Does hedging with implied volatility factors improve the hedging efficiency of barrier options? *

Szymon Borak†
Matthias R. Fengler
Wolfgang K. Härdle

CASE – Center for Applied Statistics and Economics
Humboldt-Universität zu Berlin,
Spandauer Straße 1, 10178 Berlin, Germany

February 11, 2009

Forthcoming: The Journal of Risk Model Validation

*We gratefully acknowledge financial support by the Deutsche Forschungsgemeinschaft and the Sonderforschungsbereich 649 "Ökonomisches Risiko".

†Corresponding author: borak@wiwi.hu-berlin.de, TEL ++49 30 2093 5630 FAX ++49 30 2093 5649.

Does hedging with implied volatility factors improve the hedging efficiency of barrier options?

Abstract

The price of a barrier option depends on the shape of the entire implied volatility surface which is a high-dimensional dynamic object. Barrier options are hence exposed to nontrivial volatility risk. We extract the key risk factors of implied volatility surface fluctuations by means of a semiparametric factor model. Based on the factors we define a practical hedging procedure within a local volatility framework. The hedging performance is evaluated using DAX index options.

JEL classification codes: G11

Keywords: implied volatility surface, smile, local volatility, exotic options, semiparametric factor model, hedging

1 Introduction

In equity derivative markets barrier options are appealing instruments for investors looking for a partial protection of their equity allocation. From the perspective of an institution issuing barrier options this demand raises the need of efficient hedging strategies. This is a challenging task for at least two reasons. First, reverse barrier options, such as down-and-out puts and up-and-out calls, have discontinuous payoff profiles and knock out deep in-the-money thereby loosing the maximum possible intrinsic value. Second, barrier options, as many other exotic options, are exposed to nontrivial volatility risk, since the knock-out probability strongly depends on the skew of the implied volatility smile. The latter effect also prevents simple Black-Scholes type formulae, such as those by Rubinstein and Reiner (1991), from being usable in practice.

Nowadays there is a plethora of models available that take the shape of the implied volatility surface (IVS) into account for option valuation. Potential candidates are: the local volatility (LV) model proposed by Dupire (1994), Derman and Kani (1994), and Rubinstein (1994), which introduces a nonparametric local volatility function that deterministically depends on the asset price and time; stochastic volatility models like Hull and White (1987), Stein and Stein (1991), Heston (1993), Carr et al. (2003); jump-diffusion models, such as Merton (1976), Bates (1996), and Kou (2002). When calibrated to the IVS, all these models are able to replicate the plain vanilla market to a similar extent, whereas their prices for barrier options may differ due to the different properties of the underlying asset price dynamics, see Hull and Suo (2002) and Hirsa et al. (2003) on model risk for barrier options. The more challenging part is hedging. For it is straight forward to compute derivatives for the parameters of these models, but it is intricate to give the parameter greeks a meaning by mapping them on tradable instruments provided by the plain vanilla market. More seriously, since the prices of the hedging instruments, either over-the-counter or as listed options, are given in terms of implied volatility, they necessarily follow the dynamics of the IVS. Indeed it is in question whether the IVS dynamics inherent in the model that is calibrated to a static surface and used for pricing truly match the stylized facts of IVS dynamics, see Hagan et al. (2002) and Bergomi (2005) for such a discussion in context of the LV model and the Heston model, respectively. In contrast, the dynamics of the IVS are empirically well understood, see Skiadopoulos et al. (1999), Alexander (2001), Cont and da Fonseca (2002), Fengler et al.

(2003), Hafner (2004), Fengler et al. (2007) among others. The typical approach extracts the main driving factors like level, slope, or term structure movements and models these factors. It therefore appears natural to exploit this knowledge for hedging and portfolio risk management.

The aim of this paper is to study dynamic hedges of reverse barrier options built on factor functions of empirically observed IVS dynamics. We project the complex, high dimensional dynamics of the IVS on a low and finite dimensional space spanned by the semiparametric factor model (SFM)

$$\widehat{\sigma}_t(\kappa,\tau) = \exp\left\{\sum_{l=0}^{L} Z_{t,l}\, m_l(\kappa,\tau)\right\}, \tag{1}$$

where $\widehat{\sigma}_t(\kappa,\tau)$ denotes the implied volatility of a certain moneyness κ and maturity τ observed in time t. The functions m are nonparametric components and invariant in time, while the time evolution is modelled by the latent factor series $Z_{t,l}$. In order to estimate (1) we apply an estimation technique suggested in Fengler et al. (2007). The SFM estimates the prevalent movements of the IVS in an $(L+1)$-dimensional function space.

Given the estimated factor functions $\widehat{m}$, we construct hedges for barrier options priced in a LV model. We use a LV model, since by the nonparametric nature of the local volatility function it can match any arbitrage-free set of option prices to an arbitrarily precise degree. It will hence replicate the deformations of the IVS defined by the estimated factor functions and allow for a precise computation of factor greeks not prone to calibration error. Moreover, the LV model is numerically very efficient and allows for fast and accurate price valuations using the finite difference method. The factor hedges we obtain are more general than the usual vega hedges which are defined by a parallel shift of the IVS since they will take into account nontrivial surface movements, such as nonparallel up-and-down shifts, slope and term structure risks. Depending on the payoff profile of an exotic option, these risks can be substantial. Our approach is hence similar in spirit to Diebold et al. (2006) who define factor based duration measures and study the efficacy of these measures for the insurance of bond portfolios.

We note that strictly speaking it may not be necessary to vega hedge in an LV framework,

since it defines a complete market. This however is a theoretical perspective which does not correspond to market practice. When minimizing portfolio risk, traders are likely to set up vega hedges as soon as a liquid over-the-counter or listed option markets allow them to do so. In this sense our approach is e.g. similar to the practice of hedging a long dated plain vanilla option which are priced by means of a smile-adjusted Black-Scholes model by adding a short dated option to the portfolio.

The dynamic hedging performance of plain vanilla options in a LV model is studied in Dumas et al. (1998), Coleman et al. (2001), McIntyre (2001) and Vähämaa (2004), while the case of reverse barrier options is treated in Engelmann et al. (2006). Engelmann et al. (2006) implement hedging strategies that are delta ($\partial/\partial S$), vega ($\partial/\partial\sigma$) and vanna ($\partial^2/\partial\sigma\partial S$) neutral where vega and vanna are obtained by parallel shifts of the IVS and computing the difference quotient. We complement this analysis by defining sensitivities with respect to the most prevalent IVS movements motivated by model (1), namely ($\partial/\partial Z_1$), ($\partial/\partial Z_2$) and by constructing portfolios neutral to these greeks. For this purpose we establish a portfolio containing a reverse barrier option and hedge it on a daily basis with plain vanillas and the underlying asset using DAX data from January 3rd, 2000 to June 30th, 2004. We then study the distribution of the hedging errors across the different hedging strategies.

For completeness we remark that static hedging of barrier options is a competing way of portfolio insurance, see Derman et al. (1995), Carr and Chou (1997), Carr et al. (1998), Andersen et al. (2002), Tompkins (2002), Nalholm and Poulsen (2006a), Nalholm and Poulsen (2006b). For a static hedge one sets up a portfolio of plain vanillas which replicates the payoff of the barrier option as close as possible. The hedge is unwound in case of a knock-out or at expiry and no other adjustment of the hedge is necessary. In fact, Engelmann et al. (2007) and Maruhn et al. (2008) show that there are static hedges outperforming dynamic hedges. However, the practical use of static hedges is limited, since they may not always be implementable due to insufficient market depth of listed plain vanilla options.

The paper is structured as follows. In Section 2 we present the framework on which the empirical procedure is based. Section 3 concentrates on the description of the hedging method. In Section 4 we present the data, describe the empirical hedging design and discuss the empirical results. Section 5 concludes.

2 Models

2.1 Local Volatility Model

In the LV model the risk neutral price of the underlying asset is governed by the stochastic differential equation:

$$dS_t = r_t S_t dt + \sigma(S_t, t) S_t dW_t, \tag{2}$$

where W_t is a Wiener process and r_t denotes the instantaneous interest rate. Dividends are assumed to be zero, since the DAX, on which our empirical study is based, is a performance index. $\sigma(S_t, t)$ is the local volatility function which depends on the underlying price and time. This function has a unique representation if an arbitrage-free set of call options is given for all strikes and maturities, Dupire (1994). It can be shown that

$$\sigma^2(S_t,t) = \left. \frac{2\frac{\partial \widehat{\sigma}(K,T)}{\partial T} + \frac{\widehat{\sigma}(K,T)}{T} + 2K\int_0^T r_s ds \frac{\partial \widehat{\sigma}(K,T)}{\partial K}}{K^2 \left\{ \frac{\partial^2 \widehat{\sigma}(K,T)}{\partial K^2} - d_1\sqrt{T}\left(\frac{\partial \widehat{\sigma}(K,T)}{\partial K}\right)^2 + \frac{1}{\widehat{\sigma}(K,T)}\left(\frac{1}{K\sqrt{T}} + d_1 \frac{\partial \widehat{\sigma}(K,T)}{\partial K}\right)^2 \right\}} \right|_{K=S_t, T=t} \tag{3}$$

where $d_1 = \frac{\log(S_0/K) + \int_0^T r_s ds + 0.5\widehat{\sigma}^2(K,T)T}{\widehat{\sigma}(K,T)\sqrt{T}}$ and where $\widehat{\sigma}(K,T)$ is the implied volatility at strike K and expiry T. Formula (3) gives a correspondence between local and implied volatility surfaces.

The LV model received much attention in the finance community since it achieves an almost exact fit of the observed vanilla market and is numerically and computationally very tractable. The price of the barrier option denoted by V with barrier B and expiry date T is obtained by numerically solving the partial differential equation

$$r_t V(S,t) = \frac{\partial V(S,t)}{\partial t} + \frac{1}{2}\sigma^2(S,t)S^2 \frac{\partial V(S,t)}{\partial S^2} + r_t S \frac{\partial V(S,t)}{\partial S} \tag{4}$$

with additional boundary conditions, i.e. $V(B,t) = 0$ for $t < T$ and $V(S,T)$ equal to the payoff at expiry. For calibration of the model a number of methods are available, see Bouchouev and Isakov (1999) for comprehensive review. For example one may directly apply

the formula (3). Here we adopt the approach of Andersen and Brotherton-Ratcliffe (1997) which determines r and σ so that forwards, zero coupon bonds and plain vanilla options are priced correctly on each grid point. The finite difference method then gives barrier option prices and sensitivities very efficiently.

Yet the LV is also subject to criticism, see Fengler (2005, Chapter 3.11) for the details of this discussion. The severest objection was brought forward by Hagan et al. (2002) by showing that the LV model implies unrealistic smile dynamics and consequently wrong spot greeks. In practice this problem can be addressed by enforcing the desired smile dynamics when computing the greeks. Instead of calculating model-consistent LV greeks, one fixes the IVS in strikes (sticky-strike) or in moneyness (sticky-moneyness) and recalibrates the LV surface under the spot movements. Engelmann et al. (2006) find that the empirical performance of the dynamic hedges is negligible under different stickiness assumptions, if a vega hedge is implemented. Overall they find that the sticky-strike approach, which we will adopt here, performs best. We therefore believe that the LV model serves well for the purpose of this study.

2.2 The Semiparametric Factor Model

To model the IVS dynamics we employ the SFM which yields estimates of the IVS for each day of the sample and explains its dynamic behavior by extracting a small number of key driving factors of the surface movements. For this aim one could use any other factor model like the functional principal components model of Cont and da Fonseca (2002) or the parametric model of Hafner (2004). An alternative definition of the skew shifts can be also found in Taleb (1997). Our choice for the SFM is motivated by the flexible nonparametric structure, which allows to extract the most important factors along with a dimension reduction, and its adaptedness to the expiry behavior of implied volatility data, see Fengler et al. (2007) for details.

To describe the SFM denote by $Y_{t,j}$ the log-implied volatility observed on day $t = 1, \ldots, T$. The index $j = 1, \ldots, J_t$ counts the implied volatilities observed on day t. Let $X_{t,j}$ be a two-dimensional variable containing (forward) moneyness $\kappa_{t,j}$ and time to maturity $\tau_{t,j}$. We define the moneyness $\kappa_{t,j} \stackrel{\text{def}}{=} K_{t,j}/F_{\tau_{t,j}}$, where $K_{t,j}$ is a strike and $F_{\tau_{t,j}}$ the forward price of

the underlying asset at time t. The SFM regresses $Y_{t,j}$ on $X_{t,j}$ by:

$$Y_{t,j} = \sum_{l=0}^{L} Z_{t,l}\, m_l(X_{t,j}) + \varepsilon_{t,j}, \tag{5}$$

where m_l $(l = 1, ...L)$ are nonparametric components and the $Z_{t,l}$ form a latent factor series depending on time t. The estimation error is denoted by $\varepsilon_{t,j}$. The basis functions $m_0, \ldots, m_L$ are constant in time, while the dynamic propagation of the IVS is modelled by the time varying weights $Z_{t,l}$.

The estimation procedure is based on minimizing the following least squares criterion ($\widehat{Z}_{t,0} \equiv 1$ for identification):

$$\sum_{t=1}^{T} \sum_{j=1}^{J_t} \int \left\{ Y_{t,j} - \sum_{l=0}^{L} \widehat{Z}_{t,l} \widehat{m}_l(u) \right\}^2 K_h(u - X_{t,j})\, du, \tag{6}$$

where K_h denotes a two-dimensional kernel function. A possible choice for a two-dimensional kernel 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 $k_h(v) = h^{-1}k(h^{-1}v)$ is a one dimensional kernel function. The minimization procedure searches across all functions $\widehat{m}_l : \mathbb{R}^2 \longrightarrow \mathbb{R}$ $(l = 0, ..., L)$ and time series $\widehat{Z}_{t,l} \in \mathbb{R}$ $(t = 1, ..., T; l = 1, ..., L)$. Details concerning the estimation algorithm can be found in Fengler et al. (2007) and Park et al. (2009). In the final step of the procedure one orthogonalizes the functions $\widehat{m}_1, \ldots, \widehat{m}_L$ and orders them with respect to the variance explained. As a consequence the largest portion of variance is explained by the quantity $\widehat{Z}_{t,1}\widehat{m}_1$ and the second largest by $\widehat{Z}_{t,1}\widehat{m}_1 + \widehat{Z}_{t,2}\widehat{m}_2$ and so forth.

In order to illustrate the decomposition of the IVS dynamics achieved by the SFM we present in Figure 1 the results on DAX option data from January 3rd, 2000 till June 30th, 2004. The figure presents the estimated $\widehat{Z}_{t,l}$ time series in the upper panel and the estimates of the basis functions in the lower panel. The function $\widehat{m}_0$ is not presented to save space. It has no effect on the dynamics of the IVS but has to be included to set the correct level of the surface. The function $\widehat{m}_1$ is relatively flat and corresponds to the most important shocks. Changes in $\widehat{Z}_{t,1}$ result in up-and-down type of movements of the whole surface, but the deviations from a

flat basis function give different weight for each maturity-moneyness location. This effect is illustrated in Figure 2, where we plot several surfaces and one particular smile with different values of $\widehat{Z}_{t,1}$. The second factor function can be interpreted as a tilting of the smile. This can be inferred from the shape of $\widehat{m}_2$ and its influence on the IVS in the plots. The variation in $\widehat{Z}_{t,2}$ results in changing the slope of the smile by making it steeper or flatter while keeping roughly the same implied volatility levels.

We finally remark that the SFM has spurred further research on IVS dynamics and beyond. Brüggemann et al. (2008) study the statistical properties of the estimated factor series using a vector autoregressive framework and analyze the associated movements of macroeconomic variables. Giacomini and Härdle (2008) apply the modelling idea for an explanation of the dynamics of risk neutral densities. The CO_2 allowance term structure is studied in Trück et al. (2006) and electricity forward curves in Borak and Weron (2009).

3 Hedging Framework

Dynamic hedging of the asset V, in our case the reverse barrier option, is based on frequent adjustments of the hedge portfolio. This hedging strategy requires to construct a portfolio which is to first (or higher) order neutral to the relevant risk factors. Apart from standard delta hedging, a successful strategy requires hedging the vega, and possibly higher order greeks as pointed out by Ederington and Guan (2007).

For the LV framework Engelmann et al. (2006) study delta, delta-vega and delta-vega-vanna hedges. One knock-out option is hedged with the underlying asset and a set of plain vanilla options. Let the value of the barrier option be denoted by V and let HP_1 and HP_2 be portfolios of plain vanilla options. The corresponding hedge ratios are then given by solving

$$\begin{pmatrix} 1 & \frac{\partial HP_1}{\partial S} & \frac{\partial HP_2}{\partial S} \\ 0 & \frac{\partial HP_1}{\partial \widehat{\sigma}} & \frac{\partial HP_2}{\partial \widehat{\sigma}} \\ 0 & \frac{\partial^2 HP_1}{\partial \widehat{\sigma} \partial S} & \frac{\partial^2 HP_2}{\partial \widehat{\sigma} \partial S} \end{pmatrix} \cdot \begin{pmatrix} a_0 \\ a_1 \\ a_2 \end{pmatrix} = \begin{pmatrix} \frac{\partial V}{\partial S} \\ \frac{\partial V}{\partial \widehat{\sigma}} \\ \frac{\partial^2 V}{\partial \widehat{\sigma} \partial S} \end{pmatrix}. \tag{7}$$

Equation (7) reflects the full delta-vega-vanna hedge. Putting $a_2 = 0$ reduces (7) to the

delta-vega hedge and $a_1 = a_2 = 0$ to the pure delta hedge. Since good hedges have a large exposure to the risk factors to be hedged, one could use an at-the-money plain vanilla option for the HP_1 and for HP_2 a risk reversal. A risk reversal is a combination of a long out-of-the-money call and a short out-of-the-money put (or vice versa).

In order to compute the sensitivities one reprices the option under different scenarios and computes the greeks by a finite difference quotient. Following Engelmann et al. (2006), we make a sticky strike assumption for our greeks, i.e. the IVS remains constant in strikes. Vega and vanna are computed shifting the IVS in a parallel fashion. To be more specific, we compute

$$\frac{\partial V}{\partial S} \overset{\text{def}}{\approx} \frac{V\left(S+\Delta S,\widehat{\sigma}\right)-V\left(S-\Delta S,\widehat{\sigma}\right)}{2\Delta S}, \tag{8}$$

$$\frac{\partial V}{\partial \widehat{\sigma}} \overset{\text{def}}{\approx} \frac{V\left(S,\widehat{\sigma}+\Delta\widehat{\sigma}\right)-V\left(S,\widehat{\sigma}-\Delta\widehat{\sigma}\right)}{2\Delta\widehat{\sigma}}, \tag{9}$$

$$\begin{aligned}\frac{\partial^2 V}{\partial S\partial\widehat{\sigma}} \overset{\text{def}}{\approx} & \left\{V\left(S+\Delta S,\widehat{\sigma}+\Delta\widehat{\sigma}\right)-V\left(S+\Delta S,\widehat{\sigma}\right)\right. \\ & \left.-V\left(S-\Delta S,\widehat{\sigma}+\Delta\widehat{\sigma}\right)+V\left(S-\Delta S,\widehat{\sigma}\right)\right\}/(2\Delta S\Delta\widehat{\sigma}).\end{aligned} \tag{10}$$

With small abuse of notation $V(S,\widehat{\sigma})$ denotes here the price obtained with spot S and IVS $\widehat{\sigma}$, where we omit its arguments for simplicity. $\widehat{\sigma}+\Delta\widehat{\sigma}$ means the parallel shift of the whole surface.

It is empirically widely confirmed that parallel shifts are the most prevalent movements of the IVS. It would be misleading, however, to conclude from this observation that other types of surface variations do only negligibly influence the prices of exotic derivatives, such as barrier options. Contrariwise a higher slope leads to a smaller price of an in-the-money down-and-out put. Consider an artificial example of two one year down-and-out put with strike 110, barrier 80 at the current spot level of 100. The first option is priced with the IVS observed on January 3rd, 2000 and the second one on January 2nd, 2001. Figure 3 shows the surfaces of these days. The LV prices of these options are 1.91% and 2.37% respectively (in percentage of the spot price), which is quite a difference. From the upper panel of Figure 1 one observes that the level related factor assumes similar values on these days, while the slope factor differs significantly. This price discrepancy stems mainly from the slope effect,

which is an exposure not directly hedged in traditional approaches. Our procedure will hedge such volatility shocks.

In our hedging framework we define new sensitivities with respect to the variation of the (log)-IVS, which we call ζ-greeks. Based on the results discussed in Section 2.2, the ζ_1-greek ($\partial/\partial Z_{t,1}$) reflects an adjusted up-and-down shift, while the ζ_2-greek ($\partial/\partial Z_{t,2}$) corresponds to the slope effect. Similarly to (7) we obtain the hedge ratios by

$$\begin{pmatrix} 1 & \frac{\partial HP_1}{\partial S} & \frac{\partial HP_2}{\partial S} \\ 0 & \frac{\partial HP_1}{\partial Z_{t,1}} & \frac{\partial HP_2}{\partial Z_{t,1}} \\ 0 & \frac{\partial HP_1}{\partial Z_{t,2}} & \frac{\partial HP_2}{\partial Z_{t,2}} \end{pmatrix} \cdot \begin{pmatrix} a_0 \\ a_1 \\ a_2 \end{pmatrix} = \begin{pmatrix} \frac{\partial V}{\partial S} \\ \frac{\partial V}{\partial Z_{t,1}} \\ \frac{\partial V}{\partial Z_{t,2}} \end{pmatrix}. \tag{11}$$

We call the full setting a $\zeta_1\zeta_2$-hedge, the reduced one with $a_2 = 0$ a ζ_1-hedge. As in the traditional hedge we use an at-the-money plain vanilla for HP_1, again due to the high vega. For HP_2, we employ risk reversals because they primarily respond to changes in the wings of the IVS. Moreover, by selecting appropriate strikes it can even be set up in a vega-neutral, i.e. ζ_1-neutral, way.

We calculate the ζ-greeks by means of a difference quotient. As pricing input for the barrier options we do not use the estimate of the IVS obtained by the SFM, as it is necessarily subject to an estimation error. Instead, in order to avoid mispricings, we use the truly observed ones. Thus, by the definition of the ζ-greeks, the approximations are given by

$$\frac{\partial V}{\partial Z_{t,l}} \overset{\text{def}}{\approx} \frac{V\left(S, \widehat{\sigma}\exp(\Delta Z_{t,l}\widehat{m}_l)\right) - V\left(S, \widehat{\sigma}\exp(-\Delta Z_{t,l}\widehat{m}_l)\right)}{2\Delta Z_{t,l}}. \tag{12}$$

In the practical implementation of (12) one faces a couple of numerical issues, which need to be addressed. First, the size of the $\Delta Z_{t,l}$ has to be chosen. An increment too small or too large can distort the meaning of the greeks. Moreover it cannot be unique for all $Z_{t,l}$, since the shift size depends on the basis functions $\widehat{m}_l$ and on the IVS on a particular day. Therefore we choose for each t a $\Delta Z_{t,l}$ such that the (absolute) mean upward (downward) shift amounts approximately to one volatility-point. Note that we do *not* use $\widehat{Z}_{t,l}$ for these perturbations. Another challenge is an accurate calculation of the barrier greeks. To reduce

numerical errors we employ a constant grid in the pricing algorithm for calculating the ζ-greeks. Furthermore, the IVS $\widehat{\sigma}$ needs to be arbitrage-free. However, the shifted surfaces do not necessarily possess this property. We thus additionally check no-arbitrage conditions before calculating the ζ-greeks and apply an algorithm due to Fengler (2008) in case of violations. This method estimates the option price function by means of a natural smoothing spline under no-arbitrage constraints, i.e. under convexity, monotonicity and bounds on the price function and on the first order strike derivatives. The resulting estimate is then converted back to implied volatility. The algorithm is not applied when computing vega and vanna since parallel shifts do typically not result into arbitrage violations.

The aforementioned greeks are demonstrated in Figure 4 for the down-and-out put with half a year to expiry. The plot displays the greeks as a function of spot and keeps other characteristics of the barrier option unchanged. It has to be noted that the SFM, i.e. $\widehat{Z}_{t,l}$ and $\widehat{m}_l$, can only be identified up to sign. The sign of the ζ-greeks therefore has no particular meaning. Hence vega and ζ_1 display similar patterns. For the spot values close to the barrier level vega is negative and approaches zero as it becomes a delta product. For out-of-the money options vega is positive since the option then resembles a plain vanilla contract. A similar behavior is observed for ζ_2 and vanna, but the vanna is discontinues at the barrier as it is derived from the delta.

4 Empirical Results

4.1 Data

The data set covers DAX index options traded at the EUREX from January 3rd, 2000 till June 30th, 2004 which give 1135 trading days. We use settlement prices, which are prices published by the EUREX based on the last intra-day trades. The DAX index is a capital weighted performance index comprising 30 German blue chips. Since dividends less corporate tax are reinvested into the index, they do not need to be taken into account for option valuation.

We preprocess the data by eliminating implied volatilities bigger than 80% and maturities

smaller than 10 days. Arbitrage violations in the option data are removed by the arbitrage free smoothing procedure described in Fengler (2008). After smoothing, the data are converted into a regular grid of moneyness and time to maturities. For option pricing, the zero rates from EURIBOR quotes are linearly interpolated, see Dumas et al. (1998) for this practice.

4.2 Experimental Design

In our empirical study we assume no transaction costs, no restrictions on short selling and the possibility of trading each asset at arbitrary size. Each security is priced using the LV model calibrated to daily market data. We implement the hedging strategies described in Section 3, i.e. we focus exclusively on volatility and spot risks, leaving other risks like interest rate exposure unhedged.

In the first step of our experiment we estimate the SFM. As kernel function we use a product quartic kernel, where $k(u) = 15/16(1-u^2)^2$ for $|u| < 1$ and 0 otherwise. For a data driven bandwidth choice and the model size selection, we refer to Fengler et al. (2007). The basic idea is to estimate the model for different combinations of L and h and compare various information criteria. For the moneyness direction we finally use a bandwidth of 0.04, but we slightly oversmooth the surfaces in the time to maturity direction in order to reduce numerical errors for the subsequent price computations. More precisely, we use a local bandwidth modelled by an arctangent function which increases monotonously from 0.02 to 0.15 (expressed in years). Since in the hedging procedure only two main factors are included, we set $L = 2$. With this choice the model describes sufficiently well the IVS dynamics, since the measure of explained variation is close to 98%.

For each day up to one year before the last observation date in the sample, a long position in the reverse barrier option is created. This is to evaluate all initiated hedges at market prices within the sample. We use up-and-out calls with strikes at 80% of the spot and barriers at 140% and down-and-out put with strikes at 80% and barriers at 110%. These specifications correspond to typically traded contracts. Based on the calibrated LV model, ζ-greeks, delta, vega and vanna are calculated and the hedging strategies as described in Section 3 are set up. We concentrate on vega, vanna, ζ_1 and $\zeta_1\zeta_2$ strategies since the pure delta hedge is of

inferior quality. As HP_1 we use at-the-money puts for the up-and-out calls and at-the-money calls for the down-and-out puts. The risk reversal are structured by taking 80% and 120% strikes of the current spot.

Positions that have not knocked are updated on a daily basis. This choice is motivated by the results of Engelmann et al. (2006) who do not obtain different rankings of the strategies for other re-balancing frequencies. For each day we calculate the greeks to solve (7) and (11) and adjust the hedge ratios a_0, a_1, a_2. The hedges are financed from the cash account and if the barrier is breached or the barrier option expires we unwind the hedge and record the hedging error. All positions are traded at market prices. In case of a knock-out event, the hedging error pays or earns interest until expiry in order to render the results comparable. Also the cash account bears interest or is financed at the riskless short rate of the concurrent trading day. Summing up, we have a collection of hedging errors for the two types of barrier options with four different hedging strategies for each of them.

One could object that the experimental design suffers from an in-sample problem, since the SFM is estimated on the same data set as the hedging experiment. It is however a common finding in the empirical literature, either on interest rates or on the IVS, that eigenvectors or eigenfunctions are remarkably stable across time. Formal tests on IVS data between the years 1995 to 2001 confirming this hypothesis are provided by Fengler (2005, Chapter 5.2.3). Even if we made use of a training-sample, we would therefore recover very similar factor functions. Thus the issue will not seriously affect the results.

4.3 Results

For evaluating the performance we use a pool of 885 hedging errors (1135 trading days less 250 days, since products issued thereafter would not expire within the sample). In order to make them comparable we normalize by the spot price at the time when the hedge is initiated. This normalization is common in practice and is meant to remove the dependence from the underlying's level. Another normalizing factor could be the option price itself, but since the risk reversal has a market price close to zero, measuring errors with respect to the spot appears to be more natural.

The aim of hedging is to replicate the payoff of the option. In the ideal case the hedge portfolio should have zero variance and zero mean, but for obvious reasons this cannot be realized in practice. Our aim is to give a comparative analysis of the hedging error distributions in order to check how the volatility factors affect the hedging performance. We use traditional descriptive statistics to assess the location and dispersion of the errors. Clearly, a superior method would keep these quantities close to zero in absolute terms.

The empirical results are summarized in Tables 2 and 3 for up-and-out calls and down-and-out puts respectively. We present the minimum, maximum, mean, median, standard deviation, and the absolute deviation around the median. The terminal hedging error distributions are given in the rows marked with a '0'. As can be inferred from the tables, the center of all distributions is located around zero, with means slightly below zero for the up-and-out calls and slightly above zero for the down-and-out puts. Thus the different hedges are hardly distinguishable in terms of the center of the distribution. This finding corresponds to our expectations: the volatiliy risk is removed, both for the vega and the ζ_1-hedges, and vanna and $\zeta_1\zeta_2$-hedges do not add any additional drift, since they are almost costless.

For evaluating the dispersion of the hedging errors we focus on the standard deviation and the absolute deviation around the median (madev.). The first observation is that hedges relying on higher order greeks tend to exhibit lower variance. In case of the down-and-out puts the vanna hedge has a slightly smaller dispersion than the $\zeta_1\zeta_2$-hedge, and the traditional vega hedge performs very similar to the ζ_1-hedge. For the up-and-out calls the ranking is reversed: the standard hedges are clearly outperformed by the factor hedges. How can this asymmetry be explained and how is the quality of the factor hedges to be judged?

There are two major sources of bias in the hedging strategies due to the behavior of the underlying. Observe that during the analyzed time period the DAX had a downward trend: 81% out of the down-and-out put options knocked out, but only 10% of the up-and-out call options, while 5% of the puts and 39% calls expired in-the-money, see Table 1. As a first issue consider the huge amount of up-and-out calls ending in-the-money. This gives rise to what is known among practitioners as 'theta risk'. For explanation reconsider the case in Section 3, where we demonstrated that the prices for one-year down-and-out puts with a strike of 110% and barrier at 80% were less than 3% in the two scenarios. In contrast, when the put ends in-the-money it will pay out up to 30%. Consequently, the value of an in-the-money reverse

barrier option increases sharply the nearer the expiry date draws (i.e. has a strong theta), rendering it more and more difficult for traders to earn the payoff by trading the gamma. Theta risk can thus lead to a more dispersed error distribution. A second issue is gap risk. We do not unwind the hedges at the barriers, but at the observed spots, since this is the more realistic scenario in practice. When a barrier is breached, one still owns the hedge and incurs unbalanced gains or losses. Again this leads to a more dispersed hedging error distribution. As is clear from Table 1, theta risk is dominating the risk in case of the calls and gap risk in case of the puts.

