rate: $t \in [501, 3,719]$				Bank portfolio: $t \in [501, 5,602]$			
	α	β	δ	μ	α	β	δ	μ
HYPADA	1.74	.01	.78	−.01	1.44	.01	0	0
NIGADA	1.34	.01	1.33	−0	.99	−.01	.94	.02
HYPGARCH	1.65	.02	.63	−.02	1.41	−.01	0	0
NIGGARCH	1.20	.02	1.21	−.01	.87	0	.88	−.02

of the devolatilized returns, whereas the $t(6)$ displays a heavier right tail than the benchmark. This misspecification is enlarged in the log-density comparison (right) in Figure 6. In addition, the devolatilized returns ε_t^{LC} with assumptions of the HYP and NIG distributions match the shape of the benchmark better than ε_t^{GARCH}. Based on the GARCH(1, 1) process, all of the density estimations deviate from these benchmarks. This weak performance of the GARCH devolatilization is more obvious in the German bank data (Fig. 7). The benchmarks with the GARCH devolatilization are weakly matched by all four distributional assumptions. It is possible that the German bank portfolio data are less liquid and thus are stationarily weaker than the ex-change rate data, where the local constant model is more suitable, at least theoretically, for capturing the movement of the volatility process.

A latent problem of density estimation is whether the distribution of the stochastic term is actually stationary. A small experiment indicates that the distributional parameters, like volatility, could be time-variant as well. Figure 8 shows the HYP-quantile forecasts based on 500 historical devolatilized returns of the exchange rate for each time point. It provides evidence that the quantiles vary as time passes, especially for extreme probability levels such as $p = .995$. The same phenomenon holds for the NIG distribution, which is omitted here. If the sample size

Figure 6. Estimated densities based on the HYP, NIG, standard normal, and $t(6)$ distributions compared with the nonparametric kernel density in the left panel (top left). The local constant model is applied to devolatilize R_t of the DEM/USD data on the top and the GARCH(1, 1) on the bottom. The corresponding log densities are displayed in the right panel. The distributional parameters are listed in Table 6.
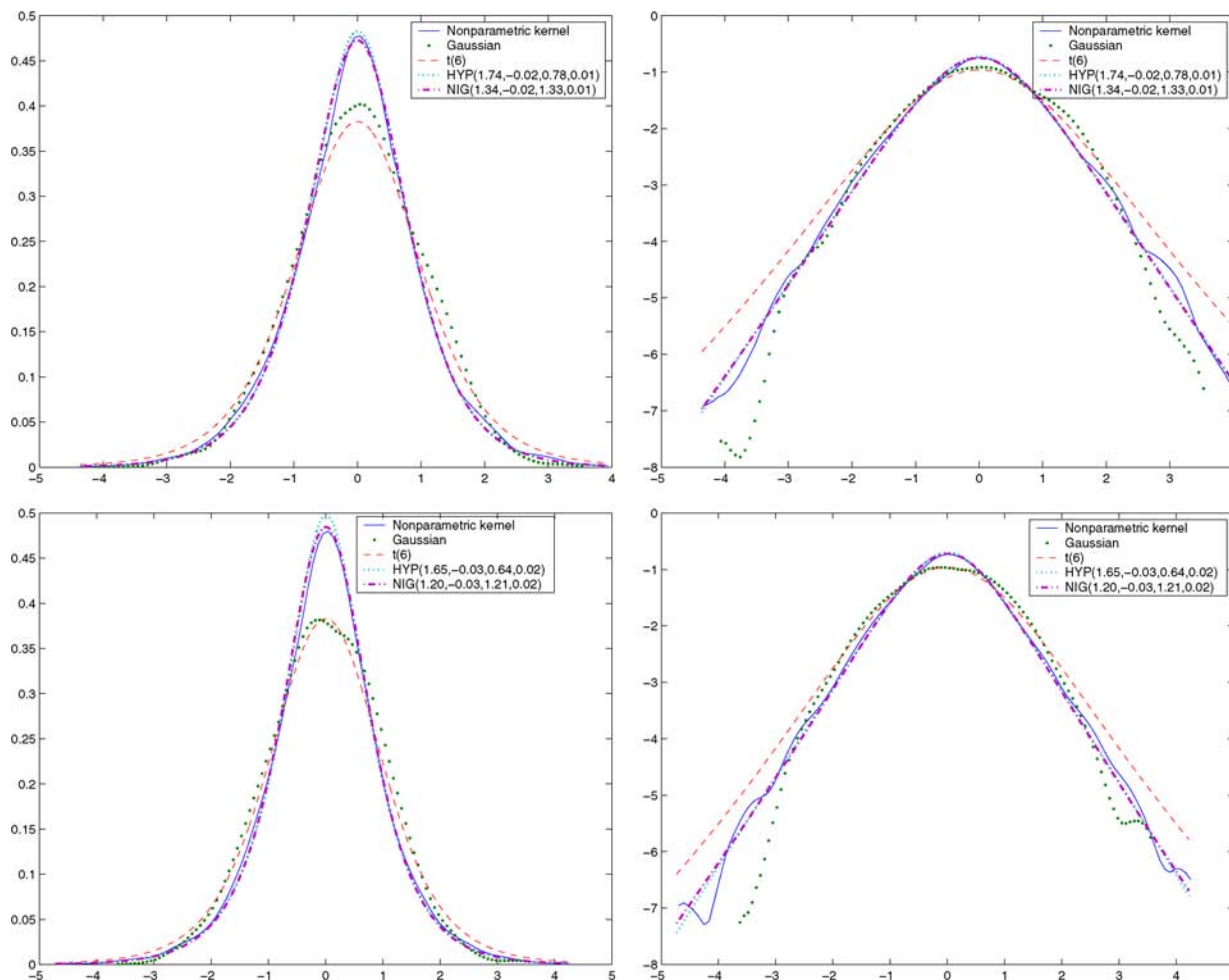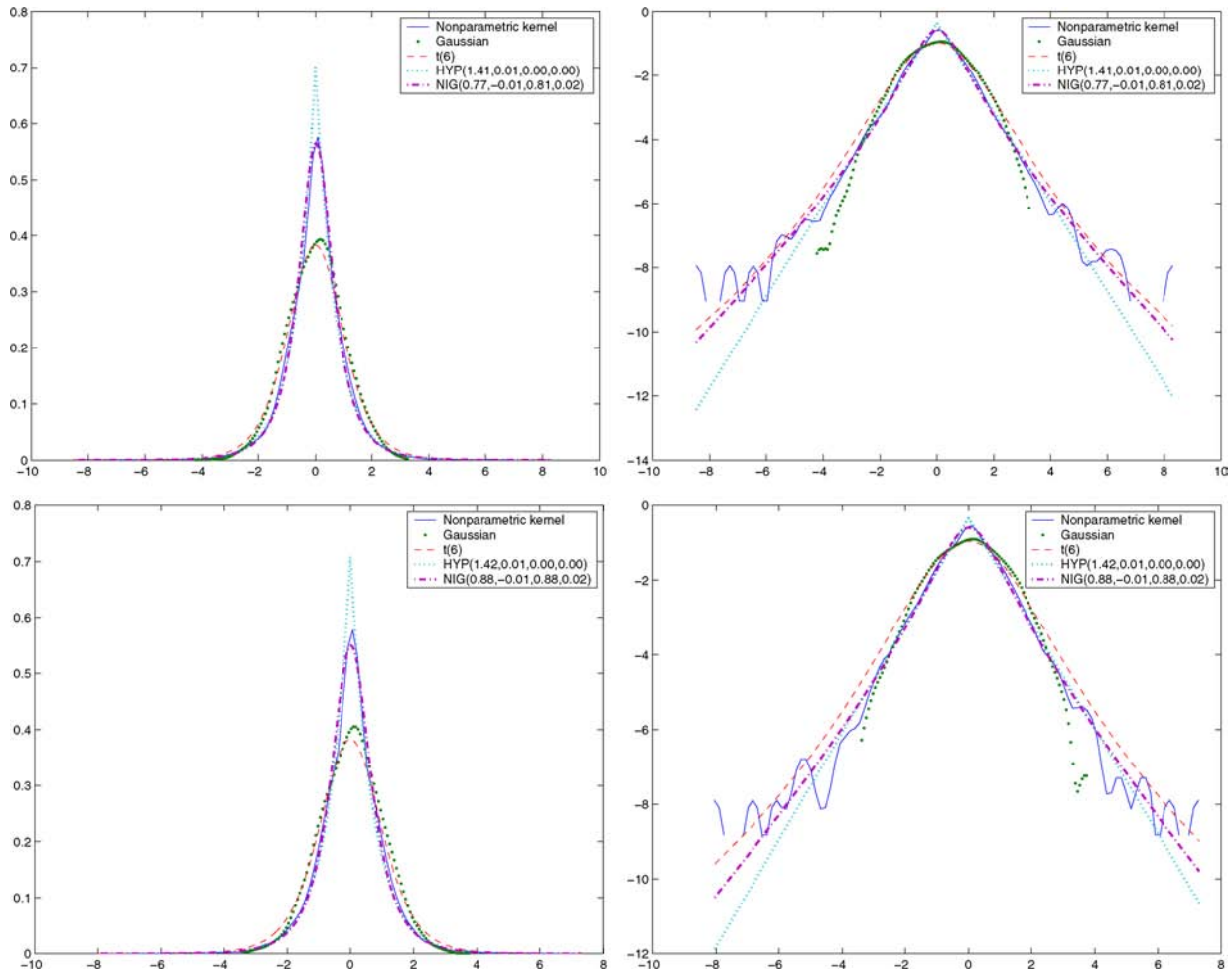
Figure 7. Estimated densities based on the HYP, NIG, standard normal, and $t(6)$ distributions compared with the nonparametric kernel density (circles) in the left panel (top left). The local constant model is applied to devolatilize R_t of the German bank data on the top and the GARCH$(1, 1)$ on the bottom. The corresponding log densities are displayed in the right panel. The distributional parameters are listed in Table 6.

is small, then we could not stick to the assumption that the innovations are identically distributed, although it assumes that the historical observations are iid as well. Instead, we should update the distributional parameters daily. For example, we may

Figure 8. Quantiles estimated based on the past 500 devolatilized returns of the exchange rate. From the top the evolving HYP quantiles for $p = .995, .99, .975, .95, .90, .10, .05, .025, .01,$ and $.005$.

estimate the local distribution based on the previous 500 data points. However, as the sample size increases to infinity, the distribution converges to the unconditional distribution asymptotically. Given the two data sets with large sample size, we assume that all of the observations have an identical distribution.

4.2 VaR

In this section we focus on the model selection from the proposals given in Table 2. Above all, the selected model should be theoretically reasonable and practically tractable to implement. Based on this criterion, we prefer the GHADA model due to its desirable properties, as discussed earlier. Another important criterion of model selection is to compare the accuracy of VaR forecasts. The VaR at the probability level p is forecast as

$$\widetilde{VaR}_{p,t+1} = \tilde{\sigma}_{t+1}\hat{q}_{\varepsilon_t}(p). \tag{19}$$

The quantile $q_{\varepsilon_t}(p)$ is computed according to the distributional estimation. The volatility in the future is unavailable, but empirical studies have found that volatility time series appear to have a unit root (Poon and Granger 2003). Therefore, the estimated volatility $\hat{\sigma}_t$ today is naturally used as the forecast of tomorrow.

Exceedance occurs when an observation exceeds the forecast VaR. A valid VaR model should neither underestimate nor overestimate the market risk. To evaluate the validation of VaR calculation, backtesting presented by Christoffersen (1998) is implemented. The empirical risk level, the ratio of exceedances in the time interval under consideration, is compared with the expected risk level p,

$$\frac{N}{T} = T^{-1} \sum_{t=1}^{T} \mathbf{1}_t, \qquad (20)$$

where $\mathbf{1}_t$ denotes the indicator of exceedances at time point t. An empirical risk level that is larger than p indicates an overestimation of the selected model. In this case, more capital requirements than necessary need to be located in the central bank, and prospective business opportunities are lost. In contrast, if the empirical risk level is much smaller than p, then punishment due to risk underestimation will occur. The null hypothesis of the backtesting is formulated as

$$H_0 : \mathsf{E}[N] = Tp. \qquad (21)$$

Under H_0 where N is a binomial random variable with parameters T and p, the likelihood ratio test statistic can be derived as

$$LR = -2\log\{(1-p)^{T-N} p^N\}$$
$$+ 2\log\{(1-N/T)^{T-N}(N/T)^N\}, \quad (22)$$

which is asymptotically $\chi^2(1)$ distributed with critical values 3.84 (95%) and 6.63 (99%) (Jorion 2001).

Table 7 summarizes the backtesting results using eight different models on the two data sets. As illustrated in Figure 6, the HYP and NIG distributional assumptions reflect the empirical distribution of the DEM/USD stochastic term better than the normal and t distributions; therefore, models based on these two distributions should give more accurate VaRs. The backtesting provides evidence that this is true, as shown in Table 7. In general, the normal distribution underestimates, and the $t(6)$ distribution overestimates the risk. Moreover, given the DEM/USD data, there is no large difference between models with the HYP and NIG distributions and the two volatility estimation techniques. Concerning the German bank data, the HYP model gives more accurate VaR forecasts than the NIG, and the local constant technique performs better than the GARCH(1, 1) setup.

5. CONCLUSION

In this article we have proposed a risk management (GHADA) model based on the adaptive volatility estimation and the GH distribution. Compared with the other proposed risk management models listed in Table 2, the GHADA technique gives more accurate VaR forecasts in real data analysis. Some interesting points are summarized:

- Above all, the two subclasses, HYP and NIG, of the GH distribution can better describe the distributional features of the risk factors than the normal and t distributions. Both distributions closely match the empirical density, especially the right-tail behavior of losses, which in turn leads to precise quantile estimations at extreme risk levels. But

Table 7. Backtesting results for the DEM/USD data and the German bank portfolio data

	DEM/USD data			German bank data		
	N/T	LR1	p value	N/T	LR1	p value
$p = .95$						
HYPADA	.9443	2.05	.15	.9515	−NAN	−NAN
NIGADA	.9443	2.05	.15	.9482	−NAN	−NAN
NADA	.9475	.41	.51	.9547	−NAN	−NAN
tADA	.9711	*35.37	0	.9680	−NAN	−NAN
HYPGARCH	.9521	.32	.57	.9512	−NAN	−NAN
NIGGARCH	.9503	0	.93	.9486	−NAN	−NAN
NGARCH	.9521	.32	.57	.9535	−NAN	−NAN
tGARCH	.9692	*28.92	0	.9696	+INF	0
$p = .975$						
HYPADA	.9742	.08	.77	.9737	.32	.56
NIGADA	.9732	.38	.53	.9698	5.28	.02
NADA	.9714	1.61	.20	.9682	−NAN	−NAN
tADA	.9888	*31.65	0	.9843	*20.91	0
HYPGARCH	.9754	.02	.86	.9772	1.10	.29
NIGGARCH	.9745	.02	.86	.9764	.46	.49
NGARCH	.9708	2.21	.13	.9706	3.83	.05
tGARCH	.9844	*13.65	0	.9856	*28.22	0
$p = .99$						
HYPADA	.9897	.02	.88	.9894	.17	.67
NIGADA	.9897	.02	.88	.9884	1.20	.27
NADA	.9854	6.02	.01	.9811	*31.81	0
tADA	.9972	*23.60	0	.9941	*10.26	0
HYPGARCH	.9906	.15	.69	.9909	.51	.47
NIGGARCH	.9906	.15	.69	.9919	2.13	.14
NGARCH	.9822	*15.71	0	.9835	*18.02	0
tGARCH	.9956	*13.17	0	.9954	*19.54	0
$p = .995$						
HYPADA	.9950	0	.98	.9945	.23	.62
NIGADA	.9950	0	.98	.9939	1.11	.29
NADA	.9897	*13.66	0	.9880	*35.62	0
tADA	.9990	*16.16	0	.9968	4.11	.04
HYPGARCH	.9950	0	.98	.9960	1.29	.25
NIGGARCH	.9950	0	.98	.9962	1.83	.17
NGARCH	.9891	*16.68	0	.9892	*25.70	0
tGARCH	.9987	*13.09	0	.9972	6.24	.01

*Indicates that the model is rejected at 99% confidence level.

we do not have sufficient evidence to say that one subclass is better than the other, although the HYP performs better in the simulation study, as well as the empirical study based on German bank portfolio. We consider that the performance of these two subclasses relies on the data set considered. However, a subjective suggestion is that the HYP is more suitable for less liquid portfolios.

- The local constant technique provides at least as good a volatility estimate as a GARCH(1, 1). The simulation study shows that sometimes the former provides more accurate volatility estimation on average. Furthermore, the difference between the detection speeds of two techniques is trivial. We want to emphasize here that although the GARCH(1, 1) performs well in the simulation study, it is weakened by its theoretical assumption; that is, the volatility estimation follows a time-constant closed form. This influence is illustrated in the German bank portfolio analysis, where even the HYP and NIG weakly represent the

empirical density of the devolatilized return based on the GARCH(1, 1) technique. In contrast, the devolatilized return based on the local constant technique displays a nice fit without problems.

- The proposed approach can be easily applied to calculate and forecast other risk measures, such as expected shortfall.

In financial markets, it is more interesting and challenging to measure the risk levels of multiple time series. Härdle, Herwartz, and Spokoiny (2003) have proposed an idea for adaptive volatility estimation with multiple time series. Prause (1999) has discussed the estimate from the high-dimensional GH distribution. But both of those studies are computationally cumbersome and infeasible as the dimension of the data set increases. Instead, we expect to apply a different approach, independent component analysis (ICA), to convert the problem of high dimensionality to univariate study through a linear transformation (Hyvärinen, Karhunen, and Oja 2001). This procedure can be formulated as

$$R_t = \mathbf{b}_t^\top \mathbf{X}_t$$
$$= \mathbf{b}_t^\top \mathbf{W} \mathbf{Y}_t$$
$$= \mathbf{b}_t^\top \mathbf{W} \operatorname{diag}(\sigma_{y1,t}, \dots, \sigma_{yd,t}) \varepsilon_t,$$

where R_t denotes the portfolio loss with d-dimensional individual instruments that have losses, $\mathbf{X}_t \in \mathbb{R}^d$. The weights of the portfolio are expressed as $\mathbf{b}_t$. After the linear transformation with the matrix $\mathbf{W}$, we obtain (approximately) independent components $\mathbf{Y}_t$. For each IC, the GHADA technique helps specify the volatility and marginal distribution. Because the components are independent, the joint density is simply the product of these marginals. A detailed study of this approach has been given by Chen, Härdle, and Spokoiny (2007).

APPENDIX: PROOF OF LEMMA

For a GH-distributed random variable, we have the following lemma.

Lemma A.1. For every $0 \le \gamma \le 1$, there exists a constant $a_\gamma > 0$ such that

$$\log \mathsf{E}[e^{u\zeta_\gamma}] \le \frac{a_\gamma u^2}{2},$$

where $\zeta_\gamma = (|\varepsilon|^\gamma - C_\gamma)/D_\gamma$ with ε from a GH distribution.

Proof. First, we show that the moment-generating function $\mathsf{E}[e^{u\zeta_\gamma}]$ exists for all $u \in \mathbb{R}$. Suppose that $\mathcal{L}(x) = \mathrm{GH}(\lambda, \alpha, \beta, \delta, \mu)$ with density function f for the transformed variable $y \overset{\text{def}}{=} |x|^\gamma$; then we have

$$P(y \le z) = P\big(-z^{1/\gamma} \le x \le z^{1/\gamma}\big)$$
$$= \int_{-\infty}^{z^{1/\gamma}} f(x)\,dx - \int_{-\infty}^{-z^{1/\gamma}} f(x)\,dx, \qquad z > 0.$$

Then the density of $y \in (0, \infty)$ is

$$g(z) = \frac{d}{dz} P(y \le z)$$
$$= \gamma^{-1}\{f(z^{1/\gamma})z^{1/\gamma-1} + f(-z^{1/\gamma})z^{1/\gamma-1}\}$$
$$= \gamma^{-1}z^{1/\gamma-1}\{f(z^{1/\gamma}) + f(-z^{1/\gamma})\}, \qquad z > 0.$$

Because $f_{GH}(x; \lambda, \alpha, \beta, \delta, \mu = 0) \sim x^{\lambda-1}e^{-(\alpha-\beta)x}$ as $x \to \pm\infty$, it follows that

$$g(z) \sim \frac{z^{1/\gamma-1}}{\gamma}\{z^{(\lambda-1)/\gamma}e^{(\beta-\alpha)z^{1/\gamma}} + z^{(\lambda-1)/\gamma}e^{-(\beta+\alpha)z^{1/\gamma}}\}$$
$$= \frac{z^{\lambda/\gamma-1}}{\gamma}\{e^{(\beta-\alpha)z^{1/\gamma}} + e^{-(\beta-\alpha)z^{1/\gamma}}\}, \qquad z \longrightarrow \infty.$$

For $\gamma < 1$, it holds that $\int_0^\infty e^{uz}g(z)\,dz < \infty \ \forall u \in \mathbb{R}$, because

$$\lim_{z \to \infty} (\beta - \alpha)z^{1/\gamma} + uz \to -\infty \qquad \forall u \in \mathbb{R}$$

and

$$\lim_{z \to \infty} -(\beta + \alpha)z^{1/\gamma} + uz \to -\infty \qquad \forall u \in \mathbb{R}.$$

Because the integration depends only on the exponential part, it also holds that

$$\int_0^\infty z^n e^{uz} g(z)\,dz = \int_0^\infty \frac{\partial^n}{\partial u^n}(e^{uz})g(z)\,dz = \frac{\partial^n}{\partial u^n}\mathsf{E}[e^{uy}] < \infty.$$

Then it can be shown that the moment-generating function and $\log(\mathsf{E}[e^{uy}])$ are smooth. It holds that for every $t > 0$,

$$\mathsf{E}[e^{uy}] = \mathsf{E}[e^{u|x|^\gamma}]$$
$$= \mathsf{E}[e^{u|x|^\gamma}\mathbf{1}(|x| \le t)] + \mathsf{E}[e^{u|x|^\gamma}\mathbf{1}(|x| > t)]$$
$$\le e^{ut^\gamma} + \mathsf{E}[e^{|x|ut^{\gamma-1}}I(|x| > t)]. \tag{A.1}$$

Without loss of generality, we assume that $\mu = 0$. Furthermore,

$$f_{GH}(x; \lambda, \alpha, \beta, \delta, \mu = 0) \sim x^{\lambda-1}e^{-(\alpha-\beta)x} \quad \text{as } x \to \infty,$$

and $\int_y^\infty x^{\lambda-1}e^{-x}\,dx \sim y^{\lambda-1}e^{-y}$ as $y \to \infty$ (Press et al. 1992).

For an arbitrary but fixed $u \in \mathbb{R}_+$ and $t_0 > 1$ so that $ut^{\gamma-1} < \alpha - \beta$, it holds that for all $t \ge t_0$,

$$f(t) \le C_1 t^{\lambda-1}e^{(\beta-\alpha)t}$$

and

$$\int_{(\alpha-\beta-ut^{\gamma-1})t}^\infty x^{\lambda-1}e^{-x}\,dx$$
$$\le C_2[(\alpha - \beta - ut^{\gamma-1})t]^{\lambda-1}e^{-(\alpha-\beta-ut^{\gamma-1})t},$$

where $C_1, C_2 > 1$.

Consequently, for $t \ge t_0$,

$$\mathsf{E}[e^{u|t|^{\gamma-1}x}\mathbf{1}(|x| > t)]$$
$$= \int_t^\infty e^{ut^{\gamma-1}x} f(x)\,dx$$
$$\le C_1 \int_t^\infty e^{ut^{\gamma-1}x} x^{\lambda-1}e^{-(\alpha-\beta)x}\,dx$$
$$= C_1 \int_t^\infty x^{\lambda-1}e^{-(\alpha-\beta-ut^{\gamma-1})x}\,dx$$
$$= C_1(\alpha - \beta - ut^{\gamma-1})^{-\lambda} \int_{(\alpha-\beta-ut^{\gamma-1})t}^\infty x^{\lambda-1}e^{-x}\,dx$$
$$\le C_1 C_2 t^{\lambda-1}e^{-(\alpha-\beta-ut^{\gamma-1})t}(\alpha - \beta - ut^{\gamma-1}t)^{-1}. \tag{A.2}$$

If u is so large that $t \overset{\text{def}}{=} (\frac{\alpha-\beta}{2})^{1/(\gamma-1)}u^c \ge t_0$ with $\frac{1}{1-\gamma} \le c$, then (A.2) holds true because $ut^{\gamma-1} = (\frac{\alpha-\beta}{2})uu^{c(\gamma-1)} \le \frac{\alpha-\beta}{2} < \alpha - \beta$.

Given $t = (\frac{\alpha-\beta}{2}u)^{1/(1-\gamma)}$, we get

$$\mathsf{E}\big[e^{ut^{\gamma-1}x}\mathbf{1}(|x| > t)\big]$$

$$\leq \frac{2C_1C_2}{\alpha-\beta}\left(\frac{\alpha-\beta}{2}u\right)^{(\lambda-1)/(1-\gamma)}e^{-(\alpha-\beta)/2((\alpha-\beta)/2u)^{1/(1-\gamma)}},$$

from which we get

$$\log\big(\mathsf{E}\big[e^{ut^{\gamma-1}}\mathbf{1}(x > t)\big]\big)$$

$$\leq C_3 + \frac{\lambda-1}{1-\gamma}\log(u) - \left(\frac{\alpha-\beta}{2}\right)^{(2-\gamma)/(1-\gamma)}u^{1/(1-\gamma)}.$$

Furthermore, $\log(\mathsf{E}[e^{ut^{\gamma-1}}\mathbf{1}(x > t)])u^{-1/(1-\gamma)}$ is also bounded for $u \to \infty$. Analogously, we can show the bounding of $\log(\mathsf{E}[e^{ut^{\gamma-1}} \times \mathbf{1}(x < -t)])u^{-1/(1-\gamma)}$. Therefore, for $\gamma < 1$, the whole term $\mathsf{E}[e^{u|x|^\gamma}\mathbf{1}(|x| > t)]u^{-1/(1-\gamma)}$ is bounded as $u \to \infty$.

Given $t = (\frac{\alpha-\beta}{2}u)^{1/(1-\gamma)}$, we have

$$e^{ut^\gamma} = e^{((\alpha-\beta)/2)^{\gamma/(1-\gamma)}u^{1/(1-\gamma)}}$$

and

$$u^{-1/(1-\gamma)}\log\big(e^{ut^\gamma}\big) = \left(\frac{\alpha-\beta}{2}\right)^{\gamma/(1-\gamma)} = \text{constant}.$$

Thus $u^{-1/(1-\gamma)}\log(\mathsf{E}[e^{u|x|^\gamma}]) \leq u^{-1/(1-\gamma)}[\log(e^{ut^\gamma}) + \log\{E[e^{ut^{\gamma-1}|x|}\mathbf{1}(|x| > t)]\}]$ is bounded for $u \to \infty$; that is, for sufficiently large u_0, there exists a constant $C_u > 0$ such that

$$\mathsf{E}\big[e^{u|x|^\gamma}\big] \leq C_u u^{1/(1-\gamma)}, \qquad u \geq u_0.$$

Proof of the Martingale Property of θ_t. Consider a predictable process p_t (e.g., the volatility σ_t or the local parameter θ_t) with respect to the information set $\mathcal{F}_{t-1}$,

$$\Upsilon_t = \exp\left(\sum_{s=1}^{t}p_s\zeta_s - (a_\gamma/2)\sum_{s=1}^{t}p_s^2\right).$$

Here Υ_t is a supermartingale, because

$$\mathsf{E}(\Upsilon_t|\mathcal{F}_{t-1}) - \Upsilon_{t-1} = \mathsf{E}(\Upsilon_t|\mathcal{F}_{t-1}) - \mathsf{E}(\Upsilon_{t-1}|\mathcal{F}_{t-1})$$

$$= \mathsf{E}\left[\exp\left(\sum_{s=1}^{t}p_s\zeta_s - (a_\gamma/2)\sum_{s=1}^{t}p_s^2\right)\right.$$

$$\left. - \exp\left(\sum_{s=1}^{t-1}p_s\zeta_s - (a_\gamma/2)\sum_{s=1}^{t-1}p_s^2\right)\bigg|\mathcal{F}_{t-1}\right]$$

$$= \mathsf{E}\left[\exp\left(\sum_{s=1}^{t-1}p_s\zeta_s - (a_\gamma/2)\sum_{s=1}^{t-1}p_s^2\right)\right.$$

$$\left. \times \{\exp(p_t\zeta_t - a_\gamma/2p_t^2) - 1\}\bigg|\mathcal{F}_{t-1}\right]$$

$$= \underbrace{\frac{\exp(p_1\zeta_1)}{\exp(a_\gamma/2p_1)}}_{\leq 1, Lemma\ A.1}\cdots\underbrace{\frac{\exp(p_{t-1}\zeta_{t-1})}{\exp(a_\gamma/2p_{t-1})}}_{\leq 1}$$

$$\times \mathsf{E}\left[\underbrace{\frac{\exp(p_t\zeta_t)}{\exp(a_\gamma/2p_t)}-1}_{\leq 1}\bigg|\mathcal{F}_{t-1}\right]$$

$$\leq 0,$$

that is, $\mathsf{E}(\Upsilon_t|\mathcal{F}_{t-1}) \leq \Upsilon_{t-1}$. By this lemma, we obtain a generalized version of theorem 3.1 of Mercurio and Spokoiny (2004) to the case where the $\boldsymbol{\varepsilon}$'s are from a GH distribution. The statistical properties of $\hat{\theta}_I$ are given in Theorem 1.

REFERENCES

Andersen, T., Bollerslev, T., Christoffersen, P., and Diebold, F. (2006), "Volatility and Correlation Forecasting," in *Handbook of Economic Forecasting*, eds. G. Elliott, C. Granger, and A. Timmermann, Amsterdam: Elsevier, pp. 777–878.

Barndorff-Nielsen, O. (1977), "Exponentially Decreasing Distributions for the Logarithm of Particle Size," *Proceedings of the Royal Society of London*, Ser. A, 353, 401–419.

——— (1997), "Normal Inverse Gaussian Distributions and Stochastic Volatility Modelling," *Scandinavian Journal of Statistics*, 24, 1–13.

Barndorff-Nielsen, O. E., and Blæsild, P. (1981), "Hyperbolic Distribution and Ramifications: Contributions to Theory and Applications," in *Statistical Distributions in Scientific Work*, Vol. 4, eds. C. Taillie, P. G. Patil, and A. Baldessari, Dordrect: Reidel, pp. 19–44.

Barndorff-Nielsen, O., and Shephard, N. (2001), "Modelling by Lévy Processes for Financial Econometrics," in *Lévy Processes: Theory and Applications*, eds. O. Barndorff-Nielsen, T. Mikosch, and S. Resnik, Boston: Birkhauser, pp. 283–318.

Bibby, B. M., and Sørensen, M. (2001), "Hyperbolic Processes in Finance," Technical Report 88, University of Aarhus, School of Business.

Bollerslev, T. (1995), "Generalized Autoregressive Conditional Heteroskedasticity," in *ARCH: Selected Readings*, ed. R. F. Engle, Oxford, U.K.: Oxford University Press, pp. 42–60.

Bollerslev, T., and Woolridge, J. (1992), "Quasi-Maximum Likelihood Estimation and Inference in Dynamic Models With Time-Varying Covariances," *Econometric Reviews*, 11, 143–172.

Chen, Y., Härdle, W., and Spokoiny, V. (2007), "Portfolio Value at Risk Based on Independent Components Analysis," *Journal of Computational and Applied Mathematics*, 1, 594–607.

Christoffersen, P. F. (1998), "Evaluating Interval Forecast," *International Economic Review*, 39, 841–862.

Eberlein, E., and Keller, U. (1995), "Hyperbolic Distributions in Finance," *Bernoulli*, 1, 281–299.

Eberlein, E., Kallsen, J., and Kristen, J. (2003), "Risk Management Based on Stochastic Volatility," *Journal of Risk*, 5, 19–44.

Embrechts, P., McNeil, A., and Straumann, D. (1999), "Correlation: Pitfalls and Alternatives," *Risk*, 12, 69–71.

Engle, R. F. (1982), "Autoregressive Conditional Heteroscedasticity With Estimates of the Variance of United Kingdom Inflation," *Econometrica*, 50, 987–1008.

Franke, J., Härdle, W., and Hafner, C. (2004), *Statistics of Financial Markets*, Berlin: Springer-Verlag.

Härdle, W., Herwartz, H., and Spokoiny, V. (2003), "Time-Inhomogeneous Multiple Volatility Modelling," *Journal of Financial Econometrics*, 1, 55–95.

Härdle, W., Müller, M., Sperlich, S., and Werwatz, A. (2004), *Nonparametric and Semiparametric Models*, Berlin: Springer-Verlag.

Harvey, A., Ruiz, E., and Shephard, N. (1995), "Multivariate Stochastic Variance Models," in *ARCH: Selected Readings*, ed. R. F. Engle, Oxford, U.K.: Oxford University Press, pp. 256–276.

Hyvärinen, A., Karhunen, J., and Oja, E. (2001), *Independent Component Analysis*, New York: Wiley.

Jaschke, S., and Jiang, Y. (2002), "Approximating Value at Risk in Conditional Gaussian Models," in *Applied Quantitative Finance*, eds. W. Härdle, T. Kleinow, and G. Stahl, Berlin: Springer-Verlag, pp. 3–33.

Jorion, P. (2001), *Value at Risk*, New York: McGraw-Hill.

Mercurio, D., and Spokoiny, V. (2004), "Statistical Inference for Time-Inhomogeneous Volatility Models," *The Annals of Statistics*, 32, 577–602.

Poon, S., and Granger, C. (2003), "Forecasting Volatility in Financial Markets: A Review," *Journal of Economic Literature*, 41, 478–539.

Prause, K. (1999), "The Generalized Hyperbolic Model: Estimation, Financial Derivatives and Risk Measures," PhD thesis, University of Freiburg.

Press, W., Teukolsky, S., Vetterling, W., and Flannery, B. (1992), *Numerical Recipes in C*, New York: Cambridge University Press.

Smoothed L-estimation of regression function

P. Čížek [a,*], J. Tamine [b], W. Härdle [c]

[a] *Department of Econometrics & OR, Tilburg University, P.O.Box 90153, 5000 LE, Tilburg, The Netherlands*
[b] *Université de la Méditerranée, GREQAM Vieille Charité, 2 rue de la Charité, 13002 Marseille, France*
[c] *Institut für Statistik und Ökonomie, Humboldt-Universität zu Berlin, Spandauer Str. 1, D-10178 Berlin, Germany*

A B S T R A C T

The Nadaraya–Watson nonparametric estimator of regression is known to be highly sensitive to the presence of outliers in data. This sensitivity can be reduced, for example, by using local L-estimates of regression. Whereas the local L-estimation is traditionally done using an empirical conditional distribution function, we propose to use instead a smoothed conditional distribution function. The asymptotic distribution of the proposed estimator is derived under mild β-mixing conditions, and additionally, we show that the smoothed L-estimation approach provides computational as well as statistical finite-sample improvements. Finally, the proposed method is applied to the modelling of implied volatility.

1. Introduction

The nonparametric estimation of regression functions has been receiving much attention in the literature (see Härdle (1990) and Pagan and Ullah (1999), for an overview) and one of the most widely used estimator is the Nadaraya–Watson (NW) estimator by Nadaraya (1964) and Watson (1964). This estimator, being a local average of the response variable, is highly sensitive to the presence of outliers in data (see Barnett and Lewis (1994), for a general discussion of the concept of an outlier). Possible outliers do not only increase the variance of the estimator, but can also create fictitious peaks and therefore structure in the estimation. In order to robustify the NW estimator, Boente and Fraiman (1994) used local L-estimates such as local α-trimmed means instead of a locally weighted average. Their procedure consists in using an empirical conditional distribution function (cdf) that allows us to estimate the amount of data to be discarded. We demonstrate that the empirical cdf is not an optimal choice and that the use of a smoothed cdf, which is shown to be a generalization of the empirical one, has substantial advantages. First, following theoretical arguments of Fernholz (1997) for unconditional L-estimates, we expect estimates based on a smoothed cdf to have better finite-sample properties. Second, the proposed procedure does not require estimating of the local cdf at every point of the sample but only on an integration grid, and consequently, it is less computationally intensive with an increasing sample size.

The paper is organised as follows. In Section 2, we describe both the empirical and smoothed estimators of cdf and explain how they can be used to estimate the conditional L-estimator of regression. In Section 3, asymptotic bounds for the smoothed cdf are given and the asymptotic distribution of the smoothed L-estimator is derived for absolutely regular stationary data. This allows for time-series applications, where outliers are particularly likely to appear as pointed out by Sakata and White (1998), for instance. In Section 4, we show why the smoothed estimator is computationally less time-consuming than the empirical one. In Section 5, we briefly recall the arguments given by Fernholz (1997) in favour of smoothing cdf estimates and perform a Monte Carlo study that demonstrates the superiority of the smoothed estimator in finite samples. Finally

---

* Corresponding author. Tel.: +31 134668723; fax: +31 134663280.
E-mail address: P.Cizek@uvt.nl (P. Čížek).

in Section 6, we provide an example of an application that significantly benefits from the use of a robust nonparametric L-estimation.

2. Robust L-estimates of regression

Here, the conditional L-estimator of regression is introduced in Section 2.1. As it depends on a generally unknown cdf, possible estimators are presented in Section 2.2 and plugged in the conditional L-estimator in Section 2.3. The bandwidths for the proposed nonparametric estimator are chosen by cross-validation (CV) as discussed in Section 2.4.

2.1. Model definition

Let $(\mathbf{X}, Y)$ be a $(d+1)$-dimensional random vector. Following Boente and Fraiman (1994), we aim to identify and evaluate the conditional L-estimate of regression and define it by

$$m_L(\mathbf{x}) = \int yJ\{F_{\mathbf{x}}(y)\} \, dF_{\mathbf{x}}(y), \tag{1}$$

where $F_{\mathbf{x}}$ denotes the cumulative distribution function of the random variable Y conditional on the event $\{\mathbf{X} = \mathbf{x}\}$ and J is an L-score function assumed to be continuously differentiable with compact support $\langle a, b \rangle \subset (0, 1)$.

Definition (1) nests several useful statistical parameters. For example, if we consider a cdf $F_{\mathbf{x}}(y)$ symmetric around the conditional expectation $m(\mathbf{x}) = E(Y|\mathbf{X} = \mathbf{x})$ and the α-trimming score function $J_\alpha(u) = I_{(\alpha,1-\alpha)}(u)/(1-2\alpha)$, $\alpha \in \langle 0, 1/2)$, the equality $m_L(\mathbf{x}) = m(\mathbf{x})$ holds. The α-trimmed conditional expectation $m_L(\mathbf{x})$ can be then used to estimate conditional expectation in a robust way. For $\alpha \to 1/2$, $m_L(\mathbf{x})$ is equal to the conditional median.

As will be described in Section 2.3, a natural way of estimating $m_L(\mathbf{x})$ consists of plugging an estimator of cdf $F_{\mathbf{x}}(y)$ in expression (1).

2.2. Estimation of conditional cumulative distribution function

Let $f(\mathbf{x}, y)$ be the joint density of the random vector $(\mathbf{X}, Y)$ and $f(\mathbf{x})$ be the marginal density of the d-dimensional random vector $\mathbf{X}$. The cdf of the random variable Y conditional on the event $\{\mathbf{X} = \mathbf{x}\}$ is defined by

$$F_{\mathbf{x}}(y) = \int_{-\infty}^{y} \frac{f(\mathbf{x}, v)}{f(\mathbf{x})} \, dv.$$

The common practice in literature (Härdle, 1990) consists in estimating this function using an empirical cdf defined by

$$\tilde{F}_{\mathbf{x}}(y) = \sum_{i=1}^{n} \frac{K_{\mathbf{x}}\left(\frac{\mathbf{x}-\mathbf{X}_i}{h_{\mathbf{x}}}\right)}{\sum_{j=1}^{n} K_{\mathbf{x}}\left(\frac{\mathbf{x}-\mathbf{X}_j}{h_{\mathbf{x}}}\right)} I(Y_i \le y), \tag{2}$$

where I denotes the indicator function and $K_{\mathbf{x}}$ is a d-dimensional product kernel. To avoid cumbersome notation, we use for simplicity the same bandwith $h_{\mathbf{x}}$ in each $\mathbf{X}$ direction: thus $K_{\mathbf{x}}(\mathbf{x}) = \prod_{i=1}^{d} K_x(x_i/h_{\mathbf{x}})$.

The empirical cdf $\tilde{F}_{\mathbf{x}}$ is a step function. As pointed out by Fernholz (1997) in the case of an unconditional distribution function, using an empirical cdf may not be the best choice for estimating quantiles, and more generally, computing L-estimates. Furthermore, the step structure of the empirical cdf and its weak regularity properties can be difficult to handle in a theoretical framework.

Therefore, we propose to apply an additional smoothing to the variable Y and to estimate $F_{\mathbf{x}}(y)$ by a smoothed cdf

$$\hat{F}_{\mathbf{x}}(y) = \sum_{i=1}^{n} \frac{K_{\mathbf{x}}\left(\frac{\mathbf{x}-\mathbf{X}_i}{h_{\mathbf{x}}}\right)}{\sum_{j=1}^{n} K_{\mathbf{x}}\left(\frac{\mathbf{x}-\mathbf{X}_j}{h_{\mathbf{x}}}\right)} \int_{-\infty}^{y} \frac{1}{h_y} K\left(\frac{u - Y_i}{h_y}\right) du, \tag{3}$$

where h_y is an additional smoothing parameter, bandwidth in the Y direction. Estimator (3) inherits the regularity properties of the univariate kernel K and may thus be used, for instance, for estimating the derivatives of $F_{\mathbf{x}}(y)$.

Note that $\hat{F}_{\mathbf{x}}(y)$ is a generalization of the empirical cdf $\tilde{F}_{\mathbf{x}}(y)$. Under regularity assumptions, $\int_{-\infty}^{y} K\{(u - Y_i)/h_y\}/h_y du \to I(Y_i \le y)$ for $h_y \to 0$. Hence, $\tilde{F}_{\mathbf{x}}(y)$ is the limit case of $\hat{F}_{\mathbf{x}}(y)$ if the bandwith h_y converges to zero with an increasing sample size n and one can 'feel' that the asymptotic properties of $\hat{F}_{\mathbf{x}}(y)$ and $\tilde{F}_{\mathbf{x}}(y)$ will be identical (see Section 3 for details). Nevertheless, and this is the core of this work, their finite-sample behavior and computational properties appear to be significantly different.

2.3. The empirical and smoothed L-estimators

In order to estimate $m_L(\mathbf{x})$, one can plug estimators (2) and (3) of $F_{\mathbf{x}}(y)$ in expression (1). If one plugs in the empirical cdf (2) as proposed by Boente and Fraiman (1994), $m_L(\mathbf{x})$ is estimated by the local empirical L-estimator

$$\tilde{m}_L(\mathbf{x}) = \sum_{i=1}^{n} \frac{Y_i K_{\mathbf{x}}\left(\frac{\mathbf{x}-\mathbf{x}_i}{h_{\mathbf{x}}}\right)}{\sum_{j=1}^{n} K_{\mathbf{x}}\left(\frac{\mathbf{x}-\mathbf{x}_j}{h_{\mathbf{x}}}\right)} J\left\{\tilde{F}_{\mathbf{x}}(Y_i)\right\}. \tag{4}$$

Instead employing the empirical cdf, we propose to plug in the smoothed cdf (3). The function $m_L(\mathbf{x})$ is then estimated by the local smoothed L-estimator

$$\hat{m}_L(\mathbf{x}) = \int y J\left\{\hat{F}_{\mathbf{x}}(y)\right\} \mathrm{d}\hat{F}_{\mathbf{x}}(y), \tag{5}$$

where the integral is approximated using classical numerical integration routines (Davis and Rabinowitz, 1984).

In Section 3, we show that the empirical and smoothed L-estimators have the same first-order asymptotic properties. However, we demonstrate that the smoothed L-estimator (5) has superior computational and finite-sample statistical properties to the empirical L-estimator (4); see Sections 4 and 5.

2.4. Cross-validation

On the one hand, we study nonparametric regression estimators that are robust to outliers, contrary to the least-square based methods such as the NW estimator. On the other hand, nonparametric smoothing depends on the choice of smoothing parameters (bandwidths), which are typically chosen by cross-validating some squared error (Park and Marron, 1990). By the same argument as in the case of NW, the standard CV based on squared errors appears to be sensitive to data contamination and outliers in particular (Leung et al., 1993).

As a remedy, Wang and Scott (1994) proposed to cross-validate the absolute value of the error, the L_1 cross-validation. This and more general CV criteria have been recently analyzed by Leung (2005). Since we primarily use the α-trimming score J_α here and the most robust choice of $\alpha \to 0.5$ (median smoothing) corresponds to the local L_1 regression, we employ the L_1 cross-validation (CV1) and compare it to the standard squared errors cross-validation (CV2). The CV1 criterion is not only simple to evaluate, but also proved its robust properties in the recent study by Čížek and Härdle (2006).

Let us note that we use CV to determine both bandwidths $h_{\mathbf{x}}$ and h_y even though this strategy might not be optimal in the case of the distribution-function smoothing (parameter h_y). This is however of limited importance here. First, the selection of bandwidth h_y aims at choosing a 'reasonable' value that is positive, but sufficiently small so that the eventual outliers cannot significantly influence estimation of $\hat{F}_{\mathbf{x}}(y)$ for $F(y) \in (\alpha, 1-\alpha)$ and some $\alpha > 0$. Since outliers with significant influence on estimation lead to pronounced estimation biases, CV suits this purpose well. Second, the optimal choice of h_y matters only to a limited extent because there is asymptotically no first-order difference between the empirical and smoothed L-estimators (see Section 3).

3. Asymptotic analysis

To establish asymptotic bounds on the smoothed cdf (3) and to derive the asymptotic distribution of $\hat{m}_L(\mathbf{x})$, we use the following assumptions (chosen to allow for an easy presentation rather than full generality). Note that the dependence of $h_{\mathbf{x}}$ and h_y on n is left implicit for the simplicity of notation.

A1: Variables $\{(\mathbf{X}_i, Y_i)\}_{i=1}^{n}$ form a strictly stationary absolutely regular sequence of random vectors.

A2: The density $f(\mathbf{x}, y)$ is compactly supported, admits continuous derivatives of order r, and the marginal density $f(\mathbf{x})$ is bounded from below by $b > 0$.

A3: The univariate kernel function K is a symmetric compactly supported kernel of order r_y such that the integral $K_I(y) = \int_{-\infty}^{y} K(u)\, \mathrm{d}u$ is defined for all y.

A4: The multivariate kernel function $K_{\mathbf{x}}$ is assumed to be a product kernel, $K_{\mathbf{x}}(\mathbf{x}) = \prod_{i=1}^{d} K_x(x_i/h_{\mathbf{x}})$, where K_x is a symmetric compactly supported kernel of order r_x, and $r = \min\{r_x, r_y\} \geq 2$.

A5: The bandwidths $h_{\mathbf{x}}$ and h_y satisfy $\lim_{n\to\infty} h_{\mathbf{x}} = 0$ and $\lim_{n\to\infty} h_y = 0$ in such a way that $\lim_{n\to\infty} n^{1/2} h_{\mathbf{x}}^{3d/2} \to \infty$ and $\lim_{n\to\infty} n^{1/2} h_{\mathbf{x}}^{d/2} \left(\max\{h_{\mathbf{x}}, h_y\}\right)^r \to 0$.

A6: The L-score function $J : \langle 0, 1 \rangle \to R$ is assumed to be non-negative, bounded, and having a bounded derivative almost everywhere.

First, the asymptotic bounds on the smoothed cdf (3) are given.

Lemma 1. *Under Assumptions* A1 – A5, *the following uniform bound holds for the smoothed cdf* $\hat{F}_{\mathbf{x}}(y)$ *defined in* (3):

$$\sup_{x,y} \left| \hat{F}_{\mathbf{x}}(y) - F_{\mathbf{x}}(y) \right| = O_p \left[n^{-1/2} h_{\mathbf{x}}^{-d} + \left(\max \{ h_{\mathbf{x}}, h_y \} \right)^r \right].$$

Proof. See Čížek et al. (2006, Lemma 1). □

Next, the asymptotic distribution of the smoothed L-estimator is derived.

Theorem 2. *Under Assumptions* A1 – A6, *the asymptotic distribution of the smoothed local L-estimator* $\hat{m}_L(\mathbf{x})$ *defined in* (5) *is*

$$n^{1/2} h_{\mathbf{x}}^{d/2} \left\{ \hat{m}_L(\mathbf{x}) - m_L(\mathbf{x}) \right\} \xrightarrow{\mathcal{L}} \mathcal{N}(0, V)$$

for $n \to \infty$, *where*

$$V = \int \left[\int_{-\infty}^{+\infty} \frac{J\{F_{\mathbf{x}}(y)\}}{f(\mathbf{x})} \left\{ I_{(-\infty, y)}(w) - F_{\mathbf{x}}(y) \right\} \mathrm{d}y \right]^2 f(\mathbf{x}, w) \mathrm{d}w \cdot \int K_{h_{\mathbf{x}}}^2(\mathbf{u}) \mathrm{d}\mathbf{u}.$$

Proof. See Čížek et al. (2006, Theorem 2). □

Consequently, the asymptotic distribution of the smoothed L-estimator (5) is equivalent to that of the empirical L-estimator (4), and therefore, the choice of bandwidth h_y does not asymptotically play any role (provided that Assumption A5 holds). This follows from the fact that the bandwidth h_y does not influence the number of observations used in (local) estimation: it is used to estimate a distribution function, which takes always into account all observations, rather than a density function, which is estimated locally.

4. Computational comparison

The asymptotic distributions of the empirical L-estimator $\tilde{m}_L(\mathbf{x})$ and the smoothed L-estimator $\hat{m}_L(\mathbf{x})$ are identical. On the other hand, the computational costs of the empirical L-estimator grow much faster with the sample size than those of the smoothed L-estimator. To show this, we will compare the computational costs for both L-estimators from the theoretical and empirical points of view.

We assume that the computation of $K_{\mathbf{x}} \{(\mathbf{x} - \mathbf{X}_i)/h_{\mathbf{x}}\}$ requires $C_{\mathbf{x}}(d)$ operations, the computation of $K_I \{(y - Y_i)/h_y\}$ requires C_I operations, and the computation of the score function $J(u)$ requires C_J operations. Since d is the number of employed explanatory variables, we consider d to be fixed here and thus $C_{\mathbf{x}}(d)$ to be a constant. Finally to determine the bandwidth parameters, we assume that CV of $h_{\mathbf{x}}$ requires $C_{cv}(n, d)$ evaluations of an estimator and CV of h_y is always done on a finite grid of C_c points (this suffices given the lesser importance of h_y being chosen optimally, see Section 2.4 and Theorem 2).

Empirical estimation. The calculation of the empirical cdf $\tilde{F}_{\mathbf{x}}(\cdot)$ at one point y using (2) needs $n(C_{\mathbf{x}}(d) + 3)$ operations. Since the computation of $\tilde{m}_L(\mathbf{x})$ requires the computation of $\tilde{F}_{\mathbf{x}}(\cdot)$ at every point Y_i, we get for the empirical L-smoother $\tilde{m}_L(\mathbf{x})$

$$n^2 \{C_{\mathbf{x}}(d) + 3\} + n \{C_J + C_{\mathbf{x}}(d) + 3\} = O(n^2) \tag{6}$$

as the cost of operations. This is eventually multiplied by $C_{cv}(n, d)$ if bandwidth(s) $h_{\mathbf{x}}$ is determined by CV.

Smoothed estimation. Analogously, the computation of the smoothed cdf $\hat{F}_{\mathbf{x}}(\cdot)$ at one point y requires $n(C_{\mathbf{x}}(d) + C_I + 2)$. We assume that the integral in expression (5) is approximated on a grid of k points by a numerical integration method of order $m \geq 2$, that is, with the integration error proportional to k^{-m}.[1] Thus, its computation needs $O(k)$ operations. Consequently, the computation of $\hat{m}_L(\mathbf{x})$ requires

$$n \{C_{\mathbf{x}}(d) + C_I + 2\} O(k) = O(nk) \tag{7}$$

operations. This is again multiplied by $C_c \cdot C_{cv}(n, d)$, if bandwidths $h_{\mathbf{x}}$ and h_y are determined by CV.

Since the costs of CV differs only by a constant factor C_c, the only crucial factor for the comparison of computational speed is how quickly k should grow with n. Since the errors made due to the numerical integration should not be larger than the errors (biases and variance) of nonparametric estimation, $k^{-m} < (nh_{\mathbf{x}}^d)^{-1/2}$ and hence $k > (nh_{\mathbf{x}}^d)^{1/(2m)}$. As $h_{\mathbf{x}}^d < n^{1/3}$ by Assumption A5, it follows that $k > n^{2/(3m)}$ in the worst case. For the simplest and least precise methods with $m = 2$, we should choose $k \sim n^{1/3}$. Consequently, the computational complexity of the smoothed L-estimator is at most $O(n^{4/3})$ and further decreases when the optimal or a smaller bandwidth is employed and when better integration methods are used.

---

[1] For example, the well-known trapezoidal and Simpson rules have precisions of orders $m = 2$ and $m = 4$, respectively, see Davis and Rabinowitz (1984).

Table 1
Relative computational times of the empirical and smoothed L-estimators for sample size n and an integration grid of size k

Method (k)	$n = 50$	$n = 100$	$n = 200$	$n = 500$	$n = 1000$	$n = 2000$
Empirical	0.4	0.8	1.4	3.5	7.0	14.0
Smoothed (50)	0.5	0.4	0.5	0.5	0.5	0.5
Smoothed (100)	1.0	1.0	1.0	1.0	1.0	1.0
Smoothed (150)	2.0	1.4	1.5	1.5	1.5	1.5
Smoothed (250)	3.0	2.5	2.4	2.5	2.5	2.5
Smoothed (500)	5.5	4.9	4.9	5.0	5.0	5.0

Computation time. For a large n, the smoothed L-estimator requires a substantially smaller number of operations than the empirical L-estimator. To corroborate this theoretical result, we performed a set of simulations and measured the computation times of both methods. We used the model

$$Y = m(X) + \varepsilon, \tag{8}$$

where the regressor X is univariate and uniformly distributed on the interval $\langle 0, 1\rangle$, $X \sim U(0, 1)$, and the regression function is given by $m(x) = -1 + x^{1/2} - x^2$. The error term ε has a symmetric distribution so that the regression function $m(x)$ and the α-trimmed expectation $m_L(x)$ are equal. The estimations are performed using the α-trimming L-score J_α with $\alpha = 0.25$.

Table 1 contains the relative times necessary for the estimation of the empirical L-estimator (4) and of the smoothed one (5) for different sample sizes n and different numbers k of points on the numerical integration grid (times are relative to the smoothed L-estimator using $k = 100$). In our opinion, choosing $250 \leq k \leq 500$ is a good choice for applications involving samples with several thousands of observations. Results in Table 1 confirm our theoretical findings that the smoothed estimator will be faster to compute for large samples. Already when sample size n is twice the size k of the integration grid, $n > 2k$, the smoothed L-estimator is faster to compute.

5. Finite-sample comparison

As explained in Section 3, the empirical L-estimator and its smoothed counterpart need not be compared from an asymptotic point of view since they share the same asymptotic distribution. However, relying on arguments established by Fernholz (1997) in an unconditional setting, we can hypothesize that the smoothed L-estimator has better finite-sample properties than the empirical one. This might seem surprising since the additional smoothing involved in estimating the smoothed cdf $\hat{F}_\mathbf{x}(\cdot)$ may cause an additional bias (asymptotically negligible but sensible in finite sample). Nevertheless, as demonstrated by Fernholz (1997), this additional bias goes along with a decrease of the variance of the estimator. Because this decrease surpasses the additional bias, the smoothing will result in a gain in terms of the mean squared error of the smoothed conditional L-estimator.

To verify this claim, we perform a set of Monte Carlo experiments and compare the Nadaraya–Watson (NW), the empirical and smoothed L-estimators, where the α-trimming score function J_α is used for $\alpha = 0.25$ and $\alpha = 0.50$ (the choice $\alpha = 0$ would correspond to NW). We use again the data-generating process (8), that is $Y = -1 + X^{1/2} - X^2 + \varepsilon$, where $X \sim U(0, 1)$, and study performance of all methods at different error distributions and sample sizes. The presented results are based on 500 replications. All estimates are computed using the quartic kernel and both bandwidths $h_\mathbf{x}$ and h_y are chosen by L_1 cross-validation (CV1) on interval $(0, 0.2)$. Hence to match the use of CV1, the estimation results are compared by means of the mean absolute errors (MAE). One could argue that CV1, used to guarantee robust bandwidth choice, could possibly lead to worse performance of all methods, and NW in particular, for Gaussian data. To address this issue, we compared the MAE of all estimators with bandwidths cross-validated by CV1 and CV2: at all points $x \in (0, 1)$, the relative differences between the CV1 and CV2 results were smaller than 5% for the NW and smoothed L-estimator and smaller than 10% for the empirical L-estimator.

In the rest of this section, we attempt to compare the finite-sample properties of the empirical and smoothed L-estimators for different sample sizes (Section 5.1), for different error-term distributions (Section 5.2), and for data contaminated by outliers (Section 5.3).

5.1. Influence of sample size

Let us now compare the empirical and smoothed L-estimators for Gaussian data, $\varepsilon \sim N(0, 0.1)$, at various sample sizes: $n = 50, 100, 200$. The results for J_α with $\alpha = 0.25$ and $\alpha = 0.50$ are summarized in Fig. 1. The first observation is that the smoothed L-estimator performs in all cases better than the empirical L-estimator and is generally very close to the NW estimator. The relative difference between the empirical and smoothed L-estimators seems to be roughly the same at various sample sizes, but, since the estimation errors are decreasing as n increases, the difference decreases in absolute terms. Moreover, it seems that the difference is smaller for $\alpha = 0.25$ than for $\alpha = 0.50$. This indicates that the smoothed cdf helps more if there are more observations close the discontinuity of the L-score function J_α.
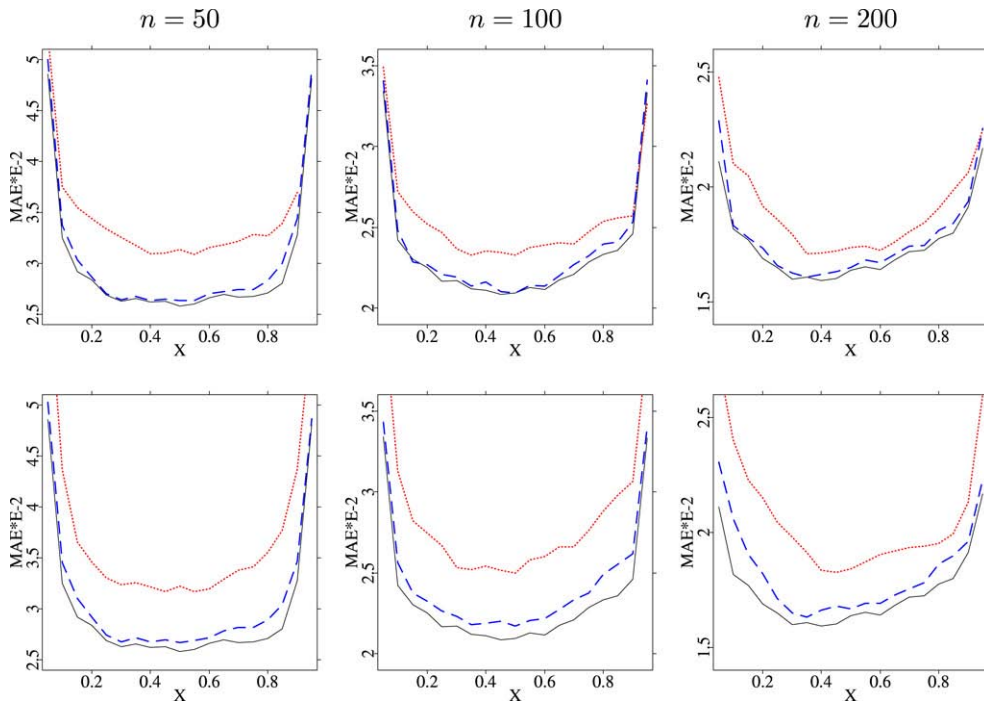
Fig. 1. The mean absolute errors of the Nadaraya–Watson (solid line), empirical L-estimator (dotted line), and smoothed L-estimator (dashed line) for different sample sizes n and the L-score J_α. The upper panels contain estimators for $\alpha = 0.25$ and the bottom panels contain estimators for $\alpha = 0.50$.

5.2. Effect of distributional model

Next, we look at the performance of all methods under various distributional models: for Gaussian $\varepsilon \sim N(0, 0.1)$, Student $\varepsilon \sim \sigma_t t_3$, and double exponential $\varepsilon \sim DExp(\sigma_e)$ errors, where $\sigma_t \doteq 0.057$ and $\sigma_e \doteq 0.071$ are chosen so that the standard deviation of errors is always equal to 0.1. Results for all error distributions and $n = 100$ are summarized in Fig. 2. We can again observe that the smoothed L-estimator generally outperforms the empirical L-estimator, although the difference is more pronounced for $\alpha = 0.50$ and is largest for the Gaussian data. Looking at the results for the Student and double-exponential errors, the empirical and smoothed L-estimators are preferable to the NW estimator for $\alpha = 0.25$ and $\alpha \in \{0.25, 0.50\}$, respectively.

5.3. Sensitivity to outliers

Finally, let us compare all methods for Gaussian samples, $\varepsilon \sim N(0, 0.1)$, of $n = 100$ observations contaminated by three (relatively small) outliers at $x = 0.25, 0.50$, and 0.75 with the values of the dependent variable y uniformly distributed on $(-1.5, 1.5)$, $y \sim U(-1.5, 1.5)$. The results using CV1 are in the upper row of Fig. 3. Obviously, NW is heavily influenced by the outliers, but this is not the case of the L-estimators. The smoothed L-estimator generally performs better than the empirical L-estimator, but due to the smoothing in the y-direction and its dependence on bandwidth h_y, the smoothed L-estimator exhibits a limited sensitivity to the outliers, especially when less trimming is used ($\alpha = 0.25$).

All presented simulations support two important observations. First of all, the smoothed L-estimator performs in all simulations better than the empirical one, although the difference does not have to be very large in some cases. Second, the performance of the smoothed L-estimator seems to be robust with respect to the error distribution and its contamination in the sense that the corresponding mean absolute errors are more or less similar under all tested distribution models.

Compared to the NW estimator, the smoothed L-estimator is slightly worse than NW under Gaussian errors, especially for $\alpha \to 0.5$, but outperforms NW for other examined distributions as well as in the presence of outliers. Hence, the smoothed L-estimator seems to be a good method for nonparametric regression even if the presence of outliers in the data is only hypothetical. On the other hand, the classical NW estimator can be computed much faster than the discussed L-estimators, and additionally, it usually renders easy generalizations to new types of estimation problems (e.g., nonparametric functional data analysis). Note however that the L-estimation is typically applicable in the same areas as NW provided that replacing the conditional mean by the conditional median is acceptable and provides valid answers in a given context. For example, the L-estimation could be applied in the nonparametric functional data analysis because nonparametric L-estimators deal with explanatory variables X in the same way as NW, see (4) and (5), and only the response Y is treated differently than in the case of NW.

Fig. 2. The mean absolute errors of the Nadaraya–Watson (solid lines), empirical L-estimator (dotted lines), and smoothed L-estimator (dashed lines) for sample size $n = 100$, the L-score J_α, and various error distributions. The upper panels contain estimators for $\alpha = 0.25$ and the bottom panels contain estimators for $\alpha = 0.50$.
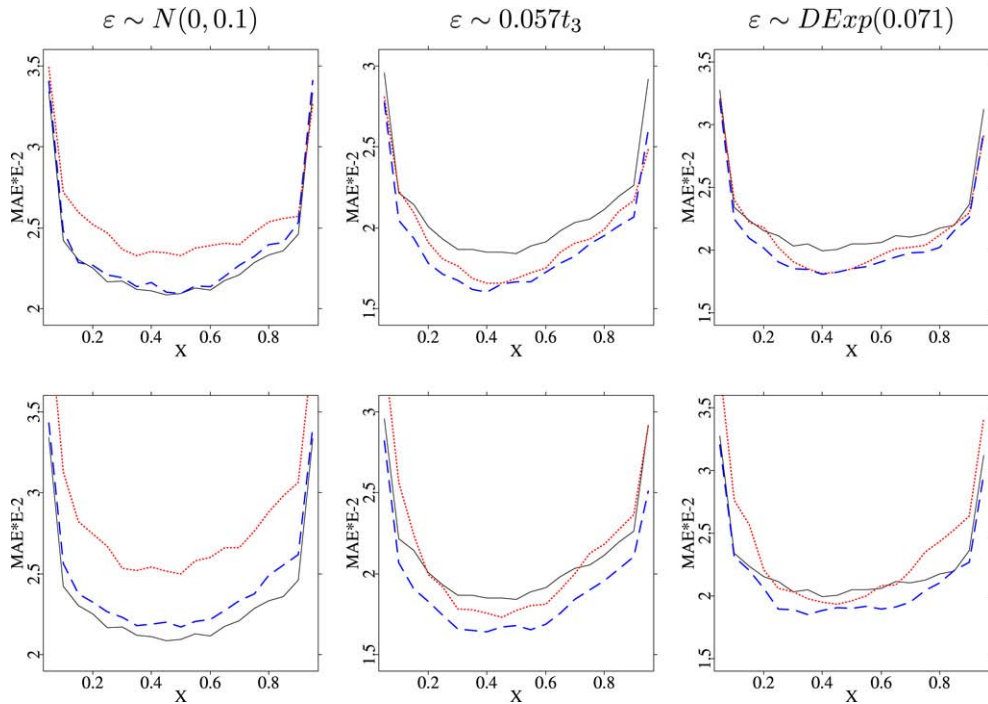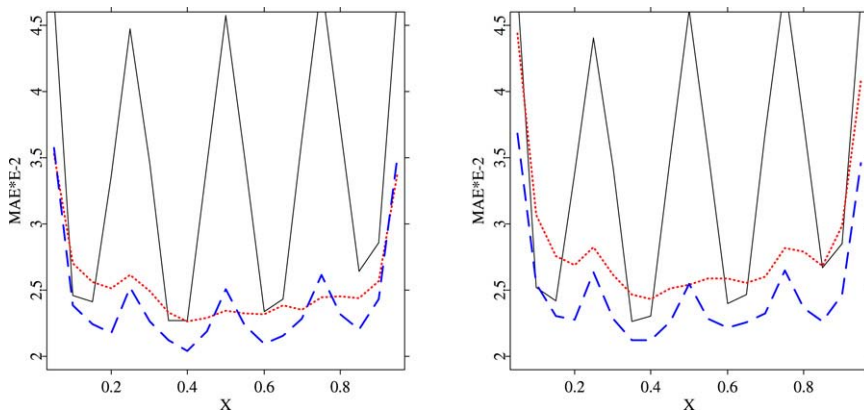

Fig. 3. The mean absolute errors of the Nadaraya–Watson (solid lines), empirical L-estimator (dotted lines), and smoothed L-estimator (dashed lines) for sample size $n = 100$ for the L-score J_α under contamination by 3 outliers. The left panels contain estimators for $\alpha = 0.25$ and the right panels contain estimators for $\alpha = 0.50$.

6. Application

An important application of robust smoothing, such as the local L-estimators, could be in the analysis of volatility, that is, the conditional variance of financial time series (e.g. exchange rates or stock prices). One of approaches, the so-called implied volatility (IV) approach, derives the implied volatility of a stock or a stock index from the currently observed prices of related financial instruments (options) on the market. Having data on individual transactions concerning options, one can estimate IV as a function of their maturity (time to expiry) and their strike price, which results in the so-called IV surface (see Härdle et al. (2002), Chapter 6). The series of such surfaces are then usually analyzed in time to reveal the dynamics of IV and to predict future developments (e.g. Fengler et al. (2003)). Thus, the nonparametric estimators of IV surfaces form an important input of further statistical analysis.

An IV surface on a specified grid is typically estimated by a nonparametric smoothing method such as the NW estimator (e.g. Aït-Sahalia and Lo (1998)); an example is given in Fig. 4 (left panel). Such estimation could benefit from using a robust nonparametric estimator for several reasons: data used to estimate IV are often scarce or of low quality for some
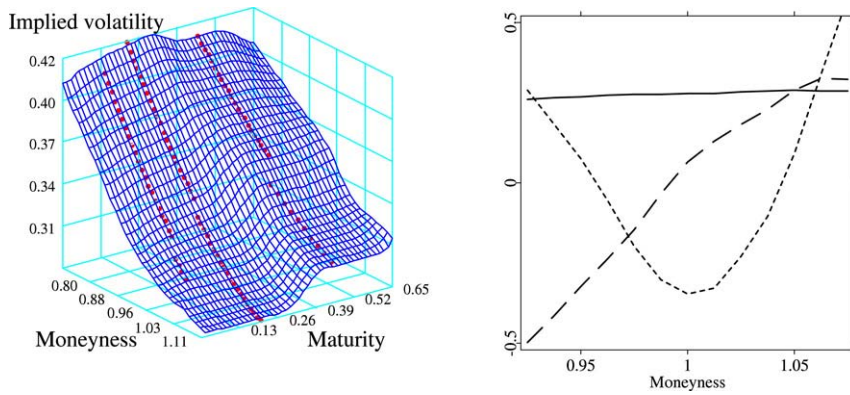
Fig. 4. Example of an implied volatility surface of the DAX index from January 4, 1999, (left panel) and the first three common principle components of DAX implied volatility surfaces (March 2001, to May, 2001).
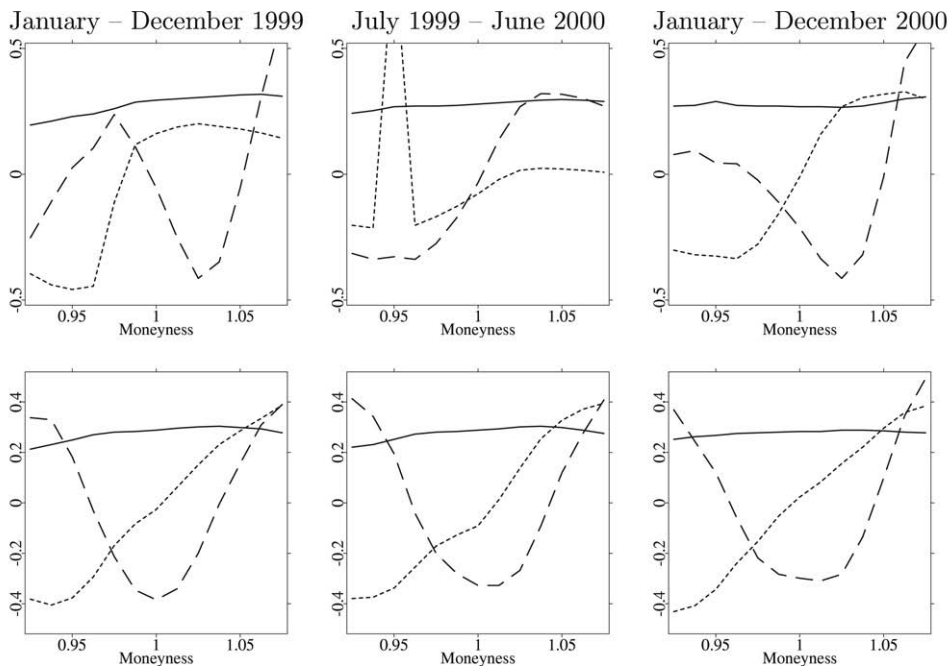

Fig. 5. The first three common principle components of implied volatility surfaces for the DAX index estimated by the Nadaraya–Watson (left panels) and smoothed local L-estimators (right panels) in different (overlapping) time periods.

combinations of maturities and strike prices (data concern individual transactions on the market); the structure of data often enforces oversmoothing (see Fengler et al. (2003)), which increases the impact area of wrong observations (outliers); and transaction data are difficult to check and clear because of their size.