To receive a deeper insight, we refer once more to Tables 2 and 3. We report the statistics of the hedging experiment stopped at 1 day, 5 days and 25 days before the expiry. As is seen the dispersion measures increase the nearer expiry draws, and the distributions become less skewed and less heavy-tailed, while the location measures prove to remain stable. In terms of dispersion the relative order of the hedging strategies across the two products remains the same: for the down-and-out puts the strategies are comparable, while factor hedging remains superior for the up-and-out calls. This finding is confirmed in Figure 5, which displays the standard deviations of the hedging errors as a function through the options' life time. It is intuitive to expect this function to increase. Moreover there is a sharp jump just before the expiry date contributing a large portion of the overall cumulative hedging error in particular for the up-and-out calls. All these observations highlight the importance of the expiry effect relative to gap risk when interpreting the data.

We overall conclude two main findings. First, factor hedging is at least of similar quality as traditional hedging approaches. In particular the hedging efficiency does not deteriorate. This is a reassuring result given the huge computational effort that must be spent and that could easily come at the costs of accuracy. This result is obtained when the barrier options expire worthless or knock out early in life time. Second, when the option needs to be hedged till expiry and ends in-the-money, the factor hedging approach dominates clearly. From a trader's perspective the first situation is the 'easy one' unless the knock-out occurs close to expiry. The second one is much more intricate, because the intrinsic value needs to be earned. This is a strong case for volatility factor hedging.

5 Conclusion

We provide an empirical study on hedging reverse barrier options in the local volatility model. The main focus of this study is on risk factors arising from a decomposition of the dynamic behavior of the implied volatility surface, which are identified with a flexible semiparametric technique. The hedging framework is constructed as a natural extension to traditional vega hedging, where the sensitivity is measured with respect to the more complex surface movements.

Our empirical investigation shows that hedging higher order risk with risk reversals brings improvements to hedging with at-the-money plain vanillas only. This is consistent across the vanna hedge and the more complex factor based hedges, thus confirming evidence of Ederington and Guan (2007). Intuitively the vega hedge resembles a single factor based hedge since the first dynamic factor corresponds to a parallel type of shift. Adding a vanna hedge or another factor to the portfolio removes similar risks as can be inferred from the comparable hedging performance.

Measured in terms of the hedging error variance, factor hedging performs at least as good as the corresponding vega and vanna hedges, in certain cases it is superior. As is confirmed by hedging up-and-out call options and down-and-out put options, the first case occurs when options knock out early in life time or expire worthless, while the second occurs when the options need to be hedged up to expiry and end in-the-money. This evidence is present not only in the terminal hedging errors but also through the option's life time. From a trader's perspective the second case is the more interesting, making factor hedging a powerful alternative to traditional hedging.

These findings, however, are not necessarily similar for other complex derivatives sensitive to IVS movements, such as cliquets or long-dated forward starting options. Also a portfolio context may yield different findings. In particular, when a book of options contains assets with several maturities it could be beneficial to consider additional factors, such as those related to the term structure of the IVS. This exposure can be hedged by constructing the corresponding calendar spreads. Another application in a portfolio context could be stress test scenarios based on the volatility factors. This would provide a good understanding of

the volatility exposure of the portfolio. We leave these issues to future research.

References

Alexander, C. (2001). Principles of the skew. *RISK*, 14(1):S29–S32.

Andersen, L. B. G., Andreasen, J., and Eliezer, D. (2002). Static replication of barrier options: Some general results. *Journal of Computational Finance*, 5(4):1–25.

Andersen, L. B. G. and Brotherton-Ratcliffe, R. (1997). The equity option volatility smile: An implicit finite-difference approach. *Journal of Computational Finance*, 1(2):5–37.

Bates, D. S. (1996). Jumps and stochastic volatility: Exchange rate processes implicit in Deutsche Mark options. *Review of Financial Studies*, 9:69–107.

Bergomi, L. (2005). Smile dynamics II. *RISK*, 18(10):67–73.

Borak, S. and Weron, R. (2009). A semiparametric factor model for electricity forward curve dynamics. *The Journal of Energy Markets*, 1(3).

Bouchouev, I. and Isakov, V. (1999). Uniqueness, stability and numerical methods for the inverse problem that arises in financial markets. *Inverse Problems*, 15:R95–R116.

Brüggemann, R., Härdle, W., Mungo, J., and Trenkler, C. (2008). VAR modeling for dynamic loadings driving volatility strings. *Journal of Financial Econometrics*, 6:361–381.

Carr, P. and Chou, A. (1997). Breaking Barriers. *Risk Magazine*, 10:139–145.

Carr, P., Ellis, K., and Gupta, V. (1998). Static hedging of exotic options. *Journal of Finance*, 53(3):1165–1190.

Carr, P., Geman, H., Madan, D., and Yor, M. (2003). Stochastic volatility for Lévy processes. *Mathematical Finance*, 13:345–382.

Coleman, T. F., Kim, Y., Li, Y., and Verma, A. (2001). Dynamic hedging with a deterministic local volatility function model. *Journal of Risk*, 4(1):63–89.

Cont, R. and da Fonseca, J. (2002). The dynamics of implied volatility surfaces. *Quantitative Finance*, 2(1):45–60.

Derman, E., Ergener, D., and Kani, I. (1995). Static options replication. *Journal of Derivatives*, 2(4):78–95.

Derman, E. and Kani, I. (1994). Riding on a smile. *RISK*, 7(2):32–39.

Diebold, F., Ji, L., and Li, C. (2006). A three-factor yield curve model: Non-affine structure, systematic risk sources, and generalized duration. In Klein, L., editor, *Long-Run Growth and Short-Run Stabilization: Essays in Memory of Albert Ando*. Edward Elgar, Cheltenham, U.K.

Dumas, B., Fleming, J., and Whaley, R. E. (1998). Implied volatility functions: Empirical tests. *Journal of Finance*, 80(6):2059–2106.

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

Ederington, L. and Guan, W. (2007). Higher order greeks. *Journal of Derivatives*, 14:7–34.

Engelmann, B., Fengler, M., Nalholm, M., and Schwendner, P. (2007). Static versus Dynamic Hedges: An Empirical Comparison for Barrier Options. *Review of Derivatives Research*, 9(3):239–264.

Engelmann, B., Fengler, M., and Schwendner, P. (2006). Better than its reputation: An empirical hedging analysis of the local volatility model for barrier options. Working paper, Available at SRRN.

Fengler, M. R. (2005). *Semiparametric Modeling of Implied Volatility*. Lecture Notes in Finance. Springer-Verlag, Berlin, Heidelberg.

Fengler, M. R. (2008). Arbitrage-free smoothing of the implied volatility surface. *Quantitative Finance*. Forthcoming.

Fengler, M. R., Härdle, W., and Mammen, E. (2007). A semiparametric factor model for implied volatility surface dynamics. *Journal of Financial Econometrics*, 5(2):189–218.

Fengler, M. R., Härdle, W., and Villa, C. (2003). The dynamics of implied volatilities: A common principle components approach. *Review of Derivatives Research*, 6:179–202.

Giacomini, E. and Härdle, W. (2008). Dynamic Semiparametric Factor Models in Pricing Kernels Estimation. In Dabo-Niang, S. and Ferraty, F., editors, *Functional and Operatorial Statistics*, pages 181–187. Physica-Verlag HD.

Hafner, R. (2004). *Stochastic Implied Volatility*. Springer, Berlin.

Hagan, P., Kumar, D., Lesniewski, A., and Woodward, D. (2002). Managing smile risk. *Wilmott magazine*, 1:84–108.

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.

Hirsa, A., Courtadon, G., and Madan, D. (2003). The effect of model risk on the valuation of barrier options. *Journal of Risk Finance*, 4:47–55.

Hull, J. and White, A. (1987). The pricing of options on assets with stochastic volatilities. *Journal of Finance*, 42:281–300.

Hull, J. C. and Suo, W. (2002). A methodology for assessing model risk and its application to the implied volatility function model. *Journal of Financial and Quantitative Analysis*, 37(2):297–318.

Kou, S. G. (2002). A jump-diffusion model for option pricing. *Management Science*, 48:1086–1101.

Maruhn, J., Nalholm, M., and Fengler, M. R. (2008). Empirically robust static uncertain skew hedges for reverse barrier options. Working paper.

McIntyre, M. L. (2001). Performance of Dupire's implied diffusion approach under sparse and incomplete data. *Journal of Computational Finance*, 4(4):33–84.

Merton, R. C. (1976). Option pricing when underlying stock returns are discontinuous. *Journal of Financial Economics*, 3:125–144.

Nalholm, M. and Poulsen, R. (2006a). Static hedging and model risk for barrier options. *Journal of Future Markets*, 26:449–463.

Nalholm, M. and Poulsen, R. (2006b). Static hedging of barrier options under general asset dynamics: Unification and application. *Journal of Derivatives*, 13:46–60.

Park, B., Mammen, E., Härdle, W., and Borak, S. (2009). Time Series Modelling with Semiparametric Factor Dynamics. *Journal of the American Statistical Association*. Forthcoming.

Rubinstein, M. (1994). Implied binomial trees. *Journal of Finance*, 49:771–818.

Rubinstein, M. and Reiner, E. (1991). Breaking down the barrier. *RISK*, 4(9):28–35.

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

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

Taleb, N. (1997). *Dynamic Hedging: Managing Vanilla and Exotic Options*. John Wiley & Sons.

Tompkins, R. (2002). Static versus dynamic hedging of exotic option: An evaluation of hedge performance via simulation. *The Journal of Risk Finance*, 3:6–34.

Trück, S., Borak, S., Härdle, W., and Weron, R. (2006). Convenience yields for CO_2 emission allowance futures contracts. Discussion Paper 2006-076, SfB 649, Humboldt-Universität zu Berlin.

Vähämaa, S. (2004). Delta hedging with the smile. *Financial Markets and Portfolio Management*, 18(3):241–255.

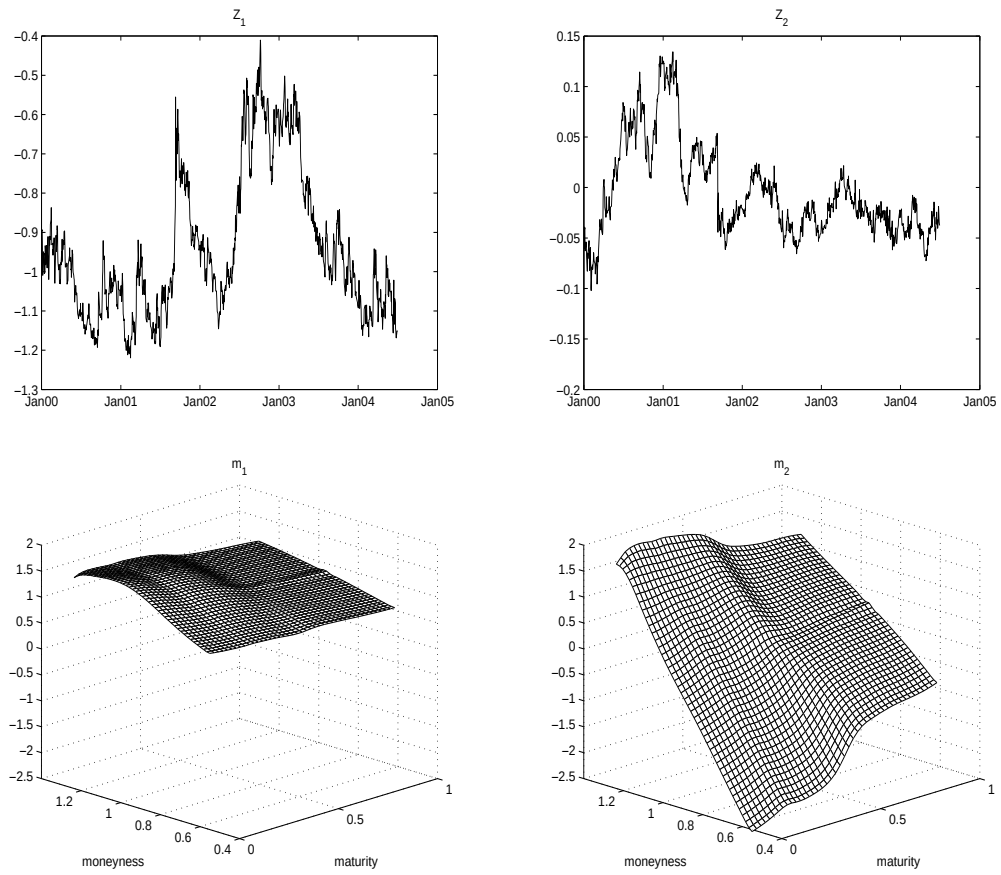


Figure 1: *The estimates of the SFM obtained from IVS data from January 3rd, 2000 till June 30th, 2004 for* $L = 2$. *Upper panel: estimated latent factor series* $\widehat{Z}_1$ *and* $\widehat{Z}_2$. *Lower panel: estimates of* $\widehat{m}_1$, *the non-uniform up-and-down shift, and* $\widehat{m}_2$, *the slope risk.*

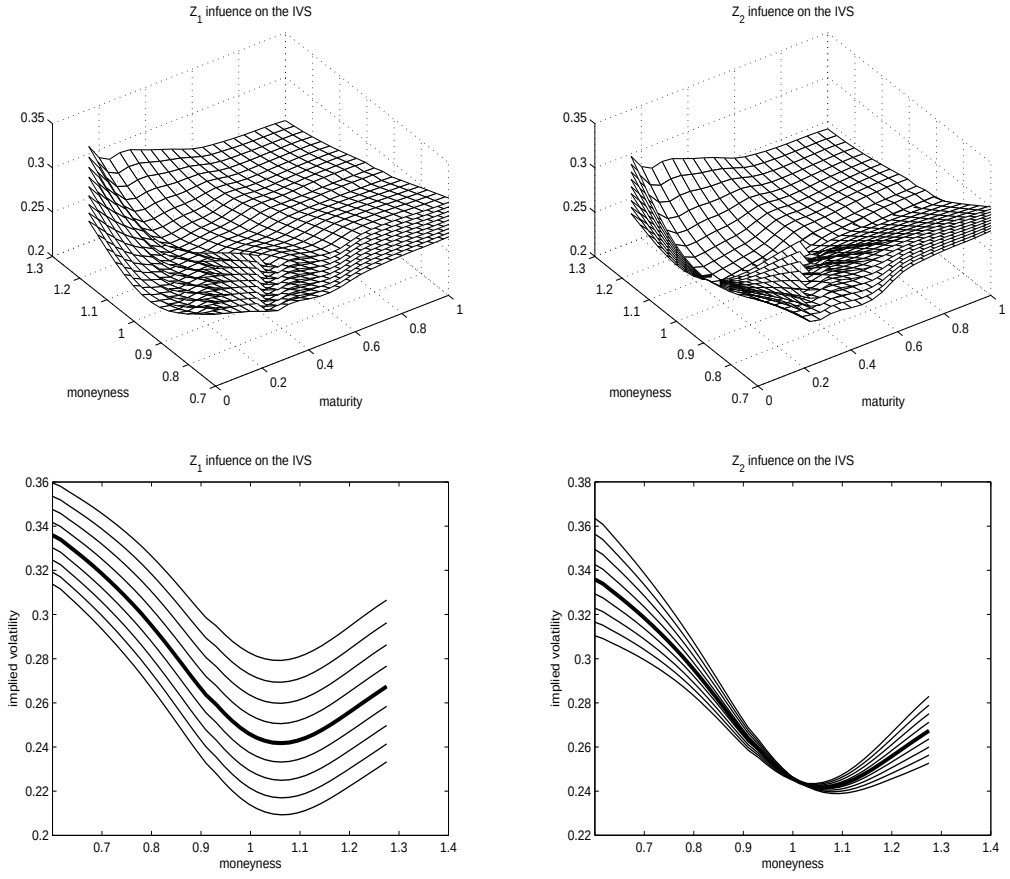


Figure 2: *Impact of $\widehat{Z}_1$ and $\widehat{Z}_2$ on the IVS. Shocks in $\widehat{Z}_1$ trigger up-and-down movements while shocks in $\widehat{Z}_2$ tilt the smile around at-the-money point. Upper panel: a visualization of the shocks for the entire surface. Lower panel: the impact presented on one particular smile.*

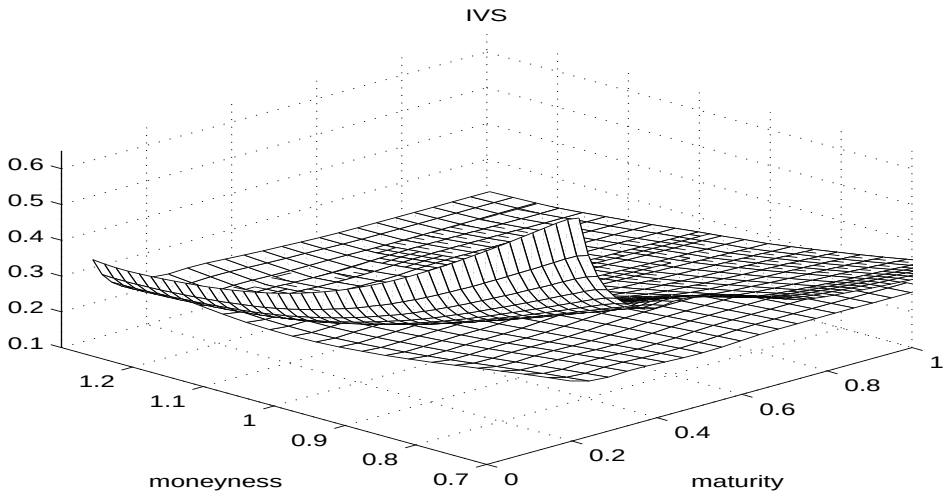


Figure 3: *IVS observed on January 3rd, 2000 (the steeper surface) and January 2nd, 2001 (the flatter one). DAX levels on these days were 6751 and 6290 respectively.*

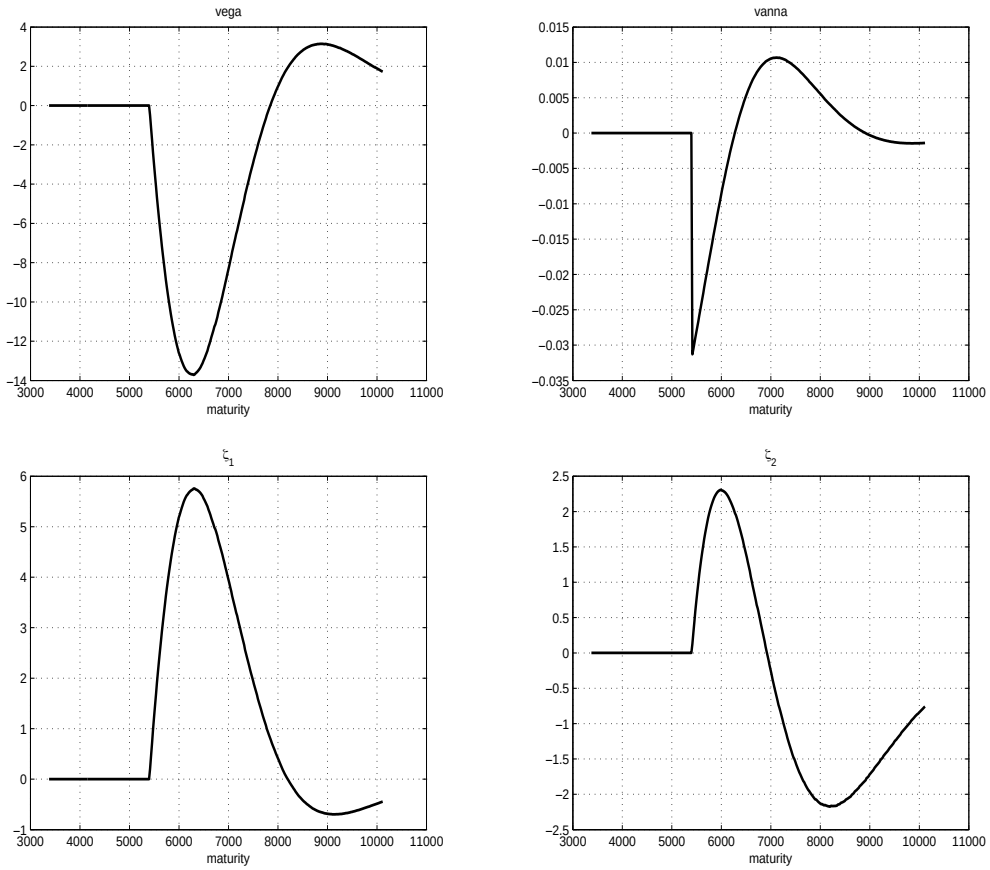


Figure 4: *Greeks for a down-and-out put option with maturity 0.5 years with barrier 5400 strike 7425 as a function of the spot. Upper left panel: vega. Upper right panel: vanna. Lower right panel:* ζ_1*. Lower right panel:* ζ_2

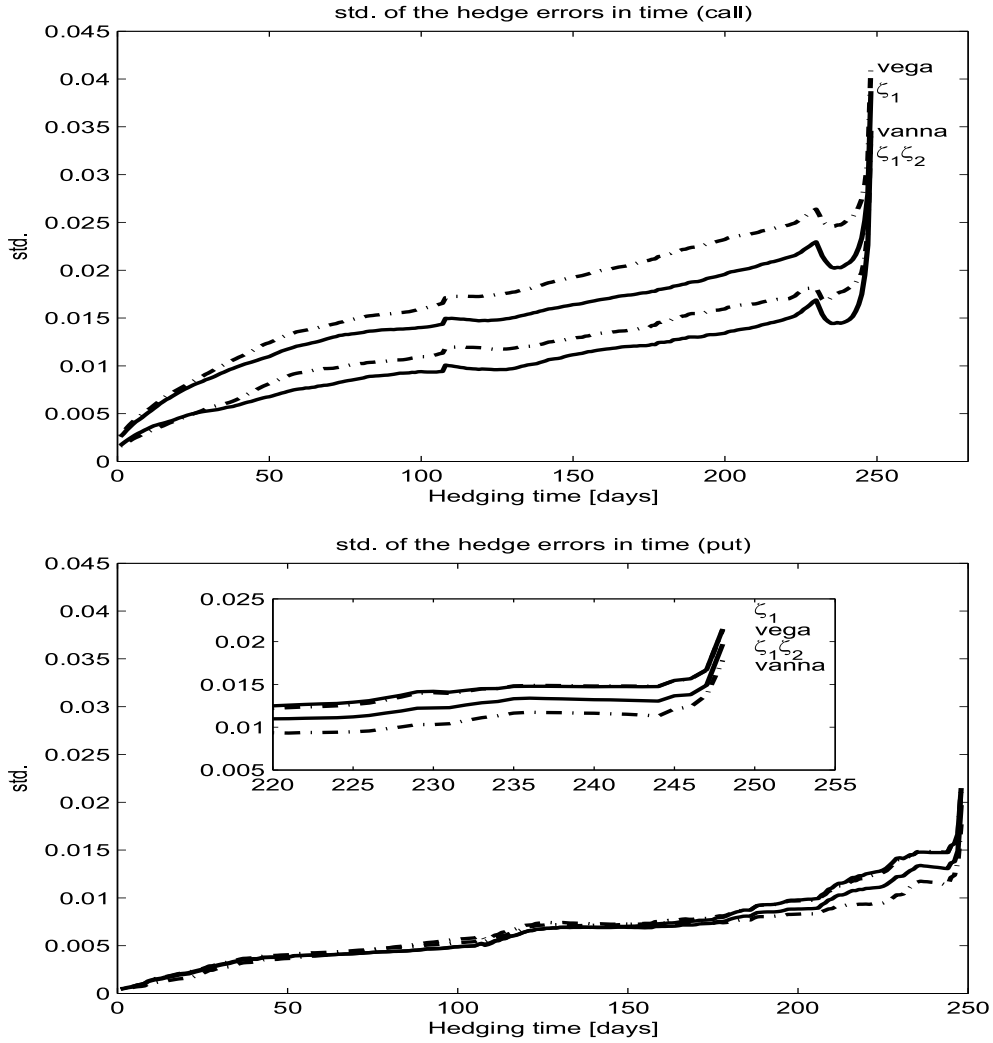


Figure 5: *Standard deviations of the hedging errors as a function of time from option issuance. Solid lines represent the factor hedging methods motivated by the SFM. Dashed lines represent the vega and vanna hedges. Upper panel: up-and-out call. Lower panel: down-and-out put.*

option type	barrier	strike	knock-outs	in-the-money
up-and-out call	140%	80%	10%	39%
down-and-out put	80%	110%	81%	5%

Table 1: *Characteristics of the analyzed barrier options. Strikes and barriers are in percentage of spot at issuance. The column 'knock-outs' refers to the contracts that breached the barrier and 'in-the-money' to those yielding a positive payoff at expiry.*

	days	min	max	mean	median	std.	madev.	skew.	kurt.
vega	0	-0.1038	0.5813	-0.0165	-0.0175	0.0413	0.0209	7.2801	97.60
	1	-0.1038	0.2581	-0.0172	-0.0174	0.0314	0.0199	2.3526	19.32
	5	-0.1037	0.0970	-0.0181	-0.0169	0.0260	0.0183	0.3636	4.91
	25	-0.0827	0.0649	-0.0174	-0.0164	0.0249	0.0178	0.0587	3.74
ζ_1	0	-0.0752	0.5768	-0.0118	-0.0136	0.0387	0.0183	8.4877	119.04
	1	-0.0751	0.2332	-0.0125	-0.0134	0.0279	0.0172	2.8026	22.27
	5	-0.0749	0.0755	-0.0134	-0.0121	0.0216	0.0155	0.2846	4.60
	25	-0.0761	0.0573	-0.0134	-0.0127	0.0215	0.0161	0.0343	3.55
vanna	0	-0.1340	0.5310	-0.0081	-0.0138	0.0345	0.0151	8.6289	124.62
	1	-0.1340	0.1842	-0.0089	-0.0136	0.0239	0.0140	2.1325	17.40
	5	-0.1339	0.0807	-0.0099	-0.0131	0.0187	0.0121	0.2157	9.54
	25	-0.0582	0.0772	-0.0096	-0.0141	0.0173	0.0118	1.3367	6.22
$\zeta_1\zeta_2$	0	-0.0830	0.5684	-0.0066	-0.0119	0.0345	0.0137	10.6470	161.00
	1	-0.0829	0.2091	-0.0073	-0.0117	0.0226	0.0126	4.0718	31.51
	5	-0.0829	0.0710	-0.0083	-0.0113	0.0157	0.0106	1.3447	7.12
	25	-0.0370	0.0629	-0.0086	-0.0118	0.0152	0.0108	1.3559	5.72

Table 2: *Hedging error distributions of the up-and-out calls. Given are descriptive statistics for the various hedging strategies. The rows present the statistics at* 0, 1, 5 *and* 25 *days before expiration.*

	days	min	max	mean	median	std.	madev.	skew.	kurt.
vega	0	-0.0264	0.2799	0.0058	-0.0004	0.0213	0.0105	5.4903	51.91
	1	-0.0756	0.1172	0.0050	-0.0004	0.0166	0.0098	2.4531	12.38
	5	-0.0187	0.0882	0.0041	-0.0008	0.0147	0.0090	2.4682	10.92
	25	-0.0186	0.0749	0.0038	-0.0004	0.0124	0.0083	2.0267	8.73
ζ_1	0	-0.0210	0.2808	0.0080	0.0016	0.0214	0.0107	5.4775	51.59
	1	-0.0702	0.1215	0.0072	0.0015	0.0167	0.0100	2.4501	12.14
	5	-0.0137	0.0882	0.0063	0.0013	0.0147	0.0091	2.4112	10.29
	25	-0.0113	0.0798	0.0059	0.0014	0.0127	0.0085	2.0632	8.54
vanna	0	-0.0608	0.2072	0.0022	-0.0016	0.0178	0.0081	5.7326	53.32
	1	-0.0955	0.1309	0.0014	-0.0016	0.0137	0.0074	2.7735	23.76
	5	-0.0323	0.0649	0.0006	-0.0018	0.0114	0.0069	1.9306	9.98
	25	-0.0205	0.0582	0.0004	-0.0016	0.0093	0.0059	2.0273	9.23
$\zeta_1\zeta_2$	0	-0.0332	0.2676	0.0065	0.0008	0.0196	0.0092	6.0258	60.47
	1	-0.0824	0.1146	0.0057	0.0008	0.0149	0.0085	2.4463	14.06
	5	-0.0234	0.0774	0.0048	0.0007	0.0130	0.0079	2.4134	10.42
	25	-0.0121	0.0727	0.0045	0.0010	0.0110	0.0069	2.3842	10.27

Table 3: *Hedging error distributions of the down-and-out puts. Given are descriptive statistics for the various hedging strategies. The rows present the statistics at* 0, 1, 5 *and* 25 *days before the expiration.*

The Econometrics Journal

Econometrics Journal (2009), volume **12**, pp. 248–271.
doi: 10.1111/j.1368-423X.2009.00292.x

Adaptive pointwise estimation in time-inhomogeneous conditional heteroscedasticity models

P. ČÍŽEK[†], W. HÄRDLE[‡] AND V. SPOKOINY[§]

[†]*Department of Econometrics & OR, Tilburg University, P.O. Box 90153, 5000LE Tilburg, The Netherlands*
E-mail: P.Cizek@uvt.nl

[‡]*Humboldt-Universität zu Berlin and CASE, Spandauerstrasse 1, 10178 Berlin, Germany*
E-mail: haerdle@wiwi.hu-berlin.de

[§]*Weierstrass-Institute, Humboldt-Universität zu Berlin and CASE, Mohrenstrasse 39, 10117 Berlin, Germany*
E-mail: spokoiny@wias-berlin.de

First version received: April 2008; final version accepted: April 2009

Summary This paper offers a new method for estimation and forecasting of the volatility of financial time series when the stationarity assumption is violated. Our general, local parametric approach particularly applies to general varying-coefficient parametric models, such as GARCH, whose coefficients may arbitrarily vary with time. Global parametric, smooth transition and change-point models are special cases. The method is based on an adaptive pointwise selection of the largest interval of homogeneity with a given right-end point by a local change-point analysis. We construct locally adaptive estimates that can perform this task and investigate them both from the theoretical point of view and by Monte Carlo simulations. In the particular case of GARCH estimation, the proposed method is applied to stock-index series and is shown to outperform the standard parametric GARCH model.

Keywords: *Adaptive pointwise estimation*, *Autoregressive models*, *Conditional heteroscedasticity models*, *Local time-homogeneity*.

1. INTRODUCTION

A growing amount of econometrical and statistical research is devoted to modelling financial time series and their volatility, which measures dispersion at a point in time (i.e. conditional variance). Although many economies and financial markets have been recently experiencing many shorter and longer periods of instability or uncertainty such as the Asian crisis (1997), the Russian crisis (1998), the start of the European currency (1999), the 'dot-Com' technology-bubble crash (2000–02) or the terrorist attacks (September, 2001), the war in Iraq (2003) and the current global recession (2008), mostly used econometric models are based on the assumption of time homogeneity. This includes linear and non-linear autoregressive (AR) and moving-average models and conditional heteroscedasticity (CH) models such as ARCH (Engel, 1982)

and GARCH (Bollerslev, 1986), stochastic volatility models (Taylor, 1986), as well as their combinations such as AR-GARCH.

On the other hand, the market and institutional changes have long been assumed to cause structural breaks in financial time series, which was confirmed, e.g. in data on stock prices (Andreou and Ghysels, 2002, and Beltratti and Morana, 2004) and exchange rates (Herwatz and Reimers, 2001). Moreover, ignoring these breaks can adversely affect the modelling, estimation and forecasting of volatility as suggested e.g. by Diebold and Inoue (2001), Mikosch and Starica (2004), Pesaran and Timmermann (2004) and Hillebrand (2005). Such findings led to the development of the change-point analysis in the context of CH models; see e.g. Chen and Gupta (1997), Kokoszka and Leipus (2000) and Andreou and Ghysels (2006).

An alternative approach lies in relaxing the assumption of time homogeneity and allowing some or all model parameters to vary over time (Chen and Tsay, 1993, Cai et al., 2000, and Fan and Zhang, 2008). Without structural assumptions about the transition of model parameters over time, time-varying coefficient models have to be estimated non-parametrically, e.g. under the identification condition that their parameters are smooth functions of time (Cai et al., 2000). In this paper, we follow a different strategy based on the assumption that a time series can be locally, i.e. over short periods of time, approximated by a parametric model. As suggested by Spokoiny (1998), such a local approximation can form a starting point in the search for the longest period of stability (homogeneity), i.e. for the longest time interval in which the series is described well by the parametric model. In the context of the local constant approximation, this strategy was employed for volatility modelling by Härdle et al. (2003), Mercurio and Spokoiny (2004) and Spokoiny (2009a). Our aim is to generalize this approach so that it can identify intervals of homogeneity for any parametric CH model regardless of its complexity.

In contrast to the local constant approximation of the volatility of a process (Mercurio and Spokoiny, 2004), the main benefit of the proposed generalization consists in the possibility to apply the methodology to a much wider class of models and to forecast over a longer time horizon. The reason is that approximating the mean or volatility process by a constant is in many cases too restrictive or even inappropriate and it is fulfilled only for short time intervals, which precludes its use for longer-term forecasting. On the contrary, parametric models like GARCH mimic the majority of stylized facts about financial time series and can reasonably fit the data over rather long periods of time in many practical situations. Allowing for time dependence of model parameters offers then much more flexibility in modelling real-life time series, which can be both with or without structural breaks since global parametric models are included as a special case.

Moreover, the proposed adaptive local parametric modelling unifies the change-point and varying-coefficient models. First, since finding the longest time-homogeneous interval for a parametric model at any point in time corresponds to detecting the most recent change-point in a time series, this approach resembles the change-point modelling as in Bai and Perron (1998) or Mikosch and Starica (1999, 2004), for instance, but it does not require prior information such as the number of changes. Additionally, the traditional structural-change tests require that the number of observations before each break point is large (and can grow to infinity) as these tests rely on asymptotic results. On the contrary, the proposed pointwise adaptive estimation does not rely on asymptotic results and does not thus place any requirements on the number of observations before, between or after any break point. Second, since the adaptively selected time-homogeneous interval used for estimation necessarily differs at each time point, the model coefficients can arbitrarily vary over time. In comparison to varying-coefficient models assuming

smooth development of parameters over time (Cai et al., 2000), our approach however allows for structural breaks in the form of sudden jumps in parameter values.

Although seemingly straightforward, extending Mercurio and Spokoiny's (2004) procedure to the local parametric modelling is a non-trivial problem, which requires new tools and techniques. We concentrate here on the change-point estimation of financial time series, which are often modelled by data-demanding models such as GARCH. While the benefits of a flexible change-point analysis for time series spanning several years are well known, its feasibility (which stands in the focus of this work) is much more difficult to achieve. The reason is thus that, at each time point, the procedure starts from a small interval, where a local parametric approximation holds, and then iteratively extends this interval and tests it for time-homogeneity until a structural break is found or data exhausted. Hence, a model has to be initially estimated on very short time intervals (e.g. 10 observations). Using standard testing methods, such a procedure might be feasible for simple parametric models, but it is hardly possible for more complex parametric models such as GARCH that generally require rather large samples for reasonably good estimates.

Therefore, we use an alternative and more robust approach to local change-point analysis that relies on a finite-sample theory of testing a growing sequence of historical time intervals on homogeneity against a change-point alternative. The proposed adaptive pointwise estimation procedure applies to a wide class of time-series models, including AR and CH models. Concentrating on the latter, we describe in details the adaptive procedure, derive its basic properties, and focusing on the feasibility of adaptive estimation for CH models, study the performance in comparison to the parametric (G)ARCH by means of simulations and real-data applications. The main conclusion is two-fold: on one hand, the adaptive pointwise estimation is feasible and beneficial also in the case of data-demanding models such as GARCH; on the other hand, the adaptive estimates based on various parametric models such as constant, ARCH or GARCH models are much closer to each other (while being better than the usual parametric estimates), which eliminates to some extent the need for using too complex models in adaptive estimation.

The rest of the paper is organized as follows. In Section 2, the parametric estimation of CH models and its finite-sample properties are introduced. In Section 3, we define the adaptive pointwise estimation procedure and discuss the choice of its parameters. Theoretical properties of the method are discussed in Section 4. In the specific case of the ARCH(1) and GARCH(1,1) models, a simulation study illustrates the performance of the new methodology with respect to the standard parametric and change-point models in Section 5. Applications to real stock-index series data are presented in Section 6. The proofs are provided in the Appendix.

2. PARAMETRIC CONDITIONAL HETEROSCEDASTICITY MODELS

Consider a time series Y_t in discrete time, $t \in N$. The CH assumption means that $Y_t = \sigma_t \varepsilon_t$, where $\{\varepsilon_t\}_{t \in N}$ is a white noise process and $\{\sigma_t\}_{t \in N}$ is a predictable volatility (conditional variance) process. Modelling of the volatility process σ_t typically relies on some parametric CH specification such as the ARCH (Engle, 1982) and GARCH (Bollerslev, 1986) models:

$$\sigma_t^2 = \omega + \sum_{i=1}^{p} \alpha_i Y_{t-i}^2 + \sum_{j=1}^{q} \beta_j \sigma_{t-j}^2, \tag{2.1}$$

where $p \in N, q \in N$ and $\boldsymbol{\theta} = (\omega, \alpha_1, \ldots, \alpha_p, \beta_1, \ldots, \beta_q)^\top$ is the parameter vector. An attractive feature of this model is that, even with very few coefficients, one can model most stylized facts of financial time series like volatility clustering or excessive kurtosis, for instance. A number of (G)ARCH extensions were proposed to make the model even more flexible; e.g. EGARCH (Nelson, 1991), QGARCH (Sentana, 1995) and TGARCH (Glosten et al., 1993) that account for asymmetries in a volatility process.

All such CH models can be put into a common class of generalized linear volatility models:

$$Y_t = \sigma_t \varepsilon_t = \sqrt{g(X_t)}\varepsilon_t, \tag{2.2}$$

$$X_t = \omega + \sum_{i=1}^{p} \alpha_i h(Y_{t-i}) + \sum_{j=1}^{q} \beta_j X_{t-j}\,, \tag{2.3}$$

where g and h are known functions and X_t is a (partially) unobserved process (structural variable) that models the volatility coefficient σ_t^2 via transformation $g : \sigma_t^2 = g(X_t)$. For example, the GARCH model (2.1) is described by $g(u) = u$ and $h(r) = r^2$.

Models (2.2)–(2.3) are time homogeneous in the sense that the process Y_t follows the same structural equation at each time point. In other words, the parameter $\boldsymbol{\theta}$ and hence the structural dependence in Y_t is constant over time. Even though models like (2.2)–(2.3) can often fit data well over a longer period of time, the assumption of homogeneity is too restrictive in practical applications: to guarantee a sufficient amount of data for sufficiently precise estimation, these models are often applied over time spans of many years. On the contrary, the strategy pursued here requires only local time homogeneity, which means that at each time point t there is a (possibly rather short) interval $[t - m, t]$, where the process Y_t is well described by models (2.2)–(2.3). This strategy aims then both at finding an interval of homogeneity (preferably as long as possible) and at the estimation of the corresponding parameter values $\boldsymbol{\theta}$, which then enable predicting Y_t and X_t.

Next, we discuss the parameter estimation for models (2.2)–(2.3) using observations Y_t from some time interval $I = [t_0, t_1]$. The conditional distribution of each observation Y_t given the past $\mathscr{F}_{t-1}$ is determined by the structural variable X_t, whose dynamics are described by the parameter vector $\boldsymbol{\theta}$: $X_t = X_t(\boldsymbol{\theta})$ for $t \in I$ due to (2.3). We denote the underlying value of $\boldsymbol{\theta}$ by $\boldsymbol{\theta}_0$.

For estimating $\boldsymbol{\theta}_0$, we apply the quasi-maximum likelihood (quasi-MLE) approach using the estimating equations generated under the assumption of Gaussian errors ε_t. This guarantees efficiency under the normality of innovations and consistency under rather general moment conditions (Hansen and Lee, 1994, and Francq and Zakoian, 2007). The log-likelihood for models (2.2)–(2.3) on an interval I can be represented in the form

$$L_I(\boldsymbol{\theta}) = \sum_{t \in I} \ell\{Y_t, g[X_t(\boldsymbol{\theta})]\}$$

with log-likelihood function $\ell(y, \upsilon) = -0.5\{\log(\upsilon) + y^2/\upsilon\}$. We define the quasi-MLE estimate $\widetilde{\boldsymbol{\theta}}_I$ of the parameter $\boldsymbol{\theta}$ by maximizing the log-likelihood $L_I(\boldsymbol{\theta})$,

$$\widetilde{\boldsymbol{\theta}}_I = \operatorname*{argmax}_{\boldsymbol{\theta} \in \Theta} L_I(\boldsymbol{\theta}) = \operatorname*{argmax}_{\boldsymbol{\theta} \in \Theta} \sum_{t \in I} \ell\{Y_t, g[X_t(\boldsymbol{\theta})]\}, \tag{2.4}$$

and denote by $L_I(\widetilde{\boldsymbol{\theta}}_I)$ the corresponding maximum.

To characterize the quality of estimating the parameter vector $\boldsymbol{\theta}_0 = (\omega, \alpha_1, \ldots, \alpha_p, \beta_1, \ldots, \beta_q)^\top$ by $\widetilde{\boldsymbol{\theta}}_I$, we now present an exact (non-asymptotic) exponential risk bound. This bound concerns the value of maximum $L_I(\widetilde{\boldsymbol{\theta}}_I) = \max_{\boldsymbol{\theta}\in\Theta} L_I(\boldsymbol{\theta})$ rather than the point of maximum $\widetilde{\boldsymbol{\theta}}_I$. More precisely, we consider the difference $L_I(\widetilde{\boldsymbol{\theta}}_I, \boldsymbol{\theta}_0) = L_I(\widetilde{\boldsymbol{\theta}}_I) - L_I(\boldsymbol{\theta}_0)$. By definition, this value is non-negative and represents the deviation of the maximum of the log-likelihood process from its value at the 'true' point $\boldsymbol{\theta}_0$. Later, we comment on how the accuracy of estimation of the parameter $\boldsymbol{\theta}_0$ by $\widetilde{\boldsymbol{\theta}}_I$ relates to the value $L_I(\widetilde{\boldsymbol{\theta}}_I, \boldsymbol{\theta}_0)$. We will also see that the bound for $L_I(\widetilde{\boldsymbol{\theta}}_I, \boldsymbol{\theta}_0)$ yields the confidence set for the parameter $\boldsymbol{\theta}_0$, which will be used for the proposed change-point test. Now, the non-asymptotic risk bound is specified in the following theorem, which formulates corollaries 4.2 and 4.3 of Spokoiny (2009b) for the case of the quasi-MLE estimation of a CH model (2.2)–(2.3) at $\boldsymbol{\theta} = \boldsymbol{\theta}_0$. The result can be viewed as an extension of the Wilks phenomenon that the distribution of $L_I(\widetilde{\boldsymbol{\theta}}_I, \boldsymbol{\theta}_0)$ for a linear Gaussian model is $\chi^2_p/2$, where p is the number of estimated parameters in the model.

THEOREM 2.1. *Assume that the process Y_t follows models (2.2)–(2.3) with the parameter $\boldsymbol{\theta}_0 \in \Theta$, where the set Θ is compact. The function $g(\cdot)$ is assumed to be continuously differentiable with the uniformly bounded first derivative and $g(x) \geq \delta > 0$ for all x. Further, let the process $X_t(\boldsymbol{\theta})$ be sub-ergodic in the sense that for any smooth function $f(\cdot)$ there exists f^* such that for any time interval I*

$$\boldsymbol{E}_{\boldsymbol{\theta}_0}\left|\sum_I \bigl\{f(X_t(\boldsymbol{\theta})) - \boldsymbol{E}_{\boldsymbol{\theta}_0} f(X_t(\boldsymbol{\theta}))\bigr\}\right|^2 \leq f^*|I|, \quad \boldsymbol{\theta} \in \Theta.$$

Finally, let $\boldsymbol{E}\exp\{\varkappa(\varepsilon_t^2 - 1)|\mathscr{F}_{t-1}\} \leq c(\varkappa)$ for some $\varkappa > 0, c(\varkappa) > 0$, and all $t \in N$. Then there are $\lambda > 0$ and $\mathfrak{e}(\lambda, \boldsymbol{\theta}_0) > 0$ such that for any interval I and $\mathfrak{z} > 0$

$$\boldsymbol{P}_{\boldsymbol{\theta}_0}\bigl(L_I(\widetilde{\boldsymbol{\theta}}_I, \boldsymbol{\theta}_0) > \mathfrak{z}\bigr) \leq \exp\{\mathfrak{e}(\lambda, \boldsymbol{\theta}_0) - \lambda\mathfrak{z}\}. \tag{2.5}$$

Moreover, for any $r > 0$, there is a constant $\mathfrak{R}_r(\boldsymbol{\theta}_0)$ such that

$$\boldsymbol{E}_{\boldsymbol{\theta}_0}\bigl|L_I(\widetilde{\boldsymbol{\theta}}_I, \boldsymbol{\theta}_0)\bigr|^r \leq \mathfrak{R}_r(\boldsymbol{\theta}_0). \tag{2.6}$$

REMARK 2.1. The condition $g(x) \geq \delta > 0$ guarantees that the variance process cannot reach zero. In the case of GARCH, it is sufficient to assume $\omega > 0$, for instance.

One attractive feature of Theorem 2.1, formulated in the following corollary, is that it enables constructing the non-asymptotic confidence sets and testing the parametric hypothesis on the basis of the fitted log-likelihood $L_I(\widetilde{\boldsymbol{\theta}}_I, \boldsymbol{\theta})$. This feature is especially important for our procedure presented in Section 3.

COROLLARY 2.1. *Under the assumptions of Theorem 2.1, let the value $\mathfrak{z}_\alpha$ fulfil $\mathfrak{e}(\lambda, \boldsymbol{\theta}_0) - \lambda\mathfrak{z}_\alpha < \log\alpha$ for some $\alpha < 1$. Then the random set $\mathscr{E}_I(\mathfrak{z}_\alpha) = \{\boldsymbol{\theta} : L_I(\widetilde{\boldsymbol{\theta}}_I, \boldsymbol{\theta}) \leq \mathfrak{z}_\alpha\}$ is an α-confidence set for $\boldsymbol{\theta}_0$ in the sense that $\boldsymbol{P}_{\boldsymbol{\theta}_0}(\boldsymbol{\theta}_0 \notin \mathscr{E}_I(\mathfrak{z}_\alpha)) \leq \alpha$.*

Theorem 2.1 also gives a non-asymptotic and fixed upper bound for the risk of estimation $L_I(\widetilde{\boldsymbol{\theta}}_I, \boldsymbol{\theta}_0)$ that applies to an arbitrary sample size $|I|$. To understand the relation of this result to the classical rate result, we can apply the standard arguments based on the quadratic expansion

of the log-likelihood $L(\widetilde{\boldsymbol{\theta}}, \boldsymbol{\theta})$. Let $\nabla^2 L(\boldsymbol{\theta})$ denote the Hessian matrix of the second derivatives of $L(\boldsymbol{\theta})$ with respect to the parameter $\boldsymbol{\theta}$. Then

$$L_I(\widetilde{\boldsymbol{\theta}}_I, \boldsymbol{\theta}_0) = 0.5(\widetilde{\boldsymbol{\theta}}_I - \boldsymbol{\theta}_0)^\top \nabla^2 L_I(\boldsymbol{\theta}'_I)(\widetilde{\boldsymbol{\theta}}_I - \boldsymbol{\theta}_0), \tag{2.7}$$

where $\boldsymbol{\theta}'_I$ is a convex combination of $\boldsymbol{\theta}_0$ and $\widetilde{\boldsymbol{\theta}}_I$. Under usual regularity assumptions and for sufficiently large $|I|$, the normalized matrix $|I|^{-1}\nabla^2 L_I(\boldsymbol{\theta})$ is close to some matrix $V(\boldsymbol{\theta})$, which depends only on the stationary distribution of Y_t and is continuous in $\boldsymbol{\theta}$. Then (2.5) approximately means that $\|\sqrt{V(\boldsymbol{\theta}_0)}(\widetilde{\boldsymbol{\theta}}_I - \boldsymbol{\theta}_0)\|^2 \le \mathfrak{z}/|I|$ with probability close to 1 for large $\mathfrak{z}$. Hence, the large deviation result of Theorem 2.1 yields the root-$|I|$ consistency of the MLE estimate $\widetilde{\boldsymbol{\theta}}_I$. See Spokoiny (2009b) for further details.

3. POINTWISE ADAPTIVE NON-PARAMETRIC ESTIMATION

An obvious feature of models (2.2)–(2.3) is that the parametric structure of the process is assumed constant over the whole sample and cannot thus incorporate changes and structural breaks at unknown times in the models. A natural generalization leads to models whose coefficients may change over time (Fan and Zhang, 2008). One can then assume that the structural process X_t satisfies the relation (2.3) at any time, but the vector of coefficients $\boldsymbol{\theta}$ may vary with the time t, $\boldsymbol{\theta} = \boldsymbol{\theta}(t)$. The estimation of the coefficients as general functions of time is possible only under some additional assumptions on these functions. Typical assumptions are (i) varying coefficients are smooth functions of time (Cai et al., 2000) and (ii) varying coefficients are piecewise constant functions (Bai and Perron, 1998, and Mikosch and Starica, 1999, 2004).

Our local parametric approach differs from the commonly used identification assumptions (i) and (ii). We assume that the observed data Y_t are described by a (partially) unobserved process X_t due to (2.2), and at each point T, there exists a historical interval $I(T) = [t_0, T]$ in which the process X_t 'nearly' follows the parametric specification (2.3) (see Section 4 for details on what 'nearly' means). This local structural assumption enables us to apply well-developed parametric estimation for data $\{Y_t\}_{t \in I(T)}$ to estimate the underlying parameter $\boldsymbol{\theta} = \boldsymbol{\theta}(T)$ by $\hat{\boldsymbol{\theta}} = \hat{\boldsymbol{\theta}}(T)$. (The estimate $\hat{\boldsymbol{\theta}} = \hat{\boldsymbol{\theta}}(T)$ can then be used for estimating the value $\hat{X}_T$ of the process X_t at T from equation (2.3) and for further modelling such as forecasting Y_{T+1}.) Moreover, this assumption includes the above-mentioned 'smooth transition' and 'switching regime' assumptions (i) and (ii) as special cases: parameters $\hat{\boldsymbol{\theta}}(T)$ vary over time as the interval $I(T)$ changes with T and, at the same time, discontinuities and jumps in $\hat{\boldsymbol{\theta}}(T)$ as a function of time are possible.

To estimate $\hat{\boldsymbol{\theta}}(T)$, we have to find the historical interval of homogeneity $I(T)$, i.e. the longest interval I with the right-end point T, where data do not contradict a specified parametric model with fixed parameter values. Starting at each time T with a very short interval $I = [t_0, T]$, we search by successive extending and testing of interval I on homogeneity against a change-point alternative: if the hypothesis of homogeneity is not rejected for a given I, a larger interval is taken and tested again. Contrary to Bai and Perron (1998) and Mikosch and Starica (1999), who detect all change points in a given time series, our approach is local: it focuses on the local change-point analysis near point T of estimation and tries to find only one change closest to the reference point.

In the rest of this section, we first discuss the test statistics employed to test the time-homogeneity of an interval I against a change-point alternative in Section 3.1. Later, we rigorously describe the pointwise adaptive estimation procedure in Section 3.2. Its

implementation and the choice of parameters entering the adaptive procedure are described in Sections 3.2–3.4. Theoretical properties of the method are studied in Section 4.

3.1. Test of homogeneity against a change-point alternative

The pointwise adaptive estimation procedure crucially relies on the test of local time-homogeneity of an interval $I = [t_0, T]$. The null hypothesis for I means that the observations $\{Y_t\}_{t \in I}$ follow the parametric models (2.2)–(2.3) with a fixed parameter $\boldsymbol{\theta}_0$, leading to the quasi-MLE estimate $\widetilde{\boldsymbol{\theta}}_I$ from (2.4) and the corresponding fitted log-likelihood $L_I(\widetilde{\boldsymbol{\theta}}_I)$.

The change-point alternative for a given change-point location $\tau \in I$ can be described as follows: process Y_t follows the parametric models (2.2)–(2.3) with a parameter $\boldsymbol{\theta}_J$ for $t \in J = [t_0, \tau]$ and with a different parameter $\boldsymbol{\theta}_{J^c}$ for $t \in J^c = [\tau + 1, T]; \boldsymbol{\theta}_J \neq \boldsymbol{\theta}_{J^c}$. The fitted log-likelihood under this alternative reads as $L_J(\widetilde{\boldsymbol{\theta}}_J) + L_{J^c}(\widetilde{\boldsymbol{\theta}}_{J^c})$. The test of homogeneity can be performed using the likelihood ratio (LR) test statistic $T_{I,\tau}$:

$$T_{I,\tau} = \max_{\boldsymbol{\theta}_J, \boldsymbol{\theta}_{J^c} \in \Theta} \{L_J(\boldsymbol{\theta}_J) + L_{J^c}(\boldsymbol{\theta}_{J^c})\} - \max_{\boldsymbol{\theta} \in \Theta} L_I(\boldsymbol{\theta}) = \left\{L_J(\widetilde{\boldsymbol{\theta}}_J) + L_{J^c}(\widetilde{\boldsymbol{\theta}}_{J^c}) - L_I(\widetilde{\boldsymbol{\theta}}_I)\right\}.$$

Since the change-point location τ is generally not known, we consider the supremum of the LR statistics $T_{I,\tau}$ over some subset $\tau \in \mathscr{T}(I)$; cf. Andrews (1993):

$$T_{I,\mathscr{T}(I)} = \sup_{\tau \in \mathscr{T}(I)} T_{I,\tau}. \tag{3.1}$$

A typical example of a set $\mathscr{T}(I)$ is $\mathscr{T}(I) = \{\tau : t_0 + m' \leq \tau \leq T - m''\}$ for some fixed $m', m'' > 0$.

3.2. Adaptive search for the longest interval of homogeneity

This section presents the proposed adaptive pointwise estimation procedure. At each point T, we aim at estimating the unknown parameters $\boldsymbol{\theta}(T)$ from historical data $Y_t, t \leq T$; this procedure repeats for every current time point T as new data arrive. At the first step, the procedure selects on the base of historical data an interval $\hat{I}(T)$ of homogeneity in which the data do not contradict the parametric models (2.2)–(2.3). Afterwards, the quasi-MLE estimation is applied using the selected historical interval $\hat{I}(T)$ to obtain estimate $\hat{\boldsymbol{\theta}}(T) = \widetilde{\boldsymbol{\theta}}_{\hat{I}(T)}$. From now on, we consider an arbitrary, but fixed time point T.

Suppose that a growing set $I_0 \subset I_1 \subset \cdots \subset I_K$ of historical interval-candidates $I_k = [T - m_k + 1, T]$ with the right-end point T is fixed. The smallest interval I_0 is accepted automatically as homogeneous. Then the procedure successively checks every larger interval I_k on homogeneity using the test statistic $T_{I_k,\mathscr{T}(I_k)}$ from (3.1). The selected interval $\hat{I}$ corresponds to the largest accepted interval $I_{\hat{k}}$ with index $\hat{k}$ such that

$$T_{I_k,\mathscr{T}(I_k)} \leq \mathfrak{z}_k, \quad k \leq \hat{k}, \tag{3.2}$$

and $T_{I_{\hat{k}+1},\mathscr{T}(I_{\hat{k}+1})} > \mathfrak{z}_{\hat{k}+1}$, where the critical values $\mathfrak{z}_k$ are discussed later in this section and specified in Section 3.3. This procedure then leads to the adaptive estimate $\hat{\boldsymbol{\theta}} = \widetilde{\boldsymbol{\theta}}_{\hat{I}}$ corresponding to the selected interval $\hat{I} = I_{\hat{k}}$.

The complete description of the procedure includes two steps. (A) Fixing the set-up and the parameters of the procedure. (B) Data-driven search for the longest interval of homogeneity.

(A) Set-up and parameters:

1 Select specific parametric models (2.2)–(2.3) [e.g. constant volatility, ARCH(1), GARCH(1,1)].
2 Select the set $\mathscr{I} = (I_0, \ldots, I_K)$ of interval-candidates, and for each $I_k \in \mathscr{I}$, the set $\mathscr{T}(I_k)$ of possible change points $\tau \in I_k$ used in the LR test (3.1).
3 Select the critical values $\mathfrak{z}_1, \ldots, \mathfrak{z}_K$ in (3.2) as described in Section 3.3.

(B) Adaptive search and estimation: Set $k = 1$, $\hat{I} = I_0$ and $\hat{\boldsymbol{\theta}} = \widetilde{\boldsymbol{\theta}}_{I_0}$.

1 Test the hypothesis $H_{0,k}$ of no change point within the interval I_k using test statistics (3.1) and the critical values $\mathfrak{z}_k$ obtained in (A3). If a change point is detected ($H_{0,k}$ is rejected), go to (B3). Otherwise proceed with (B2).
2 Set $\hat{\boldsymbol{\theta}} = \widetilde{\boldsymbol{\theta}}_{I_k}$ and $\hat{\boldsymbol{\theta}}_{I_k} = \widetilde{\boldsymbol{\theta}}_{I_k}$. Further, set $k := k + 1$. If $k \leq K$, repeat (B1); otherwise go to (B3).
3 Define $\hat{I} = I_{k-1} =$ 'the last accepted interval' and $\hat{\boldsymbol{\theta}} = \widetilde{\boldsymbol{\theta}}_{\hat{I}}$. Additionally, set $\hat{\boldsymbol{\theta}}_{I_k} = \cdots = \hat{\boldsymbol{\theta}}_{I_K} = \hat{\boldsymbol{\theta}}$ if $k \leq K$.

In step (A), one has to select three main ingredients of the procedure. First, the parametric model used locally to approximate the process Y_t has to be specified in (A1), e.g. the constant volatility or GARCH(1,1) in our context. Next, in step (A2), the set of intervals $\mathscr{I} = \{I_k\}_{k=0}^K$ is fixed, each interval with the right-end point T, length $m_k = |I_k|$, and the set $\mathscr{T}(I_k)$ of tested change points. Our default proposal is to use a geometric grid $m_k = [m_0 a^k]$, $a > 1$, and to set $I_k = [T - m_k + 1, T]$ and $\mathscr{T}(I_k) = [T - m_{k-1} + 1, T - m_{k-2}]$. Although our experiments show that the procedure is rather insensitive to the choice of m_0 and a (e.g. we use $m_0 = 10$ and $a = 1.25$ in simulations), the length m_0 of interval I_0 should take into account the parametric model selected in (A1). The reason is that I_0 is always assumed to be time-homogeneous and m_0 thus has to reflect flexibility of the parametric model; e.g. while $m_0 = 20$ might be reasonable for the GARCH(1,1) model, $m_0 = 5$ could be a reasonable choice for the locally constant approximation of a volatility process. Finally, in step (A3), one has to select the K critical values $\mathfrak{z}_k$ in (3.2) for the LR test statistics $T_{I_k, \mathscr{T}(I_k)}$ from (3.1). The critical values $\mathfrak{z}_k$ will generally depend on the parametric model describing the null hypothesis of time-homogeneity, the set $\mathscr{I}$ of intervals I_k and corresponding sets of considered change points $\mathscr{T}(I_k)$, $k \leq K$, and additionally, on two constants r and ρ that are counterparts of the usual significance level. All these determinants of the critical values can be selected in step (A) and the critical values are thus obtained before the actual estimation takes place in step (B). Due to its importance, the method of constructing critical values $\{\mathfrak{z}_k\}_{k=1}^K$ is discussed separately in Section 3.3.

The main step (B) performs the search for the longest time-homogeneous interval. Initially, I_0 is assumed to be homogeneous. If I_{k-1} is negatively tested on the presence of a change point, one continues with I_k by employing test (3.1) in step (B1), which checks for a potential change point in I_k. If no change point is found, then I_k is accepted as time-homogeneous in step (B2); otherwise the procedure terminates in step (B3). We sequentially repeat these tests until we find a change point or exhaust all intervals. The latest (longest) interval accepted as time-homogeneous is used for estimation in step (B3). Note that the estimate $\hat{\boldsymbol{\theta}}_{I_k}$ defined in (B2) and (B3) corresponds to the latest accepted interval $\hat{I}_k$ after the first k steps, or equivalently, the interval selected out of $I_1, \ldots, I_k$.

Moreover, the whole search and estimation step (B) can be repeated at different time points T without reiterating the initial step (A) as the critical values $\mathfrak{z}_k$ depend only on the approximating parametric model and interval lengths $m_k = |I_k|$, not on the time point T (see Section 3.3).

3.3. Choice of critical values $\mathfrak{z}_k$

The presented method of choosing the interval of homogeneity $\hat{I}$ can be viewed as multiple testing procedure. The critical values for this procedure are selected using the general approach of testing theory: to provide a prescribed performance of the procedure under the null hypothesis, i.e. in the pure parametric situation. This means that the procedure is trained on the data generated from the pure parametric time-homogeneous model from step (A1). The correct choice in this situation is the largest considered interval I_K and a choice $I_{\hat{k}}$ with $\hat{k} < K$ can be interpreted as a 'false alarm'. We select the minimal critical values ensuring a small probability of such a false alarm. Our condition slightly differs though from the classical level condition because we focus on parameter estimation rather than on hypothesis testing.