To demonstrate the impact of replacing the NW estimator by the smoothed L-estimator, we estimate the IV surfaces of DAX (a German stock index) from January 1999 to May 2001 both by NW and the smoothed L-estimator with the L-score $J_\alpha, \alpha = 0.25$ (the quartic kernel is used in both cases). Instead of plainly comparing them as in Čížek (2004), we use them in further analysis. More specifically, to reduce the dimension of IV surfaces, several authors analyzed the IV surfaces cut at several maturities as a set of vectors (functions) and tried to find their common principle components (CPC) (e.g. Fengler et al. (2003) and Benko and Härdle (2005)). The shapes of the first three CPC, see the right panel of Fig. 4 for estimates from March to May 2001, are well established and correspond to the main stylised facts concerning IV: they represent the level (solid flat line), skew (dashed sloped line), and turn (U-shaped finely dashed line) of the IV surface.

Now, we replicate the analysis of Fengler et al. (2003) and estimate the CPC of the IV surfaces produced by the NW and smoothed L-estimator for different periods in years 1999 and 2000. Results based on both nonparametric estimators are presented in Fig. 5. The three estimated CPC based on the IV surfaces obtained by the smoothed L-estimator are stable: they are similar to those on Fig. 4 and to each other irrespective of the time period considered. On the other hand, the estimated

CPC based on the NW estimation are somewhat similar to those on Fig. 4 and to those based on the L-smoothing in year 2000, but greatly differ for other one-year subperiods. This documents a high sensitivity of NW to various data errors.

7. Conclusion

Our theoretical and empirical results clearly point out the superiority of the smoothed L-estimator over the empirical L-estimator both for computational and finite-sample properties. Although one might argue that, with an increasing sample size, the difference between the two estimators disappears as suggested by the asymptotic results, it is necessary to keep in mind that this is implied by the first-order asymptotics. Additionally, the computational burden of using the empirical L-estimator becomes very pronounced in large samples. This is where one can benefit from the lower computational demands of the smoothed L-estimator, growing much slower with an increasing sample size than in the case of the empirical L-estimator. The use of the smoothed L-estimator is thus indicated in all cases.

Acknowledgements

References

Aït-Sahalia, Y., Lo, A.W., 1998. Nonparametric estimation of state-price densities implicit in financial assets. Journal of Finance 53, 499–547.
Barnett, V., Lewis, T., 1994. Outliers in Statistical Data. Wiley, New York.
Benko, M., Härdle, W., 2005. Common functional implied volatility analysis. In: Čížek, P., Härdle, W., Weron, R. (Eds.), Statistical Tools for Finance and Insurance. Springer, Heidelberg, pp. 115–134.
Boente, G., Fraiman, R., 1994. Local L-estimators for nonparametric regression under dependence. Journal of Nonparametric Statistics 4, 91–101.
Čížek, P., 2004. Smoothed local L-estimation with an application. In: Hubert, M., Pison, G., Struyf, A., van Aelst, S. (Eds.), Theory and Applications of Recent Robust Methods. Birkhäuser, Basel, pp. 59–70.
Čížek, P., Härdle, W., 2006. Robust estimation of dimension reduction space. Computational Statistics & Data Analysis 51, 545–555.
Čížek, P., Tamine, J., Härdle, W., Smoother L-estimation of regression function. Center discussion paper 2006/20, Tilburg University, 2006.
Davis, P.J., Rabinowitz, P., 1984. Methods of Numerical Integration. Academic Press, New York.
Fengler, M., Härdle, W., Villa, P., 2003. The dynamics of implied volatilities: A common principle components approach. Review of Derivative Research 6, 179–202.
Fernholz, L.T., 1997. Reducing the variance by smoothing. Journal of Statistical Planning and Inference 57, 29–38.
Härdle, W., 1990. Applied Nonparametric Regression. Cambridge University Press, Cambridge.
Härdle, W., Kleinow, T., Stahl, G., 2002. Applied Quantitative Finance. Springer, Heidelberg.
Leung, D., 2005. Cross-validation in nonparametric regression with outliers. The Annals of Statistics 33, 2291–2310.
Leung, D., Marriott, F., Wu, E., 1993. Bandwidth selection in robust smoothing. Journal of Nonparametric Statistics 2, 333–339.
Nadaraya, E.A., 1964. On estimating regression. Theory of Probability and its Applications 10, 186–190.
Pagan, A., Ullah, A., 1999. Nonparametric Econometrics. Cambridge University Press, Cambridge.
Park, B.U., Marron, J.S., 1990. Comparison of data-driven bandwidth selectors. Journal of the American Statistical Association 85, 66–72.
Sakata, S., White, H., 1998. High breakdown point conditional dispersion estimation with application to S&P 500 daily returns volatility. Econometrica 66, 529–567.
Wang, F., Scott, D., 1994. The L_1 method for robust nonparametric regression. Journal of the American Statistical Association 89, 65–76.
Watson, G.S., 1964. Smooth regression analysis. Sankhya, Ser. A 26, 359–372.

Calibration Design of Implied Volatility Surfaces

K. Detlefsen and W. K. Härdle
Humboldt Universität zu Berlin

Abstract: The calibration of option pricing models leads to the minimization of an error functional. We show that its usual specification as a root mean squared error implies prices of exotic options that can change abruptly when plain vanilla options expire. We propose a simple and natural method to overcome these problems, illustrate drawbacks of the usual approach and show advantages of our method. To this end, we calibrate the Heston model to a time series of DAX implied volatility surfaces and then price cliquet options.

Key words: Calibration, data design, implied volatility surface, Heston model, cliquet option.

1. Introduction

A stable calibration of option pricing models is of paramount importance for investment banks because the prices of exotic options are determined by estimated models. The parametric models are normally calibrated to the liquid market of plain vanilla options which can be represented directly by prices or implied volatilities. (The implied volatility corresponding to a price can be obtained by numerically inverting the formula of Black and Scholes (1973)). These prices or volatilities are observed only for certain strikes and maturities. Instead of being uniformly distributed in the strike-maturity region these input option data have a special design. This special data design should be taken into account in the calibration routine in order to avoid that prices of exotic options exhibit jumps because of the design of the input data used for the calibration.

We consider in Section 2 the special design of implied volatility surfaces. In Section 3 we calibrate DAX implied volatility surfaces by a standard approach and identify some problems. In Section 4 we describe our method and show how it deals with the problems of the usual approach. In Section 5 we compare the methods and see pricing differences for long times to maturity. Finally we draw our conclusions in Section 6.

2. Data design

Figure 1 shows the grid on which the EUREX reported the settlement volatil-
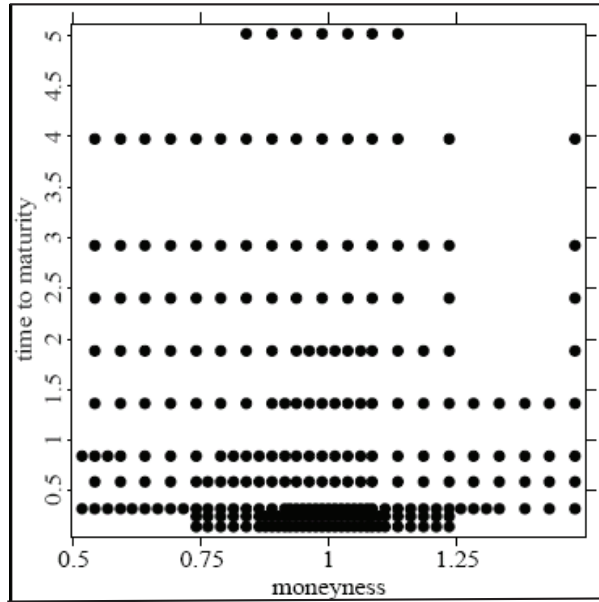ities of European DAX options on March 1st, 2004.

Figure 1: Grid of the DAX implied volatility surface of the EUREX on March
1st, 2004. (only for moneyness between 0.5 and 1.5)

The figure shows that the data have a special design. They come in strings
because only options with certain times to maturity are traded. These strings are
not uniformly distributed in the time to maturity dimension because for short
times there are more strings. The distribution within strings also differs: On
the one hand the moneyness range changes on the other hand the frequency of
the observations within a string differs. Moreover there are missing observations
for out of the money calls. In addition the maximal time to maturity varies for
different underlyings and changes even for one underlying if new products are
issued by the exchange.

Another special feature becomes apparent in time series analysis. The grid
of observations moves in time because on a new day all options have one day
less to maturity than the day before. In this way the grid moves to zero days
to maturity and whole strings disappear when their time to maturity is zero (or
under a prescribed level). Moreover new strings appear in the grid when a new
option with a long time to maturity is issued.

Thus the grid of observations has a special design that changes in time. Both
of these points should be taken care of in calibration methods in order to avoid

wrong or fluctuating prices.

3. The Mean Squared Error

Option pricing models have been calibrated for a long time. Usually the difference between the market and the model quantities is measured by a root mean squared error:

$$\sqrt{\frac{1}{n}\sum_{i=1}^{n}(Q_i^{mod}-Q_i^{mar})^2}$$

where Q denotes the option price or the corresponding implied volatility (in the Black-Scholes model) and mod (mar) refers to a model (market) quantity. (The implied volatility can be computed by numerically inverting the Black-Scholes formula with respect to the volatility parameter.) The index i runs over all observations of the surface that are considered in the calibration. Bakshi, Cao, and Chen (1997) minimize such a functional in their calibration and also Schoutens *et al.* (2004) use this error measure for their analysis. Variations such as relative errors instead of the absolute differences are preferred sometimes.

For independent and identically distributed observations this error measure converges to the L_2-norm. But the data from the EUREX are not independent and identically distributed because of the special design. Hence this error measure lacks this theoretical underpinning. Moreover its economic interpretation is not clear because it is some kind of average error at some points. This leads to two problems: The grid and therefore also the error measure are given by the exchange. The grid and hence the meaning of the error change over time. Preferable would be an error measure that has the same meaning for all surfaces and this meaning should not depend on the grid given by the exchange.

In practice, this problem is sometimes dealt with by disregarding whole strings. Despite its simplicity this method has some drawbacks: It is possible to choose on one day a grid that works well for the analysis. But this subjective approach can hardly be generalized sufficiently for a time series analysis. Moreover disregarding relevant observations cannot lead to a better statistic and often the resulting calibrations are significantly worse.

The above error measure is generally implemented by choosing a region of interest in the moneyness-maturity plane. As we consider here options with $T=1$ or $T=3$ years to maturity we use only observations with expiry between 4 months and 4 years. In the moneyness dimension we consider only option with moneyness between $0.8-0.1T$ and $1.2+0.1T$ where T denotes time to maturity. Such a cone seems to be a good choice for the input data because the longer the time T the higher is the variance of the stock price distribution $\mathcal{L}(S_T)$. We

analyze cliquet options whose strikes lie within this region of interest. In Section 5 we consider the question of wrong prices. Here we show only the fluctuating prices resulting from the movement of the grid.

Table 1: Prices of cliquet options on 24th and 25th February 2004. (max. relative std. error: 0.1%)

		24/02	25/02	25/02 with string
mean squared error	T=1	0.0621	0.0618	0.0620
	T=3	0.0730	0.0749	0.0728
volume error	T=1	0.0612	0.0615	0.0613
	T=3	0.0754	0.0750	0.0750

In table 1 we present the prices of cliquet options in the Heston model. These cliquet options and all others considered here have local floors (caps) of $f_l^i = -8\%$ ($c_l^i = 8\%$) for the periods $t_i = iT/3$, $i = 1, 2, 3$ and a global floor of zero. The payoff H of such cliquet options can be described mathematically by

$$H = \max \left[0, \sum_{i=1}^{N} \min \left\{ c_l^i, \max \left(f_l^i, \frac{S_{t_i} - S_{t_{i-1}}}{S_{t_{i-1}}} \right) \right\} \right].$$

The model has been calibrated w.r.t. the quadratic errors of implied volatilities by a global optimization routine. To this end, we used differential evolution by Storn and Price (1997), a simple population based, stochastic function minimizer. The option prices have been computed afterwards from the calibrated parameters by Monte Carlo simulations using 250 time steps a year and 1000000 simulations (max. relative std. error: 0.1%).

The table shows the prices on two subsequent days. On the second day a string drops below the maturity limit of 4 months and is not considered anymore by the above error measure. The table also gives the prices that would have been calculated if the string were still considered in the calibration. It becomes clear from the table that the prices on the first day and on the second day with the string still considered are quite close for both times to maturity. But the prices on second day without the string differ clearly from the other prices for the long time to maturity. The price of these cliquet options is about 2.5% higher. Thus these price differences are significant compared to the usual profit margin. The prices of the cliquet options with 1 year to maturity do not fluctuate when the string disappears.

Similar problems occur when a new string appears in the grid. If no region of interest is specified the same problems arise because new strings still appear and the strings of expired options disappear.

4. The Volume Error

Although the input data for the calibration are observed only for a few maturities the calibrated model is normally used for all times to maturity to price exotic options. Hence we want to measure the difference between the market and the model surface for all times to maturity and not only for a few maturities. This leads us to consider the volume between the two surfaces. This volume can be interpreted economically as a generalized sum of the price differences (between model and market) of all options in a region. This approach resolves the possibly wrong pricing because the weight is distributed (more) uniformly over all times to maturity. We consider again a region of interest in the moneyness-maturity plane and compute the volume always over this area. To this end, the error surface which is the squared difference between the market and the model surface has to be approximated. As the error measures should have the dimension price or implied volatility we divide the volume by the area of the region of interest.

Hence, we compute first the observed points of the error surface by

$$e_i \stackrel{\text{def}}{=} (Q_i^{mod} - Q_i^{mar})^2,$$

where the index i runs over all observations in the area of interest and Q denotes prices or implied volatilities. Then we approximate the volume V_e under the error surface based on the points e_i $(i = 1, \ldots, n)$ and finally we compute the error measure ϵ by

$$\epsilon \stackrel{\text{def}}{=} \sqrt{V_e/A}$$

where A denotes the area of the region of interest. This error represents the average error of all options in the considered region.

There are many ways to approximate the volume of the error surface. Non- or semiparametric methods (see e.g. Härdle *et al.*, 2004) give a good fit if there are enough data points. But in general implied volatility surfaces do not have enough data points for this approach (especially for long times to maturity). Moreover smoothing is relatively time consuming for computations and depends a lot on the smoothing parameter.

Another approach is to construct points of the implied volatility surface on a stable uniform grid and then to approximate the error in the usual way. This method depends critically on the constructed points and so far there is no generally accepted construction procedure. Arbitrage and especially calendar arbitrage make such a construction difficult (see e.g. Kahalé, 2004). Moreover a fine grid increases the calibration time significantly. Although this approach has advantages for some models (e.g. local volatility) our simpler method leads to similar results

and requires less computation time. In addition it has the precise interpretation as mean height of the volume.

We consider as above the moneyness-maturity region

$$\{(m, T) \ : \ 1/3 \leq T \leq 4 \, , 0.8 - 0.1T \leq m \leq 1.2 + 0.1T\}.$$

For this simple region of interest we use some linear interpolation in the maturity direction for calculating the volume. We compute the volume separately between the maturities. To this end, we first construct for each maturity boundary values of the error string by extrapolation. Then we approximate the area under this error string for each maturity. Afterwards, mean heights can be computed for these areas so that the area under the error string can be represented by an equivalent rectangle. Now the volume between two adjacent maturities can be approximated by the volume between the corresponding rectangles and this volume can be calculated exactly from elementary solids. For the first and last maturity we consider only the part of the solid that lies over the region of interest (by linearly interpolating the heights).

The table 1 shows the prices of cliquet options when the model is calibrated w.r.t. the volume error. The prices of the cliquets with 1 year to maturity are quite similar to the corresponding prices calculated from the parameters of the other calibration. But the prices of the cliquet with 3 years to maturity differ significantly to the prices of the other calibration. Moreover the prices do not fluctuate when an implied volatility string disappears. Hence the volume error eliminates the fluctuating prices. The price differences to the usual approach are analyzed in the next section.

5. Pricing Exotic Options

We have seen that the prices of cliquet options jump when an implied volatility string disappears and the mean squared error is used for calibration.

The error that describes the goodness of fit jumps at the same time. This is shown in figure 2 that reports the errors for 20 trading days. After 10 trading days an implied volatility string falls out of the region of interest. The blue line jumps exactly at this point of time. As the blue line represents the root mean squared error the figure shows that these errors jump. Thus this error measure is hard to interpret and does not give reliable results. The volume error which is shown as red line shows no jumps and hence has the same interpretation for different surfaces or days.
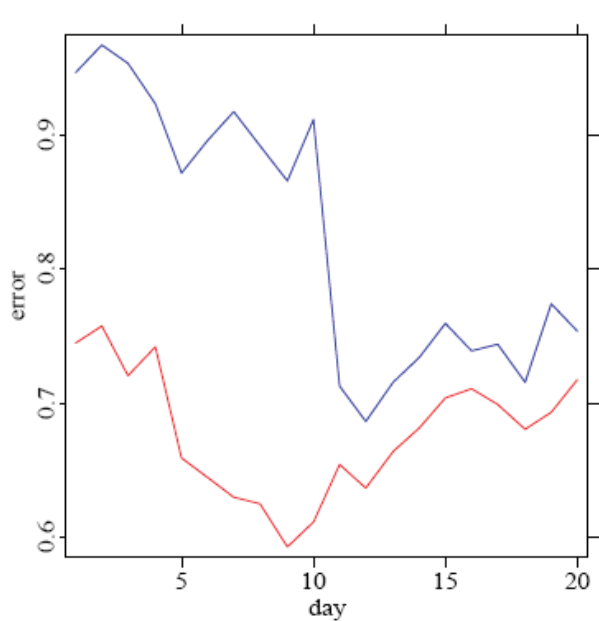
Figure 2: Calibration errors from 11/02/2004 to 09/03/2004. (blue: mean squared error, red: volume error)

When the model is calibrated w.r.t. the mean squared error the prices and the errors jump when a volatility string expires. Figure 3 shows the prices of cliquet options on 10 trading days before and after the expiration of a string. The red line describes the prices for the calibration w.r.t. the volume error. These prices show a stable downward trend with small jumps that could be attributed to changes of the input data (spot, etc). On the other hand the prices from the calibration w.r.t. the mean squared error jump. After the jump the two time series of prices are similar. But before the jump they differ about 2.5%. As the prices are similar for cliquets with small times to maturity and also for cliquets with long times to maturity after jumps the price difference before the jump can be seen as wrong prices.
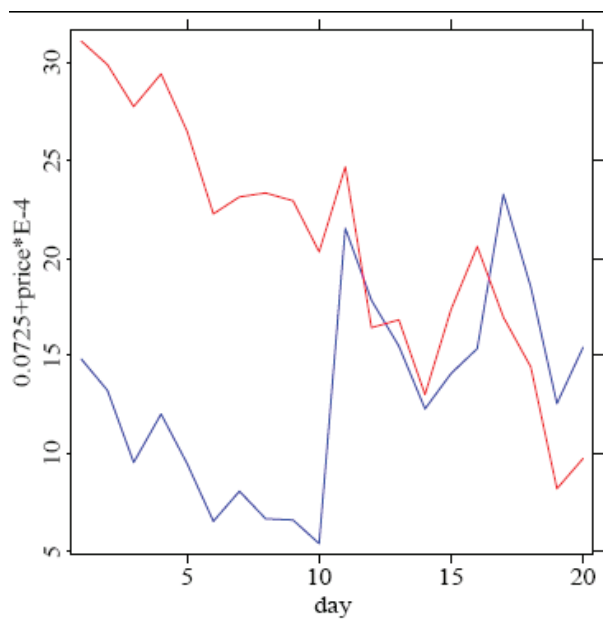
Figure 3: Prices of cliquet options from 11/02/2004 to 09/03/2004. (blue: mean squared error, red: volume error

Thus these plots show how prices jump and where possibly wrong prices appear. These two observations correspond to the movement of the strings and the special data design.

6. Conclusion

We have analyzed different calibrations of the Heston model and their implications for the pricing of exotic options. The model has been calibrated by minimization of different error functionals.

We have shown that the usual specification of this error measure as root mean squared error leads to fluctuating prices of exotic options. Moreover, the usual error has the shortcomings that it has no direct economic interpretation and this meaning is also changing over time. These problems stem from the special design of implied volatility surfaces.

In order to take care of this data design we have proposed an error measure based on the volume between the market and the model surface. Calibration w.r.t. this error gives rise to continuous changes in the price and the error itself. Thus the fluctuation have been eliminated. Moreover this measure seems to correct wrong prices before expiration of implied volatility strings. In addition this error measure has the same interpretation as mean height of the volume between the model and the market surface for different days or surfaces.

Our approach can easily be generalized by weighting the errors observed between the market and the model. Such weighting by a 2 dimensional density function can be seen as a generalization of specifying a region of interest in the moneyness-maturity plane. The approach described in Section 4 puts more weight on long maturities because we consider a wider moneyness range for long times to maturity. Such a weighting makes sense for pricing options expiring after a long time.

Another approach for measuring the error is based on the interpolation of the observed implied volatility surface. This method leads to similar results but suffers from two drawbacks: There is no unique or generally accepted way for arbitrage free interpolation and the interpolated observations slow down calibration routines significantly. Moreover it leads to similar results as our approach.

Stable calibration routines are important for equity models because they imply prices of exotic options that are continuous in time. Such prices are the basis for trading exotic options. But stable parameters are even more important for calculating the greeks that are used by traders for hedging. Hence the stable and fast calibration method proposed helps reducing losses in derivatives trading resulting.

Acknowledgement

References

Bakshi, G., Cao, C. and Chen, Z. (1997). Empirical Performance of Alternative Option Pricing Models. *The Journal of Finance* **5**, 2003-2049.

Black, F. and Scholes, M. (1973). The Pricing of Options and Corporate Liabilities. *Journal of Political Economy* **81**, 637-659.

Heston, S. (1993). A closed-form solution for options with stochastic volatility with applications to bond and currency options. *Review of Financial Studies* **6**, 327-343.

Härdle, W., Müller, M., Sperlich, S. and Werwatz, A. (2004). *Nonparametric and Semiparametric Models*. Springer.

Kahalé, N. (2004). An arbitrage-free interpolation of volatilities. *Risk Magazine* **17**, 102-106.

Schoutens, W., Simons, E. and Tistaert, J. (2004). A perfect calibration! Now what? *Wilmott magazine*, March 2004.

Storn, R. and K. Price (1997). Differential evolution - a simple and efficient heuristic for global optimization over continuous space. *Journal of Global Optimization* **11**, 341-359.

K. Detlefsen
CASE - Center for Applied Statistics and Economics
Humboldt Universität zu Berlin
Wirtschaftswissenschaftliche Fakultät
Spandauer Strasse 1, 10178 Berlin, Germany
detlefsen@wiwi.hu-berlin.de

W. K. Härdle
CASE - Center for Applied Statistics and Economics
Humboldt Universität zu Berlin
Wirtschaftswissenschaftliche Fakultät
Spandauer Strasse 1, 10178 Berlin, Germany
haerdle@wiwi.hu-berlin.de

Long Memory Persistence in the Factor of Implied Volatility Dynamics

Wolfgang Karl Härdle
Center for Applied Statistics and Economics
Humboldt-Universität zu Berlin, Germany
Tel: +49-30-2093-5654; Fax: +49-30-2093-5649

Julius Mungo
Center for Applied Statistics and Economics
Humboldt-Universität zu Berlin, Germany
E-mail: mungo@wiwi.hu-berlin.de
Tel: +49-30-2093-5654; Fax: +49-30-2093-5649

Abstract

The volatility implied by observed market prices as a function of the strike and time to maturity form an Implied Volatility Surface (IVS). Practical applications require reducing the dimension and characterize its dynamics through a small number of factors. Such dimension reduction is summarized by a Dynamic Semiparametric Factor Model (DSFM) that characterizes the IVS itself and their movements across time by a multivariate time series of factor loadings. This paper focuses on investigating long range dependence in the factor loadings series. Our result reveals that shocks to volatility persist for a very long time, affecting significantly stock prices. For appropriate representation of the series dynamics, we model the long memory in levels and absolute returns using the class of fractional integrated volatility models that provide flexible structure to capture nonlinear dynamics and the slow decaying autocorrelation function reasonably well. Our empirical results highlight the significance of long-range dependence in the factor of implied volatility. This would be useful to regulators, derivative market participants and practitioners whose interest is to reasonably forecast stock market movements as well as application to risk management.

Keywords: ImpliedVolatility, Dynamic Semiparametric Factor Modelling, Long Memory,
Fractional Integrated Volatility Models
JEL Classification Codes: C14, C32, C52, C53, G12

1. Introduction

As a measure of the standard deviation of the daily range of price movements, volatility is an important determinant of the riskiness of an asset, a crucial parameter in derivative pricing such as options. Findings across several asset markets have reported high persistence of volatility shocks and that over sufficiently long periods of time, volatility is typically stationary with "mean reverting" behavior, Bollerslev and Jubinski (1999). Such series are characterized by distinct but non-periodic cyclical patterns and their behavior is such that current values are not only influenced by immediate past values but values from previous time periods, allowing for persistence or long memory. Long memory

describes the correlation structure of a series at long lags. It is well known that the volatility implied by observed option prices as a function of the strike and time to maturity form an IVS.For each day the IVS forms a high dimensional object that has unknown stochastic behavior that needs to be analyzed. For practical applications such as in risk management, it is desirable to reduce the dimension of this object and characterize its dynamics through a small number of factors. Such dimension reduction may be summarized by a Dynamic Semiparametric Factor Models that characterize the IVS and their movements across time by a multivariate time series of factor loadings, Borak et al. (2005), Fengler et al. (2007) and Borak et al. (2007).

The DSFM approximate the implied volatility surface by regressing log-implied volatility on a two-dimensional covariate containing moneyness and time-to-maturity.

To introduce this model, denote by $Y_{t,j} = log\{\hat\sigma^2_{t,j}(\kappa,\tau)\}$, the log-implied volatility where t = 1,…,J is an index of time, in this case the number of the day, and j = 1,…, J_t is the number of IV observations on day t.

Let $X_{t,j} = (\kappa_{t,j}, \tau_{t,j})$ be a two-dimensional covariate where $K_{t,j}$ is a moneyness matrix and $\tau_{t,j}$ denotes time-to-maturity. Moneyness is defined as $\kappa_{t,j} = \dfrac{K_{t,j}}{F_{t,j}}$ where $K_{t,j}$ is a strike and $F_{t,j} = S_t e^{(r_{t,j} \cdot \tau_{t,j})}$ the underlying futures price belonging to the option trade (t, j). The model is expressed as:

$$Y_{t,j} = \sum_{k=0}^{K} z_{t,k} m_k(X_{t,j}) + \varepsilon_{t,j} \tag{1.1}$$

where $z_{t,0} = 1$, m_k are smooth basis functions (k=0,…,K) and $z_{t,k}$ are time dependent weights or factor loadings. This model can be seen as a regression model with inherent time evolution. The estimation involves nonparametric kernel regression method with no assumption on the functional form of the factors except smooth basis function. The DSFM simultaneously estimate the factor functions and fits the surface. The IVS is assumed to be a weighted sum of the smooth functional factors, m_k and its dynamics is explained by the stochastic behavior of the factor loadings $z_{t,k}$.

Approximations of the factor loadings are obtained by fitting model (1.1) to the implied volatility observations and the functions m_k are estimated by orthogonal series estimators so that they have zero correlations among each other, Borak et al. (2005). The factor loadings $z_t = (z_{t,1}, …., z_{t,K})^T$ form an unobserved multivariate time series. The estimates $\hat z_{t,k}$ and $\hat m_k$ are obtained in (1.1) as minimizers of the following least squares criterion:

$$\sum_{t=1}^{I} \sum_{j=1}^{J_t} \int \left\{ Y_{t,j} - \sum_{k=0}^{K} \hat z_{t,k} \hat m_k \right\}^2 K_h(u - X_{t,j}) du, \tag{1.2}$$

K_h denotes a two-dimension kernel function, chosen as a product of one dimensional kernels $K_h(u) = k_{h1}(u1) \times k_{h2}(u2)$ where $h = (h1, h2)^T$ are bandwidths and $k_h(v) = k(h^{-1}v)/h$ is a one-dimensional kernel function. The estimation procedure can be seen as a combination of functional principal component analysis, nonparametric curve estimation and backfitting for additive models. The choice for the number of dynamic basis function (K) is based on calculating the residual sum of squares per total variance,

$$RV(K) = \frac{\sum_t \sum_j \left\{ Y_{t,j} - \sum_{k=0}^{K} \hat z_{t,k} \hat m_k(X_{t,j}) \right\}^2}{\sum_t \sum_j (Y_{t,j} - \bar Y)^2}, \tag{1.3}$$

$\bar Y$ is the overall mean of the observations. K is chosen such that the variance explained, represented by 1−RV (K) is sufficiently high to give a good approximation to the IVS.

In our application, K = 3 is chosen with 97% of the variation explained. Details on the procedure are presented in Borak et al. (2005), Fengler et al. (2007) and Borak et al. (2007).

This paper applies the DSFM on the German DAX index market from 04.01.1999 to 25.02.2003. Figure 1 displays the implied volatility surface from the DSFM fit for the DAX-Option on 2 May 2000, with moneyness between 0.8 and 1.12 and time to maturity between 0 and 0.5 years. Figure 2 shows three volatility-driving factors that could be interpreted in terms of level, slope and curvature factor. z1 governs movements in the general level, z2 is largely associated with changes in the slope and z3 is closely related to dynamic changes in the curvature of the IVS.

Figure 1: Implied volatility surface from DSFM fit for the DAX-Option on 2 May 2000, with moneyness between 0.8 and 1.12 and time to maturity between 0 and 0.5 years.
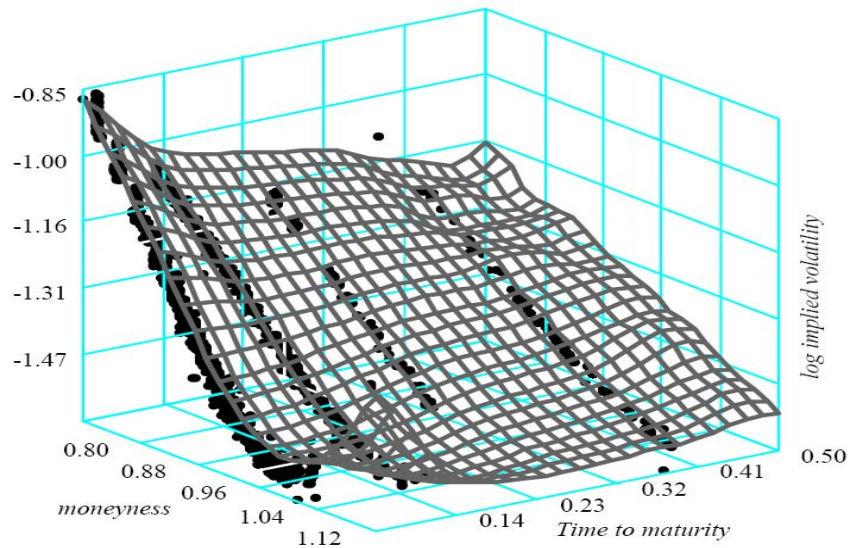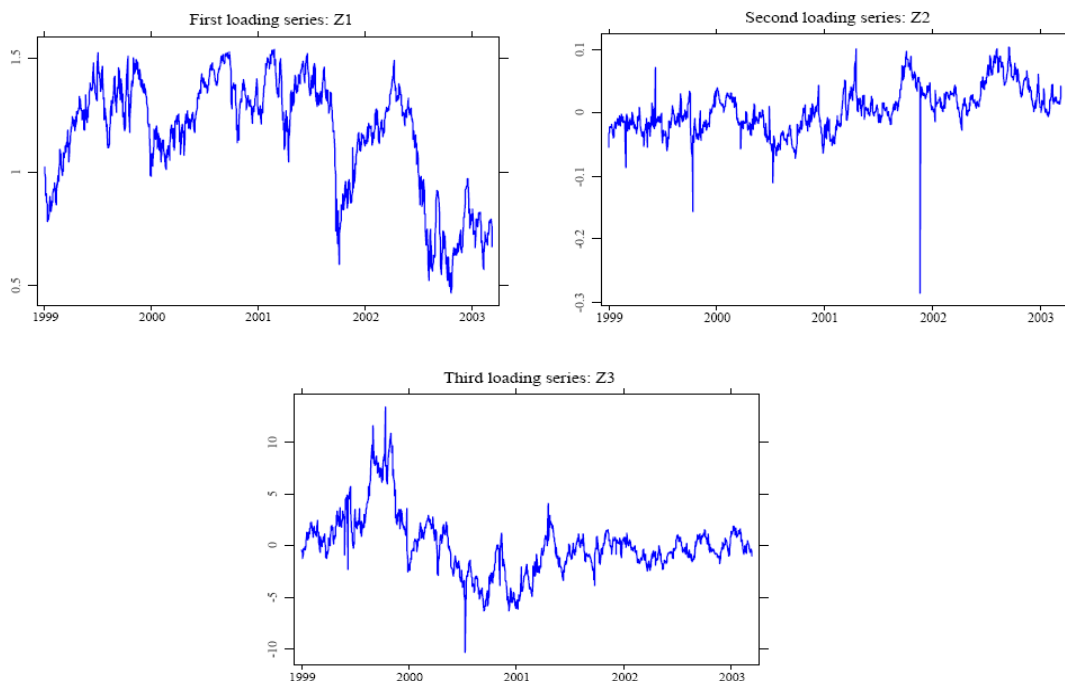

Figure 2: Time series plots in levels of three loading series from a DSFM fit for the DAX-Option from 04.01.1999 − 25.02.2003.

The aim of this paper is to investigate dependence in the factor loadings of implied volatility strings because information on persistence can guide the search for economic explanation of the movements in asset returns as well as in risk management applications. Several research involving the autocorrelation functions of various volatility measures (squared, log-squared and absolute returns) have reported decay at a very slow mean-reverting hyperbolic rate, Ding et al. (1993), Bollerslev and Wright (2000) and Sibbertsen (2004). Our analysis follows this line of research on long range dependence investigation and modelling.

First, we consider model independent tests for stationarity, I(0) against fractional alternative I(d). We apply the rescaled variance test (V/S) of Giraitis et al. (1999) that uses the heteroscedastic and autocorrelation consistent (HAC) estimator of the variance, Newey and West (1987) for normalization and the semiparametric (LobRob) test of Lobato and Robinson (1998), that does not depend on a specific parametric form of the spectrum in the neighborhood of the zero frequency. We also apply the log-periodogram regression estimator (GPH) of Geweke and Porter-Hudak (1983) and the Gaussian Semiparametric estimator (GSP) of Robinson (1995a) in estimating the degree of long memory in the factor loadings series. Results are indicative of long-range dependence in the factor loadings series in levels and absolute returns.

The first factor, z1 can be interpreted as highly persistent and influences all options similarly, irrespective of maturity. The second factor z2 gradually diminishes for longer maturities and the third factor z3 governs large volatility changes in relatively short maturities.

Second, for appropriate representation of the series dynamics and the possibility of improved forecasting, we model long memory in volatility using the ARFIMA, FIGARCH and HYGARCH models. These models provide flexible structure that captures slow decaying autocorrelation reasonably well. In comparison, models in absolute returns have better performance, confirming the findings of Ding et al. (1993), that absolute returns are most appropriate indicator to represent the long memory volatility processes. Our results imply that shocks to the volatility will persist for long time, affecting the DAX stock prices significantly. This confirms evidence for long memory in volatilities of German stock returns, Sibbertsen (2004). Such dependence or persistence will have importance economic consequences for short-term trading and long range investment strategies. Better option pricing may results from models that price and hedge derivative securities when there is prior information on long-memory volatility in terms of expectation on the potential level of volatility and the rate at which volatility changes.

In the presence of long memory, Granger and Joyeux (1980), Geweke and Porter-Hudak (1983) have shown the possibility for improved price forecasting performance within a linear time series framework than with traditional procedures. Option pricing have also been shown to be significantly different when standard models are applied as compared to models allowing for long memory. By applying the GARCH, EGARCH, FIEGARCH and EGARCH models, Bollerslev and Mikkelsen (1996) have shown that the price of an option increases with the degree of integration. This means that GARCH models give the lowest price whereas the highest option price is obtained for the IGARCH model. For long memory alternative, Herzberg and Sibbertsen (2004) have shown that prices for the FIGARCH, HYGARCH are in between the GARCH and IGARCH prices.

In addition to documented studies of the economic implications of long memory, Cheung and Lai (1995), Wilson and Okunev (1999), revealed that portfolio diversification decisions in the case of strategic asset allocation may become extremely sensitive to the investment horizon. There may be diversification benefits in the short and medium term, but not if the assets are held together over the long term if long memory is present. e.g., in a market that exhibits antipersistence, asset prices tend to reverse its trend in the short term thus creating short-term trading opportunities. In addition Mandelbrot (1971) has shown that in the presence of long memory the arrival of new market information cannot be fully arbitraged away. It is also known that the possibility of speculative profits as a result of superior long range dependence model forecast would cast doubt on the basic tenets of market efficiency.

Motivated by evidence of long range dependence in the factor loadings levels and absolute returns, we perform estimation and prediction using the ARFIMA, FIGARCH and HYGARCH models

that are known to provide flexible structure to capture slow decaying autocorrelation reasonably well than with traditional ARMA procedures. The rest of our work is structured as follows.

Section 2, introduces fractional integration and Long-memory processes. Here we examine some methodology for testing and estimating long range dependence that we apply in our analysis. Section 3 introduces the structure of the class of models we apply to analyze the long memory in the factor loading series. In section 4 we report and discuss our results for the series in levels and absolute returns. A summary of our analysis results and conclusions is given in section 5.

2. Fractional Integration and Long-Memory

The framework of fractional integration yields convenient modelling of long range dependence, Granger and Joyeux (1980), Baillie (1996). A time series process z_t is integrated of order d, I(d) if

$$(1-L)^d z_t = \varepsilon_t \tag{2.1}$$

where ε_t I(0) and L is the lag operator ($Lz_t = z_{t-1}$). The non-integer parameter d is the difference parameter and $(1-L)^d$ is the fractional filter defined by its binomial expansion, $(1-L)^d = \sum_{j=0}^{\infty} \frac{\Gamma(j-d)}{\Gamma(-d)\Gamma(j+1)} L^j$ where $\Gamma(z) = \int_0^{\infty} t^{z-1} e^{-t} dt$ is the gamma function. The autocorrelation function of such a series is given by

$$\rho_k = \frac{\Gamma(1-d)\Gamma(k+d)}{\Gamma(d)\Gamma(k+1-d)} \sim C k^{2d-1} \tag{2.2}$$

for d in the range of (0, 0.5), where C is a strictly positive constant. In such case z_t is said to exhibit long memory. For 0<d<0.5, the series is stationary. For d=0, the series is an I(0) process and said to have no long-memory. For 0.5<d<1 the process is mean reverting as there is no long run impact of an innovation to future values of the process. In the case where 0<d<1, not only the immediate past value of z_t influence the current value, but also values from previous time periods as well. The sum over the autocorrelation does not converge, so that it is a suitable model for long memory, Granger and Joyeux (1980).

2.1. Tests and estimators of long memory processes

We consider two model independent tests for stationarity, I(0) against fractional alternatives I(d). The tests include the rescaled variance test (V/S) of Giraitis et al. (1999) that uses the heteroscedastic and autocorrelation consistent (HAC) estimator of the variance, Newey and West 81987) for normalization and the semiparametric (LobRob) test of Lobato and Robinson (1998), that does not depend on a specific parametric form of the spectrum in the neighborhood of the zero frequency. The Rescaled Variance test is applied by centering the KPSS statistic based on the partial sum of the deviations from the mean:

$$V/S = \frac{1}{T^2 \hat{\sigma}_T^2(q)} \left[\sum_{k=1}^{T} \left\{ \sum_{j=1}^{k} (z_j - \overline{z}_T) \right\}^2 - \frac{1}{T} \left\{ \sum_{k=1}^{T} \sum_{j=1}^{k} (z_j - \overline{z}_T) \right\}^2 \right] \tag{2.3}$$

where $S_k = \sum_{j=1}^{k} (z_j - \overline{z}_T)$ are artial sums of the observations and $\hat{\sigma}_T^2(q) = \hat{\gamma}_0 + 2 \sum_{j=1}^{q} \left(1 - \frac{j}{1+q}\right) \hat{\gamma}_j$ is the heteroscedastic and autocorrelation consistent (HAC) estimator of the variance, (q < T). $\hat{\gamma}_0$ is the variance of the process, and the sequence $\{\hat{\gamma}_j\}_{j=1}^{q}$ denotes the autocovariances of the process up to the order q. Giraitis, Kokoszka and Leipus (2000) showed that this statistic can detect long range dependence in the volatility for the class of ARCH(1) processes.

The Semiparametric test is based on the approximation of the spectrum of a long memory process. This test allows one to discriminate between d > 0 and d < 0.

In the univariate case the test statistic for I(0) against I(d) is given by

$$t_{LobRob} = \sqrt{(m)}\,\frac{\hat{C}_1}{\hat{C}_0} \tag{2.4}$$

with $\hat{C}_k = \frac{1}{m}\sum_{j=1}^{m}\varsigma_j^k I(\lambda_j)$ and $\varsigma_j = \log(j) - 1/m\sum_{i=1}^{m}\log(i)$, $I(\lambda) = \frac{1}{2\pi T}\left|\sum_{t=1}^{T} z_t e^{it\lambda}\right|^2$, $\left(i = \sqrt{-1}\right)$ is the

estimated periodogram. $\lambda_j = \frac{2\pi j}{T}, j = 1,...,m << [T/2]$ is a degenerate band of Fourier frequencies with

bandwidth parameter m. Under the null hypothesis the test statistic is asymptotically normally distributed. If the statistic is in the lower fractile of the standardized normal distribution, the series exhibit long-memory whilst if in the upper fractile of that distribution, the series is antipersistent.

To estimate the memory parameter d, we apply two frequently used estimators, the log-periodogram regression estimator (GPH) of Geweke and Porter-Hudak (1983) and the Gaussian Semiparametric estimator (GSP) of Robinson (1995a).

The **log-periodogram regression estimator** is based on the periodogram of a time series z_t (t =1, …,T) defined by

$$I(\lambda_j) = \frac{1}{2\pi T}\left|\sum_{t=1}^{T} z_t e^{it\lambda}\right|^2 \tag{2.5}$$

where $\lambda_j = \frac{2\pi j}{T}, j = 1,...,m$ (m is a positive integer). The memory parameter d is estimated from a linear

regression of the log (λ_j) on a constant and variable $X_j = \log\{4\sin^2(\lambda_j/2)\}$:

$$\hat{d}_{GPH} = \frac{\sum_{j=1}^{m}(X_j - \bar{X})\log\{I(\lambda j)\}}{2\sum_{j=1}^{m}(X_j - \bar{X})} \tag{2.6}$$

We consider only harmonic frequencies $\lambda_j = \frac{2\pi j}{T}$ (the jth Fourier frequency), $j \in (l,m)$, where l is a trimming parameter discarding the lowest frequencies and m is a bandwidth parameter. The cut-off parameter ensures robustness of the estimator. For the Gaussian case with $d \in (-0.5, 0.5)$, the estimator

is consistent and asymptotically normal with standard error of $\frac{\pi}{\sqrt{24m}}$ Robinson (1995b).

Validity of the GPH estimator for an enlarged interval has been demonstrated by Velasco (1999). More precisely, he shows that in the interval [0.5, 0.75), where the time series is nonstationary, asymptotic Normality and consistency is preserved as in the original interval (−0.5, 0.5), while for values of d in the interval [0.75, 1) the estimator is still consistent. Deo and Hurvich (2001) have shown that this estimator is also valid for some non-Gaussian time series.

The **Gaussian Semiparametric estimator** is based on the approximation, $\lim_{\lambda_i \to 0^+} f(\lambda_i) = C\,\lambda_i^{-2d}$ of a long memory process in the Whittle approximate maximum likelihood

estimator, $L_w(\theta)$. For $m^* = \left[\frac{T}{2}\right]$ an approximation to the Gaussian likelihood, Beran (1994) is given

by

$$L_w(\theta) = -\frac{1}{2\pi}\sum_{j=1}^{m^*}\log f_\theta(\lambda_j) + \frac{I_T(\lambda_j)}{f_\theta(\lambda_j)} \tag{2.7}$$

for a given parametric spectral density $f_\theta(\lambda)$. Estimating d is by solving the minimization,,

$$\arg\min_{C,d} L(C,d) = \frac{1}{m}\sum_{j=1}^{m*}\left\{\log(C\lambda_j^{-2d}) + \frac{I(\lambda_j)}{C\lambda_j^{-2d}}\right\}$$ $I(\lambda_j)$ is the periodogram evaluated for a

degenerated range of m harmonic frequencies $\lambda_j = \frac{2\pi j}{T}, j = 1,..., m \ll [T/2]$, j = 1,. .. ,m << [T/2],

where [.] represents the integer part operator, bounded by the bandwidth parameter m, which increases

with the sample size T but more slowly. This bandwidth m must satisfy $\frac{1}{m} + \frac{m}{T} \to 0$ as $T \to \infty$. If

$m = \left[\frac{T}{2}\right]$ this estimator is a Gaussian estimator for the parametric model $f(\lambda) = C\lambda^{-2d}$ An estimator

for d is given by

$$\hat{d}_{GSP} = \arg\min_d\left\{\log\left(\frac{1}{m}\sum_{j=1}^{m}\frac{I(\lambda_j)}{C\lambda_j^{-2d}}\right) - \frac{2d}{m}\sum_{j=1}^{m}\log(\lambda_j)\right\} \tag{2.8}$$

Robinson (1995a) showed that $\sqrt{m}\left(\hat{d}_{GSP} - d\right)\xrightarrow{d} N\left(0, \frac{1}{4}\right)$ and is valid in the presence of

some form of conditional heteroscedasticity, (Robinson and Henry 1999). Phillips and Shimotsu (2004) have shown that the ranges of consistency and asymptotic normality for the model type in (2.1) are the same as those of the GPH estimator.

3. Long Memory Models

Several studies have dealt with models that provide useful ways of analyzing the relationships between the conditional mean and variance of a process exhibiting long memory and slow decay in its levels. In our analysis we consider an autoregressive fractional integrated moving average (ARFIMA) process model, Granger and Joyeux (1980), Hosking (1981), the (FIGARCH), Baillie et al. (1996), a combination of fractional integrated process for the mean with regular GARCH process for conditional variance and the Hyperbolic GARCH (HYGARCH), Davidson (2004).

The **ARFIMA(p,d,q)** model represented as

$$\Phi(L)(1-L)^d(z_t - \mu) = \Theta(L)\varepsilon_t \tag{3.1}$$

$\varepsilon_t \sim i.i.d(0,\sigma_\varepsilon^2)$ extend the integration order of the conventional ARMA model to a on-integer value

between 0 and 1. $\Phi(L) = 1 - \sum_{i=1}^{p}\phi_i L^i$ and $\Theta(L) = 1 + \sum_{j=1}^{q}\phi_j Lj$ are the autoregressive and moving

average polynomials in the lag operator L respectively. d is the long memory parameter and $(1-L)^d$ is the fractional difference filter as defined in equation (2.1). The ARFIMA process displays persistence for 0<d<0.5 and anti-persistence for −0.5 < d < 0. For |d| > 0.5, the process is non-stationary as it has finite variance. In our application, the values for p and q are chosen such that the ordered pair (p, q) minimizes the AIC criterion. We estimate the model parameters by maximum likelihood approach, see Doornik and Ooms (2004) that allows for (break-) regressors in the mean and structural changes in the variance, and by non-linear least squares estimation method of Beran (1994) that is asymptotically efficient in the presence of GARCH errors

The **FIGARCH(p,δ,q)** model of (Baillie et al. 1996) given as

$$\Phi(L)(1-L)^\delta \varepsilon_t^2 = \omega + \Theta(L)\nu_t \tag{3.2}$$

where $\nu_t = \varepsilon_t^2 - \sigma_t^2$ combine the fractional integrated process for the mean with regular GARCH process for the conditional variance. The conditional variance can be represented as

$$\sigma_t^2 = \frac{\omega}{1-\theta(L)} + \left[1 - \frac{\phi(L)(1-L)^\delta}{1-\theta(L)}\right]\varepsilon_t^2 \tag{3.3}$$

with $0 \leq \delta \leq 1$. The δ in FIGARCH does not have the same interpretation of persistence as d in ARFIMA. The fractional differencing operator in the ARFIMA model applies to the constant term in the mean equation while in FIGARCH it does not apply to ω in the variance equation.

We base our analysis on the FIGARCH parameterization proposed by (Chung 1999)

$$\Phi(L)(1-L)^{\delta} \varepsilon_t^2 = \Theta(L)(\varepsilon_t^2 - \sigma^2) \tag{3.4}$$

where σ^2 is the unconditional variance of ε_t. The conditional variance is formulated as

$$\sigma_t^2 = \sigma^2 + \left[1 - \frac{\phi(L)(1-L)^{\delta}}{1-\theta(L)} \right](\varepsilon_t^2 - \sigma^2) \tag{3.5}$$

For p=q=1, (Chung 1999) shows that $\sigma^2 > 0$ and $0 \leq \Phi1 \leq \theta_1 \leq_1$ is a sufficient condition for positive σ_t^2. When $\delta=0$ or 1, the FIGARCH model nests the GARCH(p,q) and IGARCH processes respectively. The IGARCH model is a short memory process having no variance and while the FIGARCH has shortest memory with $\delta > 0$ closest to 1. If $\delta > 0$ the FIGARCH is a non-stationary long memory process, otherw ise is a stationary long memory process, Laurent and Peters (2002).

The hyperbolic GARCH model, **HYGARCH(p,α,d,q)** of Davidson (2004) extends the conditional variance of the FIGARCH(p,δ,q) model by introducing weights to the difference operator in equation (3.2) such that $(1-L)^d = \left[(1-\alpha) + \alpha(1-L)^d\right]$.

The model parameterization is given by

$$\sigma_t^2 = \frac{\omega}{1-\theta(L)} + \left[1 - \frac{\phi(L)\{1 + \alpha(1-L)^d\}}{1-\theta(L)} \right]\varepsilon_t^2 \tag{3.6}$$

where α are weights to $(1-L)^d$. The parameters α and d are assumed positive. The HYGARCH nest GARCH models (for $\alpha=0$), IGARCH (for $\alpha=d=1$) and FIGARCH (for $\alpha=1$ or log $\alpha=0$). When $\alpha < 1$ (log $\alpha < 0$) the process is stationary.

4. Empirical Analysis

The factor loadings series data are obtained from DSFM for implied volatility on the German DAX index market from January 1999 to February 2003 (data available at, http://sfb649.wiwi.hu-berlin.de/fedc). Table1 presents descriptive statistics. Plots of sample autocorrelation functions, spectrum and periodogram (on log-log plane) are shown in Figure 3. The autocorrelation are positive and decay hyperbolically to zero as the lag increases. A linear relationship in the periodogram on log-log plane indicates the presence of self-similarities, the fluctuations in a power-law fashion. Figure 4 shows time series plots in absolute returns of three factor loadings series. Since unit root tests are known to perform relatively poorly in distinguishing between I(1) and the I(d) alternatives for d<1, Diebold and Rudebusch (1991), we apply model independent tests, V/S and LobRob for I(0) against I(d) alternatives. With no data driven guideline for the choice of truncation lags m, we use different values (m = 2, 3, 5, 7, 10, 20, 50) in the V/S test and (m = 30, 50, 150, 200, 300) in LobRob test.

Table 1: Summary statistics for factor loadings times series on the German DAX index market from January 1999 to February 2003, a sample of 1039 observations.

series	mean	std.dev.	min.	max.	skew.	kurt	J.Bera(p-value)
z1	1.165	0.253	0.465	1.538	-0.820	2.698	0.000
z2	0.005	0.035	-0.286	0.104	-0.251	6.958	0.000
z3	0.000	0.028	-0.103	0.134	0.933	5.452	0.000

We observe that long range dependence is different for each factor loading such that it could be interpreted in terms of a long term, middle long term and short term impact on the dynamics of IVS. The first factor loading, z1 is highly persistent and influences all options similarly, irrespective of maturity. The impact of the second factor loading, z2 gradually diminishes for longer maturities, while the third factor governs large volatility changes in relatively short maturities. Results in Tables 2 and 3

for the V/S and LobRob tests respectively indicate long-range dependence in all three factor loadings levels. For absolute returns, the tests indicate long memory in |z1| and |z3| while antipersistence could not be rejected in |z2|. Both tests results reject long memory for all factor loadings returns. Table 4 shows the $\hat{d}_{GPH}$ and $\hat{d}_{GPH}$ estimates of d for the series in levels, returns and absolute returns. To evaluate the sensitivity of results for the $\hat{d}_{GPH}$ estimator, we report estimates of d for bandwidth $m = T^{\alpha}$ where α = 0.5, 0.525, 0.575, 0.60, 0.80 and T=1052 is the sample size. For the GSP estimator the bandwidth is chosen such that m = [T /4], [T/8], [T/16], [T/32], [T/64]. Results for series in levels show 0.5<d<1; for the return series most estimates from $\hat{d}_{GPH}$ and $\hat{d}_{GPH}$ are in −0.5<d<0 while estimates of d for the absolute returns are within 0≤ d<0.5.

Figure 3: Plots of sample autocorrelation functions (lag length 300), spectrum and periodogram (in the log-log plane) of the factor loadings series.
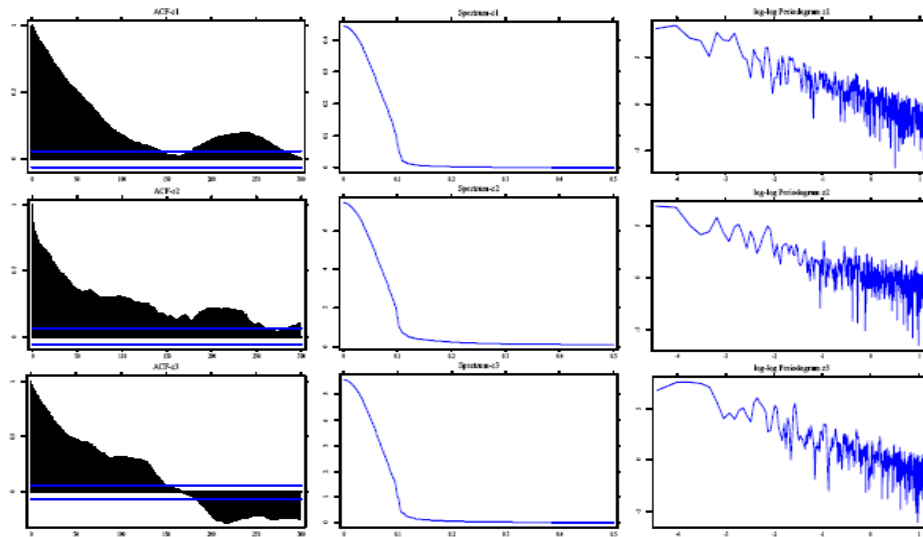

Figure 4: Time series plots of the three factor loading series in absolute returns from 04.01.1999 − 25.02.2003.
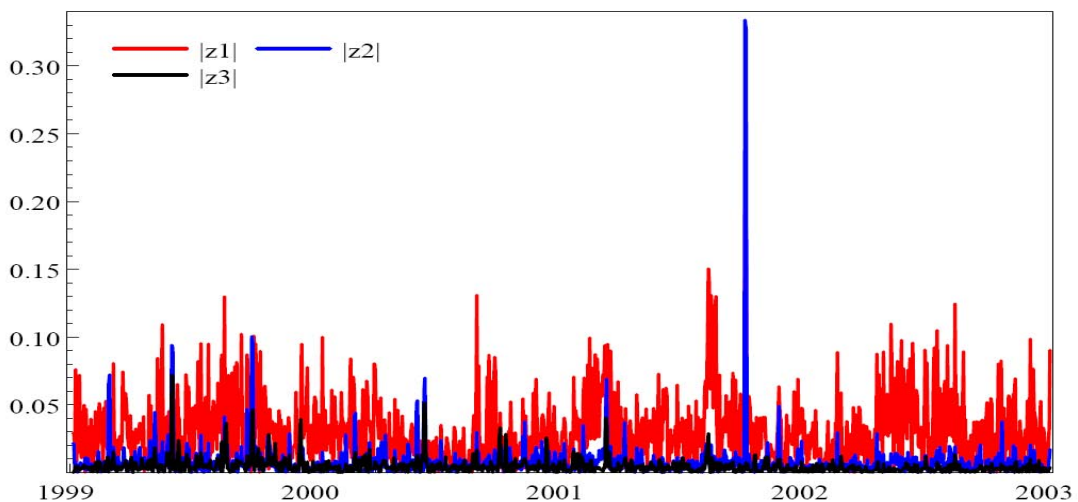
Table 2: Rescaled variance V/S test for I(0) against I(d) for series in levels, return (rt) and absolute return ($|r_t|$). m is the truncation lag. If the evaluated statistics are over the critical value, 0.1869 for I(0), we fail to reject the alternative hypothesis that the series display long memory.

	level						
m	**2**	**3**	**5**	**7**	**10**	**20**	**50**
Z1	4.63	3.29	2.22	1.68	1.24	0.68	0.32
Z2	3.77	2.89	1.54	1.54	1.15	0.65	0.31
Z3	3.54	1.82	1.38	1.38	1.02	0.57	0.27
	r_t						
Z1	0.02	0.02	0.03	0.03	0.04	0.05	0.06
Z2	0.00	0.01	0.01	0.01	0.01	0.02	0.03
Z3	0.01	0.01	0.01	0.02	0.03	0.03	0.04
m	$|r_t|$						
Z1	0.24	0.21	0.19	0.15	0.13	0.09	0.07
Z2	0.05	0.05	0.05	0.04	0.05	0.04	0.04
Z3	0.84	0.78	0.72	0.64	0.56	0.44	0.28

Table 3: LobRob: Semiparametric test for I(0) against long-memory and antipersistence for series in levels, return (r_t) and absolute return ($|r_t|$). Short memory is rejected against long-memory if the test statistic is in the lower tail of the standard normal distribution. If the test statistic is in upper tail of the standard normal distribution, short memory is rejected against antipersistent.

	level				
m	**30**	**50**	**150**	**200**	**300**
Z1	-6.08	-10.60	-38.46	-38.46	-53.60
Z2	-6.50	-11.06	-28.83	-35.98	-49.20
Z3	-5.72	-9.86	-28.33	-36.33	-50.70
	r_t				
Z1	0.65	0.96	2.69	2.58	2.83
Z2	2.02	1.47	5.14	5.76	7.03
Z3	1.40	1.61	2.91	2.99	4.25
m	$|r_t|$				
Z1	-1.25	-1.88	-7.76	-10.04	-12.42
Z2	-0.72	0.37	-0.01	-0.52	-1.00
Z3	-4.83	-6.81	-9.08	-11.13	-9.96

To guarantee that the long memory diagnosis is not a consequence of occasional r structural break such as the 11th September, 2001 terrorist attack on the World Trade Center, we use subsamples of the data to examine whether long run dependencies can be uncovered, Anderson and Bollerslev (1998). This approach is possible given that the value of d is not affected by temporal aggregation, Bollerslev and Wright (2000). Results, not presented here indicate long memory for short span of the data in levels and absolute returns. This suggests that long memory is an inherent characteristic of the factor loading series.

Table 4: d estimates for levels, returns and absolute returns: log periodogram ($\hat{d}_{GSP}$) estimates, bandwidth m=T^{α} with α=0.5, 0.525, 0.575, 0.60, 0.8 and the Gaussian semiparametric ($\hat{d}_{GSP}$) estimates, bandwidth m=[T/4], [T/8], [T/16], [T/32], [T/64]. Sample size, T=1052.

$\hat{d}_{GPH}$	z1			$\hat{d}_{GPH}$	z1		
m	**level**	**r_t**	**$\lvert r_t \rvert$**	**m**	**level**	**r_t**	**$\lvert r_t \rvert$**
32	0.95	-0.20	0.31	16	0.77	-0.36	-0.21
38	0.90	-0.25	0.18	32	0.83	-0.20	0.28
54	0.98	-0.15	0.17	65	0.95	-0.09	0.28
65	0.96	-0.14	0.19	131	0.89	-0.17	0.31
261	0.97	-0.10	0.27	263	0.93	-0.12	0.27
$\hat{d}_{GSP}$	Z2			$\hat{d}_{GSP}$	Z2		
m	**level**	**r_t**	**$\lvert r_t \rvert$**	**m**	**level**	**r_t**	**$\lvert r_t \rvert$**
32	0.69	-0.63	-0.09	16	0.79	-0.86	-0.38
38	0.73	-0.52	0.02	32	0.68	-0.62	-0.10
54	0.84	-0.35	-0.04	65	0.75	-0.34	-0.01
65	0.85	-0.39	-0.05	131	0.67	-0.39	-0.02
261	0.70	-0.40	-0.00	263	0.66	-0.37	-0.03
$\hat{d}_{GSP}$	Z3			$\hat{d}_{GSP}$	Z3		
M	**level**	**r_t**	**$\lvert r_t \rvert$**	**m**	**level**	**r_t**	**$\lvert r_t \rvert$**
32	0.95	-0.24	0.33	16	0.68	-0.50	0.69
38	0.85	-0.32	0.31	32	0.86	-0.27	0.38
54	0.99	-0.16	0.36	65	0.84	-0.24	0.35
65	0.90	-0.23	0.35	131	0.83	-0.22	0.21
261	0.70	-0.22	0.17	263	0.85	-0.17	0.17

To summarize, our analysis suggest long-range dependence in loading levels as well as in absolute returns for z1 and z3. In general no long memory in returns was detected and evidence of antipersistence in the absolute returns for z2 could not be ruled out. We therefore interpret that there is quite some correlation structure in the loadings in levels as well as in absolute returns. This implies some degree of persistence and the expectation of a slow decay in impulse responses. We also observe that the long range dependence is different for each factor loading such that it could be interpreted in terms of a long term, middle long term and short term impact on the dynamics of IVS. The first factor loading, z1 is highly persistent and influences all options similarly, irrespective of maturity. The impact of the second factor loading, z2 gradually diminishes for longer maturities, while the third factor governs large volatility changes in relatively short maturities.

4.1 Long Memory Models Application

For appropriate representation of the factor loadings series dynamics and the possibility of improved forecasting, we model the long memory in levels and absolute returns using the ARFIMA, FIGARCH and HYGARCH models. These models are known to describe volatility reasonably well and provide flexible structure that captures slow decaying autocorrelation functions. Table 5 shows estimates, $\hat{d}$ of the long-memory parameter with ARFIMA models for series in levels and absolute returns. Estimates of d are highly significant across all time series in levels. The t-statistic are highly significant to reject the null hypothesis (H0: d = 0) at 1% significance level. Results for absolute returns |z2| does confirm earlier tests results in that antipersitence in |z2| may not be rejected.

Table 5: ARFIMA long-memory parameter estimation for factor loading seriesin levels and absolute returns, from 04.01.1999 to 25.02.2003. t-value of the estimated parameter in brackets, $\log(\ell)$ is the log-likelihood and (AIC) Akaike Information Criterion.

Series	ARFIMA(p,d,q)	d	$\log(\ell)$	AIC
Z1	(5,d,4)	0.29 (1.78)	1892.33	-3764.66
Z2	(3,d,3)	0.53 (4.87)	2755.75	-5453.51
Z3	(1,d,5)	0.29 (0.74)	3885.06	-7152.12
|z1|	(2,d,2)	0.30 (4.73)	2381.53	-4753.06
|z2|	(1,d,5)	-0.32 (-0.71)	2913.99	-5813.98
|z3|	(1,d,2)	0.24 (2.34)	3927.21	-7846.42

Estimation results for the FIGARCH(1, δ, 1) and HYGARCH(1, d, 1) models are reported in Tables 6 and 7 respectively. We assume the Student-t distribution because it can appropriately account for leptokurticity exhibited by high frequency financial data, (Pagan 1996). Under student-t distributed innovations, the FIGARCH and HYGARCH long memory parameter estimates indicate long memory in levels and absolute returns. Besides, the student-t distribution parameters, ν are significantly different from zero, indicating strong fat tail phenomena. Estimates, $\delta > 0$ in FIGARCH models suggest non stationary long memory characteristics in levels and absolute returns for the first and third factor loadings series, whereas the second series (d< 0) indicate a stationary long memory behavior.

We assess models fit through the log-likelihood, the Akaike Information Criterion, and the performance of the Box-Pierce statistic for testing remaining serial correlation in the squared standardized residuals. The FIGARCH and HYGARCH models perform well in describing the high persistence existing in the conditional variance. The $Q^2(24)$ statistics suggests that the FIGARCH model can better capture the autocorrelations in the conditional variance for the series in levels while the HYGARCH is more appropriate in the case of absolute returns. The FIGARCH models report higher log likelihood values for the series in levels while the HYGARCH values are higher for absolute returns. Moreover, models in absolute returns produce better fit than those in levels, which confirms the findings of Ding et al. (1993), that absolute returns are the most appropriate indicators to represent the long memory volatility processes.

Table 6: FIGARCH estimation of the factor loading series in levels and absolute returns with t statistics in parentheses. Significance is at 5% level. Estimation is with the Student distribution with ν degrees of freedom. $Ln(\ell)$ is the value of the maximized likelihood. $Q(24)$ and $Q^2(24)$ are the Box-Pierce statistic for remaining serial correlation in the standardized and squared standardized residuals respectively, using 24 lags with p-values in square brackets. The critical value at significant level of 5% is 36.4.

	z1	*z2*	*z3*	*|z1|*	*|z2|*	*|z3|*
μ	1.329 (11.190)	-0.002 (-0.257)	-0.003 (-3.733)	0.024 (25.460)	0.005 (22.660)	0.003 (21.240)
ω	25.011 (2.368)	5.156 (2.566)	3.534 (0.538)	8.264 (2.336)	5.482 (3.363)	0.590 (4.203)
δ	0.460 (4.280)	-0.031 (-1.364)	0.716 (18.330)	0.291 (2.693)	-0.040 (-1.344)	0.172 (2.181)
Φ_1	0.161 (1.580)	0.742 (10.100)	-0.865 (-14.110)	0.085 (0.636)	0.823 (14.870)	0.309 (1.931)
β_1	0.510 (3.214)	-0.012 (-1.111)	-0.261 (-3.965)	0.248 (1.586)	-0.031 (-1.830)	-0.032 (-0.415)
ν	10.870 (3.185)	2.560 (11.320)	30.965 (1.349)	4.919 (6.016)	2.243 (22.710)	2.520 (12.660)
$Ln(\ell)$	1938.359	3302.076	3250.264	2424.146	3786.055	4378.002
AIC	-3.717	-6.404	-6.243	-4.654	-7.274	-8.413
Q(24)	58.793 [0.000]	19.167 [0.691]	4095.820 [0.000]	164.692 [0.000]	11.872 [0.972]	36.460 [0.036]
$Q^2(24)$	15.226 [0.852]	0.097 [1.000]	7.319 [0.998]	11.789 [0.961]	0.092 [1.000]	2.156 [1.000]

Table 7: HYGARCH estimation of the factor loading series in levels and absolute returns with t statistics in parentheses. Significance is at 5% level. Estimation is with the Studentt distribution with ν degrees of freedom. $log(\alpha)$ is the log of weight α, to the difference operator $(1-L)^d$. $Ln(\ell)$ is the value of the maximized likelihood. $Q(24)$ and $Q^2(24)$ are the Box-Pierce statistic for remaining serial correlation in the standardized and squared standardized residuals respectively, using 24 lags with p-values in square brackets.

	z1	*z2*	*z3*	*|z1|*	*|z2|*	*|z3|*
μ	1.304 (11.520)	-0.001 (0.703)	-0.003 (-2.061)	0.024 (26.290)	0.004 (20.230)	0.003 (19.900)
ω	0.000 (-0.203)	0.184 (5.055)	0.156 (3.344)	0.790 (1.885)	0.344 (2.267)	0.069 (2.120)
d	0.082 (1.164)	0.980 (32.630)	0.919 (13.220)	0.975 (7.073)	0.985 (46.840)	0.001 (11.610)
Φ_1	-0.861 (-20.720)	0.456 (9.161)	-0.893 (-18.660)	-0.024 (-0.256)	0.421 (2.820)	0.271 (1.663)
β_1	-0.897 (-26.300)	0.321 (0.037)	-0.107 (-1.791)	0.730 (5.715)	0.548 (4.431)	-0.050 (-0.776)
V	11.065 (3.059)	3.106 (7.918)	24.200 (1.413)	5.175 (5.538)	2.414 (9.835)	2.668 (8.027)
log (α)	0.836 (1.003)	0.083 (1.586)	-1.661 (-7.471)	-0.120 (-1.699)	-0.504 (-2.553)	4.953 (10.200)
$Ln(\ell)$	1934.661	3301.990	3256.190	2428.212	3785.100	4378.439
AIC	-3.708	-4.562	-6.252	-4.660	-7.270	-8.412
Q(24)	59.270 [0.000]	7315.110 [0.000]	4096.210 [0.000]	153.997 [0.000]	10.412 [0.988]	37.739 [0.027]
$Q^2(24)$	19.704 [0.610]	13.140 [0.930]	4.499 [0.999]	11.364 [0.940]	0.084 [1.000]	2.042 [1.000]

We examined in-sample fit of the models for the series in levels and absolute returns as well as their conditional variance forecast. Results for the FIGARCH and HYGARCH models forecasts are presented in Figures 5 and 6. The models in-sample forecast performances, Table 8 are evaluated on the basis of the Root Mean Square Error (RMSE) and the Mean Absolute Prediction Error (MAPE). We find that the FIGARCH and HYGARCH model show better forecast performance than the ARFIMA model and seems to successfully achieve the aim of modelling the long memory behavior of volatility in a parsimonious way.