In the pure parametric case, the 'ideal' estimate corresponds to the largest considered interval I_K. Due to Theorem 2.1, the quality of estimation of the parameter $\boldsymbol{\theta}_0$ by $\widetilde{\boldsymbol{\theta}}_{I_K}$ can be measured by the log-likelihood 'loss' $L_{I_K}(\widetilde{\boldsymbol{\theta}}_{I_K}, \boldsymbol{\theta}_0)$, which is stochastically bounded with exponential and polynomial moments: $\boldsymbol{E}_{\boldsymbol{\theta}_0}|L_{I_K}(\widetilde{\boldsymbol{\theta}}_{I_K}, \boldsymbol{\theta}_0)|^r \le \mathfrak{R}_r(\boldsymbol{\theta}_0)$. If the adaptive procedure stops earlier at some intermediate step $k < K$, we select instead of $\widetilde{\boldsymbol{\theta}}_{I_K}$ another estimate $\hat{\boldsymbol{\theta}} = \widetilde{\boldsymbol{\theta}}_{I_k}$ with a larger variability. The loss associated with such a false alarm can be measured by the value $L_{I_K}(\widetilde{\boldsymbol{\theta}}_{I_K}, \hat{\boldsymbol{\theta}}) = L_{I_K}(\widetilde{\boldsymbol{\theta}}_{I_K}) - L_{I_K}(\hat{\boldsymbol{\theta}})$. The corresponding condition bounding the loss due to the adaptive estimation reads as

$$\boldsymbol{E}_{\boldsymbol{\theta}_0}\big|L_{I_K}(\widetilde{\boldsymbol{\theta}}_{I_K}, \hat{\boldsymbol{\theta}})\big|^r \le \rho\mathfrak{R}_r(\boldsymbol{\theta}_0). \tag{3.3}$$

This is in fact an implicit condition on the critical values $\{\mathfrak{z}_k\}_{k=1}^K$, which ensures that the loss associated with the false alarm is at most the ρ-fraction of the log-likelihood loss of the 'ideal' or 'oracle' estimate $\widetilde{\boldsymbol{\theta}}_{I_K}$ for the parametric situation. The constant r corresponds to the power of the loss in (3.3), while ρ is similar in meaning to the test level. In the limit case when r tends to zero, this condition (3.3) becomes the usual level condition: $\boldsymbol{P}_{\boldsymbol{\theta}_0}(I_K \text{ is rejected}) = \boldsymbol{P}_{\boldsymbol{\theta}_0}(\widetilde{\boldsymbol{\theta}}_{I_K} \ne \hat{\boldsymbol{\theta}}) \le \rho$. The choice of the metaparameters r and ρ is discussed in Section 3.4.

A condition similar to (3.3) is imposed at each step of the adaptive procedure. The estimate $\hat{\boldsymbol{\theta}}_{I_k}$ coming after the k steps of the procedure should satisfy

$$\boldsymbol{E}_{\boldsymbol{\theta}_0}\big|L_{I_k}(\widetilde{\boldsymbol{\theta}}_{I_k}, \hat{\boldsymbol{\theta}}_{I_k})\big|^r \le \rho_k\mathfrak{R}_r(\boldsymbol{\theta}_0), \quad k = 1, \dots, K, \tag{3.4}$$

where $\rho_k = \rho\, k/K \le \rho$. The following theorem presents some sufficient conditions on the critical values $\{\mathfrak{z}_k\}_{k=1}^K$ ensuring (3.4); recall that $m_k = |I_k|$ denotes the length of I_k.

THEOREM 3.1. *Suppose that $r > 0, \rho > 0$. Under the assumptions of Theorem 2.1, there are constants a_0, a_1, a_2 such that the condition (3.4) is fulfilled with the choice*

$$\mathfrak{z}_k = a_0 r \log(\rho^{-1}) + a_1 r \log(m_K/m_{k-1}) + a_2 \log(m_k), \quad k = 1, \dots, K.$$

Since K and $\{m_k\}_{k=1}^K$ are fixed, the $\mathfrak{z}_k$'s in Theorem 3.1 have a form $\mathfrak{z}_k = C + D\log(m_k)$ for $k = 1, \dots, K$ with some constant C and D. However, a practically relevant choice of these constants has to be done by Monte Carlo simulations. Note first that every particular choice of the coefficients C and D determines the whole set of the critical values $\{\mathfrak{z}_k\}_{k=1}^K$ and thus the local change-point procedure. For the critical values given by fixed (C, D), one can run the procedure and observe its performance on the simulated data using the data-generating process (2.2)–(2.3); in particular, one can check whether the condition (3.4) is fulfilled. For any (sufficiently large) fixed value of C, one can thus find the minimal value $D(C) < 0$ of D that ensures (3.4).

Every corresponding set of critical values in the form $\mathfrak{z}_k = C + D(C)\log(m_k)$ is admissible. The condition $D(C) < 0$ ensures that the critical values decreases with k. This reflects the fact that a false alarm at an early stage of the algorithm is more crucial because it leads to the choice of a highly variable estimate. The critical values $\mathfrak{z}_k$ for small k should thus be rather conservative to provide the stability of the algorithm in the parametric situation. To determine C, the value $\mathfrak{z}_1$ can be fixed by considering the false alarm at the first step of the procedure, which leads to estimation using the smallest interval I_0 instead of the 'ideal' largest interval I_K. The related condition (used in Section 5.1) reads as

$$\boldsymbol{E}_{\theta_0}\big|L_{I_K}(\widetilde{\boldsymbol{\theta}}_{I_K}, \widetilde{\boldsymbol{\theta}}_{I_0})\big|^r \mathbf{1}(T_{I_1,\mathscr{T}(I_1)} > \mathfrak{z}_1) \le \rho\mathfrak{R}_r(\boldsymbol{\theta}_0)/K. \tag{3.5}$$

Alternatively, one could select a pair (C, D) that minimizes the resulting prediction error; see Section 3.4.

3.4. Selecting parameters r and ρ

The choice of critical values using inequality (3.4) additionally depends on two 'metaparameters' r and ρ. A simple strategy is to use conservative values for these parameters and the corresponding set of critical values (e.g. our default is $r = 1$ and $\rho = 1$). On the other hand, the two parameters are global in the sense that they are independent of T. Hence, one can also determine them in a data-driven way by minimizing some global forecasting error (Cheng et al., 2003). Different values of r and ρ may lead to different sets of critical values and hence to different estimates $\hat{\boldsymbol{\theta}}^{(r,\rho)}(T)$ and to different forecasts $\hat{Y}^{(r,\rho)}_{T+h|T}$ of the future values Y_{T+h}, where h is the forecasting horizon. Now, a data-driven choice of r and ρ can be done by minimizing the following objective function:

$$(\hat{r}, \hat{\rho}) = \operatorname*{arg\,min}_{r\ge 0,\rho>0} PE_{\Lambda,\mathscr{H}}(r,\rho) = \operatorname*{arg\,min}_{r,\rho} \sum_T \sum_{h\in\mathscr{H}} \Lambda\big(Y_{T+h}, \hat{Y}^{(r,\rho)}_{T+h|T}\big), \tag{3.6}$$

where Λ is a loss function and $\mathscr{H}$ is the forecasting horizon set. For example, one can take $\Lambda_r(\upsilon, \upsilon') = |\upsilon - \upsilon'|^r$ for $r \in [1/2, 2]$. For daily data, the forecasting horizon could be one day, $\mathscr{H} = \{1\}$, or two weeks, $\mathscr{H} = \{1, \ldots, 10\}$.

4. THEORETIC PROPERTIES

In this section, we collect basic results describing the quality of the proposed adaptive procedure. First, the definition of the procedure ensures the performance prescribed by (3.4) in the parametric situation. We however claimed that the adaptive pointwise estimation applies even if the process Y_t is only locally approximated by a parametric model. Therefore, we now define a locally 'nearly parametric' process, for which we derive an analogy of Theorem 2.1 (Section 4.1). Later, we prove certain 'oracle' properties of the proposed method (Section 4.2).

4.1. Small modelling bias condition

This section discusses the concept of a 'nearly parametric' case. To define it rigorously, we have to quantify the quality of approximating the true latent process X_t, which drives the observed data Y_t due to (2.2), by the parametric process $X_t(\boldsymbol{\theta})$ described by (2.3) for some $\boldsymbol{\theta} \in \Theta$. Below

we assume that the innovations ε_t in the model (2.2) are independent and identically distributed and denote the distribution of $\sqrt{\upsilon}\varepsilon_t$ by P_υ so that the conditional distribution of Y_t given $\mathscr{F}_{t-1}$ is $P_{g(X_t)}$. To measure the distance of a data-generating process from a parametric model, we introduce for every interval $I_k \in \mathscr{I}$ and every parameter $\boldsymbol{\theta} \in \Theta$ the random quantity

$$\Delta_{I_k}(\boldsymbol{\theta}) = \sum_{t \in I_k} \mathscr{K}\{g(X_t), g[X_t(\boldsymbol{\theta})]\},$$

where $\mathscr{K}(\upsilon, \upsilon')$ denotes the Kullback–Leibler distance between P_υ and $P_{\upsilon'}$. For CH models with Gaussian innovations ε_t, $\mathscr{K}(\upsilon, \upsilon') = -0.5\{\log(\upsilon/\upsilon') + 1 - \upsilon/\upsilon'\}$. In the parametric case with $X_t = X_t(\boldsymbol{\theta}_0)$, we clearly have $\Delta_{I_k}(\boldsymbol{\theta}_0) = 0$. To characterize the 'nearly parametric case', we introduce a $\{$small modelling bias$\}$ (SMB) condition, which simply means that, for some $\boldsymbol{\theta} \in \Theta$, $\Delta_{I_k}(\boldsymbol{\theta})$ is bounded by a small constant with a high probability. Informally, this means that the 'true' model can be well approximated on the interval I_k by the parametric one with the parameter $\boldsymbol{\theta}$. The best parametric fit (2.3) to the underlying model (2.2) on I_k can be defined by minimizing the value $\boldsymbol{E}\Delta_{I_k}(\boldsymbol{\theta})$ over $\boldsymbol{\theta} \in \Theta$ and $\widetilde{\boldsymbol{\theta}}_{I_k}$ can be viewed as its estimate.

The following theorem claims that the results on the accuracy of estimation given in Theorem 2.1 can be extended from the parametric case to the general non-parametric situation under the SMB condition. Let $\varrho(\hat{\boldsymbol{\theta}}, \boldsymbol{\theta})$ be any loss function for an estimate $\hat{\boldsymbol{\theta}}$.

THEOREM 4.1. *Let for some $\boldsymbol{\theta} \in \Theta$ and some $\Delta \geq 0$*

$$\boldsymbol{E}\Delta_{I_k}(\boldsymbol{\theta}) \leq \Delta. \tag{4.1}$$

Then it holds for an estimate $\hat{\boldsymbol{\theta}}$ constructed from the observations $\{Y_t\}_{t \in I_k}$ that

$$\boldsymbol{E}\log\bigl(1 + \varrho(\hat{\boldsymbol{\theta}}, \boldsymbol{\theta})/\boldsymbol{E}_{\boldsymbol{\theta}}\varrho(\hat{\boldsymbol{\theta}}, \boldsymbol{\theta})\bigr) \leq 1 + \Delta.$$

This general result applied to the quasi-MLE estimation with the loss function $L_I(\widetilde{\boldsymbol{\theta}}_I, \boldsymbol{\theta})$ yields the following corollary.

COROLLARY 4.1. *Let the SMB condition (4.1) hold for some interval I_k and $\boldsymbol{\theta} \in \Theta$. Then*

$$\boldsymbol{E}\log\Bigl(1 + \bigl|L_{I_k}(\widetilde{\boldsymbol{\theta}}_{I_k}, \boldsymbol{\theta})\bigr|^r/\mathfrak{R}_r(\boldsymbol{\theta})\Bigr) \leq 1 + \Delta,$$

where $\mathfrak{R}_r(\boldsymbol{\theta})$ is the parametric risk bound from (2.6).

This result shows that the estimation loss $|L_I(\widetilde{\boldsymbol{\theta}}_I, \boldsymbol{\theta})|^r$ normalized by the parametric risk $\mathfrak{R}_r(\boldsymbol{\theta})$ is stochastically bounded by a constant proportional to e^Δ. If Δ is not large, this result extends the parametric risk bound (Theorem 2.1) to the non-parametric situation under the SMB condition. Another implication of Corollary 4.1 is that the confidence set built for the parametric model (Corollary 2.1) continues to hold, with a slightly smaller coverage probability, under SMB.

4.2. The 'oracle' choice and the 'oracle' result

Corollary 4.1 suggests that the 'optimal' or 'oracle' choice of the interval I_k from the set $I_1, \ldots, I_K$ can be defined as the largest interval for which the SMB condition (4.1) still holds (for a given small $\Delta > 0$). For such an interval, one can neglect deviations of the underlying

process from a parametric model with a fixed parameter $\boldsymbol{\theta}$. Therefore, we say that the choice k^* is the 'oracle' choice if there exists $\boldsymbol{\theta} \in \Theta$ such that

$$\boldsymbol{E}\,\Delta_{I_{k^*}}(\boldsymbol{\theta}) \le \Delta \tag{4.2}$$

for a fixed $\Delta > 0$ and that (4.2) does not hold for $k > k^*$. Unfortunately, the underlying process X_t and, hence, the value Δ_{I_k} is unknown and the oracle choice cannot be implemented. The proposed adaptive procedure tries to mimic this oracle on the basis of available data using the sequential test of homogeneity. The final oracle result claims that the adaptive estimate provides the same (in order) accuracy as the oracle one.

By construction, the pointwise adaptive procedure described in Section 3 provides the prescribed performance if the underlying process follows the parametric model (2.2). Now, condition (3.4) combined with Theorem 4.1 implies similar performance in the first k^* steps of the adaptive estimation procedure.

THEOREM 4.2. *Let $\boldsymbol{\theta} \in \Theta$ and $\Delta > 0$ be such that $\boldsymbol{E}\,\Delta_{I_{k^*}}(\boldsymbol{\theta}) \le \Delta$ for some $k^* \le K$. Also let* $\max_{k \le k^*} \boldsymbol{E}_{\boldsymbol{\theta}} |L_{I_k}(\widetilde{\boldsymbol{\theta}}_{I_k}, \boldsymbol{\theta})|^r \le \mathfrak{R}_r(\boldsymbol{\theta})$. *Then*

$$\boldsymbol{E}\log\left(1 + \frac{\left|L_{I_{k^*}}\left(\widetilde{\boldsymbol{\theta}}_{I_{k^*}}, \boldsymbol{\theta}\right)\right|^r}{\mathfrak{R}_r(\boldsymbol{\theta})}\right) \le 1 + \Delta \quad \textit{and} \quad \boldsymbol{E}\log\left(1 + \frac{\left|L_{I_{k^*}}\left(\widetilde{\boldsymbol{\theta}}_{I_{k^*}}, \hat{\boldsymbol{\theta}}_{I_{k^*}}\right)\right|^r}{\mathfrak{R}_r(\boldsymbol{\theta})}\right) \le \rho + \Delta.$$

Similarly to the parametric case, under the SMB condition $\boldsymbol{E}\,\Delta_{I_{k^*}}(\boldsymbol{\theta}) \le \Delta$, any choice $\hat{k} < k^*$ can be viewed as a false alarm. Theorem 4.2 documents that the loss induced by such a false alarm at the first k^* steps and measured by $L_{I_{k^*}}(\widetilde{\boldsymbol{\theta}}_{I_{k^*}}, \hat{\boldsymbol{\theta}}_{I_{k^*}})$ is of the same magnitude as the loss $L_{I_{k^*}}(\widetilde{\boldsymbol{\theta}}_{I_{k^*}}, \boldsymbol{\theta})$ of estimating the parameter $\boldsymbol{\theta}$ from the SMB (4.2) by $\widetilde{\boldsymbol{\theta}}_{I_{k^*}}$. Thus, under (4.2) the adaptive estimation during steps $k \le k^*$ does not induce larger errors into estimation than the quasi-MLE estimation itself.

For further steps of the algorithm with $k > k^*$, where (4.2) does not hold, the value $\Delta' = \boldsymbol{E}\,\Delta_{I_k}(\boldsymbol{\theta})$ can be large and the bound for the risk becomes meaningless due to the factor $e^{\Delta'}$. To establish the result about the quality of the final estimate, we thus have to show that the quality of estimation cannot be destroyed at the steps $k > k^*$. The next 'oracle' result states the final quality of our adaptive estimate $\hat{\boldsymbol{\theta}}$.

THEOREM 4.3. *Let $\boldsymbol{E}\,\Delta_{I_{k^*}}(\boldsymbol{\theta}) \le \Delta$ for some $k^* \le K$. Then $L_{I_{k^*}}(\widetilde{\boldsymbol{\theta}}_{I_{k^*}}, \hat{\boldsymbol{\theta}})\mathbf{1}(\hat{k} \ge k^*) \le \mathfrak{z}_{k^*}$ yielding*

$$\boldsymbol{E}\log\left(1 + \frac{\left|L_{I_{k^*}}\left(\widetilde{\boldsymbol{\theta}}_{I_{k^*}}, \hat{\boldsymbol{\theta}}\right)\right|^r}{\mathfrak{R}_r(\boldsymbol{\theta})}\right) \le \rho + \Delta + \log\left(1 + \frac{\mathfrak{z}_{k^*}^r}{\mathfrak{R}_r(\boldsymbol{\theta})}\right).$$

Due to this result, the value $L_{I_{k^*}}(\widetilde{\boldsymbol{\theta}}_{I_{k^*}}, \hat{\boldsymbol{\theta}})$ is stochastically bounded. This can be interpreted as the oracle property of $\hat{\boldsymbol{\theta}}$ because it means that the adaptive estimate $\hat{\boldsymbol{\theta}}$ belongs with a high probability to the confidence set of the oracle estimate $\widetilde{\boldsymbol{\theta}}_{I_{k^*}}$.

5. SIMULATION STUDY

In the last two sections, we present simulation study (Section 5) and real data applications (Section 6) documenting the performance of the proposed adaptive estimation procedure. To verify the practical applicability of the method in a complex setting, we concentrate on the volatility estimation using parametric and adaptive pointwise estimation of constant volatility, ARCH(1) and GARCH(1,1) models (for the sake of brevity, referred to as the local constant,

local ARCH and local GARCH). The reason is that the estimation of GARCH models requires generally hundreds of observations for a reasonable quality of estimation, which puts the adaptive procedure working with samples as small as 10 or 20 observations to a hard test. Additionally, the critical values obtained as described in Section 3.3 depend on the underlying parameter values in the case of (G)ARCH.

Here we first study the finite-sample critical values for the test of homogeneity by means of Monte Carlo simulations and discuss practical implementation details (Section 5.1). Later, we demonstrate the performance of the proposed adaptive pointwise estimation procedure in simulated samples (Section 5.2). Note that, throughout this section, we identify the GARCH(1,1) models by triplets (ω, α, β): e.g. (1, 0.1, 0.3)-model. Constant volatility and ARCH(1) are then indicated by $\alpha = \beta = 0$ and $\beta = 0$, respectively. The GARCH estimation is done using the GARCH 3.0 package (Laurent and Peters, 2006) and Ox 3.30 (Doornik, 2002). Finally, since the focus is on modelling the volatility σ_t^2 in (2.2), the performance measurement and comparison of all models at time t is done by the absolute prediction error (PE) of the volatility process over a prediction horizon $\mathscr{H}$: $\mathrm{APE}(t) = \sum_{h\in\mathscr{H}} |\sigma_{t+h}^2 - \hat{\sigma}_{t+h|t}^2|/|\mathscr{H}|$, where $\hat{\sigma}_{t+h|t}^2$ represents the volatility prediction by a particular model.

5.1. *Finite-sample critical values for the test of homogeneity*

A practical application of the pointwise adaptive procedure requires critical values for the test of local homogeneity of a time series. Since they are obtained under the null hypothesis that a chosen parametric model (locally) describes the data, see Section 3, we need to obtain the critical values for the constant volatility, ARCH(1) and GARCH(1,1) models. Furthermore, for given r and ρ, the average risk (3.4) between the adaptive and oracle estimates can be bounded for critical values that linearly depend on the logarithm of interval length $|I_k| : \mathfrak{z}(|I_k|) = \mathfrak{z}_k = C + D\log(|I_k|)$ (see Theorem 3.1). As described in Section 3.3, we choose here the smallest C satisfying (3.5) and the corresponding minimum admissible value $D = D(C) < 0$ that guarantees the conditions (3.4).

We simulated the critical values for ARCH(1) and GARCH(1,1) models with different values of underlying parameters; see Table 1 for the critical values corresponding to $r = 1$ and $\rho = 1$. Their simulation was performed sequentially on intervals with lengths ranging from $|I_0| = m_0 = 10$ to $|I_K| = 570$ observations using a geometric grid with multiplier $a = 1.25$; see Section 3.2. (The results are, however, not sensitive to the choice of a.)

Unfortunately, the critical values depend on the parameters of the underlying (G)ARCH model (in contrast to the constant-volatility model). They generally seem to increase with the values of the ARCH and GARCH parameters keeping the other one fixed; see Table 1. To deal with this dependence on the underlying model parameters, we propose to choose the largest (most conservative) critical values corresponding to any estimated parameter in the analysed data. For example, if the largest estimated parameters of GARCH(1,1) are $\hat{\alpha} = 0.3$ and $\hat{\beta} = 0.8$, one should use $\mathfrak{z}(10) = 26.4$ and $\mathfrak{z}(570) = 14.5$, which are the largest critical values for models with $\alpha = 0.3, \beta \le 0.8$ and with $\alpha \le 0.3$, $\beta = 0.8$. (The proposed procedure is, however, not overly sensitive to this choice, as we shall see later.)

Finally, let us have a look at the influence of the tuning constants r and ρ in (3.4) on the critical values for several selected models (Table 2). The influence is significant, but can be classified in the following way. Whereas increasing ρ generally leads to an overall decrease of critical values (cf. Theorem 3.1), but primarily for the longer intervals, increasing r leads to an increase of

Table 1. Critical values $\mathfrak{z}_k = \mathfrak{z}(|I_k|)$ of the supremum LR test.

$\mathfrak{z}(\lvert I_k\rvert)$		β									
α	$\lvert I_k\rvert$	0.0	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9
0.0	10	15.5	15.5	16.4	16.8	17.9	17.3	17.0	17.0	16.9	16.0
	570	5.5	7.2	7.0	7.0	7.5	7.5	7.4	7.3	7.0	6.7
0.1	10	16.3	14.5	15.1	15.9	16.4	15.9	16.1	16.0	16.0	
	570	8.6	9.0	9.1	9.6	9.8	10.7	11.5	12.5	14.0	
0.2	10	16.7	15.2	15.7	16.2	16.9	18.9	20.1	25.1		
	570	9.4	10.6	11.2	11.4	11.4	12.5	13.3	14.2		
0.3	10	18.5	16.4	16.7	16.9	18.1	21.8	26.4			
	570	9.7	10.8	12.0	12.4	12.9	13.5	14.5			
0.4	10	22.1	16.5	18.3	19.3	22.8	30.9				
	570	9.9	12.0	13.0	13.4	13.9	14.7				
0.5	10	26.2	19.1	19.5	25.4	38.1					
	570	10.7	12.6	13.8	14.0	14.6					
0.6	10	33.0	22.8	25.9	32.4						
	570	12.7	12.7	13.9	15.3						
0.7	10	41.1	24.8	29.1							
	570	16.8	14.7	16.1							
0.8	10	66.2	26.4								
	570	31.5	15.8								
0.9	10	88.6									
	570	60.9									

Note: $\omega = 1, r = 1$ and $\rho = 1$.

critical values mainly for the shorter intervals; cf. (3.4). In simulations and real applications, we verified that a fixed choice such as $r = 1$ and $\rho = 1$ performs well. To optimize the performance of the adaptive methods, one can however determine constants r and ρ in a data-dependent way as described in Section 3.3. We use here this strategy for a small grid of $r \in \{0.5, 1.0\}$ and $\rho \in \{0.5, 1.0, 1.5\}$ and find globally optimal r and ρ. We will document, though, that the differences in the average absolute PE (3.6) for various values of r and ρ are relatively small.

5.2. Simulation study

We aim (i) to examine how well the proposed estimation method is able to adapt to long stable (time-homogeneous) periods and to less stable periods with more frequent volatility changes and (ii) to see which adaptively estimated model—local volatility, local ARCH or local GARCH—performs best in different regimes. To this end, we simulated 100 series from two change-point GARCH models with a low GARCH effect (ω, 0.2, 0.1) and a high GARCH effect (ω, 0.2, 0.7). Changes in constant ω are spread over a time span of 1000 days; see Figure 1. There is a long stable period at the beginning (500 days $\approx$ 2 years) and end (250 days $\approx$ 1 year) of time series with several volatility changes between them.

Table 2. Critical values $\mathfrak{z}(|I_k|)$ of the supremum LR test for various values r and ρ.

Model (ω, α, β)		(0.1, 0.0, 0.0)		(0.1, 0.2, 0.0)		(0.1, 0.1, 0.8)	
r	ρ	$\mathfrak{z}(10)$	$\mathfrak{z}(570)$	$\mathfrak{z}(10)$	$\mathfrak{z}(570)$	$\mathfrak{z}(10)$	$\mathfrak{z}(570)$
1.0	0.5	16.3	7.3	17.4	11.2	18.7	17.1
1.0	1.0	15.4	5.5	16.7	9.4	16.0	14.0
1.0	1.5	14.9	4.5	15.9	8.3	15.2	13.4
0.5	0.5	10.7	7.1	11.7	10.1	11.7	10.1
0.5	1.0	8.9	5.5	10.3	8.5	10.3	8.5
0.5	1.5	7.7	4.6	9.3	7.5	9.3	7.5

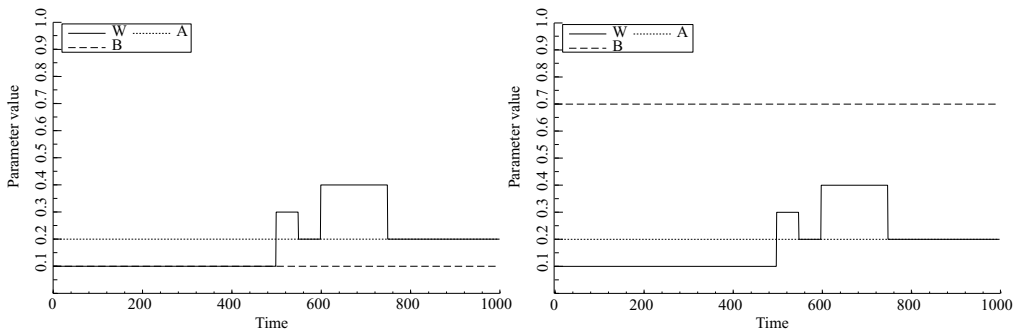


Figure 1. GARCH(1,1) parameters of low (left panel) and high (right panel) GARCH-effect simulations.

5.2.1. Low GARCH effect. Let us now discuss simulation results from the low GARCH-effect model. First, we mention the effect of structural changes in time series on the parameter estimation. Later, we compare the performance of all methods in terms of absolute PE.

Estimating a parametric model from data containing a change point will necessarily lead to various biases in estimation. For example, Hillebrand (2005) demonstrates that a change in volatility level ω within a sample drives the GARCH parameter β very close to 1. This is confirmed when we analyse the parameter estimates for parametric and adaptive GARCH at each time point $t \in [250, 1000]$ as depicted on Figure 2, where the mean (solid line), the 10% and 90% quantiles (dotted lines), and the true values (thick dotted line) of the model parameters are provided. The parametric estimates are consistent before breaks starting at $t = 500$, but the GARCH parameter β becomes inconsistent and converges to 1 once data contain breaks, $t > 500$. The locally adaptive estimates are similar to parametric ones before the breaks and become rather imprecise after the first change point, but they are not too far from the true value on average and stay consistent (in the sense that the confidence interval covers the true values). The low precision of estimation can be attributed to rather short intervals used for estimation (cf. Figure 2 for $t <$ 500).

Next, we would like to compare the performance of parametric and adaptive estimation methods by means of absolute PE: first for the prediction horizon of one day, $\mathscr{H} = \{1\}$, and later for prediction two weeks ahead, $\mathscr{H} = \{1, \ldots, 10\}$. To make the results easier to decipher,

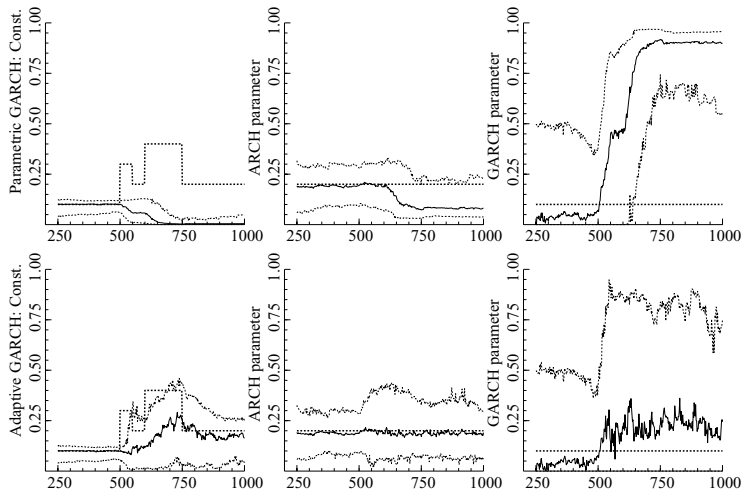


Figure 2. Parameter values estimated by the parametric (top row) and locally adaptive (bottom row) GARCH methods.

we present in what follows PEs averaged over the past month (21 days). The absolute-PE criterion was also used to determine the optimal values of parameters r and ρ (jointly across all simulations and for all $t = 250, \ldots, 1000$). The results differ for different models: $r = 0.5$, $\rho = 0.5$ for local constant, $r = 0.5$, $\rho = 1.0$ for local ARCH, and $r = 0.5$, $\rho = 1.5$ for local GARCH.

Let us now compare the adaptively estimated local constant, local ARCH and local GARCH models with the parametric GARCH, which is the best performing parametric model in this set-up. Forecasting one period ahead, the average PEs for all methods and the median lengths of the selected time-homogeneous intervals for adaptive methods are presented on Figure 3 for $t \in [250, 1000]$. First of all, let us observe in the case of the simplest local constant model that even the (median) estimated interval of homogeneity at the end of the first homogeneous period, $1 \leq t < 500$, can actually be shorter than the true one. The reason is that the probability of some 5 or 10 subsequent observations used as I_0 having their sample variance very different from the underlying one increases with the length of the series.

Next, one can notice that all methods are sensitive to jumps in volatility, especially to the first one at $t = 500$: the parametric ones because they ignore a structural break, the adaptive ones because they use a small amount of data after a structural change. In general, the local GARCH performs rather similarly to the parametric GARCH for $t < 650$ because it uses all historical data. After initial volatility jumps, the local GARCH, however, outperforms the parametric one, $650 < t < 775$. Following the last jump at $t = 750$, where the volatility level returns closer to the initial one, the parametric GARCH is best of all methods for some time, $775 < t < 850$, until the adaptive estimation procedure detects the (last) break, and after it, 'collects' enough observations for estimation. Then the local GARCH and local ARCH become preferable to the parametric model again, $850 < t$. Interestingly, the local ARCH approximation performs almost as well as both GARCH methods and even outperforms them shortly after structural breaks (except for break at $t = 750$), $600 < t < 775$ and $850 < t < 1000$. Finally, the local constant

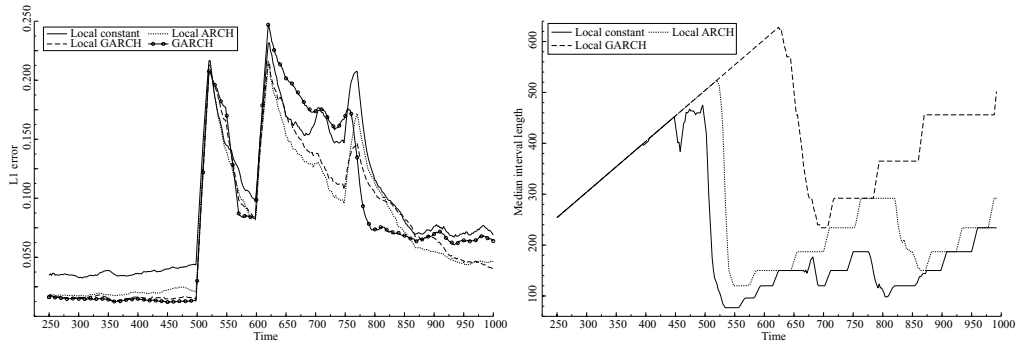


Figure 3. Left-hand panel: Low GARCH-effect simulations—absolute prediction errors one period ahead. Right-hand panel: The median lengths of the adaptively selected intervals.

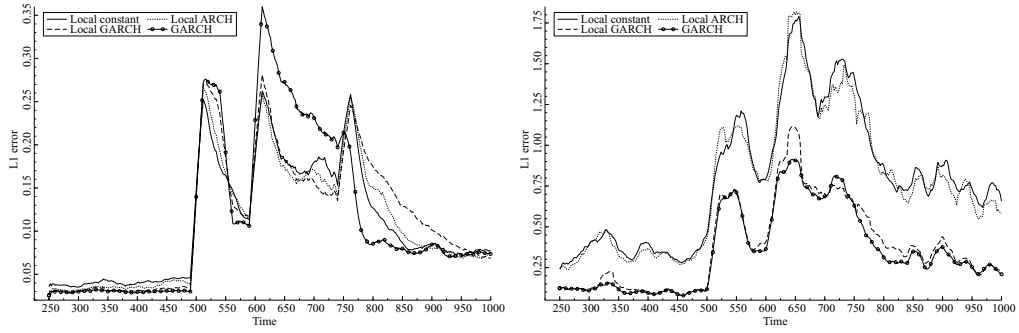


Figure 4. Left-hand panel: Low GARCH-effect simulations—absolute prediction errors 10 periods ahead. Right-hand panel: High GARCH-effect simulations—absolute prediction errors one period ahead.

volatility is lacking behind the other two adaptive methods whenever there is a longer time period without a structural break, but keeps up with them in periods with frequent volatility changes, $500 < t < 650$. All these observations can be documented also by the absolute PE averaged over the whole period $250 \leq t \leq 1000$ (we refer to it as the global PE from now on): the smallest PE is achieved by local ARCH (0.075), then by local GARCH (0.079) and the worst result is from local constant (0.094).

Additionally, all models are compared using the forecasting horizon of 10 days. Most of the results are the same (e.g. parameter estimates) or similar (e.g. absolute PE) to forecasting one period ahead due to the fact that all models rely on at most one past observation. The absolute PEs averaged over one month are summarized for $t \in [250, 1000]$ on Figure 4, which reveals that the difference between local constant volatility, local ARCH and local GARCH models are smaller in this case. As a result, it is interesting to note that: (i) the local constant model becomes a viable alternative to the other methods (it has in fact the smallest global PE 0.107 from all adaptive methods) and (ii) the local ARCH model still outperforms the local GARCH (global

PEs are 0.108 and 0.116, respectively) even though the underlying model is GARCH (with a small value of $\beta = 0.1$ however).

5.2.2. High GARCH effect. Let us now discuss the high GARCH-effect model. One would expect much more prevalent behaviour of both GARCH models, since the underlying GARCH parameter is higher and the changes in the volatility level ω are likely to be small compared to overall volatility fluctuations. Note that the optimal values of tuning constant r and ρ differ from the low GARCH-effect simulations: $r = 0.5$, $\rho = 1.5$ for local constant; $r = 0.5$, $\rho = 1.5$ for local ARCH; and $r = 1.0$, $\rho = 0.5$ for local GARCH.

Comparing the absolute PEs for the one-period-ahead forecast at each time point (Figure 4) indicates that the adaptive and parametric GARCH estimations perform approximately equally well. On the other hand, both the parametric and adaptively estimated ARCH and constant volatility models are lacking significantly. Unreported results confirm, similarly to the low GARCH-effect simulations, that the differences among method are much smaller once a longer prediction horizon of 10 days is used.

6. APPLICATIONS

The proposed adaptive pointwise estimation method will be now applied to real time series consisting of the log-returns of the DAX and S&P 500 stock indices (Sections 6.1 and 6.2). We will again summarize the results concerning both parametric and adaptive methods by the absolute PEs one day ahead averaged over one month. As a benchmark, we employ the parametric GARCH estimated using the last two years of data (500 observations). Since we however do not have the underlying volatility process now, it is approximated by squared returns. Despite being noisy, this approximation is unbiased and provides usually the correct ranking of methods (Andersen and Bollerslev, 1998).

6.1. DAX analysis

Let us now analyse the log-returns of the German stock index DAX from January 1990 till December 2002 depicted at the top of Figure 5. Several periods interesting for comparing the performance of parametric and adaptive pointwise estimates are selected since results for the whole period might be hard to decipher at once.

First, consider the estimation results for years 1991 to 1996. Contrary to later periods, there are structural breaks practically immediately detected by all adaptive methods (July 1991 and June 1992; cf. Stapf and Werner, 2003). For the local GARCH, this differs from less pronounced structural changes discussed later, which are typically detected only with delays of several months. One additional break detected by all methods occurs in October 1994. Note that parameters r and ρ were $r = 0.5$, $\rho = 1.5$ for local constant, $r = 1.0$, $\rho = 1.0$ for local ARCH, and $r = 0.5$, $\rho = 1.5$ for local GARCH.

The results for the period 1991–96 are summarized in the left bottom panel of Figure 5, which depicts the PEs of each adaptive method relative to the PEs of parametric GARCH. First, one can notice that the local constant and local ARCH approximations are preferable till July 1991, where we have less than 500 observations. After the detection of the structural change in June 1991, all adaptive methods are shortly worse than the parametric GARCH due to the limited amount of data used, but then outperform the parametric GARCH till the next structural break in the second half of 1992. A similar behaviour can be observed after the break detected in October 1994,

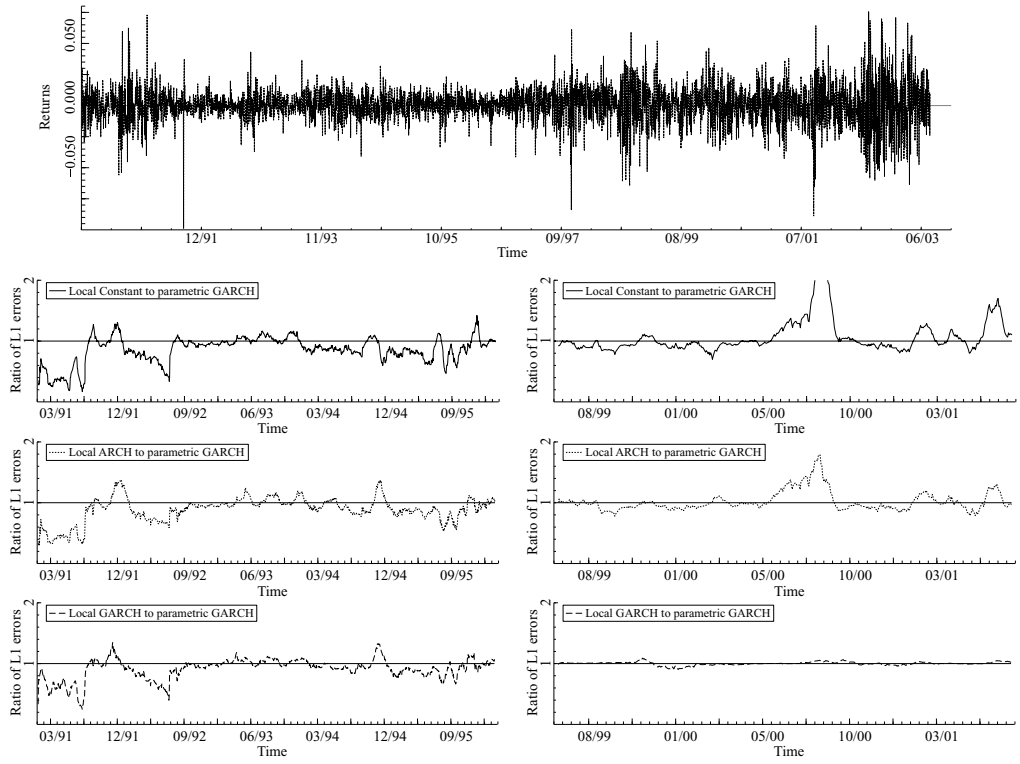


Figure 5. Top panel: The log-returns of DAX series. Bottom panels: The absolute prediction errors of the pointwise adaptive methods relative to the parametric GARCH errors for predictions one period ahead.

where the local constant and local ARCH models actually outperform both the parametric and adaptive GARCH. In the other parts of the data, the performance of all methods is approximately the same, and even though the adaptive GARCH is overall better than the parametric one, the most interesting fact is that the adaptively estimated local constant and local ARCH models perform equally well. In terms of the global PE, the local constant is best (0.829), followed by the local ARCH (0.844) and local GARCH (0.869). This closely corresponds to our findings in simulation study with low GARCH effect in Section 5.2. Note that for other choices of r and ρ, the global PEs are at most 0.835 and 0.851 for the local constant and local ARCH, respectively. This indicates low sensitivity to the choice of these parameters.

Next, we discuss the estimation results for years 1999 to 2001 ($r = 1.0$ for all methods now). After the financial markets were hit by the Asian crisis in 1997 and the Russian crisis in 1998, the market headed to a more stable state in year 1999. The adaptive methods detected the structural breaks in the autumn of 1997 and 1998. The local GARCH detected them, however, with more than a one-year delay—only during 1999. The results in Figure 5 (right bottom panel) confirm that the benefits of the adaptive GARCH are practically negligible compared to the parametric GARCH in such a case. On the other hand, the local constant and ARCH methods perform slightly better than both GARCH methods during the first presented year (July 1999 to June 2000). From July 2000, the situation becomes just the opposite and the performance

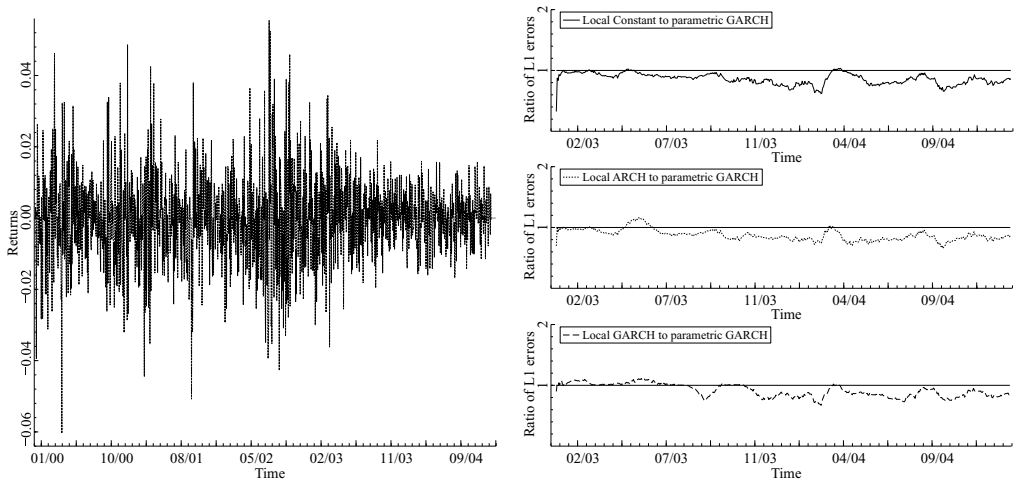


Figure 6. Left-hand panel: The log-returns of S&P 500. Right-hand panel: The absolute prediction errors of the pointwise adaptive methods relative to the parametric GARCH errors for predictions one period ahead.

of the GARCH models is better (parametric and adaptive GARCH estimates are practically the same in this period since the last detected structural change occurred approximately two years ago). Together with previous results, this opens the question of model selection among adaptive procedures as different parametric approximations might be preferred in different time periods. Judging by the global PE, the local ARCH provides slightly better predictions on average than the local constant and local GARCH—despite the 'peak' of the PE ratio in the second half of year 2000 (see Figure 5). This, however, depends on the specific choice of loss Λ in (3.6).

Finally, let us mention that the relatively similar behaviour of the local constant and local ARCH methods is probably due to the use of ARCH(1) model, which is not sufficient to capture more complex time developments. Hence, ARCH(p) might be a more appropriate interim step between the local constant and GARCH models.

6.2. S&P 500

Now we turn our attention to more recent data regarding the S&P 500 stock index considered from January 2000 to December 2004; see Figure 6. This period is marked by many substantial events affecting the financial markets, ranging from September 11, 2001, terrorist attacks and the war in Iraq (2003) to the crash of the technology stock-market bubble (2000–02). For the sake of simplicity, a particular time period is again selected: year 2003 representing a more volatile period (the war in Iraq) and year 2004 being a less volatile period. All adaptive methods detected rather quickly a structural break at the beginning of 2003, and additionally they detected a structural break in the second half of 2003, although the adaptive GARCH did so with a delay of more than eight months. The ratios of monthly PE of all adaptive methods to those of the parametric GARCH from January 2003 to December 2004 are summarized on Figure 6 ($r = 0.5$ and $\rho = 1.5$ for all methods).

In the beginning of year 2003, corresponding with 2002 to a more volatile period (see Figure 6), all adaptive methods perform as well as the parametric GARCH. In the middle of year 2003, the local constant and local ARCH models are able to detect another structural change (possibly less pronounced than the one at the beginning of 2003 because of its late detection by the adaptive GARCH). Around this period, the local ARCH shortly performs worse than the parametric GARCH. From the end of 2003 and in year 2004, all adaptive methods starts to outperform the parametric GARCH, where the reduction of the PEs due to the adaptive estimation amounts to 20% on average. All adaptive pointwise estimates exhibit a short period of instability in the first months of 2004, where their performance temporarily worsens to the level of parametric GARCH. This corresponds to 'uncertainty' of the adaptive methods about the length of the interval of homogeneity. After this short period, the performance of all adaptive methods is comparable, although the local constant performs overall best of all methods (closely followed by local ARCH) judged by the global PE.

Similarly to the low GARCH-effect simulations and to the analysis of DAX in Section 6.1, it seems that the benefit of pointwise adaptive estimation is most pronounced during periods of stability that follow an unstable period (i.e. year 2004) rather than during a presumably rapidly changing environment. The reason is that, despite possible inconsistency of parametric methods under change points, the adaptive methods tend to have a rather large variance when the intervals of time homogeneity become very short.

7. CONCLUSION

We extend the idea of adaptive pointwise estimation to parametric CH models. In the specific case of ARCH and GARCH, which represent particularly difficult cases due to high data demands and dependence of critical values on underlying parameters, we demonstrate the use and feasibility of the proposed procedure: on the one hand, the adaptive procedure, which itself depends on a number of auxiliary parameters, is shown to be rather insensitive to their choice, and on the other hand, it facilitates the global selection of these parameters by means of fit or forecasting criteria. The real-data applications highlight the flexibility of the proposed time-inhomogeneous models since even simple varying-coefficients models such as constant volatility and ARCH(1) can outperform standard parametric methods such as GARCH(1,1). Finally, the relatively small differences among the adaptive estimates based on different parametric approximations indicate that, in the context of adaptive pointwise estimation, it is sufficient to concentrate on simpler and less data-intensive models such as ARCH(p), $0 \leq p \leq 3$, to achieve good forecasts.

ACKNOWLEDGMENTS

This research was supported by the Deutsche Forschungsgemeinschaft through the SFB 649 'Economic Risk'.

REFERENCES

Andersen, T. G. and T. Bollerslev (1998). Answering the skeptics: yes, standard volatility models do provide accurate forecasts. *International Economic Review 39*, 885–905.

Andreou, E. and E. Ghysels (2002). Detecting multiple breaks in financial market volatility dynamics. *Journal of Applied Econometrics 17*, 579–600.

Andreou, E. and E. Ghysels (2006). Monitoring disruptions in financial markets. *Journal of Econometrics 135*, 77–124.

Andrews, D. W. K. (1993). Tests for parameter instability and structural change with unknown change point. *Econometrica 61*, 821–56.

Bai, J. and P. Perron (1998). Estimating and testing linear models with multiple structural changes. *Econometrica 66*, 47–78.

Beltratti, A. and C. Morana (2004). Structural change and long-range dependence in volatility of exchange rates: either, neither or both? *Journal of Empirical Finance 11*, 629–58.

Bollerslev, T. (1986). Generalized autoregressive conditional heteroskedasticity. *Journal of Econometrics 31*, 307–27.

Cai, Z., J. Fan and Q. Yao (2000). Functional coefficient regression models for nonlinear time series. *Journal of the American Statistical Association 95*, 941–56.

Chen, J. and A. K. Gupta (1997). Testing and locating variance changepoints with application to stock prices. *Journal of the American Statistical Association 92*, 739–47.

Chen, R. and R. J. Tsay (1993). Functional-coefficient autoregressive models. *Journal of the American Statistical Association 88*, 298–308.

Cheng, M.-Y., J. Fan and V. Spokoiny (2003). Dynamic nonparametric filtering with application to volatility estimation. In M. G. Akritas and D. N. Politis (Eds.), *Recent Advances and Trends in Nonparametric Statistics*, 315–33. Amsterdam: Elsevier.

Diebold, F. X. and A. Inoue (2001). Long memory and regime switching. *Journal of Econometrics 105*, 131–59.

Doornik, J. A. (2002). Object-oriented programming in econometrics and statistics using Ox: a comparison with C++, Java and C#. In S. S. Nielsen (Ed.), *Programming Languages and Systems in Computational Economics and Finance*, 115–47. Dordrecht: Kluwer.

Engle, R. F. (1982). Autoregressive conditional heteroscedasticity with estimates of the variance of United Kingdom inflation. *Econometrica 50*, 987–1008.

Fan, J. and W. Zhang (2008). Statistical models with varying coefficient models. *Statistics and Its Interface 1*, 179–95.

Francq, C. and J.-M. Zakoian (2007). Quasi-maximum likelihood estimation in GARCH processes when some coefficients are equal to zero. *Stochastic Processes and their Applications 117*, 1265–84.

Glosten, L. R., R. Jagannathan and D. E. Runkle (1993). On the relation between the expected value and the volatility of the nominal excess return on stocks. *Journal of Finance 48*, 1779–801.

Hansen, B. and S.-W. Lee (1994). Asymptotic theory for the GARCH(1,1) quasi-maximum likelihood estimator. *Econometric Theory 10*, 29–53.

Härdle, W., H. Herwatz and V. Spokoiny (2003). Time inhomogeneous multiple volatility modelling. *Journal of Financial Econometrics 1*, 55–99.

Herwatz, H. and H. E. Reimers (2001). Empirical modeling of the DEM/USD and DEM/JPY foreign exchange rate: structural shifts in GARCH-models and their implications. 2001–83, Discussion Paper SFB 373, Humboldt-Univerzität zu Berlin, Germany.

Hillebrand, E. (2005). Neglecting parameter changes in GARCH models. *Journal of Econometrics 129*, 121–38.

Kokoszka, P. and R. Leipus (2000). Change-point estimation in ARCH models. *Bernoulli 6*, 513–39.

Laurent, S. and J.-P. Peters (2006). *G@RCH 4.2, Estimating and Forecasting ARCH Models*. London: Timberlake Consultants Press.

Mercurio, D. and V. Spokoiny (2004). Statistical inference for time-inhomogeneous volatility models. *Annals of Statistics 32*, 577–602.

Mikosch, T. and C. Starica (1999). Change of structure in financial time series, long range dependence and the GARCH model. Working Paper, Department of Statistics, University of Pennsylvania. See `http://citeseer.ist.psu.edu/mikosch99change.html`.

Mikosch, T. and C. Starica (2004). Changes of structure in financial time series and the GARCH model. *Revstat Statistical Journal 2*, 41–73.

Nelson, D. B. (1991). Conditional heteroskedasticity in asset returns: a new approach. *Econometrica 59*, 347–70.

Pesaran, M. H. and A. Timmermann (2004). How costly is it to ignore breaks when forecasting the direction of a time series? *International Journal of Forecasting 20*, 411–25.

Sentana, E. (1995). Quadratic ARCH models. *Review of Economic Studies 62*, 639–61.

Spokoiny, V. (1998). Estimation of a function with discontinuities via local polynomial fit with an adaptive window choice. *Annals of Statistics 26*, 1356–78.

Spokoiny, V. (2009a). Multiscale local change-point detection with applications to value-at-risk. *Annals of Statistics 37*, 1405–36.

Spokoiny, V. (2009b). Parameter estimation in time series analysis. WIAS Preprint No. 1404, Weierstrass Institute for Applied Analysis and Stochastics, Berlin, Germany.

Stapf, J. and T. Werner (2003). How wacky is DAX? The changing structure of German stock market volatility. Discussion Paper 2003/18, Deutsche Bundesbank, Germany.

Taylor, S. J. (1986). *Modeling Financial Time Series*. Chichester: Wiley.

APPENDIX: PROOFS

Proof of Corollary 2.1: Given the choice of $\mathfrak{z}_\alpha$, it directly follows from (2.5). □

Proof of Theorem 3.1: Consider the event $\mathscr{B}_k = \{\hat{I} = I_{k-1}\}$ for some $k \le K$. This particularly means that I_{k-1} is accepted while $I_k = [T - m_k + 1, T]$ is rejected; i.e. there is $I' = [t', T] \subseteq I_k$ and $\tau \in \mathscr{T}(I_k)$ such that $T_{I_k,\tau} > \mathfrak{z}_k = \mathfrak{z}_{I_k,\mathscr{T}(I_k)}$. For every fixed $\tau \in \mathscr{T}(I_k)$ and $J = I_k \setminus [\tau + 1, T]$, $J^c = [\tau + 1, T]$, it holds by definition of $T_{I_k,\tau}$ that

$$T_{I_k,\tau} \le L_J(\widetilde{\boldsymbol{\theta}}_J) + L_{J^c}(\widetilde{\boldsymbol{\theta}}_{J^c}) - L_I(\boldsymbol{\theta}_0) = L_J(\widetilde{\boldsymbol{\theta}}_J, \boldsymbol{\theta}_0) + L_{J^c}(\widetilde{\boldsymbol{\theta}}_{J^c}, \boldsymbol{\theta}_0).$$

This implies by Theorem 2.1 that $\boldsymbol{P}_{\boldsymbol{\theta}_0}(T_{I_k,\tau} > 2\mathfrak{z}) \le \exp\{\mathfrak{e}(\lambda, \boldsymbol{\theta}_0) - \lambda\mathfrak{z}\}$. Now,

$$\boldsymbol{P}_{\boldsymbol{\theta}_0}(\mathscr{B}_k) \le \sum_{t'=T-m_k+1}^{T-m_0} \sum_{\tau=t'+1}^{T-m_0+1} 2\exp\{\mathfrak{e}(\lambda, \boldsymbol{\theta}_0) - \lambda\mathfrak{z}_k/2\} \le 2\frac{m_k^2}{2}\exp\{\mathfrak{e}(\lambda, \boldsymbol{\theta}_0) - \lambda\mathfrak{z}_k/2\}.$$

Next, by the Cauchy–Schwartz inequality

$$\begin{aligned}\boldsymbol{E}_{\boldsymbol{\theta}_0}|L_{I_K}(\widetilde{\boldsymbol{\theta}}_{I_K}, \hat{\boldsymbol{\theta}})|^r &= \sum_{k=1}^{K} \boldsymbol{E}_{\boldsymbol{\theta}_0}[|L_{I_K}(\widetilde{\boldsymbol{\theta}}_{I_K}, \widetilde{\boldsymbol{\theta}}_{k-1})|^r \mathbf{1}(\mathscr{B}_k)] \\ &\le \sum_{k=1}^{K} \boldsymbol{E}_{\boldsymbol{\theta}_0}^{1/2}|L_{I_K}(\widetilde{\boldsymbol{\theta}}_{I_K}, \widetilde{\boldsymbol{\theta}}_{k-1})|^{2r} \boldsymbol{P}_{\boldsymbol{\theta}_0}^{1/2}(\mathscr{B}_k).\end{aligned}$$

Under the conditions of Theorem 2.1, it follows similarly to (2.6) that

$$\boldsymbol{E}_{\boldsymbol{\theta}_0}|L_{I_K}(\widetilde{\boldsymbol{\theta}}_{I_K}, \widetilde{\boldsymbol{\theta}}_{k-1})|^{2r} \le (m_K/m_{k-1})^{2r}\mathfrak{R}_{2r}^*(\boldsymbol{\theta}_0)$$

for some constant $\mathfrak{R}^*_{2r}(\boldsymbol{\theta}_0)$ and $k = 1, \ldots, K$, and therefore,

$$\boldsymbol{E}_{\boldsymbol{\theta}_0}|L_{I_K}(\widetilde{\boldsymbol{\theta}}_{I_K}, \hat{\boldsymbol{\theta}})|^r \le [\mathfrak{R}^*_{2r}(\boldsymbol{\theta}_0)]^{1/2} \sum_{k=1}^{K} m_k (m_K/m_{k-1})^r \exp\{\mathfrak{e}(\lambda, \boldsymbol{\theta}_0)/2 - \lambda \mathfrak{z}_k/4\}$$

and the result follows by simple algebra provided that $a_1\lambda/4 \ge 1$ and $a_2\lambda/4 > 2$. □

LEMMA A.1. *Let $\boldsymbol{P}$ and $\boldsymbol{P}_0$ be two measures such that the Kullback–Leibler divergence $\boldsymbol{E}\log(d\boldsymbol{P}/d\boldsymbol{P}_0)$, satisfies $\boldsymbol{E}\log(d\boldsymbol{P}/d\boldsymbol{P}_0) \le \Delta < \infty$. Then for any random variable ζ with $\boldsymbol{E}_0\zeta < \infty$, it holds that $\boldsymbol{E}\log(1+\zeta) \le \Delta + \boldsymbol{E}_0\zeta$.*

Proof: By simple algebra one can check that for any fixed y the maximum of the function $f(x) = xy - x\log x + x$ is attained at $x = e^y$ leading to the inequality $xy \le x\log x - x + e^y$. Using this inequality and the representation $\boldsymbol{E}\log(1+\zeta) = \boldsymbol{E}_0\{Z\log(1+\zeta)\}$ with $Z = d\boldsymbol{P}/d\boldsymbol{P}_0$ we obtain

$$\begin{aligned}\boldsymbol{E}\log(1+\zeta) &= \boldsymbol{E}_0\{Z\log(1+\zeta)\} \le \boldsymbol{E}_0(Z\log Z - Z) + \boldsymbol{E}_0(1+\zeta)\\ &= \boldsymbol{E}_0(Z\log Z) + \boldsymbol{E}_0\zeta - \boldsymbol{E}_0 Z + 1.\end{aligned}$$

It remains to note that $\boldsymbol{E}_0 Z = 1$ and $\boldsymbol{E}_0(Z\log Z) = \boldsymbol{E}\log Z$. □

Proof of Theorem 4.1: Lemma A.1 applied with $\zeta = \varrho(\hat{\boldsymbol{\theta}}, \boldsymbol{\theta})/\boldsymbol{E}_{\boldsymbol{\theta}}\varrho(\hat{\boldsymbol{\theta}}, \boldsymbol{\theta})$ yields the result in the view of

$$\begin{aligned}\boldsymbol{E}_{\boldsymbol{\theta}}(Z_{I,\boldsymbol{\theta}}\log Z_{I,\boldsymbol{\theta}}) &= \boldsymbol{E}\log Z_{I,\boldsymbol{\theta}} = \boldsymbol{E}\sum_{t\in I}\log\frac{p[Y_t, g(X_t)]}{p[Y_t, g(X_t(\boldsymbol{\theta}))]}\\ &= \boldsymbol{E}\sum_{t\in I}\boldsymbol{E}\left\{\log\frac{p[Y_t, g(X_t)]}{p[Y_t, g(X_t(\boldsymbol{\theta}))]}\,\middle|\,\mathscr{F}_{t-1}\right\} = \boldsymbol{E}\Delta_{I_k}(\boldsymbol{\theta}).\end{aligned}$$ □

Proof of Corollary 4.1: It is Theorem 4.1 formulated for $\varrho(\boldsymbol{\theta}', \boldsymbol{\theta}) = L_I(\boldsymbol{\theta}', \boldsymbol{\theta})$. □

Proof of Theorem 4.2: The first inequality follows from Corollary 4.1, the second one from condition (3.4) and the property $x \ge \log x$ for $x > 0$. □

Proof of Theorem 4.3: Let $\hat{k} = k > k^*$. This means that I_k is not rejected as homogeneous. Next, we show that for every $k > k^*$ the inequality $T_{I_k,\tau} \le T_{I_k,\mathscr{T}(I_k)} \le \mathfrak{z}_k$ with $\tau = T - m_{k^*} = T - |I_{k^*}|$ implies $L_{I_{k^*}}(\widetilde{\boldsymbol{\theta}}_{I_{k^*}}, \widetilde{\boldsymbol{\theta}}_{I_k}) \le \mathfrak{z}_{k^*}$. Indeed with $J = I_k \backslash I_{k^*}$, this means that, by construction, $\mathfrak{z}_k \le \mathfrak{z}_{k^*}$ for $k > k^*$ and

$$\mathfrak{z}_k \ge T_{I_k,\tau} = L_{I_{k^*}}(\widetilde{\boldsymbol{\theta}}_{I_{k^*}}, \widetilde{\boldsymbol{\theta}}_{I_k}) + L_J(\widetilde{\boldsymbol{\theta}}_J, \widetilde{\boldsymbol{\theta}}_{I_k}) \ge L_{I_{k^*}}(\widetilde{\boldsymbol{\theta}}_{I_{k^*}}, \widetilde{\boldsymbol{\theta}}_{I_k}).$$

It remains to note that

$$|L_{I_{k^*}}(\widetilde{\boldsymbol{\theta}}_{I_{k^*}}, \hat{\boldsymbol{\theta}})|^r \le |L_{I_{k^*}}(\widetilde{\boldsymbol{\theta}}_{I_{k^*}}, \hat{\boldsymbol{\theta}}_{I_{k^*}})|^r \mathbf{1}(\hat{k} < k^*) + \mathfrak{z}^r_{k^*}\mathbf{1}(\hat{k} > k^*),$$

which obviously yields the assertion. □

AStA Adv Stat Anal (2009) 93: 387–402
DOI 10.1007/s10182-009-0115-4

ORIGINAL PAPER

Dynamic semiparametric factor models in risk neutral density estimation

Enzo Giacomini · Wolfgang Härdle · Volker Krätschmer

Received: 1 March 2009 / Accepted: 31 August 2009 / Published online: 18 September 2009

Abstract Dynamic semiparametric factor models (DSFM) simultaneously smooth in space and are parametric in time, approximating complex dynamic structures by time invariant basis functions and low dimensional time series. In contrast to traditional dimension reduction techniques, DSFM allows the access of the dynamics embedded in high dimensional data through the lower dimensional time series. In this paper, we study the time behavior of risk assessments from investors facing random financial payoffs. We use DSFM to estimate risk neutral densities from a dataset of option prices on the German stock index DAX. The dynamics and term structure of risk neutral densities are investigated by Vector Autoregressive (VAR) methods applied on the estimated lower dimensional time series.