Figure 5: Left-Right panels: FIGARCH and HY GARCH conditional variance forecast of factor loading in levels. Time interval from 04.01.1999 − 25.02.2003, with 1039 observations.
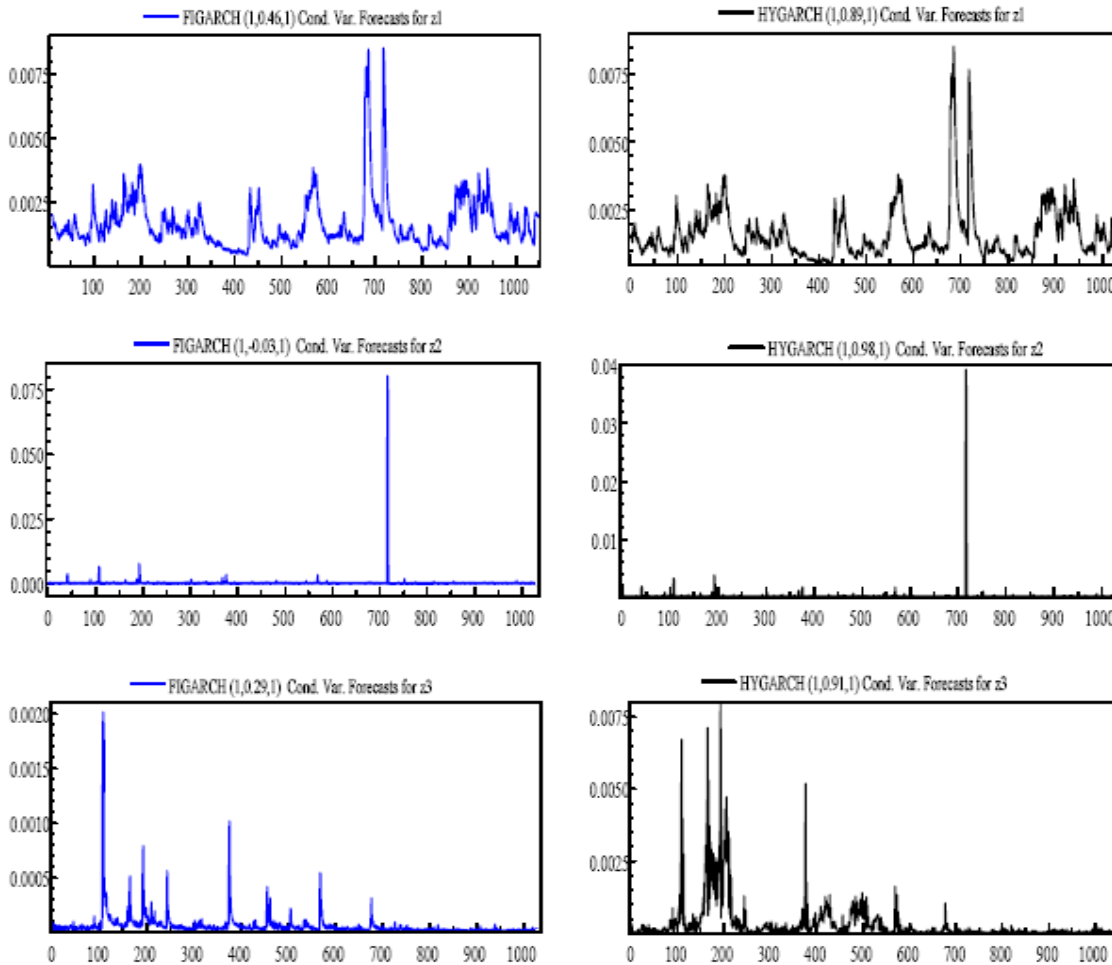
Figure 6:　Left-Right panels: FIGARCH and HYGARCH conditional variance forecast of absolute returns from 04.01.1999−25.02.2003.
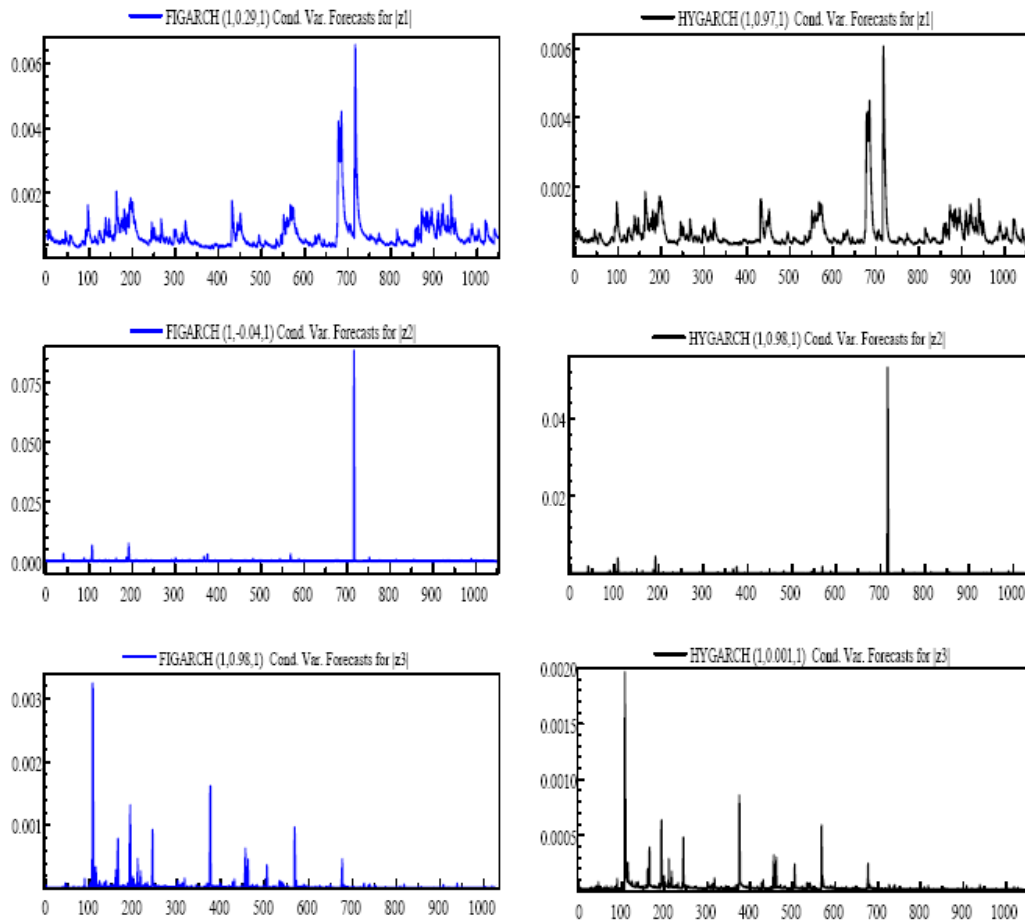

Table 8:　In-sample performance of the five-step ahead forecast of the estimated ARFIMA, FIGARCH and HYGARCH models for the factor loading series in levels and absolute returns. The measures of forecast accuracy are the Root Mean Square Error (RMSE) and the Mean Absolute Prediction Error (MAPE).

RMSE	z1	z2	z3	|z1|	|z2|	|z3|
ARFIMA	0.026	0.006	0.005	0.019	0.007	0.002
FIGARCH	0.171	0.001	0.001	0.001	0.001	0.001
HYGARCH	0.171	0.001	0.001	0.001	0.001	0.001
MAPE	z1	z2	z3	|z1|	|z2|	|z3|
ARFIMA	3.180	27.600	68.720	52.657	70.577	41.706
F IGARCH	0.991	1.166	2.837	22.910	1.892	9.134
HYGARCH	0.991	0.860	9.413	16.380	2.409	10.39

5. Conclusions

This paper presents an empirical investigation of long memory dynamics in the factors of Implied Volatility Strings. The factor loadings series were obtained by applying a Dynamic Semiparametric Factor Model (DSFM) for implied volatility strings on the German DAX index market. Long range dependence in the factor loadings series is tested using the rescaled variance V/S and the semiparametric LobRob tests. We estimated the degree of long memory based on the log-periodogram GPH regression estimator and the GSP estimator based on the Whittle approximate maximum

likelihood estimate. Results are indicative of long-range dependence in the factor loading series in levels and absolute returns.

The factors can be interpreted in terms of a long term, middle long term and short term impact on the dynamics of IV S. The first factor loading, z1 is highly persistent and influences all options similarly, irrespective of maturity. The impact of the second factor loading, z2 gradually diminishes for longer maturities and the third factor governs large volatility changes in relatively short maturities. Such dependence or persistence has importance implications for short-term trading and long range investment strategies. As a consequence, hedging strategies of a long position should take into consideration the long-memory effects in a short position in a call option. This would certainly provide more secure protection against negative effects of long-range persistence in volatility. On the other hand, better results could be obtained for models that price and hedge derivative securities when there is prior information on long-memory volatility in terms of expectation on the potential level of volatility and the rate at which volatility changes.

For an appropriate representation of the series dynamics and the possibility of improved forecasting, we model the long memory in volatility via the class of flexible processes, the ARFIMA, FIGARCH and HYGARCH models. Our results indicate that these models appear to capture the slow decaying autocorrelation function and therefore are applicable in mimicking the dynamics of the factor loadings. In comparison, models in absolute returns have better performance, confirming the findings of Ding et al. (1993), that absolute returns are the most appropriate indicator to represent the long memory volatility processes.

It would be interesting to find out if there are persistent time scales that are of local importance or influence the factor loading time series. Therefore, a possible extension for future research would include studies on spectral analysis or wavelet transform to identify such persistent time scales. In addition, the resulting long range dependence and evidence of fat tail phenomenon also provides a natural extension to risk management in terms of Value-at-Risk and Expected Shortfall. This would be useful to regulators, derivative market participants and practitioners whose interest is to reasonably forecast stock market movements as well as application to risk management.

Acknowledgement

References

[1]   Anderson, T. and Bollerslev, T., (1998). Deutsche Mark-Dollar Volatility: Intraday Activity Patterns, Macroeconomic Announcements, and Long Run Dependencies. Journal of Finance, 53: 219–265.

[2]   Baillie, R. T., (1996). Long memory processes and fractional integration economics. Journal of Econometrics, 73: 5–59.

[3]   Baillie, R. T., Bollerslev, T., Mikkelson, H., (1996). Fractionally Integrated Generalized Autoregressive Conditional Heteroskedasticity.Journal of Econometrics,14:3–30.

[4]   Beran, J., (1994): Statistics for long memory processes. Chapman & Hall, New York.

[5]   Bollerslev,T.,(1987). A conditionally heteroskedastic time series model for speculative prices and rates of returns. Review of Economics and Statistics, 69: 542-547.

[6]   Bollerslev, T. and Jubinski, D., (1999). Equity trading volume and volatility: Latent information arrivals and common long-run dependencies. Journal of Business and Economic Statistics, 17: 9 - 21.

[7]   Bollerslev, T. and Mikkelsen, H., (1996): Modelling and Pricing Long Memory in Stock Market Volatility. Journal of Econometrics, 73: 151 - 184.

[8]   Bollerslev, T., and Wright, J.H., (2000). Semiparametric estimation of Long memory volatility dependencies: The role of high-frequency data. Journal of Econometrics, Vol.98, No.1, pp.81-106.

[9]   Borak, S., Härdle, W. and Fengler, M., (2005). DSFM fitting of Implied Volatility Surfaces, Proceedings 5th International Conference on Intelligent System Design and Applications, IEEE Computer Society Number P2286, Library of Congress Number 2005930524.

[10]   Borak, S., Härdle, W., Mammen, E. and Park, B. U. (2007). Time series modelling with semiparametric factor dynamics. SFB 649 Discussion Paper 2007-023, Humboldt-Universit¨at zu Berlin.

[11]   Cheung, Y. W., and Lai, K. S., (1995). A search for long memory in international stock market returns. Journal of International Money and Finance, 14, 597-615.

[12]   Chung, C. F., (1999). Estimating the Fractionally Integrated GARCH Model. Working paper, National Täıwan University.

[13]   Davidson, J., (2004). Moments and memory properties of linear conditional heteroskedasticity models, and a new model. Journal of Business and Economic Statistics, 22, 16-29.

[14]   Deo, R. S. and Hurvich, C. M., (2001). On the Log Periodogram Regression Estimator of the Memory Parameter in Long Memory Stochastic Volatility Models. Econometric Theory, 17(686-710).

[15]   Diebold, F.X. and Rudebusch, G.D., (1991). On the Power of the Dickey-Fuller Test Against Fractional Alternatives. Economic letters, 35, 155–160.

[16]   Ding, Z. and Granger, C.W.J., (1996). Modelling Volatility Persistence of Speculative Returns. A New Approach. Journal of Econometrics, 73, 185–215.

[17]   Ding, Z., Granger, C.W.J. and Engle, R.F, (1993): A long memory property of stock market returns and a new model. Journal of Empirical Finance, 1, 83 -106.

[18]   Doornik, J.A., (2006a). An Introduction to OxMetrics 4. A software System for Data Analysis and Forecasting. Timberlake Consultant Ltd., first edn.

[19]   Doornik, J.A., Ooms, M., (2004). Inference and forecasting for ARFIMA models with an Application to US and UK inflation. Studies in Nonlinear Dynamics and Econometrics 8, No. 2, Article 14.

[20]   Fengler, M., Härdle, W. and Mammen, E., (2007). A Semiparametric Factor Model for Implied Volatility Surface Dynamics. Journal of Financial Econometrics, Vol.5, No.2,000-000.

[21]   Geweke, Porter-Hudak (1983). The Estimation and Application of Long-Memory Time Series Models. Journal of Time Series Analysis, 4, 221-238.

[22]   Giraitis, L., Kokoszka, P. and Leipus, R., (2000). Stationary ARCH models: Dependence structure and central limit theorem. Econometric Theory 16, 3-22.

[23]   Giraitis, L., Kokoszka, P. and Leipus, R., (2001). Testing for long memory in the presence of a general trend. Journal of Applied Probability, 38, 1033–1054.

[24]   Giraitis, L., Kokoszka, P., Leipus, R. and Teyssi`ere, G., (1999). Semiparametric Estimation of the Intensity of Long-Memory in Conditional Heteroskedasticity. G.R.E.Q.A.M., 99a24, Universite Aix-Marseille III.

[25]   Granger,C.W.J.,(1980). Long memory relationships and the aggregation of dynamic models. Journal of Econometrics, 14: 261–279.

[26]   Granger, C.W.J. and Ding, Z., (1996). Varieties of Long-Memory Models. Journal of Econometrics, 73, 61–77.

[27]   Granger, C. W. J. and Joyeux, R., (1995). An introduction to long memory times series models and fractional differencing. Journal of Time Series Analysis, 1,15–39.

[28]   Herzberg, M., Sibbertsen, P., (2004). Pricing of options under different volatility models.Technical Report/ Universit¨at Dortmund, SFB 475, Komplexitäts reduktion in Multivariaten Datenstrukturen, 2004, 62.

[29]   Hosking, J.R.M., (1981): Fractional Differencing. Biometrika, 68(1), 165-176.

[30]   Laurent, S. and Peters, J. P., (2002). Garch 2.2: An Ox Package for estimating and forecasting Various ARCH Models. Journal of Economic Surveys, 16(3), 447–485.

[31]   Lobato, I. and P.M. Robinson (1998). A Nonparametric Test for I(0). Review of Economic Studies, 65(3), 475495.

[32]   McLeod, A. I. and Li, W. K., (1983). Diagnostic checking ARMA time series models using squared residual autocorrelations. Journal of Time Series Analysis, 4, 269-273.

[33]   Mandelbrot, B. B., (1971). When a price be arbitraged efficiently? a limit to the validity of the random walk and Martingale models. Review of Economics and Statistics, 53, 22-236.

[34]   Newey, W.K. and West, K.D., (1987). A Simple Positive Definite, Heteroskedasticity and Autocorrelation Consistent Covariance Matrix. Econometrica, 55, 703–705.

[35]   Pagan, A., (1996). The econometrics of financial markets. Journal of Empirical Finance, 3, 15–102.

[36]   Phillips, P.C.B and Shimotsu, K., (2004). Local Whittle estimation in nonstationary and unit root cases. Annals of Statistics, 32, 656–692.

[37]   Robinson, P.M., (1995a). Gaussian Semiparametric Estimation of Long-Range Dependence. Annals of Statistics, 23, 1630-1661.

[38]   Robinson, P.M., (1995b). Log-Periodogram regression of time Series with Long Range Dependence. Annals of Statistics, 23(3), 1048-1072.

[39]   Robinson, P.M., (2003). Long memory time series. In Robinson, P.M. (Ed.), Times series with Long Memory. Oxford University Press, Oxford.

[40]   Robinson, P.M., Henry, M.,(1999).Long and Short Memory Conditional Heteroskedasticity in Estimating the Memory Parameter in Levels. Economic Theory, 15, 299-336.

[41]   Sibbertsen, P., (2004). Long-Memory in Volatilities of German Stock Returns. Empirical Economics, 29, 477 - 488.

[42]   Velasco, C., (1999). Gaussian Semiparametric Estimation of Non-Stationary Time Series. Journal of Time Series Analysis, 20: 87 127(41).

[43]   Velasco, C., (19999. Non-Stationary Log-periodogram Regression. Journal of Econometrics, 91: 325 - 371.

[44]   Wilson, P. J., and Okunev, J., (1999). Long-Term Dependencies and Long Run Non-Periodic co-Cycles: Real Estate and Stock Markets. Journal of Real Estate Research, 18, 257 - 278.

Calibration of Parametric CAT Bonds.
A Case Study of Mexican Earthquakes*

By Wolfgang Karl Härdle and Brenda López Cabrera

Abstract

The study of natural catastrophe models plays an important role in the prevention and mitigation of disasters. After the occurrence of a natural disaster, the reconstruction can be financed with catastrophe bonds (CAT bonds) or reinsurance. This paper examines the calibration of a real parametric CAT bond for earthquakes that was sponsored by the Mexican government, which is based on the estimation of the intensity rate of the arrival process of earthquake (which would trigger this particular CAT bond) from the two sides of the contract: the reinsurance and the capital markets. Additionally, the intensity rate from the historical data was estimated to conduct a comparative analysis. The results demonstrate that, under specific conditions, the financial strategy of the government, a mix of reinsurance and CAT bond, is optimal in the sense that it provides coverage of USD 450 million for a lower cost than the reinsurance itself.

JEL Classification: G19, G29, N26, N56, Q29, Q54

1. Introduction

By its geographical position, Mexico finds itself under a great variety of natural phenomena which can cause disasters, like earthquakes, eruptions, hurricanes, burning forest, floods and aridity (dryness). In case of disaster, the effects on financial and natural resources are huge and volatile. In Mexico the first risk to transfer is the seismic risk, because although it is the less recurrent, it has the biggest impact on the population and country. For example, an earthquake of magnitude 8.1 *Mw* Richter scale that hit Mexico in 1985, destroyed hundreds of buildings and caused thousands of deaths. Figure 1 depicts the number of earthquakes higher than 6.5 *Mw* occurred in Mexico during the years 1900 – 2003.

After the occurrence of a natural disaster, the reconstruction can be financed with catastrophe bonds (CAT bonds) or reinsurance. For insurers, rein-

---

surers and other corporations CAT bonds are hedging instruments that offer multi year protection without the credit risk present in reinsurance by providing full collateral for the risk limits offered through the transaction. For investors CAT bonds offer attractive returns and reduction of portfolio risk, since CAT bonds defaults are uncorrelated with defaults of other securities.
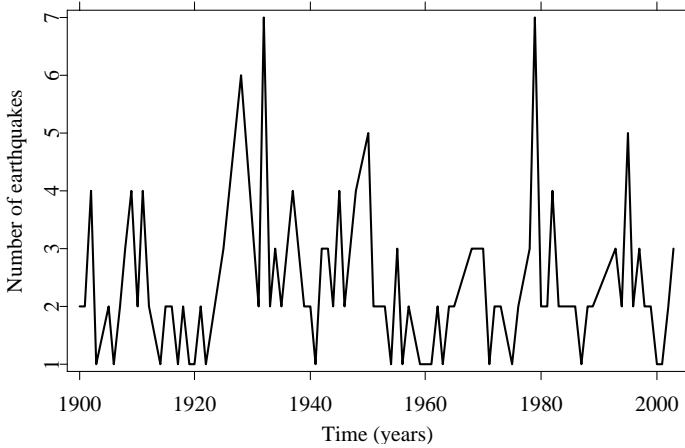

Figure 1: *Number of Mexican earthquakes occurred during 1900 – 2003*. In this figure, we plot the number of earthquakes higher than 6.5 *Mw* occurred in Mexico during the years 1900-2003. Earthquakes less than 6.5 *Mw* were not taken into account because of their high frequency and low loss impact. The data was provided by the National Institute of Seismology in Mexico, SSN (2006)

Baryshnikov / Mayo / Taylor (2001) present an arbitrage free solution to the pricing of CAT bonds under conditions of continous trading and according to the statistical characteristics of the dominant underlying processes. Instead of pricing, Anderson / Bendimerad / Canabarro / Finkemeier (2000) devoted to the CAT bond benefits by providing an extensive relative value analysis. Others, like Croson / Kunreuther (2000) focus on the CAT management and their combination with reinsurance. Lee / Yu (2002) analyze default risk on CAT bonds and therefore their pricing methodology is focused only on CAT bonds that are issued by insurers. Also under an arbitrage-free framework, Vaugirard (2003) valuates catastrophe bonds by Monte Carlo simulation and stochastic interest rates. Burnecki / Kukla (2003) correct and apply the results of Baryshnikov et al. (2001) to calculate non-arbitrage prices of a zero coupon and coupon CAT bond. Lee / Yu (2006) examine how a reinsurance company can increase the value of reinsurance contract and reduce its default risk by issuing CAT bonds.

As the study of natural catastrophe models plays an important role in the prevention and mitigation of disasters, the main motivation of this paper is the

calibration of CAT bonds. In particular, we examine the calibration of a pure parametric CAT bond for earthquakes that was sponsored by the Mexican government. The advantage of the pure parametric CAT bond is that the CAT bond payment is based on some physical parameters of the underlying event, e.g. the magnitude Mw of the earthquake. The calibration of the CAT bond is based on the estimation of the intensity rate that describes the event process (earthquakes that trigger the CAT bond's payoff) from the two sides of the contract: from the reinsurance market that consists of the sponsor company (the Mexican government) and the issuer of reinsurance coverage (in this case Swiss Re) and from the capital market, which is formed by the issuer of the CAT bond (CAT-MEX Ltd.) and the investors. In addition to these intensity estimates, the historical intensity rate is computed to conduct a comparative analysis between the intensity rates to know whether the sponsor company is getting protection at a fair price or whether the CAT bond is sold to the investors for a reasonable price. Our results demonstrate that the reinsurance market estimates a probability of an earthquake lower than the one estimated from historical data. Under specific conditions, the financial strategy of the government, a mix of reinsurance and CAT bond is optimal in the sense that it provides coverage of USD 450 million for a lower cost than the reinsurance itself.

Our paper is structured as follows. In the next section we discuss fundamentals of CAT bonds. Section 3 is devoted to the calibration of the real pure parametric CAT bond for earthquakes in Mexico. Section 4 summarizes the article. All quotations of money in this paper will be in USD and therefore we will omit the explicit notion of the currency.

2. CAT Bonds

In the mid-1990's catastrophe bonds (CAT bonds), also named as *Act of God or Insurance-linked bond,* were developed to ease the transfer of catastrophe based insurance risk from insurers, reinsurers and corporations (sponsors) to capital market investors. CAT bonds are bonds whose coupons and principal payments depend on the performance of a pool or index of natural catastrophe risks, or on the presence of specified trigger conditions. They protect sponsor companies from financial losses caused by large natural disasters by offering an alternative or complement to traditional reinsurance.

The transaction involves four parties: the sponsor or ceding company (government agencies, insurers, reinsurers), the special purpose vehicle SPV (or issuer), the collateral and the investors (institutional investors, insurers, reinsurers, and hedge funds). The basic structure is shown in Figure 2. The sponsor sets up a SPV as an issuer of the bond and a source of reinsurance protection. The issuer sells bonds to capital market investors and the proceeds are deposited in a collateral account, in which earnings from assets are collected

and from which a floating rate is paid to the SPV. The sponsor enters into a reinsurance or derivative contract with the issuer and pays him a premium. The SPV usually gives quarterly coupon payments to the investors. The premium and the investment bond proceeds that the SPV received from the collateral are a source of interest or coupons paid to investors. If there is no trigger event during the life of the bonds, the SPV gives the principal back to the investors with the final coupon or the generous interest; otherwise the SPV pays the ceding according to the terms of the reinsurance contract and sometimes pays nothing or partially the principal and interest to the investors.

There is a variety of trigger mechanisms to determine when the losses of a natural catastrophe should be covered by the CAT bond. These include the indemnity, the modeled loss, the industry index, the parametric index, the pure parametric and the hybrid trigger. Each of these mechanisms shows a range of levels of basis risks and transparency to investors.
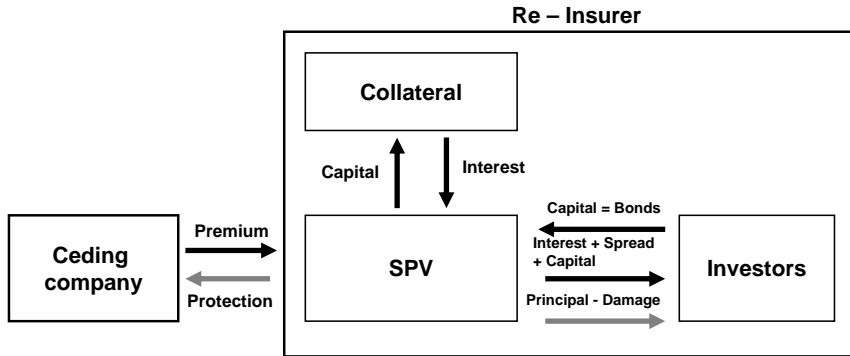

Figure 2: *Cash flows diagram of a CAT bond*. In case of the occurrence of an event (gray arrow), the SPV gives the principal back to the investors with the final coupon or the generous interest. In case of no event (black arrow), the SPV pays the ceding according to the terms of the reinsurance contract and sometimes pays nothing or partially the principal and interest to the investors

The *Indemnity trigger* involves the actual loss of the ceding company. The ceding company receives reimbursement for its actual losses from the covered event, above the predetermined level of losses. This trigger closely replicates the traditional reinsurance, but it is exposed to catastrophic and operational risk of the ceding company. In a *Modeled loss trigger* mechanism, after a catastrophe occurs the physical parameters of the catastrophe are used by a modelling firm to estimate the expected losses to the ceding company's portfolio. Instead of dealing with the company's actual claims, the transaction is based on the estimates of the model. If the modeled losses are above a specified threshold, the bond is triggered. With an *Industry index trigger,* the ceding

company recovers a proportion of total industry losses in excess of a predetermined point to the extent of the remainder of the principal. The *Parametric index trigger* uses different weighted boxes to reflect the ceding company's exposure to events in different areas. The *Pure parametric index* payouts are triggered by the occurrence of a catastrophic event with certain defined physical parameters, e.g. wind speed and location of a hurricane or the magnitude or location of an earthquake. A *Hybrid trigger* uses more than one trigger type in a single transaction.

The pricing of CAT bonds reveals some similarities to the defaultable bonds, but CAT bonds offer higher returns because of the unfixed stochastic nature of the catastrophe process. The similarity between catastrophe und default in the log-normal context has been commented in Kau / Keenan (1996).

3. Calibrating a Mexican Parametric CAT Bond

In 1996, the Mexican government established the Mexico's Fund for Natural Disasters (FONDEN) in order to reduce the exposure to the impact of natural catastrophes and to recover quickly as soon as they occur. However, FONDEN is funded by fiscal resources which are limited and have been insufficient to meet the government's obligations. Faced with the shortage of the FONDEN's resources and the high probability of earthquake occurrence, in May 2006 the Mexican government sponsored a parametric CAT bond against earthquake risk. The decision was taken because the instrument design protects and magnifies, with a degree of transparency, the resources of the trust. The CAT bond payment is based on some physical parameters of the underlying event (e.g. the magnitude Mw), thereby there is no justification of losses. The parametric CAT bond helps the government with emergency services and rebuilding after a big earthquake.

The CAT bond was issued by a SPV Cayman Islands CAT-MEX Ltd. and structured by Swiss Reinsurance Company (SRC) together with Deutsche Bank Securities. The 160 million CAT bond pays a tranche equal to the London Inter-Bank Offered Rate (LIBOR) plus 235 basis points. The CAT bond is part of a total coverage of 450 million against earthquake risk for three years and the total premium paid by the government is equal to 26 million. The payment of losses is conditional upon confirmation by a leading independent consulting firm which develops catastrophe risk assessment. This event verification agent (Applied Insurance Research Worldwide Corporation – AIR) modeled the seismic risk and detected nine seismic zones, see Figure 3. Given the federal governmental budget plan, just three out of these nine zones were insured in the transaction: zone 1, zone 2 and zone 5, with coverage of 150 million in each case, Secretaría de Hacienda y Crédito Público México (2004). The CAT bond payment would be triggered if there is an *event,* i.e. an earth-

quake higher or equal than 8 *Mw* hitting zone 1 or zone 2, or an earthquake higher or equal than 7.5 *Mw* hitting zone 5.
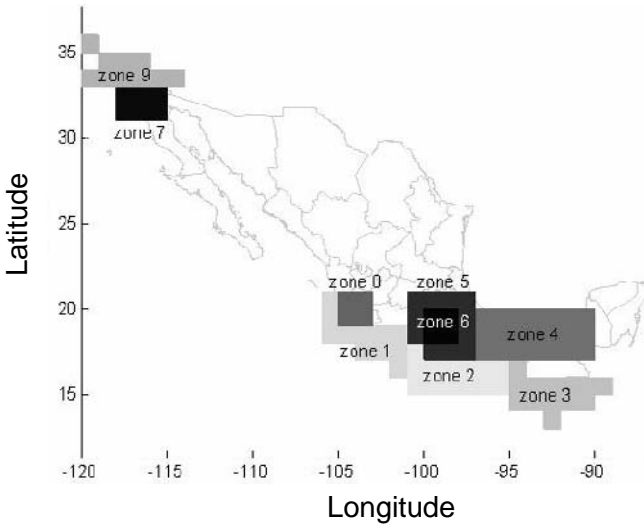

Figure 3: *Map of seismic regions in Mexico, source: SHCP (2004).* The event verification agent modeled the seismic risk and detected nine seismic zones in Mexico. Only zone 1, zone 2 and zone 5 were insured in the transaction with coverage of 150 million in each case

The cash flows diagram for the Mexican CAT bond is described in Figure 4, Secretaría de Hacienda y Crédito Público México (2004). CAT-MEX Ltd. issues the bond that is placed among investors and invests the proceeds in high quality assets within a collateral account. Simultaneous to the issuance of the bond, CAT-MEX Ltd. enters into a reinsurance contract with SRC. The proceeds of the bond will also serve to provide SRC coverage for earthquakes in Mexico in connection with an insurance agreement that FONDEN has entered with the European Finance Reinsurance Co. Ltd., an indirect wholly-owned subsidiary of SRC. A separate Event Payment Account was established with the Bank of New York providing FONDEN the ability to receive loss payments directly from CAT-MEX Ltd., subject to the terms and conditions of the insurance agreement. In case of occurrence of a trigger event, an earthquake with a certain magnitude in any of the three defined zones in Mexico, SRC pays the covered insured amount to the government, which stops paying premiums at that time and investors sacrifices their full principal and coupons.

Assuming perfect financial market, the calibration of the parametric CAT bond is based on the estimation of the intensity rate that describes the flow

process of events (earthquakes that trigger the CAT bond's payoff) from the two sides of the contract: from the reinsurance and the capital markets.
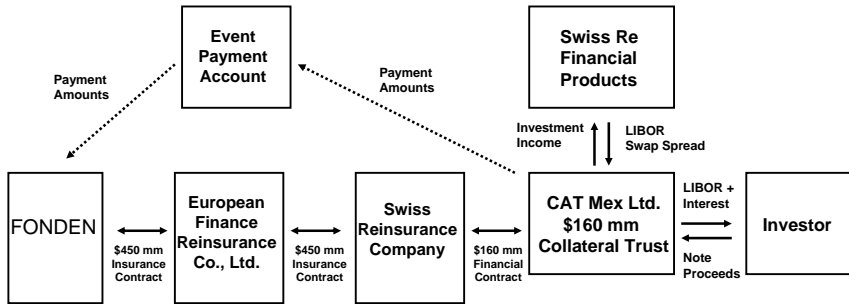

Figure 4: *The cash flows diagram for the Mexican CAT bond, source: SHCP (2004).* CAT-MEX Ltd. issues the bond and invests the proceeds in a collateral account. Simultaneous, CAT-MEX Ltd. enters into a reinsurance contract with SRC. The proceeds of the bond will also serve to provide SRC coverage for earthquakes in Mexico in connection with an insurance agreement that FONDEN has entered with the European Finance Reinsurance Co. Ltd. A separate Event Payment Account was established providing FONDEN the ability to receive loss payments directly from CAT-MEX Ltd. In case of occurrence of a trigger event, an earthquake with a certain magnitude in any of the three defined zones in Mexico, SRC pays the covered insured amount to the government and investors sacrifices their full principal and coupons

The arrival process of earthquakes or the number of earthquakes in the interval $(0, t]$ is described by the process $N_t, t \geq 0$. This process uses the times T_i when the ith earthquake occurs or the times between earthquakes $\tau_i = T_i - T_{i-1}$. The earthquake process N_t in terms of τ_i's is defined as:

$$(1) \qquad N_t = \sum_{n=1}^{\infty} 1(T_n < t).$$

Since earthquakes can strike at any time during the year with the same probability, the traditional approach in seismology is to model earthquake recurrence as a random process, in which the earthquakes suffer the *loss of memory property* $P(X > x + y | X > y) = P(X > x)$, where X is a random variable. The arrival process of earthquakes N_t can be characterized with a Homogeneous Poisson Process (HPP), with intensity rate $\lambda > 0$ if N_t is a point process governed by the Poisson law and the waiting times τ_i are exponentially distributed with intensity λ. Hence, the probability of occurrence of an earthquake is:

$$(2) \qquad P(\tau_i < t) = 1 - P(\tau_i \geq t) = 1 - e^{-\lambda t}.$$

In fact, we are interested in the occurrence of the first event. We define the first waiting time as the *stopping time* equal to $\tau = \min\{t : N_t > 0\}$, with cdf $F_\tau(t) = \mathrm{P}(\tau < t) = \mathrm{P}(N_t > 0) = 1 - e^{-\lambda t}$ and $f_\tau(t) = \lambda e^{-\lambda t}$.

3.1 Calibration in the Reinsurance Market

Let the random variable $J = 450 \cdot \mathbf{1}(\tau < 3)$ with density function $f_\tau(t)$ be the payoff of the covered insured amount to the government in case of occurrence of an event over a three years period $T = 3$. Denote H as the total premium paid by the government equal to 26 million. Suppose a flat term structure of continuously compounded discount interest rates and a HPP with intensity λ_1 that describes the arrival process of earthquakes which would trigger the CAT bond's payoff. Under the risk neutral pricing measure, a compounded discount *actuarially fair insurance price at time $t = 0$* in the reinsurance market is defined as:

$$
(3) \qquad\qquad H = \mathrm{E}\left[Je^{-\tau r_\tau}\right]
$$

$$
= \mathrm{E}\left[450 \cdot \mathbf{1}(\tau < 3)e^{-\tau r_\tau}\right]
$$

$$
= 450 \int_0^3 e^{-r_t t} f_\tau(t)\, dt
$$

$$
= 450 \int_0^3 e^{-r_t t} \lambda_1 e^{-\lambda_1 t}\, dt
$$

i.e. the insurance premium is equal to the value of the expected discounted loss from earthquake. Substituing the value of H and assuming an annual continuously compounded discount interest rate $r_t = \log(1.0541)$ constant and equal to the LIBOR in May 2006, we get:

$$
(4) \qquad\qquad 26 = 450 \int_0^3 e^{-\log(1.0541)t} \lambda_1 e^{-\lambda_1 t}\, dt
$$

where $1 - e^{-\lambda_1 t}$ is the probability of occurrence of an event. The estimation of the intensity rate of events from the reinsurance market λ_1 is equal to 0.0214. That means that the premium paid by the government to the insurance company considers a probability of occurrence of an event in three years equal to 0.0624 or the insurer expects 2.15 events in one hundred years.

3.2 Calibration in the Capital Market

For computing the intensity rate of events in the capital market λ_2, we suppose that the contract structure defines a coupon CAT bond that pays to the investors the principal P equal to 160 million at time to maturity $T = 3$ and

gives coupons C every 3 months during the bond's life in case of no event. If there is an event, the investors sacrifice their principal and coupons. These coupon bonds offered by CAT-MEX Ltd. pay to the investors a fixed spread rate z equal to 235 basis points over LIBOR. We consider the annual discretely compounded discount interest rate $r_t = 5.4139\%$ to be constant and equal to LIBOR in May 2006. The fixed coupons payments C have a value of:

(5) $$C = \left(\frac{r_t + z}{4}\right)P = \left(\frac{5.4139\% + 2.35\%}{4}\right)160 = 3.1055 .$$

Let the random variable G be the investors' gain from investing in the bond, which consists of the principal and coupons. Moreover, assume that the arrival process of earthquakes, which would trigger this particular bond, follows a HPP with intensity λ_2. Under the risk neutral pricing measure, the discretely discount *fair bond price at time $t = 0$* is given by:

(6) $$P = E\left[G\left(\frac{1}{1+r_\tau}\right)^\tau\right]$$

$$= E\left[\sum_{t=1}^{12} C \cdot \mathbf{1}(\tau > \frac{t}{4})\left(\frac{1}{1+r_t}\right)^{\frac{t}{4}} + P \cdot \mathbf{1}(\tau > 3)\left(\frac{1}{1+r_t}\right)^3\right]$$

$$= \sum_{t=1}^{12} Ce^{-\lambda_2 \frac{t}{4}}\left(\frac{1}{1+r_t}\right)^{\frac{t}{4}} + Pe^{-3\lambda_2}\left(\frac{1}{1+r_t}\right)^3 .$$

In this case, the investors receive 12 coupons during 3 years and its principal P at maturity $T = 3$. Hence, substituting the values of the principal $P = 160$ million and the coupons $C = 3.1055$ million in equation (6), it follows:

(7) $$160 = \sum_{t=1}^{12} 3.06\left(\frac{e^{-\lambda_2}}{1.0541}\right)^{\frac{t}{4}} + \frac{160e^{-3\lambda_2}}{(1.0541)^3} .$$

Solving equation (7), the intensity rate of events from the capital market λ_2 is equal to 0.0241. In other words, the capital market estimates a probability of occurrence of an event equal to 0.0699, equivalently to 2.4 events in one hundred years.

3.3 Calibration via Historical Data

In addition to the intensity estimates in the Reinsurance and Capital Market, the historical intensity rate λ_3 that describes the flow process of events (earthquakes that trigger the CAT bond's payoff) is calculated to conduct a comparative analysis between them. The data was provided by the National Institute of Seismology in Mexico, Servicio Sismológico Nacional Instituto de Geosifísica

(2006). It describes the time t, the depth d, the magnitude Mw and the epicenters of 192 earthquakes higher than 6.5 Mw occurred in the country during 1900 to 2003. Earthquakes less than 6.5 Mw were not taken into account because of their high frequency and low loss impact. Table 1 shows that almost 50 % of the earthquakes has occurred in the insured zones that were defined in the CAT bond contract, mainly in zone 2.

Table 1

Frequency of the earthquake location for the 1900 – 2003 earthquake data

Zone	Frequency	Percent	% Cumulative
1	30	16 %	16 %
2	42	22 %	38 %
5	18	9 %	47 %
Other	102	53 %	100 %

Let Y_i be i.i.d. random variables, indicating the magnitude Mw of the ith earthquake at time t. Define $\bar{u}$ as the threshold magnitude for a specific location. The estimation of the historical intensity rate λ_3 is based on the *intensity model*. This model assumes that there exist i.i.d. random variables ε_i called *trigger events* that characterize earthquakes with magnitude Y_i higher than a defined threshold $\bar{u}$ for a specific location, i.e. $\varepsilon_i = \mathbf{1}(Y_i \geq \bar{u})$. Then the trigger event process B_t is characterized as:

$$(8) \qquad B_t = \sum_{i=1}^{N_t} \varepsilon_i$$

where N_t is a HPP describing the arrival process of earthquakes with intensity $\lambda > 0$. B_t is a process which counts only earthquakes that trigger the CAT bond's payoff. However, the dataset contains only three such events, what leads to the calibration of the intensity of the process B_t be based on only two waiting times. Therefore in order to compute λ_3, consider the process B_t and define p as the probability of occurrence of a trigger event conditional on the occurrence of the earthquake. Then the probability of no event up to time t is equal to:

$$(9) \qquad P(B_t = 0) = P(N_t = 0) + P(N_t = 1)(1 - p) + P(N_t = 2)(1 - p)^2 + \ldots$$

$$= \sum_{k=0}^{\infty} P(N_t = k)(1 - p)^k = \sum_{k=0}^{\infty} \frac{(\lambda t)^k}{k!} e^{(-\lambda t)}(1 - p)^k$$

$$= \sum_{k=0}^{\infty} \frac{\{\lambda(1 - p)t\}^k}{k!} e^{(-\lambda t)} e^{-\lambda(1-p)t} e^{\lambda(1-p)t} = e^{-\lambda p t} = e^{-\lambda_3 t}$$

by definition of the Poisson distribution and since $\sum_{k=0}^{\infty} \frac{\{\lambda(1-p)t\}^k}{k!} e^{-\lambda(1-p)t}$ $= 1$. Now the calibration of the λ_3 can be decomposed into the calibration of the intensity of all earthquakes with a magnitude higher than 6.5 *Mw* and the estimation of the probability of the trigger event.

Since the historical data contains three earthquakes with magnitude *Mw* higher than the defined thresholds by the modelling company, the probability of occurrence of the trigger event is equal to $p = \left(\frac{3}{192}\right)$. The estimation of the annual intensity is obtained by taking the mean of the daily number of earthquakes times 360 i.e. $\lambda = (0.005140)(360) = 1.8504$. Consequently the annual historical intensity rate for a trigger event is equal to $\lambda_3 = \lambda p = 1.8504\left(\frac{3}{192}\right) = 0.0289$. This means that approximately 2.89 events are expected to occur in the insured areas of the country within one hundred years. The magnitude of earthquakes above 6.5 *Mw* that occurred in Mexico during the period 1900 to 2003 are illustrated in Figure 5. It also indicates earthquakes that occurred in the insured zones and trigger events.

Table 2 summarizes the values of the intensity rates $\lambda's$ and the probabilities of occurrence of a trigger event in one and three years. Whereas the reinsurance market expects approximately 2.15 events to occur in one hundred years, the capital market anticipates 2.42 events and the historical data predicts 2.89 events. In other words, the CAT bond has approximatly 6.99 % chance of default (or event) within 3 years, what makes the bond to be rated in the range of "BB" by Standard and Poor's (2007) or "Ba" by Moody's Investors Services (2007). Table 3 shows the one-to-ten year cumulative default rates from two sources of data.

Table 2

Calibration of intensity rates: the intensity rate from the reinsurance market λ_1, the intensity rate from the capital market λ_2 and the historical intensity rate λ_3

	λ_1	λ_2	λ_3
Intensity	0.0214	0.0241	0.0289
Probability of event in 1 year	0.0212	0.0238	0.0284
Probability of event in 3 year	0.0624	0.0699	0.0830
No. expected events in 100 years	2.1482	2.4171	2.8912

In Table 2, we also observe that the value of the λ_3 depends on the time period of the historical data and it is not very accurate since it is based only on three events. The intensity rate λ_3 could be interpreted as the real intensity rate describing the flow process of events, however its confidence intervals in Table 4 are very large that its value might change with a different period of the historical data. Nevertheless, the estimation of the risk neutral intensity rates of event λ_1 or λ_2 are reliable since their values lie within the confidence intervals of λ_3.
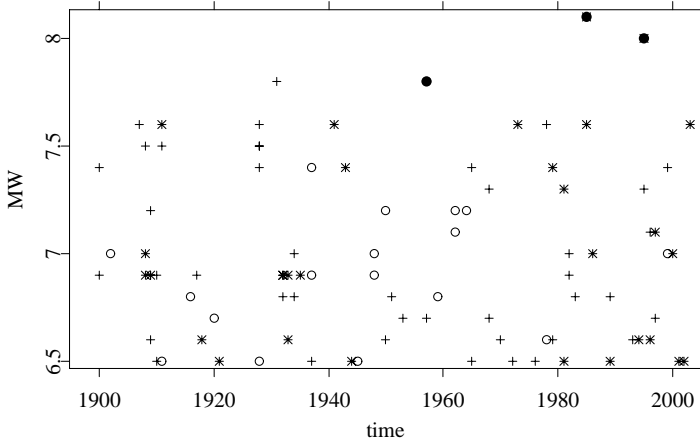
Figure 5: *Magnitude of trigger events and Mexican earthquakes occurred during the years 1900-2003*. The plot shows the magnitude of trigger events (filled circles) occurred in Mexico during 1900-2003, earthquakes occurred in zone 1 (stars), in zone 2 (crosses), in zone 5 (circles) and earthquakes out of insured zones (triangles). Observe that there were more earthquakes with high magnitude but did not occurred in insured zones

Table 3

Cumulative Default Rate comparison (in % for up to 10 years)

Rating	Data	1	2	3	4	5	6	7	8	9	10
AAA / Aaa	Moody's	0.00	0.00	0.00	0.04	0.08	0.13	0.19	0.19	0.19	0.19
	S&P	0.00	0.00	0.09	0.19	0.29	0.43	0.50	0.62	0.66	0.70
AA / Aaa	Moody's	0.01	0.02	0.05	0.12	0.18	0.23	0.26	0.29	0.31	0.37
	S&P	0.01	0.05	0.10	0.20	0.32	0.43	0.56	0.68	0.78	0.89
A / A	Moody's	0.20	0.10	0.25	0.38	0.51	0.64	0.76	0.89	1.01	1.09
	S&P	0.06	0.17	0.31	0.47	0.68	0.91	1.19	1.41	1.64	1.90
BBB / Baa	Moody's	0.21	0.57	1.00	1.53	2.06	2.57	3.70	3.51	3.92	4.37
	S&P	0.24	0.71	1.23	1.92	2.61	3.28	3.82	4.38	4.89	5.42
BB / Ba	Moody's	1.27	3.50	6.20	8.89	11.26	13.37	15.26	16.95	18.44	19.83
	S&P	1.07	3.14	5.61	7.97	10.10	12.12	13.73	15.15	16.47	17.49
B / B	Moody's	5.26	11.44	17.31	22.41	27.26	31.59	35.50	38.80	41.59	43.80
	S&P	4.99	10.92	15.90	19.76	22.55	24.72	26.54	28.00	29.20	30.42
CCC / Caa	Moody's	17.14	28.13	37.62	45.34	50.89	55.00	57.76	60.65	64.79	71.27
	S&P	26.29	34.73	39.96	43.19	46.22	47.49	48.61	49.23	50.95	51.83

Apparently the difference between the intensity rates for a trigger event λ_1, λ_2 and λ_3 seems to be insignificant, but for the government it has a financial and social repercussion since the intensity rate of the flow process of events influences the price of the parametric CAT bond that will help the government to obtain resources after a big earthquake. Particularly after a catastrophic event occurred, the reinsurance market suffers from a shortage of capital but this gives reinsurance firms the ability to gain more market power that will enable them to charge higher premiums than expected. Our estimation of intensity rates, contrary to the theory predictions, show that the Mexican government paid total premiums of 26 million that is 0.75 times the real actuarially fair one (34.605 million), which is obtained by substituting the historical intensity rate λ_3 in equation (4). At first glance, it appears that either the government saves 8.605 million (24.86) % from transferring the seismic risk with a reinsurance contract, but in fact this difference is explained by the market price of risk, which in this case is negative. Since λ_3 is only 50 % confident, no further analysis about market price of risk will be done in our analysis.

Table 4

Confidence Intervals for the intensity rate
of events from the historical data λ_3

$100\,(1-\alpha)\,\%$	Confidence intervals
99 %	0 – 0.0791
95 %	0 – 0.0666
90 %	0.0083 – 0.0594
50 %	0.0183 – 0.0392

The difference between λ_1 and λ_2 might be explained by the absence of the public and liquid market of earthquake risk in the reinsurance market, since just limited information is available. This might cause the pricing in the reinsurance market to be less transparent than pricing in the capital market. Another argument to this difference might be that contracts in the capital market are more expensive than contracts in the reinsurance market, e.g. when one considers the default risk or the cost of risk capital (the required return necessary to make a capital budgeting project): the cost of risk capital in the capital market is usually higher than that in the reinsurance market and a CAT bond presents no credit risk as the proceeds of the bond are held in a SPV, a transaction off the insurer's balance sheet.

Since the insured loss faced by the government is independent of the state of insolvency of the reinsurer, our results indicate that the probability that the resinsurer will default in this transaction over the next three years could be approximately equal to the price discount that the government gets in the

risk transfer of earthquake risk (≈ 10.7) %, which is the difference in premiums computed with the corresponding intensity rates in the reinsurance and capital markets, λ_1 and λ_2. Transferring the earthquake risk with reinsurance in the three years implies that the exposure at default (the amount the reinsurer will loose as a result of the default) is approximately equal to 10.7 % of 450 million coverage and the expected loss is equal to 10.7 % of 450 million coverage times the loss given default (1-recovery rate). Since recovery rates are not very accurately estimated, we neglect the computation of the expected loss.

However, the best explanation of the low premiums for covering the seismic risk paid by the government might be the mix of the reinsurance contract and the CAT bond. Since the 160 million CAT bond is part of a total coverage of 450 million, the reinsurance company transfers 35 % of the total seismic risk to the investors, who effectively are betting that a trigger event will not hit specified regions in Mexico in the next three years. If there is an event, the reinsurer must pay to the government 290 million from the reinsurance part and 160 million from the CAT bond to cover the insured loss of 450 million. The value of the premium for 290 million coverage with intensity rate of event λ_2 is $\int_0^3 290\lambda_2 e^{-t(r_t+\lambda_1)}dt = 18.799$. Therefore the total paid premium of 26 million might consist of 18.799 million premium from the reinsurance and the CAT bond layers and 7.221 million for transaction costs or the management added value or for coupon payments. This time the exposure at default is approximately equal to 10.7 % of 290 million coverage and the expected loss is equal to 10.7 % $\times$ 290 million $\times$ (1-recovery rate). This government's financial strategy is optimal in the sense that it provides coverage of 450 million against seismic risk for a lower cost and lower exposure at default than the reinsurance itself. However, this financial strategy does not eliminate completely the costs imposed by market imperfections. Lee / Yu (2006) get similar results, when examining how a reinsurance company can increase the value of reinsurance contract and reduce its default risk by issuing CAT bonds.

4. Conclusion

The occurrence of disfavoured extreme natural events like earthquakes, hurricanes, long cold winter, heat, drought, freeze, etc. may cause substantial financial losses. In the presence of this, many sponsor companies have turned to the capital markets to cover costs of potential catastrophes by issuing CAT bonds that passes the risk on to investors. This paper calibrates a real parametric CAT bonds that was sponsored in 2006 by the Mexican government. The decision of the government to issue a parametric CAT bond relies on the fact that it triggers immediately when an earthquake meets the defined physical parameters. The parametric CAT bond especially helps the government with fast emergency services and rebuilding after a big earthquake.

Under the assumption of perfect markets, the calibration of the bond is based on the estimation of the intensity rate that describes the flow process of events (earthquakes that trigger the CAT bond's payoff) from the two sides of the contract: from the reinsurance and the capital markets. This intensity rate reflects the CAT bond's spread rate. Additionally, we estimate the historical intensity rate using the *intensity model* that accounts only earthquakes that trigger the CAT bond's payoff. The results indicate that in presence of catastrophe risk, the default risk is substantial for the valuation of the reinsurance premium. However, the argument to the low premiums payed by the government for covering the seismic risk might be the mix of reinsurance and CAT bond, where 35 % of the total seismic risk is transferred to the investors and the exposure at default is reduced.

References

Anderson, R. R. / *Bendimerad,* F. / *Canabarro,* E. / *Finkemeier,* M. (2000): Analyzing Insurance-Linked Securities, Journal of Finance 1(2): 49 – 78.

Baryshnikov, Y. / *Mayo,* A. / *Taylor,* D. (2001): Pricing of CAT bonds, Preprint .

Burnecki, K. / *Kukla,* G. (2003): Pricing of Zero-Coupon and Coupon CAT Bonds, Appl. Math. 30: 315 – 324.

Burnecki, K. / *Kukla,* G. / *Weron,* R. (2000): Property insurance loss distributions, Physica A 287: 269 – 278.

Clarke, R. / *Collura,* J. / *McGhee,* C. (2007): The catastrophe bond market at year-end 2006: Ripple into waves, Technical report, Guy Carpenter and Company, Inc.

Clarke, R. / *Faust,* J. / *McGhee,* C. (2005): The Catastrophe Bond Market at Year-End 2004: The growing appetite for catastrophic risk, Technical report, Guy Carpenter and Company, Inc.

Clarke, R. / *Faust,* J. / *McGhee,* C. (2006): The Catastrophe Bond Market at Year-End 2005: Ripple effects from record storms, Technical report, Guy Carpenter and Company, Inc.

Croson, D. / *Kunreuther,* H. (2000): Customizing indemnity contracts and indexed cat bonds for natural hazard risks, Journal of Risk Finance 1(3): 24 – 41.

Dubinsky, W. / *Laster,* D. (2003): Insurance Link Securities, Technical report, Swiss Re Capital Markets Corporation.

Froot, K. (2001): The market for catastrophe risk: A clinical examination, Journal of Risk Financial economics 60: 529 – 571.

Grandell, J. (1991): Aspects of Risk Theory, Springer, New York.

Group on California Earthquake Probabilities, W. (1988): Probabilities of large earthquakes occurring in California on the San Andreas fault, U.S. Geol. Open-file report 398: 1 – 62.

International Association of Insurance Supervisors, I. (2003): Non-life insurance securitisation, Technical report, IAIS Issues papers and other reports.

Kau, J. / *Keenan,* D. (1996): An option-theoretic model of catastrophes applied to mortage insurance, Journal of Risk and Insurance 63(4): 639 – 656.

Lee, J. / *Yu,* M. (2002): Pricing default risky CAT bonds with moral hazard and basis risk, Journal of Risk and Insurance 69 (1): 25 – 44.

Lee, J. / *Yu,* M. (2006): Valuation of catastrophe reinsurance with catastrophe bonds, Insurance: Mathematics and Economics 41: 264 – 278.

McGhee, C. (2004): The Catastrophe Bond Market at Year-End 2003: Market update, Technical report, Guy Carpenter and Company, Inc.

Moody's Investors Service (2007): Default and recovery rates: 1920 – 2006, Technical report, Moody's Special Report, New York.

Mooney, S. (2005): The World Catastrophe reinsurance market 2005, Technical report, Guy Carpenter and Company, Inc.

Secretaría de Hacienda y Crédito Público México (2001): Acuerdo que establece las reglas de operación del fondo de desastres naturales FONDEN, Technical report, SHCP, Mexico.

Secretaría de Hacienda y Crédito Público México (2004): Administración de riesgos catastróficos del FONDEN, Technical report, SHCP, Mexico.

Servicio Sismológico Nacional Instituto de Geosifísica UNAM (2006): Earthquakes data base 1900 – 2003, Technical report, UNAM, Mexico.

Solutions, R. M. (2006): Mexico earthquake, Technical report, RMS, www.rms.com / Catastrophe / Models / Mexico.asp.

Standard and Poor's (2007): Rating performance 2006: Stability and transition, Technical report, S&P, New York.

Vaugirard, V. E. (2003): Valuing catastrophe bonds by monte carlo simulations, Applied Mathematical Finance 10: 75 – 90.

Variable Selection and Oversampling in the Use of Smooth Support Vector Machines for Predicting the Default Risk of Companies

WOLFGANG HÄRDLE,[1] YUH-JYE LEE,[2]
DOROTHEA SCHÄFER[3] AND YI-REN YEH[2]

[1] *CASE, Humboldt University, Berlin, Germany*
[2] *Department of Computer Science Information Engineering, National Taiwan University of Science and Technology, Taipei, Taiwan*
[3] *German Institute of Economic Research, Berlin, Germany*

ABSTRACT

In the era of Basel II a powerful tool for bankruptcy prognosis is vital for banks. The tool must be precise but also easily adaptable to the bank's objectives regarding the relation of false acceptances (Type I error) and false rejections (Type II error). We explore the suitability of smooth support vector machines (SSVM), and investigate how important factors such as the selection of appropriate accounting ratios (predictors), length of training period and structure of the training sample influence the precision of prediction. Moreover, we show that oversampling can be employed to control the trade-off between error types, and we compare SSVM with both logistic and discriminant analysis. Finally, we illustrate graphically how different models can be used jointly to support the decision-making process of loan officers.   Copyright © 2008 John Wiley & Sons, Ltd.

KEY WORDS insolvency prognosis; support vector machines; statistical learning theory; non-parametric classification

INTRODUCTION

Default prediction is at the core of credit risk management and has therefore always attracted special attention. It has become even more important since the Basel Committee on Banking Supervision (Basel II) established borrowers' rating as the crucial criterion for minimum capital requirements of banks. The methods for generating rating figures have developed significantly over the last 10 years (Krahnen and Weber, 2001). The rationale behind the increased sophistication in predicting borrowers' default risk is the aim of banks to minimize their cost of capital and to mitigate their own bankruptcy risks.

* Correspondence to: Dorothea Schäfer, German Institute for Economic Research (DIW) Berlin, Mohrenstrasse 58, 10117 Berlin, Germany. E-mail: dschaefer@diw.de

In this paper we intend to contribute to the increasing sophistication by exploring the predicting power of smooth support vector machines (SSVM). SSVM are a variant of the conventional support vector machines (SVM). The working principle of SVM in general can be described very easily. Imagine a group of observations in distinct classes such as balance sheet data from solvent and insolvent companies. Assume that the observations are such that they cannot be separated by a linear function. Rather than fitting nonlinear curves to the data, SVM handle this problem by using a specific transformation function—the kernel function—that maps the data from the original space into a higher-dimensional space where a hyperplane can do the separation linearly. The constrained optimization calculus of SVM gives a unique optimal separating hyperplane and adjusts it in such a way that the elements of distinct classes possess the largest distance to the hyperplane. By re-transforming the separating hyperplane into the original space of variables, the typical nonlinear separating function emerges (Vapnik, 1995). The main difference between SSVM and SVM is the following: the SSVM technique formulates the problem as an unconstrained minimization problem. This formulation has nice mathematical properties such as strong convexity and infinitely often differentiability.

Our aim is threefold when using SSVM. Firstly, we examine the power of the SSVM in predicting company defaults; secondly, we investigate how important factors that are exogenous to the model, such as selecting the appropriate set of accounting ratios, length of training period and structure of the training sample, influence the precision; and thirdly, we explore how oversampling and downsampling affect the trade-off between Type I and Type II errors. In addition, we illustrate graphically how loan officers can benefit from jointly considering the prediction results of different SSVM variants and different models.

There are basically three distinct approaches in predicting the risk of default: option theory-based approaches, parametric models and non-parametric methods. While the first class relies on the rule of no arbitrage, the latter two are based purely on statistic principles. The popular (Merton, 1974) model treats the company's equity as the underlying asset of a call option held by shareholders. In case of insolvency shareholders deny exercising. The probability of default is derived from an adapted Black–Scholes formula. Later, several authors (e.g., Longstaff and Schwartz, 1995; Mella-Barral and Perraudin, 1997; Leland and Toft, 1996; Zhou, 2001; to name only a few) proposed variations to ease the strict assumptions on the structure of the data imposed by the Merton model. These approaches are frequently denoted as structural models. However, the most challenging requirement is the knowledge of market values of debt and equity. This precondition is a severe obstacle to using the Merton model adequately as it is only satisfied in a minority of cases.

Parametric statistical models can be applied to any type of data, whether they are market based or book based. The first model introduced was discriminant analysis (DA) for univariate (Beaver, 1966) and multivariate models (Altman, 1968). After DA usage of the logit and probit approach for predicting default was proposed in Martin (1977) and Ohlson (1980). These approaches rely on the a priori assumed functional dependence between risk of default and predictor. DA requires a linear functional dependence, or a pre-shaped polynomial functional dependence in advanced versions. Logit and probit tools work with monotonic relationships between default event and predictors such as accounting ratios. However, such restrictions often fail to meet the reality of observed data. This fact makes it clear that there is a need for an approach that, in contrast to conventional methods, relaxes the requirements on data and/or lowers the dependence on heuristics. Semi-parametric models as in Hwang *et al.* (2007) are between conventional linear models and non-parametric approaches. Nonlinear classification methods such as support vector machines (SVM) or neural networks are even stronger candidates to meet these demands as they go beyond conventional

discrimination methods. Tam and Kiang (1992) and Altman *et al.* (1994) focus on neural networks. In contrast, we concentrate on SVM exclusively.

The SVM method is a relatively new technique and builds on the principles of statistical learning theory. It is easier to handle compared to neural networks. Furthermore, SVM have a wider scope of application as the class of SVM models includes neural networks (Schölkopf and Smola, 2002). The power of SVM technology becomes evident in a situation as depicted in Figure 1 where operating profit margin and equity ratio are used as explanatory variables. A separating function similar to a parabola (in black) appears in the two-dimensional space. The accompanying light-grey lines represent the margin boundaries whose shape and location determine the distance of elements from the separating function. In contrast, the logit approach and discriminant DA yield the (white) linear separating function (Härdle *et al.*, 2007a).

Selecting the best accounting ratios for executing the task of predicting is an important issue in practice but has not received appropriate attention in research. We address this issue of how important the chosen set of predictors is for the outcome. For this purpose we explore the prediction potential of SSVM within a two-step approach. First, we derive alternative sets of accounting ratios that are used as predictors. The benchmark set comes from Chen *et al.* (2006). A second set is defined by a 1-norm SVM, and the third set is based on the principle of adding only those variables that contain the most contrary information with respect to an initial set that is a priori chosen. We call the latter procedure the incremental forward selection of variables. As a result we are working with three variants of SSVM. In the second step, these variants are compared with respect to their prediction power. We also compare SSVM with two traditional methods: the logit model and linear discriminant analysis.

The analysis is built on 28 accounting ratios of 20,000 solvent and 1000 insolvent German companies. Our findings show that the different SSVM types have an overall good performance with the means of correct predictions ranging from 70% to 78%. The SSVM on the basis of incremental
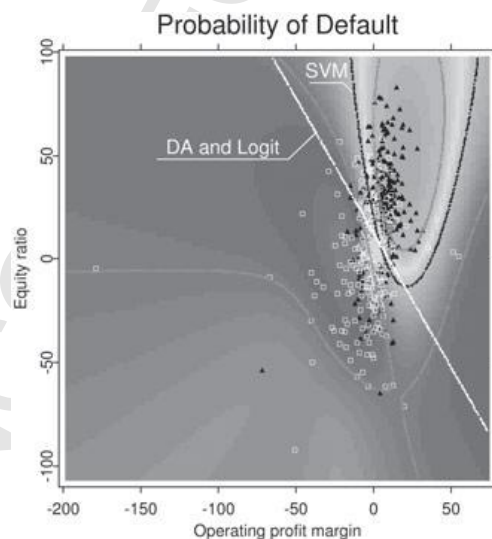

Figure 1. SVM-separating function (black) with margin in a two-dimensional space

forward selection clearly outperform the SSVM based on predictors selected by the 1-norm SVM. It is also found that oversampling influences the trade-off between Type I and Type II errors. Thus, oversampling can be used to make the relation of the two error types an issue of bank policy.

The rest of the paper is organized as follows. The following two sections describe the data, performance measures and SVM methodology. In the fourth section the variable selection technique and outcome are explained. The fifth section presents the experimental settings, estimation procedure and findings, and illustrates selected results. The sixth section concludes.

DATA AND MEASURES OF ACCURACY

In this study of the potential virtues of SVM in insolvency prognosis the CreditReform database is employed. The database consists of 20,000 financially and economically solvent and 1000 insolvent German companies observed once in the period from 1997 to 2002. Although the companies were randomly selected, accounting information dates most frequently in 2001 and 2002. Approximately 50% of the observations come from this period. The industry distribution of the insolvent companies is as follows: manufacturing 25.7%, wholesale and retail trade 20.1%, real estate 9.4%, construction 39.7% and others 5.1%. The latter includes businesses in agriculture, mining, electricity, gas and water supply, transport and communication, financial intermediation social service activities and hotels and restaurants. The 20,000 solvent companies belong to manufacturing (27.4%), wholesale and retail trade (24.8%), real estate (16.9%), construction (13.9%) and others (17.1%). There is only low coincidence between the industries represented in the insolvent and the solvent group of 'others'. The latter comprises many companies in industries such as publication administration and defense, education and health. Figure 2 shows the distribution of solvent and insolvent companies across industries. A set of balance sheet and income statement items describes each company. The ones we use for further analysis are described below:

- AD (amortization and depreciation)
- AP (accounts payable)
- AR (account receivable)
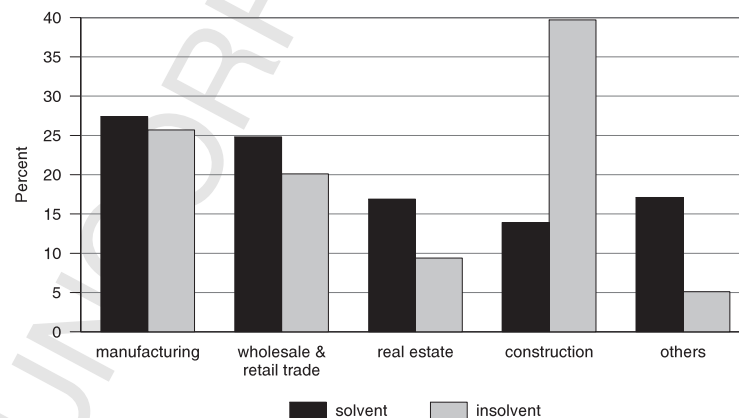
Figure 2. The distribution of solvent and insolvent companies across industries

- CA (current assets)
- CASH (cash and cash equivalents)
- CL (current liabilities)
- DEBT (debt)
- EBIT (earnings before interest and tax)
- EQUITY (equity)
- IDINV (growth of inventories)
- IDL (growth of liabilities)
- INTE (interest expense)
- INV (inventories)
- ITGA (intangible assets)
- LB (lands and buildings)
- NI (net income)
- OI (operating income)
- QA (quick assets)
- SALE (sales)
- TA (total assets)
- TL (total liabilities)
- WC (working capital (= CA − CL))

The companies appear in the database several times in different years; however, each year of balance sheet information is treated as a single observation. The data of the insolvent companies were collected 2 years prior to insolvency. The company sizes are measured by total assets. We construct 28 ratios to condense the balance sheet information (see Table I). However, before dealing with the CreditReform dataset, some companies whose behavior is very different from other ones are filtered out in order to make the dataset more compact. The data pre-processing procedure is described as follows:

1. We excluded companies whose total assets were not in the range of 10^5–10^7 EUR (remaining insolvent: 967; solvent: 15,834).
2. In order to compute the accounting ratios AP/SALE, OI/TA, TL/TA, CASH/TA, IDINV/INV, INV/SALE, EBIT/TA and NI/SALE, we have removed companies with zero denominators (remaining insolvent: 816; solvent 11,005).
3. We dropped outliers, that is, in the insolvent class companies with extreme values of financial indices have been removed (remaining insolvent: 811; solvent: 10,468).

After pre-processing, the dataset consists of 11,279 companies (811 insolvent and 10,468 solvent). In the following analysis, we focus on the revised dataset.