Keywords Dynamic factor models · Dimension reduction · Risk neutral density

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., 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.

E. Giacomini (✉) · W. Härdle · V. Krätschmer
CASE—Center for Applied Statistics and Economics, Humboldt-Universität zu Berlin, Spandauerstr. 1, 10178 Berlin, Germany
e-mail: enzogiacomini@gmail.com

V. Krätschmer
Institute of Mathematics, Technische Universität Berlin, Straße des 17. Juni 136, 10623 Berlin, Germany

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 as

$$Y_{jt} = \mathcal{F}_t(X_{jt}) + \varepsilon_{jt}, \tag{1.1}$$

where ε_{jt} denote unknown zero-mean error terms and $\{\mathcal{F}_t\}$ are realizations of $\mathcal{F}$. Each sample $\mathcal{S}_t = \{(Y_{jt}, X_{jt}) : j = 1, \ldots, J_t\}$, $t = 1, \ldots, T$, may correspond to observations on, e.g., different individuals, time periods or experimental conditions. Examples in biomedicine are measurements of growth curves and brain potentials across individuals, see Kneip and Gasser (1992) and Gasser and Kneip (1995), in econometrics such are expenditures across households and implied volatilities across trading days, see Kneip (1994) and Fengler et al. (2007).

A large branch of functional data analysis concentrates on approximating $\mathcal{F}$ by lower dimensional objects. Distributions on function spaces are highly complex objects and dimension reduction techniques present a feasible and interpretable approach for investigating 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 Rice and Silverman (1991) and Ramsay and Dalzell (1991).

Asymptotic results on FPCA have been obtained by Dauxois et al. (1982) and Hall et al. (2006) for observed functional data $\{\mathcal{F}_t\}$. For non-observable data, the standard approach is to perform FPCA on presmoothed $\{\widehat{\mathcal{F}}_t\}$, see Benko et al. (2009) for recent developments. In practical applications, however, presmoothing may suffer from design-sparseness, see Cont and 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 the dynamics of the unobserved random functions, calling for dimension reduction techniques that smooth in space and are parametric in time.

In this paper, we investigate the dynamics of $\{\mathcal{F}_t\}$ by reducing dimensionality without presmoothing. $\mathcal{F}_t$ is considered as a linear combination of $L + 1 \ll T$ unknown smooth basis functions $m_l \in L_2(\mathcal{X})$, $l = 0, \ldots, L$:

$$\mathcal{F}_t(X_{jt}) = \sum_{l=0}^{L} Z_{lt} m_l(X_{jt}), \tag{1.2}$$

where $Z_t = (Z_{0t}, \ldots, Z_{Lt})^\top$ is an unobservable random vector taking values on $\mathbb{R}^{L+1}$ with $Z_{0t} = 1$. Defining the tuple of functions $m = (m_0, \ldots, m_L)^\top$, the Dynamic Semiparametric Factor Model (DSFM) reads as

$$Y_{jt} = Z_t^\top m(X_{jt}) + \varepsilon_{jt}. \tag{1.3}$$

The basis functions are estimated nonparametrically avoiding specification issues. Their estimation is performed simultaneously with Z_t, i.e., the smoothing is transferred directly to m_l and design-sparseness issues become secondary. In addition, the random process $\{Z_t\}$ is allowed to be non-stationary. Park et al. (2009) show that

under (1.2) the autocorrelation structures of $\{\widehat{Z}_t\}$ and $\{Z_t\}$ are asymptotically equivalent; therefore, no loss is incurred by inferring the dynamics from the estimated $\{\widehat{Z}_t\}$, and there is no payment for not knowing the true $\{Z_t\}$. This result is essential for investigating cointegration between dynamical systems, see Brüggemann et al. (2008) for an econometric application.

Note that the common regressors model, Kneip (1994), also represents unobservable functions by (1.2). There are, however, crucial differences between the DSFM and common regressors:

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.

Thus DSFM goes beyond traditional dimension reductions techniques (FPCA and common regressors) as it captures structural dynamics embedded in the observations.

In economics, there is substantial interest in the behavior (over time) of investors facing risks and its relation to macroeconomic and financial indicators. The knowledge about the dynamics of risk assessments from investors is essential for many applications ranging from pricing of illiquid instruments to risk management.

Option prices contain information on risk assessments from investors facing future financial payoffs, summarized in the risk neutral densities q, see Ait-Sahalia and Lo (1998). An European call option with price C_t at time $0 \le t \le T$, maturity date $T > 0$ and strike $K > 0$ is a financial instrument that delivers the random payoff $(S_T - K)^+$ at time T where S_t is the price of an underlying asset at time 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

$$q_{t,T}(s_T|s_t) = e^{r(T-t)} \frac{\partial^2 C_t(s_t, r, K, T-t)}{\partial K^2}\bigg|_{K=s_T}, \tag{1.4}$$

where $r > 0$ is interest rate, see Sect. 4 for details.

We estimate risk neutral densities based on observed intraday prices of calls on the German stock index (DAX). Each observation consists of a price Y_{jt} on a design point $X_{jt} = (\kappa_{jt}, \tau_{jt})^\top$ where $j = 1, \ldots, J_t$, denote the transactions at day $t = 1, \ldots, T$, κ is the moneyness, a monotone transformation of strikes K, and $\tau = T - t$ is the time to maturity associated with the option. Stock exchange regulations impose prespecified values for tradable maturities resulting in degenerated designs, see Fig. 1.

Following Ait-Sahalia and Lo (1998) and Fengler et al. (2007), call prices are transformed into log-implied volatilities $\widetilde{Y}_{jt} = \log C_{\mathrm{BS}}^{-1}(Y_{jt})$, where C_{BS} is the Black–Scholes call price function defined in Sect. 4. These are assumed as discretized noisy values of the log-implied volatility surface evaluated at $\{X_{jt}\}$:

$$\widetilde{Y}_{jt} = \log \mathcal{V}_t(X_{jt}) + \varepsilon_{jt}, \tag{1.5}$$

where $\mathcal{V} \in L_2(\mathcal{X})$, $\mathcal{X} \subset \mathbb{R}^2_+$, is a smooth random function, called the implied volatility surface, and ε_{jt} is an error term. The realizations $\{\mathcal{V}_t\}$ are filtered out from the data with DSFM and, remarking that C_{BS} is a function of K, the risk neutral densities are

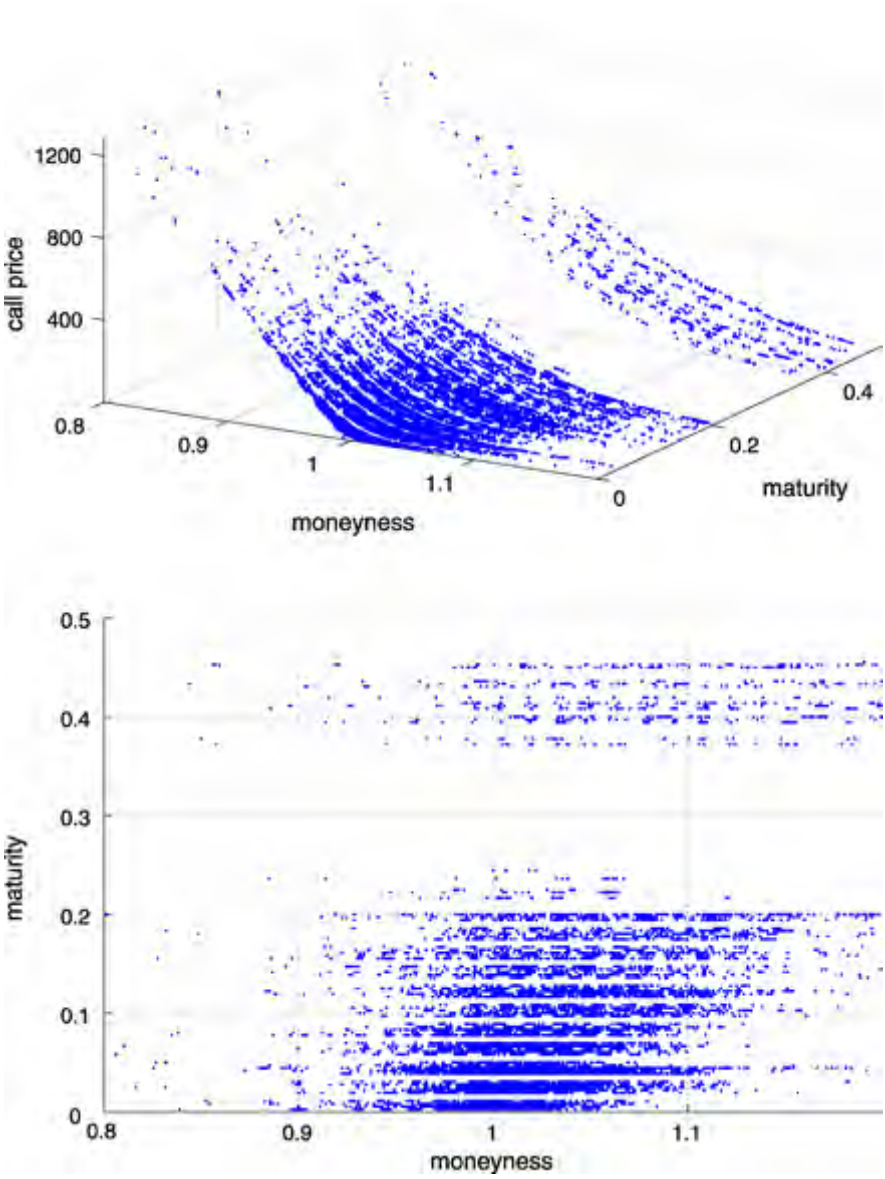


Fig. 1 Samples S_t, $t = 1, \ldots, 22$, of DAX call prices traded on January 2001 (*left*). Corresponding unbalanced design $\{X_{jt}\}$ (*right*)

obtained by (1.4) with $C_{\mathrm{BS}}(\widehat{\mathcal{V}})$ as an estimator for C_t. The dynamics of the estimated $\{\widehat{q}_{t,T}\}$ is analyzed based on the autocorrelation structure of $\{\widehat{Z}_t\}$.

In the sequel, the DSFM estimation method and its asymptotic properties are described (Sect. 2). In Sect. 3, the risk neutral densities are defined, and in Sect. 4 they are estimated from observed prices of European call options on the DAX index

(ODAX dataset). Their dynamic structure is then analyzed 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} are unknown error terms with $E[\varepsilon_{jt}] = 0$ and $E[\varepsilon_{jt}^2] < \infty$. The variables $X_{11}, \ldots, X_{T,J_T}$, $\varepsilon_{1,1}, \ldots, \varepsilon_{T,J_T}$ 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 Park et al. (2009), 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 an $((L+1) \times K)$ 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 squares estimators as

$$(\widehat{\Gamma}, \widehat{Z}) = \arg \min_{\Gamma \in \mathcal{G}, Z \in \mathcal{Z}} \sum_{t=1}^{T} \sum_{j=1}^{J} \{Y_{jt} - Z_t^\top \Gamma \psi(X_{jt})\}^2, \tag{2.2}$$

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 behavior of the least squares estimators $(\widehat{\Gamma}, \widehat{Z})$. See Park et al. (2009) for the proof.

Theorem 2.1 *Suppose that DSFM holds and that* $(\widehat{\Gamma}, \widehat{Z})$ *is defined by* (2.2). *Under Assumptions* (A1)–(A8), *see Appendix, it holds for* $K, J \to \infty$:

$$\frac{1}{T} \sum_{1 \le t \le T} \left\| \widehat{Z}_t^\top \widehat{\Gamma} - Z_t^\top \Gamma^* \right\|^2 = \mathcal{O}_P\left(\delta_K^2 + \xi^2\right).$$

See (A5) and (A8) for the definitions of δ_K and ξ. Note that the model (2.1) is only identifiable up to linear transformations. Consider an $((L+1) \times (L+1))$ regular matrix $B = (b_{ij})$ with $b_{1j} = \delta_{1j}$ and $b_{i1} = \delta_{i1}$ for $i, j = 1, \ldots, L+1$, where $\delta_{ij} = \mathbf{1}(i = j)$. Define $Z_t^* = B^\top Z_t$, $m^* = B^{-1} m$. Then from (1.2)

$$\mathcal{F}_t(X) = Z_t^\top m(X) = Z_t^\top B B^{-1} m(X) = Z_t^{*\top} m^*(X)$$

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 $\{\widetilde{Z}_t\}$, $\widetilde{Z}_t = B^\top \widehat{Z}_t$, are asymptotically equivalent to the autocovariances of the true unobservable $\{Z_t\}$. This equivalence is transferred to classical estimation and testing procedures in the context of, e.g., vector autoregressive models and, in particular, justifies inference based on $\{\widetilde{Z}_t\}$ 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, 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* $T \to \infty$:

$$\frac{1}{T} \sum_{t=h_d}^{h_u} \widetilde{Z}_{c,t} (\widetilde{Z}_{c,t+h} - \widetilde{Z}_{c,t})^\top - \frac{1}{T} \sum_{t=h_d}^{h_u} Z_{c,t} (Z_{c,t+h} - Z_{c,t})^\top = \mathcal{O}_P\big(T^{-1/2}\big),$$

where $\widetilde{Z}_t = B^\top \widehat{Z}_t$. *Moreover,*

$$\frac{1}{T} \sum_{t=h_d}^{h_u} \widetilde{Z}_{n,t} \widetilde{Z}_{n,t+h}^\top - \frac{1}{T} \sum_{t=h_d}^{h_u} Z_{n,t} Z_{n,t+h}^\top = \mathcal{O}_P\big(T^{-1/2}\big).$$

See Park et al. (2009) for the proof. Note that, in contrast to FPCA, DSFM does not require stationarity neither for $\{Z_t\}$ nor for $\{\varepsilon_t\}$, but only weak assumptions on the average behavior of Z_t, like being a martingale difference, see Appendix.

3 Risk neutral density estimation

3.1 Risk neutral densities

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

$$C_t = e^{-r(T-t)} E^{\mathbb{Q}}\big[\Psi(S_T) | \mathcal{F}_t\big]. \tag{3.1}$$

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

$$C_t = e^{-r(T-t)} E\big[\Psi(S_T) \mathcal{K}_\pi^t(S_t, S_T) | S_t\big],$$

where E denotes the expectation under $\mathbb{P}$ and $\mathcal{K}_\pi^t(S_t, S_T) \overset{\text{def.}}{=} \frac{E[\pi|S_t,S_T]}{E[\pi|S_t]}$. The conditional risk neutral distribution of S_T is defined as

$$Q_{S_T|S_t=s_t}([S_T \le x]) \overset{\text{def.}}{=} \int_{-\infty}^{x} \mathcal{K}_\pi^t(s_t, \cdot)\, dP_{S_T|S_t=s_t}, \tag{3.2}$$

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 modeled by

$$dY_t = g(Y_t) + \rho\, dW_t^1 + \overline{\rho}\, dW_t^2,$$

where $\rho \in [-1, 1]$ is some correlation factor, $\overline{\rho} \overset{\text{def.}}{=} \sqrt{1-\rho^2}$ 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 an Ornstein–Uhlenbeck dynamics with mean reversion term $g(y) = \iota(\theta - y)$ for constants $\theta \ge 0$ and $\iota > 0$. 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, recalling (3.1), the call prices can be expressed as

$$C_t(s_t, r, K, T-t) = e^{-r(T-t)} \int (s_T - K)^+ q_{t,T}(s_T|s_t)\, ds_T.$$

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 the risk neutral density function $q_{t,T}(\cdot|s_t)$, implied by $\mathbb{Q}$ through (3.2).

3.2 Estimation

Adapting Breeden and Litzenberger (1978), one can show that the risk neutral density function $q_{t,T}(\cdot|s_t)$ is obtained as the second derivative of the call price function C_t with respect to strike K

$$q_{t,T}(s_T|s_t) = e^{r\tau} \left.\frac{\partial^2 C_t(s_t, r, K, \tau)}{\partial K^2}\right|_{K=s_T}, \tag{3.3}$$

where $\tau = T - t$ is the time to maturity.

The unknown price function C_t might be smoothed out of price observations and used in (3.3) to recover risk neutral densities. Here we follow the semiparametric approach from Ait-Sahalia and Lo (1998) where the smoothing is carried out in the space of implied volatilities.

The implied volatility surface is the function $v_t : \mathbb{R}_+^2 \to \mathbb{R}_+$ satisfying for all $(K, \tau) \in \mathbb{R}_+^2$

$$C_t(s_t, r, K, \tau) = C_{\mathrm{BS}}\big\{s_t, r, K, \tau, v_t(K, \tau)\big\}, \tag{3.4}$$

where $C_{\mathrm{BS}}(s, r, K, \tau, v) = s\Phi(d_1) - Ke^{-r\tau}\Phi(d_2)$ is the Black–Scholes price of Ψ with strike K and maturity τ, $\Phi(\cdot)$ is the cumulative distribution function of the standard normal distribution, $d_1 = \{\log(\frac{s}{K}) + (r + \frac{1}{2}v^2)\tau\}/(v\sqrt{\tau})$ and $d_2 = d_1 - v\sqrt{\tau}$.

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 (3.3) and (3.4), the functional random variable $\mathcal{H} \in L_2(\mathcal{X})$, called the risk neutral (RN) surface, is defined as

$$\begin{aligned}\mathcal{H}(s, r, K, \tau, \mathcal{V}) &= e^{r\tau} D^2 C_{\mathrm{BS}}(s, r, K, \tau, \mathcal{V}) \\ &= \varphi(d_2)\left\{\frac{1}{K\sqrt{\tau}\mathcal{V}} + \frac{2d_1}{\mathcal{V}} D\mathcal{V} + K\sqrt{\tau}\frac{d_1 d_2}{\mathcal{V}}(D\mathcal{V})^2 + K\sqrt{\tau} D^2\mathcal{V}\right\},\end{aligned} \tag{3.5}$$

where D^m denotes the mth partial derivative with respect to K and $\varphi(\cdot)$ the probability density function of the standard normal distribution. The explicit derivation of (3.5) and a detailed treatment of implied volatilities can be found in Hafner (2004) and Fengler (2005). Clearly, lower dimension objects describing $\mathcal{V}$ may be used to analyze the RN surface $\mathcal{H}$.

A functional dataset containing realizations of the implied volatility surface $\mathcal{V}$ is, however, not available, as in an exchange only discretized values of $\mathcal{V}_t$ corrupted by noise are registered from trades. On each day $t = 1, \ldots, T$ there are J_t options traded, each intraday trade $j = 1, \ldots, J_t$ corresponds to an observed option price Y_{jt} at a pair of moneyness κ and maturities τ, $X_{jt} = (\kappa_{jt}, \tau_{jt})^\top$ where $\kappa = e^{r\tau}K/s_t$. Let $C_{\mathrm{BS}}(v) = C_{\mathrm{BS}}(v; s, r, K, \tau)$ denote the Black–Scholes price as a function of v with all other arguments held constant. As $C_{\mathrm{BS}}(v)$ is continuous and monotone in v with inverse C_{BS}^{-1}, the observed implied volatility associated with trade j at day t is then $v_{jt} = C_{\mathrm{BS}}^{-1}(Y_{jt})$. Figure 2 shows the implied volatilities from options on the German Stock Index DAX traded on 2 May 2000, the sparse and degenerated design is caused by regulation imposed by stock exchanges on the tradable maturities from call options.

For numerical tractability, see Fengler et al. (2007), observations v_{jt} are transformed into log-implied volatilities $\widetilde{Y}_{jt} = \log v_{jt}$ and based on $\{(\widetilde{Y}_{jt}, X_{jt})\}$, we use DSFM to model

$$\widetilde{Y}_{jt} = Z_t^\top m(X_{jt}) + \varepsilon_{jt}. \tag{3.6}$$

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 (3.5) by $\widehat{\mathcal{H}}_t = \mathcal{H}(s_t, r, K, \tau, \widehat{\mathcal{V}}_t)$. The dynamics of the unobservable sequence of RN surfaces $\{\mathcal{H}_t\}$ implied in the observations may be investigated by analyzing the lower dimensional $\{\widehat{Z}_t\}$.

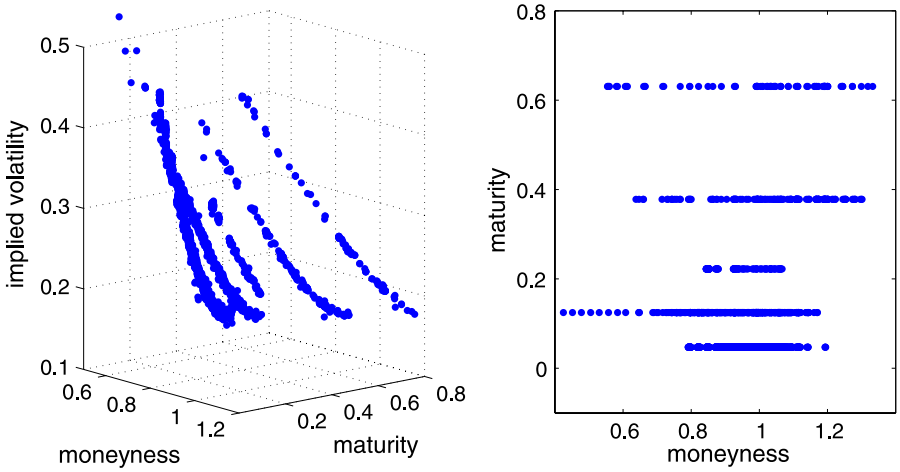


Fig. 2 Implied volatilities (*left*) and data design (*right*), ODAX on 2 May 2000

Table 1 Descriptive statistics, number of intraday observations J_t, $t = 1, \ldots, 253$

Mean	Std. dev.	Max	Min
2845.92	1589.90	11298	616

4 Application

In this section, the implied volatility and risk neutral surfaces are estimated with DSFM from intraday prices of calls on the DAX index, i.e., S_t represents the value of the DAX index at time t. The dataset contains prices observed from 1 Jan. 2001 to 1 Jan. 2002 corresponding to $T = 253$ trading days. The descriptive statistics of the number of intraday observations J_t are in Table 1, the total number of intraday observations across days is $\sum_{t=1}^{T} J_t = 720017$.

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 number of basis functions is chosen based on

$$\mathrm{EV}(L) = 1 - \frac{\sum_{t=1}^{T} \sum_{j=1}^{J_t} \{\widetilde{Y}_{jt} - \widehat{Z}_t^{\top} \widehat{m}(X_{jt})\}^2}{\sum_{t=1}^{T} \sum_{j=1}^{J_t} (\widetilde{Y}_{jt} - \overline{Y})^2},$$

where $\overline{Y} = (\sum_{t=1}^{T} \sum_{j=1}^{J_t} \widetilde{Y}_{jt}) / \sum_{t=1}^{T} J_t$. The value $\mathrm{EV}(L)$ may be interpreted as the ratio of variation explained by the model to total variation. As established by numerous simulations in Park et al. (2009), the order of the splines and number of knots have negligible influence on $\mathrm{EV}(L)$.

4.1 Simulation

The choice of the number of basis functions based on the explained variation criteria is validated by a small simulation study. Datasets $\{(Y_{jt}, X_{jt})\}$ are generated following

$$\begin{aligned} Y_{jt} &= \sum_{l=0}^{L^*} Z_{lt} m_l(X_{jt}) + \varepsilon_{jt}, \quad j = 1, \ldots, J,\ t = 1, \ldots, T, \\ \varepsilon_{jt} &\sim N\big(0, \sigma_\varepsilon^2\big), \\ X_{jt} &\sim U\big([0,1]^2\big), \end{aligned} \tag{4.1}$$

where ε_{jt} and X_{jt} are i.i.d. For $\zeta_t = (Z_{1t}, \ldots, Z_{L^*t})^\top$, with 0_d denoting the $(d \times 1)$ vector of zeros and I_d the d identity matrix we define

$$\begin{aligned} Z_t &= (1, \zeta_t)^\top, \\ \zeta_t &= A_{L^*}\zeta_{t-1} + u_t, \\ u_t &\sim N\big(0_{L^*}, \sigma_u^2 I_{L^*}\big), \end{aligned}$$

where u_t is i.i.d. and A_{L^*} is a square matrix containing the first L^* rows and L^* columns from A,

$$A = \begin{pmatrix} 0.95 & -0.2 & 0 & 0.1 \\ 0 & 0.8 & 0.1 & 0.2 \\ 0.1 & 0 & 0.6 & -0.1 \\ 0 & 0.1 & -0.2 & 0.5 \end{pmatrix}.$$

The basis functions are defined as

$$\begin{aligned} m_0(\kappa, \tau) &= 1, \\ m_1(\kappa, \tau) &= 3.46(\kappa - 0.5), \\ m_2(\kappa, \tau) &= 9.45\big\{(\kappa - 0.5)^2 + (\tau - 0.5)^2\big\} - 1.6, \\ m_3(\kappa, \tau) &= 1.41 \sin(2\pi\tau), \\ m_4(\kappa, \tau) &= 1.41 \cos(2\pi\kappa), \end{aligned}$$

and are close to orthogonal, enhancing similar choice from Park et al. (2009). The value L^* denotes the true number of dynamic basis functions.

Setting $T = 500$, $J = 100$, $\sigma_\varepsilon = 0.05$, and $\sigma_u = 0.1$, $i = 1, \ldots, 100$ samples following (4.1) are generated with $L^* = 2, 3$ and 4. Each of them is estimated by DSFM with $L = 1, \ldots, 6$, and the corresponding $\mathrm{EV}_i(L)$ is computed. The average explained variation under the true L^*, defined as $\mathcal{EV}(L; L^*) = \frac{1}{100}\sum_i \mathrm{EV}_i(L)$, is also calculated.

Table 2 shows $\mathcal{EV}(L; L^*)$ and indicates that the increase in the average explained variation between estimation with L^* and $L^* + 1$ dynamic basis functions, $\mathcal{EV}(L^* + 1; L^*) - \mathcal{EV}(L^*; L^*)$, is close to zero across values of L^*. Therefore,

Table 2 Average explained variation $\mathcal{EV}(L; L^*)$ based on 100 samples from (4.1), across number of dynamic basis functions used in the estimation L and the true L^*

$\mathcal{EV}(L; L^*)$		L^*		
		2	3	4
L	1	0.86	0.75	0.71
	2	0.99	0.90	0.89
	3	0.99	0.99	0.97
	4	0.99	0.99	0.99
	5	0.99	0.99	0.99

Table 3 Number of basis functions and explained variation

L	1	2	3	4	5
EV(L)	0.77	0.97	0.98	0.98	0.98

for DSFM estimation, we select the smallest L such that $\text{EV}(L-1) < \text{EV}(L) \approx \text{EV}(L+1)$.

4.2 Results

The implied volatility and RN surfaces are estimated with DSFM as in (3.6) with $L = 3$. Table 3 shows that the addition of the fourth or fifth dynamic basis function results in negligible increase in EV(L).

Following Fengler et al. (2007) and Park et al. (2009), the estimated $\widehat{Z}_t$ and $\widehat{m}$ are respectively transformed and orthonormalized so that $\{\widehat{Z}_{lt}^\top \widehat{m}_l\}$ has a larger contribution than $\{\widehat{Z}_{(l+1)t}^\top \widehat{m}_{l+1}\}$, $l = 1, \ldots, L-1$, to the total variation $\sum_{t=1}^{T} \int \widehat{Z}_t^\top \widehat{m}$. This transformation aims to improve the interpretation of the basis functions in the analysis of the dynamics of implied volatility surfaces. In the analysis of risk neutral surfaces dynamics, however, it does not present a clear advantage. The covariance structures from $\{\widehat{Z}_t\}$ and $\{Z_t\}$ are then asymptotically equivalent up to orthogonal transformations.

Figures 3 and 4 depict the estimated loading factors series $\{\widehat{Z}_t\}$ and basis functions $\widehat{m}_l$. The upward and downward peaks observed in $\widehat{Z}_{2t}$ occur on days 6 Feb. 2001 and 5 Nov. 2001 and are caused respectively by extremely unbalanced design and low price levels. The first day has $J_t = 1697$ observations concentrated on short maturities, while the latter has $J_t = 3268$ with very low prices at high maturities.

From (3.5), we obtain a sequence of RN surfaces $\{\widehat{\mathcal{H}}_t\}$, $t = 1, \ldots, 253$. We define $\widehat{\mathcal{H}}_t(\kappa, \tau)$ as $\mathcal{H}(\kappa, \tau; s_t, r, \widehat{\mathcal{V}}_t)$ where $\kappa = e^{r\tau} K/s_t$. Figure 5 shows $\widehat{\mathcal{H}}_t(\kappa, \tau)$ across moneyness κ and maturity τ at t corresponding to 10 Jul. 2001.

In a first step, we investigate the covariance structure of $\{\widehat{Z}_t\}$ by means of VAR analysis. Table 4 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 5. Moreover, the VAR(2) model is stationary as the roots of the characteristic polynomial lie inside of the unit circle.

A natural issue is to analyze 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

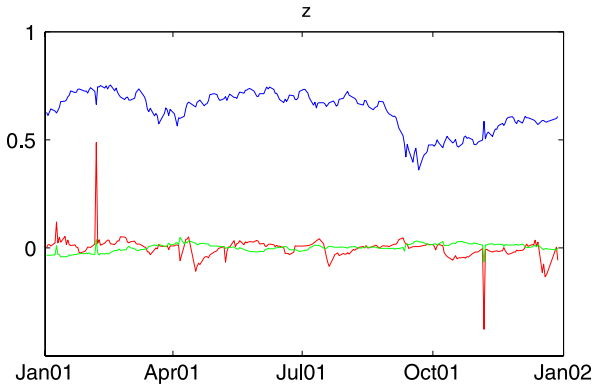


Fig. 3 Estimated $\{\widehat{Z}_{lt}\}$, $l = 1, 2, 3$ (*top to bottom*)

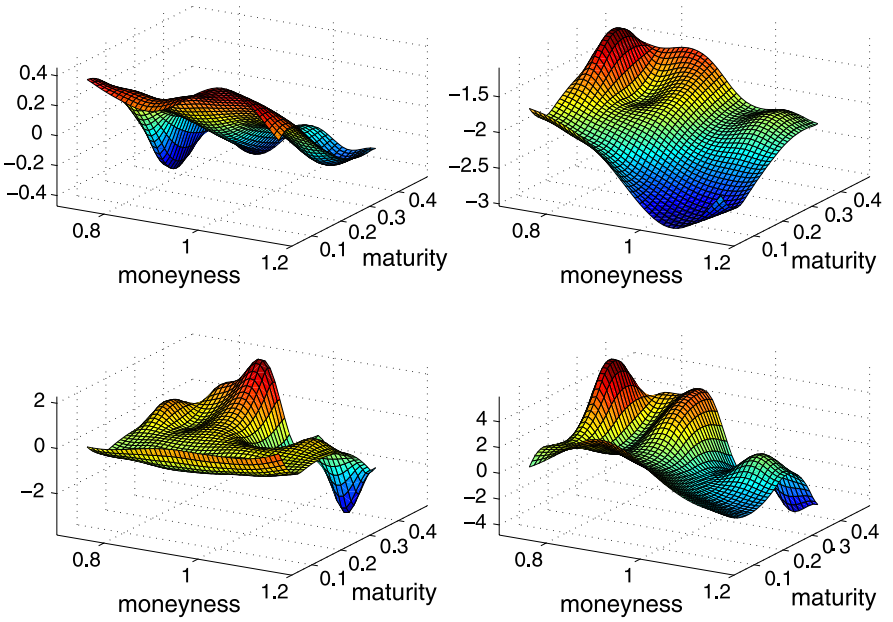


Fig. 4 Estimated basis functions $\widehat{m}_l$, $l = 0, \ldots, 3$, clockwise

γ and excess kurtosis η of $\widehat{q}_{t,T}(\cdot|s_t)$ across t for a maturity τ where $\widehat{q}_{t,T}(\cdot|s_t) = \widehat{\mathcal{H}}_t(\cdot, \tau)$. Figure 6 displays the 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 $\{\widehat{Z}_{3t}\}$, motivating 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 modeled 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.