The performance of the SSVM is evaluated on the basis of three measures of accuracy: Type I error rate (%), Type II error rate (%) and total error rate (%). The Type I error is the ratio of the number of insolvent companies predicted as solvent ones to the number of insolvent companies. The Type II error is the ratio of the number of solvent companies predicted as insolvent ones to the number of solvent companies. Accordingly, the error-type rates (in percentage) are defined as follows

- Type I error rate = FN/(FN + TP) × 100 (%);
- Type II error rate = FP/(FP+ TN) × 100 (%);
- Total error rate = (FN + FP)/(TP + TN + FP + FN) × 100 (%);

1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46

Table I. Definitions of accounting ratios used in the analysis

Variable	Ratio	Indicator for
X1	NI/TA	Profitability
X2	NI/SALE	Profitability
X3	OI/TA	Profitability
X4	OI/SALE	Profitability
X5	EBIT/TA	Profitability
X6	(EBIT + AD)/TA	Profitability
X7	EBIT/SALE	Profitability
X8	EQUITY/TA	Leverage
X9	(EQUITY-ITGA)/	Leverage
	(TA-ITGA-CASH-LB)	Leverage
X10	CL/TA	Leverage
X11	(CL-CASH)/TA	Leverage
X12	TL/TA	Leverage
X13	DEBT/TA	Leverage
X14	EBIT/INTE	Leverage
X15	CASH/TA	Liquidity
X16	CASH/CL	Liquidity
X17	QA/CL	Liquidity
X18	CA/CL	Liquidity
X19	WC/TA	Liquidity
X20	CL/TL	Liquidity
X21	TA/SALE	Activity
X22	INV/SALE	Activity
X23	AR/SALE	Activity
X24	AP/SALE	Activity
X25	Log(TA)	Size
X26	IDINV/INV	Growth
X27	IDL/TL	Growth
X28	IDCASH/CASH	Growth

where

True positive (*TP*):   Predict insolvent companies as insolvent ones
False positive (*FP*):   Predict solvent companies as insolvent ones
True negative (*TN*):   Predict solvent companies as solvent ones
False negative (*FN*):   Predict insolvent companies as solvent ones

The following matrix explains the terms used in the definition of error rates:

		Predicted class	
		Positive	Negative
Actual	Positive	*True positive* (*TP*)	*False negative* (*FN*)
Class	Negative	*False positive* (*FP*)	*True negative* (*TN*)

SVM METHODOLOGY

In recent years, the so-called support vector machines (SVM), which have their roots in the theory of statistical learning (Burges, 1998; Christianini and Shawe-Taylor, 2000; Vapnik, 1995) have

become one of the most successful learning algorithms for classification as well as for regression (Drucker *et al.*, 1997; Mangasarian and Musicant, 2000; Smola and Schölkopf, 2004). Some features of SVM make them particularly attractive for predicting the default risk of companies. SVM are a non-parametric technique that learn the separating function from the data; they are based on a sound theoretical concept, do not require a particular distribution of the data, and deliver an optimal solution for the expected loss from misclassification. SVM estimate the separating hyperplane between defaulting and non-defaulting companies under the constraint of a maximal margin between the two classes (Vapnik, 1995; Schölkopf and Smola, 2002).

SVM can be formulated differently. However, in all variants either a constrained minimization problem or an unconstrained minimization problem is solved. The objective function in these optimization problems basically consists of two parts: a misclassification penalty part which stands for *model bias* and a regularization part which controls the *model variance*. We briefly introduce three different models: the smooth support vector machines (SSVM) (Lee and Mangasarian, 2001), the smooth support vector machines with reduced kernel technique (RSVM) and the 1-norm SVM. The SSVM will be used for classification and the 1-norm SVM will be employed for variable selection. The RSVM are applied for oversampling in order to mitigate the computational burden due to increasing the number of instances in the training sample.

Smooth support vector machines

The aim of the SVM technique is to find the separating hyperplane with the largest margin from the training data. This hyperplane is 'optimal' in the sense of statistical learning: it strikes a balance between overfitting and underfitting. Overfitting means that the classification boundary is too curved and therefore has less ability to classify unseen data correctly. Underfitting, on the other hand, gives a too simple classification boundary and leaves too many misclassified observations (Vapnik, 1995). We begin with linear support vector machines. Given a training dataset $S = \{(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_n, y_n)\} \subseteq \mathbb{R}^d \times \mathbb{R}$, where $\mathbf{x}_i \in \mathbb{R}^d$ is the input data and $y_i \in \{-1, 1\}$ is the corresponding class label, a conventional SVM separating hyperplane is generated by solving a convex optimization problem given as follows:

$$
\min_{(w, b, \xi) \in \mathbb{R}^{d+1+n}} C \sum_{i=1}^{b} \xi_i + \frac{1}{2} \|w\|_2^2
$$
$$
\text{s.t. } y_i(w^\top \mathbf{x}_i + b) + \xi_i \geq 1
$$
$$
\xi_i \geq 0, \quad \text{for } i = 1, 2, \ldots, n
$$
(1)

where C is a positive parameter controlling the trade-off between the training error (model bias) and the part of maximizing the margin (model variance) that is achieved by minimizing $\|w\|_2^2$. In contrast to the conventional SVM of (1), smooth support vector machines minimize the square of the slack vector ξ with weight $\frac{C}{2}$. In addition, the SSVM methodology appends $\frac{b^2}{2}$ to the term that is to be minimized. This expansion results in the following minimization problem:

$$
\min_{(w, b, \xi) \in \mathbb{R}^{d+1+n}} \frac{C}{2} \sum_{i=1}^{n} \xi_i^2 + \frac{1}{2}(\|w\|_2^2 + b^2)
$$
$$
\text{s.t. } y_i(w^\top \mathbf{x}_i + b) + \xi_i \geq 1
$$
$$
\xi_i \geq 0, \quad \text{for } i = 1, 2, \ldots, n
$$
(2)

In a solution of (2), ξ is given by $\xi_i = \{1 - y_i(w^\top \mathbf{x}_i + b)\}_+$ for all i where the *plus* function x_+ is defined as $x_+ = \max\{0, x\}$. Thus, we can replace ξ_i in (2) by $\{1 - y_i(w^\top \mathbf{x}_i + b)\}_+$. This will convert the problem (2) into an unconstrained minimization problem as follows:

$$\min_{(w,\,b)\in\mathbb{R}^{d+1}} \frac{C}{2} \sum_{i=1}^{n} \{1 - y_i(w^\top \mathbf{x}_i + b)\}_+^2 + \frac{1}{2}(\|w\|_2^2 + b^2) \tag{3}$$

This formulation reduces the number of variables from $d + 1 + n$ to $d + 1$. However, the objective function to be minimized is not twice differentiable, which precludes the use of a fast Newton method. In the SSVM, the plus function x_+ is approximated by a smooth *p-function*, $p(x, \alpha) = x + \frac{1}{\alpha} \log(1 + e^{-\alpha x}), \alpha > 0$. Replacing the plus function with a very accurate smooth approximation *p*-function gives the smooth support vector machine formulation:

$$\min_{(w,\,b)\in\mathbb{R}^{d+1}} \frac{C}{2} \sum_{i=1}^{n} p(\{1 - y_i(w^\top \mathbf{x}_i + b)\}, \alpha)^2 + \frac{1}{2}(\|w\|_2^2 + b^2) \tag{4}$$

where $\alpha > 0$ is the smooth parameter. The objective function in problem (4) is strongly convex and infinitely differentiable. Hence, it has a unique solution and can be solved by using a fast Newton–Armijo algorithm. For the nonlinear case, this formulation can be extended to the nonlinear SVM by using the kernel trick as follows:

$$\min_{(u,\,b)\in\mathbb{R}^{n+1}} \frac{C}{2} \sum_{i=1}^{n} p\left(\left[1 - y_i\left\{\sum_{j=1}^{n} u_j K(\mathbf{x}_i, \mathbf{x}_j) + b\right\}\right], \alpha\right)^2 + \frac{1}{2}(\|u\|_2^2 + b^2) \tag{5}$$

where $K(\mathbf{x}_i, \mathbf{x}_j)$ is a kernel function. This kernel function represents the inner product of $\phi(\mathbf{x}_i)$ and $\phi(\mathbf{x}_j)$, where ϕ is a certain mapping from input space $\mathbb{R}^d$ to a feature space $\mathcal{F}$. We do not need to know the mapping of ϕ explicitly. This is the so-called kernel trick. The nonlinear SSVM classifier can be expressed in matrix form as follows:

$$\sum_{u_j \neq 0} u_j K(A_j^\top, \mathbf{x}) + b = K(\mathbf{x}, A^\top)u + b \tag{6}$$

where $A = [\mathbf{x}_1^\top; \ldots; \mathbf{x}_n^\top]$ and $A_j = \mathbf{x}_j^\top$.

Reduced support vector machine

In large-scale problems, the full kernel matrix will be very large so it may not be appropriate to use the full kernel matrix when dealing with (5). In order to avoid facing such a big full kernel matrix, we brought in the reduced kernel technique (Lee and Huang, 2007). The key idea of the reduced kernel technique is to randomly select a portion of data and to generate a thin rectangular kernel matrix, then to use this much smaller rectangular kernel matrix to replace the full kernel matrix. In the process of replacing the full kernel matrix by a reduced kernel, we use the Nyström approximation (Smola and Schölkopf, 2000) for the full kernel matrix:

$$K(A, A^\top) \approx K(A, \tilde{A}^\top)K(\tilde{A}, \tilde{A}^\top)^{-1}K(\tilde{A}, A^\top) \tag{7}$$

where $K(A, A^\top) = K_{n \times n}$, $\tilde{A}_{\tilde{n} \times d}$ is a subset of A and $K(A, \tilde{A}) = \tilde{K}_{n \times \tilde{n}}$ is a reduced kernel. Thus, we have

$$K(A, A^\top)u \approx K(A, \tilde{A}^\top)K(\tilde{A}, \tilde{A}^\top)^{-1}K(\tilde{A}^\top, A)u = K(A, \tilde{A}^\top)\tilde{u} \tag{8}$$

where $\tilde{u} \in \mathbb{R}^{\tilde{n}}$ is an approximated solution of u via the reduced kernel technique. The reduced kernel method constructs a compressed model and cuts down the computational cost from $\mathcal{O}(n^3)$ to $\mathcal{O}(\tilde{n}^3)$. It has been shown that the solution of reduced kernel matrix approximates the solution of full kernel matrix well. The SSVM with the reduced kernel are called RSVM.

1-Norm support vector machine
The 1-norm support vector machine replaces the regularization term $\|w\|_2^2$ in (1) with the ℓ_1-norm of w. The ℓ_1-norm regularization term is also called the LASSO penalty (Tibshirani, 1996). It tends to shrink the coefficients w's towards zeros in particular for those coefficients corresponding to redundant noise features (Zhu *et al.*, 2003; Williams and Seeger, 2001). This nice feature will lead to a way of selecting the important ratios in our prediction model. The formulation of 1-norm SVM is described as follows:

$$\min_{(w, b, \xi) \in \mathbb{R}^{d+1+n}} C\sum_{i=1}^{n} \xi_i + \|w\|_1$$
$$\text{s.t.} \quad y_i(w^\top \mathbf{x}_i + b) + \xi_i \geq 1 \tag{9}$$
$$\xi_i \geq 0, \quad \text{for } i = 1, 2, \ldots, n.$$

The objective function of (9) is a piecewise linear convex function. We can reformulate it as the following linear programming problem:

$$\min_{(w, s, b, \xi) \in \mathbb{R}^{d+d+1+n}} C\sum_{i=1}^{n} \xi_i + \sum_{j=1}^{d} s_j$$
$$\text{s.t.} \quad y_i(w^\top \mathbf{x}_i + b) + \xi_i \geq 1 \tag{10}$$
$$-s_j \leq w_j \leq s_j, \quad \text{for } j = 1, 2, \ldots, d,$$
$$\xi_i \geq 0, \quad \text{for } i = 1, 2, \ldots, n$$

where s_j is the upper bound of the absolute value of w_j. In the optimal solution of (10) the sum of s_j is equal to $\|w\|_1$.

The 1-norm SVM can generate a very sparse solution w and lead to a parsimonious model. In a linear SVM classifier, solution sparsity means that the separating function $f(\mathbf{x}) = w^\top \mathbf{x} + b$ depends on very few input attributes. This characteristic can significantly suppress the number of nonzero coefficient w's, especially when there are many redundant noise features (Fung and Mangasarian, 2004; Zhu *et al.*, 2003). Therefore the 1-norm SVM can be a very promising tool for the variable selection tasks. We will use it to choose the important financial indices for our bankruptcy prognosis model.

SELECTION OF ACCOUNTING RATIOS

In principle any possible combination of accounting ratios could be used as explanatory variables in a bankruptcy prognosis model. Therefore, appropriate performance measures are needed to gear the process of variable selection towards picking the ratios with the highest separating power. In

1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46

Chen *et al.* (2006) accuracy ratio (AR) and conditional information entropy ratio (CIER) determine the selection procedure's outcome. It turned out that the ratio 'accounts payable divided by sales', X24 (AP/SALE), has the best performance values for a univariate SVM model. The second selected variable was the one combined with X24 that had the best performance in a bivariate SVM model. This is the analogue of forward selection in linear regression modeling. Typically, improvement declines if new variables are added consecutively. In Chen *et al.* (2006) the performance indicators started to decrease after the model included eight variables. The described selection procedure is quite lengthy, since there are at least 216 accounting ratio combinations to be considered. We will not employ the procedure here but use the chosen set of eight variables as the benchmark set V1. Table II presents V1 in the first column.

We propose two different approaches for variable selection that will simplify the selection procedure. The first one is based on 1-norm SVM introduced above. The SVM were applied to the period from 1997 to 1999. We selected the variables according to the size of the absolute values of the coefficients *w* from the solution of the 1-norm SVM. Table II displays the eight selected variables as V2. We obtain eight variables out of 28. Note that five variables, X2, X3, X5, X15 and X24, are also in the benchmark set V1.

The second variable selection scheme is incremental forward variable selection. The intuition behind this scheme is that a new variable will be added into the already selected set, if it brings in the most extra information. We measure the extra information for an accounting ratio using the distance between this new ratio vector and the space spanned by the current selected ratio subset. This distance can be computed by solving a least-squares problem (Lee *et al.*, 2008). The ratio with the farthest distance will be added into the selected accounting ratio set. We repeat this procedure until a certain stopping criterion is satisfied. The accounting ratio X24 (AP/SALE) is used as the initial selected accounting ratio. Then we follow the procedure seven times to select seven more extra accounting ratios. The variable set generated is called V3. We will use these three variable sets, V1, V2 and V3, for further data analysis in the next section. The symbol [+] denotes the variables that are common to all sets: X2, X3, X5 and X24.

Table II. Selected variables

Variable	Definition	V1	V2	V3
X2[+]	NI/SALE	x	x	x
X3[+]	OI/TA	x	x	x
X4	OI/SALE			x
X5[+]	EBIT/TA	x	x	x
X6	(EBIT + AD)/TA		x	
X7	EBIT/SALE			x
X8	EQUITY/TA		x	
X12	TL/TA	x		
X13	DEBT/TA			x
X15	CASH/TA	x	x	
X21	TA/SALE			x
X22	INV/SALE	x		
X23	AR/SALE		x	
X24[+]	AP/SALE	x	x	x
X26	IDINV/INV	x		

EXPERIMENTAL SETTING AND RESULTS

In this section we present our experimental setting and results. We compare the performance of three sets of accounting ratios, V1, V2 and V3, in our SSVM-based insolvency prognosis model. The performance is measured by Type I error rate, Type II error rate and total error rate. Fortunately, in reality, there is only a small number of insolvent companies compared to the number of solvent companies. Due to the small share in a sample that reflects reality, a simple classification such as naive Bayesian or a decision tree tends to classify every company as solvent. Such a classification would imply accepting all companies' loan applications and would thus lead to a very high Type I error rate while the total error rate and the Type II error rate are very small. Such models are useless in practice.

Our cleaned dataset consists of around 10% of insolvent companies. Thus, the sample is fairly unbalanced although the share of insolvent companies is higher than in reality. In order to deal with this problem, insolvency prognosis models usually start off with more balanced training and testing samples than reality can provide. For example, Härdle *et al.* (2007b) employ a downsampling strategy and work with balanced (50%/50%) samples. The chosen bootstrap procedure repeatedly randomly selects a fixed number of insolvent companies from the training set and adds the same number of randomly selected solvent companies. However, in this paper we adopt an oversampling strategy, to balance the size between the solvent and the insolvent companies, and refer to the downsampling procedure primarily for reasons of reference.

Oversampling duplicates the number of insolvent companies a certain number of times. In this experiment, we duplicate in each scenario the number of insolvent companies as many times as necessary to reach a balanced sample. Note that in our oversampling scheme every solvent and insolvent company's information is utilized. This increases the computational burden due to increasing the number of training instances. We employ the reduced kernel technique introduced above to mediate this problem.

All classifiers we need in these experiments are reduced SSVM with the Gaussian kernel, which is defined as

$$K(\mathbf{x}, \mathbf{z}) = e^{-\gamma \|\mathbf{x} - \mathbf{z}\|_2^2}$$

where γ is the width parameter. In nonlinear SSVM, we need to determine two parameters: the penalty term C and γ. The 2D grid search will consume a lot of time. In order to cut down the search time, we adopt the uniform design model selection method (Huang *et al.*, 2007) to search an appropriate pair of parameters.

Performance of SSVM
We conduct the experiments in a scenario in which we always train the SSVM bankruptcy prognosis model from the data at hand and then use the trained SSVM to predict the following year's cases. This strategy simulates the real task of prediction which binds the analyst to use past data for forecasting future outcomes. The experimental setting is described in Table III. The number of periods which enter the training set changes from 1 year (S1) to 5 years (S5).

In Tables IV and V we report the results for the oversampling and downsampling strategy respectively. Mean and standard deviation of Type I, Type II and total error rates (misclassification rates) are shown. We perform these experiments for the three variable sets, V1 to V3, and compare the oversampling and downsampling scheme in each experiment. All experiments are repeated 30 times

Table III. The scenario of our experiments

Scenario	Observation period of training set	Observation period of testing set
S1	1997	1998
S2	1997–1998	1999
S3	1997–1999	2000
S4	1997–2000	2001
S5	1997–2001	2002

Table IV. Results as a percentage (%) of oversampling for three variable sets (RSVM)

Set of accounting ratios	Scenario	Type I error rate		Type II error rate		Total error rate	
		Mean	SD	Mean	SD	Mean	SD
V1	S1	33.16	0.55	26.15	0.13	26.75	0.12
	S2	31.58	0.01	29.10	0.07	29.35	0.07
	S3	28.11	0.73	26.73	0.16	26.83	0.16
	S4	30.14	0.62	25.66	0.17	25.93	0.15
	S5	24.24	0.56	23.44	0.13	23.48	0.13
V2	S1	29.28	0.92	27.20	0.24	27.38	0.23
	S2	28.20	0.29	30.18	0.18	29.98	0.16
	S3	27.41	0.61	29.67	0.19	29.50	0.17
	S4	28.12	0.74	28.32	0.19	28.31	0.15
	S5	23.91	0.62	24.99	0.10	24.94	0.10
V3	S1	29.28	0.83	25.11	0.25	25.46	0.21
	S2	31.27	0.62	29.79	0.34	29.94	0.35
	S3	30.91	0.13	27.21	0.19	27.48	0.18
	S4	32.00	0.54	25.19	0.17	25.61	0.14
	S5	26.98	0.42	22.90	0.11	23.08	0.11

Table V. Results as a percentage (%) of downsampling for three variable sets (SSVM with Gaussian kernel)

Set of accounting ratios	Scenario	Type I error rate		Type II error rate		Total error rate	
		Mean	SD	Mean	SD	Mean	SD
V1	S1	32.20	3.12	28.98	1.70	29.26	1.46
	S2	29.74	2.29	28.77	1.97	28.87	1.57
	S3	30.46	1.88	26.23	1.33	26.54	1.17
	S4	31.55	1.52	23.89	0.97	24.37	0.87
	S5	28.81	1.53	23.09	0.73	23.34	0.69
V2	S1	29.94	2.91	28.07	2.15	28.23	1.79
	S2	28.77	2.58	29.80	1.89	29.70	1.52
	S3	29.88	1.88	27.19	1.32	27.39	1.19
	S4	29.06	1.68	26.26	1.00	26.43	0.86
	S5	26.92	1.94	25.30	1.17	25.37	1.06
V3	S1	30.87	3.25	26.61	2.45	26.98	2.11
	S2	33.31	2.16	28.60	2.01	29.08	1.65
	S3	31.82	1.52	26.41	1.45	26.80	1.31
	S4	35.0	2.13	24.29	0.77	24.96	0.68
	S5	30.66	1.60	21.92	0.96	22.30	0.92

because of the randomness in the experiments. The randomness is very obvious in the downsampling scheme (see Table V). Each time we only choose negative instances with the same size of the whole positive instances. The observed randomness in our oversampling scheme (Table IV) is due to applying the reduced kernel technique to solving the problem. We use the training set in the downsampling scheme as the reduced set. That is, we use all the insolvent instances and the equal number of solvent instances as our reduced set in generating the reduced kernel. Then we duplicate the insolvent part of the kernel matrix to balance the size of insolvent and solvent companies.

Both tables reveal that different variable selection schemes produce dissimilar results with respect to both precision and deviation of predicting. The oversampling scheme shows better results in the Type I error rate but has slightly bigger total error rates. It is also obvious that in almost all models a longer training period works in favor of accuracy of prediction. Clearly, the oversampling schemes have much smaller standard deviations in the Type I error rate, Type II error rate, and total error rate than the downsampling one. According to this observation, we conclude that the oversampling scheme will generate a more robust model than the downsampling scheme.

Figure 3 illustrates the development (learning curve) of the Type I error rate and total error rate with regard to variable set V3 for both oversampling and downsampling. The bullets on the lines
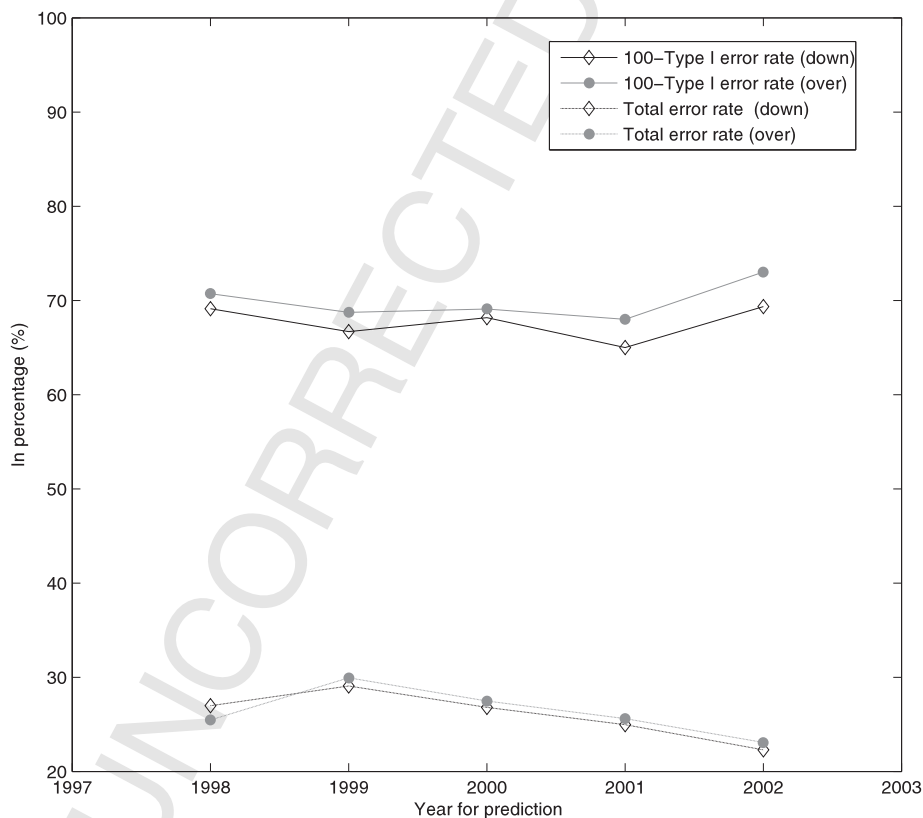
Figure 3. Learning curve for variables set V3

mark the different training scenarios. For example, the first bullets from the left represent S1 (training set from 1997, testing set from 1998), the second bullets illustrate S2 (training set from 1997 to 1998, testing set from 1999) etc. For the purpose of better visibility, the Type I error rate is only indirectly displayed as 100 − Type I error rate. The upper solid line in gray represents the oversampling scheme and the black solid line the downsampling one. Note that the performance in terms of the Type I error rate is worse the higher the distance between the upper end of the diagram and the solid lines. The learning curve over the time frame the training sample covers shows an upward tendency between S1 and S5 for the number 100 − Type 1 error rate. However, the curves are non-monotonic. There is a disturbance for the forecast of year 1999 that is based on training samples that cover 1997 to 1998, and also one for the forecast of year 2001 based on training samples covering 1997 to 2000. Both disturbances may have been caused by the reform of the German insolvency code that came into force in 1999. The most important objective of the reform was to allow for more company restructuring and less liquidation than before. This reform considerably changed the behavior of German companies towards declaring insolvency, and thus most likely the nature of balance sheets that are associated with insolvent companies.

The disturbances are less visible with respect to the overall performance. The dashed lines near the lower edge of the diagram box show total error rates, gray for the oversampling and black for the downsampling scheme. There is a clear tendency towards a lower total error rate from S2 to S5 for both schemes. The downsampling line is slightly below the oversampling one, representing a slightly better performance in terms of the mean of the total error rate. However, this result has to be seen in the light of the trade-off between magnitude and stability of results, as oversampling yields much more stable results. The standard deviations for V3 are only a small portion of the numbers generated by the downsampling procedure across all training scenarios (Tables IV and V).

Table VI presents the comparison between the sets by focusing on the total error rate. It indicates by an asterisk whether the differences in means are significant at the 10% level via *t*-test and, in addition, gives the set which is superior in the dual comparison. Variable set V2 is nearly absent in Table VI. Thus V2 is clearly outperformed by both sets V1 and V3. There is no clear distinction between V1 and V3 except for Scenario S5. Given the long training period V3 is superior in both the downsampling and oversampling scenarios and generates the lowest total error rate in absolute terms.

In order to investigate the effect of the oversampling versus the downsampling scheme we follow the setting as above, but we use the V3 variable set. For each training–testing pair, we carry out oversampling for positive instances from 6 to 15 time. We show the trend and effect in Figure 4. It

Table VI. Statistical significance in differences in means (10% level) between the three variable sets: total error

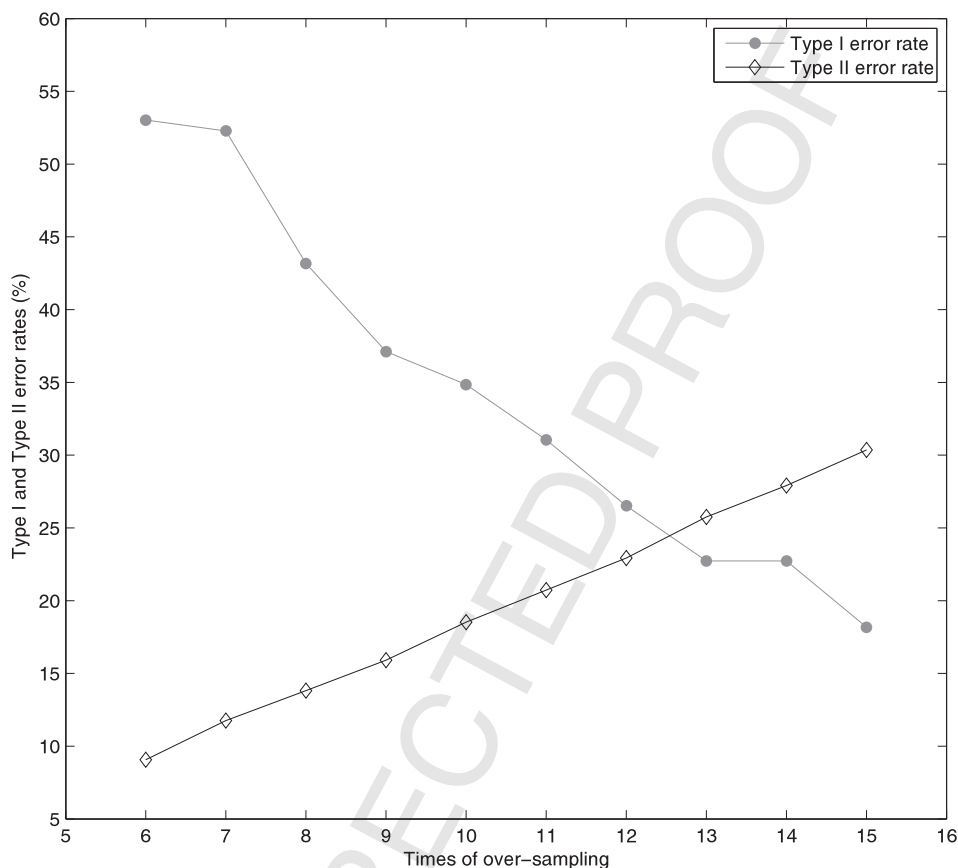| Sets	S1	S2	S3	S4	S5
Oversampling					
V1 vs. V2	V1*	V1*	V1*	V1*	V1*
V1 vs. V3	V3*	V1*	V1*	V3*	V3*
V2 vs. V3	V3*		V3*	V3*	V3*
Downsampling					
V1 vs. V2	V2*	V1*	V1*	V1*	V1*
V1 vs. V3	V3*			V1*	V3*
V2 vs. V3	V3*		V3*	V3*	V3*

Figure 4. The effect of oversampling on Type I and Type II error rates for scenario S5 and variables set V3

is easy to see that the Type I (II) error rate decreases (increases) as the oversampling time increases. This feature implies that the machine would have a tendency of classifying all companies as solvent if the training sample had realistic shares of insolvent and solvent companies. Such behavior would produce a Type I error rate of 100%. The more balanced the sample is, the higher the penalty for classifying insolvent companies as solvent. This fact is illustrated in Figure 4 by the decreasing curve with respect to the number of duplications of insolvent companies.

Often banks favor a strategy that allows them to minimize the Type II errors for a given number of Type I errors. The impact of oversampling on the trade-off between the two types of errors—shown in Figure 4—implies that the number of oversampling times is a strategic variable in training the machine. This number can be determined by the bank's aim regarding the relation of Type I and Type II errors.

Comparison with logit and linear discriminant analysis

The examination of SSVM is incomplete without comparing it to highly used traditional methods such as the logistic model (LM) and linear discriminant analysis (DA). Therefore, we replicate the research design of the previous section with both traditional models. In addition, we test whether the difference in means in the total error rate is statistically significant. The comparison of means with regard to the total error rate is presented in Tables VII and VIII for the oversampling and downsampling strategy respectively. Table IX summarizes the comparison of the approaches and displays the statistical significance of their mean differences. Asterisks indicate the out-performance

Table VII. Comparison of the total error rate (%) as generated by SSVM with LM and DA: oversampling for three variable sets

Set of accounting ratios	Scenario	SSVM Mean	LM Mean	DA Mean
V1	S1	26.75	26.50	25.60
	S2	29.35	28.96	27.22
	S3	26.83	28.94	27.42
	S4	25.93	26.20	25.55
	S5	23.48	26.95	28.23
V2	S1	27.38	26.80	26.20
	S2	29.98	28.63	28.70
	S3	29.50	29.52	29.46
	S4	28.31	28.43	28.08
	S5	24.94	29.22	31.42
V3	S1	25.46	25.07	23.65
	S2	29.94	28.29	27.02
	S3	27.48	27.89	25.84
	S4	25.61	26.60	24.85
	S5	23.08	25.32	26.15

Table VIII. Comparison of the total error rate (%) as generated by SSVM with LM and DA: downsampling for three variable sets

Set of accounting ratios	Scenario	SSVM Mean	LM Mean	DA Mean
V1	S1	29.26	26.86	27.34
	S2	28.87	28.62	28.26
	S3	26.54	27.54	28.22
	S4	24.37	24.80	25.47
	S5	23.34	24.81	25.86
V2	S1	28.23	27.28	28.62
	S2	29.70	29.29	29.65
	S3	27.39	28.56	29.58
	S4	26.43	26.41	27.96
	S5	25.37	26.52	29.69
V3	S1	26.98	26.03	25.47
	S2	29.08	28.04	27.22
	S3	26.80	26.60	26.51
	S4	24.96	25.25	25.44
	S5	22.30	23.96	24.31

Table IX. Statistical significance in differences of means (10% level) between SSVM and LM and SSVM and DA, respectively, for the sets V1 to V3: total error rate

V1	S1	S2	S3	S4	S5
Oversampling					
SSVM vs. LM			*	*	*
SSVM vs. DA			*		*
Downsampling					
SSVM vs. LM			*	*	*
SSVM vs. DA			*	*	*
V2	**S1**	**S2**	**S3**	**S4**	**S5**
Oversampling					
SSVM vs. Logit				*	*
SSVM vs. DA					*
Downsampling					
SSVM vs. Logit			*		*
SSVM vs. DA			*	*	*
V3	**S1**	**S2**	**S3**	**S4**	**S5**
Oversampling					
SSVM vs. LM			*	*	*
SSVM vs. DA					*
Downsampling					
SSVM vs. LM					*
SSVM vs. DA				*	*

of the logistic model or discriminant analysis by SSVMs at the 10% level via *t*-test. It is obvious that the SSVM technique yields the better results, the longer the period is from which the training observations are taken. In fact, the results show that the SSVM works significantly better than LM and DA in most cases in S3 to S5, with the clearest advantage for testing sets S4 and S5, where the accounting information of the predicted companies dates most frequently in 2001 and 2002.

We also investigate the effect of oversampling on LM and DA. We follow the same setting in the previous section, doing oversampling for positive instances from 6 to 15 times. Unlike the SSVM-based insolvency prognosis model, the DA approach is insensitive in both Type I and Type II error rates to the replication of positive instances. The result for DA is illustrated in Figure 5. The LM approach has very similar results to the SSVM model. We will not show the result here.

More data visualization
Each SSVM model has its own output value. We use this output to construct 2D coordinate systems. Figure 6 shows an example for scenario S5 where the scores of the $SSVM_{V3}$ model ($SSVM_{V1}$ model) are represented by the horizontal (vertical) line. A positive (negative) value indicates predicted solvency (insolvency). We then map all insolvent companies in the testing set onto the coordinate systems. There are 132 insolvent companies and 2866 solvent companies in this testing set. We also randomly choose the same amount of solvent companies from the testing set.

The plus points in the lower left quadrant and the circle points in the upper right quadrant show the number of Type I errors and Type II errors, respectively, in both models. Plus points in the upper right quadrant and circle points in the lower left quadrant reflect those companies that are predicted
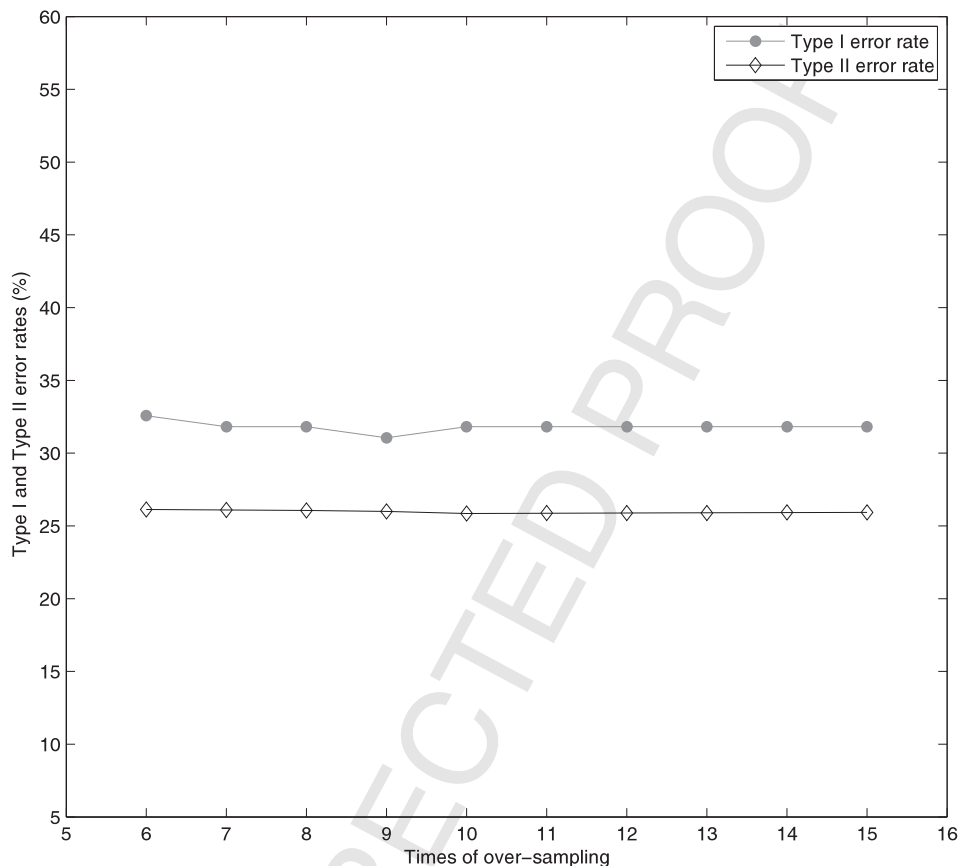
Figure 5. The effect of oversampling on Type I and Type II error rates for scenario S5 and variables set V3 in DA

correctly by both models. Circles and plus points in the lower right quadrant (upper left quadrant) represent conflicting prognoses. We also report the number of insolvent companies and the number of solvent companies in each quadrant of Figure 6. The two different insolvency prognosis models based on V1 and V3, respectively, can be considered as alternative experts. The two forecasts for each instance in the testing set is plotted in the diagram. The proposed visualization scheme could be used to support loan officers in their final decision on accepting or rejecting a client's application. Furthermore, this data visualization scheme can also be applied to two different learning algorithms, such as $SSVM_{V3}$ vs. LM_{V3} and $SSVM_{V3}$ vs. DA_{V3}. We show these data visualization plots in Figures 7 and 8. If the loan application has been classified as solvent or insolvent by alternative machines, it is most likely that the prognosis meets reality (the plus points in the upper right quadrant and the circle points in the lower left quadrant). Opposing forecasts, however, should be taken as a hint to evaluate the particular company more thoroughly, for example by employing an expert team, or even by using a third model.
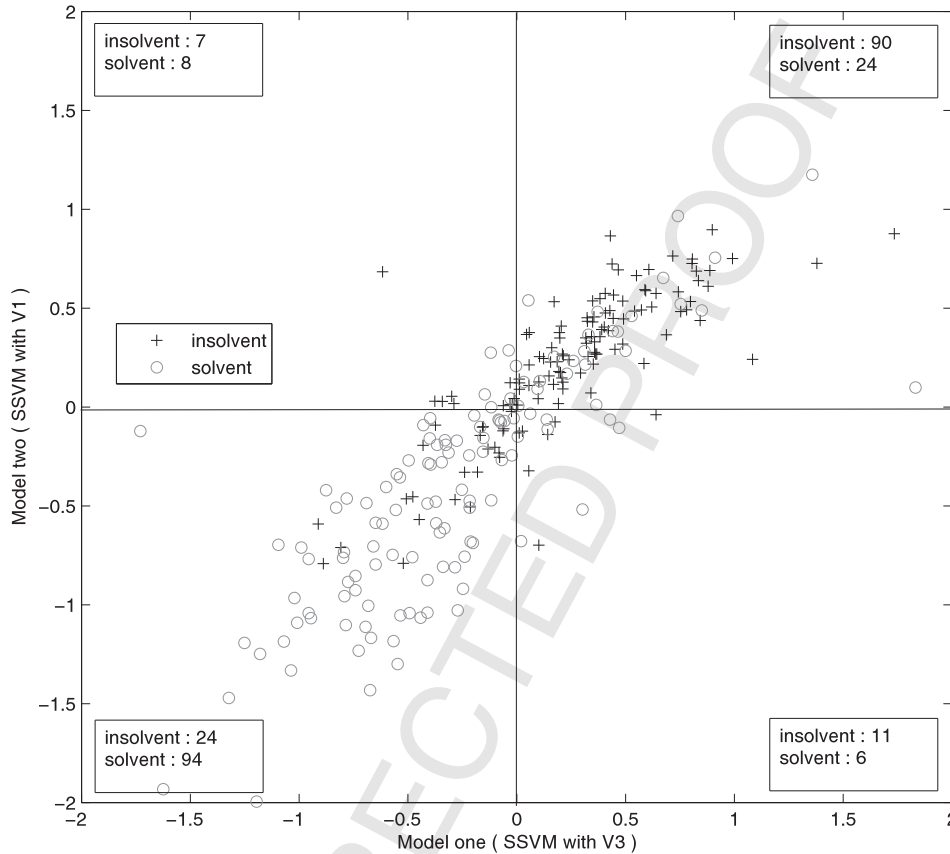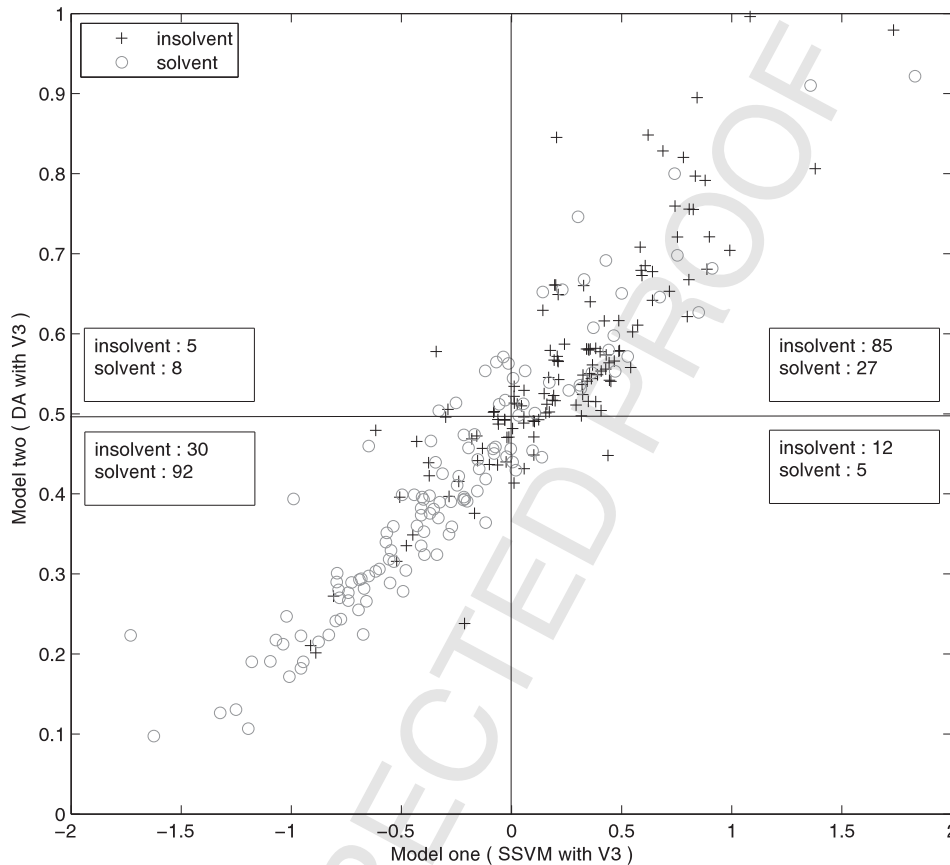
Figure 6. Data visualization via model one (generated by SSVM with V3) and model two (generated by SSVM with V1) in scenario S5

CONCLUSION

In this paper we apply different variants of support vector machines to a unique dataset of German solvent and insolvent companies. We use a priori a given set of predictors as a benchmark, and suggest two further variable selection procedures; the first procedure uses the 1-norm SVM and the second, incremental way consecutively selects the variable that is the farthest one from the column space of the current variable set. Given the three SSVM based on distinct variable sets, the relative performance of the types of smooth support vector machines is tested. The performance is measured by error rates. The two sets of variables newly created here lead to a dissimilar performance of SSVM. The selection of variables by the 1-norm SVM clearly underperforms compared to the incremental selection scheme. This difference in accuracy hints at the need for further research with respect to the variable selection. The training period makes a clear difference, though. Results improve considerably if more years of observation are used in training the machine. The SSVM
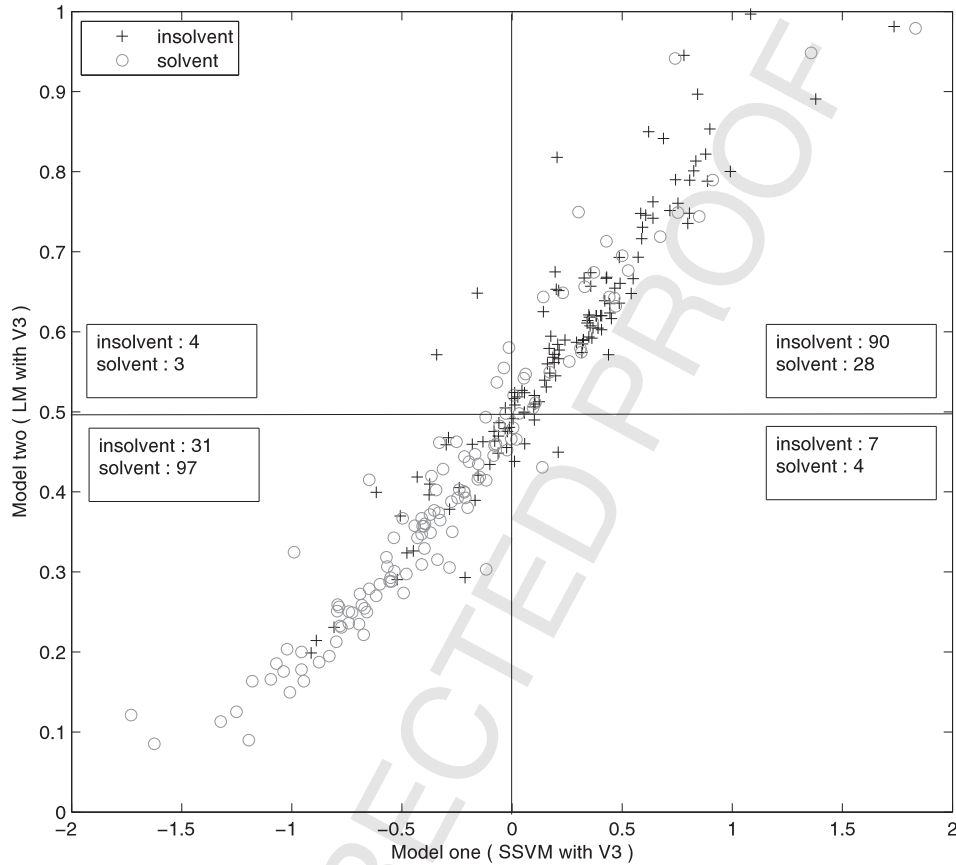
Figure 7. Data visualization via model one (generated by SSVM with V3) and model two (generated by LM with V3) in scenario S5

model benefits more from longer training periods than traditional methods do. As a consequence the logit model and discriminant analysis are both outperformed by the SSVM in long-term training scenarios. Moreover, the oversampling scheme works very well in dealing with unbalanced datasets. It provides flexibility to control the trade-off between Type I and Type II errors, and is therefore a strategic instrument in a bank's hand. The results generated are very stable in terms of small deviations of Type I, Type II and total error rates.

Finally, we want to stress that SSVM should be considered not as a substitute for traditional methods but rather as a complement which, when employed side by side with either the logit model or discriminant analysis, can generate new information that helps practitioners select those companies that are difficult to predict and, therefore, need more attention and further treatment.

Figure 8, Data visualization via model one (generated by SSVM with V3) and model two (generated by DA with V3) in scenario S5

REFERENCES

Altman E. 1968. Financial ratios, discriminant analysis and the prediction of corporate bankruptcy. *Journal of Finance* **23**(4): 589–609.
Altman E, Marco G, Varetto F. 1994. Corporate distress diagnosis: comparisons using linear discriminant analysis and neural networks (the italian experience). *Journal of Banking and Finance* **18**: 505–529.

J. Forecast. (2008)
                                                                      DOI: 10.1002/for

Beaver W. 1966. Financial ratios as predictors of failures: empirical research in accounting: selected studies. *Journal of Accounting Research* **4**: 71–111.

Burges CJC. 1998. A tutorial on support vector machines for pattern recognition. *Data Mining and Knowledge Discovery* **2**(2): 121–167.

Chen S, Härdle W, Moro RA. 2006. Estimation of default probabilities with support vector machines. *SFB 649 Discussion Paper 2006-077*.

Cristianini N, Shawe-Taylor J. 2000. *An Introduction to Support Vector Machines*. Cambridge University Press: Cambridge, UK.

Drucker H, Burges CJC, Kaufman L, Smola A, Vapnik V. 1997. Support vector regression machines. In *Advances in Neural Information Processing Systems 9*, Mozer MC, Jordan MI, Petsche T (eds). MIT Press: Cambridge, MA; ••–••.

Fung G, Mangasarian OL. 2004. A feature selection Newton method for support vector machine classification. *Computational Optimization and Applications* **28**(2): 185–202.

Härdle W, Moro R, Schäfer D. 2007a. Graphical data representation in bankruptcy analysis based on support vector machines. In *Handbook of Data Visualization*, Chen C, Hrdle W, Unwin A (eds). Springer: Heidelberg; ••–••.

Härdle W, Moro RA, Schäfer D. 2007b. Estimating probabilities of default with support vector machines. *SFB 649 Discussion Paper 2007-035*.

Huang CM, Lee YJ, Lin DKJ, Huang SY. 2007. Model selection for support vector machines via uniform design. *Computational Statistics and Data Analysis*, Special Issue on Machine Learning and Robust Data Mining (to appear).

Hwang RC, Cheng KF, Lee JC. 2007. A semiparametric method for predicting bankruptcy. *Journal of Forecasting* **26**(5): 317–342.

Krahnen JP, Weber M. 2001. Generally accepted rating principles: a primer. *Journal of Banking and Finance* **25**(1): 3–23.

Lee YJ, Huang SY. 2007. Reduced support vector machines: a statistical theory. *IEEE Transactions on Neural Networks* **18**: 1–13.

Lee YJ, Mangasarian OL. 2001. SSVM: a smooth support vector machine. *Computational Optimization and Applications* **20**: 5–22.

Lee YJ, Chang CC, Chao CH. 2008. Incremental forward feature selection with application to microarray gene expression. *Journal of Biopharmaceutical Statistics* (to appear).

Leland H, Toft K. 1996. Optimal capital structure, endogenous bankruptcy, and the term structure of credit spreads. *Journal of Finance* **51**: 987–1019.

Longstaff FA, Schwartz ES. 1995. A simple approach to valuating risky fixed and floating rate debt. *Journal of Finance* **50**: 789–819.

Mangasarian OL, Musicant DR. 2000. Robust linear and support vector regression. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **22**(9): 950–955.

Martin D. 1977. Early warning of bank failure: a logit regression approach. *Journal of Banking and Finance* **1**: 249–276.

Mella-Barral P, Perraudin W. 1997. Strategic debt service. *Journal of Finance* **52**: 531–556.

Merton R. 1974. On the pricing of corporate debt: the risk structure of interest rates. *Journal of Finance* **29**(2): 449–470.

Ohlson J. 1980. Financial ratios and the probabilistic prediction of bankruptcy. *Journal of Accounting Research* **18**(1): 109–131.

Schölkopf B, Smola AJ. 2002. *Learning with Kernels*. MIT Press: Cambridge, MA.

Smola A, Schölkopf B. 2000. Sparse greedy matrix approximation for machine learning. In *Proceedings of the 17th International Conference on Machine Learning*, San Francisco, CA.

Smola A, Schölkopf B. 2004. A tutorial on support vector regression. *Statistics and Computing* **14**: 199–222.

Tam K, Kiang M. 1992. Managerial application of neural networks: the case of bank failure prediction. *Management Science* **38**(7): 926–947.

Tibshirani R. 1996. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society* **58**(1): 267–288.

Vapnik VN. 1995. *The Nature of Statistical Learning Theory*. Springer: New York.

Williams CKI, Seeger M. 2001. Using the Nyström method to speed up kernel machines. *Advances in Neural Information Processing Systems* **13**: 682–688.

Zhou C. 2001. The term structure of credit spreads with jump risk. *Journal of Banking and Finance* **25**: 2015–2040.

Zhu J, Rosset S, Hastie T, Tibshirani R. 2003. 1-Norm support vector machines. In *Advances in Neural Information Processing Systems* ••: ••–••.

Authors' biographies:
Wolfgang Härdle ••.

Yuh-Jye Lee ••.

Dorothea Schäfer ••.

Yi-Ren Yeh ••.

Authors' addresses:
Wolfgang Härdle, Center for Applied Statistics and Economics, Humboldt-Universität zu Berlin, Spandauer Straße 1, 10178 Berlin, Germany.

Yuh-Jye Lee and **Yi-Ren Yeh**, Department of Computer Science Information Engineering, National Taiwan University of Science and Technology, Taipei 106, Taiwan.

Dorothea Schäfer, German Institute for Economic Research (DIW) Berlin, Mohrenstrasse 58, 10117 Berlin, Germany.

7

9

1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46

robust econometrics

Econometric data are often obtained under conditions that cannot be well controlled, and so partial departures from the model assumptions in use (data contamination) occur relatively frequently. To address this, we first introduce concepts of robust statistics for qualifying and quantifying sensitivity of estimation methods to data contamination as well as important approaches to robust estimation. Later, we discuss how robust estimation methods have been adapted to various areas of econometrics, including time series analysis and general GMM-based estimation.

Econometrics often deals with data under, from the statistical point of view, non-standard conditions such as heteroscedasticity or measurement errors, and the estimation methods thus need either to be adapted to such conditions or to be at least insensitive to them. Methods insensitive to violation of certain assumptions – for example, insensitive to the presence of heteroscedasticity – are in a broad sense referred to as 'robust' (for example, robust to heteroscedasticity). On the other hand, there is also a more specific meaning of the word 'robust', which stems from the field of robust statistics. This latter notion defines robustness rigorously in terms of the behaviour of an estimator both at the assumed (parametric) model and in its neighbourhood in the space of probability distributions. Even though the methods of robust statistics have been used only in the simplest settings, such as estimation of location, scale, or linear regression for a long time, they have motivated a range of new econometric methods, which we focus on in this article.

The concepts and measures of robustness are introduced first (Section 1), followed by a most common types of estimation methods and their properties (Section 2). Various econometric methods based on these common estimators are discussed in Section 3, covering tasks from time-series regression over GMM estimation to simulation-based methods.

1 Measures of robustness

Robustness properties can be formulated within two frameworks: qualitative and quantitative robustness. *Qualitative robustness* is concerned with the situation in which the shape of the underlying (true) data distribution deviates slightly from the assumed model. It focuses on questions like stability and performance loss over a family of such slightly deviating distributions. *Quantitative robustness* is involved when the sensitivity of estimators to a proportion of aberrant observations is studied.

A simple example can make this clear. Suppose one has collected a sample on an individual's income (after say ten years of schooling) and one is interested in estimating the mean income. If $\{x_i\}_{i=1}^n$ denotes the logarithm of this data and we suppose that they have a cumulative distribution function (cdf) F, assumed to be $N(\mu, \sigma^2)$, the maximum likelihood estimator (MLE) is $\bar{x} = \int u dF_n(u) = T(F_n)$, where $F_n(u) = n^{-1} \sum_{i=1}^n I(x_i \leq u)$, and $\mu = \int u dF(u) = T(F)$. Qualitative robustness asks here the question: how well will μ be estimated if the true distribution is in some neighbourhood of F? Quantitative robustness would concentrate on the question: will $T(F_n)$ be bounded if some observations $x_i \to \infty$? In fact, the latter question is easy to answer: if $x_i \to \infty$ for some i, $T(F_n) = \bar{x} \to \infty$ as well. So we can say here in a loose sense that $\bar{x}$ is not quantitatively robust.

1.1 Formalities

In the following we present a mathematical set-up that allows us to formalize our thoughts on robustness.

The notion of the sensitivity of an estimator T is put into theory by considering a model characterized by a cdf F and its neighbourhood $\mathscr{F}_{\varepsilon,G}$: distributions $(1 - \varepsilon)F + \varepsilon G$, where $\varepsilon \in (0,1/2)$ and G is an arbitrary probability distribution, which represents data contamination. Hence, not all data necessarily follow the pre-specified distribution, but the ε-part of data can come from a different distribution G. If $H \in \mathscr{F}_{\varepsilon,G}$, the estimation method T is then judged by how sensitive or robust the estimates $T(H)$ are to the size of $\mathscr{F}_{\varepsilon,G}$, or alternatively, to the distance from the assumed cdf F. Two main concepts for robust measures analyse the sensitivity of an estimator to infinitesimal deviations, $\varepsilon \to 0$, and to finite (large) deviations, $\varepsilon > 0$, respectively. Despite generality of the concept, easy interpretation and technical difficulties often limit our choice to point-mass distributions (Dirac measures) $G = \delta_x, x \in \mathbb{R}$, which simply represents an (erroneous) observation at point $x \in \mathbb{R}$. This simplification is also used in the following text.

1.2 Qualitative robustness

The influence of infinitesimal contamination on an estimator is characterized by the *influence function*, which measures the relative change in estimates caused by an infinitesimally small amount ε of contamination at x (Hampel et al., 1986). More formally,

$$IF(x; T, F) = \lim_{\varepsilon \to 0} \frac{T\{(1 - \varepsilon)F + \varepsilon\delta_x\} - T(F)}{\varepsilon}. \tag{1}$$

For each point x, the influence function reveals the rate at which the estimator T changes if a wrong observation appears at x. In the case of sample mean $\bar{x} = T(F_n)$ for $\{x_i\}_{i=1}^n$, we obtain

$$IF(x; T, F_n) = \lim_{\varepsilon \to 0} \left[(1 - \varepsilon) \int u dF_n(u) + \varepsilon \int u d\delta_x(u) - \int u dF_n(u) \right] / \varepsilon$$

$$= \lim_{\varepsilon \to 0} \left[-\int u dF_n(u) + \int u d\delta_x(u) \right] = x - \bar{x}.$$

The influence function allows us to define various desirable properties of an estimation method. First, the largest influence of contamination on estimates can be formalized by the *gross-error sensitivity*,

$$\gamma(T, F) = \sup_{x \in \mathbb{R}} IF(x; T, F), \tag{2}$$

which under robustness considerations is finite and small. Even though such a measure can depend on F in general, the qualitative results (for example, $\gamma(T,F)$ being bounded) are typically independent of F. Second, the sensitivity to small changes in data, for example moving an observation from x to $y \in \mathbb{R}$, can be measured by the *local-shift sensitivity*

$$\lambda(T, F) = \sup_{x \neq y} \frac{||IF(x; T, F) - IF(y; T, F)||}{||x - y||}. \tag{3}$$

Also, this quantity should be relatively small since we generally do not expect that small changes in data cause extreme changes in values or sensitivity of estimates. Third, as unlikely large or distant observations may represent data errors, their influence on estimates should become zero. Such a property is characterized by the *rejection point*,

$$\rho(T, F) = \inf_{r>0} \{r : IF(x; T, F) = 0, ||x|| \geq r\}, \tag{4}$$

which indicates the non-influence of large observations.

1.3 Quantitative robustness

Alternatively, the behaviour of the estimator T can be studied for any finite amount ε of contamination. The most common property looked at in this context is the estimator's bias $b(T, H) = E_H\{T(H)\} - E_F\{T(F)\}$, which measures a distance between the estimates for clean data, $T(F)$, and contaminated data, $T(H), H \in \mathscr{F}_{\varepsilon,G}$. The corresponding *maximum-bias curve* measures the maximum bias of T on $\mathscr{F}_{\varepsilon,G}$ at any ε:

$$B(\varepsilon, T) = \sup_{x \in \mathbb{R}} b\{T, (1 - \varepsilon)F + \varepsilon\delta_x\}. \tag{5}$$

Although the computation of this curve is rather complex, Berrendero and Zamar (2001) provide general methodology for its computation in the context of linear regression.

The maximum-bias curve is not only useful on its own, but allows us to define further scalar measures of robustness. The most prominent is the *breakdown point* (Hampel, 1971), which is defined as the smallest amount ε of contamination that can cause an infinite bias:

$$\varepsilon^*(T) = \inf_{\varepsilon \geq 0} \{\varepsilon : B(\varepsilon, T) = \infty\}. \tag{6}$$

The intuitive aim of this definition specifies the breakdown point $\varepsilon^*(T)$ as the smallest amount of contamination that makes the estimator T useless. Note that in most cases $\varepsilon^*(T) \leq 0.5$ (He and Simpson, 1993). This definition and the upper bound, however, apply only in simple cases, such as location or linear regression estimation (Davies and Gather, 2005). The most general definition of breakdown point formalizes the idea of 'useless' estimates in the following way: an estimator is said to break down if, under contamination, it is not random anymore, or, more precisely, it can achieve only a finite set of values (Genton and Lucas, 2003). This definition is based on the fact that estimates are functions of observed random samples and are thus random quantities themselves unless they fail. Although the latter definition includes the first one, the latter one may generally depend on the underlying model F, for example in time-series context.

2 Estimation approaches

Denote by F_n an empirical distribution function (edf) corresponding to a sample $\{x_i\}_{i=1}^n \in \mathbb{R}$ drawn from a model based on probability distribution F. Most estimation methods can be defined as an extremum problem, minimizing a contrast $\int h(z, \theta)dF(z)$ over θ in a parameter space, or as a solution of an equation, $\int g(z, \theta)dF(z) = 0$ in θ. The estimation for a given sample utilizes finite-sample equivalents of these integrals, $\int h(z, \theta)dF_n(z)$ and $\int g(z, \theta)dF_n(z)$, respectively.

Consider the pure location model $X_i = \mu + \sigma\varepsilon_i, i = 1, \ldots, n$, with a known scale σ and $\varepsilon \sim F$. The cdf of X is then $F\{(x - \mu)/\sigma\}$. With a quadratic contrast function $h(x, \theta) = (x - \theta)^2$, the estimation problem is to minimize $\int (x - \theta)^2 dF\{(x - \mu)/\sigma\}$ with respect to θ. For known F, this leads to $\theta = \mu$ and one sees that, without loss of generality, one can assume $\mu = 0$ and

$\sigma = 1$. For the sample $\{x_i\}_{i=1}^n$ characterized by edf F_n, the location parameter μ is estimated by

$$\hat{\mu} = \arg\min_\theta \int (x - \theta)^2 dF_n(x) = n^{-1} \sum_{i=1}^n x_i = \bar{x}.$$

Note that for $g(x, \theta) = x - \theta$, the parameter μ is the solution to $\int g(x, \theta) dF(x) = 0$. The estimator may therefore be alternatively defined through $\mu = T(F) = \int u dF(u)$.

As indicated in the introduction, this standard estimator of location performs unfortunately rather poorly under the sketched contamination model. Estimating a population mean by the least squares (LS) or sample mean $\bar{x} = T(F_n)$ has the following properties. First, the influence function (1)

$$\begin{aligned} IF(x; T, F) &= \lim_{\varepsilon \to 0} \frac{T\{(1 - \varepsilon)F + \varepsilon\delta_x\} - T(F)}{\varepsilon} \\ &= \lim_{\varepsilon \to 0} \frac{\{(1 - \varepsilon) \int u dF(u) + \varepsilon x\} - \int u dF(u)}{\varepsilon} \\ &= \lim_{\varepsilon \to 0} \varepsilon^{-1} \left\{ -\varepsilon \int u dF(u) + \varepsilon x \right\} \\ &= x - \int u dF(u) = x - T(F). \end{aligned}$$

Hence, the gross-error sensitivity (2) $\gamma(T, F) = \infty$, the local-shift sensitivity (3) $\lambda(T, F) = 0$, and the rejection point (4) $\rho(T, F) = \infty$. Second, the maximum-bias (5) is infinite for any $\varepsilon > 0$ since

$$\sup_{x \in \mathbb{R}} \|T\{(1 - \varepsilon)F + \varepsilon\delta_x\} - T(F)\| = \sup_{x \in \mathbb{R}} \| - \varepsilon T(F) + \varepsilon x\| = \infty.$$

Consequently, the breakdown point (6) of the sample mean $\bar{x} = T(F_n)$ is zero, $\varepsilon^*(T) = 0$.

Thus, none of robustness measures characterizing the change of T under contamination of data (even infinitesimally small) is finite. This behaviour, typical for LS-based methods, motivated alternative estimators that have the desirable robust properties. In this section, the M-estimators, S-estimators, and τ-estimators are discussed as well as some extensions and combination of these approaches. Even though there is a much wider range of robust estimation principles, we focus on those already studied and adopted in various areas of econometrics.

2.1 M-estimators

To achieve more flexibility in accommodating requirements on robustness, Huber (1964) proposed the *M-estimator* by considering a general extremum estimator based on $\int \rho(z, \theta) dF(z)$, thus minimizing $\int \rho(z, \theta) dF_n(z)$ in finite samples. Providing that the first derivative $\psi(z, \theta) = \partial\rho(z, \theta)/\partial\theta$ exists, an M-estimator can be also defined by an implicit equation $\int \psi(z, \theta) dF_n(z) = 0$.

This extremely general definition is usually adapted to a specific estimation problem such as location, scale, or regression estimation. In a univariate location model, $F(z)$ can be parameterized as $F(z - \theta)$ and hence one limits $\rho(z, \theta)$ and $\psi(z, \theta)$ to $\rho(z - \theta)$ and $\psi(z - \theta)$. In the case of scale estimation, $F(z) = F(z/\theta)$ and consequently $\rho(z, \theta) = \rho(z/\theta)$ and $\psi(z, \theta) = \psi(z/\theta)$. In linear regression, $z = (x, y)$ and a zero-mean error term $\varepsilon = y - x^\top\theta$. Analogously to the location case, one can then consider $\rho(z, \theta) = \rho(y - x^\top\theta)$ and $\psi(z, \theta) = \psi(y - x^\top\theta)x$, or more generally, $\rho(z, \theta) = \rho(y - x^\top\theta, x)$ and

$\psi(z,\theta) = \psi(y - x^\top\theta, x)$ (*GM*-estimators). Generally, we can express $\rho(z,\theta)$ as $\rho\{\eta(z,\theta)\}$, $\psi\{\eta(z,\theta)\}$, where $\eta(z,\theta) \sim F$.

Some well-known choices of univariate objective functions ρ and ψ are given in Table 1; functions $\rho(t)$ are usually assumed to be non-constant, non-negative, even, and continuously increasing in $|t|$. This documents flexibility of the concept of *M*-estimators, which include LS and quantile regression as special cases.

On the other hand, many of the ρ and ψ functions in Table 1 depend on one or more constants $a, b, c \in \mathbb{R}$. If an estimator T is to be invariant to the scale of data, one can apply the estimator to rescaled data, that is, to minimize $\int \rho\{(z - \theta)/s\}dF_n(z)$ or to solve $\int \psi\{(z - \theta)/s\}dF_n(z) = 0$ for a scale estimate s like the median absolute deviation (MAD). Alternatively, one may also estimate parameters θ and scale s simultaneously by considering $\rho(z, \{\theta, s\}) = \rho\{(z - \theta)/s\}$ or

$$\psi(z, \{\theta, s\}) = \{\psi_1(z, \{\theta, s\}), \psi_s(z, \{\theta, s\})\}.$$

Let us now turn to the question how the choice of functions ρ and ψ determines the robust properties of *M*-estimators. First, the influence function of an *M*-estimator can generally depend on several quantities such as its asymptotic variance or the position of explanatory variables in the regression case, but the influence function is always proportional to function $\psi(z,b)$. Thus, the finite gross-error sensitivity, $\gamma(T, F) < \infty$, requires bounded $\psi(t)$ (which is not the case with LS). Similarly, the finite rejection point, $\rho(T, F) < \infty$, leads to $\psi(t)$ being zero for all sufficiently large t (the *M*-estimators defined by such a ψ-function are called redescending). Hampel et al. (1986) shows how, for a given bound on $\gamma(T, F)$, one can determine the most efficient choice of ψ function (for example, the skipped median, $\psi(t) = sign(t)I(|t| < K), K > 0$, in the location case).

More formally, the optimality of *M*-estimators in the context of qualitative robustness can be studied by the *asymptotic relative efficiency* (ARE) of an estimator $\hat{\theta}^1$ relative to another estimator $\hat{\theta}^2$:

$$ARE(\hat{\theta}^1, \hat{\theta}^2) = \frac{\text{as. } var(\hat{\theta}^1)}{\text{as. } var(\hat{\theta}^2)}. \tag{7}$$

For example, at the normal distribution with $\hat{\theta}^1$ and $\hat{\theta}^2$ being the least absolute deviation (LAD) and LS estimators, ARE equals $2/\pi \approx 0.64$. Under the Student cdf t_5, the ARE of the two estimators climbs up to ≈ 0.96. For Huber's M-estimator, we see that its limit cases are the median for $c \to 0$ and the mean for $c \to \infty$. At the normal distribution and for $c = 1.345$, we have ARE of about 0.95. This means that this *M*-estimator is almost as efficient as

Table 1 Examples of ρ and ψ functions used with *M*-estimators

	$\rho(t)$	$\psi(t)$
Least squares	t_2	$2t$
Least absolute deviation	$|t|$	Sign (t)
Quantile estimation	$\{\tau - I(x < 0)\}x$	$\tau - I(x < 0)$
Huber: for $|t| \leq c$	t^2	$2t$
for $c < |t|$	$c|t|$	$c \, sign(t)$
Hample: for $|t| \leq c$	t^2	$2t$
for $a < |t| \leq b$	$a|t|$	$a \, sign(t)$
for $b < |t| \leq c$	$\frac{ac}{c-b}t - \frac{a}{c-b}t^2 sign(t)$	$a(c - |t|)/(c - b)$
for $c < |t|$	$a|t|$	0
Biweight (Tukey)	$-(c^2 - t^2)^3 I(|t| \leq c)/6$	$t(c^2 - t^2)^2 I(|t| \leq c)/6$
Sine (Andrews)	$-c \, \cos(x/c)I(|t| \leq \pi c)$	$\sin(x/c)I(|t| \leq \pi c)$

MLE, but does not lose so drastically in performance as the standard mean under contamination because of the bounded influence function.

Whereas the influence function of M-estimators is closely related to the choice of its objective function, the global robustness of M-estimators is in a certain sense independent of this choice. Maronna, Bustos and Yohai (1979) showed in linear regression that the breakdown point of M-estimators is bounded by $1/p$, where p is the number of estimated parameters. As a remedy, several authors proposed one-step M-estimators that are defined, for example, as the first step of the iterative Newton–Raphson procedure, used to minimize $\int \rho(z, \theta) dF(z)$, starting from initial robust estimators $\hat{\theta}^0$ of parameters and $\hat{s}^0$ of scale (see Welsh and Ronchetti, 2002, for an overview). Possible initial estimators can be those discussed in subsections 2.2 and 3.3. For example for an M-estimator of location $\hat{\theta}$ defined by a function $\psi(x, \theta) = \psi(x - \theta)$, its one-step counterpart can be defined at sample $\{x_i\}_{i=1}^n$ by

$$\hat{\theta} = \hat{\theta}^0 + \hat{s}^0 \sum_{i=1}^n \psi\left(\frac{x_i - \hat{\theta}^0}{\hat{s}^0}\right) \bigg/ \sum_{i=1}^n \psi'\left(\frac{x_i - \hat{\theta}^0}{\hat{s}^0}\right),$$

where $\hat{\theta}^0$ and $\hat{s}^0$ represent initial robust estimators of location and scale like the median and MAD, respectively. Such one-step estimators, under certain conditions on the initial estimators, preserve the breakdown point of the initial estimators, and at the same time have the same first-order asymptotic distribution as the original M-estimator (Simpson, Ruppert and Carroll, 1992; Welsh and Ronchetti, 2002). Further development of such ideas includes an adaptive choice of parameters of function ψ in the iterative step (Gervini and Yohai, 2002).