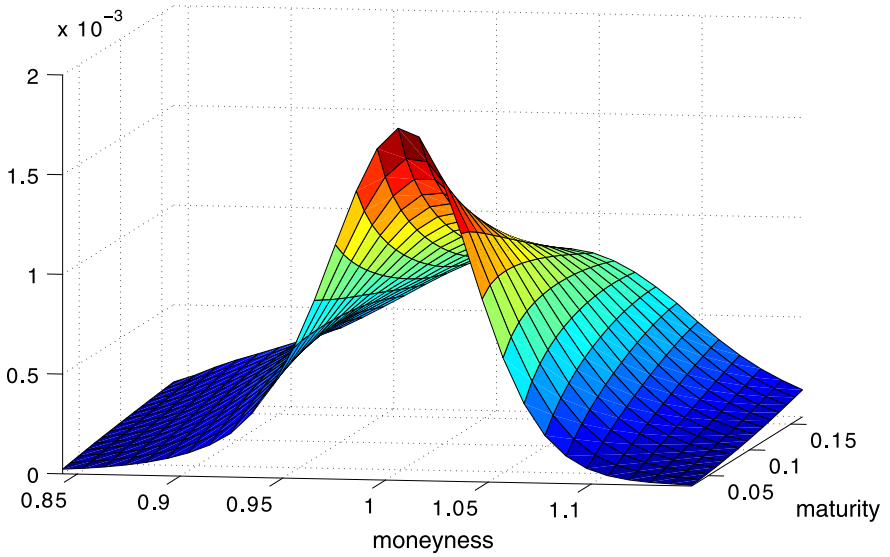


Fig. 5 Estimated RN surface, $\widehat{\mathcal{H}}_t$ at t corresponding to 10 Jul. 2001

Table 4 Estimated parameters for the VAR(2) model on $\{\widehat{Z}_t\}$

	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 5 Lag selection criteria for VAR models on $\{\widehat{Z}_t\}$. The asterisks denote the smallest value for each criterion

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

Modeling the dynamics of risk neutral densities using DSFM allows quantifying the mechanisms governing risk perceptions from agents acting in a market. Insights are obtained in two directions, concerning the autocovariance structure of $\{\widehat{Z}_t\}$, i.e., the time behavior 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 6 and 7 the excess kurtosis and skewness from $\widehat{q}_{t,T}$ at maturity $\tau = 18$ are determined by the corresponding lagged values of $\widehat{Z}_t$.

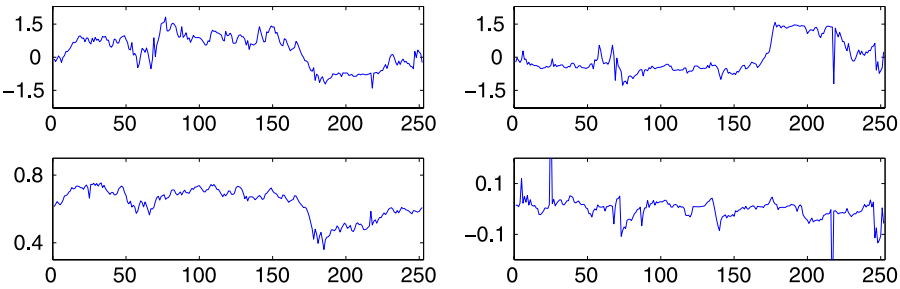


Fig. 6 Left: RN excess kurtosis $\{\eta_t\}$, $\tau = 18$ (*top*), $\{\widehat{Z}_{1t}\}$ (*bottom*). Right: RN skewness $\{\gamma_t\}$, $\tau = 18$ (*top*), $\{\widehat{Z}_{2t}\}$ (*bottom*)

Table 6 Estimated parameters for the VAR(2) model on $\{(\widehat{Z}_{1t}, \eta_t)\}$

	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 7 Estimated parameters for the VAR(2) model on $\{(\widehat{Z}_{3t}, \gamma_t)\}$

	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

The presented methodology allows the investigation of the dynamics from risk neutral skewness and excess kurtosis based on statistical inference on $\{\widehat{Z}_t\}$. A natural further step is to perform econometric analysis on 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.

Acknowledgements Financial support from the Deutsche Forschungsgemeinschaft via SFB 649 "Economic Risk" is gratefully acknowledged. The authors also thank the editor, an associate editor and two referees for their helpful comments.

Appendix: Assumptions

The results from Theorems 2.1 and 2.2, see Park et al. (2009), rely on the following assumptions:

(A1) The variables $X_{11}, \ldots, X_{JT}$, $\varepsilon_{11}, \ldots, \varepsilon_{JT}$ and $Z_1, \ldots, Z_T$ are independent. The process Z_t is allowed to be nonrandom.

(A2) For $t = 1, \ldots, T$, the variables $X_{1t}, \ldots, X_{Jt}$ 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 $t = 1, \ldots, T$.

(A3) We assume that $E[\varepsilon_{jt}] = 0$ for $t = 1, \ldots, T$ and $j = 1, \ldots, J$, and

$$\sup_{t=1,\ldots,T,\, j=1,\ldots,J} E \exp\bigl[c\varepsilon_{jt}^2\bigr] < \infty$$

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, $\sup_{x\in[0,1]} \|\psi(x)\| = \mathcal{O}(K^{1/2})$.

(A5) The components $m_0, \ldots, m_L$ can be approximated by $\psi_1, \ldots, \psi_K$, i.e.,

$$\delta_K = \sup_{x\in[0,1]^d} \inf_{\Gamma\in\mathcal{G}} \bigl|m(x) - \Gamma\psi(x)\bigr| \to 0 \tag{A.1}$$

for $l = 0, \ldots, L$ and $K \to \infty$. We denote by Γ^* the matrix that fulfills

$$\sup_{x\in[0,1]^d} \bigl|m(x) - \Gamma\psi(x)\bigr| \le 2\delta_K.$$

(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

$$\sup_{x\in[0,1]} \max_{1\le t\le T} \bigl\| Z_t^\top \Gamma\psi(x)\bigr\| \le M_T,$$

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

$$\xi^2 = (K+T)M_T^2 \log(JTM_T)(JT)^{-1} \to 0, \tag{A.2}$$

where 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 $X_{11}, \ldots, X_{TJ}$ and $\varepsilon_{11}, \ldots, \varepsilon_{TJ}$.

(A10) The functions $m_0, \ldots, m_L$ are linearly independent. In particular, no function is equal to 0.

(A11) It holds that $(K^{1/2}M_T + T^{1/4})(\xi + \delta_K) = \mathcal{O}(1)$.

References

Ait-Sahalia, Y., Lo, A.: Nonparametric estimation of state-price densities implicit in financial asset prices. J. Finance **53**, 499–547 (1998)

Benko, M., Kneip, A., Härdle, W.: Common functional principal components. Ann. Stat. **37**(1), 1–34 (2009)

Breeden, D., Litzenberger, R.: Prices of state-contingent claims implicit in options prices. J. Bus. **51**, 621–651 (1978)

Brüggemann, R., Härdle, W., Mungo, J., Trenkler, C.: VAR modeling for dynamic loadings driving volatility strings. J. Financ. Econ. **6**, 361–381 (2008)
Cont, R., da Fonseca, J.: The dynamics of implied volatility surfaces. Quant. Finance **2**, 45–60 (2002)
Dauxois, J., Pousse, A., Romain, Y.: Asymptotic theory for the principal component analysis of a vector random function: some applications to statistical inference. J. Multivar. Anal. **12**, 136–154 (1982)
Fengler, M.: Semiparametric Modeling of Implied Volatility. Springer, Heidelberg (2005)
Fengler, M., Härdle, W., Mammen, E.: A semiparametric factor model for implied volatility surface dynamics. J. Financ. Econ. **5**, 189–218 (2007)
Gasser, T., Kneip, A.: Searching for structure in curve samples. J. Am. Stat. Assoc. **90**(432), 1179–1188 (1995)
Hafner, R.: Stochastic Implied Volatility. Springer, Heidelberg (2004)
Hall, P., Müller, H., Wang, J.: Properties of principal component methods for functional and longitudinal data analysis. Ann. Stat. **34**(3), 1493–1517 (2006)
Hernández-Hernández, D., Schied, A.: A control approach to robust maximization with logarithmic utility and time-consistent penalties. Stoch. Process. Appl. **117**(8), 980–1000 (2007)
Kneip, A.: Nonparametric estimation of common regressors for similar curve data. Ann. Stat. **22**(3), 1386–1427 (1994)
Kneip, A., Gasser, T.: Statistical tools to analyse data representing a sample of curves. Ann. Stat. **20**(3), 1266–1305 (1992)
Park, B., Mammen, E., Härdle, W., Borak, S.: Time series modelling with semiparametric factor dynamics. J. Am. Stat. Assoc. **104**(485), 284–298 (2009)
Ramsay, J.O., Dalzell, C.T.: Some tools for functional data analysis. J. R. Stat. Soc. B **53**(3), 539–572 (1991)
Rice, J., Silverman, B.W.: Estimating the mean and covariance structure nonparametrically when the data are curves. J. R. Stat. Soc. B **53**, 233–243 (1991)

JOBNAME: jbes 00#0 2009 PAGE: 1 OUTPUT: Wednesday October 22 04:22:51 2008
asa/jbes/164302/06-114

Inhomogeneous Dependence Modeling With Time-Varying Copulae

Enzo GIACOMINI
Center for Applied Statistics and Economics, Humboldt-Universität zu Berlin, 10178 Berlin, Germany; and Weierstrass Institute for Applied Analysis and Stochastics, 10117 Berlin, Germany (*giacomini@wiwi.hu-berlin.de*)

Wolfgang HÄRDLE
AU1 Weierstrass Institute for Applied Analysis and Stochastics, 10117 Berlin, Germany

Vladimir SPOKOINY
Center for Applied Statistics and Economics, Humboldt-Universität zu Berlin, 10178 Berlin, Germany; and Weierstrass Institute for Applied Analysis and Stochastics, 10117 Berlin, Germany

Measuring dependence in multivariate time series is tantamount to modeling its dynamic structure in space and time. In risk management, the nonnormal behavior of most financial time series calls for non-Gaussian dependences. The correct modeling of non-Gaussian dependences is, therefore, a key issue in the analysis of multivariate time series. In this article we use copula functions with adaptively estimated time-varying parameters for modeling the distribution of returns. Furthermore, we apply copulae to the estimation of value-at-risk of portfolios and show their better performance over the *RiskMetrics* approach.

KEY WORDS: Adaptive estimation; Nonparametric estimation; Value-at-risk.

1. INTRODUCTION

Time series of financial data are high dimensional and typically have a non-Gaussian behavior. The standard modeling approach based on properties of the multivariate normal distribution therefore often fails to reproduce the stylized facts (i.e., fat tails, asymmetry) observed in returns from financial assets.

A correct understanding of the time-varying multivariate (conditional) distribution of returns is vital to many standard applications in finance such as portfolio selection, asset pricing, and value-at-risk (var) calculation. Empirical evidence from asymmetric return distributions have been reported in the recent literature. Longin and Solnik (2001) investigate the distribution of joint extremes from international equity returns and reject multivariate normality in their lower orthant; Ang and Chen (2002) test for conditional correlation asymmetries in U.S. equity data, rejecting multivariate normality at daily, weekly, and monthly frequencies; and Hu (2006) models the distribution of index returns with mixtures of copulae, finding asymmetries in the dependence structure across markets. For a concise survey on stylized empirical facts from financial returns see Cont (2001) and Granger (2003).

Modeling distributions with copulae has drawn attention from many researchers because it avoids the ''procrustean bed'' of normality assumptions, producing better fits of the empirical characteristics of financial returns. A natural extension is to apply copulae in a dynamic framework with conditional distributions modeled by copulae with time-varying parameters. The question, though, is how to steer the time-varying copulae parameters. This question is the focus of this article.

A possible approach is to estimate the parameter from structurally invariant periods. There is a broad field of econometric literature on structural breaks. Tests for unit root in macroeconomic series against stationarity with a structural break at a known change point have been investigated by Perron (1989), and for an unknown change point by Zivot and Andrews (1992), Stock (1994) and Hansen (2001); Andrews (1993) tests for parameter instability in nonlinear models; Andrews and Ploberger (1994) construct asymptotic optimal tests for multiple structural breaks. In a different set up, Quintos, Fan, and Philips (2001) test for a constant tail index coefficient in Asian equity data against a break at an unknown point.

Time-varying copulae and structural breaks are combined in Patton (2006). The dependence structure across exchange rates is modeled with time-varying copulae with a parameter specified to evolve as an ARMA-type process. Tests for a AU2 structural break in the ARMA coefficients at a known change point have been performed, and strong evidence of a break was found. In a similar fashion, Rodriguez (2007) models the dependence across sets of Asian and Latin American stock indexes using time-varying copula where the parameter follows regime-switching dynamics. Common to these articles is that they use a fixed (parametric) structure for the pattern of changes in the copula parameter.

In this article we follow a semiparametric approach, because we are not specifying the parameter changing scheme. Rather, we select locally the time-varying copula parameter. The choice is performed via an adaptive estimation under the assumption of local homogeneity: For every time point there exists an interval of time homogeneity in which the copula parameter can be well approximated by a constant. This interval is recovered from the data using local change point analysis. This does not imply that the model follows a change

Journal of Business & Economic Statistics
January 2009, Vol. 00, No. 0
DOI 10.1198/jbes.2009.0000

point structure. The adaptive estimation also applies when the parameter varies smoothly from one value to another (see Spokoiny 2008).

F1 Figure 1 shows the time-varying copula parameter determined by our procedure for a portfolio composed of daily prices of six German equities and the ''global'' copula parameter, shown by a constant horizontal line. The absence of parametric specification for time variations in the dependence structure (its dynamics is obtained adaptively from the data) allows for flexibility in estimating dependence shifts across time.

The obtained time-varying dependence structure can be used in financial engineering applications, the most prominent being the calculation of the var of a portfolio. Using copulae with adaptively estimated dependence parameters we estimate the AU3 var from DAX portfolios over time. As a benchmark procedure we choose *RiskMetrics*, a widely used methodology based on conditional normal distributions with a GARCH specification for the covariance matrix. Backtesting underlines the improved performance of the proposed adaptive time-varying copulae fitting.

This article is organized as follows: Section 2 presents the basic copulae definitions, Section 3 discusses the var and its estimation procedure. The adaptive copula estimation is exposed in Section 4 and is applied to simulated data in Section 5. In Section 6, the var from DAX portfolios is estimated based on adaptive time-varying copulae. The estimation performance is compared with the *RiskMetrics* approach by means of backtesting. Section 7 concludes.

2. COPULAE

Copulae merge marginally into joint distributions, providing a natural way for measuring the dependence structure between random variables. Copulae are present in the literature since Sklar (1959), although related concepts originate in Hoeffding (1940) and Fréchet (1951), and have been widely studied in the statistical literature (see Joe 1997, Nelsen 1998, and Mari and Kotz 2001). Applications of copulae in finance, insurance, and econometrics have been investigated in Embrechts, McNeil, and Straumann (2002); Embrechts, Hoeing, and Juri (2003a); Franke, Härdle, and Hafner (2004); and Patton (2004) among others. Cherubini, Luciano, and Vecchiato (2004) and McNeil, Frey, and Embrechts (2005) provide an overview of copulae for practical problems in finance and insurance.

Assuming absolutely continuous distributions and continuous marginals throughout this article, we have from Sklar's theorem that for a d-dimensional distribution function F with marginal cdf's $F_1, \ldots, F_d$ there exists a unique copula $C : [0, 1]^d \to [0, 1]$ satisfying AU4

$$F(x_1, \ldots, x_d) = C\{F_1(x_1), \ldots, F_d(x_d)\} \tag{2.1}$$

for every $\boldsymbol{x} = (x_1, \ldots, x_d)^{\mathrm{T}} \in \mathbb{R}^d$. Conversely, for a random vector $\boldsymbol{X} = (X_1, \ldots, X_d)^{\mathrm{T}}$ with cdf $F_{\boldsymbol{X}}$, the copula of $\boldsymbol{X}$ may be written as $C_X(u_1, \ldots, u_d) = F_X\{F_1^{-1}(u_1), \ldots, F_d^{-1}(u_d)\}$, where $u_j = F_j(x_j)$, F_j is the cdf of X_j, and $F_j^{-1}(\alpha) = \inf\{x_j : F_j(x_j) \geq \alpha\}$ its generalized inverse, $j = 1, \ldots, d$. A prominent copula is the Gaussian

$$C_{\Psi}^{Ga}(u_1, \ldots, u_d) = F_Y\{\Phi^{-1}(u_1), \ldots, \Phi^{-1}(u_d)\} \tag{2.2}$$

where $\Phi(s)$, $s \in \mathbb{R}$ stands for the one-dimensional standard normal cdf, $F_{\boldsymbol{Y}}$ is the cdf of $\boldsymbol{Y} = (Y_1, \ldots, Y_d)^{\top} \sim N_d(\mathbf{0}, \boldsymbol{\Psi})$, $\mathbf{0}$ is the $(d \times 1)$ vector of zeros, and $\boldsymbol{\Psi}$ is a correlation matrix. The Gaussian copula represents the dependence structure of the multivariate normal distribution. In contrast, the Clayton copula given by

$$C_\theta(u_1, \ldots, u_d) = \left\{ \left(\sum_{j=1}^{d} u_j^{-\theta} \right) - d + 1 \right\}^{-\theta^{-1}} \tag{2.3}$$

for $\theta > 0$, expresses asymmetric dependence structures.

The dependence at upper and lower orthants of a copula C may be expressed by the upper and lower tail dependence coefficients $\lambda_U = \lim_{u \to 0} \widehat{C}(u, \ldots, u)/u$ and $\lambda_L = \lim_{u \to 0} C(u, \ldots, u)/u$, where $u \in (0, 1]$ and $\widehat{C}$ is the survival copula of C (see Joe 1997 and Embrechts, Lindskog, and McNeil 2003b). Although Gaussian copulae are asymptotically independent at the tails ($\lambda_L = \lambda_U = 0$), the d-dimensional Clayton copulae exhibit lower tail dependence ($\lambda_L = d^{-1/\theta}$) but are asymptotically independent at the upper tail ($\lambda_U = 0$). Joe (1997) provides a summary of diverse copula families and detailed description of their properties.

For estimating the copula parameter, consider a sample $\{\boldsymbol{x}_t\}_{t=1}^{T}$ of realizations from $\boldsymbol{X}$ where the copula of $\boldsymbol{X}$ belongs to a parametric family $\mathrm{C} = \{C_\theta, \theta \in \Theta\}$. Using Equation (2.1), the log-likelihood reads as $L(\theta; \boldsymbol{x}_1, \ldots, \boldsymbol{x}_T) = \sum_{t=1}^{T} [\log c\, F_1(x_{t,1}), \ldots, F_d(x_{t,d}); \theta\} + \sum_{j=1}^{d} \log f_j(x_{t,j})]$, where $c(u_1, \ldots, u_d) = \partial^d C(u_1, \ldots, u_d)/\partial u_1 \ldots \partial u_d$ is the density of the copula C and f_j is the probability density function of F_j. The canonical maximum likelihood estimator $\hat{\theta}$ maximizes the pseudo log-likelihood with empirical marginal cdf's $\tilde{L}(\theta) = \sum_{t=1}^{T} \log c\, \{\widehat{F}_1(x_{t,1}), \ldots, \widehat{F}_d(x_{t,d}); \theta$, where

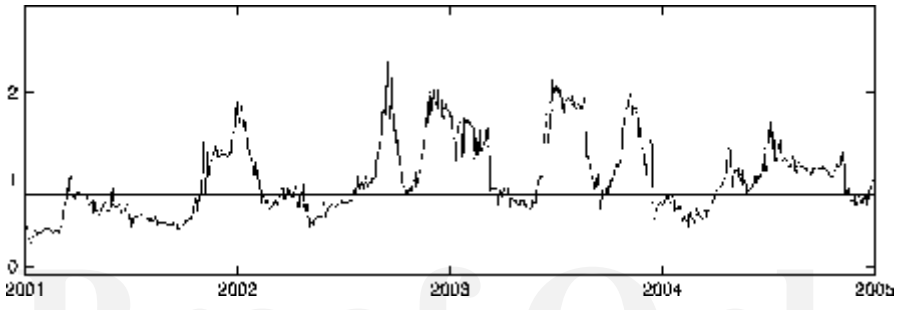


Figure 1. Time-varying dependence. Time-varying dependence parameter and global parameter (horizontal line) estimated with Clayton copula, stock returns from Allianz, Münchener Rückversicherung, BASF, Bayer, DaimlerChrysler, and Volkswagen.

$$\widehat{F}_j(s) = \frac{1}{T+1}\sum_{k=1}^{T} 1_{\{x_{kj}\le s\}} \tag{2.4}$$

for $j = 1, \ldots, d$. Note that $\widehat{F}_j$ differs from the usual empirical cdf by the denominator $T+1$. This ensures that $\{\widehat{F}_1(x_{t,1}), \ldots, \widehat{F}_d(x_{t,d})\}^\top \in (0,1)^d$ and avoids infinite values the copula density may take on the boundary of the unit cube (see McNeil, Frey, and Embrechts 2005). Joe (1997); Cherubini, Luciano, and Vecchiato (2004); and Chen and Fan (2006) provide a detailed exposition of inference methods for copulae.

3. VALUE-AT-RISK AND COPULAE

The dependence (over time) between asset returns is especially important in risk management, because the profit and loss (P&L) function determines the var. More precisely, the var of a portfolio is determined by the multivariate distribution of risk factor increments. If $\boldsymbol{w} = (w_1, \ldots, w_d)^\top \in \mathbb{R}^d$ denotes a portfolio of positions on d assets and $\boldsymbol{S}_t = (S_{t,1}, \ldots, S_{t,d})^\top$ a nonnegative random vector representing the prices of the assets at time t, the value V_t of the portfolio w is given by $V_t = \sum_{j=1}^d w_j S_{t,j}$. The random variable

$$L_t = (V_t - V_{t-1}), \tag{3.1}$$

called the profit and loss (P&L) function, expresses the change in the portfolio value between two subsequent time points. Defining the log-returns $\boldsymbol{X}_t = (X_{t,1}, \ldots, X_{t,d})^\top$, where $X_{t,j} = \log S_{t,j} - \log S_{t-1,j}$ and $\log S_{0,j} = 0, j = 1, \ldots, d$, Equation (3.1) can be written as

$$L_t = \sum_{j=1}^{d} w_j S_{t-1,j}\{\exp(X_{t,j}) - 1\}. \tag{3.2}$$

The cdf of L_t is given by $F_{t,L_t}(x) = P_t(L_t \le x)$. The var at level α from a portfolio w is defined as the α quantile from F_{t,L_t}:

$$\mathrm{var}_t(\alpha) = F^{-1}_{t,L_t}(\alpha). \tag{3.3}$$

It follows from Equation (3.2) that F_{t,L_t} depends on the specification of the d-dimensional distribution of the risk factors $\boldsymbol{X}_t$. Thus, modeling their distribution over time is essential for obtaining the quantiles (Eq. 3.3).

The *RiskMetrics* technique, a widely used methodology for var estimation, assumes that risk factors $\boldsymbol{X}_t$ follow a conditional multivariate normal distribution $\mathcal{L}(\boldsymbol{X}_t|\mathcal{F}_{t-1}) = N(\boldsymbol{0}, \boldsymbol{\Sigma}_t)$, where $\mathcal{F}_{t-1} = \sigma(\boldsymbol{X}_1, \ldots, \boldsymbol{X}_{t-1})$ is the σ field generated by the first $t-1$ observations, and estimates the covariance matrix $\boldsymbol{\Sigma}_t$ for one period return as

$$\widehat{\boldsymbol{\Sigma}}_t = \lambda\widehat{\boldsymbol{\Sigma}}_{t-1} + (1-\lambda)\boldsymbol{X}_{t-1}\boldsymbol{X}_{t-1}^\top, \tag{3.4}$$

where the parameter λ is the so-called decay factor. $\lambda = 0.94$ provides the best backtesting results for daily returns according to Morgan (1996). Using the copulae-based approach, one first corrects the contemporaneous mean and volatility in the log-returns process:

$$X_{t,j} = \mu_{t,j} + \sigma_{t,j}\varepsilon_{t,j}, \tag{3.5}$$

where $\mu_{t,j} = E[X_{t,j}|\mathcal{F}_{t-1}]$ is the conditional mean and $\sigma^2_{t,j} = E[(X_{t,j} - \mu_{t,j})^2|\mathcal{F}_{t-1}]$ is the conditional variance of $X_{t,j}$. The standardized innovations $\boldsymbol{\varepsilon}_t = (\varepsilon_{t,1}, \ldots, \varepsilon_{t,d})^\top$ have joint cdf F_{ε_t} given by

$$F_{\varepsilon_t}(x_1, \ldots, x_d) = C_\theta\{F_{t,1}(x_1), \ldots, F_{t,d}(x_d)\}, \tag{3.6}$$

where $F_{t,j}$ is the cdf of $\varepsilon_{t,j}$ and C_θ is a copula belonging to a parametric family $C = C_\theta, \theta \in \Theta\}$. For details on the previous model specification, see Chen and Fan (2006) and Chen, Fan, and Tsyrennikov (2006). For the Gaussian copula with Gaussian marginals, we recover the conditional Gaussian *RiskMetrics* framework.

To obtain the var in this setup, the dependence parameter and cdf's from residuals are estimated from a sample of log-returns and are used to generate P&L Monte Carlo samples. Their quantiles at different levels are the estimators for the var (see Embrechts, McNeil, and Straumann 2002).

The whole procedure can be summarized as follows (see Härdle, Kleinow, and Stahl 2002; and Giacomini and Härdle 2005): For a portfolio $\boldsymbol{w} \in \mathbb{R}^d$ and a sample $\{x_{t,j}\}_{t=1}^T, j = 1, \ldots, d$ of log-returns, the var at level α is estimated according to the following steps:

1. Determination of innovations $\{\hat{\varepsilon}_t\}_{t=1}^T$ by, for example, "deGARCHing"
2. Specification and estimation of marginal cdf's $F_j(\hat{\varepsilon}_j)$
3. Specification of a parametric copula family C and estimation of the dependence parameter θ
4. Generation of Monte Carlo sample of innovations ε and losses L
5. Estimation of $\widehat{\mathrm{var}}(\alpha)$, the empirical α quantile of F_L

4. MODELING WITH TIME-VARYING COPULAE

Similar to the *RiskMetrics* procedure, one can perform a moving (fixed-length) window estimation of the copula parameter. This procedure, though, does not fine-tune local changes in dependences. In fact, the cdf F_{ε_t} from Equation (3.6) is modeled as $F_{t,\varepsilon_t} = C_{\theta_t}\{F_{t,1}(\cdot), \ldots, F_{t,d}(\cdot)\}$ with probability measure P_{θ_t}. The moving window of fixed width will estimate a θ_t for each t, but it has clear limitations. The choice of a small window results in a high pass filtering and, hence, in a very unstable estimate with huge variability. The choice of a large window leads to a poor sensitivity of the estimation procedure

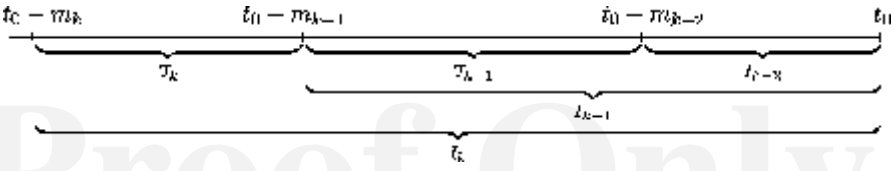


Figure 2. Local change point procedure. Choice of intervals I_k and I_k.

JOBNAME: jbes 00#0 2009 PAGE: 4 OUTPUT: Wednesday October 22 04:22:58 2008
asa/jbes/164302/06-114

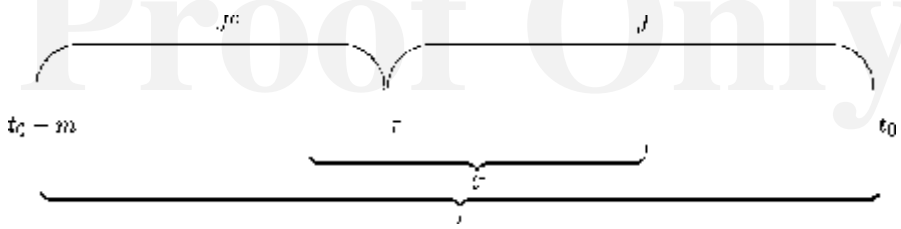


Figure 3. Homogeneity test. Testing interval I, tested interval I, and subintervals J and J^c for a point $\tau \in$ I.

and to a high delay in the reaction to changes in dependence measured by the parameter θ_t.

To choose an interval of homogeneity, we use a local parametric fitting approach as introduced by Polzehl and Spokoiny (2006), Belomestny and Spokoiny (2007) and Spokoiny (2008). The basic idea is to select for each time point t_0 an interval $I_{t_0} = [t_0 - m_{t_0}, t_0]$ of length m_{t_0} in such a way that the time-varying copula parameter θ_t can be well approximated by a constant value θ. The question is, of course, how to select m_{t_0} in an online situation from historical data. The aim should be to select I_{t_0} as close as possible to the so-called "oracle" choice interval. The oracle choice is defined as the largest interval $I = [t_0 - m^*_{t_0}, t_0]$, for which the small modeling bias condition

$$\Delta_I(\theta) = \sum_{t \in I} \mathcal{K}(P_{\theta_t}, P_\theta) \le \Delta \tag{4.1}$$

for some $\Delta \ge 0$ holds. Here, θ is constant and $\mathcal{K}(P_\vartheta, P_{\vartheta'}) = E_\vartheta \log\{p(y,\vartheta)/p(y,\vartheta')\}$ denotes the Kullback-Leibler divergence. In such an oracle choice interval, the parameter $\theta_{t_0} = \theta_t|_{t=t_0}$ can be "optimally" estimated from $I = [t_0 - m^*_{t_0}, t_0]$. The error and risk bounds are calculated in Spokoiny (2008). It is important to mention that the concept of local parametric approximation allows one to treat in a unified way the case of "switching regime" models with spontaneous changes of parameters and the "smooth transition" case when the parameter varies smoothly in time.