2.2 S-estimators

An alternative approach to *M*-estimators achieving high breakdown point (HBP) was proposed by Rousseeuw and Yohai (1984). The S-estimators are defined by minimization of a scale statistics $s^2(z, b) = s\{\eta(z, b)\}$ defined as the *M*-estimate of scale,

$$\int \rho[\eta(z, b)/s\{\eta(z, b)\}]dF_n(z) = K = \int \rho(t)dF(t),$$

at the model distribution F; the functions ρ and η are those defining *M*-estimators in subsection 2.1. More generally, one can define *S*-estimators by means of any scale-equivariant statistics s^2, that is, $s\{c\eta(z, b)\} = |c|s\{\eta(z, b)\}$. Under this more general definition, *S*-estimators include as special cases LS and LAD estimators. Further, they encompass several well-known robust methods including least median of squares (LMS) and least trimmed squares (LTS): whereas the first defines the scale statistics $s^2\{\eta(z, b)\}$ as the median of squared residuals $\eta(z, b)$, the latter uses the scale defined by the sum of h smallest residuals $\eta(z, b)$. In order to appreciate the difference to *M*-estimators, it is worth pausing for a moment and to present LMS, the most prominent representative of *S*-estimators, in the location case:

$$\arg \min_{\theta} \ \text{med} \ \{(x_1 - \theta)^2, \ldots, (x_n - \theta)^2\}.$$

Due to its definition, the *S*-estimators have the same influence function as the *M*-estimator constructed from the same function ρ. Contrary to *M*-estimators, they can achieve the highest possible breakdown point $\varepsilon^* = 0.5$. For example, this is the case of LMS and LTS. For Gaussian data, the most

efficient (in the sense of ARE (7)) among the S-estimators with $\varepsilon^* = 0.5$ is, however, the one corresponding to $K = 1.548$ and ρ being the Tukey biweight function (see Table 1). Given the HBP of S-estimators, their maximum-bias behaviour is of interest too. Although it depends on the function ρ and constant K (Berrendero and Zamar, 2001), Yohai and Zamar (1993) proved that LMS minimizes maximum bias among a large class of (residual admissible) estimators, which includes most robust methods.

An important shortcoming of HBP S-estimation is, however, its low ARE: under Gaussian data, efficiency relative to LS varies from zero per cent to 27 per cent. Thus, S-estimators are often used as initial estimators for other, more efficient methods. Nevertheless, if an S-estimator is not applied directly to sample observations, but rather to the set of all pairwise differences of sample observations, the resulting generalized S-estimator exhibits higher relative efficiency for Gaussian data, while preserving its robust properties (Croux, Rousseeuw and Hossjer, 1994; Stromberg, Hossjer and Hawkins, 2000).

2.3 τ-estimators

The S-estimators improve upon M-estimators in terms of their breakdown-point properties, but at the cost of low Gaussian efficiency. Although one-step M-estimators based on an initial S-estimate can remedy this deficiency to a large extent, their exact breakdown properties are not known. One alternative approach, proposed by Yohai and Zamar (1988), extends the principle of S-estimation in the following way. Assuming that ρ_1 and ρ_2 are non-negative, even, and continuous functions, the M-estimate $s^2(z, \theta) = s^2\{\eta(z, \theta)\}$ of scale can be defined as in the case of S-estimation,

$$\int \rho_1[\eta(z, \theta)/s\{\eta(z, \theta)\}]dF_n(z) = K = \int \rho_1(t)dF(t).$$

Next, the τ-estimate of scale is defined by

$$\tau^2(z, \theta) = s^2\{\eta(z, \theta)\} \int \rho_2[\eta(z, \theta)/s\{\eta(z, \theta)\}]dF_n(z)$$

and the corresponding τ-*estimator* of parameters θ is then defined by minimizing the τ-estimate of scale, $\tau^2(z, \theta)$.

As a generalization of S-estimation, the τ-estimators include S-estimators as a special case for $\rho_1 = \rho_2$ because then $\tau^2(x, \theta) = \theta s^2(z, \theta)$. On the other hand, if $\rho_2(t) = t^2$, $\tau^2(z, \theta) = \int \eta^2(z, \theta)dF_n(z)$ is just the standard deviation of model residuals. Compared with S-estimators, the class of τ estimators can improve in terms of relative Gaussian efficiency because its breakdown point depends only on function ρ_1, whereas its asymptotic variance is a function of both ρ_1 and ρ_2. Thus, ρ_1 can be defined to achieve the breakdown point equal to 0.5 and ρ_2 consequently adjusted to reach a pre-specified relative efficiency for Gaussian data (for example, 95 per cent).

3 Methods of robust econometrics

The concepts and methods of robust estimation discussed in Sections 1 and 2 are typically proposed in the context of a simple location or linear regression models, on the assumption of independent, continuous, and identically distributed random variables. This, however, rarely corresponds to assumptions typical for most econometric models. In this section, we therefore present an

overview of developments and extensions of robust methods to various econometric models. As the M-estimators are closest to the commonly used LS and MLE, most of the extensions employ M-estimation. The HBP techniques are not that frequently found in the economics literature (Zaman, Rousseeuw and Orhan, 2001; Sapra, 2003) and are mostly applied only as a diagnostic tool.

In the rest of this section, robust estimation is first discussed in the context of models with discrete explanatory variables, models with time-dependent observations, and models involving multiple equations. Later, robust alternatives to general estimation principles, such as MLE and generalized method of moments (GMM), are discussed. Before doing so, let us mention that dangers of data contamination are not studied only from the theoretical point of view. There is a number of studies that check the presence of outliers in real data and their influence on estimation methods. For example, there is evidence of data contamination and its adverse effects on LS and MLE in the case of macroeconomic time series (Balke and Fomby, 1994; Atkinson, Koopman and Shephard, 1997), in financial time series (Sakata and White, 1998; Franses, van Dijk and Lucas, 2004), marketing data (Franses, Kloek and Lucas, 1999), and many other areas. These adverse effects include biased estimates, masking of structural changes, and creating seemingly nonlinear structures, for instance.

3.1 Discrete variables

To achieve a HBP, many robust methods such as LMS often eliminate a large portion of observations from the calculation of their objective function. This can cause non-identification of parameters associated with categorical variables. For example, having data on income $\{y_i\}_{i=1}^n$ of men and women, where gender is indicated by $\{d_i\}_{i=1}^n \in \{0, 1\}$, one can estimate the mean income of men and women by a simple regression model $y_i = a + bd_i$. If a HBP method such as LMS or LTS is used to estimate the model and it eliminates a large portion of observations from the calculation (for example, one half of them), the remaining data could easily contain only income of men or only income of women, and consequently the mean income of one of the groups could not be then identified. Even though this seems unlikely in our simple example, it becomes more pronounced as the number of discrete variables grows (see Hubert and Rousseeuw, 1997, for an example).

A common strategy employs a robust estimator with a HBP for a model with only continuous variables, and using this initial estimate, the model with all variables is estimated by an M-estimator. Such a combined procedure preserves the breakdown point of the HBP estimator: even though a misclassified values of categorical explanatory variables can bias the estimates, this bias will be bounded in common models as the categorical variables are bounded as well. See Hubert and Rousseeuw (1997) and Maronna and Yohai (2000), who combine an initial S-estimator with an M-estimator.

3.2 Time series

In time series, there are several issues not addressed by the standard theory of robust estimation because of time-dependency of observations. First, the asymptotic behaviour of various robust methods has to be established; see Koenker and Machado (1999), Koenker and Xiao (2002) for L_1 regression, Künsch (1984) and Bai and Wu (1997) for M-estimators and Sakata and

White (2001), Zinde-Walsh (2002), and Čížek (2006) for various S-type estimators. In these cases, the results are usually established for general non-linear models.

Second, the effects of data contamination are more complex and wide-spread due to time dependency: an error in one observation is transferred, by means of a model, to others close in time. The possible effects of outliers in time series are elaborated by Chen and Liu (1993) and Tsay, Pena and Pankratz (2000), for instance. The first work also offers a sequential identification of outliers (an alternative procedure based on τ-estimators is offered by Bianco et al., 2001). Consequently, the robust properties in time series differ from those experienced in cross-sectional data. For example, the breakdown point is asymptotically zero in the case of M-estimators (Sakata and White, 1995) and can be much below 0.5 for various S-estimators (Genton and Lucas, 2003).

A further issue specific to time series is testing for stationarity of a series. Effects of outliers are in this respect similar to those of neglected structural changes. To differentiate between random outliers and real structural changes, robust tests for change-point detection have been proposed by Gagliardini, Trojani and Urga (2005) and Fiteni (2002; 2004); the last of these papers uses τ-estimation. The asymptotics of M-estimators under unit-root assumption and the corresponding tests have been established, for example, by Lucas (1995), Koenker and Xiao (2004), and Haldrup, Montans and Sanso (2005). An early reference is Franke, Härdle and Martin (1984).

3.3 Multivariate regression

An important application of robust methods in economics concerns the multivariate regression case. This is relatively straightforward with exogenous explanatory variables only, see Koenker and Portnoy (1990), Bilodeau and Duchesne (2000), and Lopuhaä (1992) for the M-, S-, and τ-estimation, respectively. Estimating general simultaneous equations models has to mimic either three-stage LS or full information MLE (Marrona and Yohai, 1997). Whereas Koenker and Portnoy (1990) follow with the weighted LAD the first approach, Krishnakumar and Ronchetti (1997) use M-estimation together with the second strategy.

3.4 General estimation principles

There are naturally many more model classes for which one can construct robust estimation procedures. Since most econometric models can be estimated by means of MLE or GMM, it is however easier to concentrate on robust counterparts of these two estimation principles. There are other estimation concepts, such as nonparametric smoothing, that can employ robust estimation (Härdle, 1982), but they go beyond the scope of this article.

First, recent contributions to robust MLE can be split to two groups. One simply defines a weighted maximum likelihood, where weights are computed from an initial robust fit (Windham, 1995; Markartou, Basu and Lindsay, 1997). Alternatively, some erroneous observations can be excluded completely from the likelihood function (Clarke, 2000; Marazzi and Yohai, 2004). This approach requires existence of an initial robust estimate, and thus it is not useful for models for which there are no robust methods available. The second approach is motivated by the S-estimation, namely, LTS, and defines the maximum trimmed likelihood as an estimator maximizing the

product of the h largest likelihood contribution, that is, those corresponding only to h most likely observations (Hadi and Luceño, 1997). This estimator has been studied mainly in the context of generalized linear models (Müller and Neykov, 2003), but its consistency is established in a much wider class of models (Čížek, 2004).

Second, more widely used GMM has also attracted attention from its robustness point of view. A special case – instrumental variable estimation – has been studied, for example, by Wagenvoor and Waldman (2002) and Kim and Muller (2006). See also Chernozhukov and Hansen (2006) for instrumental variable quantile regression. More generally, Ronchetti and Trojani (2001) have proposed an M-estimation-based generalization of GMM, studied its robust properties, and designed corresponding tests. This work became a starting point for others, who have extended the methodology of Ronchetti and Trojani (2001) to robustify simulation-based methods of moments (Genton and Ronchetti, 2003; Ortelli and Trojani, 2005).

P. Čížek and W. Härdle

See also

< xref = A000235 > adaptive estimation;
< xref = C000561 > categorical data;
< xref = C000559 > computational methods;
< xref = xyyyyyy > generalized method of moments;
< xref = xyyyyyy > least squares;
< xref = M000300 > maximum likelihood;
< xref = xyyyyyy > measurement error models;
< xref = xyyyyyy > time series analysis.

Bibliography

Atkinson, A., Koopman, S. and Shephard, N. 1997. Detecting shocks: outliers and breaks in time series. *Journal of Econometrics* 80, 387–422.

Bai, Z.D. and Wu, Y. 1997. General M-Estimation. *Journal of Multivariate Analysis* 63, 119–35.

Balke, N. and Fomby, T. 1994. Large shocks, small shocks, and economic fluctuations: outliers in macroeconomic time series. *Journal of Applied Econometrics* 9, 181–200.

Berrendero, J. and Zamar, R. 2001. Maximum bias curves for robust regression with non-elliptical regressors. *Annals of Statistics* 29, 224–51.

Bianco, A., Ben, M., Martínez, E. and Yohai, V. 2001. Regression models with ARIMA errors. *Journal of Forecasting* 20, 565–79.

Bilodeau, M. and Duchesne, P. 2000. Robust estimation of the SUR model. *Canadian Journal of Statistics* 28, 277–88.

Chen, C. and Liu, L.-M. 1993. Joint estimation of model parameters and outlier effects in time series. *Journal of the American Statistical Association* 88, 284–97.

Chernozhukov, V. and Hansen, C. 2006. Instrumental quantile regression inference for structural and treatment effect models. *Journal of Econometrics* 132, 491–525.

Čížek, P. 2004. General trimmed estimation: robust approach to nonlinear and limited dependent variable models. Discussion Paper No. 2004/130, CentER ,Tilburg University.

Čížek, P. 2006. Least trimmed squares in nonlinear regression under dependence. *Journal of Statistical Planning and Inference* 136, 3967–88.

Clarke, B. 2000. An adaptive method of estimation and outlier detection in regression applicable for small to moderate sample sizes. *Probability and Statistics* 20, 25–50.

Croux, C., Rousseeuw, P. and Hossjer, O. 1993. Generalized S-estimators. *Journal of the American Statistical Association* 89, 1271–81.

Davies, L. and Gather, U. 2005. Breakdown and groups. *Annals of Statistics* 33, 988–93.

Fiteni, I. 2002. Robust estimation of structural break points. *Econometric Theory* 18, 349–386.

Fiteni, I. 2004. τ-estimators of regression models with structural change of unknown location. *Journal of Econometrics* 119, 19–44.

Franses, P., Kloek, T. and Lucas, A. 1999. Outlier robust analysis of longrun marketing effects for weekly scanning data. *Journal of Econometrics* 89, 293–315.

Franses, P., van Dijk, D. and Lucas, A. 2004. Short patches of outliers, ARCH and volatility modelling. *Applied Financial Economics* 14, 221–31.

Franke, J., Härdle, W. and Martin, R. 1984. *Robust and Nonlinear Time Series Analysis*. Berlin: Springer.

Gagliardini, P., Trojani, F. and Urga, G. 2005. Robust GMM tests for structural breaks. *Journal of Econometrics* 129, 139–82.

Genton, M. and Lucas, A. 2003. Comprehensive definitions of breakdown-points for independent and dependent observations. *Journal of the Royal Statistical Society, Series B* 65, 81–94.

Genton, M. and Ronchetti, E. 2003. Robust indirect inference. *Journal of the American Statistical Association* 98, 67–76.

Gervini, D. and Yohai, V. 2002. A class of robust and fully efficient regression estimators. *Annals of Statistics* 30, 583–616.

Hadi, A. and Luceño, A. 1997. Maximum trimmed likelihood estimators: a unified approach, examples, and algorithms. *Computational Statistics & Data Analysis* 25, 251–72.

Haldrup, N., Montans, A. and Sanso, A. 2005. Measurement errors and outliers in seasonal unit root testing. *Journal of Econometrics* 127, 103–28.

Hampel, F. 1971. A general qualitative definition of robustness. *Annals of Mathematical Statistics* 42, 1887–96.

Hampel, F., Ronchetti, E., Rousseeuw, P. and Stahel, W. 1986. *Robust Statistics: The Approach Based on Influence Function*. New York: Wiley.

Härdle, W. 1982. Robust regression function estimation. *Journal of Multivariate Analysis* 14, 169–80.

He, X. and Simpson, D. 1993. Lower bounds for contamination bias: globally minimax versus locally linear estimation. *Annals of Statistics* 21, 314–37.

Huber, P. 1964. Robust estimation of a location parameter. *Annals of Mathematical Statistics* 35, 73–101.

Hubert, M. and Rousseeuw, P. 1997. Robust regression with both continuous and binary regressors. *Journal of Statistical Planning and Inference* 57, 153–63.

Kim, T.-H. and Muller, C. 2006. Two-stage Huber estimation. *Journal of Statistical Planning and Inference* (forthcoming).

Koenker, R. and Machado, J. 1999. Goodness of fit and related inference processes for quantile regression. *Journal of the American Statistical Association* 94, 1296–310.

Koenker, R. and Portnoy, S. 1990. M-estimation of multivariate regressions. *Journal of the American Statistical Association* 85, 1060–8.

Koenker, R. and Xiao, Z. 2002. Inference on the quantile regression process. *Econometrica* 70, 1583–612.

Koenker, R. and Xiao, Z. 2004. Unit root quantile autoregression inference. *Journal of the American Statistical Association* 99, 775–87.

Krishnakumar, J. and Ronchetti, E. 1997. Robust estimators for simultaneous equations models. *Journal of Econometrics* 78, 295–314.

Kunsch, H. 1984. Infinitesimal robustness for autoregressive processes. *Annals of Statistics* 12, 843–63.

Lopuhaä, H. 1992. Multivariate τ-estimators. *Canadian Journal of Statistics* 19, 307–21.

Lucas, A. 1995. An outlier robust unit root test with an application to the extended Nelson–Plosser data. *Journal of Econometrics* 66, 153–73.

Marazzi, A. and Yohai, V. 2004. Adaptively truncated maximum likelihood regression with asymmetric errors. *Journal of Statistical Planning and Inference* 122, 271–91.

Markatou, M., Basu, A. and Lindsay, B. 1997. Weighted likelihood estimating equations: the discrete case with applications to logistic regression. *Journal of Statistical Planning and Inference* 57, 215–32.

Maronna, R., Bustos, O. and Yohai, V. 1979. Bias- and efficiency-robustness of general *M*-estimators for regression with random carriers. In *Smoothing Techniques for Curve Estimation*, ed. T. Gasser and M. Rossenblatt. Berlin: Springer.

Maronna, R. and Yohai, V. 1997. Robust estimation in simultaneous equations models. *Journal of Statistical Planning and Inference* 57, 233–44.

Maronna, R. and Yohai, V. 2000. Robust regression with both continuous and categorical predictors. *Journal of Statistical Planning and Inference* 89, 197–214.

Müller, C. and Neykov, N. 2003. Breakdown points of trimmed likelihood estimators and related estimators in generalized linear models. *Journal of Statistical Planning and Inference* 116, 503–19.

Ortelli, C. and Trojani, F. 2005. Robust efficient method of moments. *Journal of Econometrics* 128, 69–97.

Ronchetti, E. and Trojani, F. 2001. Robust inference with GMM estimators. *Journal of Econometrics* 101, 37–69.

Rousseeuw, P. and Yohai, V. 1984. Robust regression by means of *S*-estimators. In *Robust and Nonlinear Time Series Analysis*, ed. J. Franke, W. Härdle and R. Martin. Heidelberg: Springer.

Sakata, S. and White, H. 1995. An alternative definition of finite-sample breakdown point with application to regression model estimators. *Journal of the American Statistical Association* 90, 1099–106.

Sakata, S. and White, H. 1998. High breakdown point conditional dispersion estimation with application to S&P 500 daily returns volatility. *Econometrica* 66, 529–67.

Sakata, S. and White, H. 2001. S-estimation of nonlinear regression models with dependent and heterogeneous observations. *Journal of Econometrics* 103, 5–72.

Sapra, S. 2003. High-breakdown point estimation of some regression models. *Applied Economics Letters* 10, 875–78.

Simpson, D., Ruppert, D. and Carroll, R. 1992. On one-step GM estimates and stability of inferences in linear regression. *Journal of the American Statistical Association* 87, 439–50.

Stromberg, A., Hossjer, O. and Hawkins, D. 2000. The least trimmed differences regression estimator and alternatives. *Journal of the American Statistical Association* 95, 853–64.

Tsay, R., Pena, D. and Pankratz, A. 2000. Outliers in multivariate time series. *Biometrika* 87, 789–804.

Wagenvoor, R. and Waldman, R. 2002. On B-robust instrumental variable estimation of the linear model with panel data. *Journal of Econometrics* 106, 297–324.

Welsh, A. and Ronchetti, E. 2002. A journey in single steps: robust one-step M-estimation in linear regression. *Journal of Statistical Planning and Inference* 103, 287–310.

Windham, M. 1995. Robustifying model fitting. *Journal of the Royal Statistical Society, Series B* 57, 599–609.

Yohai, V. and Zamar, R. 1988. High breakdown-point estimates of regression by means of the minimization of an efficient scale. *Journal of the American Statistical Association* 83, 406–13.

Yohai, V. and Zamar, R. 1993. A minimax-bias property of the least a-quantile estimates. *Annals of Statistics* 21, 1824–42.

Zaman, A., Rousseeuw, P. and Orhan, M. 2001. Econometric applications of high-breakdown robust regression techniques. *Economics Letters* 71, 1–8.

Zinde-Walsh, V. 2002. Asymptotic theory for some high breakdown point estimators. *Econometric Theory* 18, 1172–96.

Index terms

generalized method of moments
heteroscedasticity
influence function
instrumental variables
least squares
linear regression
maximum likelihood
maximum-bias curve
measurement errors
M-estimation
Newton–Raphson procedure
nonlinear models
probability distribution
qualitative and quantitative robustness
quantile regression
random variables
robust econometrics
robust statistics
S-estimation
simultaneous equations models
time-series analysis

Index terms not found:

time-series analysis

Time Dependent Relative Risk Aversion

Enzo Giacomini[1], Michael Handel[2] and Wolfgang K. Härdle[1]

[1] CASE - Center for Applied Statistics and Economics
Humboldt-Universität zu Berlin
Spandauer Straße 1, 10178 Berlin, Germany

[2] Dr. Nagler & Company GmbH
Maximilanstr. 47, 80538 Munich, Germany

Abstract

Risk management and the thorough understanding of the relations between financial markets and the standard theory of macroeconomics have always been among the topics most addressed by researchers, both financial mathematicians and economists. This work aims at explaining investors' behavior from a macroeconomic aspect (modeled by the investors' pricing kernel and their relative risk aversion) using stocks and options data. Daily estimates of investors' pricing kernel and relative risk aversion are obtained and used to construct and analyze a three-year long time-series. The first four moments of these time-series as well as their values at the money are the starting point of a principal component analysis. The relation between changes in a major index level and implied volatility at the money and between the principal components of the changes in relative risk aversion is found to be linear. The relation of the same explanatory variables to the principal components of the changes in pricing kernels is found to be log-linear, although this relation is not significant for all of the examined maturities.

1 Introduction

Risk management has developed in the recent decades to be one of the most fundamental issues in quantitative finance. Various models are being developed and applied by researchers as well as financial institutions. By modeling price fluctuations of assets in a portfolio, the loss can be estimated using statistical methods. Different measures of risk, such as standard deviation of returns or confidence interval Value at Risk, have been suggested. These measures are based on the probability distributions of assets' returns extracted from the data-generating process of the asset.

However, an actual one dollar loss is not always valued in practice as a one dollar loss. Purely statistical estimation of loss has the disadvantage of ignoring the circumstances of the loss. Hence the notion of an investor's *utility* has been introduced. Arrow (1964) and Debreu (1959) were the first to introduce elementary securities to formalize economics of uncertainty. The so-called *Arrow-Debreu securities* are the starting point of all modern financial asset pricing theories. Arrow-Debreu securities entitle their holder to a payoff of 1$ in one specific state of the world, and 0 in all other states of the world. The price of such a security is determined by the market, on which it is tradable, and is subsequent to a supply and demand equilibrium. Moreover, these prices contain information about investors' preferences due to their dependence on the conditional probabilities of the state of the world at maturity and due to the imposition of market-clearing and general equilibrium conditions. The prices reflect investors' beliefs about the future, and the fact that they are priced differently in different states of the world implies, that a one-dollar gain is not always worth the same, in fact its value is exactly the price of the security.

A very simple security that demonstrates the concept of Arrow-Debreu securities is a European option. The payoff function of a call option at maturity T is

$$\psi(S_T) = (S_T - K)^+ \stackrel{\text{def}}{=} \max(S_T - K, 0) \tag{1}$$

where K is the strike price, T is maturity and S_T is the asset's price at maturity.

Since an option is a state-dependent contingent claim, it can be valued using the concept of Arrow-Debreu securities. Bearing in mind, that Arrow-Debreu prices can be perceived as a distribution (when the interest rate is 0, they are non negative and sum up to one), the option price is the discounted expectation of random payoffs received at maturity. Since the payoff equals

the value of the claim at maturity time (to eliminate arbitrage opportunities), the value process is by definition a martingale. Introducing a new probability measure Q, such that the discounted value process is a martingale, we can write

$$C_t = e^{-r(T-t)} \operatorname{E}_t^Q[\psi(S_T)] \stackrel{\text{def}}{=} e^{-r(T-t)} \sum_s q_s \psi_s(S_T) \qquad (2)$$

where r is the interest rate and q_s is the price of an Arrow-Debreu security if $r = 0$, paying 1\$ in state s and nothing in any other state. The superscript Q denotes the expectation based on the risk neutral probability measure, the subscript t means that the expectation is conditioned on the information known at time t. The continuous counterpart of the Arrow-Debreu state contingent claims will be defined in the next section as the risk-neutral density or in its more commonly used name, the *State Price Density* (SPD).

Based on the relations between the actual data generating process of a major stock index and its risk-neutral probability measure, we can derive measures that help us learn a lot about investors' beliefs and get an idea of the forces which drive them. This work aims at investigating the dynamics of investors' beliefs.

2 Black and Scholes and Macroeconomic Asset-Pricing Models

The distinction between the actual data generating process of an asset and the market valuations is the essence of macroeconomic dynamic equilibrium asset-pricing models, in which market forces and investors' beliefs are key factors to value an asset with uncertain payoffs.

A standard dynamic exchange economy as discussed by Lucas (1978), Rubinstein (1976) and many others, imposes that securities markets are complete, that they consist of one consumption good and that the investors, which have no exogenous income other than from trading the goods, seek to maximize their state-dependent utility function. There is one risky stock S_t in the economy, corresponding to the market portfolio in a total normalized supply. In addition, the economy is endowed by a riskless bond with a continuously compounded rate of return r. The stock price follows the stochastic process

$$\frac{dS_t}{S_t} = \mu dt + \sigma dW_t \qquad (3)$$

where μ denotes the drift, σ is the volatility and W_t is a standard Brownian motion. The drift and volatility could be functions of the asset price, time and many other factors. However, for simplicity, they are considered constant in this section. The conditional density of the stock price, which is implied by equation (3), is denoted by $p_t(S_T|S_t)$. In this setting, due to continuous dividend payments, the discounted process with cumulative dividend reinvestments should be a martingale and is denoted by

$$\widetilde{S}_t \stackrel{\text{def}}{=} e^{-(r+\delta)t} S_t \tag{4}$$

Since we are dealing with corrected data and in order to simplify the theoretic explanations, we will consider $\delta = 0$ from now on and omit the dividends from the equations.

Taking the total differential yields

$$\begin{aligned}
d\widetilde{S}_t &= d(e^{-rt} S_t) \\
&= -re^{-rt} S_t dt + e^{-rt} dS_t \\
&= -re^{-rt} S_t dt + e^{-rt} [\mu S_t dt + \sigma S_t dW_t] \\
&= (\mu - r)\widetilde{S}_t dt + \sigma \widetilde{S}_t dW_t \\
&= \sigma \widetilde{S}_t d\overline{W}_t
\end{aligned} \tag{5}$$

where $\overline{W}_t \stackrel{\text{def}}{=} W_t + \frac{\mu - r}{\sigma} t$ can be perceived as a Brownian motion on the probability space corresponding to the risk-neutral measure Q. The term $\frac{\mu - r}{\sigma}$ is called the market price of risk, it measures the excess return per unit of risk borne by the investor and hence it vanishes under Q, justifying the name risk-neutral pricing. Risk-neutral pricing can be understood as the pricing done by a risk-neutral investor, an investor who is indifferent to risk and hence not willing to pay the extra premium. The conditional risk-neutral density of the stock price under Q, implied by equation (5) and denoted as $q_t(S_T|S_t)$, is the state-price density which was described as the continuous counterpart of the Arrow-Debreu prices from equation (2). The basic theorem of asset pricing states, that absence of arbitrage implies the existence of a positive linear pricing rule (Cochrane (2001)), and if the market is complete and indeed arbitrage-free, it can be shown that the risk-neutral measure Q is unique.

In order to relate the subjective and risk-neutral densities to macroeconomic factors, we first need to review some of the basic concepts and definitions of macroeconomic theory. Under some specific assumptions, it is well known that a representative agent exists. The original representative agent model includes utility functions which are based on consumption (see, for example,

Mas-Colell et al. (1995)). However, introducing labor income or intermediate consumption do not affect the results significantly and hence, without loss of generality, we review the concept of marginal rate of substitution with the help of a simple consumption based asset pricing model. The fundamental desire for more consumption is described by an intertemporal two-periods utility function as

$$U(c_t, c_{s_{t+1}}) = u(c_t) + \beta \, \mathrm{E}_t[u(c_{s_{t+1}})] \stackrel{\text{def}}{=} u(c_t) + \beta \sum_s u(c_{s_{t+1}}) p_t(s_{t+1}|s_t) \quad (6)$$

where s_t denotes the state of the world at time t, c_t denotes the consumption at time t, $c_{s_{t+1}}$ denotes consumption at the unknown state of the world at time $t+1$, $p_t(s_{t+1}|s_t)$ is the probability of the state of the world at time $t+1$ conditioned on information at time t, $u(c)$ is the one-period utility of consumption and β is a subjective discount factor. We further assume that an agent can buy or sell as much as he wants from an asset with payoff $\psi_{s_{t+1}}$ at price P_t. If Y_t is the agent's wealth (endowment) at t and ξ is the amount of asset he chooses to buy, then the optimization problem is

$$\max_{\{\xi\}}\{u(c_t) + \mathrm{E}_t[\beta u(c_{s_{t+1}})]\}$$

subject to

$$\begin{aligned} c_t &= Y_t - P_t \cdot \xi \\ c_{s_{t+1}} &= Y_{s_{t+1}} + \psi_{s_{t+1}} \cdot \xi \end{aligned}$$

The first constraint is the budget contraint at time t, the agent's endowment at time t is divided between his consumption and the amount of asset he chooses to buy. The budget constraint at time $t+1$ sustains the Walrasian property, i.e. the agent consumes all of his endowment and asset's payoff at the last period. The first order condition of this problem yields

$$P_t = \mathrm{E}_t\left[\beta \frac{u'(c_{s_{t+1}})}{u'(c_t)} \psi_{s_{t+1}}\right] \quad (7)$$

We define $\mathrm{MRS}_t \stackrel{\text{def}}{=} \beta \, \mathrm{E}_t\left[\frac{u'(c_{s_{t+1}})}{u'(c_t)}\right]$ as the *Marginal Rate of Substitution* at t, meaning the rate at which the investor is willing to substitute consumption at $t+1$ for consumption at t. If consumption at $t+1$ depends on the state of the world (which is the case discussed here), the MRS is also referred to as a *stochastic discount factor*.

Famous works like Lucas (1978) or Merton (1973) address the asset pricing models in a more general manner. The utility function depends on the agent's

wealth Y_t at time t and the payoff function depends on the underlying asset S_t. According to Merton (1973), in equilibrium, the optimal solution is to invest in the risky stock at every $t < T$ and then consume the final value of the stock, i.e. $Y_t = S_t$ for $\forall t < T$ and $Y_T = S_T = c_T$. This is a multi-period generalization of the model introduced before (equation (6)), where period T corresponds to $t+1$ in the previous section. Defining time to maturity as $\tau \overset{\text{def}}{=} T - t$, the date t price of an asset with a liquidating payoff of $\psi(S_T)$ is path independent, as the marginal utilities in the periods prior to maturity cancel out. This price is given by

$$P_t = e^{-r\tau} \int_0^\infty \psi(S_T) \lambda \frac{U'(S_T)}{U'(S_t)} p_t(S_T|S_t) dS_T \tag{8}$$

where $\lambda e^{-r\tau} = \beta$ to correspond to equation (7) and λ being a constant independent of index level, for scaling purposes.

Considering the call option price under the unique risk-neutral probability measure in equation (2) and the existence of a positive linear pricing rule in the absence of arbitrage, we argue that the price of any asset can be expressed as a discounted expected payoff (discounted at the risk-free rate) as long as we calculate the expectation with respect to the *risk-neutral density*. Since a risk-neutral agent always has the same marginal utility of wealth, the ratio of marginal utilities in equation (8) vanishes under Q, and equation (8) can be rewritten as

$$P_t = e^{-r\tau} \int_0^\infty \psi(S_T) q_t(S_T|S_t) dS_T = e^{-r\tau} \mathrm{E}_t^Q[\psi(S_T)] \tag{9}$$

where $q_t(S_T|S_t)$ is the State Price Density and the expectation $\mathrm{E}_t^Q[\psi(S_T)]$ is taken with respect to the risk-neutral probability measure Q and not the subjective probability measure, thus reflecting an objective belief about the future states of the world.

Combining equations (8) and (9) we can define the *pricing kernel $M_t(S_T)$*, which relates to the the state price density $q_t(S_T|S_t)$, the subjective probability and the utility function as

$$M_t(S_T) \overset{\text{def}}{=} \frac{q_t(S_T|S_t)}{p_t(S_T|S_t)} = \lambda \frac{U'(S_T)}{U'(S_t)} \tag{10}$$

and therefore $\mathrm{MRS}_t = e^{-r\tau} \mathrm{E}_t[M_t(S_T)]$. Substituting out the $q_t(S_T|S_t)$ in

equation (9) using equation (10) yields the *Lucas asset pricing equation*:

$$\begin{aligned} P_t &= e^{-r\tau} \, \mathrm{E}_t^Q[\psi(S_T)] \\ &= e^{-r\tau} \int_0^\infty M_t(S_T) \cdot \psi(S_T) p_t(S_T|S_t) dS_T \\ &= e^{-r\tau} \, \mathrm{E}_t[M_t(S_T) \cdot \psi(S_T)] \end{aligned} \tag{11}$$

The dependence of the pricing kernel on the investor's utility function has urged researchers to try and estimate distributions based on various utility functions. Arrow (1965) and Pratt (1964) showed a connection between the pricing kernel and the representative agent's measure of risk aversion. The agent's risk aversion is a measure of the curvature of the agent's utility function. The higher the agent's risk aversion is, the more curved his utility function becomes. If the agent were risk-neutral, the utility function would be linear. In order to keep a fixed scale in measuring the risk aversion, the curvature is multiplied by the level of the asset (the argument of the utility function), i.e. the representative agent's *coefficient of Relative Risk Aversion* (RRA) is defined as

$$\rho_t(S_T) \stackrel{\text{def}}{=} -\frac{S_T u''(S_T)}{u'(S_T)} \tag{12}$$

According to equation (10) the pricing kernel is related to the marginal utilities as

$$M_t(S_T) = \lambda \frac{U'(S_T)}{U'(S_t)}$$

$$\Rightarrow \quad M_t'(S_T) = \lambda \frac{U''(S_T)}{U'(S_t)} \tag{13}$$

Substituting out the first and second derivatives of the utility function in equation (12) using equation (13) yields

$$\rho_t(S_T) = -\frac{S_T \lambda M_t'(S_T) U'(S_t)}{\lambda M_t(S_T) U'(S_t)} = -\frac{S_T M_t'(S_T)}{M_t(S_T)} \tag{14}$$

Using equation (10) we can express the RRA as

$$\begin{aligned} \rho_t(S_T) &= -\frac{S_T[q_t(S_T|S_t)/p_t(S_T|S_t)]'}{q_t(S_T|S_t)/p_t(S_T|S_t)} \\ &= -S_T \frac{[q_t'(S_T|S_t)p_t(S_T|S_t) - p_t'(S_T|S_t)q_t(S_T|S_t)]/p_t^2(S_T|S_t)}{q_t(S_T|S_t)/p_t(S_T|S_t)} \\ &= -S_T \frac{q_t'(S_T|S_t)p_t(S_T|S_t) - p_t'(S_T|S_t)q_t(S_T|S_t)}{q_t(S_T|S_t)p_t(S_T|S_t)} \\ &= S_T \left[\frac{p_t'(S_T|S_t)}{p_t(S_T|S_t)} - \frac{q_t'(S_T|S_t)}{q_t(S_T|S_t)} \right] \end{aligned} \tag{15}$$

We now have a method of deriving the investor's pricing kernel and his risk aversion just by knowing, or being able to estimate, the subjective and the risk-neutral densities. As an example, we consider the popular power utility function

$$u(c_t) = \begin{cases} \frac{1}{1-\gamma} c_t^{1-\gamma} & \text{for } 0 < \gamma \neq 1 \\ \log(c_t) & \text{for } \gamma = 1 \end{cases} \tag{16}$$

Rubinstein (1976) showed, that for such a utility function, aggregate consumption is proportional to aggregate wealth, corresponding to the utility of wealth or asset prices discussed above. It can be seen, that as $\gamma \to 0$ the utility is reduced to a linear function. The logarithmic utility function when $\gamma = 1$ is obtained by applying the L'Hospital rule.

The marginal rate of substitution of an investor with a power utility function is

$$\text{MRS}_t = \beta \, \text{E}_t \left[\frac{u'(c_T)}{u'(c_t)} \right] = \beta \, \text{E}_t \left[\left(\frac{c_T}{c_t} \right)^{-\gamma} \right] \tag{17}$$

which means, that it is a function of consumption growth and it is easy to relate it to empirical data. The relative risk aversion of an investor with a power utility can be calculated using equation (12), with consumption instead of wealth as an argument, as the utility function is utility of consumption

$$\rho(c_T) = -c_T \frac{-\gamma (c_T)^{-\gamma-1}}{(c_T)^{-\gamma}} = \gamma \tag{18}$$

This equation shows that the RRA turns out to be a constant, and for the logarithmic utility case, the risk aversion is 1.

Jackwerth (2000) argues that due to the risk aversion of the investor with a power utility function, the pricing kernel is a monotonically decreasing function of aggregate wealth. He estimates q and p using data on the S&P500 index returns, as it is common to assume that this index represents the aggregate wealth held by investors, and computes the pricing kernel according to equation (10). However, he finds out that the pricing kernel is not a monotonically decreasing function as expected. Plotted against the return on the S&P500, the pricing kernel according to Jackwerth (2000) is locally increasing, implying an increasing marginal utility and a convex utility function. It is referred to as the *Pricing Kernel Puzzle*. The shape of the pricing kernel does not correspond to the basic assumption of asset pricing theory. Although Jackwerth (2000) tends to rule out methodological errors, he never proves that the ratio of two estimators equals the estimate of the ratio. He assumes that if q and p are estimated correctly, then their ratio should yield

a good estimator for the pricing kernel. This assumption still needs to be proved, but dealing with it is beyond the scope of this work.

Under the assumptions of the well-known Black & Scholes (1973) model, the price of a plain vanilla call option with a payoff function as in equation (1) is given by the *Black and Scholes formula*

$$C^{BS}(S_t, t, K, T, \sigma, r, \delta) = e^{-\delta\tau} S_t \Phi(d_1) - e^{-r\tau} K \Phi(d_2) \qquad (19)$$

where δ is the continuous dividend rate, r is a constant riskless interest rate, τ is time to maturity, $\Phi(u)$ is the cumulative standard normal distribution function and

$$d_1 = \frac{\ln(S_t/K) + (r - \delta + 0.5\sigma^2)\tau}{\sigma\sqrt{\tau}} \text{ and } d_2 = d_1 - \sigma\sqrt{\tau} \qquad (20)$$

where we assume $\delta = 0$ for the remaining of this work, as mentioned before. Furthermore, the Black & Scholes (1973) implied volatility is assumed to be constant and the corresponding risk-neutral density is log-normal with mean $(r - 0.5\sigma^2)\tau$ and variance $\sigma^2\tau$.

A famous work by Breeden & Litzenberger (1978) proved the following relation, which also holds when the assumptions of the Black & Scholes (1973) model do not:

$$e^{r\tau} \left. \frac{\partial^2 C(S_t, K, \tau)}{\partial K^2} \right|_{K=S_T} = q_t(S_T) = \text{SPD} \qquad (21)$$

Sustaining the assumptions of the Black & Scholes (1973) model and plugging equation (19) into equation (21) yields

$$q^{BS}(S_T | S_t) = \frac{1}{S_T \sqrt{2\pi\sigma^2\tau}} \cdot e^{-\frac{[\ln(S_T/S_t) - (r - 0.5\sigma^2)\tau]^2}{2\sigma^2\tau}} \qquad (22)$$

meaning that the underlying asset price follows the stochastic process

$$\frac{dS_t}{S_t} = r \cdot dt + \sigma \cdot dW_t \qquad (23)$$

i.e., the stock price in a Black & Scholes (1973) world follows a geometric Brownian motion under both probability measures, only with different drifts. Since the subjective probability under the Black & Scholes (1973) is also log-normal but with drift μ, plugging the SPD from equation (22) and the

log-normal subjective density into equation (10) yields a closed-form solution for the investor's pricing kernel

$$M_t^{BS}(S_T) = \left(\frac{S_T}{S_t}\right)^{-\frac{\mu-r}{\sigma^2}} \cdot e^{\frac{(\mu-r)(\mu+r-\sigma^2)\tau}{2\sigma^2}} \tag{24}$$

The only non constant term in this expression is $\frac{S_T}{S_t}$, which corresponds to consumption growth in a pure exchange economy. Since the pricing kernel in equation (24) is also the ratio of the marginal utility functions (equation (10)), the investor's utility function can be derived by solving the differential equation. If we consider the following constants

$$\begin{aligned}\gamma &= \frac{\mu-r}{\sigma^2} \\ \lambda &= e^{\frac{(\mu-r)(\mu+r-\sigma^2)\tau}{2\sigma^2}}\end{aligned} \tag{25}$$

we can rewrite equation (24) as

$$M_t^{BS}(S_T) = \lambda \left(\frac{S_T}{S_t}\right)^{-\gamma} \tag{26}$$

which corresponds to a power utility function. The B&S utility function is therefore

$$u^{BS}(S_t) = \left(1 - \frac{\mu-r}{\sigma^2}\right)^{-1} \cdot S_t^{\left(1-\frac{\mu-r}{\sigma^2}\right)} \tag{27}$$

the subjective discount factor of intertemporal utility is

$$\beta^{BS} = \lambda e^{-r\tau} = e^{\frac{(\mu-r)(\mu+r-\sigma^2)\tau}{2\sigma^2} - r\tau} \tag{28}$$

and the relative risk aversion is constant

$$\rho_t^{BS}(S_T) = \gamma = \frac{\mu-r}{\sigma^2} \tag{29}$$

The above equations prove that a constant RRA utility function sustains the Black & Scholes (1973) model, as was shown by Rubinstein (1976), Breeden & Litzenberger (1978) and many others.

Referring again to the stochastic process in equation (5), in which the Brownian motion $\overline{W}_t$ is defined on the probability space corresponding to the risk-neutral measure, the Brownian motion under the assumptions of the Black & Scholes (1973) model with a constant RRA can be expressed as

$$\overline{W}_t = W_t + \frac{\mu-r}{\sigma}t = W_t + \sigma\gamma t \tag{30}$$

whereas the stochastic process of the corrected stock price can be expressed as a direct function of the investor's relative risk aversion

$$d\widetilde{S}_t = \sigma\widetilde{S}_t d\overline{W}_t = \sigma\widetilde{S}_t dW_t + \sigma^2\widetilde{S}_t\gamma dt \tag{31}$$

10

3 A Static Model: Daily Estimation

It is well known that the assumptions of the Black & Scholes (1973) model do not hold in practice. Transaction costs, taxes, restrictions on short-selling and non-continuous trading violate the model's assumptions. Moreover, the stochastic process does not necessarily follow a Brownian motion and the implied volatility is not constant and experiences a smile. Consequently, the SPD does not have a closed form solution and has to be estimated numerically. Rubinstein (1994) showed, that an estimated subjective probability together with a good estimation of the SPD enable an assessment of the representative agent's preferences. Hence, the model presented in this section aims at estimating the pricing kernel using the ratio between the subjective density and the SPD, and it disregards the issue of whether a ratio of two estimates is a good approximation for the estimated ratio itself.

This section is divided into four parts. The first part provides a short description of the database used in this work. The static model for estimating the pricing kernel and relative risk aversion on a daily basis is introduced in the following parts of this section. When the densities and preferences are known for every day, the dynamics of the time-series can be examined. The results of this examination are reported in the next section.

3.1 The Database

The database used for this work consists of intraday DAX and options data which has undergone a thorough preparation scheme. The data was obtained from the MD*Base, maintained at the Center for Applied Statistics and Economics (CASE) at the Humboldt-University of Berlin. The first trading day in the database is January 4th 1999 and the last one is April 30th 2002, i.e. more than three years of intraday data and 2,921,181 observations. The options data contains tick statistics on the DAX index options and is provided by the German-Swiss Futures Exchange EUREX. Each single contract is documented and contains the future value of the DAX (corresponding to the maturity and corrected for dividends according to equation (4)), the strike, the interest rate (linearly interpolated to approximate a "riskless" interest rate for the specific option's time to maturity), the maturity of the contract, the closing price, the type of the option, calculated future moneyness, calculated Black and Scholes implied volatility, the exact time of the trade (in hundredths of seconds after midnight), the number of contacts and the date.

In order to exclude outliers at the boundaries, only observations with a maturity of more than one day, implied volatility of less than 0.7 and future moneyness between 0.74 and 1.22 are considered, remaining with 2,719,640 observations on 843 trading days. For every single trading day starting April 1999, the static model described in the following section is run and the results are collected. The daily estimation begins three months after the first trading day in the database because part of the estimation process is conducted on historical data, and the history "window" is chosen to be three months, as explained in the next section.

3.2   Subjective Density Estimation

The subjective density is estimated using a simulated GARCH model, the parameters of which are estimated based on historical data. This method was shown by Jackwerth (2000) and others to resemble the actual subjective density.

The first step is to extract the data from the three months preceding the date of the daily assessment. That is the reason for starting the daily process in April instead of January 1999. The intraday options data from the preceding three months are replaced by daily averages of the stock index and the interest rate, averaged over the specific day. When we have a three months history of daily asset prices, we can fit a GARCH (1,1) model to the data. A strong GARCH (1,1) model is described by

$$
\begin{aligned}
\varepsilon_t &= \sigma_t Z_t \\
\sigma_t^2 &= \omega + \alpha \varepsilon_{t-1}^2 + \beta \sigma_{t-1}^2
\end{aligned}
\tag{32}
$$

where Z_t is an independent identically distributed innovation with a standard normal distribution. The logarithmic returns of the daily asset prices are calculated according to $\varepsilon_t = \Delta \log(S_t) = \log(S_t) - \log(S_{t-1})$, and this time series together with its daily standard deviation σ_t are the input of the GARCH estimation. The parameters ω, α and β are estimated using the quasi maximum likelihood method, which is an extension of the maximum likelihood measure, when the estimator is not efficient.

After the parameters of the GARCH process have been estimated, a simulation of a new GARCH process is conducted, starting on the date of the daily assessment. Equations (32) are used for the simulation, but this time the unknown variables are the time series σ_t and ε_t, while the parameters ω, α and β are the ones estimated from the historical data. The simulation

creates a T days long time series, and is run N times. The simulated DAX is calculated as

$$S_t = S_{t-1}e^{\varepsilon_t} \quad \forall t \in \{1, \ldots, T\} \tag{33}$$

where S_0 is the present level of the index on the day of the daily assessment.

Our aim is to estimate the subjective density in some fixed time points, which correspond to specific maturities used for the SPD estimation discussed next. Therefore, after the simulation has been completed, the simulated data on the dates, which correspond to the desired maturities, is extracted, and the daily subjective density is estimated using a kernel regression on the desired moneyness grid, which corresponds to the asset's gross return. The transformation from the simulated S_t to the moneyness grid is achieved using $e^{-rT}\frac{S_T}{S_0}$ for each desired horizon T, where r is the daily average risk-free rate on the present day. The subjective density is estimated for every trading day included in the database. In figure 1 we plot the simulated subjective densities on four different trading days for four different maturities.

It can be seen in figure 1, that the distribution resembles a log normal distribution, which is more spread the longer the maturity is. A well known feature of financial data is that equity index return volatility is stochastic, mean-reverting and responds asymmetrically to positive and negative returns, due to the leverage effect. Therefore, this GARCH (1,1) model estimation, which experiences a slight positive skewness, is an adequate measure for the index returns, and it resembles the nonparametric subjective densities, which were estimated by Aït Sahalia & Lo (2000) and Brown & Jackwerth (2004).

3.3 State-Price Density Estimation

There is a vast literature on estimating the SPD using nonparametric and semiparametric methods. Aït Sahalia & Lo (2000), for example, suggest a semiparametric approach using the nonparametric kernel regression discussed in Härdle (1990). They propose a call pricing function according to Black & Scholes (1973), but with a nonparametric function for the volatility. The volatility is estimated using a two dimensional kernel estimator

$$\widehat{\sigma}(\kappa, \tau) = \frac{\sum_{i=1}^{n} k_\kappa\left(\frac{\kappa - \kappa_i}{h_\kappa}\right) k_\tau\left(\frac{\tau - \tau_i}{h_\tau}\right) \sigma_i}{\sum_{i=1}^{n} k_\kappa\left(\frac{\kappa - \kappa_i}{h_\kappa}\right) k_\tau\left(\frac{\tau - \tau_i}{h_\tau}\right)} \tag{34}$$

where $\kappa \stackrel{\text{def}}{=} \frac{K}{e^{r\tau}S_t}$ is future moneyness, τ is time to maturity and σ_i is the implied volatility. The kernel functions k_κ and k_τ together with the appropriate bandwidths h_κ and h_τ are chosen such that the asymptotic properties
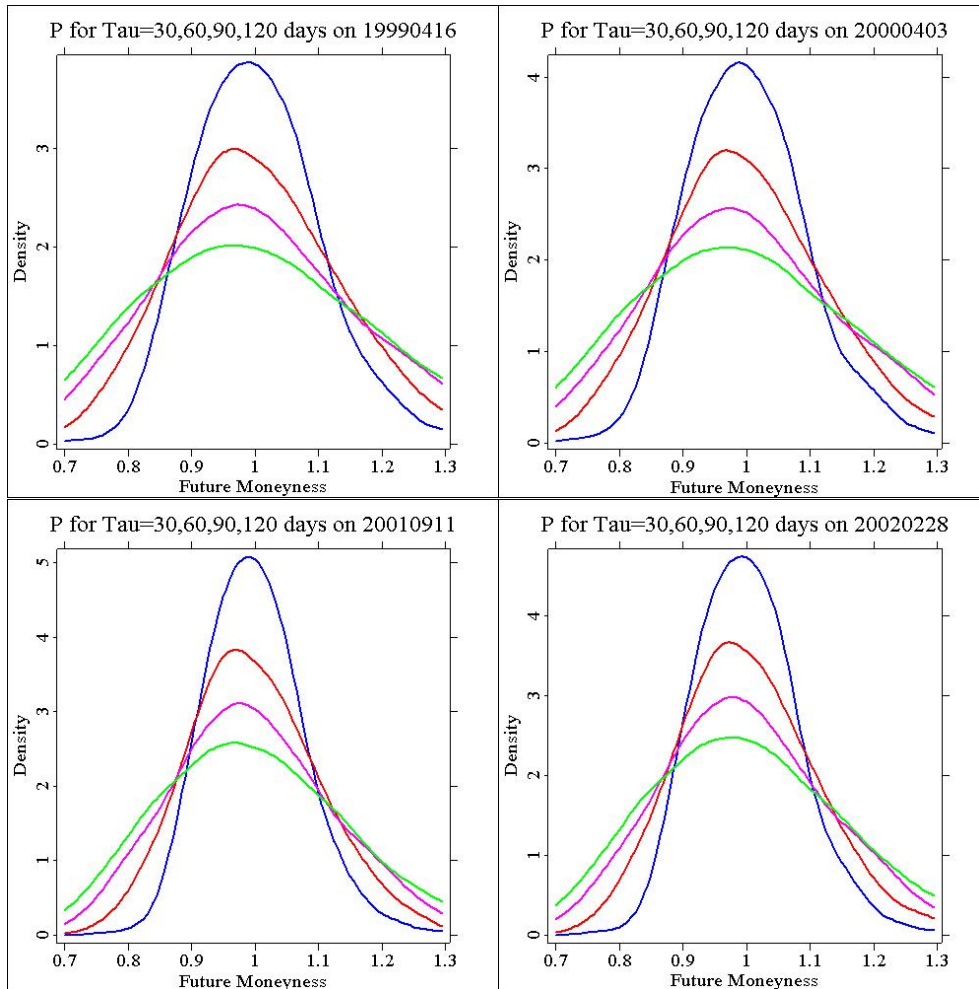
Figure 1: Subjective density for different maturities (30,60,90,120 days) on different trading days.

🔍 EPKdailyprocess.xpl

of the second derivative of the call price are optimized. The kernel function measures the drop of likelihood, that the true density function goes through a certain point, when it does not coincide with a certain observation. The price of the call is then calculated using the Black & Scholes (1973) formula but with the estimated volatility, and the SPD is estimated using equation (21).

A major advantage of such a method comparing to nonparametric ones is that only the volatility needs to be estimated using a nonparametric regression. The other variables are parametric, thus reducing the size of the problem

14

significantly. Other important qualities of kernel estimators are a well developed and tractable statistical inference and the fact that kernel estimators take advantage of past data, as well as future data, when estimating the current distribution. The problem of kernel based SPDs is that they could, for certain dates, yield a poor fit to the cross-section of option prices, although for other dates the fit could be quite good.

The state-price density in this work is estimated using a local polynomial regression as proposed by Rookley (1997) and described thoroughly in Huynh et al. (2002). The choice of Nadaraya-Watson type smoothers, used by Aït Sahalia & Lo (2000), is inferior to local polynomial kernel smoothing. More accurately, the Nadaraya-Watson estimator is actually a local polynomial kernel smoother of degree 0. If we use higher order polynomial smoothing methods, we can obtain better estimates of the functions. Local polynomial kernel smoothing also provides a convenient and effective way to estimate the partial derivatives of a function of interest, which is exactly what we look for when estimating SPDs.

The first step is to calculate the implied volatility for each given maturity and moneyness in the daily data (based on the B&S formula when prices are given and σ is the unknown). Then a local polynomial regression is used to smooth the implied volatility points and to create the implied volatility surface from which the SPD can be derived. The basic idea of local polynomial regression is based on a locally weighted least squares regression, where the weights are determined by the choice of a kernel function, the distance of an observation from a certain estimated point defining the surface/line at this coordinate and the chosen bandwidth vector. The use of the moneyness measure and time to maturity reduces the regression to two dimensions and enables freedom in estimating the surface in fictional points that do not exist in the database.

The concept of local polynomial estimation is quite straightforward. The input data at this stage is a trivariate data, a given grid of moneyness (κ), time to maturity (τ) and the implied volatility ($\sigma^{BS}(\kappa, \tau)$). We now consider the following process for the implied volatility surface

$$\widehat{\sigma} = \phi(\kappa, \tau) + \sigma^{BS}(\kappa, \tau) * \varepsilon \tag{35}$$

where $\phi(\kappa, \tau)$ is an unknown function, which is three times continuously differentiable, and ε is a Gaussian white noise. Then a Taylor expansion for

15

the function $\phi(\kappa, \tau)$ in the neighborhood of (κ_0, τ_0) is

$$
\begin{aligned}
\phi(\kappa, \tau) \approx \phi(\kappa_0, \tau_0) \quad &+ \quad \left.\frac{\partial \phi}{\partial \kappa}\right|_{\kappa_0, \tau_0} (\kappa - \kappa_0) + \frac{1}{2} \left.\frac{\partial^2 \phi}{\partial \kappa^2}\right|_{\kappa_0, \tau_0} (\kappa - \kappa_0)^2 \\
&+ \quad \left.\frac{\partial \phi}{\partial \tau}\right|_{\kappa_0, \tau_0} (\tau - \tau_0) + \frac{1}{2} \left.\frac{\partial^2 \phi}{\partial \tau^2}\right|_{\kappa_0, \tau_0} (\tau - \tau_0)^2 \\
&+ \quad \frac{1}{2} \left.\frac{\partial^2 \phi}{\partial \kappa \partial \tau}\right|_{\kappa_0, \tau_0} (\kappa - \kappa_0)(\tau - \tau_0) \quad\quad\quad (36)
\end{aligned}
$$

Minimizing the expression

$$
\begin{aligned}
\sum_{j=1}^{n} \Big\{ \sigma^{BS}(\kappa_j, \tau_j) &- [\beta_0 + \beta_1(\kappa_j - \kappa_0) + \beta_2(\kappa_j - \kappa_0)^2 + \beta_3(\tau_j - \tau_0) \\
&+ \beta_4(\tau_j - \tau_0)^2 + \beta_5(\kappa_j - \kappa_0)(\tau_j - \tau_0)]\Big\}^2 K_h(\kappa - \kappa_0)(\tau - \tau_0) \quad (37)
\end{aligned}
$$

yields the estimated implied volatility surface and its first two derivatives at the same time, as $\left.\widehat{\frac{\partial \phi}{\partial \kappa}}\right|_{\kappa_0, \tau_0} = \widehat{\beta_1}$ and $\left.\widehat{\frac{\partial^2 \phi}{\partial \kappa^2}}\right|_{\kappa_0, \tau_0} = 2\widehat{\beta_2}$. This is a very useful feature, as the second derivative is used to calculate the SPD for a certain fixed maturity. A detailed derivation of $\frac{\partial^2 C}{\partial K^2}$ (used for the SPD according to Breeden & Litzenberger (1978)) as a function of $\frac{\partial \sigma}{\partial \kappa}$ and $\frac{\partial^2 \sigma}{\partial \kappa^2}$ (which are obtained from the implied volatility surface estimation) is given, for example, by Huynh et al. (2002).

The estimated risk neutral densities for the same dates and the same maturities as in figure 1 are depicted in figure 2. The SPD is estimated on a future moneyness scale, thus reducing the number of parameters that need to be estimated.

One of the trading days plotted in figure 2 is September 11th 2001. It is interesting to see that the options data on this trading day reflects some increased investors' beliefs, that the market will go down in the long run. Similar behavior is found in the trading days following that particular day as well as in other days of crisis. The highly volatile SPD for negative returns, which could be explained, for example, by the leverage effect or the correlation effect, could reflect a dynamic demand for insurance against a market crash. This phenomenon is more apparent in days of crisis and was reported by Jackwerth (2000) as well.
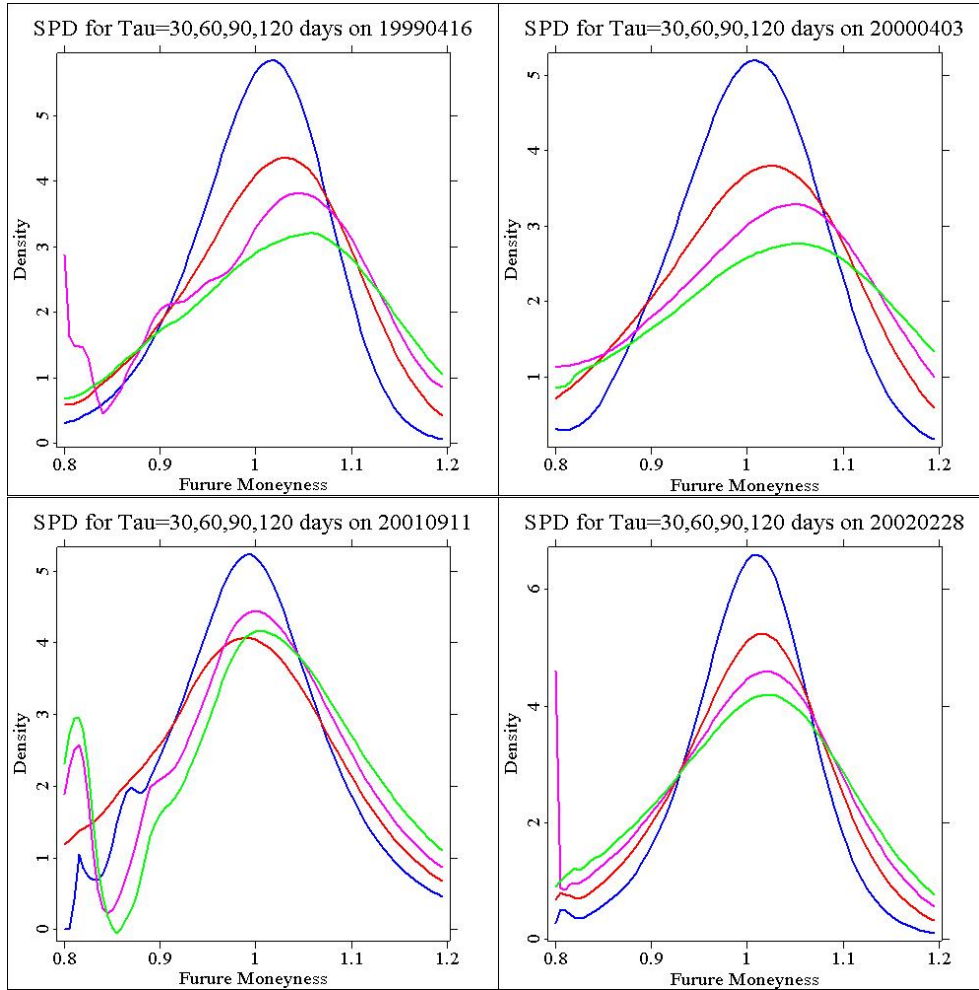
Figure 2: State-Price density for different maturities (30,60,90,120 days) on different trading days.

3.4 Deriving the Pricing Kernel and Risk Aversion

At this stage, we have the estimated subjective and state-price densities for the same maturities and spread over the same grid. The next step is to calculate the daily estimates for the pricing kernel and risk aversion.

The pricing kernel is calculated using equation (10), where the estimated subjective density and the estimated SPD replace $p(S_T|S_t)$ and $q(S_T|S_t)$ in the equation respectively. Since the grid is a moneyness grid, and the p and q are estimated on the moneyness grid, the estimated pricing kernel is

actually $M_t(\kappa_T)$. The coefficient of relative risk aversion is then computed by numerically estimating the derivative of the estimated pricing kernel with respect to the moneyness and then according to equation (14).

The estimated pricing kernels depicted in figure 3 for different trading days and different maturities bear similar characteristics to those reported by Aït Sahalia & Lo (2000), Jackwerth (2000), Rosenberg & Engle (2002) and others, who conducted a similar process on the S&P500 index. The pricing kernel is not a monotonically decreasing function, as suggested in classic macroeconomic theory. It is more volatile and steeply upward sloping for large negative return states, and moderately downward sloping for large positive return states. Moreover, the pricing kernel contains a region of increasing marginal utility at the money (around $\kappa = 1$), implying a negative risk aversion. This feature can clearly be seen in figure 4, which depicts the coefficient of relative risk aversion and shows clearly, that the minimal risk aversion is obtained around the ATM region and the relative risk aversion is negative. The negative risk aversion around the ATM region implies the possible existence of risk seeking investors, whose utility functions are locally convex.

Jackwerth (2000) named this phenomenon the *pricing kernel puzzle* and suggested some possible explanations to it. One possible explanation is that, a broad index (DAX in this work, S&P500 in his work) might not be a good proxy for the market portfolio and as such, the results are significantly different than those implied in the standard macroeconomic theory. In addition to the poor fit of the index, the assumptions for the existence of a representative agent might not hold, meaning that markets are not complete or the utility function is not strictly state-independent or time-separable.

Another possibility is that historically realized returns are not reliable indicators for subjective probabilities, or that the subjective distribution is not well approximated by the actual one. This deviation stems from the fact that investors first observe historical returns without considering crash possibilities, and only afterwards incorporate crash possibilities, which make their subjective distribution look quite different than the one estimated here. The historical estimation or the log-normal distribution assumptions ignore the well known volatility clustering of financial data.

Looking from another interesting point of view, investors might make mistakes in deriving their own subjective distributions from the actual objective one, thus leading to mispricing of options. Jackwerth (2000) claims, that mispricing of options in the market is the most plausible explanation to the negative risk aversion and increasing marginal utility function.

18
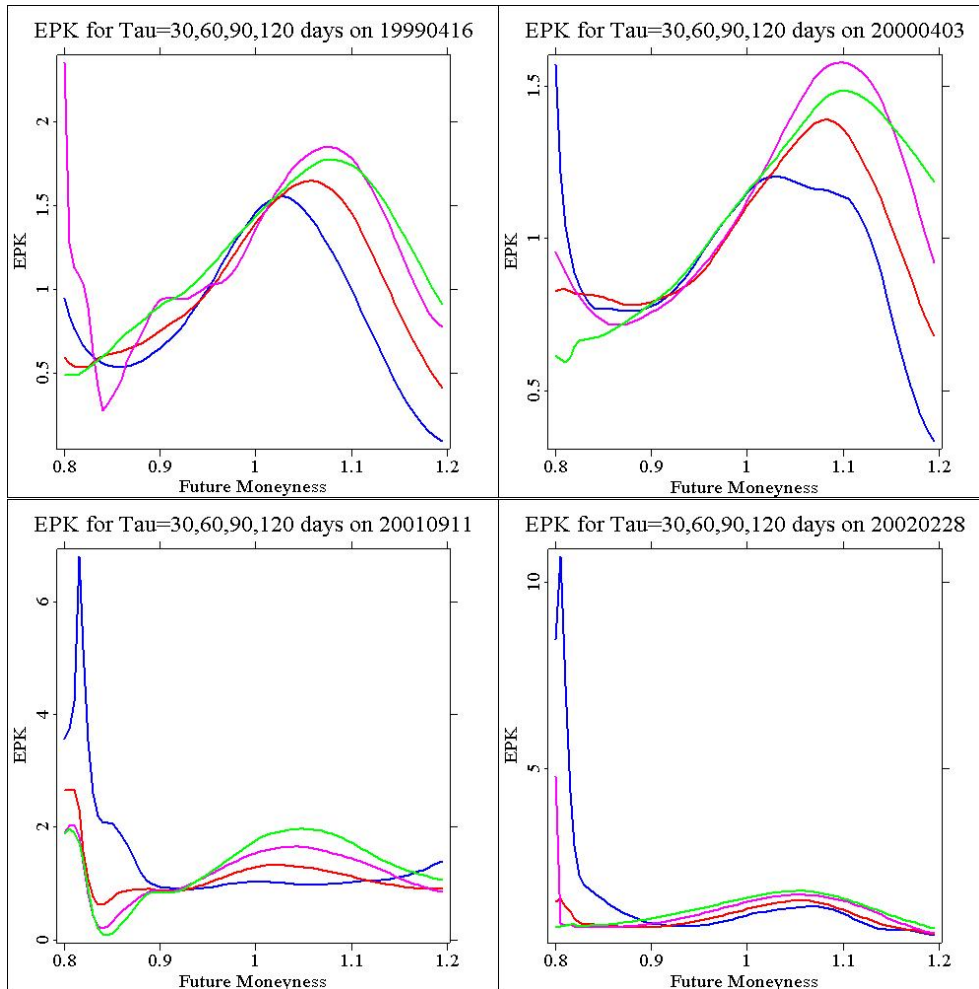
Figure 3: Estimated Pricing Kernel for different maturities (30,60,90,120 days) on different trading days.

EPKdailyprocess.xpl

This work does not aim, however, at finding a solution to the pricing kernel puzzle. The implicit assumption in this work is that some frictions in the market lead to the contradicting of standard macroeconomic theory, resulting in a region of increasing marginal utility. In the following section, a dynamic analysis of the pricing kernel and relative risk aversion is conducted along the three-year time frame.
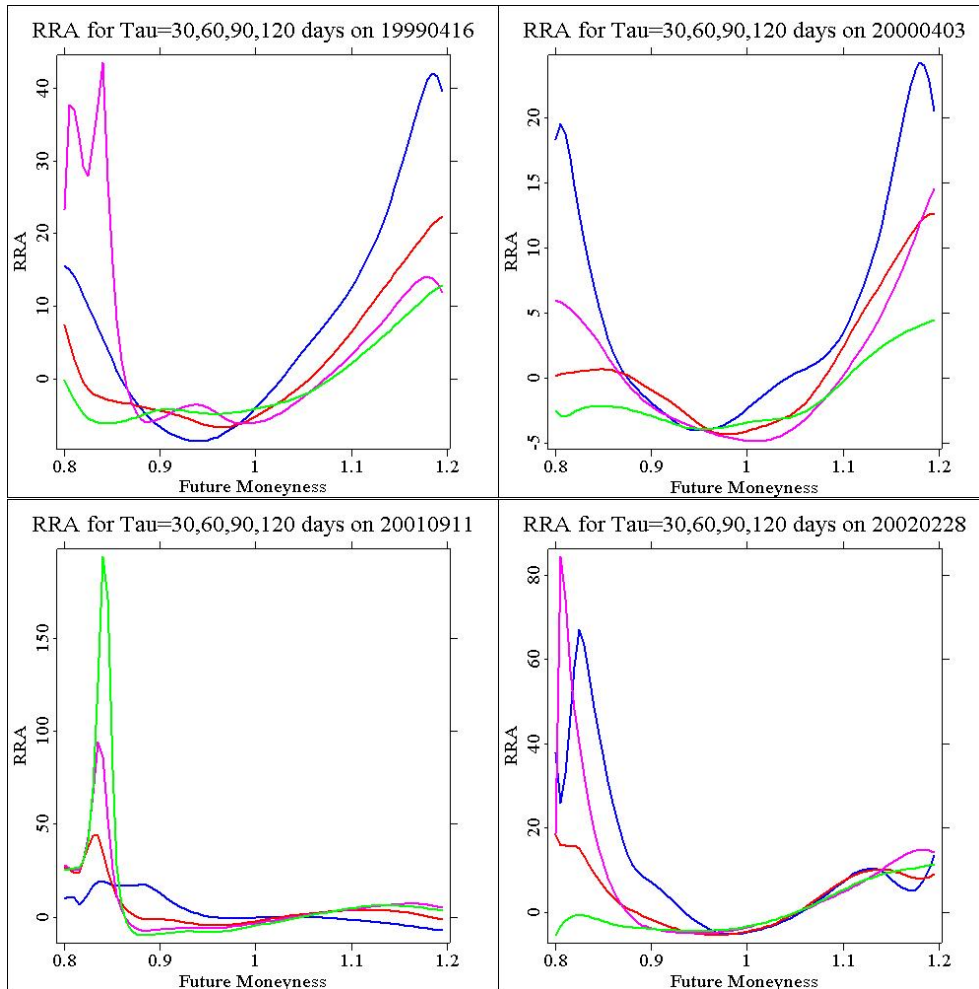
Figure 4: Estimated relative risk aversion for different maturities (30,60,90,120 days) on different trading days.

EPKdailyprocess.xpl

4 A Dynamic Model: Time-Series Analysis

Since the process described above is conducted on a daily basis and in most of the trading days, the GARCH and local polynomial estimations produce a good fit to the data, three-year long time-series data of pricing kernel and relative risk aversion are obtained. In this section we will analyze these time-series and show their moments. A principal component analysis will be conducted on the stationary series and the principal components will be tested as response variables in a GLS regression.

4.1 Moments of the Pricing Kernel and Relative Risk Aversion

In order to explore the characteristics of the pricing kernel and the relative risk aversion, their first four moments at any trading day have to be computed, i.e. the mean (μ_t), standard deviation (σ_t), skewness $(Skew_t)$ and kurtosis $(Kurt_t)$ of the functions across the moneyness grid. In addition, the daily values of the estimated functions at the money (ATM) are calculated and analyzed. Including this additional moment could prove essential as it was shown before that the functions behave quite differently at the money than in other regions. Each of the estimates (pricing kernel and relative risk aversion) is a function of moneyness and time to maturity, which was chosen to be a vector of four predetermined maturities, and as in the previous section we concentrate on $\tau = (30, 60, 90, 120)^\top$ days.

The figures in the following pages depict the time-series of the ATM values and mean values of the pricing kernel and the relative risk aversion, each estimated for four different maturities on 589 trading days between April 1999 and April 2002. The trading days, on which the GARCH model does not fit the data, or the local polynomial estimation experiences some negative volatilities, were dropped. Time-series of the daily standard deviation, skewness and kurtosis, as well as the differences time-series, were collected but not included in this paper.

The plots in the next pages show, that the pricing kernel at the money (figure 5) behaves similarly across different maturities and bears similar characteristics to its general mean (figure 6). This result implies, that characterizing the pricing kernel using the four first moments of its distribution is adequate. Contrary to the pricing kernel, the relative risk aversion at the money (figure 7) looks quite different than its general mean (figure 8). The ATM relative risk aversion is mostly negative, as detected already in the daily estimated relative risk aversion. The mean relative risk aversion, however, is mostly positive. Another feature of the relative risk aversion is that it becomes less volatile the longer the maturity is, implying the existence of more nervous investors for assets with short maturities. The main conclusion we can draw from the relative risk aversion plots is that the four first moments of the distribution do not necessarily represent all the features of the relative risk aversion correctly, and the collection of the extra details regarding the ATM behavior is justified, as it will be shown by the principal component analysis.
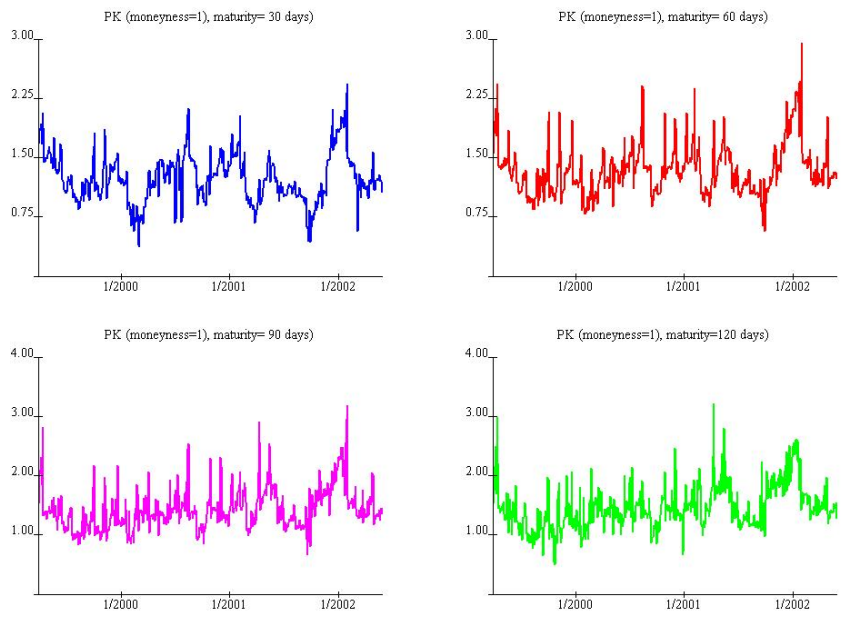
Figure 5: ATM Pricing Kernel for different maturities (30,60,90,120 days).
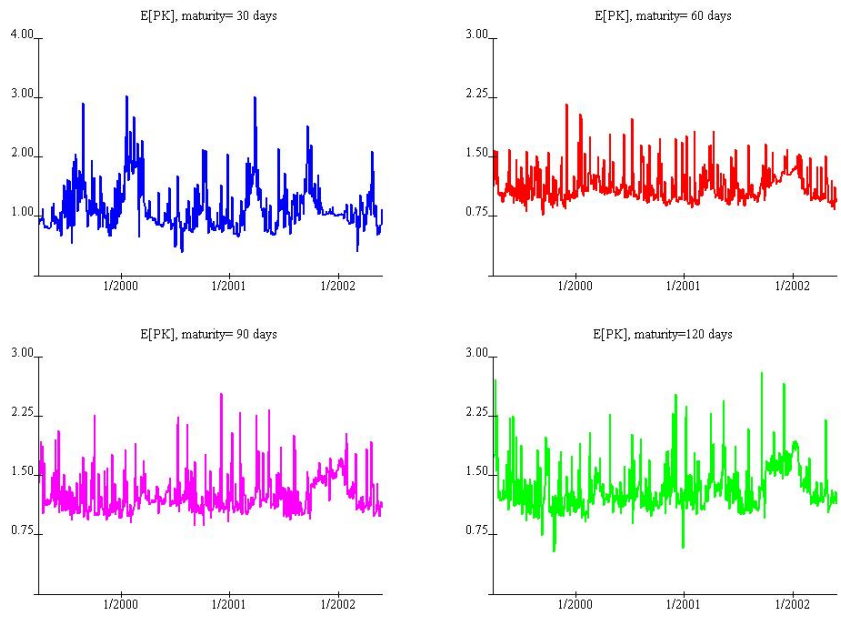
EPKtimeseries.xpl

Figure 6: Mean of Pricing Kernel for different maturities (30,60,90,120 days).
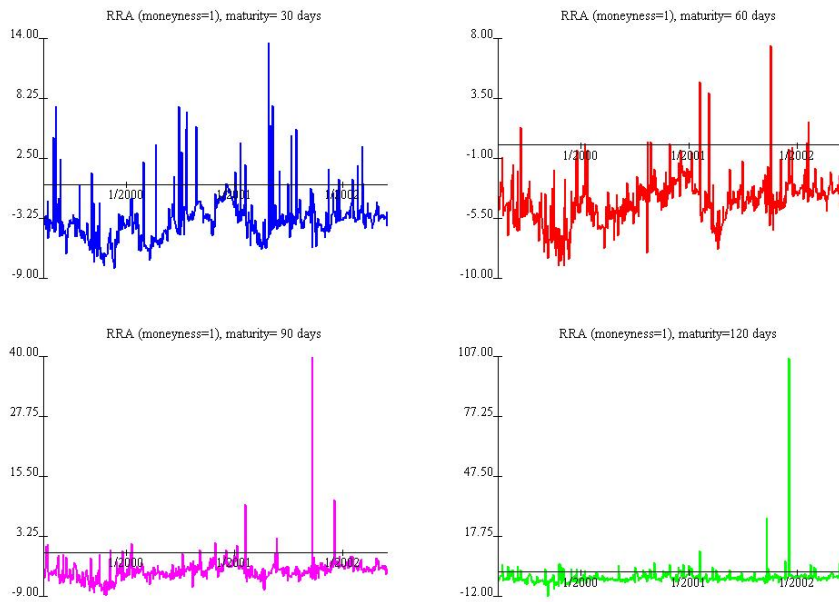
EPKtimeseries.xpl

Figure 7: ATM Relative Risk Aversion for different maturities (30,60,90,120 days).
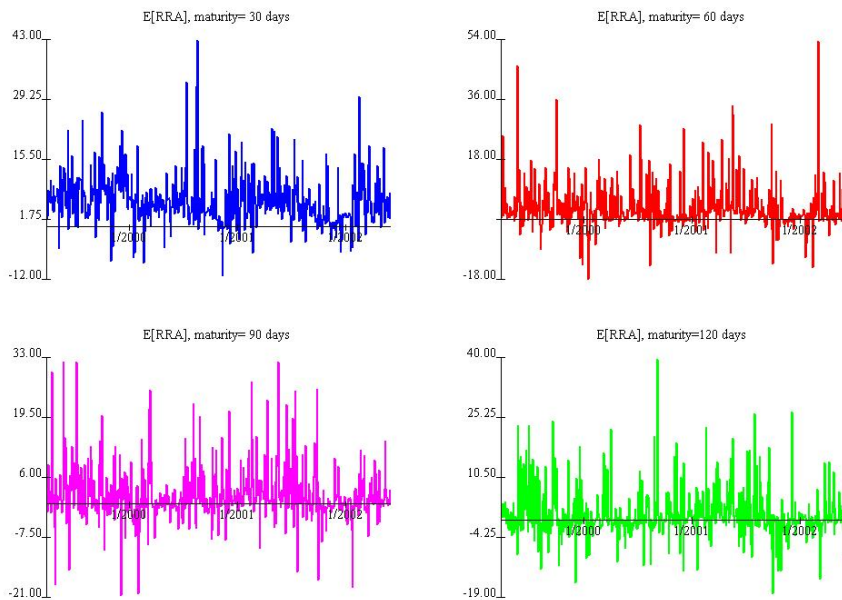
EPKtimeseries.xpl

Figure 8: Mean of Relative Risk Aversion for different maturities (30,60,90,120 days).

EPKtimeseries.xpl

After describing the characteristics of the different time-series, and before we concentrate on specific time-series for further analysis, it is essential to determine which of the time-series are stationary. The test chosen to check for stationarity is the KPSS test, originally suggested by Kwiatkowski et al. (1992).

Conducting stationarity tests for the various functions has shown, that the moments of the time-series themselves are in most of the cases not stationary, and the logarithmic differences of the moments are not always defined, due to the existence of negative values. Contrary to that, the absolute differences of all moments and across all maturities were found to be stationary. Therefore, we concentrate from now on only on the absolute differences of the moments.

4.2 Principal Component Analysis

In the following, we will focus on a principal component analysis (PCA) of the time-series in order to try and explain the variation of the time-series using a small number of influential factors. As stated before, the only time-series to be considered are the differences of the moments, found to be stationary. The PCA process starts with the definition of the following data matrix for pricing kernel differences

$$
\mathcal{X} = \begin{pmatrix} \Delta PK_2^{ATM} & \Delta\mu_2 & \Delta\sigma_2 & \Delta Skew_2 & \Delta Kurt_2 \\ \Delta PK_3^{ATM} & \Delta\mu_3 & \Delta\sigma_3 & \Delta Skew_3 & \Delta Kurt_3 \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ \Delta PK_n^{ATM} & \Delta\mu_n & \Delta\sigma_n & \Delta Skew_n & \Delta Kurt_n \end{pmatrix} \tag{38}
$$

for each maturity 30, 60 and 90 days, where the differences are defined e.g. as $\Delta\mu_t \stackrel{\text{def}}{=} \mu_t - \mu_{t-1}$ and similarly for the other columns of the matrix $\mathcal{X}$. A similar matrix is defined for the differences of the relative risk aversion. PCA can be conducted either on the covariance matrix of the variables or on their correlation matrix. If the variation were of the same scale, the covariance matrix could be used for the PCA. However, the data is not scale-invariant, hence a standardized PCA must be applied, i.e. conducting the PCA on the correlation matrix.

The principal components can explain the variability of the data. The proportion of variance explained by a certain principal component is the ratio of the corresponding eigenvalue of the correlation matrix to the sum of all eigenvalues, whereas the proportion of variance explained by the first few

principal components is the sum of the proportions of variance explained by each of them.

The principal component analysis shows, that three principal components could explain about 85% of the total variability. Nevertheless, the second and third principal components were found to be correlated, and in order to perform a univariate analysis on the principal components, they have to be orthogonal to each other. Therefore, only the first two principal components of the pricing kernel and relative risk aversion differences are considered from now on. The first two principal components explain approximately 80% of the variability of the pricing kernel differences (the first factor explains 60% and the second explains 20%), and approximately 70% of the variability of the relative risk aversion differences (divided equally among the two factors).

The jth eigenvector expresses the weights used in the linear combination of the original data in the jth principal component. Since we are considering only two principal components, the first two eigenvectors are of interest. More specifically, we can construct the first principal components for each of the examined time-series. The following demonstrates the weights of the moments in the principal components of the differences of the pricing kernel with a maturity of 60 days

$$
\begin{aligned}
y_{1,t}(\tau = 60) &= 0.06\Delta PK_t^{AT} + 0.92\Delta\mu_t + 0.38\Delta\sigma_t + 0.05\Delta Skew_t \\
&\quad -0.03\Delta Kurt_t \\
y_{2,t}(\tau = 60) &= 0.47\Delta PK_t^{ATM} + 0.24\Delta\mu_t - 0.58\Delta\sigma_t - 0.54\Delta Skew_t \\
&\quad +0.29\Delta Kurt_t
\end{aligned}
$$

It can clearly be seen, that the dominant factors in the first principal component are the changes in mean and standard deviation, whereas the dominant factors in the second principal component are the changes in skewness and standard deviation. The equations do not change much when other maturities are considered. As for the moments of the relative risk aversion, the first principal component is dominated solely by the changes in standard deviation and the second principal component is mainly dominated by the change in relative risk aversion at the money.

We conclude therefore, that the variation of the pricing kernel and relative risk aversion differences can be explained by two factors. The first factor of pricing kernel differences explains 60% of the variability and can be perceived as a central mass movement factor, consisting of the changes in expectation

and standard deviation. The second factor explains additional 20% of the variability and can be perceived as a change of tendency factor, consisting of changes in skewness and standard deviation. The principal components of the relative risk aversion are a little different. The first one explains approximately 35% of the variability and can be perceived as a dispersion change factor, dominated by the change in standard deviation. The contribution of the second principal component to the total variability is 35% as well and it is dominated by the change in relative risk aversion of the investors at the money. The mean of relative risk aversion differences seems to play no role in examining the variability of the relative risk aversion.

The correlation between the ith moment and the jth principal component is calculated as

$$r_{X_i, Y_j} = g_{ij} \sqrt{\frac{l_j}{s_{X_i X_i}}} \qquad (39)$$

where g_{ij} is the ith element of the jth eigenvector, l_j is the corresponding eigenvalue and $s_{X_i X_i}$ is the standard deviation of the ith moment X_i.

Descriptive statistics of the principal components time-series and their correlations with the moments are given in tables 1 and 2 for the pricing kernel and relative risk aversion respectively. The means of the principal components are very close to zero, as they are linear combinations of the differences of the moments, which are themselves approximately zero mean.

Principal Component	Mean $\times 10^4$	Standard Deviation	Correlation with				
			ΔPK_t^{ATM}	$\Delta \mu_t$	$\Delta \sigma_t$	$\Delta Skew_t$	$\Delta Kurt_t$
$\tau = 30$							
$y_{1,t}$	-2.46	0.76	-0.02	0.42	0.62	0.02	-0.02
$y_{2,t}$	-4.39	4.15	0.21	0.25	-0.16	0.29	0.08
$\tau = 60$							
$y_{1,t}$	4.34	0.44	0.06	0.74	0.30	0.04	-0.03
$y_{2,t}$	8.53	4.06	0.22	0.11	-0.27	-0.25	0.13
$\tau = 90$							
$y_{1,t}$	2.80	0.55	0.09	-0.61	0.46	0.11	-0.05
$y_{2,t}$	9.20	2.04	0.23	-0.19	-0.21	-0.32	0.11

Table 1: Descriptive statistics, principal components of the pricing kernel differences.

The moments highly correlated with the principal components are, not surprisingly, the ones which were reported to be dominant when constructing the

Principal	Mean	Standard	Correlation with				
Component	$\times 10^3$	Deviation	ΔRRA_t^{ATM}	$\Delta\mu_t$	$\Delta\sigma_t$	$\Delta Skew_t$	$\Delta Kurt_t$
$\tau = 30$							
$y_{1,t}$	11.5	14.75	0.03	0.04	0.61	0.00	0.01
$y_{2,t}$	0.55	9.36	0.33	-0.22	-0.02	-0.32	0.26
$\tau = 60$							
$y_{1,t}$	-2.57	26.90	0.10	0.04	0.60	-0.02	0.03
$y_{2,t}$	1.60	13.75	0.36	0.20	-0.06	-0.24	-0.35
$\tau = 90$							
$y_{1,t}$	1.72	28.60	-0.08	0.15	0.63	0.05	0.04
$y_{2,t}$	3.71	9.22	0.18	0.36	-0.05	-0.27	0.20

Table 2: Descriptive statistics, principal components of the relative risk aversion differences.

principal components. Nevertheless, table 1 implies an inconsistent behavior of the different moments across maturities. The first principal components of the pricing kernel differences (the first rows for each of the maturities in table 1) are positively correlated with the changes in mean and standard deviation (the dominating moments) for short term maturities, but negatively correlated with the mean differences of 90 days maturity pricing kernels. The second principal components of pricing kernel differences (the second rows for each of the maturities in table 1) are negatively correlated with the change of standard deviation for all maturities, but their correlations with the change of skewness are not consistent across maturities, implying a bad fit. Since the first principal component of the pricing kernel differences could explain approximately 60% of the variability, whereas the second factor can explain only 20%, the inconsistent behavior could be justified by the poor contribution of the second principal component to the total variability.

The correlations of the first and second principal components of the relative risk aversion differences with their dominant factors (table 2) are found to be consistent across maturities. The first principal component is positively correlated with its most dominant moment, the changes in the relative risk aversion standard deviation. This correlation means essentially, that the less homoscedastic the relative risk aversion is, i.e. the larger the changes in standard deviation are, the larger the first principal component of the relative risk aversion differences becomes. The second principal component of the relative risk aversion differences is positively correlated with its most dominant moment, the behavior at the money. The more volatile the relative risk
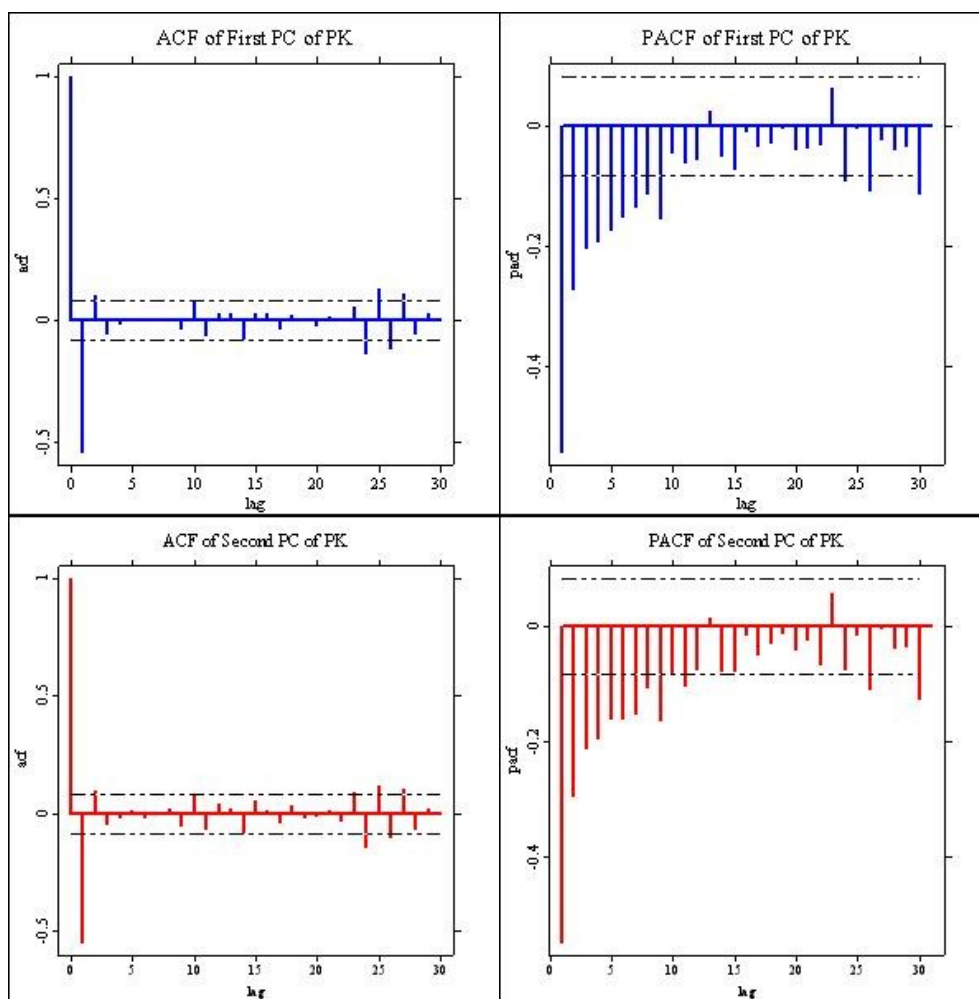
Figure 9: Autocorrelation function (left panel) and partial autocorrelation function (right panel) of the principal components of pricing kernel differences ($\tau = 60$ days). The autocorrelation functions of the principal components of relative risk aversion differences behave similarly exhibiting a MA(1) process.

EPKtimeseries.xpl

aversion at the money is, the higher the second principal component is. Both principal components of the relative risk aversion differences contribute more than 30% of the variability and imply a good fit of the principal components to the data.

After constructing principal components, which explain the variability of the time-series, it is essential to check the autocorrelation and the partial auto-correlation functions of the time-dependent principal components. This is illustrated in figure 9 for the pricing kernel differences. The same functions for the principal components of the relative risk aversion differences have similar characteristics and hence not reported here. Since the principal components have similar autocorrelation and partial autocorrelation functions for all different maturities, a maturity of 60 days was arbitrarily chosen to be presented. It can be seen, that the autocorrelation function drops abruptly after the first order autocorrelation whereas the partial autocorrelation function decays gradually. These characteristics imply a MA(1) behavior (Chapter 11 in Franke et al. (2004)) and we therefore concentrate on fitting a model with a moving average component to the principal components. A calculation of the Akaike and Schwarz information criteria confirms, that the best-fitted models for the first principal components are ARMA (1,1), whereas the second principal components follow a MA(1) process. As expected, all principal components have an autocorrelated error term.

4.3 GLS Regression Model for the Principal Components

The last test conducted in this work is to detect a possible relation between the principal components and easily observed data, such as changes in the DAX level and in implied volatility at the money. It is well known, that the simplest relation between an explanatory variable and a response variable can be described and examined using a simple linear regression model

$$y = X\beta + \epsilon \tag{40}$$

where y is a $n \times 1$ response vector, X is a $n \times p$ explanatory matrix, β is a $p \times 1$ vector of parameters to estimate and ϵ is a $n \times 1$ vector of errors. If the errors were normally distributed and uncorrelated, i.e. $\epsilon \sim \mathrm{N}_n(0, \sigma^2 I_n)$ then the regression would result in the familiar ordinary least squares (OLS) estimator

$$\widehat{\beta}_{OLS} = (X'X)^{-1}X'y \tag{41}$$

with a covariance matrix

$$\mathrm{Cov}(\widehat{\beta}_{OLS}) = \sigma^2(X'X)^{-1} \tag{42}$$

Introducing autocorrelated errors as described above, the relation between the explanatory variable and the response variable can be modeled using the

generalized least squares (GLS) estimator. In the previous section, we found evidence of autocorrelated errors of order 1, meaning that the error process could be modeled using the following AR(1) process

$$\epsilon_t = \rho\epsilon_{t-1} + u_t \tag{43}$$

for all $t \in \{1, \ldots, n\}$ with $u_t \sim N_n(0, \sigma_u^2 I_n)$ as i.i.d. white noise and $|\rho| < 1$ for stability. We could choose autoregressive processes of higher order, but since most principal components were found to have an autocorrelated error term of order 1, we concentrate here on AR(1) processes.

Iterating equation (43) from time 0 onwards yields

$$\epsilon_t = \lim_{n\to\infty} \left(\rho^{n+1}\epsilon_{t-n-1} + \sum_{s=0}^{n} \rho^s u_{t-s}\right) = \sum_{s=0}^{\infty} \rho^s u_{t-s} \tag{44}$$

and hence $E[\epsilon_t] = 0$ and the covariance matrix of the error term is

$$\text{Cov}(\epsilon) = \sigma_u^2 \Omega = \frac{\sigma_u^2}{1 - \rho^2} \begin{pmatrix} 1 & \rho & \rho^2 & \ldots & \rho^{n-1} \\ \rho & 1 & \rho & \ldots & \rho^{n-2} \\ \rho^2 & \rho & 1 & \ldots & \rho^{n-3} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \rho^{n-1} & \rho^{n-2} & \rho^{n-3} & \ldots & 1 \end{pmatrix} \tag{45}$$

However, in a real application like the model discussed in this work, the error-covariance matrix is not known and must be estimated from the data along with the regression coefficients $\widehat{\beta}$. If the generating process is stationary, which is the case in the model discussed here, a commonly used algorithm for estimating these errors is normally referred to as the Prais & Winsten (1954) procedure. This algorithm begins with running a standard OLS regression and examining the residuals. The errors vector of the OLS regression is obtained simply by plugging $\widehat{\beta}$ in equation (40). Considering the residuals' first order autocorrelations from the preliminary OLS regression can suggest a reasonable form for the error-generating process. These first order autocorrelations can be estimated as

$$\widehat{\rho} = \frac{\sum_{t=2}^{n} \epsilon_t \epsilon_{t-1}}{\sum_{t=1}^{n} \epsilon_t^2} \tag{46}$$

Replacing the ρ's in equation (45) with the $\widehat{\rho}$'s from equation (46) results in the estimated matrix $\widehat{\Omega}$. The best linear unbiased estimator in that case would be the estimated generalized least squares estimator

$$\widehat{\beta}_{GLS} = (X'\widehat{\Omega}^{-1}X)^{-1}X'\widehat{\Omega}^{-1}y \tag{47}$$

30

The Prais & Winsten (1954) algorithm may seem to be a simple model, but it involves a computationally challenging estimation of $\widehat{\Omega}$. Therefore, an alternative algorithm, suggested by Sen & Srivastava (1990) is presented here. We define the following matrix as

$$
\widehat{\Psi} = \begin{pmatrix}
\sqrt{1-\widehat{\rho}^2} & 0 & 0 & \dots & 0 & 0 & 0 \\
-\widehat{\rho} & 1 & 0 & \dots & 0 & 0 & 0 \\
0 & -\widehat{\rho} & 1 & \dots & 0 & 0 & 0 \\
\vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots \\
0 & 0 & 0 & \dots & -\widehat{\rho} & 1 & 0 \\
0 & 0 & 0 & \dots & 0 & -\widehat{\rho} & 1
\end{pmatrix}
\tag{48}
$$

It can be shown, that this matrix, multiplied by its transpose and the matrix $\widehat{\Omega}$ (which is defined by equation (45)) is proportional to the unit matrix

$$
\frac{1}{1-\widehat{\rho}^2} \widehat{\Psi}'\widehat{\Psi}\widehat{\Omega} = I_n
$$

and hence the matrix $\widehat{\Psi}$ has the following property

$$
\widehat{\Psi}'\widehat{\Psi} = (1-\widehat{\rho}^2)\widehat{\Omega}^{-1}
\tag{49}
$$

Since least squares estimation is not affected by scalar multiplication, we multiply the regression model by $\sqrt{1-\widehat{\rho}^2}$. Expressing $\widehat{\Omega}^{-1}$ in equation (47) using equation (49) leads to the the following GLS estimator

$$
\widehat{\beta}_{GLS} = (X'\widehat{\Psi}'\widehat{\Psi}X)^{-1}X'\widehat{\Psi}'\widehat{\Psi}y = [(\widehat{\Psi}X)'(\widehat{\Psi}X)]^{-1}[(\widehat{\Psi}X)]'(\widehat{\Psi}y)
\tag{50}
$$

which is actually an OLS estimator of the original variables multiplied by a scalar. The transformed model can be described as

$$
y_t - \widehat{\rho}y_{t-1} = \sum_{j=0}^{p}(x_{tj} - \widehat{\rho}x_{t-1,j})\widehat{\beta}_j + u_t
\tag{51}
$$

for $t \in \{2,\dots,n\}$, u_t being a Gaussian noise. For $t = 1$ it is simply

$$
\sqrt{1-\widehat{\rho}^2}y_1 = \sqrt{1-\widehat{\rho}^2}\sum_{j=0}^{p}\widehat{\beta}_j x_{1j} + \sqrt{1-\widehat{\rho}^2}\epsilon_1
\tag{52}
$$

As stated in the beginning of the current section, the changes in the DAX level (S_t) and the changes of ATM implied volatility (IV_t^{ATM}) were chosen to be tested as explanatory variables (X), whereas the first two principal

31

components of the pricing kernel and relative risk aversion differences for different maturities were the dependent variables for the different models (y). Since the dependency on the explanatory variable does not have to be linear, different functions of the explanatory variables were tested. For each of the explanatory variables the differences, the squared differences, the logarithmic differences and the squared logarithmic differences were tested. The examined models consisted of all possible combinations between the functions stated above, as well as checking for interactions in each of the proposed models. Since no interaction was ever found to be significant, they were dropped from the model. The criterion for choosing the best model was a maximal value of the F-statistic.

Table 3 describes the best fitted models for each of the principal components (based on equation (51)). For this analysis, we consider a confidence level of 95%, i.e. any regression or regression coefficient yielding a Pvalue > 5% is regarded as non significant. The Pvalues for the regressions' coefficients appear in brackets.

The first principal component of the pricing kernel differences, which was described before as a central mass movement factor, dominated by the changes in the mean pricing kernel and the pricing kernel's standard deviation, is found to depend significantly on the logarithmic differences of ATM implied volatility. This regression is only significant for short term maturities, and the impact of the explanatory variables is positive and log-linear. The impact of the DAX log return is not significant for a short term maturity, meaning the first principal component of the pricing kernel differences is mainly influenced by the logarithmic changes in the implied volatility at the money. Therefore, we can deduce the following: The larger the changes in ATM implied volatility are and the higher the DAX log returns are (only for maturities of 60 days), the more volatile the pricing kernel becomes, with bigger daily changes in its mean and standard deviation.

We can not find a significant relationship between the second principal component of the pricing kernel differences and the explanatory variables (other than for very short maturities), a result that supports the second principal component's smaller contribution to the variability of pricing kernel differences. The pricing kernel differences have one dominant factor which explains approximately 60% of their variance and depends mainly on the logarithmic changes of the ATM implied volatility. The regression coefficients are positive, as are the correlations of the first principal component with $\Delta\mu_t(PK)$ and $\Delta\sigma_t(PK)$ for the respective maturities.

Pricing Kernel Differences

PC	Maturity	$\widehat{\rho}$	$\widehat{\beta}_1$	$x_{t,1}$	$\widehat{\beta}_2$	$x_{t,2}$	F
1	30	-0.426	-1.80	$\log \frac{S_t}{S_{t-1}}$	1.76	$\log \frac{IV_t^{ATM}}{IV_{t-1}^{ATM}}$	18.958
			(0.289)		(0.000)		(0.000)
1	60	-0.468	2.71	$\log \frac{S_t}{S_{t-1}}$	0.98	$\log \frac{IV_t^{ATM}}{IV_{t-1}^{ATM}}$	10.784
			(0.005)		(0.001)		(0.000)
1	90		N o t		S i g n i f i c a n t		
2	30	-0.473	30.21	$\log \frac{S_t}{S_{t-1}}$	12.77	$\log \frac{IV_t^{ATM}}{IV_{t-1}^{ATM}}$	20.718
			(0.001)		(0.000)		(0.000)
2	60		N o t		S i g n i f i c a n t		
2	90		N o t		S i g n i f i c a n t		

Relative Risk Aversion Differences

PC	Maturity	$\widehat{\rho}$	$\widehat{\beta}_1$	$x_{t,1}$	$\widehat{\beta}_2$	$x_{t,2}$	F
1	30	-0.544	0.03	ΔS_t	145.34	ΔIV_t^{ATM}	11.562
			(0.000)		(0.000)		(0.000)
	60	-0.457	0.03	ΔS_t	286.43	ΔIV_t^{ATM}	18.048
			(0.001)		(0.000)		(0.000)
	90	-0.510	0.02	ΔS_t	224.27	ΔIV_t^{ATM}	10.666
			(0.028)		(0.000)		(0.000)
2	30		N o t		S i g n i f i c a n t		
	60	-0.460	-0.01	ΔS_t	-92.15	ΔIV_t^{ATM}	7.217
			(0.042)		(0.000)		(0.000)
	90	-0.497	0.01	ΔS_t	35.72	ΔIV_t^{ATM}	4.026
			(0.020)		(0.011)		(0.018)

Table 3: The estimated parameters of the suggested regression model in equation (51): $y_t - \widehat{\rho} y_{t-1} = \widehat{\beta}_0 + \widehat{\beta}_1(x_{t,1} - \widehat{\rho} x_{t-1,1}) + \widehat{\beta}_2(x_{t,2} - \widehat{\rho} x_{t-1,2}) + u_t$ where $\Delta IV_t^{ATM} \stackrel{\text{def}}{=} IV_t^{ATM} - IV_{t-1}^{ATM}$, $\Delta S_t \stackrel{\text{def}}{=} S_t - S_{t-1}$ and $\widehat{\beta}_0 = 0$ because the constant is never significant due to the zero mean property of the principal components.

The results regarding the principal components of the relative risk aversion differences are quite different. These principal components are related to the absolute changes in the DAX level and in ATM implied volatility. The dependence is not log-linear, but strictly linear.

According to table 2 in the previous section, the correlations of the first principal components of the relative risk aversion differences with their dominant moments are positive. The first principal component is a dispersion factor, dominated by the change in the relative risk aversion standard deviation. According to the regression, large changes in the DAX level and the ATM implied volatility yield a larger principal component, which is associated with a larger change in risk aversion standard deviation. This result implies the existence of more uncertain investors with a more heteroscedastic risk aversion, when the DAX level and ATM implied volatility are more time-varying. This relation could be explained by the dispersion of information sets among investors. Veldkamp (2005) examines the impact of information markets on assets prices. She basically claims, that information markets, not assets markets, are the source of frenzies and herds in assets prices. However, the price fluctuations on the market affect these information sets and determine the information prices, which are incorporated in the investors' subjective beliefs. More volatile markets lead necessarily to a higher risk and to less information, which increases the demand for information in a competitive market. Hence, more volatile markets cause more information to be provided at a lower price. When less information is involved, individual agents are willing to pay for information, and the information sets of the individual agents become more dispersed. More dispersed information sets could increase heteroscedasticity of the aggregate relative risk aversion as a function of assets' returns.

The results regarding the second principal component of the relative risk aversion differences are slightly different. The second principal components are positively correlated to the change of relative risk aversion at the money. Nevertheless, the linear regression is not significant for a very short term maturity of 30 days. For long term maturities the coefficients of the regression are positive, whereas for medium term maturities, they are negative. That could be interpreted as follows: When the changes in DAX level and ATM implied volatility are larger, the relative risk aversion at the money is more volatile for long term maturities, but is less volatile for the medium term maturities.

From this section we can conclude, that the principal components model fits the relative risk aversion differences better than it fits the pricing kernel differences. We were able to fit an autocorrelated regression model to the first

34

principal component of pricing kernel differences for short and medium term maturities, and to both principal components of relative risk aversion differences. The autocorrelation is indeed found to be quite large (approximately -0.5) for all of the above models, implying the existence of an autocorrelated error as detected already.

5   Final Statements

This work focused on estimating the subjective density and the state-price density of the stochastic process associated with the DAX. Based on the work of Rubinstein (1994), a good estimation of those two measures is sufficient for deriving the investors' preferences. However, this work did not include a direct approximation of the utility function based on empirical data, but rather an estimation of the pricing kernel and the relative risk aversion as functions of the return states. The utility function could be approximated numerically by solving the differential equations discussed in section 2, after the pricing kernel and relative risk aversion function have been estimated. Nevertheless, this work aimed at examining the dynamics of these two measures, characterizing the investors' behavior, rather than deriving their implied utility function.

The daily estimated pricing kernel and relative risk aversion were found to have similar characteristics to those reported by Jackwerth (2000) and Aït Sahalia & Lo (2000). The pricing kernel was shown not to be a strictly decreasing function as suggested by classical macroeconomic theory, and the relative risk aversion experienced some negative values at the money. These findings were apparent throughout the three year long database, implying existence of risk seeking investors with a locally convex utility function, possibly due to some frictions in the representative agent's model.

The variability of the stationary daily changes in pricing kernel and relative risk aversion was found to be well explained by two factors. Since the factors experienced some evident autocorrelation, the principal components were tested as the response variable in a GLS regression model, which regressed each of the principal components on the daily changes in the DAX and in ATM implied volatility.

We found that large changes in ATM implied volatility lead to a more volatile and time-varying pricing kernel. The absence of a significant fitted regression model for the second principal component of the pricing kernel differences was in accordance with its smaller contribution to the explained variability.

35

In addition, we found evidence for the existence of more uncertain investors with a more heteroscedastic risk aversion, when the daily changes in the DAX and the ATM implied volatility were larger. This result was explained by possibly more dispersed information sets among investors.

References

Aït Sahalia, Y. & Lo, A. W. (2000), 'Nonparametric risk management and implied risk aversion', *Journal of Econometrics* **94**, 9–51.

Arrow, K. J. (1964), 'The role of securities in the optimal allocation of risk bearing', *Review of Economic Studies* **31**, 91–96.

Arrow, K. J. (1965), 'Aspects of the theory of risk-bearing', *Yrjö Hahnsson Foundation, Helsinki* .

Black, F. & Scholes, M. (1973), 'The pricing of options and corporate liabililties', *Journal of Political Economy* **81**, 637–654.

Breeden, D. & Litzenberger, R. (1978), 'Prices of state-contingent claims implicit in option prices', *Journal of Business* **51**, 621–651.

Brown, D. P. & Jackwerth, J. C. (2004), 'The pricing kernel puzzle: Reconciling index option data and economic theory', *Working Paper, University of Konstanz / University of Wisconsin* .

Cochrane, J. H. (2001), 'Asset pricing', *Princeton University Press* .

Debreu, G. (1959), 'The theory of value', *Wiley, New York* .

Franke, J., Härdle, W. & Hafner, C. (2004), 'Statistics of financial markets', *Springer Verlag, Heidelberg* .

Härdle, W. (1990), 'Applied nonparametric regression', *Cambridge University Press, Cambridge* .

Huynh, K., Kervella, P. & Zheng, J. (2002), 'Estimating state-price densities with nonparametric regression', *in: W. Härdle, T. Kleinow and G. Stahl (eds.), Applied Quantitative Finance, Springer, Ch. 8* .

Jackwerth, J. C. (2000), 'Recovering risk aversion from option prices and realized returns', *Review of Financial Studies* **13**, 433–451.

Kwiatkowski, D., Phillips, P., Schmidt, P. & Shin, Y. (1992), 'Testing the null hypothesis of stationarity against the alternative of a unit root: How sure are we that economic series have a unit root', *Journal of Econometrics* **54**, 159–178.

Lucas, R. E. (1978), 'Asset prices in an exchange economy', *Econometrica* **46**, 1429–1446.

Mas-Colell, A., Whinston, M. & Green, J. (1995), 'Microeconomic theory', *Oxford University Press* .

Merton, R. (1973), 'Rational theory of option pricing', *Journal of Economics and Management Science* **4**, 141–183.

Prais, S. J. & Winsten, C. B. (1954), 'Trend estimators and serial correlation', *Cowles Commission Discussion Paper 383, Chicago* **383**.

Pratt, J. (1964), 'Risk aversion in the small and in the large', *Econometrica* **32**.

Rookley, C. (1997), 'Fully exploiting the information content of intra day option quotes: Applications in option pricing and risk management', *University of Arizona* .

Rosenberg, J. V. & Engle, R. F. (2002), 'Empirical pricing kernels', *Journal of Financial Economics* **64**, 341–372.

Rubinstein, M. (1976), 'The valuation of uncertain income streams and the pricing of options', *Bell Journal of Economics* **7**, 407–425.

Rubinstein, M. (1994), 'Implied binomial trees', *Journal of Finance* **49**, 771–818.

Sen, A. & Srivastava, M. (1990), 'Regression analysis: Theory, methods and applications', *Springer Verlag, New-York* .

Veldkamp, L. (2005), 'Media frenzies in markets for financial information', *American Economic Review, forthcoming* .

Dynamic Semiparametric Factor Models in Risk Neutral Density Estimation

Enzo Giacomini [a,*], Wolfgang Karl Härdle [a],
Volker Krätschmer [a,b]

[a] *Humboldt-Universität zu Berlin,*
CASE - Center for Applied Statistics and Economics
Spandauerstr. 1, 10178 Berlin, Germany

[b] *Technische Universität Berlin, Institute of Mathematics*
Straße des 17. Juni 136, 10623 Berlin, Germany

Abstract

Dimension reduction techniques for functional data analysis model and approximate smooth random functions by lower dimensional objects. In many applications the focus of interest lies not only in dimension reduction but also in the dynamic behaviour of the lower dimensional objects. The most prominent dimension reduction technique - functional principal components analysis - however, does not model time dependences embedded in functional data. In this paper we use dynamic semiparametric factor models (DSFM) to reduce dimensionality and analyse the dynamic structure of unknown random functions by means of inference based on their lower dimensional representation. We apply DSFM to estimate the dynamic structure of risk neutral densities implied by prices of option on the DAX stock index.

Key words: dynamic factor models, dimension reduction, risk neutral density
AMS classification: 62G08, 62M10, 62P05
JEL classification: C14, C32, G12

1 Introduction

Large datasets containing various samples of high dimensional observations became common in diverse fields of science with advances in measurement and computational techniques. In many applications the data come in curves, i.e.

---

* Corresponding author. Tel. + 49 30 2093 5721, Fax. + 49 30 2093 5649
  Email address: `giacomini@wiwi.hu-berlin.de` (Enzo Giacomini).

as observations of discretized values of smooth random functions, presenting evident functional structure. In these cases it is natural to perform statistical inference using functional data analysis techniques.

Consider a dataset $\{(Y_{jt}, X_{jt})\}$, $j = 1, \ldots, J_t$, $t = 1, \ldots, T$ containing noisy samples of a real valued smooth random function $\mathcal{F} \in L_2(\mathcal{X})$, $\mathcal{X} \subseteq \mathbb{R}^d$, $d \in \mathbb{N}$ evaluated at unbalanced design points X_{jt} as

$$Y_{jt} = \mathcal{F}_t(X_{jt}) + \varepsilon_{jt} \tag{1.1}$$

where ε_{jt} denotes unknown zero-mean error terms and $\{\mathcal{F}_t\}$ are i.i.d. realizations of $\mathcal{F}$. Each sample $S_t = \{(Y_{jt}, X_{jt}) : j = 1, \ldots, J_t\}$, $t = 1, \ldots, T$ may correspond to a different observation on e.g. a different individual, time period or experimental condition.

Examples in biomedicine are measurements across individuals of growth curves or of brain potentials obtained from EEG, see Kneip and Gasser (1992) and Gasser and Kneip (1995). In econometrics data may originate from expenditures on commodities across households, Kneip (1994), implied volatilities observed on different trading days, Fengler, Härdle, and Mammen (2007) or measurements on stock liquidity across time, Dähne, Härdle, and Hautsch (2008). Ramsay and Silverman (2005) and Ferraty and Vieu (2006) provide an extensive list of functional datasets encountered in various applications.

A large branch of functional data analysis concentrates on modelling and approximating the random function $\mathcal{F}$ by lower dimensional objects. Distributions on function spaces are highly complex objects and dimension reduction techniques present a feasible and interpretable approach to investigate them. Functional principal components analysis (FPCA), based on the Karhunen-Loève expansion of $\mathcal{F}$, is the most prominent and widely used dimension reduction technique, see Rao (1958), Rice and Silverman (1991) and Ramsay and Dalzell (1991).

Asymptotic results on FPCA have been obtained by Dauxois, Pousse, and Romain (1982) for directly observed functional data $\{\mathcal{F}_t(X_{jt})\}$. In cases where functional data is not directly observable a preliminary smoothing is required at each S_t and FPCA is performed on the smoothed $\{\widehat{\mathcal{F}}_t\}$, see Besse, Cardot, and Ferraty (1997) and Benko, Kneip, and Härdle (2008) for recent developments. In practical applications the previous fits may suffer, however, from design-sparseness at each S_t, Cont and da Fonseca (2002) and Fengler et al. (2007).

In general lines, previous literature combines PCA and dimension reduction with presmoothing for effective dimensional space at fixed time horizon. Various applications, however, involve modelling the dynamics of the unobserved

random functions and call for dimension reduction techniques that smooth in space and are parametric in time.

In this paper we simultaneously reduce dimensionality avoiding preliminary smoothing and consider the time structure in the sequence $\{\mathcal{F}_t\}$. We describe $\mathcal{F}_t$ as a linear combination of $L + 1 << T$ unknown smooth functions, called basis functions, $m_l \in L_2(\mathcal{X})$, $l = 0, \ldots, L$:

$$\mathcal{F}_t(X_{jt}) = \sum_{l=0}^{L} Z_{lt} m_l(X_{jt}). \tag{1.2}$$

Here $Z_t = (Z_{0t}, \ldots, Z_{Lt})^\top$ is an unobservable random vector taking values on $\mathbb{R}^{L+1}$. Defining the tuple of functions $m = (m_0, \ldots, m_L)^\top$, the dynamic semiparametric factor model (DSFM), Borak, Härdle, Mammen, and Park (2008), reads as

$$Y_{jt} = Z_t^\top m(X_{jt}) + \varepsilon_{jt}. \tag{1.3}$$

The basis functions are estimated nonparametrically avoiding specification issues, i.e. the shape of m_l is extracted from the data. As m and Z_t are estimated simultaneously, the smoothing is transfered directly to m_l, $l = 0, \ldots, L$ and design-sparseness becomes secondary. In addition, and essential for investigating dynamics, the random process $\{Z_t\}$ may be non-stationary.

The form (1.2) is justified when prior knowledge about the available dataset leads to expect some common structure generating each $\mathcal{F}_t$. Approximating it by a lower dimensional linear combination of common factors is a natural way to handle and describe the unobservable generating mechanism.

In many applications the index t reflects time evolution and the unobservable $\{Z_t\}$ contains the dynamics of $\{\mathcal{F}_t\}$. Borak et al. (2008) show that under (1.2) the autocorrelation structures of estimated $\{\widehat{Z}_t\}$ and true $\{Z_t\}$ are asymptotically equivalent. This result implies that no loss is incurred by inferring the dynamic structure from $\{\widehat{Z}_t\}$, i.e. there is no payment for not knowing the true $\{Z_t\}$. This fact is essential for investigating cointegration with another dynamical systems. An application using financial data is provided by Brüggemann, Härdle, Mungo, and Trenkler (2008) where the cointegration between estimates $\{\widehat{Z}_t\}$ and macroeconomic time series is analysed.

Note that in the very similar common regressors model from Kneip (1994) the unobservable functions are also considered a linear combination of unspecified common functions as in (1.2) and are estimated from the data. There are however crucial differences between DSFM and common regressors model:

3
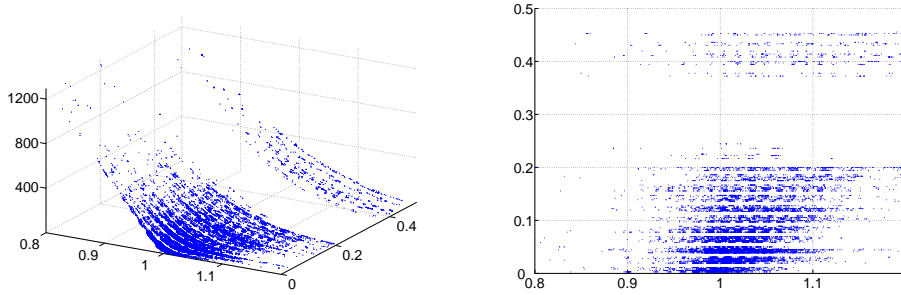
Fig. 1. Samples S_t, $t = 1, \ldots, 22$ of DAX call prices traded on January 2001 (left). Corresponding unbalanced design $\{X_{jt}\}$ (right)

(1) in DSFM $\{Z_t\}$ is a (non-stationary) random process with autocovariance structure inferable from $\{\widehat{Z}_t\}$
(2) DSFM is implementable in unbalanced designs
(3) DSFM avoids presmoothing by transferring the smoothing to the basis functions

Hence, in contrast to the dimension reductions techniques FPCA and common regressors, DSFM reduces dimensionality capturing the structural dynamics embedded in the observations.

We apply DSFM on a financial dataset of option prices in order to investigate the dynamics of risk assessments from investors acting in a market. Option prices are a valuable source of information concerning risk assessments from investors about future financial payoffs. The information is summarized in the risk neutral densities q, the continuous counterpart from Arrow-Debreu security prices, see Ait-Sahalia and Lo (1998). Under no arbitrage assumptions the risk neutral densities - corresponding to a risk neutral measure $\mathbb{Q}$ - are derived from prices of European call options.

An European call option on S_t with maturity date $T > 0$ and strike $K > 0$ is a financial instrument that delivers at time T the random payoff $(S_T - K)^+$ where S_t is the price of the underlying asset at time $0 \leq t \leq T$. Breeden and Litzenberger (1978) show that under no arbitrage assumptions the risk neutral density is obtained from the European call price function C_t through the relation

$$q_{t,T}(s_T | s_t) = e^{r(T-t)} \frac{\partial^2 C_t(s_t, K, T-t)}{\partial K^2} \bigg|_{K = s_T} \tag{1.4}$$

where $r > 0$ is interest rate, see Section (3) for details.

The knowledge about the risk neutral densities delivers (time dependent) risk assessments from investors concerning future payoffs and is essential for appli-

4

cations in financial economics, option pricing and risk management as, e.g. estimated risk neutral densities allow less liquid instruments or contracts traded over-the-counter to be priced. From the economic point of view it is natural to investigate the dynamics of the risk neutral densities and the dependence between risk assessments and macroeconomic and financial indicators. This calls for a method that reduces dimensionality and allows for dynamic analysis of the lower dimensionals. Both features are present in DSFM.

We estimate the risk neutral densies based on a dataset containing intraday prices of calls on the German stock index (DAX) from 2001. Each observation consists of a call price Y_{jt} on a design point $X_{jt} = (\kappa_{jt}, \tau_{jt})^\top$ where $j = 1, \ldots, J_t$ denotes the transactions at day $t = 1, \ldots, T$ and $\kappa = g(K)$ is the moneyness, a monotone transformation of strikes K. Exchange regulations impose prespecified values for tradable maturities $\tau = T - t$ and are responsible for the observed degenerated design, see figure 1.

Following Ait-Sahalia and Lo (1998) and Fengler et al. (2007) the observations are transformed into log-implied volatilities $\widetilde{Y}_{jt} = \log C_{BS}^{-1}(Y_{jt})$, see Section (3) for details. These are assumed as discretized noisy values of the log-implied volatility surface evaluated at design points $\{X_{jt}\}$:

$$\widetilde{Y}_{jt} = \log \mathcal{V}_t(X_{jt}) + \varepsilon_{jt}. \tag{1.5}$$

where the smooth random function $\mathcal{V} \in L_2(\mathcal{X})$, $\mathcal{X} \subset \mathbb{R}_+^2$, is called the implied volatility surface and ε_{jt} is an error term. The realizations $\{\mathcal{V}_t\}$ are filtered out from the transformed data with DSFM and the risk neutral densities estimated using (1.4) with $C_{BS}(\widehat{\mathcal{V}}_t)$ as an estimator for C_t. The dynamics of the estimated $\{\widehat{q}_{t,T}\}$ is analysed based on the autocorrelation structure of $\{\widehat{Z}_t\}$.

In the sequel the DSFM estimation method and its asymptotic properties are described (Section 2). In the application part (Section 3), risk neutral densities are defined and estimated from observed prices of European call options on the DAX index (ODAX dataset). Their dynamic structure is analysed by vector autoregressive models.

2 Estimation Method

Consider a dataset $\{(Y_{jt}, X_{jt})\}$, $j = 1, \ldots, J_t$, $t = 1, \ldots, T$ such that

$$Y_{jt} = \sum_{l=0}^{L} Z_{lt} m_l(X_{jt}) + \varepsilon_{jt} \tag{2.1}$$

where ε_{jt} is an unknown error terms with $E[\varepsilon_{jt}] = 0$ and $E[\varepsilon_{jt}^2] < \infty$ and $\{\varepsilon_{jt}\}$ are independent. Here $Z_t = (Z_{0t}, \ldots, Z_{Lt})^\top$ is an unobservable random vector taking values on $\mathbb{R}^{L+1}$ with $Z_{0t} = 1$ and $m_l \in L_2(\mathcal{X})$, $l = 0, \ldots, L$ are unknown smooth functions, called basis functions, mapping $\mathcal{X} \subseteq \mathbb{R}^d$, $d \in \mathbb{N}$ into real values.

Following Borak et al. (2008), the basis functions are estimated using a series expansion. Defining K normed functions $\psi_k : \mathcal{X} \to \mathbb{R}$, $\int_{\mathcal{X}} \psi_k^2(x)dx = 1$, $k = 1, \ldots, K$ and a $(L+1 \times K)$ matrix of coefficients $\Gamma = (\gamma_{l,k})$, $\gamma_{l,k} \in \mathbb{R}$, the tuple of functions $m = (m_0, \ldots, m_L)^\top$ is approximated by $\Gamma^\top \psi$ where $\psi = (\psi_1, \ldots, \psi_K)^\top$. For simplicity of notation we assume that $J_t = J$ does not depend on t. We define the least square estimators as

$$(\widehat{\Gamma}, \widehat{Z}) = \arg \min_{\Gamma \in \mathcal{G}, Z \in \mathcal{Z}} \sum_{t=1}^{T} \sum_{j=1}^{J} \left\{ Y_{jt} - Z_t^\top \Gamma \psi(X_{jt}) \right\}^2 \tag{2.2}$$

where $\mathcal{G} = \mathcal{M}(L+1, K)$, $\mathcal{Z} = \{Z \in \mathcal{M}(T, L+1) : Z_{0t} = 1\}$ and $\mathcal{M}(a,b)$ is the set of all $(a \times b)$ matrices. The basis functions m are estimated by $\widehat{m} = \widehat{\Gamma}\psi$.

Theorem (2.1) gives the asymptotic behaviour of the least squares estimators $(\widehat{\Gamma}, \widehat{Z})$.

Theorem 2.1 *Suppose that DSFM holds and that $(\widehat{\Gamma}, \widehat{Z})$ is defined by (2.2). Under assumptions (A1)-(A8), see Appendix A, it holds for $K, J \to \infty$:*

$$\frac{1}{T} \sum_{1 \le t \le T} \left\| \widehat{Z}_t^\top \widehat{\Gamma} - Z_t^\top \Gamma^* \right\|^2 = \mathcal{O}_P(\rho^2 + \delta_K^2)$$

See Borak et al. (2008) for the proof. Note that the model (2.1) is only identifiable up to linear transformations. Consider a $(L+1 \times L+1)$ regular matrix $B = (b_{mn})$ with $b_{1n} = \delta_{1n}$ and $b_{m1} = \delta_{m1}$ for $m, n = 1, \ldots, L+1$ where $\delta_{mn} = \mathbf{1}(m = n)$. Define $Z_t^* = B^\top Z_t$, $m^* = B^{-1}m$, then from (1.2)

$$\begin{aligned}
\mathcal{F}_t(X) &= Z_t^\top m(X) \\
&= Z_t^\top B B^{-1} m(X) \\
&= Z_t^{*\top} m^*(X)
\end{aligned}$$

for $X \in \mathcal{X}$. On the other hand it is always possible to chose orthonormal basis functions by setting $m^* = Hm$ where H is an orthogonal matrix.

Theorem (2.2) states that for any $\widehat{Z}_t$ there exists a random matrix B such that the autocovariances of $\{\widehat{Z}_t\}$, $\widetilde{Z}_t = B^\top \widehat{Z}_t$ are asymptotically equivalent to the

6

autocovariances of the true unobservable $\{Z_t\}$. This equivalence is transfered to classical estimation and testing procedures in the context of e.g. vector autoregressive models and in particular justifies inference based on $\{\widehat{Z}_t\}$ when $\{Z_t\}$ is a VAR process. Define for $H_t \in \mathcal{Z}$, $t = 1, \ldots, T$: $\overline{H} = T^{-1} \sum_{t=1}^{T} H_t$, $H_{c,t} = H_t - \overline{H}$ and $H_{n,t} = (T^{-1} \sum_{s=1}^{T} H_{c,s} H_{c,s}^{\top})^{-1/2} H_{c,t}$.

Theorem 2.2 *Suppose that DSFM holds and that $(\widehat{\Gamma}, \widehat{Z})$ is defined by (2.2). Under assumptions (A1)-(A11), see Appendix A, there exists a random matrix B such that for $h \neq 0$, $h_d = \max(1, 1 - h)$, $h_u = \max(T, T - h)$ and $T \to \infty$:*

$$\frac{1}{T} \sum_{t=h_d}^{h_u} \widetilde{Z}_{c,t} \left(\widetilde{Z}_{c,t+h} - \widetilde{Z}_{c,t} \right)^{\top} - \frac{1}{T} \sum_{t=h_d}^{h_u} Z_{c,t} \left(Z_{c,t+h} - Z_{c,t} \right)^{\top} = \mathcal{O}_P(T^{-1/2})$$

where $\widetilde{Z}_t = B^{\top} \widehat{Z}_t$. Moreover

$$\frac{1}{T} \sum_{t=h_d}^{h_u} \widetilde{Z}_{n,t} \widetilde{Z}_{n,t+h}^{\top} - \frac{1}{T} \sum_{t=h_d}^{h_u} Z_{n,t} Z_{n,t+h}^{\top} = \mathcal{O}_P(T^{-1/2})$$

See Borak et al. (2008) for the proof. Note that, in contrast to FPCA, DSFM does not require stationarity, neither for $\{Z_t\}$ nor $\{\varepsilon_t\}$, but only weak assumptions on the average behaviour of Z_t, like being a martingale difference, see Appendix A.

3 Application

Consider a financial market with one risky asset and one riskless bond with constant interest rate $r > 0$. Let the price of the asset traded on the market be described by the real valued random process $\{S_t\}$, $t = [0, T]$, $T < \infty$ on a filtered probability space $(\Omega, \{\mathcal{F}_t\}, \mathbb{P})$ with $\mathcal{F}_t = \sigma(S_u, u \leq t)$ and $\mathcal{F}_0 = \{\emptyset, \Omega\}$. Assume further no arbitrage in the financial market in the sense that there exists a (risk neutral) probability measure $\mathbb{Q}$ equivalent to $\mathbb{P}$ under which the discounted price process $\{e^{-rt} S_t\}$ is a martingale.

A European call option at strike $K > 0$ is a financial instrument that pays $\Psi(S_T) = (S_T - K)^+$ at time T. By the risk-neutral valuation principle w.r.t. $\mathbb{Q}$ the price C_t of a European call option at time t is defined to be

$$C_t = e^{-r(T-t)} E^{\mathbb{Q}}[\Psi(S_T) | \mathcal{F}_t]$$

Assuming that $\{S_t\}$ is a $\mathbb{Q}$-Markov process and denoting the $\mathbb{P}$-density of $\mathbb{Q}$ by π, the price can be rewritten as

$$
\begin{aligned}
C_t(S_t) &= e^{-r(T-t)} E^{\mathbb{Q}} \left[\Psi(S_T) | S_t \right] \\
&= e^{-r(T-t)} E \left[\Psi(S_T) \mathcal{K}_\pi^t(S_t, S_T) | S_t \right]
\end{aligned}
$$

where E denotes the expectation under $\mathbb{P}$ and $\mathcal{K}_\pi^t(S_t, S_T) \stackrel{\text{def.}}{=} \frac{E[\pi | S_t, S_T]}{E[\pi | S_t]}$. The conditional risk neutral distribution of S_T is defined as

$$
Q_{S_T | S_t = s_t}([S_T \leq x]) \stackrel{\text{def.}}{=} \int_{-\infty}^{x} \mathcal{K}_\pi^t(s_t, \cdot) \, dP_{S_T | S_t = s_t} \tag{3.1}
$$

where $P_{S_T | S_t = s_t}$ is the conditional distribution of S_T under $S_t = s_t$. Specializing to the following two factor model we assume that the price process has dynamics given by

$$
dS_t = S_t \mu(Y_t) dt + S_t \sigma(Y_t) dW_t^1
$$

here W^1 is a standard $\mathbb{P}$-Brownian motion and Y denotes an external economic factor process modelled by

$$
dY_t = g(Y_t) + \rho dW_t^1 + \bar{\rho} dW_t^2
$$

where $\rho \in [-1, 1]$ is some correlation factor, $\bar{\rho} \stackrel{\text{def.}}{=} \sqrt{1 - \rho^2}$ and W^2 is a standard $\mathbb{P}$-Brownian motion independent of W^1 under $\mathbb{P}$. Market models of this type are popular in mathematical finance and economics, in particular if Y follows and Ornstein-Uhlenbeck dynamic with mean reversion term $g(y) = \kappa(\theta - y)$ for constants $\theta \geq 0$ and $\kappa > 0$. Moreover, $\{S_t\}$ is a $\mathbb{Q}$-Markov process for any $\mathbb{Q}$, see Hernández-Hernández and Schied (2007) and the conditional risk neutral distribution $Q_{S_T | S_t = s_t}$ has a density function denoted by $q_{t,T}(\cdot | s_t)$. Hence, the call prices C can be expressed as

$$
C_t(s_t, K, T - t) = e^{r(T-t)} \int (s_T - K)^+ q_{t,T}(s_T | s_t) ds_T.
$$

We assume that the observed prices in the financial market are built based on the risk neutral valuation principle w.r.t. an unknown risk neutral measure $\mathbb{Q}$. Our interest lies in estimating the conditional risk neutral distribution $Q_{S_T | S_t = s_t}$, or equivalently its density function $q_{t,T}(\cdot | s_t)$, implied by $\mathbb{Q}$ through (3.1).

Adapting Breeden and Litzenberger (1978), one can show that the density function $q_{t,T}(\cdot|s_t)$, called from now on the risk neutral density, is obtained as the second derivative of the call price function C_t with respect to strike K

$$q_{t,T}(s_T|s_t) = e^{r\tau} \frac{\partial^2 C_t(s_t, K, \tau)}{\partial K^2}\bigg|_{K=s_T} \tag{3.2}$$

here $\tau = T - t$ is the time to maturity. The price function can be smoothed out of observations from call prices and used in (3.2) to recover state price densities implied on the data.

Following Ait-Sahalia and Lo (1998) smoothing is carried out in the space of implied volatilities. In general the exact form of C_t is hard to determine and based on historical data an estimate of C_t can be obtained by non-parametric regression. In order to cope with the course of dimensionality, Ait-Sahalia and Lo (1998) proposed a semiparametric approach using the Black-Scholes formula evaluated with a non-parametric estimator for the implied volatility in place of the true non-observable volatility.

The implied volatility surface is the function $\sigma_t : \mathbb{R}_+^2 \to \mathbb{R}_+$ satisfying for all $(K, \tau) \in \mathbb{R}_+^2$

$$C_t(s_t, K, \tau) = C_{BS}\{s_t, r_t, K, \tau, \sigma_t(K, \tau)\} \tag{3.3}$$

where $C_{BS}(s, r, K, \tau, \sigma) = s\Phi(d_1) - Ke^{-r\tau}\Phi(d_2)$ is the Black-Scholes price of Ψ with strike K and maturity τ. Here $\Phi(x)$ is the standard normal cdf, $d_1(\sigma) = \left\{\log\left(\frac{s}{K}\right) + (r + \frac{1}{2}\sigma^2)\tau\right\}/(\sigma\sqrt{\tau})$ and $d_2(\sigma) = d_1 - \sigma\sqrt{\tau}$. Note that $C_{BS}(v) = C_{BS}(s, r, K, \tau, v)$ is a continuous increasing function of v, hence $\sigma_t(K, \tau) = C_{BS}^{-1}\{C_t(s_t, K, \tau)\}$.
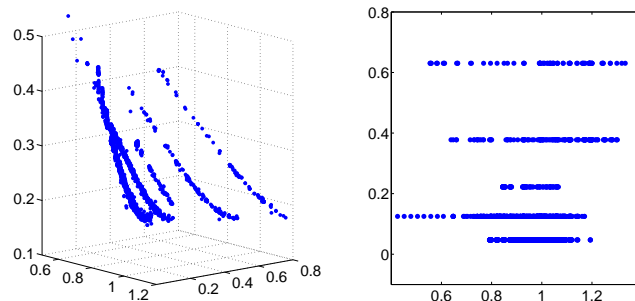
Fig. 2. Call and put implied volatilities observed (left), data design (right), ODAX on 20000502

More generally, the implied volatility surface is considered a smooth random function $\mathcal{V} \in L_2(\mathcal{X})$ on the space $\mathcal{X} \subset \mathbb{R}^2$ of strikes K and maturities τ. Combining (3.2) and (3.3) the functional random variable $\mathcal{H} \in L_2(\mathcal{X})$ called the risk neutral (RN) surface is defined as

$$\mathcal{H} = e^{r\tau} D^2 C_{BS}(\mathcal{V}) \tag{3.4}$$

where D^m denotes the m-th derivative with respect to K. Clearly lower dimension objects describing $\mathcal{V}$ may be used to analyse $\mathcal{H}$.

A functional dataset containing realizations of $\mathcal{V}$ is however not available, as in an exchange only discretized values of $\mathcal{V}_t$ corrupted by noise are registered from trades. More specifically, on each day $t = 1, \ldots, T$ there are J_t options traded, each intra-day trade $j = 1, \ldots, J_t$ corresponds to an observed option price Y_{jt} at a pair of moneyness and maturities $X_{jt} = (\kappa_{jt}, \tau_{jt})^\top$ where $\kappa = e^{r\tau} K/S$. The observed implied volatilities at the respective design points are obtained by $\sigma_{jt} = C_{BS}^{-1}(Y_{jt})$, see figure (2). For numerical tractability we perform a further transformation into log-implied volatilities $\widetilde{Y}_{jt} = \log \sigma_{jt}$ and use DSFM to model

$$\widetilde{Y}_{jt} = \log \mathcal{V}_t(X_{jt}) + \varepsilon_{jt}$$

where $\log \mathcal{V}_t = Z_t^\top m$. The implied volatility surface at t is estimated by $\widehat{\mathcal{V}}_t = \exp(\widehat{Z}_t^\top \widehat{\Gamma} \psi)$, recall (2.2). The RN surface is estimated using (3.4) by $\widehat{\mathcal{H}}_t = \mathcal{H}(\widehat{\mathcal{V}}_t)$ where

$$\mathcal{H}(\mathcal{V}) = \varphi(d_2) \left\{ \frac{1}{K\sqrt{\tau}\mathcal{V}} + 2d_1 \frac{D\mathcal{V}}{\mathcal{V}} + K\sqrt{\tau} d_1 d_2 \frac{(D\mathcal{V})^2}{\mathcal{V}} + K\sqrt{\tau} D^2 \mathcal{V} \right\} \tag{3.5}$$

and $d_1 = d_1(\mathcal{V})$ and $d_2 = d_2(\mathcal{V})$. The dynamics of the unobservable sequence of RN surfaces $\{\mathcal{H}_t\}$ implied in the observations $\{(Y_{jt}, X_{jt})\}$ may be investigated by analysing the lower dimensional $\{\widehat{Z}_t\}$.

3.2 Results

Here implied volatility and RN surfaces are estimated with DSFM from in-traday call prices on the DAX index, i.e. S_t represents the value of the DAX index at time t. The time ranges from 20010101 to 20020101 (dates are written as year, month, day) corresponding to 253 trading days.

Tensor B-splines, quadratic in τ and cubic in κ directions placed on 8×6 knots, are used for the series estimators of m. The dimension L of Z_t is chosen based on

$$RV(L) = \frac{\sum_{t=1}^{T} \sum_{j=1}^{J_t} \left\{ \widetilde{Y}_{jt} - \widehat{Z}_t^\top \widehat{m}(X_{jt}) \right\}^2}{\sum_{t=1}^{T} \sum_{j=1}^{J_t} (\widetilde{Y}_{jt} - \overline{Y})^2}$$

where $\overline{Y} = (\sum_{t=1}^{T} \sum_{j=1}^{J_t} \widetilde{Y}_{jt}) / \sum_{t=1}^{T} J_t$. The value $1 - RV(L)$ may be interpreted as the ratio of variation explained by the model to total variation. The order of the splines and number of knots have negligible influence on $RV(L)$, as established by inumerous simulations in Borak et al. (2008). The implied volatility and RN surfaces are estimated with DSFM as in (3.5) with $L = 3$. Table 1 shows that the addition of the fourth or fifth basis function results in negligible model fit improvement.

L	1	2	3	4	5
$1 - RV(L)$	0.77	0.97	0.98	0.98	0.98

Table 1
Number of basis functions and explained variation

Figures 3 and 4 depict the estimated loading factors series $\{\widehat{Z}_t\}$ and basis functions $\widehat{m}_l$. From (3.5) we obtain a sequence of RN surfaces $\{\widehat{\mathcal{H}}_t\}$, $t = 1, \ldots, 253$, figure 5 shows $\widehat{\mathcal{H}}_t$ at t corresponding to day 20010710.
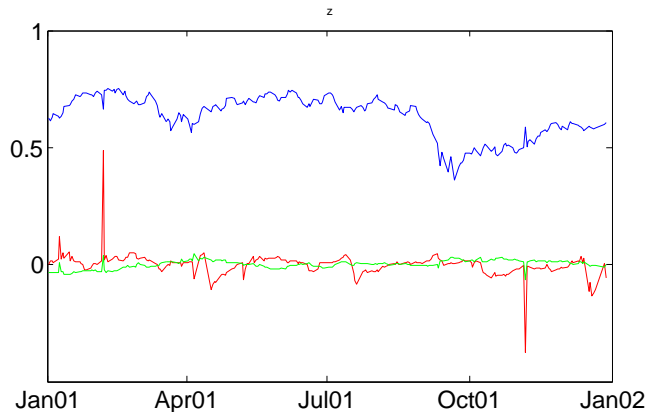
Fig. 3. Estimated $\{\widehat{Z}_{lt}\}$, $l = 1, 2, 3$ (top to bottom)

In a first step we investigate the covariance structure of $\{\widehat{Z}_t\}$ by means of VAR analysis. Table 2 presents the parameters from the VAR(2) model fitted on $\{\widehat{Z}_t\}$. The order 2 is selected based on Akaike (AIC), Schwarz (SC) and Hannan-Quinn (HQ) criteria, see table 3. Moreover the VAR(2) model satisfies stationarity as the roots of the characteristic polynomial lie inside of the unit circle.

A natural issue is to analyse the dependences between $\{Z_t\}$ and the shape of the RN surfaces $\{\widehat{\mathcal{H}}_t\}$. In order to investigate this relation we compute the skewness γ and excess kurtosis η of $\widehat{q}_{t,T}(\cdot|s_t)$ across t and maturities τ. Here
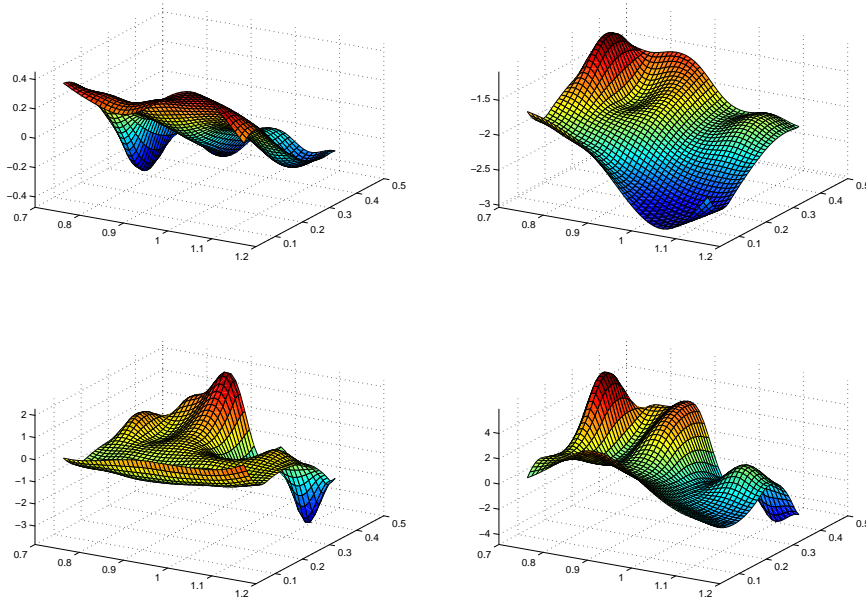
Fig. 4. Estimated basis functions $\widehat{m}_l$, $l = 0, \ldots, 3$ clockwise
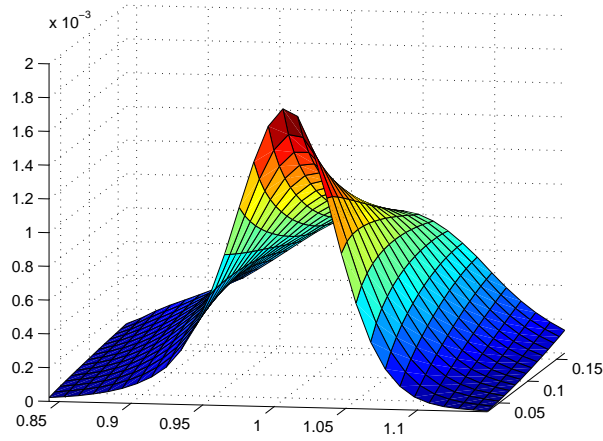
Fig. 5. Estimated RN surface, $\widehat{\mathcal{H}}_t$ at t corresponding to day 20010710

		VAR(2)					
	const	$\widehat{Z}_{1,t-1}$	$\widehat{Z}_{1,t-2}$	$\widehat{Z}_{2,t-1}$	$\widehat{Z}_{2,t-2}$	$\widehat{Z}_{3,t-1}$	$\widehat{Z}_{3,t-2}$
$\widehat{Z}_{1t}$	0.01	1.09	-0.16	0.10	-0.36	0.32	-0.23
$\widehat{Z}_{2t}$	0.01	-0.27	0.26	0.31	0.12	-1.14	0.33
$\widehat{Z}_{3t}$	0.01	-0.08	0.62	-0.05	-0.04	0.41	0.35

Table 2
Estimated parameters for the VAR(2) model on $\{\widehat{Z}_t\}$

order	AIC	SC	HQ
1	-11.03	-10.99	-11.01
2	-15.71	-15.54*	-15.64
3	-15.77*	-15.46	-15.64*
4	-15.76	-15.32	-15.58
5	-15.72	-15.16	-15.45

Table 3
Lag selection criteria for VAR models on $\{\widehat{Z}_t\}$
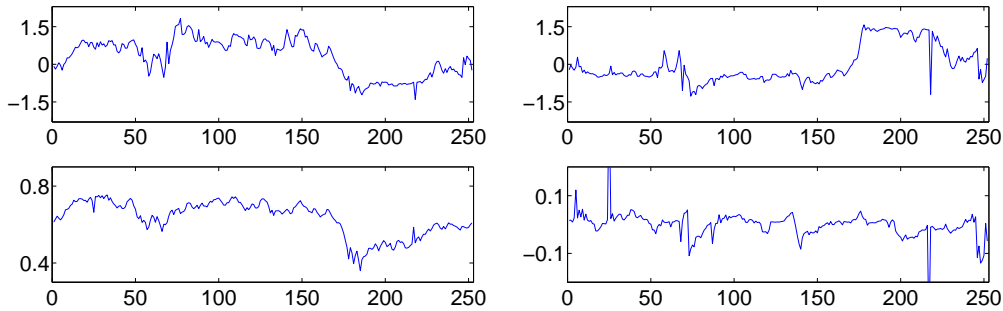
Fig. 6. Left: RN excess kurtosis $\{\eta_t\}$, $\tau = 18$ (top), $\{\widehat{Z}_{1t}\}$ (bottom) Right: RN skewness $\{\gamma_t\}$, $\tau = 18$ (top), $\{\widehat{Z}_{2t}\}$ (bottom)

$\widehat{q}_{t,T}(\cdot|s_t) = \widehat{\mathcal{H}}_t(\cdot, \tau)$. Figure 6 displays skewness $\{\gamma_t\}$ and excess kurtosis $\{\eta_t\}$ associated with $\widehat{q}_{t,T}$ for maturity $\tau = 18$ days together with $\{\widehat{Z}_{1t}\}$ and $\{\widehat{Z}_{3t}\}$ and motivates the investigation of their joint autocovariance structure.

The dynamic structure of the pairs $\{(\widehat{Z}_{1t}, \eta_t)\}$ and $\{(\widehat{Z}_{3t}, \gamma_t)\}$ for $\tau = 18$ is modelled by VAR(2) models. The choice of the VAR order is again based on AIC, SC, and HQ selection criteria. Portmanteau and LM tests on VAR residuals reject autocorrelations up to lag 12 and the roots of the characteristic polynomial lie inside of the unit circle.

	VAR(2)				
	const	$\widehat{Z}_{1,t-1}$	$\widehat{Z}_{1,t-2}$	η_{t-1}	η_{t-2}
$\widehat{Z}_{1t}$	0.04	0.86	0.08	0.01	0.00
η_t	-0.51	2.63	-1.75	0.67	0.19

Table 4
Estimated parameters for the VAR(2) model on $\{(\widehat{Z}_{1t}, \eta_t)\}$

Modelling the dynamics of risk neutral densities using DSFM allows to quantify the mechanisms governing risk perceptions from agents acting in a market. Insights are obtained in two directions, concerning the autocovariance

		VAR(2)			
	const	$\widehat{Z}_{3,t-1}$	$\widehat{Z}_{3,t-2}$	γ_{t-1}	γ_{t-2}
$\widehat{Z}_{3t}$	0.00	0.20	0.27	0.01	-0.02
γ_t	0.00	-1.69	0.68	0.81	0.24

Table 5
Estimated parameters for the VAR(2) model on $\{(\widehat{Z}_{3t}, \gamma_t)\}$

structure of $\{\widehat{Z}_t\}$, i.e. the time behaviour of the RN surfaces and their cross-correlation with the skewness and excess kurtosis from the estimated risk neutral densities, i.e. the relation between the dynamics and shape of the obtained RN surfaces. As seen in tables (4) and (5) the excess kurtosis and skewness from $\widehat{q}_{t,T}$ at maturity $\tau = 18$ are determined by the corresponding lagged values of $\widehat{Z}_t$.

Based on the presented methodology it is possible to investigate the dynamics of the risk neutral skewness and excess kurtosis based on statistical inference on $\{\widehat{Z}_t\}$. A natural further step is to perform econometric analysis to investigate the cointegration between the lower dimensional time series and macroeconomic and financial indicators. This could provide deeper insights into the relation between risk assessments from investors acting in a market and the flow of economic information at which they are exposed.

4 Acknowledgements

A Assumptions

The results from Theorems (2.1) and (2.2), see Borak et al. (2008) rely on the following assumptions:

(A1) The variables $X_{11},...,X_{JT}$, $\varepsilon_{11},...,\varepsilon_{JT}$ and $Z_1,\ldots,Z_T$ are independent. The process Z_t is allowed to be nonrandom.

(A2) For $t = 1,...,T$ the variables $X_{1t},...,X_{Jt}$ are identically distributed, have support $[0,1]^d$ and a density f_t that is bounded from below and above on $[0,1]^d$, uniformly over $t = 1,...,T$.

14

(A3) We assume that $E[\varepsilon_{jt}] = 0$ for $t = 1, ..., T$ and $j = 1, ..., J$ and

$$\sup_{t=1,...,T,j=1,...,J} E \exp[c\varepsilon_{jt}^2] < \infty$$

for $c > 0$ small enough.

(A4) The functions ψ_k may depend on the increasing indices T and J and are normed so that $\int_{[0,1]^d} \psi_k^2(x) \, dx = 1$ for $k = 1, \ldots, K$. Furthermore it holds that $\sup_{x \in [0,1]} ||\psi(x)|| = \mathcal{O}(K^{1/2})$

(A5) The components $m_0,...,m_L$ can be approximated by $\psi_1,...,\psi_K$, i.e.

$$\delta_K = \sup_{x \in [0,1]^d} \inf_{\Gamma \in \mathcal{G}} |m(x) - \Gamma\psi(x)| \to 0$$

for $l = 0, ..., L$ and $K \to \infty$. We denote by Γ^* the matrix that fulfills

$$\sup_{x \in [0,1]^d} |m(x) - \Gamma\psi(x)| \le 2\delta_K$$

(A6) There exist constants $0 < C_L < C_U < \infty$ such that all eigenvalues of the random matrix $T^{-1} \sum_{t=1}^T Z_t Z_t^\top$ lie in the interval $[C_L, C_U]$ with probability tending to one.

(A7) The minimization (2.2) runs over all values of (Γ, z) with

$$\sup_{x \in [0,1]} \max_{1 \le t \le T} ||Z_t^\top \Gamma\psi(x)|| \le M_T$$

where M_T fulfills $\max_{1 \le t \le T} ||Z_t|| \le M_T/C_m$ (with probability tending to one) for a constant $C_m > \sup_{x \in [0,1]} ||m(x)||$.

(A8) It holds that $\rho^2 = (K + T)M_T^2 \log(JTM_T)/(JT) \to 0$. The dimension L is fixed.

(A9) Z_t is a martingale difference with $E[Z_t|Z_1, \ldots, Z_{t_1}] = 0$ and for some $C > 0$ $E[||Z_t||^2|Z_1, \ldots, Z_{t_1}] < C$ (a.s). The matrix $E[Z_t Z_t^\top]$ has full rank. The process Z_t is independent of $X_{11}, ..., X_{TJ}$ and $\varepsilon_{11}, ..., \varepsilon_{TJ}$.

(A10) The functions $m_0, ..., m_L$ are linearly independent. In particular, no function is equal to 0.

(A11) It holds that $(K^{1/2}M_T + T^{1/4})(\rho + \delta_K) = \mathcal{o}(1)$.

References

Ait-Sahalia, Y., and Lo, A. (1998), "Nonparametric estimation of state-price densities implicit in financial asset prices," *Journal of Finance*, 53, 499–547.

Benko, M., Kneip, A., and Härdle, W. (2008), "Common Functional Principal Components," *Annals of Statistics*, forthcoming.

Besse, P., Cardot, H., and Ferraty, F. (1997), "Some Tools for Functional Data Analysis," *Computational Statistics and Data Analysis*, 24, 255–270.

Borak, S., Härdle, W., Mammen, E., and Park, B. (2008), "Time series modelling with semiparametric factor dynamics," *Journal of the American Statistical Association*, submitted.

Breeden, D., and Litzenberger, R. (1978), "Prices of state-contingent claims implicit in options prices," *Journal of Business*, 51, 621–651.

Brüggemann, R., Härdle, W., Mungo, J., and Trenkler, C. (2008), "VAR Modeling for Dynamic Loadings Driving Volatility Strings," *Journal of Financial Econometrics*, Advance Access.

Cont, R., and da Fonseca, J. (2002), "The Dynamics of Implied Volatility Surfaces," *Quantitative Finance*, 2, 45–60.

Dähne, S., Härdle, W., and Hautsch, H. (2008), "Quantification of Liquidity Costs: Forecasting the Ordebook," *Discussion Paper Series SFB 649, Humboldt-Universität zu Berlin*.

Dauxois, J., Pousse, A., and Romain, Y. (1982), "Asymptotic Theory for the Principal Component Analysis of a Vector Random Function: Some Applications to Statisticsl Inference," *Journal of Multivariate Analysis*, 12, 136–154.

Fengler, M., Härdle, W., and Mammen, E. (2007), "A semiparametric factor model for implied volatility surface dynamics," *Journal of Financial Econometrics*, 5, 189–218.

Ferraty, F., and Vieu, P. (2006), *Nonparametric Functional Data Analysis*, Springer, New York.

Gasser, T., and Kneip, A. (1995), "Searching for Structure in Curve Samples," *Journal of the American Statistical Association*, 90(432), 1179–1188.

Hernández-Hernández, D., and Schied, A. (2007), "A control approach to robust maximization with logarithmic utility and time-consistent penalties," *Stochastic Processes and Their Applications*, 117(8), 980–1000.

Kneip, A. (1994), "Nonparametric Estimation of Common Regressors for Similar Curve Data," *Annals of Statistics*, 22(3), 1386–1427.

Kneip, A., and Gasser, T. (1992), "Statistical Tools to Analyse Data Representing a Sample of Curves," *Annals of Statistics*, 20(3), 1266–1305.

Ramsay, J. O., and Dalzell, C. T. (1991), "Some Tools for Functional Data Analysis," *Journal of the Royal Statistical Society B*, 53(3), 539–572.

Ramsay, J. O., and Silverman, B. W. (2005), *Functional Data Analysis*, Springer, New York.

Rao, C. (1958), "Some Statistical Methods for Comparison of Growth Curves," *Biometrics*, 14, 434–471.

Rice, J., and Silverman, B. W. (1991), "Estimating the Mean and Covariance Structure Nonparametrically when the Data are Curves," *Journal of Royal Statistics Society B*, 53, 233–243.

Graphical Data Representation in Bankruptcy Analysis

W. K. Härdle[1], R. A. Moro[2], and D. Schäfer[3]

[1] C.A.S.E., Humboldt-Universität zu Berlin, Spandauer Str. 1, 10178 Berlin
   `stat@wiwi.hu-berlin.de`
[2] C.A.S.E., Humboldt-Universität zu Berlin, Spandauer Str. 1, 10178 Berlin and
   German Institute for Economic Research, Königin-Luise-Straße 5, 14195 Berlin
   `rmoro@diw.de`
[3] German Institute for Economic Research, Königin-Luise-Straße 5, 14195 Berlin
   `dschaefer@diw.de`

Graphical data representation is an important tool for model selection in bankruptcy analysis since the problem is highly non-linear and its numerical representation is much less transparent. In classical rating models a convenient representation of ratings in a closed form is possible reducing the need for graphical tools. In contrast to that non-linear non-parametric models achieving better accuracy often rely on visualisation. We demonstrate an application of visualisation techniques at different stages of corporate default analysis based on Support Vector Machines (SVM). These stages are the selection of variables (predictors), probability of default (PD) estimation and the representation of PDs for two and higher dimensional models with colour coding. It is at this stage when the selection of a proper colour scheme becomes essential for a correct visualisation of PDs. The mapping of scores into PDs is done as a non-parametric regression with monotonisation. The SVM learns a non-parametric score function that is, in its turn, non-parametrically transformed into PDs. Since PDs cannot be represented in a closed form, some other ways of displaying them must be found. Graphical tools give this possibility.

Keywords: company rating, default probability, support vector machines, colour coding

JEL classification: C14, G33, C45

1 Company Rating Methodology

Application of statistical techniques to corporate bankruptcy started in the 60's with the development of computers.[1] The first technique introduced was discriminant analysis (DA) for univariate [3] and multivariate models [1]. After DA the logit and probit models were introduced in [14] and [16]. Nowadays these models are widely used in practice, i.e. they are at the core of the rating solutions at most European central banks. The solution in the traditional framework is a linear function (a hyperplane in a multidimensional feature space) separating successful and failing companies. A company score is computed as a value of that function. In the case of the probit and logit models the score can be directly transformed into a probability of default (PD), which denotes the probability with which a company can go bankrupt within a certain period. The major disadvantages of these popular approaches is the linearity of the solution and, in the case of logit and probit models, the prespecified form of the link function between PDs and the linear combination of predictors (Figure 1).

In Figure 1 successful and failing companies are denoted with black triangles and white quadrangles respectively. There is an equal number of companies of both classes in the sample. Following the DA and logit classification rule, which give virtually the same result, we are more likely to find a failing company above and to the right from the straight line. This may lead to a conclusion that companies with significantly negative values of operating profit margin and equity ratio can be classified as successful. This, for example, allows for companies with liabilities much greater than total assets to be classified as successful. Such a situation is avoided by using a non-linear classification method, such as the SVM, which produces a non-linear boundary.

Following a traditional approach we would expect a monotonic relationship between predictors and PDs, like the falling relation for the interest coverage ratio (Figure 2). However, in reality this dependence is often non-monotonic as for such important indicators as the company size or net income change. In the latter case companies that grow too fast or too slow have a higher probability of default. That is the reason for contemplating non-linear techniques as alternatives. Two prominent examples are recursive partitioning [4] and neural networks [17]. Despite the strength of the two approaches they have visible drawbacks: orthogonal division of the data space in recursive partitioning that is usually not justified and heuristic model specification in neural networks.
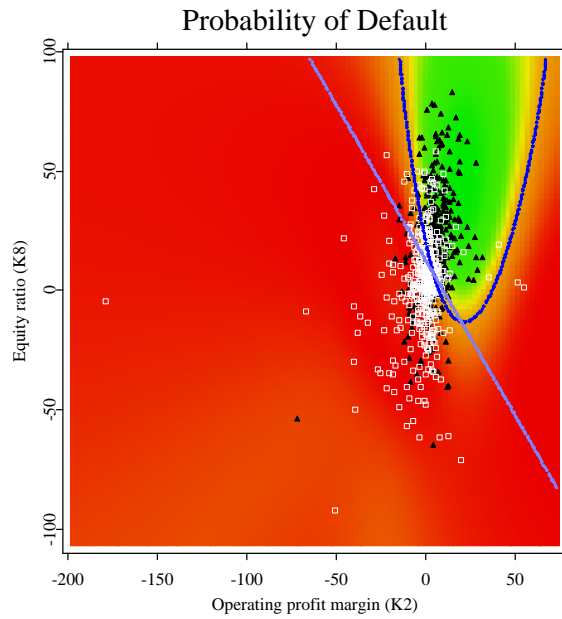
Fig. 1. A classification example. The boundary between the classes of solvent and insolvent companies was estimated using DA and logit regression (two indistinguishable linear boundaries) and an SVM (a non-linear boundary).
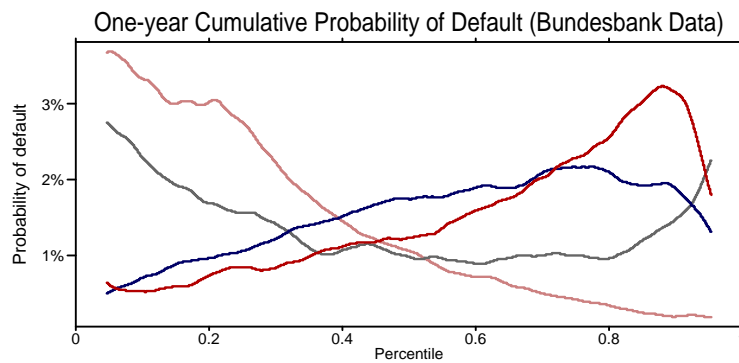
Fig. 2. One year cumulative PDs evaluated for several financial ratios on the German Bundesbank data. The ratios are net income change, K21 (gray), net interest ratio, K24 (red), interest coverage ratio, K29 (pink) and logarithm of total assets, K33 (blue). The k-nearest-neighbours procedure was used with the size of the window being around 8% of all observations. The total number of observations is 553500.

Recursive partitioning, also known as classification and regression trees (**CART**) performs classification by orthogonally dividing the data space. At each step only a division (split) along one of the axes is possible. The axis is chosen such, that a split along it reduces the variance in each of the subspaces and maximises the variance between them. Entropy based criteria can also be used. The visible drawback is the orthogonal division itself which imposes severe restrictions on the smoothness of the classifying function and may not adequately capture the correlation structure between the variables. Orthogonal division means that the separating hyperplane can only consist of orthogonal segments parallel to the coordinate grid, whereas the boundary between the classes has a smoothly changing gradient.

The **neural network** (NN) is a network of linnear classifiers (neurons) that are connected with one another in a prespecified way. The outputs of some of the neurons are inputs for others. The performance of a NN greatly depends on its structure that must be adapted for solving different problems. The network must be designed manually that requires a substantial experience from the operator. Moreover, NNs mostly do not povide a global solution but only a local one. This feature, as well as too much heuristics create many obstacles on the way of using NNs at the rating departments of banks.

We would like to have a model that is able to select a classifying function based on very general criteria. The SVM is a statistical technique that in many applications, such as optical character recognition and medical diagnostics, showed very good performance. It has a flexible solution and is controlled by adjusting only few parameters. Its overall good performance and flexibility make the SVM a suitable candidate [9].

Within a rating methodology each company is described by a set of variables x, such as financial ratios. Financial ratios, such as debt ratio (leverage) or interest coverage (earnings before interest and taxes to interest) characterise different sides of company operation. They are constructed on the basis of balance sheets and income statements. For example, the Bundesbank uses 32 ratios (predictors) computed using the company statements from its corporate bankruptcy data base. The predictors and basic statistics are given in Table 1. The whole Bundesbank data base covers the period 1987–2005 and consists of 553500 anonymised statements of solvent and insolvent companies. Most companies appear in the database several times in different years.

The class y of a company can be either $y = -1$ ('successful') or $y = 1$ ('bankrupt'). Initially, an unknown classifier function $f : x \rightarrow y$ is estimated on a training set of companies (x_i, y_i), $i = 1, ..., n$. The training set represents the data for companies which are known to have survived or gone bankrupt. In order to obtain PDs from the estimated scores f, rating practitioners usually rely on prespecified rating classes (i.e. BBB, C, AA, etc.). A certain range of scores and PDs belong to each rating class. The ranges are computed on the basis of historical data. To derive a PD for a newly scored company its score f is compared with the historical values of f's for each class. Basing on the

similarity of the scores a company is assigned to one particular class. The PD of this class becomes the PD of the company.
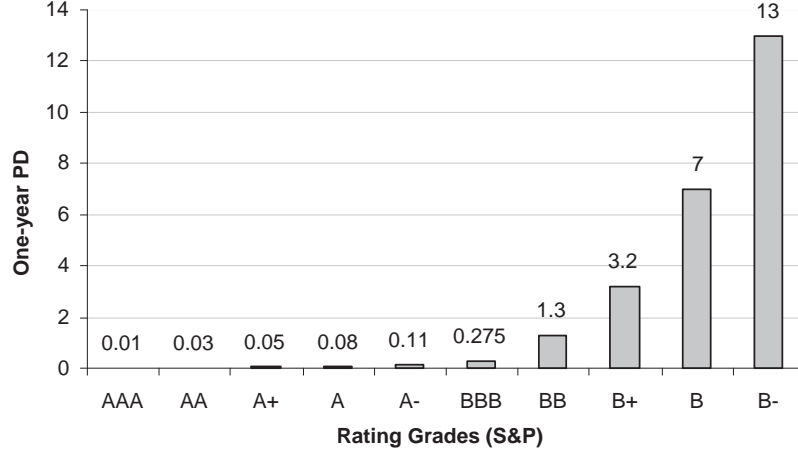
Fig. 3. One year probabilities of default for different rating grades [5].

Company bond ratings play an important role in determining the cost of debt refinancing since they reflect the probability of defaulting on the debt (Figure 3). One can notice that the differences between the classes in terms of PDs are not the same. For example, the PD increases by 6.7% or 24 times between classes BBB and B, but only by 0.07 or 8 times between classes AAA und A. The colours for coding PDs must be selected so that perceptually the classes would look equidistant, no matter what their absolute PD is. This can be achieved by using an appropriate colour scheme and colour distance scaling. The use of the HLS colour scheme in combination with a logarithmic colour scaling will be demonstarated in section 6.

2 The SVM Approach

The SVM [18] is a regression (and classification) technique that is based on margin maximisation (Figure 4) between two data classes. The margin is the distance between the hyperplanes bounding each class where in a linear perfectly separable case no observation may lie. The classifier function used by the linear SVM is a hyperplane symmetrically surrounded with a margin zone. It can be shown [9] that by maximising the margin one reduces the complexity of such a classifier. By applying kernel techniques the SVM can be extended to learn non-linear classifying functions (Figure 5).

In Figure 4 misclassifications are unavoidable when using linear classifying functions (linearly non-separable case). To account for misclassifications
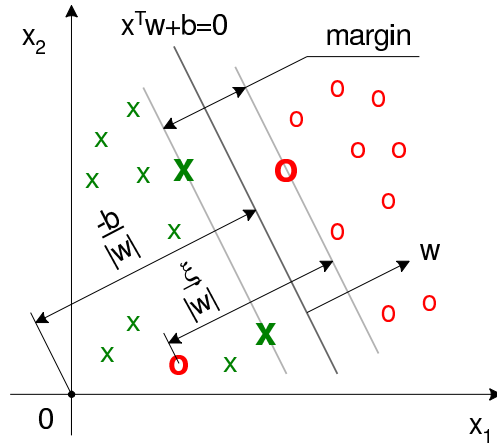
Fig. 4. The separating hyperplane $x^\top w + b = 0$ and the margin in a non-separable case. The observations marked with bold crosses and zeros are support vectors. The hyperplanes bounding the margin zone equidistant from the separating hyperplane are represented as $x^\top w + b = 1$ and $x^\top w + b = -1$.
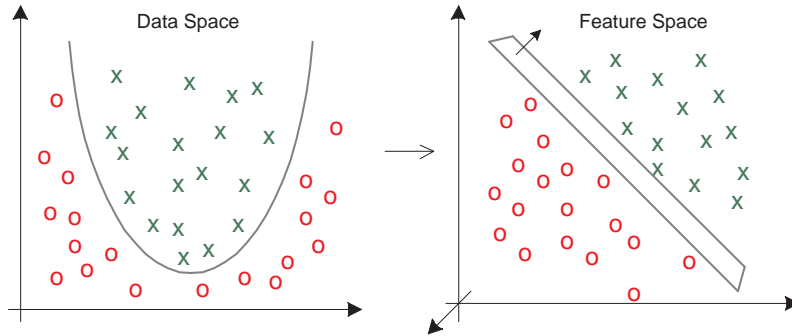
Fig. 5. Mapping from a two-dimensional data space into a three-dimensional space of features $\mathbb{R}^2 \mapsto \mathbb{R}^3$ using a quadratic kernel function $K(x_i, x_j) = (x_i^\top x_j)^2$. The three features correspond to the three components of a quadratic form: $\tilde{x}_1 = x_1^2$, $\tilde{x}_2 = \sqrt{2}x_1x_2$ and $\tilde{x}_3 = x_2^2$, thus, the transformation is $\Psi(x_1, x_2) = (x_1^2, \sqrt{2}x_1x_2, x_2^2)^\top$. The data separable in the data space with a quadratic function will be separable in the feature space with a linear function. A non-linear SVM in the data space is equivalent to a linear SVM in the feature space. The number of features will grow fast with d and the degree of the polynomial kernel p, which equals 2 in our example, making the closed-form representation of Ψ such as here practically impossible

the penalty ξ_i is introduced, which is related to the distance from the hyperplane bounding observations of the same class to observation i. $\xi_i > 0$ if a misclassification occurs. All observations satisfy the following two constraints:

$$y_i(x_i^\top w + b) \geq 1 - \xi_i, \tag{1}$$

$$\xi_i \geq 0. \tag{2}$$

With the normalisation of w, b and ξ_i as in (1) the margin equals to $2/\|w\|$. The convex objective function to be minimised given the constraints (1) and (2) is:

$$\frac{1}{2}\|w\|^2 + C\sum_{i=1}^{n}\xi_i. \tag{3}$$

The parameter C called capacity is related to the width of the margin zone. The smaller the C is, the bigger margins are possible. Using well established theory for optimisation of convex functions [6] we can derive the dual Lagrangian:

$$L_D = \frac{1}{2}w(\alpha)^\top w(\alpha) - \sum_{i=1}^{n}\alpha_i - \sum_{i=1}^{n}\delta_i\alpha_i + \sum_{i=1}^{n}\gamma_i(\alpha_i - C) - \beta\sum_{i=1}^{n}\alpha_i y_i \tag{4}$$

for the dual problem:

$$\min_{\alpha_i,\delta_i,\gamma_i,\beta}\max_{w_k,b,\xi_i} L_D, \tag{5}$$

Here for a linear SVM:

$$w(\alpha)^\top w(\alpha) = \sum_{i=1}^{n}\sum_{j=1}^{n}\alpha_i\alpha_j y_i y_j x_i^\top x_j. \tag{6}$$

For obtaining non-linear classifying functions in the data space a more general form is applicable:

$$w(\alpha)^\top w(\alpha) = \sum_{i=1}^{n}\sum_{j=1}^{n}\alpha_i\alpha_j y_i y_j K(x_i, x_j). \tag{7}$$

The function $K(x_i, x_j)$ is called a kernel function. Since it has a closed form representation, the kernel is a convenient way of mapping low dimensional data into a highly dimensional (often infinitely dimensional) space of features. It must satisfy the Mercer conditions [15], i.e. be symmetric and semipositive definite or, in other words, represent a scalar product in some Hilbert space [19].

In our study we applied an SVM with an anisotropic Gaussian kernel

$$K(x_i, x_j) = \exp\left\{-(x_i - x_j)^\top r^{-2}\Sigma^{-1}(x_i - x_j)/2\right\}, \tag{8}$$

where r is a coefficient and Σ is a variance-covariance matrix. The coefficient r is related to the complexity of classifying functions: the hgher the r is, the lower is the complexity. If kernel functions allow for sufficiently rich feature spaces, the performances of SVMs are comparable in terms of out-of-sample forecasting accuracy [18].

3 Company Score Evaluation

The company score is computed as:

$$f(x) = x^\top w + b, \tag{9}$$

where $w = \sum_{i=1}^{n} \alpha_i y_i x_i$ and $b = \frac{1}{2}(x_+ + x_-)^\top w$; x_+ and x_- are the observations from the opposite classes for which constraint (1) becomes equality. By substituting the scalar product with a kernel function we will derive a non-linear score function:

$$f(x) = \sum_{i=1}^{n} K(x_i, x)\alpha_i y_i + b. \tag{10}$$

The non-parametric score function (10) does not have a compact closed form representation. This necessitates the use of graphical tools for its visualisation.

4 Variable Selection

In this section we describe the procedure and the graphical tools for selecting the variables of the SVM model used in forecasts. We have two most important criteria of model accuracy: the accuracy ratio (AR), which will be used here as a criterion for model selection, (Figure 6) and the percentage of correctly classified out-of-sample observations. Higher values indicate better model accuracy.
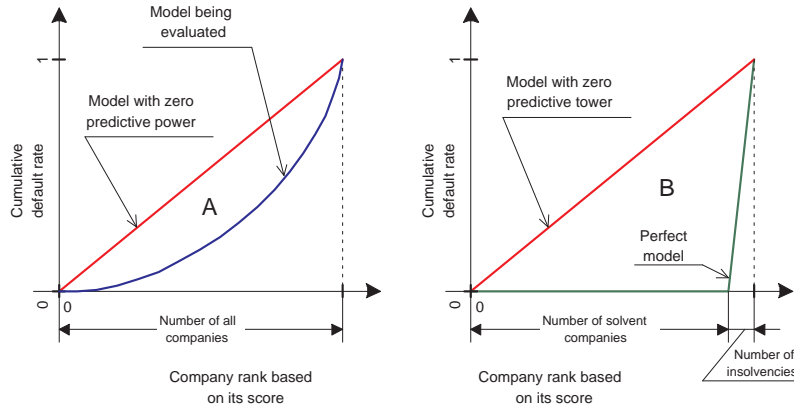
Fig. 6. The power curves for a perfect (green), random (red) and some real (blue) classification models. The AR is the ratio of two areas A/B. It lies between 0 for a random model with no predictive power and 1 for a perfect model.

We start model selection from the simplest, i.e. univariate models and then pick up the one with the highest AR. The problem that arises is how to determine the variable which provides the highest AR across possible data samples. For a parametric model we would need to estimate the distribution of the coefficients at the variables and, hence, their confidence intervals. This approach, however, is practically irrelevant for non-parametric models.

Instead we can compare goodness of models with respect to some accuracy measure, in our case AR. Firstly we will estimate the distributions of AR for different models. This can be done using bootstrapping [12]. We randomly select training and validation sets as subsamples of 500 solvent and 500 insolvent companies each. We used the 50/50 ratio since this is the worst case with the minimum AR. The two sets are not overlapping, i.e. do not contain common observations. For each of these sets we apply the SVM with parameters that provide the highest AR for bivariate models (Figure 7) and estimate ARs. Then we perform a Monte Carlo experiment: repeat the generation of subsamples and computing of ARs 100 times. Each time we will record the ARs and then estimate their distribution.

At the end of this procedure we obtain an empirically estimated distribution of AR on bootstrapped subsamples. The median AR provides a robust measure to compare different variables as predictors. The same approach can be used for comparing SVM with DA and logit regression in terms of predictive power. We compute AR for the same subsamples with the SVM, DA and logit models. The median improvements in AR for the SVM over DA and the SVM over the logistic regression are also reported below.

We will start this procedure with all univariate models with 33 variables K1-K9, K11-K33 as they are denoted at the Bundesbank and variable K10, which is a standard normal random variable used as a reference (Table 1). For each model the resulting distribution of ARs will be represented as box plots (Figure 8). The red line depicts medians. The box within each box plot shows the interquartile range (IQR), while the whiskers span to the distance of 3/2 IQR in each direction from the median. Outliers beyond that range are denoted with circles.

Basing on Figure 8 we can conclude that variables K5 (Debt Cover) and K29 (Interest Coverage Ratio) provide the highest median AR around 50%. We can also notice that variables K12, K26 and K28 have a very low accuracy: their median ARs do not exceed 11.5%. The model based on random variable K10 has AR equal zero, in other words, it has no predictive power whatsoever. For the next step we will select variable K5 that was included in the best univariate model.

For bivariate models we will select the best predictor from the univariate models (K5) and one of the rest that delivers the highest AR (K29) (Figure 9). This procedure will be repeated for each new variable added. The AR is growing until the model has eight variables, then it slowly declines. Median ARs for the models with eight variables are shown in Figure 10.
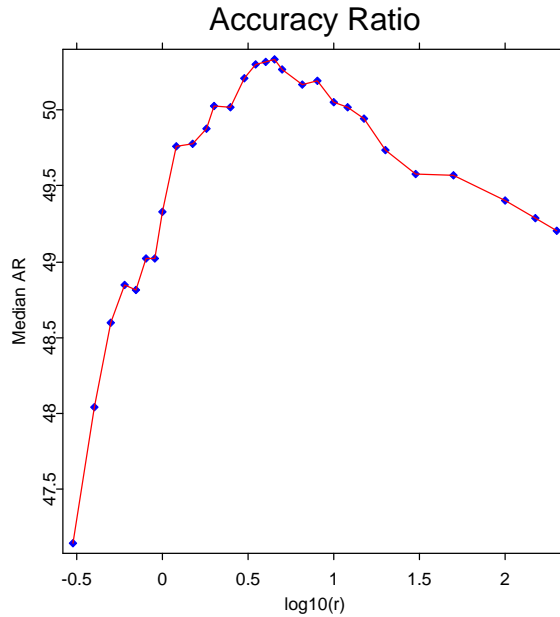
Fig. 7. The relationship between an accuracy measure (AR) and the coefficient r in the SVM formulation. Higher r's correspond to less complex models. The median ARs were estimated on 100 bootstrapped subsamples of 500 solvent and 500 insolvent companies both in the training and validation sets. A bivariate SVM with the variables K5 and K29 was used. We will be using $r = 4$ in all SVMs used in this chapter.

We have also conducted experiments with subsamples of the size of 5000 observations. The change of median was extremely small (one–two orders of magnitude smaller than the interquartile range). The interquartile range got narrower as it was expected, i.e. the difference between models with bigger samples is only more statistically significant. Thus, proving that if the difference is significant on a sample of 1000 observations, it can be guaranteed that this will remain so for bigger samples.

The SVM based on variables K5, K29, K7, K33, K18, K21, K24, K33 and K9 attains the highest median AR of around 60.0%. For comparison we plot an improvement in AR for the SVM vs. DA and logit regression on the same 100 subsamples. The data used in the DA and logit models were processed as following: if $x_i < q_{0.05}(x_i)$ then $x_i = q_{0.05}(x)$ and if $x_i > q_{0.95}(x_i)$ then $x_i = q_{0.95}(x_i)$; $i = 1, 2, \ldots, 8$; $q_\alpha(x_i)$ is an α quantile of x_i. Thus, the DA and logit regression applied were *robust versions* not sensitive to outliers. Without such a procedure the improvement would be much higher.

Table 1. Summary Statistics. q_α is an α quantile. IQR is the interquartile range.

Var.	Name	Group	$q_{0.01}$	Median	$q_{0.99}$	IQR
K1	Pre-tax profit margin	Profitability	-26.9	2.3	78.5	5.9
K2	Operating profit margin	Profitability	-24.6	3.8	64.8	6.3
K3	Cash flow ratio	Liquidity	-22.6	5.0	120.7	9.4
K4	Capital recovery ratio	Liquidity	-24.4	11.0	85.1	17.1
K5	Debt cover	Liquidity	-42.0	17.1	507.8	34.8
K6	Days receivable	Activity	0.0	31.1	184.0	32.7
K7	Days payable	Activity	0.0	23.2	248.2	33.2
K8	Equity ratio	Financing	0.3	14.2	82.0	21.4
K9	Equity ratio (adj.)	Financing	0.5	19.3	86.0	26.2
K10	Random Variable	Test	-2.3	0.0	2.3	1.4
K11	Net income ratio	Profitability	-29.2	2.3	76.5	5.9
K12	Leverage ratio	Leverage	0.0	0.0	164.3	4.1
K13	Debt ratio	Liquidity	-54.8	1.0	80.5	21.6
K14	Liquidity ratio	Liquidity	0.0	2.0	47.9	7.1
K15	Liquidity 1	Liquidity	0.0	3.8	184.4	14.8
K16	Liquidity 2	Liquidity	2.7	63.5	503.2	58.3
K17	Liquidity 3	Liquidity	8.4	116.9	696.2	60.8
K18	Short term debt ratio	Financing	2.4	47.8	95.3	38.4
K19	Inventories ratio	Investment	0.0	28.0	83.3	34.3
K20	Fixed assets ownership r.	Leverage	1.1	60.6	3750.0	110.3
K21	Net income change	Growth	-50.6	3.9	165.6	20.1
K22	Own funds yield	Profitability	-510.5	32.7	1998.5	81.9
K23	Capital yield	Profitability	-16.7	8.4	63.1	11.0
K24	Net interest ratio	Cost struct.	-3.7	1.1	36.0	1.9
K25	Own funds/pension prov. r.	Financing	0.4	17.6	84.0	25.4
K26	Tangible asset growth	Growth	0.0	24.2	108.5	32.6
K27	Own funds/provisions ratio	Financing	1.7	24.7	89.6	30.0
K28	Tangible asset retirement	Growth	1.0	21.8	77.8	18.1
K29	Interest coverage ratio	Cost struct.	-1338.6	159.0	34350.0	563.2
K30	Cash flow ratio	Liquidity	-14.1	5.2	116.4	8.9
K31	Days of inventories	Activity	0.0	42.9	342.0	55.8
K32	Current liabilities ratio	Financing	0.3	58.4	98.5	48.4
K33	Log of total assets	Other	4.9	7.9	13.0	2.1

Figure 11 represents the absolute improvement for SVM over robust DA (upper line) and SVM over robust logit regression (lower line). We can see that for all models containing variables K5, K29, K7, K33, K18, K21, K24 and one of the remaining variables the median AR was always higher for the SVM. Thus, the SVM model is always dominating in accuracy DA and logit regression with regard to AR.
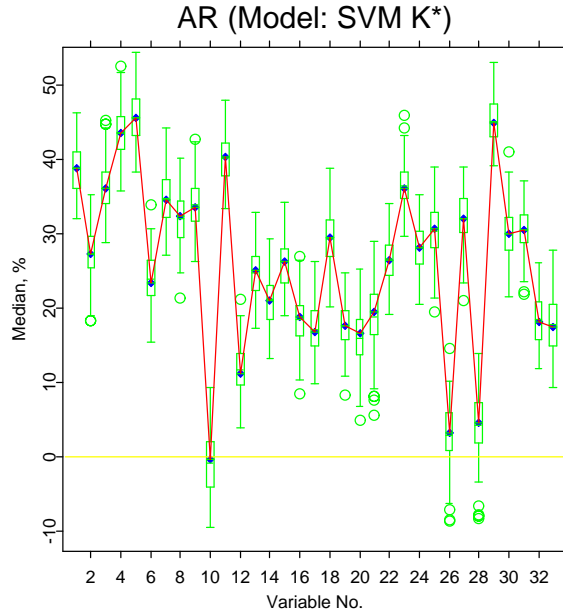
Fig. 8. Accuracy ratios for univariate SVM models. Box-plots are estimated basing on 100 random subsamples. The AR for the model containing only random variable K10 is zero.

5 Conversion of Scores into PDs

There is another way to look at the company score. It defines the distance between companies in terms of the distance to the boundary between the classes. The lower is the score, the farther is a company from the class of bankrupt companies, therefore, we can assume, the lower PD it must have. This means that the dependence between scores and PDs is assumed to be monotonous. This is the only kind of dependence that was assumed in all rating models mentioned in this chapter and the only one we use for PD calibration.

The conversion procedure consists of the estimation of PDs for the observation of the training set with a subsequent monotonisation (step one and two) and the computation of a PD for some new company (step three).

Step one is the estimation of PDs for the companies of the training set. We used kernel techniques to preliminary evaluate PDs for observation i from the training set, $i = 1, 2, \ldots, n$:

$$\widetilde{PD}(x_i) = \frac{\sum_{j=1}^{n} K_h(x_i, x_j) I_{\{y_j=1\}}}{\sum_{j=1}^{n} K_h(x_i, x_j)} \tag{11}$$
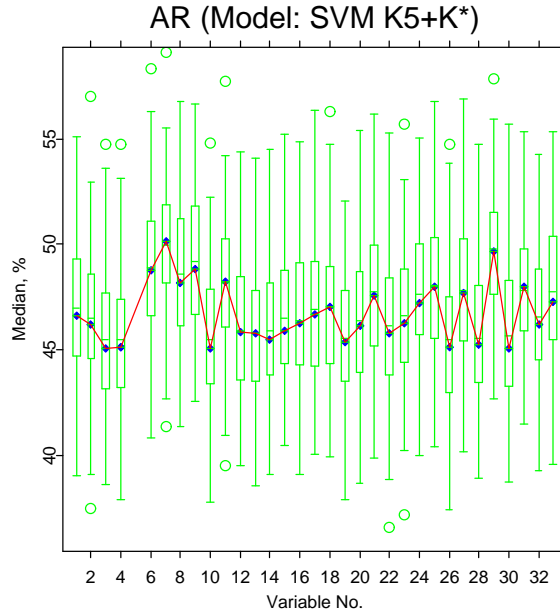
Fig. 9. Accuracy ratios for bivariate SVM models. Each model includes variable K5 and one of the remaining. Box-plots are estimated basing on 100 random subsamples.

Here a k-nearest-neighbour Gaussian kernel was used. h is the kernel bandwidth.

The preliminary PDs evaluated in this way are not necessarily a monotonical function of the score. The monotonisation of $\widetilde{PD}_i$, $i = 1, 2, \ldots, n$ is achieved at step two using the Pool Adjacent Violator (PAV) algorithm ([2] and [13]). As a result we obtain monotonised probabilities of default $PD(x_i)$ for the observations of the training set.

Finally, at step three the PDs are computed for any observation described with x as an interpolation between two PDs of the neighbouring, in terms of the score, observations from the training set, x_i and x_{i-1}, $i = 2, 3, \ldots, n$:

$$PD(x) = PD(x_i) + \frac{f(x) - f(x_{i-1})}{f(x_i) - f(x_{i-1})} \left\{ PD(x_i) - PD(x_{i-1}) \right\}. \quad (12)$$

If the score for an observation x lies beyond the range of scores for the training set, then $PD(x)$ equals to the score of the first neighbouring obseration of the training set.

Figure 12 is an example of the cumulative PD curve (power curve) and estimated PDs for a subsample of 200 companies. The PD curve has a plateau area for the observations with a high score. Default probabilities can change from 15% to 80% depending on the score.

AR (SVM K5+K29+K7+K33+K18+K21+K24+K*)
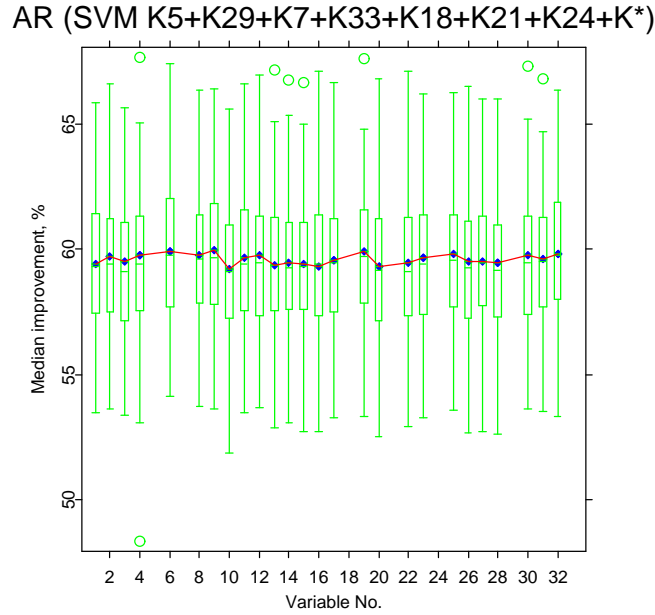
Fig. 10. Accuracy ratios for SVM models with eight variables. Each model includes variables K5, K29, K7, K33, K18, K21, K24 and one of the remaining. Box-plots are estimated basing on 100 random subsamples.

6 Colour Coding of PDs

The RGB colour space is based on three primary colours, red, green and blue, that are mixed to produce other ones. It is the colour coding scheme that is used in monitors and TV. It is, however, inconvenient for colour coding since we would like to make adjustments only to the channel responsible for colour while keeping lightness and saturation constant. This can be achieved with the HLS colour space.

We will represent the probability of default (PD) estimated with the SVM as two-dimensional plots where each colour represents a specific PD. The PD is a number that can be represented on a gray scale, e.g. in the RGB encoding as (i, i, i) with i changing from 0 to 255, e.g. the colour R=255, G=0, B=0 corresponds to red, and R=255, G=0, B=255 to violet, etc..

The HLS stands for hue, lightness (or luminance) and saturation. By adjusting only hues and keeping luminance and saturation fixed we can generate simulated colours from the range shown in Figure 14. The pure red colour corresponds to H=0 or 360, the pure green to H=120 and the pure blue to H=240.
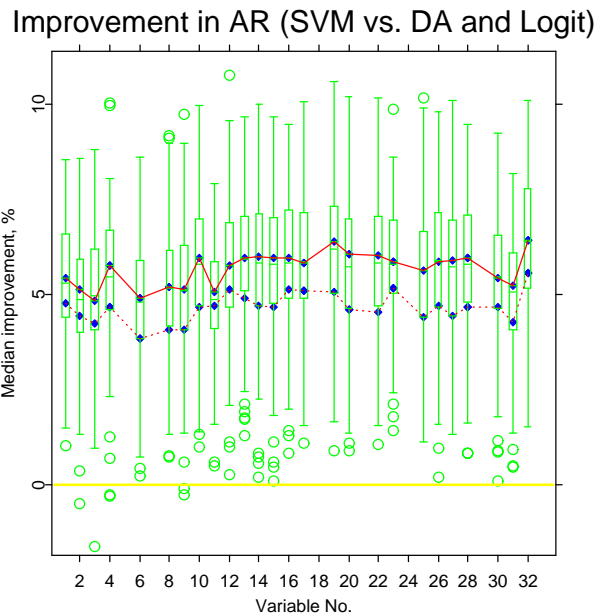
Fig. 11. Median improvement in AR. SVM vs. DA (the upper line) and SVM vs. logit regression (the lower line). Box-plots are estimated basing on 100 random subsamples for the case of DA. Each model includes variables K5, K29, K7, K33, K18, K21, K24 and one of the remaining

Red colour is often used in finance to highlight negative information, while green and blue are used to convey positive information. Therefore, we would like to code PDs with colours ranging from red for the highest PD to blue-green for the most solvent company. For this end we normalise PDs so, that the lowest PD corresponds to the hue equal 180 (green-blue) or 120 (green) while the highest PD corresponds to the hue equal 0 (red). The resulting graph that shows the data and PDs in the dimensions of variables K33 and K29 is shown in Figures 15–17. The three figures correspond consequently to three SVMs with high, average and high complexity. The saturation was fixed at 0.85 to make colours look more noble and the luminance was fixed at 0.46, the maximum possible value for the chosen saturation. The HLS colours obtained in this way were transformed into RGB ones and plotted by the XploRe ([8], [7]) as a contour plot. The outliers that lie beyond the 5% and 95% quantiles are plotted at the rand.

To produce the plot a grid was generated with 101 steps both in horizontal and vertical directions. For each point of the grid a PD was estimated and represented in the HLS colour encoding. Then the HLS colour was converted
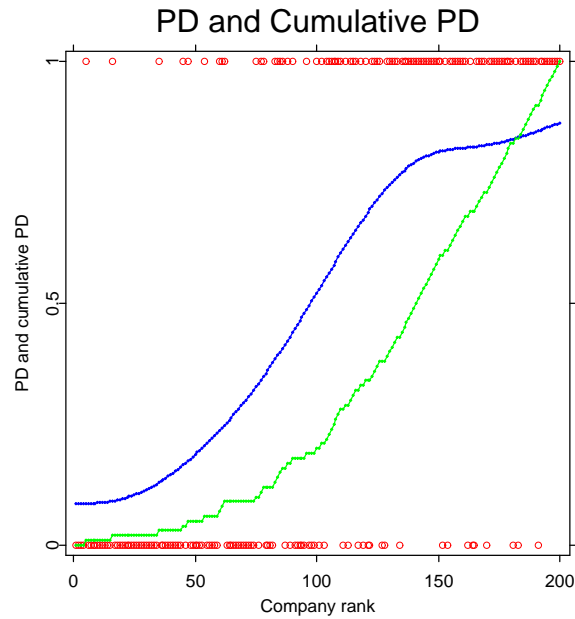
Fig. 12. PD (blue line) and cumulative PD (green line) estimated with the SVM for a subsample of 200 observation from the Bundesbank data. The variables were included into the model that achieved the highest AR: K5, K29, K7, K33, K18, K21, K24 and K9. The higher is the score, the higher is the rank of a company.

into an RGB colour and plotted by the XploRe as a small filled quadrangle. The quadranges evenly cover the whole area giving a continuous PD representation. The contour lines can be also added to the graph as illustrated by Figure 18.

7 Conclusion

In this chapter we demonstrated the application of graphical tools for variable selection, data visualisation and financial information representation and discussed such essential aspects of graphical analysis as colour coding. We believe that graphical analysis will have an increased importance as non-parametric models, such as SVM, are becoming more and more popular. On the other hand graphical representation can facilitate the acceptance of non-parametric models in various areas, e.g. finance, medicine, sound and image processing, etc. This will contribute to the development of those areas since non-linear non-parametric models better represent reality and provide higher forecasting accuracy.
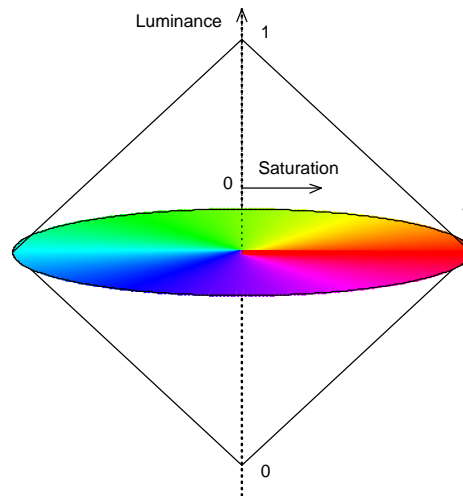
Fig. 13. The luminance and saturation dimensions of the HLS colour space. We will keep luminance and saturation constant and encode the information about PDs with hue.

References

1. Altman, E.: 1968, 'Financial Ratios, Discriminant Analysis and the Prediction of Corporate Bankruptcy'. *The Journal of Finance* pp. 589–609.
2. Barlow, R. E., J. M. Bartholomew, J. M. Bremmer, and H. D. Brunk: 1972, *Statistical Inference Under Order Restrictions*. John Wiley & Sons, New York, NY.
3. Beaver, W.: 1966, 'Financial Ratios as Predictors of Failures. Empirical Research in Accounting: Selected Studies'. *Journal of Accounting Research* pp. 71–111. supplement to vol. 5.
4. Frydman, H., E. Altman, and D.-L. Kao: 1985, 'Introducing Recursive Partitioning for Financial Classification: The Case of Financial Distress'. *The Journal of Finance* **40**(1), 269–291.
5. Füser, K.: 2002, 'Basel II – was muß der Mittelstand tun?'. *http://www.ey.com/global/download.nsf/Germany/Mittelstandsrating/ $file/Mittelstandsrating.pdf.*
6. Gale, D., H. W. Kuhn, and A. W. Tucker: 1951, *Linear Programming and the Theory of Games, in Activity Analysis of Production and Allocation, T. C. Koopmans (ed.)*. John Wiley & Sons, New York, NY.
7. Härdle, W., Z. Hlávka, and S. Klinke: 2000a, *XploRe Application Guide*. Springer Verlag, Berlin.
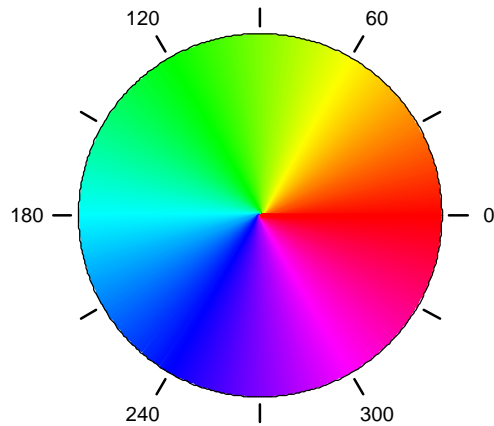
Fig. 14. The hue dimension of the HLS colour space.

8. Härdle, W., S. Klinke, and M. Müller: 2000b, *XploRe Learning Guide.* Springer Verlag, Berlin.
9. Härdle, W., R. A. Moro, and D. Schäfer: 2004a, *Predicting Bankruptcy with Support Vector Machines, W. Härdle (ed.).* Springer Verlag, Berlin.
10. Härdle, W., M. Müller, S. Sperlich, and A. Werwatz: 2004b, *Nonparametric and Semiparametric Models.* Springer Verlag, Berlin.
11. Härdle, W. and L. Simar: 2003, *Applied Multivariate Statistical Analysis.* Springer Verlag.
12. Horowitz, J. L.: 2001, *The Bootstrap, J. J. Heckman and E. E. Leamer (eds.)*, Vol. 5. Elsevier Science B.V.
13. Mammen, E.: 1991, 'Estimating a Smooth Monotone Regression Function'. *Annals of Statistics* **19**, 724–740.
14. Martin, D.: 1977, 'Early Warning of Bank Failure: A Logit Regression Approach'. *Journal of Banking and Finance* (1), 249–276.
15. Mercer, J.: 1909, 'Functions of Positive and Negative Type and Their Connection with the Theory of Integral Equations'. *Philosophical Transactions of the Royal Society of London* **209**, 415–446.
16. Ohlson, J.: 1980, 'Financial Ratios and the Probabilistic Prediction of Bankruptcy'. *Journal of Accounting Research* pp. 109–131.
17. Tam, K. and M. Kiang: 1992, 'Managerial Application of Neural Networks: the Case of Bank Failure Prediction'. *Management Science* **38**(7), 926–947.
18. Vapnik, V. N.: 1995, *The Nature of Statistical Learning Theory.* Springer, New York.
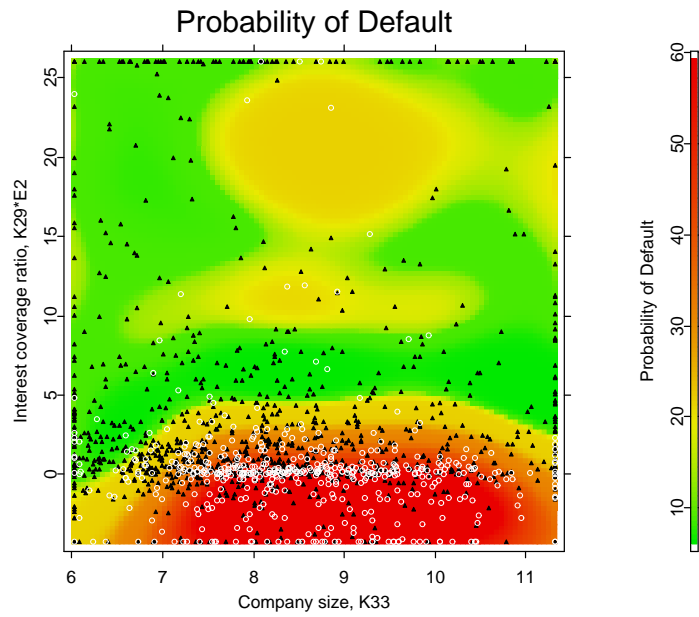
Fig. 15. Probability of default estimated for a random subsample of 500 failing and 500 surviving companies plotted for the variables K33 and K29. An SVM of high complexity with the radial basis kernel $0.5\Sigma^{1/2}$ was used.

19. Weyl, H.: 1928, *Gruppentheorie und Quantenmechanik*. Hirzel, Leipzig.
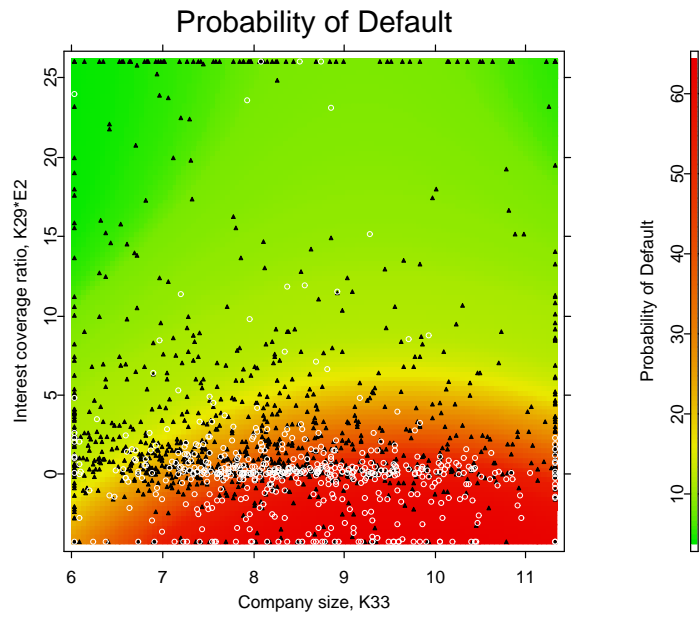
Fig. 16. Probability of default estimated for a random subsample of 500 failing and 500 surviving companies plotted for the variables K33 and K29. An SVM of average complexity with the radial basis kernel $4\Sigma^{1/2}$ was used.
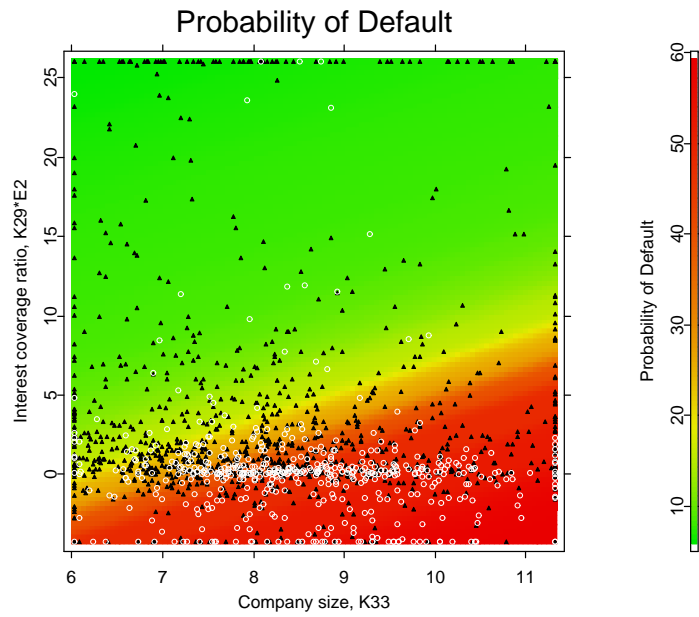
Fig. 17. Probability of default estimated for a random subsample of 500 failing and 500 surviving companies plotted for the variables K33 and K29. An SVM of low complexity with the radial basis kernel $100\Sigma^{1/2}$ was used.
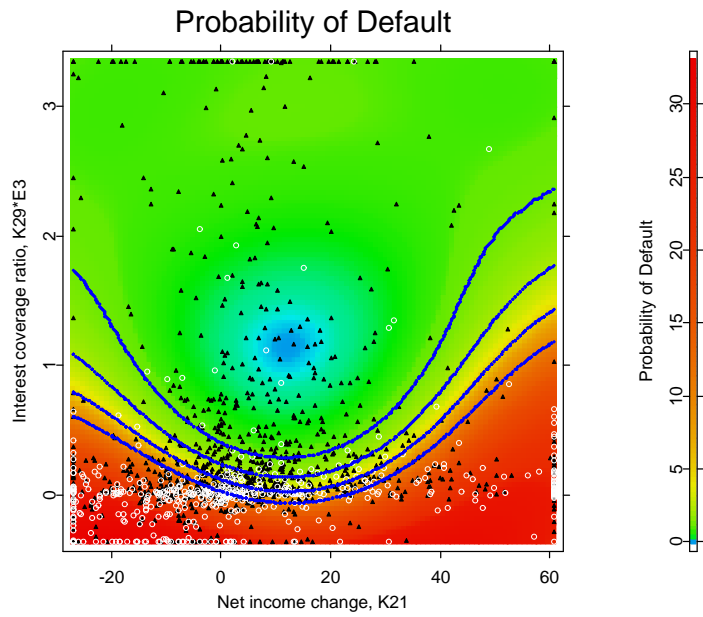
Fig. 18. Probability of default plotted for the variables K21 and K29. The boundaries of five risk classes are shown in blue, which correspond to the rating classes: BBB and above (investment grade), BB, B+, B, B- and lower.

Introduction

<div style="text-align: right">**I.1**</div>

Antony Unwin, Chun-houh Chen, Wolfgang K. Härdle

Computational Statistics
and Data Visualization

This book is the third volume of the Handbook of Computational Statistics and covers the field of data visualization. In line with the companion volumes, it contains a collection of chapters by experts in the field to present readers with an up-to-date and comprehensive overview of the state of the art. Data visualization is an active area of application and research, and this is a good time to gather together a summary of current knowledge.

Graphic displays are often very effective at communicating information. They are also very often not effective at communicating information. Two important reasons for this state of affairs are that graphics can be produced with a few clicks of the mouse without any thought and the design of graphics is not taken seriously in many scientific textbooks. Some people seem to think that preparing good graphics is just a matter of common sense (in which case their common sense cannot be in good shape), while others believe that preparing graphics is a low-level task, not appropriate for scientific attention. This volume of the Handbook of Computational Statistics takes graphics for data visualization seriously.

Data Visualization and Theory

Graphics provide an excellent approach for exploring data and are essential for presenting results. Although graphics have been used extensively in statistics for a long time, there is not a substantive body of theory about the topic. Quite a lot of attention has been paid to graphics for presentation, particularly since the superb books of Edward Tufte. However, this knowledge is expressed in principles to be followed and not in formal theories. Bertin's work from the 1960s is often cited but has not been developed further. This is a curious state of affairs. Graphics are used a great deal in many different fields, and one might expect more progress to have been made along theoretical lines.

Sometimes in science the theoretical literature for a subject is considerable while there is little applied literature to be found. The literature on data visualization is very much the opposite. Examples abound in almost every issue of every scientific journal concerned with quantitative analysis. There are occasionally articles published in a more theoretical vein about specific graphical forms, but little else. Although there is a respected statistics journal called the Journal of Computational and Graphical Statistics, most of the papers submitted there are in computational statistics. Perhaps this is because it is easier to publish a study of a technical computational problem than it is to publish work on improving a graphic display.

Presentation and Exploratory Graphics

The differences between graphics for presentation and graphics for exploration lie in both form and practice. Presentation graphics are generally static, and a single
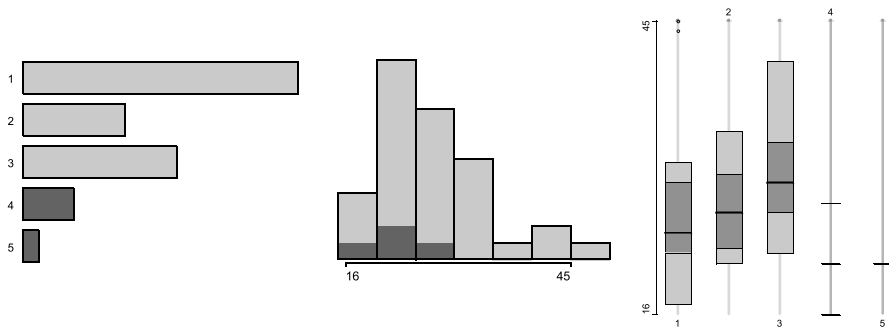
Figure 1.1. A barchart of the number of authors per paper, a histogram of the number of pages per paper, and parallel boxplots of length by number of authors. Papers with more than three authors have been selected

graphic is drawn to summarize the information to be presented. These displays should be of high quality and include complete definitions and explanations of the variables shown and of the form of the graphic. Presentation graphics are like proofs of mathematical theorems; they may give no hint as to how a result was reached, but they should offer convincing support for its conclusion. Exploratory graphics, on the other hand, are used for looking for results. Very many of them may be used, and they should be fast and informative rather than slow and precise. They are not intended for presentation, so that detailed legends and captions are unnecessary. One presentation graphic will be drawn for viewing by potentially thousands of readers while thousands of exploratory graphics may be drawn to support the data investigations of one analyst.

Books on visualization should make use of graphics. Figure 1.1 shows some simple summaries of data about the chapters in this volume, revealing that over half the chapters had more than one author and that more authors does not always mean longer papers.

Graphics and Computing 1.1.3

Developments in computing power have been of great benefit to graphics in recent years. It has become possible to draw precise, complex displays with great ease and to print them with impressive quality at high resolution. That was not always the case, and initially computers were more a disadvantage for graphics. Computing screens and printers could at best produce clumsy line-driven displays of low resolution without colour. These offered no competition to careful, hand-drawn displays. Furthermore, even early computers made many calculations much easier than before and allowed fitting of more complicated models. This directed attention away from graphics, and it is only in the last 20 years that graphics have come into their own again.

These comments relate to presentation graphics, that is, graphics drawn for the purpose of illustrating and explaining results. Computing advances have benefitted exploratory graphics, that is, graphics drawn to support exploring data, far more. Not just the quality of graphic representation has improved but also the quantity. It is now trivial to draw many different displays of the same data or to riffle through many different versions interactively to look for information in data. These capabilities are only gradually becoming appreciated and capitalized on.

The importance of software availability and popularity in determining what analyses are carried out and how they are presented will be an interesting research topic for future historians of science. In the business world, no one seems to be able to do without the spreadsheet Excel. If Excel does not offer a particular graphic form, then that form will not be used. (In fact Excel offers many graphic forms, though not all that a statistician would want.) Many scientists, who only rarely need access to computational power, also rely on Excel and its options. In the world of statistics itself, the packages SAS and SPSS were long dominant. In the last 15 years, first S and S-plus and now R have emerged as important competitors. None of these packages currently provide effective interactive tools for exploratory graphics, though they are all moving slowly in that direction as well as extending the range and flexibility of the presentation graphics they offer.

Data visualization is a new term. It expresses the idea that it involves more than just representing data in a graphical form (instead of using a table). The information behind the data should also be revealed in a good display; the graphic should aid readers or viewers in seeing the structure in the data. The term data visualization is related to the new field of information visualization. This includes visualization of all kinds of information, not just of data, and is closely associated with research by computer scientists. Up till now the work in this area has tended to concentrate just on presenting information, rather than on what may be deduced from it. Statisticians tend to be concerned more with variability and to emphasize the statistical properties of results. The closer linking of graphics with statistical modelling can make this more explicit and is a promising research direction that is facilitated by the flexible nature of current computing software. Statisticians have an important role to play here.

1.2   The Chapters

Needless to say, each Handbook chapter uses a lot of graphic displays. Figure 1.2 is a scatterplot of the number of figures against the number of pages. There is an approximate linear relationship with a couple of papers having somewhat more figures per page and one somewhat less. The scales have been chosen to maximize the data-ink ratio. An alternative version with equal scales makes clearer that the number of figures per page is almost always less than one.

The Handbook has been divided into three sections: Principles, Methodology, and Applications. Needless to say, the sections overlap. Figure 1.3 is a binary matrix visualization using Jaccard coefficients for both chapters (rows) and index entries
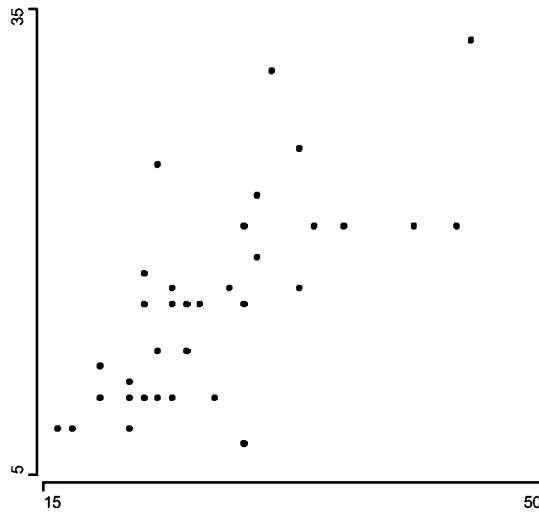
Figure 1.2. A scatterplot of the number of figures against the number of pages for the Handbook's chapters

(columns) to explore links between chapters. In the raw data map (lower-left portion of Fig. 1.3) there is a banding of black dots from the lower-left to upper-right corners indicating a possible transition of chapter/index combinations. In the proximity map of indices (upper portion of Fig. 1.3), index groups A, B, C, D, and E are overlapped with each other and are dominated by chapters of Good Graphics, History, Functional Data, Matrix Visualization, and Regression by Parts respectively.

Summary and Overview; Part II

1.2.1

The ten chapters in Part II are concerned with principles of data visualization. First there is an historical overview by Michael Friendly, the custodian of the Internet Gallery of Data Visualization, outlining the developments in graphical displays over the last few hundred years and including many fine examples.

In the next chapter Antony Unwin discusses some of the guidelines for the preparation of sound and attractive data graphics. The question mark in the chapter title sums it up well: whatever principles or recommendations are followed, the success of a graphic is a matter of taste; there are no fixed rules.

The importance of software for producing graphics is incontrovertible. Paul Murrell in his chapter summarizes the requirements for producing accurate and exact static graphics. He emphasizes both the need for flexibility in customizing standard plots and the need for tools that permit the drawing of new plot types.

Structure in data may be represented by mathematical graphs. George Michailidis pursues this idea in his chapter and shows how this leads to another class of graphic displays associated with multivariate analysis methods.
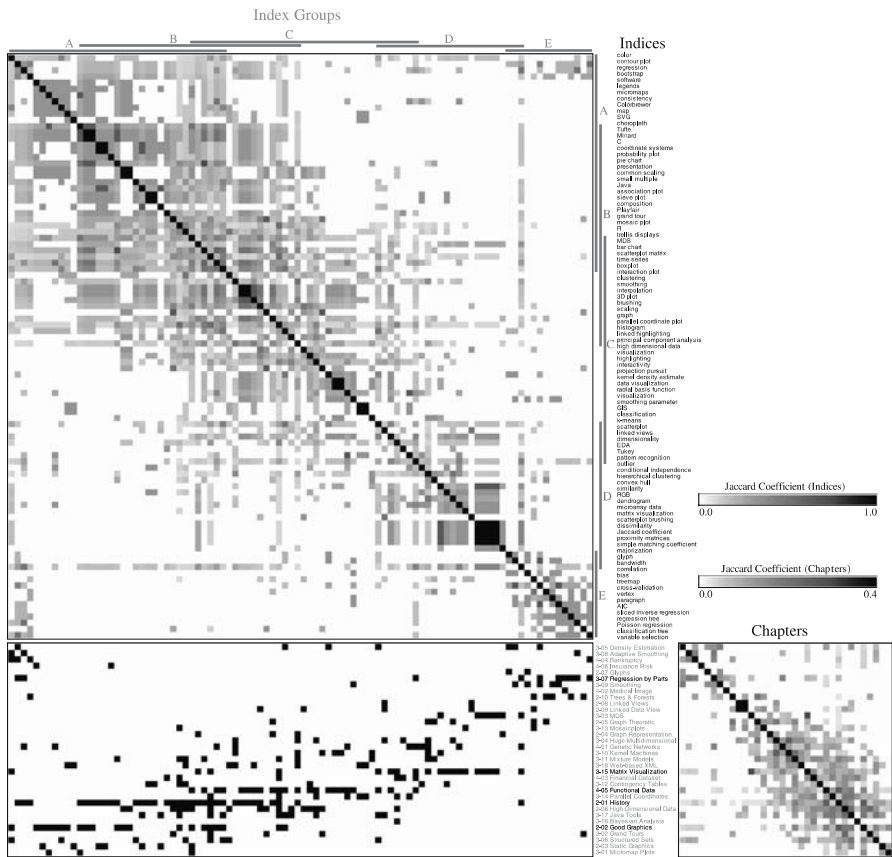
Figure 1.3. Matrix visualizations of the Handbook with chapters in the rows and index entries in the columns

Lee Wilkinson approaches graph-theoretic visualizations from another point of view, and his displays are concerned predominantly, though by no means exclusively, with trees, directed graphs and geometric graphs. He also covers the layout of graphs, a tricky problem for large numbers of vertices, and raises the intriguing issue of graph matching.

Most data displays concentrate on one or two dimensions. This is frequently sufficient to reveal striking information about a dataset. To gain insight into multivariate structure, higher-dimensional representations are required. Martin Theus discusses the main statistical graphics of this kind that do not involve dimension reduction and compares their possible range of application.

Everyone knows about Chernoff faces, though not many ever use them. The potential of data glyphs for representing cases in informative and productive ways has not been fully realized. Matt Ward gives an overview of the wide variety of possible forms and of the different ways they can be utilized.

There are two chapters on linking. Adalbert Wilhelm describes a formal model for linked graphics and the conceptual structure underlying it. He is able to encompass different types of linking and different representations. Graham Wills looks at linking in a more applied context and stresses the importance of distinguishing between views of individual cases and aggregated views. He also highlights the variety of selection possibilities there are in interactive graphics. Both chapters point out the value of linking simple data views over linking complicated ones.

The final chapter in this section is by Simon Urbanek. He describes the graphics that have been introduced to support tree models in statistics. The close association between graphics and the models (and collections of models in forests) is particularly interesting and has relevance for building closer links between graphics and models in other fields.

Summary and Overview; Part III   1.2.2

The middle and largest section of the Handbook concentrates on individual area of graphics research.

Geographical data can obviously benefit from visualization. Much of Bertin's work was directed at this kind of data. Juergen Symanzik and Daniel Carr write about micromaps (multiple small images of the same area displaying different parts of the data) and their interactive extension.

Projection pursuit and the grand tour are well known but not easy to use. Despite the availability of attractive free software, it is still a difficult task to analyse datasets in depth with this approach. Dianne Cook, Andreas Buja, Eun-Kyung Lee and Hadley Wickham describe the issues involved and outline some of the progress that has been made.

Multidimensional scaling has been around for a long time. Michael Cox and Trevor Cox (no relation, but an MDS would doubtless place them close together) review the current state of research.

Advances in high-throughput techniques in industrial projects, academic studies and biomedical experiments and the increasing power of computers for data collection have inevitably changed the practice of modern data analysis. Real-life datasets become larger and larger in both sample size and numbers of variables. Francesco Palumbo, Alain Morineau and Domenico Vistocco illustrate principles of visualization for such situations.

Some areas of statistics benefit more directly from visualization than others. Density estimation is hard to imagine without visualization. Michael Minnotte, Steve Sain and David Scott examine estimation methods in up to three dimensions. Interestingly there has not been much progress with density estimation in even three dimensions.

Sets of graphs can be particularly useful for revealing the structure in datasets and complement modelling efforts. Richard Heiberger and Burt Holland describe an approach primarily making use of Cartesian products and the Trellis paradigm. Wei-Yin Loh describes the use of visualization to support the use of regression models, in particular with the use of regression trees.

Instead of visualizing the structure of samples or variables in a given dataset, researchers may be interested in visualizing images collected with certain formats. Usually the target images are collected with various types of noise pattern and it is necessary to apply statistical or mathematical modelling to remove or diminish the noise structure before the possible genuine images can be visualized. Jörg Polzehl and Vladimir Spokoiny present one such novel adaptive smoothing procedure in reconstructing noisy images for better visualization.

The continuing increase in computer power has had many different impacts on statistics. Computationally intensive smoothing methods are now commonplace, although they were impossible only a few years ago. Adrian Bowman gives an overview of the relations between smoothing and visualization. Yuan-chin Chang, Yuh-Jye Lee, Hsing-Kuo Pao, Mei-Hsien Lee and Su-Yun Huang investigate the impact of kernel machine methods on a number of classical techniques: principal components, canonical correlation and cluster analysis. They use visualizations to compare their results with those from the original methods.

Cluster analyses have often been a bit suspect to statisticians. The lack of formal models in the past and the difficulty of judging the success of the clusterings were both negative factors. Fritz Leisch considers the graphical evaluation of clusterings and some of the possibilities for a sounder methodological approach.

Multivariate categorical data were difficult to visualize in the past. The chapter by David Meyer, Achim Zeileis and Kurt Hornik describes fairly classical approaches for low dimensions and emphasizes the link to model building. Heike Hofmann describes the powerful tools of interactive mosaicplots that have become available in recent years, not least through her own efforts, and discusses how different variations of the plot form can be used for gaining insight into multivariate data features.

Alfred Inselberg, the original proposer of parallel coordinate plots, offers an overview of this approach to multivariate data in his usual distinctive style. Here he considers in particular classification problems and how parallel coordinate views can be adapted and amended to support this kind of analysis.

Most analyses using graphics make use of a standard set of graphical tools, for example, scatterplots, barcharts, and histograms. Han-Ming Wu, ShengLi Tzeng and Chun-houh Chen describe a different approach, built around using colour approximations for individual values in a data matrix and applying cluster analyses to order the matrix rows and columns in informative ways.

For many years Bayesians were primarily theoreticians. Thanks to MCMC methods they are now able to also apply their ideas to great effect. This has led to new demands in assessing model fit and the quality of the results. Jouni Kerman, Andrew Gelman, Tian Zheng and Yuejing Ding discuss graphical approaches for tackling these issues in a Bayesian framework.

Without software to draw the displays, graphic analyis is almost impossible nowadays. Junji Nakano, Yamamoto Yoshikazu and Keisuke Honda are working on Java-based software to provide support for new developments, and they outline their approach here. Many researchers are interested in providing tools via the Web. Yoshiro Yamamoto, Masaya Iizuka and Tomokazu Fujino discuss using XML for interactive statistical graphics and explain the issues involved.

Summary and Overview; Part IV

The final section contains seven chapters on specific applications of data visualization. There are, of course, individual applications discussed in earlier chapters, but here the emphasis is on the application rather than principles or methodology.

Genetic networks are obviously a promising area for informative graphic displays. Grace Shieh and Chin-Yuan Guo describe some of the progress made so far and make clear the potential for further research.

Modern medical imaging systems have made significant contributions to diagnoses and treatments. Henry Lu discusses the visualization of data from positron emission tomography, ultrasound and magnetic resonance.

Two chapters examine company bankruptcy datasets. In the first one, Antony Unwin, Martin Theus and Wolfgang Härdle use a broad range of visualization tools to carry out an extensive exploratory data analysis. No large dataset can be analysed cold, and this chapter shows how effective data visualization can be in assessing data quality and revealing features of a dataset. The other bankruptcy chapter employs graphics to visualize SVM modelling. Wolfgang Härdle, Rouslan Moro and Dorothea Schäfer use graphics to display results that cannot be presented in a closed analytic form.

The astonishing growth of eBay has been one of the big success stories of recent years. Wolfgang Jank, Galit Shmueli, Catherine Plaisant and Ben Shneiderman have studied data from eBay auctions and describe the role graphics played in their analyses.

Krzysztof Burnecki and Rafal Weron consider the application of visualization in insurance. This is another example of how the value of graphics lies in providing insight into the output of complex models.

The Authors

The editors would like to thank the authors of the chapters for their contributions. It is important for a collective work of this kind to cover a broad range and to gather many experts with different interests together. We have been fortunate in receiving the assistance of so many excellent contributors.

The mixture at the end remains, of course, a mixture. Different authors take different approaches and have different styles. It early became apparent that even the term data visualization means different things to different people! We hope that the Handbook gains rather than loses by this eclecticism.

Figures 1.1 and 1.2 earlier in the chapter showed that the chapter form varied between authors in various ways. Figure 1.4 reveals another aspect. The scatterplot shows an outlier with a very large number of references (the historical survey of Michael Friendly) and that some papers referenced the work of their own authors more than others. The histogram is for the rate of self-referencing.
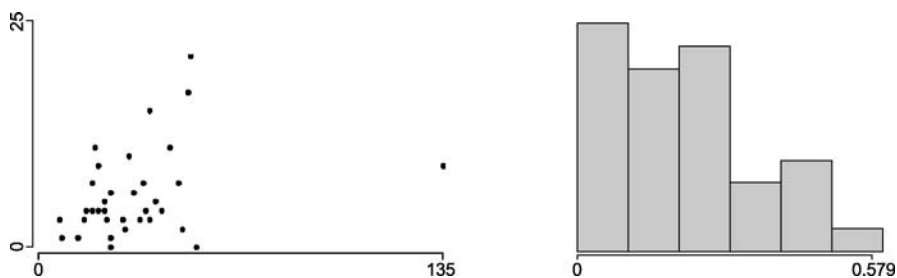
Figure 1.4. A scatterplot of the number of references to papers by a chapter's authors against the total number of references and a histogram of the rate of self-referencing

1.3 Outlook

There are many open issues in data visualization and many challenging research problems. The datasets to be analysed tend to be more complex and are certainly becoming larger all the time. The potential of graphical tools for exploratory data analysis has not been fully realized, and the complementary interplay between statistical modelling and graphics has not yet been fully exploited. Advances in computer software and hardware have made producing graphics easier, but they have also contributed to raising the standards expected.

Future developments will undoubtedly include more flexible and powerful software and better integration of modelling and graphics. There will probably be individual new and innovative graphics and some improvements in the general design of displays. Gradual gains in knowledge about the perception of graphics and the psychological aspects of visualization will lead to improved effectiveness of graphic displays. Ideally there should be progress in the formal theory of data visualization, but that is perhaps the biggest challenge of all.

Exploratory Graphics of a Financial Dataset

Antony Unwin[1], Martin Theus[2] and Wolfgang Härdle[3]

[1] Augsburg University, Germany `unwin@math.uni-augsburg.de`
[2] Augsburg University, Germany `martin.theus@math.uni-augsburg.de`
[3] Humboldt-Universität zu Berlin, Germany `haerdle@wiwi.hu-berlin.de`

1 Introduction

The first stages of any data analysis are to get to know the aims of the study and to get to know the data. In this study the main goal is to predict a company's chances of going bankrupt based on its recent financial returns. In another chapter of the Handbook, some sophisticated prediction models based on support vector machines are discussed for a similar dataset. Here, visualization methods are used to explore the large dataset of American company accounts that was made available for predicting bankruptcy, to get to know the data, and to assess the quality of the dataset. This is an initial exploratory analysis not using any expert accounting knowledge.

Exploratory Data Analysis (EDA) has been a well-known term in Statistics since Tukey's historical book (Tukey; 1977). While everyone acknowledges importance of EDA, little else has been written about it, and modern methods, including interactive graphics (Unwin; 1999), are not as much applied in practice as they might be. Interactive graphics were used extensively in the exploratory work for this chapter. The dataset is by no means particularly big but does contain more than 80,000 cases. Ways of graphically displaying large datasets are discussed in detail in Unwin, Theus and Hofmann (2006).

In considering graphic displays, it is necessary to distinguish between presentation and exploratory graphics. Graphics for displaying single variables or pairs of variables are often used for presenting distributions of data or for presenting results. Care must be taken with scales, with aspect ratios, with legends, and with every graphic property that may affect the success of the display in conveying information to others. Graphics for exploration are very different. They are more likely to be multivariate and there is no need to be particularly concerned about the aesthetic features of the graphic; the important thing is that they give a clear representation of the data. Presentation graphics are drawn to be inspected by many people (possibly millions of people, if they are used on television) and can be long-lived. Playfair's plots (Playfair; 2005) of English trade data, for example, are over 200 years old but

are still informative. Exploratory graphics are drawn for one or two people and may be very short-lived. A data analyst may examine a large number of graphics, before finding one that reveals information, and, having found that information, may decide that another kind of display is better for presenting it than the display or displays that led to its discovery.

The graphics shown in this chapter are a subset of those used in the study. They are not supposed to be "pretty", they have been drawn to do a job. Detailed scales have not been included, as it is always the distributional form that is of primary interest. Exploratory analyses are subjective and each analyst may choose different graphics and different combinations of them to uncover information from data. The only thing that can be said with certainty is that analysts who do not use graphics to get to know their data will have a poor understanding of the properties of the data they are working with. If nothing else, graphics are extremely useful for assessing the quality of data and for cleaning data.

2 Description of the Data

There are 82,626 records in the dataset. Each record contains financial information for a company for one year. Table 1 gives a list of the variables in the dataset.

Variable	Description
Cash.TA	Cash/Total Assets
Inv.TA	Inventories/Total Assets
CA.TA	Current Assets/Total Assets
Kap.TA	Property, Plant and Equipment/Total Assets
Intg.TA	Intangibles/Total Assets
LogTA	log(Total Assets)
CL.TA	Current Liabilities/Total Assets
TL.TA	Total Liabilities/Total Assets
Eq.TA	Equity/Total Assets
S.TA	Sales/Total Assets
Ebit.TA	EBIT/Total Assets
Ebit.Int	EBIT/Interest Payments
NI.TA	Net Income/Total Assets
CA.CL.TA	(Current Assets - Current Liabilities)/Total Assets
BANKR	status (Bankrupt or OK)
Year	year of accounts
State	where the company was registered
NAICS	North American Industry Classification System

Table 1. *Variables in the Bankruptcy dataset*

For each company there are 13 ratios, one size variable (log transform of assets) and a binary variable, which records whether the company went bankrupt within three years of the financial returns or not. There is also information on the State the company is registered in, on the industry sector and on the year of the accounts. The term "bankruptcy" includes Chapter 11 reorganization as well as liquidation under Chapter 7 of the US Bankruptcy Code. Financial ratios are commonly used so that data are comparable across years. Sometimes it is helpful to look at the raw data as well, particularly if there are data cleaning issues. An unusual value, or cluster of unusual values, in *Total Assets*, would affect all twelve ratios dependent on this variable. There were no missing values. Some of the ratio variables can be grouped into categories: Profit Measures (*Ebit.TA*, *NI.TA*); Leverage ratios (*Kap.TA*, *TL.TA*, and *Eq.TA*); Liquidity ratios (*Cash.TA*, *CA.TA* and *CA.CL.TA*); and Activity/Turnover ratios (*Inv.TA*, *S.TA*, and *Ebit.Int*).

To be able to generalise results from statistical models, the dataset analysed has to be a random sample from the population under study. For the bankruptcy dataset, there is a very large number of cases, but it is not clear how far they can be regarded as a random sample from anything. Apart from anything else, it is not at all clear that they can be considered to be homogeneous. Most companies are rather small and a few are very large. Can the same financial ratios really be used for companies that are so different in scale? This is the kind of question that exploratory analysis can help answer by looking at the distributions of values of data for the different groups.

The assumption is that results from this dataset can be used on datasets collected in similar fashion in the future.

Fig. 1. A bankruptcy barchart. 506 of the 82626 records refer to bankruptcy.

3 First Graphics

Figure 1 is a barchart for the bankruptcy variable. Only a small proportion of companies actually went bankrupt, which is a good thing for the companies, but makes any statistical analysis more difficult.

The displays in Figure 2 and Figure 3 show that the companies are fairly equally distributed across the regions but that some of the data are surprisingly old with a few cases prior to the mid 1960's. In the 1950's there is one record per year and linking to the geographic data shows it was not always the same company, as might have been thought.

The geographic information was originally provided by State, so there were a large number of small counts and only a few big ones. Grouping by region gives a good overview, though other groupings (e.g., by population) could also be tried. The regional classification used here is one from the FBI. Selecting the foreign group, a little more than 10% of the cases, and linking to a spinogram (Hofmann and Theus; 2006) of years, (see Figure 4), shows that the percentage of foreign registered companies has increased over the years. Querying shows this to be from about 4% in the early 1970s to 15% in 2002. In the most recent year, 2003, the rate falls to just under 11%. It is expensive for foreign firms to be listed on the US exchanges and opinion has changed as to what benefits it brings. The Sorban-Oxley act has also made it less attractive to be listed in the US.

Information was also available on industry sector in two different ways, with one classification of 426 categories by name and numerical coding and the NAICS classification by number with 925 categories. Both are far too detailed for graphical analysis and a hierarchical grouping similar to the spatial grouping of the States can be tried. The six-digit NAICS code can be combined by their first two digits and then further grouped by sector to give Figure 5. The manufacturing sector clearly dominates. To check for associations between the two classifications a crude scatterplot approach was tried. A fairly random spread of points was obtained but no particular pattern.
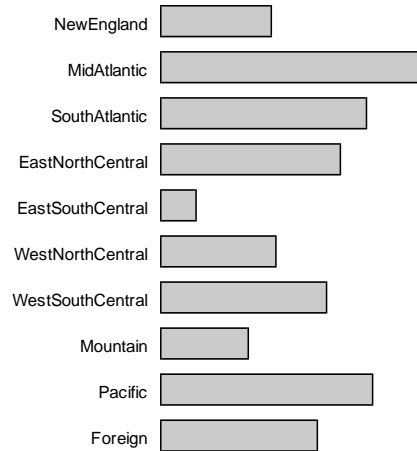
Fig. 2. A barchart of the numbers of records by US region.
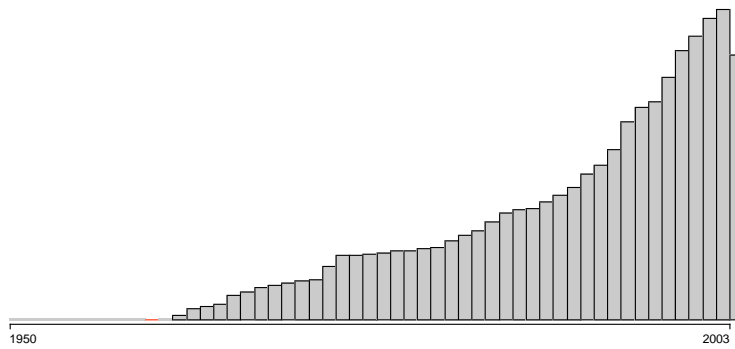
Fig. 3. A histogram of the numbers of records per year.
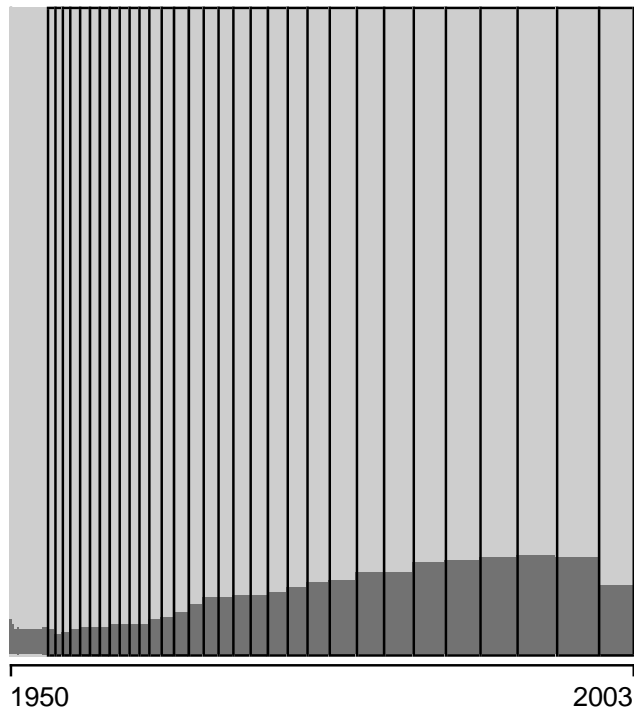
Fig. 4. A spinogram of the numbers of records per year with foreign registered companies selected. The width of a bar in a spinogram is proportional to the height of the corresponding bar in the original histogram.
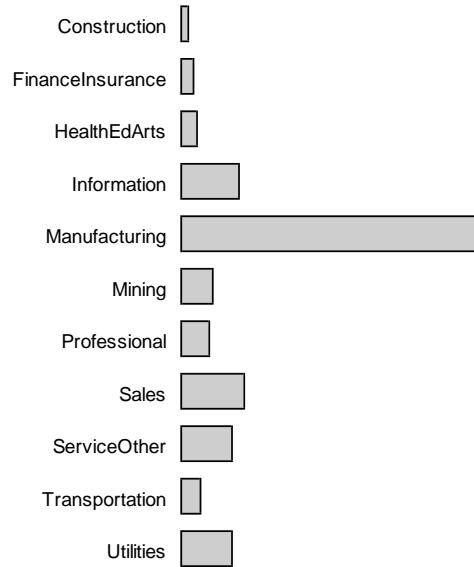
Fig. 5. A barchart of the numbers of records by NAICS groups.

Histograms might be drawn for the continuous variables, but they take up quite a lot of room, so boxplots are more efficient for displaying many variables at once. Since *logTA*, the log of total assets, is a different kind of variable from the others and important in its own right as well (because it groups the companies by size), a histogram has been drawn just for it (see Figure 6). This shows a roughly symmetric distribution with a few very low values. Selecting the foreign-registered companies again and linking to a spinogram of *logTA* reveals that the percentage of these companies rose steadily for the biggest 25% of the companies, from 7% up to more than 50%.

A set of parallel boxplots for the ratio variables is shown in Figure 7. The boxplots reveal that several of the ratios are highly skewed and this may affect whether they can be of much help in discriminating between the companies which went bankrupt and those which did not. It is possible that many of these outliers are errors of some kind and it may be that most of the outliers on individual variables are actually outliers on several variables.

4 Outliers

Outliers can be checked with a parallel coordinates display (Inselberg; 1999), as in Figure 8, where the eight ratios with highly skew distributions have been plotted and seven of the worst outliers selected. It is easy to see that several
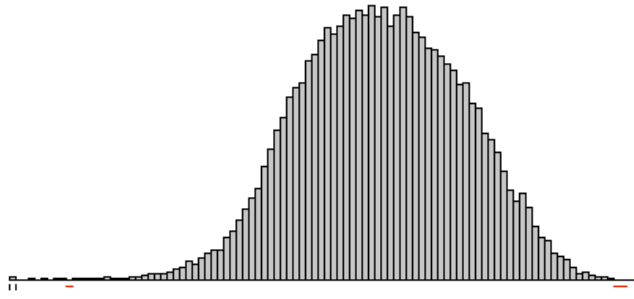
Fig. 6. A histogram of log(total assets),*logTA*, for the companies. The marks to the left under the axis are interactive controls for the anchorpoint and binwidth. The horizontal (red) marks record bins where the count is too small to be drawn.
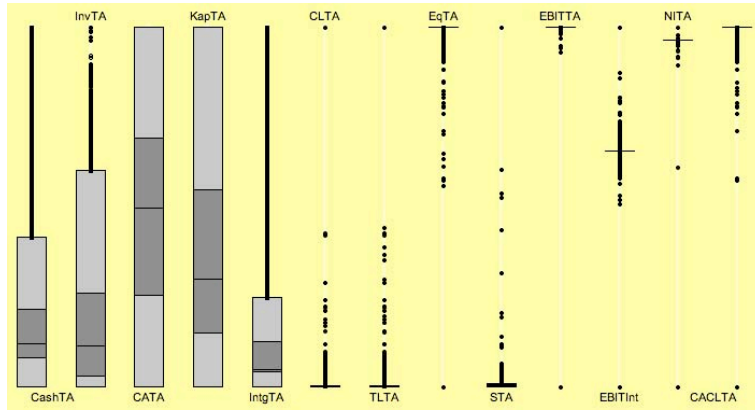
Fig. 7. Parallel boxplots of financial ratios. Each boxplot is individually scaled.

are outliers on more than one variable. It is also apparent that the ratio of equity to total assets *Eq.TA* is perfectly inversely correlated with the ratio of total liabilities to total assets *TL.TA*. This is a matter of definition with $TL.TA + Eq.TA = 1$. Although the equation looks innocuous it masks the fact that in this dataset *TL.TA* ranges from 0 to 5838. Not surprisingly, the high values of *TL.TA* only arise for low values of *Total Assets* as the L-shaped scatterplot on the left of Figure 9 shows. This plot is somewhat misleading. The zoomed version on the right reveals that there is more variability amongst the low values than the default plot suggests, though the most extreme values of *TL.TA* are still only for very low values of *Total Assets*. The bulk of the pattern suggests that very small companies have a broader range of possible liability ratio values than small companies. The low density region to the lower right implies that companies exceeding a certain size must have some liabilities. The distorting effect of the extreme values is demonstrated by the

fact that the zoomed plot, which is 10^{-5} the size of the default plot, contains 87% of the data. α-blending has been used to make the distributional structure more visible. α-blending weights the shading of each object by a specified fraction. The darkness of a pixel is that fraction times the number of objects overlapping the pixel or 1, whichever is smaller.
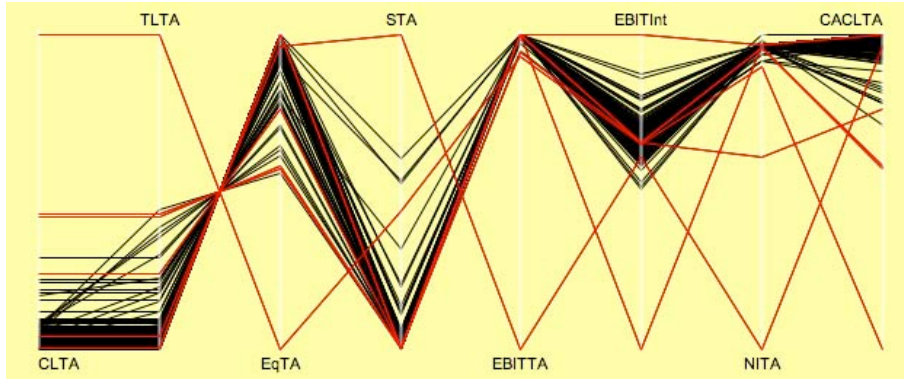
Fig. 8. Parallel coordinate plot of financial ratios with skew distributions. Seven outliers have been selected.
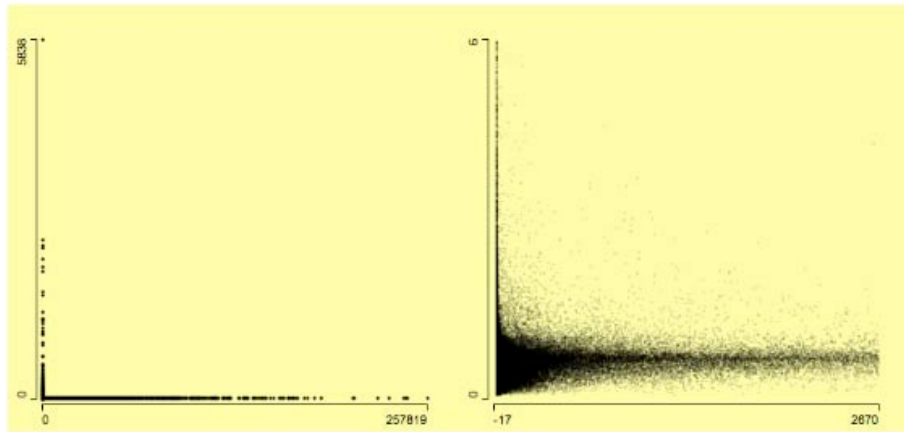
Fig. 9. Scatterplots of *TL.TA*, the ratio of total liabilities to total assets, plotted against *Total Assets*. On the left all the data, showing that all high values of liabilities are associated with low values of assets. On the right, a zoom of about 10^{-2} on the x-axis by 10^{-3} on the y-axis with some α-blending.
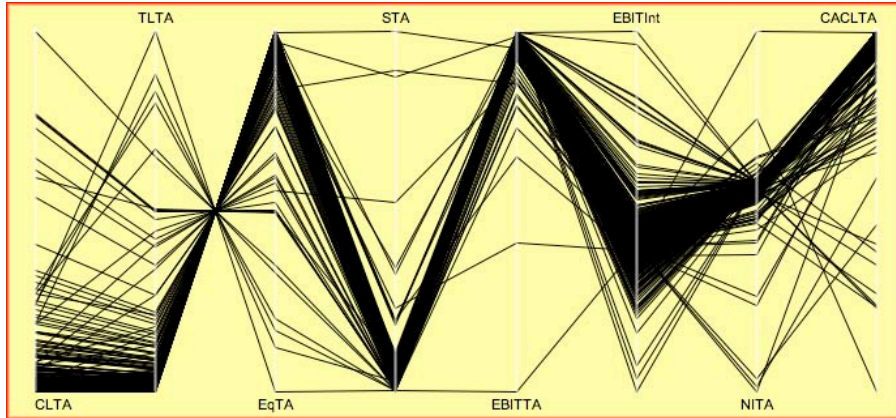
Fig. 10. Parallel coordinate plot of the financial ratios with skew distributions. The seven outliers selected in Figure 8 have been removed. The plot's (red) border is a sign that not all data are displayed.

Setting aside the seven worst outliers, Figure 8 rescales to Figure 10. Some of the variables become potentially more informative, the scales of some others are dominated by newly visible outliers.

Outliers can be dealt with in several ways. A transformation might help (but that is not always appropriate, in this case several ratios have some negative or zero values). The outliers could be trimmed or discarded (but that depends on what kind of outlier they are). With ratio variables, as here, the component variables of the ratios can be examined. Figure 11 shows a scatterplot of the variable *Sales* against *Total Assets* with the same seven outliers still highlighted. All turn out to be small companies, in terms both of *Sales* and of *Total Assets*, and small companies should probably be treated separately anyway. The scatterplot also reveals that there are some bivariate outliers. Querying shows that the six cases with high *Sales* but low *Total Assets* (the group upper left) are all the same big retail company over several successive years. The seven cases with low *Sales* but very high *Total Assets* (the group lower right) represent three companies from the information and communication sector. These two groups make up a tiny proportion of the dataset, even if they are all very large on either *Sales* or *Total Assets*.

For exploratory work it is distributional structure that is of interest, not precise values. The minima and maxima of the variables are used to determine the limits of the axes. The lower limit for *Sales* in Figure 11 shows that the lowest *Sales* values are actually negative and that there are plenty which are zero or almost zero as well. The negative values could be an accounting quirk, but should surely be discarded from the dataset, though there are only five of them. The low *Sales* values are also worth considering, what kind of companies are these and should they be kept in the dataset? They may be very new companies or companies on their last legs. There are 1412 companies
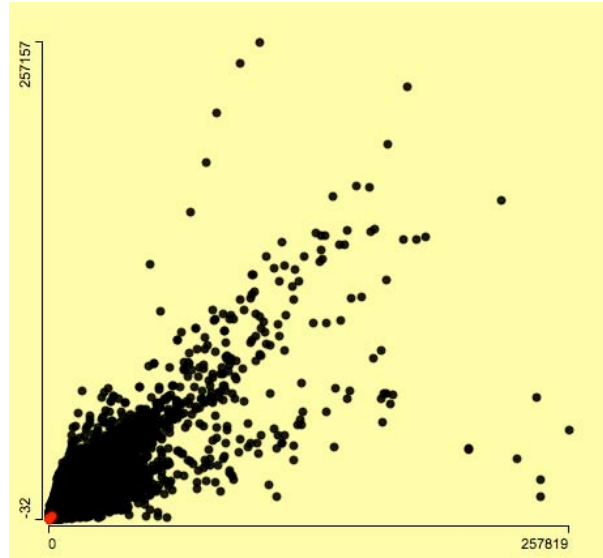
Fig. 11. Scatterplot of *Sales* v. *Total Assets* with the 7 outlying companies highlighted, the lighter (red) blob in the lower left corner.

with zero *Sales* and another 3674 with *Sales* more than zero but less than 1. These data could in principle have been obtained by zooming into a histogram for *Sales* and querying the cells, but for queries with precise boundaries it is quicker to calculate the appropriate frequency table. Graphics are better for more general, qualitative insights, while tables and statistical summaries are better for exact details.

The form of empirical distributions can be examined in many ways. When cases are of different sizes or weights, it can be revealing to look at weighted distributions. For instance, Figure 12 shows a histogram of *CA.TA*, the ratio of current assets to total assets, on the left and a histogram of the same variable weighted by *Total Assets* on the right. Companies with the highest current assets ratios clearly have low *Total Assets*.

Outliers and negative values are some of the data cleaning problems that can arise, there may be others as well.

Some statistical modelling approaches are little affected by individual gross errors and it may be that whether these cases are excluded, or adjusted, or just included in the analysis anyway, will not affect the model fit. Even when this is the case, it is useful to know what kinds of errors or unusual values can occur. It is also an opportunity to talk to people who know the context of the data well, and to get them to provide more background information. Analysing data blind without having any information is just reckless. It is surprising (and sometimes shocking) how much useful information is known about datasets that is not incorporated into analysis, be it information about
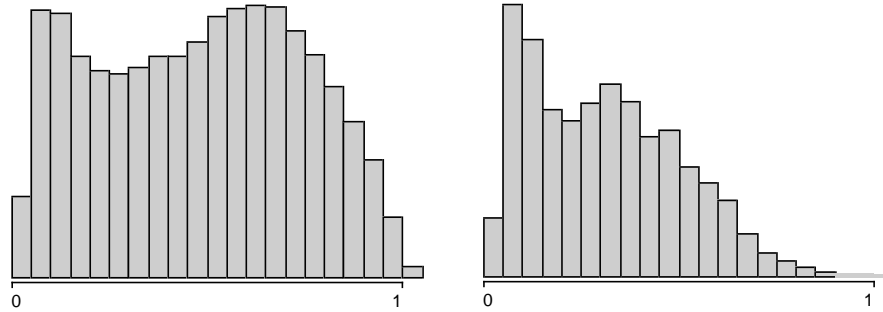
Fig. 12. A histogram of the current assets ratio on the left and a weighted histogram of the same variable, weighted by *Total Assets*, on the right.

the selection of variables collected, about the way the sample was chosen, or about the details of the data collection and data recording. Dataset owners may assume that analysts know this information but it is useful to check.

5 Scatterplots

There is often a temptation with large numbers of continuous variables to calculate correlation coefficients for all possible pairs of variables. The trouble is that correlations can be high because of outliers, or low because the association between the two variables is non-linear. Fourteen variables (the original ratios plus *logTA*) are too many to draw a scatterplot matrix, but it is interesting to look at the associations between subsets of the ratios. Scatterplots for two pairs of the financial variables *Cash.TA*, *Inv.TA*, *CA.TA* and *Kap.TA* are shown in Figure 13. The lower left triangular shapes show that the sum of the corresponding ratios is less than a limit, in these cases 1. (Lower right triangular shapes are obtained when the y variable is always lower in values than the x variable, as when the cash or inventory ratios are plotted against the current assets ratio.) The negative cash values that have already been discussed are easily seen to the left of the bulk of the data in the plot of cash and inventory ratios. In the plot of the current assets and property ratios there are a few companies which surprisingly lie above the bounding diagonal — more outlying cases to investigate. The correlation coefficients for the variables are given in Table 5. Some of them are quite high, they are certainly all significantly different from zero, but they hint little at the structure shown. The highest value in absolute terms is the correlation between current assets and property (-0.752). This deserves further investigation and Figure 14 shows the same scatterplots as in Figure 13, this time using the smallest pointsize and some α-blending instead of the defaults, providing a rough bivariate density estimate. It is now possible to see that for many of the companies, the sum of current assets and property assets is almost equal to *Total Assets*.
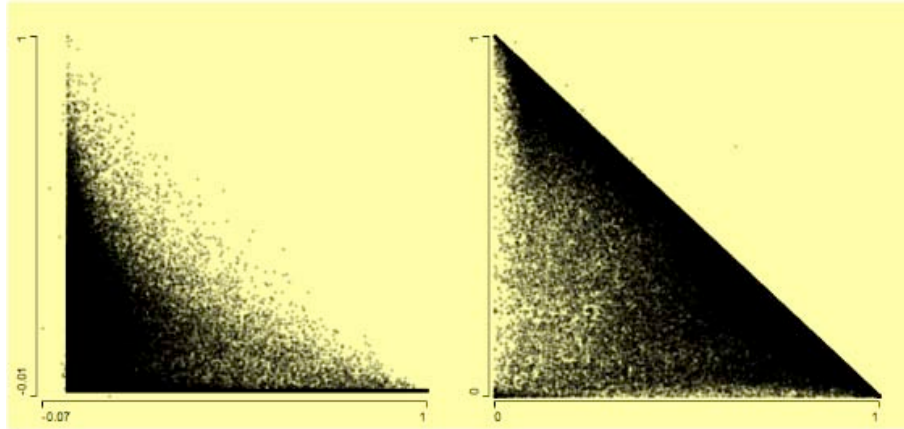
Fig. 13. Scatterplots of *Cash.TA* and *Inv.TA*, the ratios of cash and inventories to total assets, (left), and of *CA.TA* and *Kap.TA*, current assets and property to total assets (right).

	Cash.TA	Inv.TA	CA.TA	Kap.TA
Cash.TA	1.000			
Inv.TA	-0.224	1.000		
CA.TA	0.482	0.553	1.000	
Kap.TA	-0.378	-0.352	-0.752	1.000

Table 2. Correlations between the four ratio variables in Figure 13.

The scatterplots of the corresponding raw data variables provide another view of the associations between variables in the dataset, and more structure
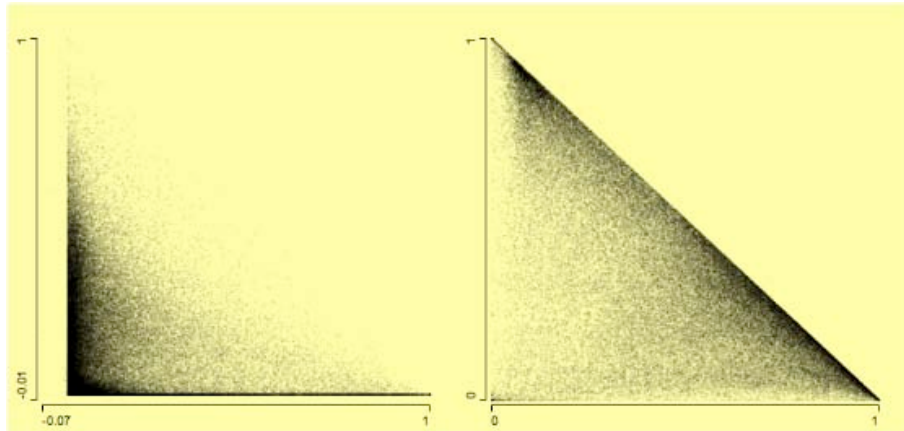
Fig. 14. The same scatterplots as in Figure 13 but with a smaller pointsize and α-blending to display the bivariate structures better.

can be seen than with scatterplots of the ratios. A few of the raw scatterplots exhibit a funnel structure, such as the plot of inventories against fixed assets and that of cash against current assets (Figure 15). Querying and linking can be used to identify specific sectors or outliers. The companies with high assets and low inventories are in the information and communication sector. Companies in the retail sector have higher inventories and lower assets. The at first sight unusual three points to the top left in the inventories/fixed assets plot are from investment banking and are consistent with other companies in that sector — except for being considerably bigger. The biggest companies on both variables in the left-hand plot are car manufacturers (higher inventories) and oil companies (higher assets).

Sometimes features stand out in a parallel coordinate plot and sometimes they are more visible in raw data scatterplots. As always, a range of different graphics should be explored.
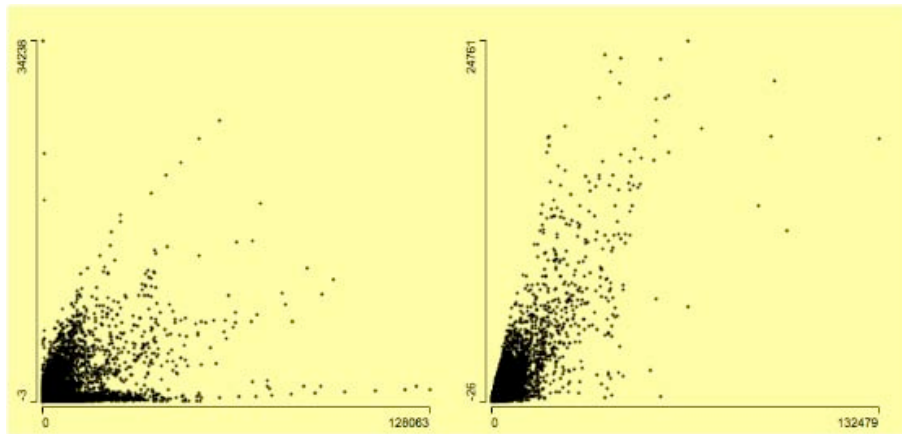
Fig. 15. Scatterplots of inventories against fixed assets (left) and of cash against current assets (right).

6 Mosaic Plots

Combinations of categorical variables may be displayed in one form or other of mosaic plots (Hofmann; 2000). For this dataset there are two difficulties with drawing these plots. Firstly, all but one of the categorical variables have very many categories (there are 54 states and 925 NAICS categories). Secondly, the one categorical variable that is binary ($BANKR$) has less than 1% of the cases in one category, so that highlighting is rarely visible. The first problem can be partly got round by grouping, combining states into regions and using

a less detailed form of the NAICS. The second problem could be solved by a special form of zooming.

Figure 16 shows a fluctuation diagram of the numbers of companies by industry sector and region. Classical mosaic plots try to make the most efficient use of the space available by making each cell as large as possible. This can make the plot difficult to interpret, especially when there are many cells. Fluctuation diagrams preserve a grid structure, so that comparisons by rows and columns are easier. The dominance of the manufacturing sector stands out in Figure 16, as well as individual details, such as the concentration of mining companies in WestSouthCentral.
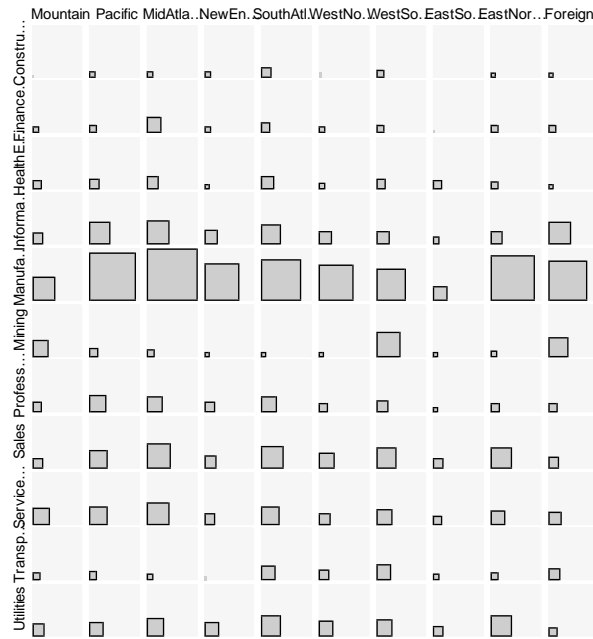
Fig. 16. Fluctuation diagram of industry sectors by regions.

All cases are treated equally in Figure 16. Of course, some companies are much bigger than others and Figure 17 shows the same cases weighted by *Total Assets*. The differences between the two figures are striking. Foreign registered companies in the manufacturing sector are much bigger than was apparent before. The regional distribution of companies in the professional sector changes a lot.
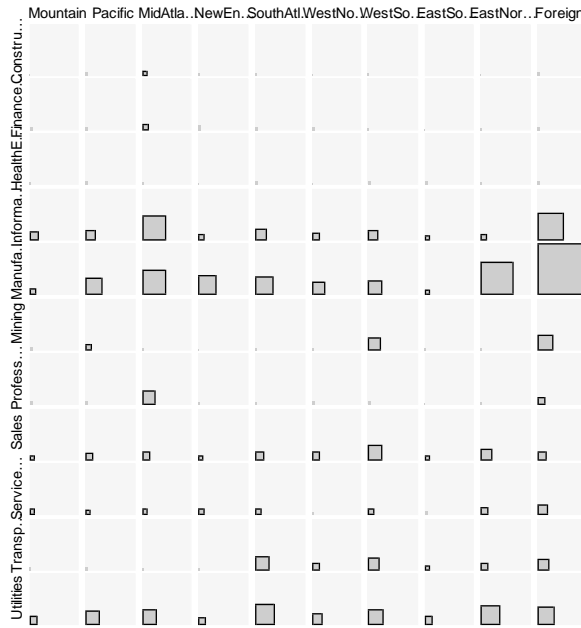
Fig. 17. Fluctuation diagram of sectors by regions weighted by *Total Assets*.

7 Initial Comparisons of Bankrupt Companies

The analysis up till now has considered all the data as one group and a number of outliers, distributional features, and specific properties have been identified. The main aim of the study remains the investigation of how the companies which went bankrupt differ from the rest. A first approach would be to look at the information available. Any company whose total liabilities exceed their total assets ($TL.TA > 1$) is likely to be in trouble. This variable is bizarrely skew and so difficult to visualize with standard plots. In this situation a simple table is better.

	Bankrupt	OK
TL.TA > 1	272	5856
	4.44%	95.6%
TL.TA ≤ 1	234	76264
	0.31%	99.7%

Table 3. Cases with more liabilities than assets and their bankruptcy status.

Clearly the variable *TL.TA* on its own is very informative. The choice of the boundary limit of 1 is determined by the context. Varying the limit interactively using a slider confirms that it is a good choice, in that both the difference in bankruptcy rates and the number of bankrupt companies are high.

To investigate the effects of more than one variable, two kinds of parallel plots can be used with the bankruptcy cases highlighted. Parallel boxplots give univariate summary comparisons, while parallel coordinate plots potentially offer multivariate comparisons.

Figure 18 shows the default parallel coordinate plot of the ratios (without *Eq.TA* but with *logTA*), drawn for all the data except 55 outliers removed in an initial data cleaning. The companies that went bankrupt have been selected. The heavy overplotting may be obscuring information and the fact that every line is drawn the same whether it represents one case or many may also mislead. Nevertheless some features can be identified: all bankrupt companies had low values of *CL.TA*, high values of *Ebit.TA* and medium values of *Ebit.Int*; there are two unusual bankrupt companies (one a high outlier on *TL.TA* and the other a high outlier on *S.TA*).

One solution to get better discriminatory power is to use α-blending. A factor of 0.1 has been used in Figure 19, and it is now possible to see that the concentration of values for bankrupt companies on some of the variables applies to the bulk of the rest of the data as well, so that variables like *CL.TA* and *Ebit.TA* will not be as informative as might have been hoped. A final step can be to apply α-blending to the highlighting as well and that has been tried in Figure 20. This suggests that *Cash.TA* and *Intg.TA* might be more helpful than at first thought.
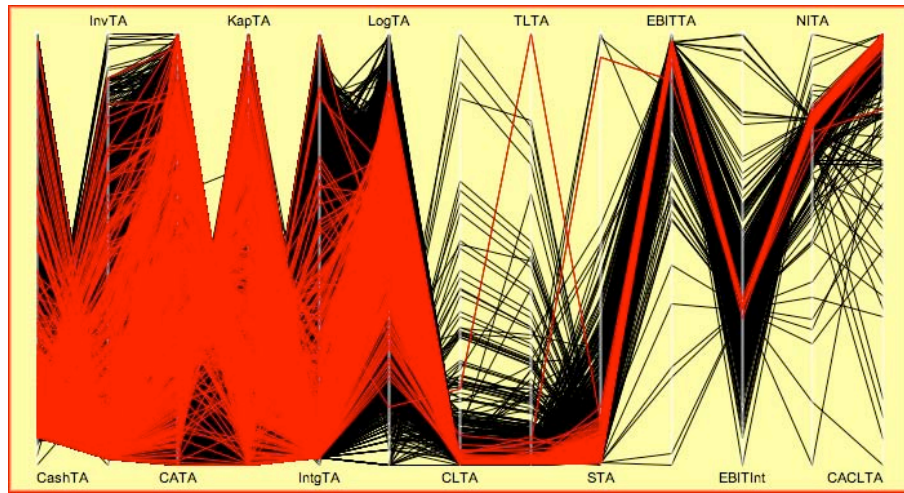

Fig. 18. Parallel coordinate plot of financial ratios and *logTA*, excluding 55 outliers, with bankrupt companies highlighted.
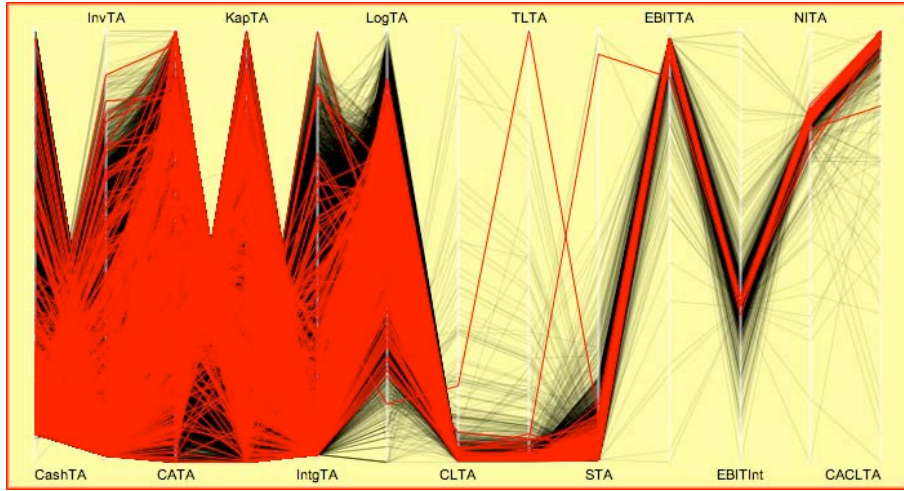
Fig. 19. Parallel coordinate plot of financial ratios and *logTA*, excluding 55 outliers, with bankrupt companies highlighted, α-blending=0.1 only for unselected data.
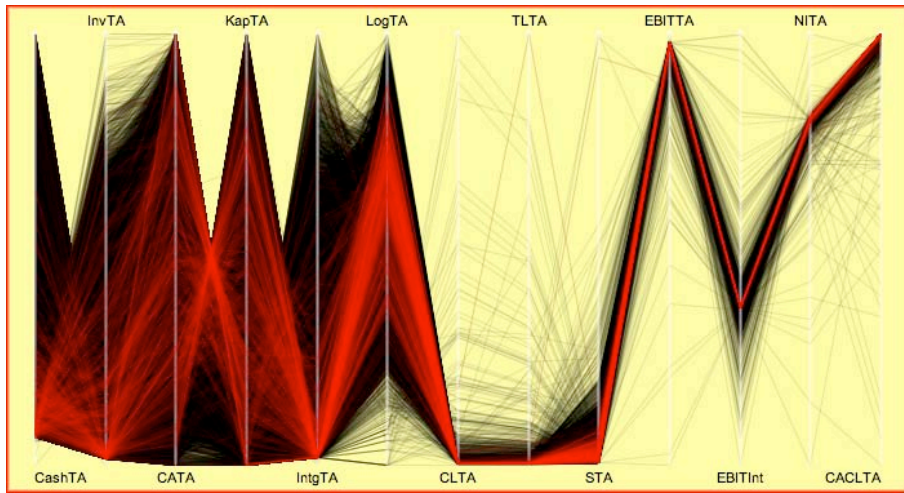

Fig. 20. Parallel coordinate plot of financial ratios and *logTA*, excluding 55 outliers, with bankrupt companies highlighted, α-blending=0.1 for selected and unselected data.

Using selection and linking for scatterplots can work well, but is dependent on the data distribution. Consider the two variables just mentioned above, *Cash.TA* and *Intg.TA*, and their distribution related to the companies which went bankrupt. A scatterplot could be drawn with a little α-blending, as on the left of Figure 21, or with a lot of α-blending as on the right. (α-blending has not been used for the highlighted cases, as on this scale they then disappear.)

Both plots contribute information but neither is fully satisfactory. Another alternative would be to draw a pair of scatterplots, one for each of the two groups of companies, but this is difficult to interpret. The success of particular plots depends on there being clear-cut information to find. The differences between the companies that went bankrupt and the others that did not are more complicated than can be displayed in a scatterplot of two variables.

Another alternative would be to use spinograms (as in Figure 24), but the numbers of bankruptcy companies are so small relative to the total number of companies that little can be seen. Fitted smooths would be better, though they require more computation.

The parallel coordinate plots employed here are a selection of many that might have been shown. The choice and order of variables influences what might be seen. The decision to discard some extreme outliers does too. The level of α-blending is also an influential factor. In other words, a parallel coordinate plot is like any other multivariate analysis, the user has a lot of freedom of choice. Careful thought helps, and so does statistical modelling. Having explored the data and built a model, parallel coordinate plots can again be useful, this time as a way to understand the model's relationship to the dataset.

8 Investigating Bigger Companies

Financial data for small companies is highly variable and could well be more unreliable than data for large companies, though this is difficult to assess. Certainly, large companies are different from small companies, and so studying
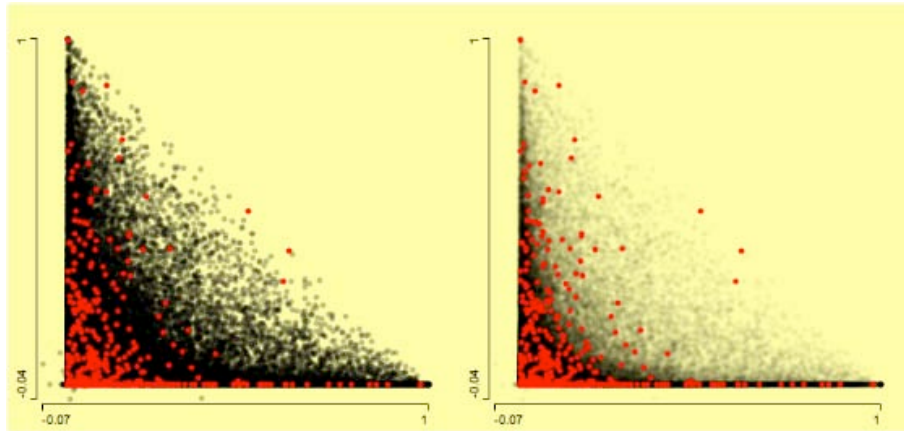

Fig. 21. Scatterplots of the ratios of intangibles and cash to total assets with companies that went bankrupt selected. More α-blending has been used in the right-hand plot.

them separately makes sense. In one important way they are not different: the bankruptcy rate for the biggest 7,690 companies (each with *Total assets* > 4000), 0.59%, is close to the rate for the rest of the dataset, 0.62% for the remaining 74,936 companies. On the other hand, using a looser definition of big (*Total assets* > 1000) gives bankruptcy rates of 0.73% for the 18610 'big' companies and 0.58% for the rest, a significant difference. Given the many different limits that might be used, a wide range of results is possible. However, it is not modest variations in bankruptcy rates by company size that are of interest, it is identifying which companies might go bankrupt.

Over forty years it would be reasonable to expect that the size of companies has increased. Curiously, for this dataset, the effect on *logTA* is negligible as Figure 22 shows.
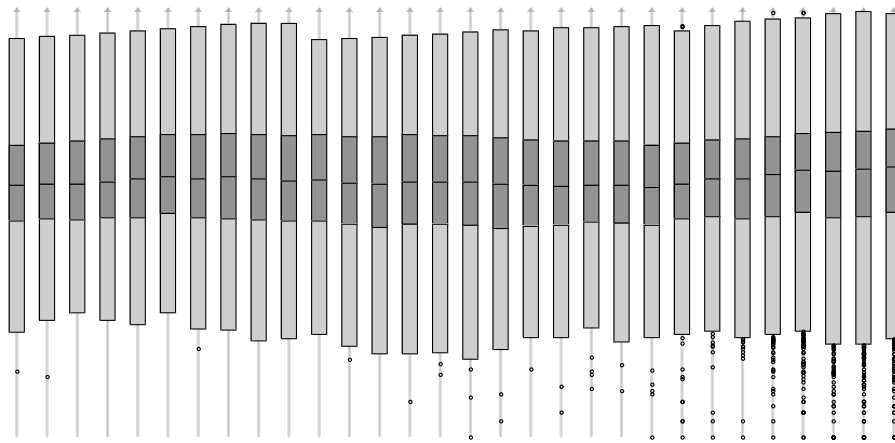

Fig. 22. Parallel boxplots of *logTA* by year from 1974 to 2003, all on the same scale.

Figure 23 shows boxplots of the original ratio variables for all the data with the group of big companies highlighted. Only the first five financial ratios are shown, as the distributions of the others are, as Figure 7 showed, far too skew to be informative. *logTA* is included to show the size distribution. Although the medians for the bigger companies differ noticeably from the median ratios for all companies (for cash, inventories and, hence, obviously, current assets, they are lower, and for the capital assets ratio the median ratio is higher), the distributions overlap substantially.

A more effective way of looking at the ratios is to use spinograms. In Figure 24 there are plots of the ratios *Cash.TA*, *Inv.TA*, *Kap.TA*, and *Intg.TA* with the big companies selected. The proportion of big companies declines as the cash ratio increases; it also declines as the inventory ratio increases, apart from the lowest group (*Inv.TA* < 0.01), where the proportion is relatively low; the proportion increases as the fixed capital ratio increases; the proportion is fairly constant for the intangibles ratio.
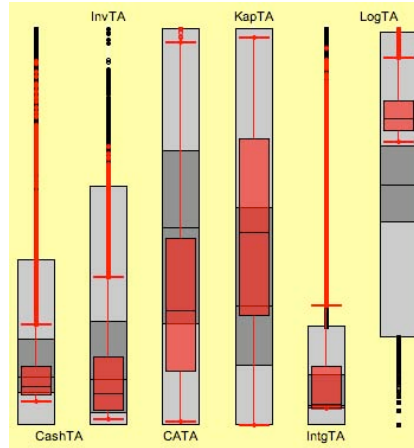
Fig. 23. Parallel boxplots of financial ratios and *logTA* for all companies. The background boxplots are for all the data and the superimposed standard boxplots are for the selected cases, companies with *Total Assets* > 1000.

Figure 25 shows just the data for the bigger companies for all variables with the bankrupt companies selected. Outliers are still affecting most of the second group of ratios, but two features stand out: *TL.TA* is generally higher, as we would expect from Table 7 for the whole dataset, and *Ebit.TA* is generally lower. It is tempting to draw conclusions about the fact that none of the biggest companies went bankrupt and that none of the companies that went bankrupt had a high cash ratio (even though the median is higher for these companies), but the selected cases make up such a small percentage of the total that caution should be exercised before drawing any conclusions.

9 Summary

Every data analysis is unique, because the data are always different. In the study reported here, there were mainly continuous variables (so parallel co-ordinate plots were useful); the few categorical variables had mostly large numbers of categories (so these had to be combined into groups); there was a fairly large number of cases (so approximating density estimations were helpful); and there were some variables that were highly skew (so that outliers and transformations were issues). A variety of plots were used, including barcharts, histograms, spinograms, boxplots, scatterplots, mosaic plots and parallel coordinate plots. Weighted versions of some plots also contributed. Trellis displays might have been tried, but then shingling of the conditional variables would have been required. That would be more appropriate after modelling. Interactivity, primarily selection, querying and linking, was used extensively, as is clear from the plots, but zooming and reformatting were also
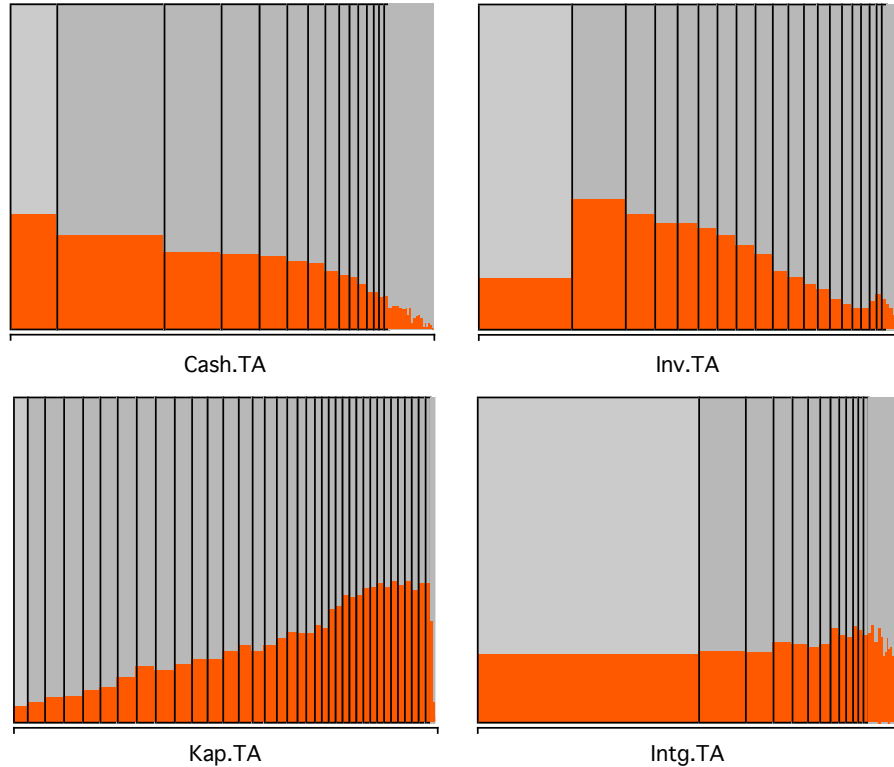
Fig. 24. Spinograms of the ratios *Cash.TA*, *Inv.TA*, *Kap.TA*, and *Intg.TA*. The companies with *Total Assets* > 1000 have been selected.

used a lot in the exploratory analyses. It is not easy to illustrate EDA in print and the chapter can only convey a pale shadow of the actual process.

Exploring the data is an important part of any data analysis. It is necessary to learn about the data, to check data quality and to carry out the data cleaning that is needed (and data cleaning is always needed with real datasets). EDA revealed here that there were some extreme outliers and some suspicious negative values. It underlined the need to transform some of the variables and it highlighted the geographic and sectoral structure of the dataset. It also revealed the surprising age of some of the data and the unexpected stability of the size distribution over time. Several interesting associations between variables were uncovered. Investigating the factors influencing bankruptcy gave further insights into the data and prepared the ground for statistical modelling.

Applying statistical models before exploring data is inefficient. Problems may arise because of some peculiarity of the data. Features are revealed that could have been found much more easily by just looking. The only possible
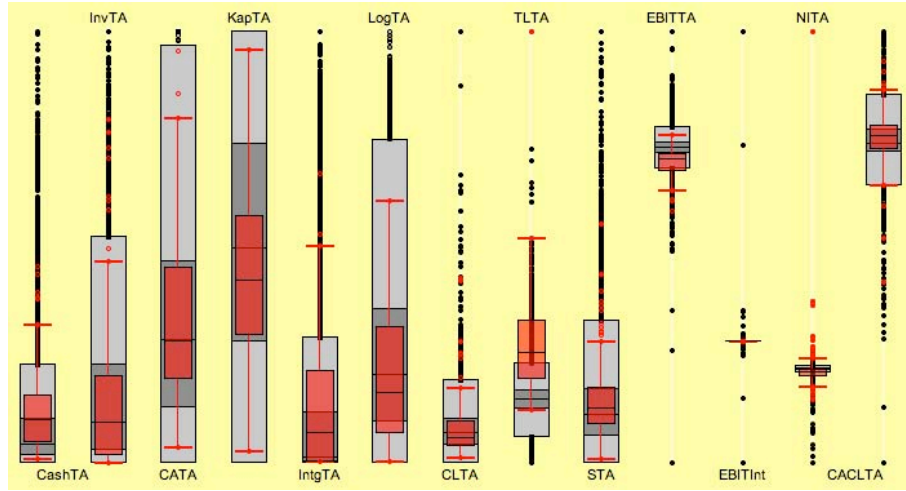
Fig. 25. Parallel boxplots of financial ratios and *logTA* for the 18610 companies with *Total Assets* > 1000. Companies that went bankrupt have been selected.

advantage of modelling without taking a look at the data first is the purist one of satisfying the idealist prerequisite for a hypothesis test — although whether that is relevant for any real analysis is moot.

Graphics are essential for exploratory work. They provide overviews and insights that complement statistical summaries. On the other hand, graphical analyses without follow-up analytic support remain inconclusive. Visualization results tend to be qualitative rather than quantitative, general rather than precise. Statistical modelling can assess the strength of evidence supporting ideas generated from graphical EDA and help to define those ideas more exactly. On top of that, statistical modelling can tease out more complex relationships, which are not immediately visually apparent. However, there is not much point in looking for complex relationships if the quality of the data is in doubt, and that is one of the reasons why modelling benefits from prior data visualization. Modelling also benefits from graphical support after analysis, both in investigating residual patterns for individual models and in comparing and combining groups of models. That is discussed in other chapters in the Handbook.

Software
The software used for most of the displays in this paper was Martin Theus's Mondrian (http://stats.math.uni-augsburg.de/Mondrian/). Other software was used at various stages to assist with data cleaning and restructuring.

References

Hofmann, H. (2000). Exploring categorical data: interactive mosaic plots, *Metrika* **51**(1): 11–26.

Hofmann, H. and Theus, M. (2006). Interactive graphics for visualizing conditional densities, *Journal of Computational and Graphical Statistics* .

Inselberg, A. (1999). Dont panic ... do it in parallel, *Computational Statistics* **14**(1): 53–77.

Playfair, W. (2005). *Playfair's Commercial and Political Atlas and Statistical Breviary*, Cambridge, London.

Tukey, J. (1977). *Exploratory Data Analysis*, Addison-Wesley, London.

Unwin, A. (1999). Requirements for interactive graphics software for exploratory data analysis, *Computational Statistics* **14**: 7–22.

Unwin, A. R., Theus, M. and Hofmann, H. (2006). *Graphics of Large Datasets*, Springer.