The oracle choice of the interval of homogeneity depends on the unknown time-varying copula parameter θ_t. The next section presents an adaptive (data-driven) procedure that mimics the oracle in the sense that it delivers the same accuracy of estimation as the oracle one. The trick is to find the largest interval in which the hypothesis of a local constant copula parameter is supported. The local change point (LCP) detection procedure originates from Mercurio and Spokoiny (2004) and sequentially tests the hypothesis: θ_t is constant (i.e., $\theta_t = \theta$) within some interval I (local parametric assumption).

The LCP procedure for a given point t_0 starts with a family of nested intervals $I_0 \subset I_1 \subset I_2 \subset \ldots \subset I_K = I_{K+1}$ of the form $I_k = [t_0 - m_k, t_0]$. The sequence m_k determines the length of these interval "candidates" (see Section 4.2). Every interval I_k leads to an estimate $\tilde\theta_k$ of the copula parameter θ_{t_0}. The procedure selects one interval $\hat I$ out of the given family and, therefore, the corresponding estimate $\hat\theta = \tilde\theta_{\hat I}$.

The idea of the procedure is to screen each interval $\mathcal{I}_k = [t_0 - m_k, t_0 - m_{k-1}]$ sequentially and check each point $\tau \in \mathcal{I}_k$ as a possible change point location (see Section 4.1 for more details). The family of intervals I_k and $\mathcal{I}_k$ are illustrated in Figure 2. The interval I_k is accepted if no change point is F2
detected within $\mathcal{I}_1, \ldots, \mathcal{I}_k$. If the hypothesis of homogeneity is rejected for an interval candidate I_k, the procedure stops and selects the latest accepted interval. The formal description reads as follows:

Start the procedure with $k = 1$ and test the hypothesis $H_{0,k}$ AU5
of no structural changes within $\mathcal{I}_k$ using the larger testing interval I_{k+1}. If no change points were found in $\mathcal{I}_k$, then I_k is accepted. Take the next interval $\mathcal{I}_{k+1}$ and repeat the previous step until homogeneity is rejected or the largest possible interval $I_K = [t_0 - m_K, t_0]$ is accepted. If $H_{0,k}$ is rejected for $\mathcal{I}_k$, the estimated interval of homogeneity is the last accepted interval $\hat I = I_{k-1}$. If the largest possible interval I_K is accepted, we take $\hat I = I_K$. We estimate the copula dependence parameter θ at time instant t_0 from observations in $\hat I$, assuming the homogeneous model within $\hat I$ (i.e., we define $\hat\theta_{t_0} = \tilde\theta_{\hat I}$). We also denote by $\hat I_k$ the largest accepted interval after k steps of

Table 1. Critical values $\mathfrak{z}_k$ (ρ, θ^*)

	$\theta^* = 0.5$			$\theta^* = 1.0$			$\theta^* = 1.5$		
k	$\rho = 0.2$	$\rho = 0.5$	$\rho = 1.0$	ρ =0.2	$\rho = 0.5$	$\rho = 1.0$	$\rho = 0.2$	$\rho = 0.5$	$\rho = 1.0$
1	3.64	3.29	2.88	3.69	3.29	2.84	3.95	3.49	2.96
2	3.61	3.14	2.56	3.43	2.91	2.35	3.69	3.02	2.78
3	3.31	2.86	2.29	3.32	2.76	2.21	3.34	2.80	2.09
4	3.19	2.69	2.07	3.04	2.57	1.80	3.14	2.55	1.86
5	3.05	2.53	1.89	2.92	2.22	1.53	2.95	2.65	1.49
6	2.87	2.26	1.48	2.92	2.17	1.19	2.83	2.04	0.94
7	2.51	1.88	1.02	2.64	1.82	0.56	2.62	1.79	0.31
8	2.49	1.72	0.35	2.33	1.39	0.00	2.35	1.33	0.00
9	2.18	1.23	0.00	2.03	0.81	0.00	2.10	0.60	0.00
10	0.92	0.00	0.00	0.82	0.00	0.00	0.79	0.00	0.00

NOTE: Critical values are obtained according to Equation (4.2), based on 5,000 simulations. Clayton copula, $m_0 = 20$ and $c = 1.25$.

JOBNAME: jbes 00#0 2009 PAGE: 5 OUTPUT: Wednesday October 22 04:23:01 2008
asa/jbes/164302/06-114

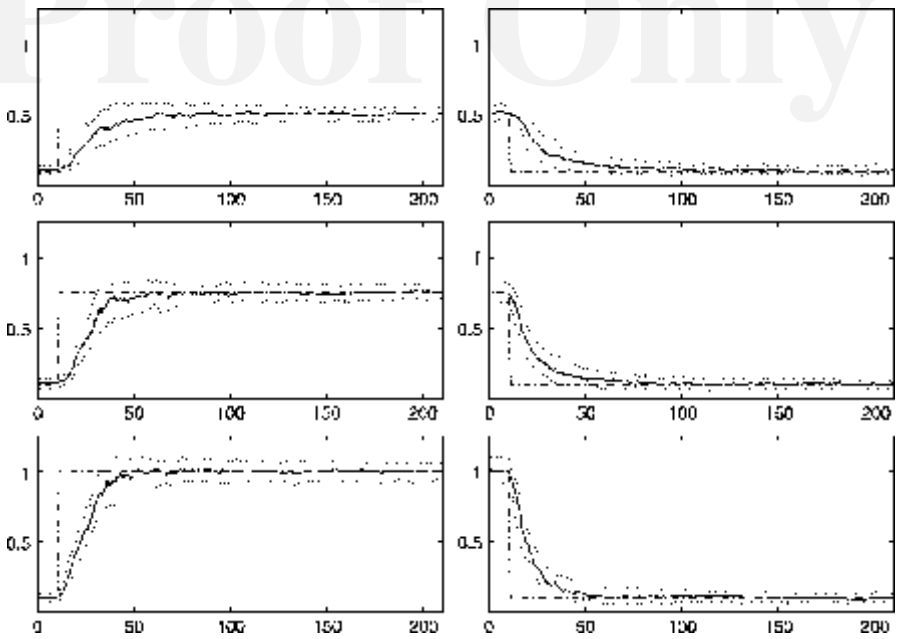


Figure 4. LCP and sudden jump in copula parameter. Pointwise median (full), and 0.25 and 0.75 quantiles (dotted) from $\hat{\theta}_t$. True parameter θ_t (dashed) with $\vartheta_a = 0.10$, $\vartheta_b = 0.50$, 0.75, and 1.00 (left, top to bottom); and $\vartheta_b = 0.10$, $\vartheta_a = 0.50$, 0.75, and 1.00 (right, top to bottom). Based on 100 simulations from Clayton copula, estimated with LCP, $m_0 = 20$, $c = 1.25$, and $\rho = 0.5$.

the algorithm and, by $\hat{\theta}_k$ the corresponding estimate of the copula parameter.

It is worth mentioning that the objective of the described estimation algorithm is not to detect the points of change for the copula parameter, but rather to determine the current dependence structure from historical data by selecting an interval of time homogeneity. This distinguishes our approach from other procedures for estimating a time-varying parameter by change point detection. A visible advantage of our approach is that it equally applies to the case of spontaneous changes in the dependence structure and in the case of smooth transition in the copula parameter. The obtained dependence structure can be used for different purposes in financial engineering, the most prominent being the calculation of the var (see also Section 6).

The theoretical results from Spokoiny and Chen (2007) and Spokoiny (2008) indicate that the proposed procedure provides the rate optimal estimation of the underlying parameter when this varies smoothly with time. It has also been shown that the procedure is very sensitive to structural breaks and provides the minimal possible delay in detection of changes, where the delay depends on the size of change in terms of Kullback-Leibler divergence.

4.1 Test of Homogeneity Against a Change Point Alternative

In the homogeneity test against a change point alternative we want to check every point of an interval I (recall Fig. 2), here called the "tested interval," on a possible change in the dependence structure at this moment. To perform this check, we assume a larger testing interval I of form $I = [t_0 - m, t_0]$, so that I is an internal subset within I. The null hypothesis H_0 means that $\forall t \in I$, $\theta_t = \theta$ (i.e., the observations in I follow the model with dependence parameter θ). The alternative hypothesis H_1 claims that $\exists\ \tau \in \mathrm{I}$ such that $\theta_t = \theta_1$ for $t \in J = [\tau, t_0]$ and $\theta_t = \theta_2 \neq \theta_1$ for $t \in J^c = [t_0 - m, \tau)$ (i.e., the parameter θ changes spontaneously in some point $\tau \in \mathrm{I}$). Figure 3 depicts I, I, and the subintervals J and J^c determined by the point $\tau \in \mathrm{I}$. F3

Let $L_I(\theta)$ be the log-likelihood and $\tilde{\theta}_I$ the maximum likelihood estimate for the interval I. The log-likelihood functions corresponding to H_0 and H_1 are $L_I(\theta)$ and $L_J(\theta_1) + L_{J^c}(\theta_2)$, respectively. The likelihood ratio test for the single change point with known fixed location τ can be written as

Table 2. Detection delay statistics

$(\vartheta_a, \vartheta_b)$	r	Mean	SD	Max	Min
	0.25	9.06	7.28	56	0
(0.50, 0.10)	0.50	13.64	9.80	60	0
	0.75	21.87	14.52	89	3
	0.25	5.16	4.24	21	0
(0.75, 0.10)	0.50	8.85	5.55	25	0
	0.75	16.72	10.37	64	3
	0.25	4.47	2.94	12	0
(1.00, 0.10)	0.50	7.94	4.28	22	0
	0.75	14.79	7.38	62	5
	0.25	8.94	6.65	36	0
(0.10, 0.50)	0.50	14.21	9.06	53	0
	0.75	21.43	12.15	68	0
	0.25	9.00	4.80	25	0
(0.10, 0.75)	0.50	14.30	5.96	40	3
	0.75	21.00	10.97	75	6
	0.25	7.39	3.67	19	0
(0.10, 1.00)	0.50	13.10	4.13	22	2
	0.75	20.13	7.34	55	10

NOTE: The detection delays δ are calculated as in Equation (5.1), with the statistics based on 100 simulations. Clayton copula, $m_0 = 20$, $c = 1.25$, and $\rho = .5$. SD, standard deviation.

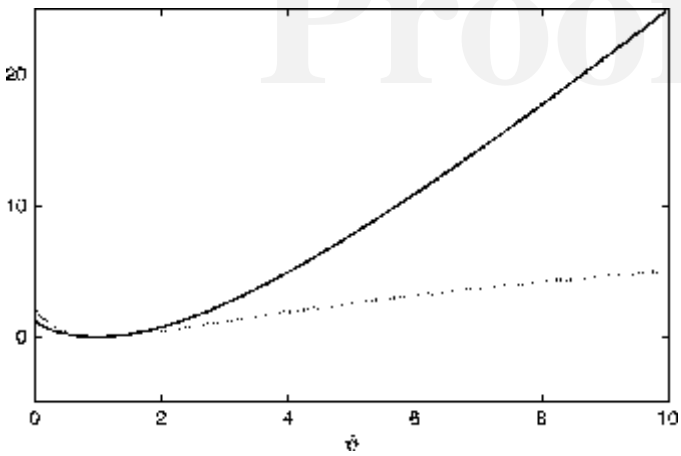


Figure 5. Divergences for upward and downward jumps. Kullback-Leibler divergences $\mathcal{K}(0.10, \vartheta)$ (full) and $\mathcal{K}(\vartheta, 0.10)$ (dashed) for Clayton copula.

$$T_{I,\tau} = \max_{\theta_1,\theta_2} \{L_J(\theta_1) + L_{J^c}(\theta_2)\} - \max_{\theta} L_I(\theta)$$
$$= L_J(\tilde{\theta}_J) + L_{J^c}(\tilde{\theta}_{J^c}) - L_I(\tilde{\theta}_I).$$

The test statistic for an unknown change point location is defined as $T_I = \max_{\tau \in \mathrm{I}} T_{I,\tau}$. The change point test compares this test statistic with a critical value $\mathfrak{z}_I$, which may depend on the interval I. One rejects the hypothesis of homogeneity if $T_I > \mathfrak{z}_I$.

4.2 Parameters of the LCP Procedure

To apply the LCP testing procedure for local homogeneity, we have to specify some parameters. This includes selecting interval candidates I_k or, equivalently, of the tested intervals $\mathcal{I}_k$ and choosing respective critical values $\mathfrak{z}_k$. One possible parameter set that has been used successfully in simulations is presented in the following section.

4.2.1 Selection of interval candidates $\mathcal{I}_k$ and internal points I_k. It is useful to take the set of numbers m_k defining the length of I_k and $\mathcal{I}_k$ in the form of a geometric grid. We fix the value m_0 and define $m_k = [m_0 c^k]$ for $k = 1, 2, \ldots, K$ and $c > 1$ where $[x]$ means the integer part of x. We set $I_k = [t_0 - m_k, t_0]$ and $\mathcal{I}_k = [t_0 - m_k, t_0 - m_{k-1}]$ for $k = 1, 2, \ldots, K$ (see Fig. 2).

4.2.2 Choice of the critical values $\mathfrak{z}_k$. The algorithm is in fact a multiple testing procedure. Mercurio and Spokoiny (2004) suggested selecting the critical value z_k to provide the overall first type error probability of rejecting the hypothesis of homogeneity in the homogeneous situation. Here we follow another proposal from Spokoiny and Chen (2007), which focuses on estimation losses caused by the "false alarm"—in our case obtaining a homogeneity interval that is too small—rather than on its probability.

In the homogeneous situation with $\theta_t \equiv \theta^*$ for all $t \in I_{k+1}$, the desirable behavior of the procedure is that after the first k steps the selected interval $\hat{I}_k$ coincides with I_k and the corresponding estimate $\hat{\theta}_k$ coincides with $\tilde{\theta}_k$, which means there is no false alarm. On the contrary, in the case of a false alarm, the selected interval $\hat{I}_k$ is smaller than I_k and, hence, the corresponding estimate $\hat{\theta}_k$ has larger variability than $\tilde{\theta}_k$. This means that the false alarm during the early steps of the procedure is more critical than during the final steps, because it may lead to selecting an estimate with very high variance. The difference between $\hat{\theta}_k$ and $\tilde{\theta}_k$ can naturally be measured by the value $L_{I_k}(\tilde{\theta}_k, \hat{\theta}_k) = L_{I_k}(\tilde{\theta}_k) - L_{I_k}(\hat{\theta}_k)$ normalized by the risk of the nonadaptive estimate $\tilde{\theta}_k$, $\mathfrak{R}(\theta^*) = \max_{k \geq 1} \boldsymbol{E}_{\theta^*} \bigl|L_{I_k}(\tilde{\theta}_k, \theta^*)\bigr|^{1/2}$. The conditions we impose read as

$$\boldsymbol{E}_{\theta^*} \bigl|L_{I_k}(\tilde{\theta}_k, \hat{\theta}_k)\bigr|^{1/2} \leq \rho \mathfrak{R}(\theta^*), \qquad k = 1, \ldots, K, \qquad \theta^* \in \Theta. \tag{4.2}$$

The critical values $\mathfrak{z}_k$ are selected as minimal values providing these constraints. In total we have K conditions to select K critical values $\mathfrak{z}_1, \ldots, \mathfrak{z}_K$. The values $\mathfrak{z}_k$ can be selected sequentially by Monte Carlo simulation, where one simulates under $H_0 : \theta_t = \theta^*$, $\forall t \in I_K$. The parameter ρ defines how conservative the procedure is. A small ρ value leads to larger critical values and hence to a conservative and nonsensitive procedure, whereas an increase in ρ results in more sensitiveness at cost of stability. For details, see Spokoiny and Chen (2007) or Spokoiny (2008).

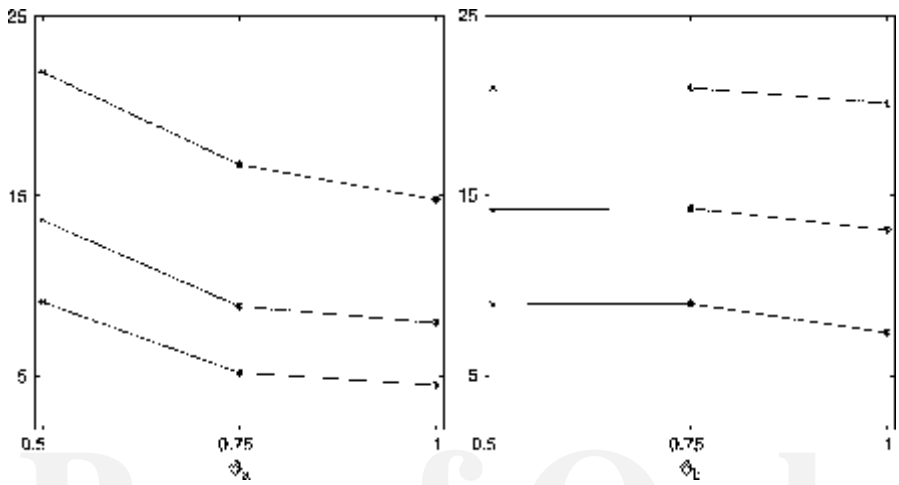


Figure 6. Mean detection delay and parameter jumps. Mean detection delays (dots) at rule $r = 0.75$, 0.50, and 0.25 from top to bottom. Left: $\vartheta_b = 0.10$ (upward jump). Right: $\vartheta_a = 0.10$ (downward jump), based on 100 simulations from Clayton copula, $m_0 = 20$, $c = 1.25$, and $\rho = 0.5$.

JOBNAME: jbes 00#0 2009 PAGE: 7 OUTPUT: Wednesday October 22 04:23:08 2008
asa/jbes/164302/06-114

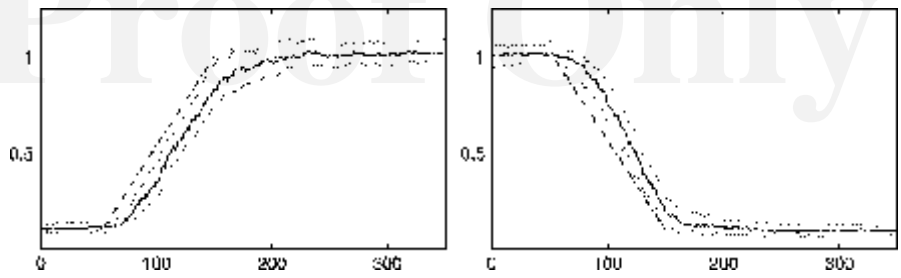


Figure 7. LCP and smooth change in copula parameter. Pointwise median (full), 0.25 and 0.75 quantiles (dotted) from $\hat{\theta}_t$ and true parameter θ_t (dashed) with $\vartheta_a = 0.10$ and $\vartheta_b = 1.00$ (left), and $\vartheta_a = 1.00$ and $\vartheta_b = 0.10$ (right). Based on 100 simulations from Clayton copula, estimated with LCP, $m_0 = 20$, $c = 1.25$, and $\rho = 0.5$.

5. SIMULATED EXAMPLES

In this section we apply the LCP procedure on simulated data with a dependence structure given by the Clayton copula. We generate sets of six-dimensional data with a sudden jump in the dependence parameter given by

$$\theta_t = \begin{cases} \vartheta_a & \text{if } -390 \le t \le 10 \\ \vartheta_b & \text{if } 10 < t \le 210 \end{cases}$$

for different values of $(\vartheta_a, \vartheta_b)$: One of them is fixed at .1 (close to independence) and the other is set to larger values.

The LCP procedure is implemented with the family of interval candidates in form of a geometric grid defined by $m_0 = 20$ and $c = 1.25$. The critical values, selected according to T1 Equation (4.2) for different ρ and θ^*, are displayed in Table 1. The choice of θ^* has negligible influence in the critical values for fixed ρ, therefore we use $\mathfrak{z}_1, \ldots, \mathfrak{z}_K$ obtained with $\theta^* = 1.0$. Based on our experience, see Spokoiny and Chen (2007) and Spokoiny (2008), the default choice for ρ is 0.5.

F4 Figure 4 shows the pointwise median and quantiles of the estimated parameter $\hat{\theta}_t$ for distinct values of $(\vartheta_a,\vartheta_b)$ based on 100 simulations. The detection delay δ at rule $r \in [0, 1]$ to jump of size $\gamma = \theta_t - \theta_{t-1}$ at t is expressed by

$$\delta(t, \gamma, r) = \min\{u \ge t : \hat{\theta}_u = \theta_{t-1} + r\gamma\} - t \qquad (5.1)$$

and represents the number of steps necessary for the estimated parameter to reach the r fraction of a jump in the true parameter.

Detection delays are proportional to the probability of error of type II (i.e., the probability of accepting homogeneity in case of a jump). Thus, tests with higher power correspond to lower delays δ. Moreover, because the Kullback-Leibler divergences for upward and downward jumps are proportional to the power of the respective homogeneity tests, larger divergences result in faster jump detections.

The descriptive statistics for detection delays to jumps at $t =$ T2 11 for different values of $(\vartheta_a,\vartheta_b)$ are in Table 2. The mean detection delay decreases with $\gamma = \vartheta_b - \vartheta_a$ and are higher for F5 downward jumps than for upward jumps. Figure 5 shows that for Clayton copulae the Kullback-Leibler divergence is higher F6 for upward jumps than for downward jumps. Figure 6 displays the mean detection delays against jump size for upward and downward jumps.

The LCP procedure is also applied on simulated data with smooth transition in the dependence parameter given by

$$\theta_t = \begin{cases} \vartheta_a & \text{if } -350 \le t \le 50 \\ \vartheta_a + \dfrac{t-50}{100}(\vartheta_b - \vartheta_a) & \text{if } 50 < t \le 150 \\ \vartheta_b & \text{if } 150 < t \le 350. \end{cases}$$

Figure 7 depicts the pointwise median and quantiles of the F7 estimated parameter $\hat{\theta}_t$ and the true parameter θ_t for $(\vartheta_a, \vartheta_b)$ set to (0.10, 1.00) and (1.00, 0.10).

6. EMPIRICAL RESULTS

In this section the var from German stock portfolios is estimated based on time-varying copulae and *RiskMetrics* approaches. The time-varying copula parameters are selected by local change point (LCP) and moving window procedures. Backtesting is used to evaluate the performances of the three methods in var estimation.

Two groups of six stocks listed on DAX are used to compose the portfolios. Stocks from group 1 belong to three different industries: automotive (Volkswagen and DaimlerChrysler), insurance (Allianz and Münchener Rückversicherung), and chemical (Bayer and BASF). Group 2 is composed of stocks from six industries: electrical (Siemens), energy (E.ON), metallurgical (ThyssenKrupp), airlines (Lufthansa), pharmaceutical (Schering), and chemical (Henkel). The portfolio values are calculated using 1,270 observations, from January 1, 2000 to December 31, 2004, of the daily stock prices (data available at http://sfb649.wiwi.hu-berlin.de/fedc).

The selected copula belongs to the Clayton family (Eq. 2.3). Clayton copulae have a natural interpretation and are well advocated in risk management applications. In line with the stylized facts for financial returns, Clayton copulae are asymmetric and present lower tail dependence, modeling joint

Table 3. p Values from tests on residuals $\hat{\varepsilon}_{t,j}$

	Ljung-Box		ARCH	
j	Group 1	Group 2	Group 1	Group 2
1	0.33	0.52	0.15	0.04
2	0.13	0.35	0.15	0.98
3	0.21	0.08	0.34	0.72
4	0.99	0.05	0.10	0.18
5	0.90	0.07	0.91	0.77
6	0.28	0.81	0.28	0.94

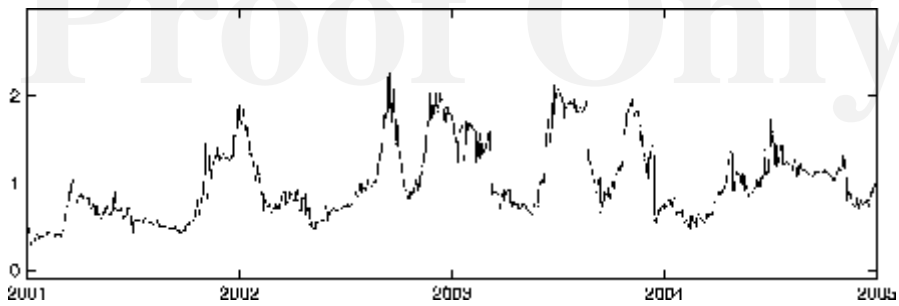


Figure 8. Time-varying dependence, group 1. Copula parameter $\hat{\theta}_t$ estimated with LCP method, Clayton copula, $m_0 = 20$, $c = 1.25$, and $\rho = 0.5$.

extreme events at lower orthants with higher probability than Gaussian copulae for the same correlation, see McNeil, Frey, and Embrechts (2005). This fact is essential for var calculations and is illustrated by the ratio between Equations (2.2) and (2.3) for off-diagonal elements of $\mathbf{\Psi}$ set to 0.25 and $\theta = 0.5$. For the quantiles $u_i = 0.05$, $i = 1, \ldots, 6$ the ratio $C_{\Psi}^{Ga}(u_1, \ldots, u_6)/C_{\theta}(u_1, \ldots, u_6)$ equals 2.3×10^{-2}, whereas for the 0.01 quantiles it equals 1.3×10^{-3}.

The var estimation follows the steps described in Section 3. Using the *RiskMetrics* approach, the log-returns $\boldsymbol{X}_t$ are assumed conditionally normal distributed with zero mean and covariance matrix following a GARCH specification with fixed decay factor $\lambda = 0.94$ as in Equation (3.4).

In the time-varying copulae estimation, the log-returns are modeled as in Equation (3.5), where the innovations ε_t have cdf $F_{t,\varepsilon_t}(x_1, \ldots, x_d) = C_{\theta_t}\{F_{t,1}(x_1), \ldots, F_{t,d}(x_d)\}$ and C_θ is the Clayton copula. The univariate log-returns $X_{t,\,j}$ corresponding to stock j are devolatized according to *RiskMetrics* (i.e., with zero conditional means and conditional variances $\sigma^2_{t,j}$ estimated by the univariate version of Equation (3.4) with a decay factor equal to 0.94). We note that this choice sets the same specification for the dynamics of the univariate returns across all methods (*RiskMetrics*, moving windows, and LCP), making their performances in var estimation comparable. Moreover, as the means from daily returns are clearly dominated by the variances and are approximately independent on the available information sets (see Jorion 1995; Fleming, Kirby, and Ostdiek 2001; and Christoffersen and Diebold 2006), their specification is very unlikely to cause a perceptible bias in the estimated variances and dependence parameters. Therefore, the zero mean assumption is, as pointed out by Kim, Malz, and Mina (1999), as good as any other choice. Daily returns are also modeled with zero conditional means in Fan and Gu (2003) and Härdle, Herwartz, and Spokoiny (2003) among others.

The GARCH specification (Eq. 3.4) with $\lambda = .94$ optimizes variance forecasts across a large number of assets (Morgan 1996), and is widely used in the financial industry. Different choices for the decay factor (like 0.85 or 0.98) result in negligible changes (about 3%) in the estimated dependence parameter.

The p values from the Ljung-Box test for serial correlation and from ARCH test for heteroscedasticity effects in the obtained residuals $\hat{\varepsilon}_{t,j}$ are in Table 3. Normality is rejected by Jarque-Bera test, with p values approximately 0.00 for all residuals in both groups. The empirical cdf's of residuals as defined in Equation (2.4) are used for the copula estimation. T3

With the moving windows approach, the size of the estimating window is fixed as 250 days corresponding to 1 business year (the same size is used in, for example, Fan and Gu (2003)). For the LCP procedure, following Section 4.2, we set the family of interval candidates as a geometric grid with $m_0 = 20$, $c = 1.25$, and $\rho = 0.5$. We have chosen these parameters from our experience in simulations (for details on robustness of the reported results with respect to the choice of m_0 and c, refer to Spokoiny (2008)).

The performance of the var estimation is evaluated based on backtesting. At each time t, the estimated var at level α for a portfolio $\boldsymbol{w}$ is compared with the realization l_t of the corresponding P&L function (see Eq. 3.2), with an exceedance occurring for each l_t less than $\widehat{\text{var}}_t(\alpha)$. The ratio of the number of exceedances to the number of observations gives the exceedance ratio

$$\hat{\alpha}_{\boldsymbol{w}}(\alpha) = \frac{1}{T}\sum_{t=1}^{T} \mathbf{1}_{\{l_t < \widehat{\text{var}}_t(\alpha)\}}.$$

Because the first 250 observations are used for estimation, $T = 1{,}020$. The difference between $\hat{\alpha}$ and the desired level α is expressed by the relative exceedance error

Figure 9. Time-varying dependence, group 2. Copula parameter $\hat{\theta}_t$ estimated with LCP method, Clayton copula, $m_0 = 20$, $c = 1.25$, and $\rho = 0.5$.

JOBNAME: jbes 00#0 2009 PAGE: 9 OUTPUT: Wednesday October 22 04:23:13 2008
asa/jbes/164302/06-114

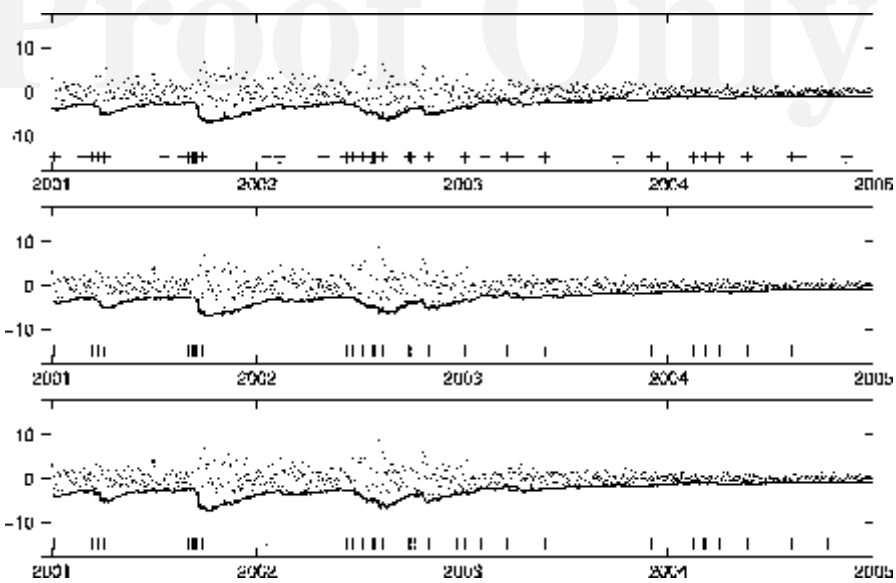


Figure 10. Estimated var across methods, group 1. P&L realizations l_t (dots), $\widehat{\text{var}}_t(\alpha)$ (line), and exceedance times (crosses). Estimated with LCP (top), moving windows (middle), and *RiskMetrics* (bottom) for equally weighted portfolio $\boldsymbol{w}^*$ at level $\alpha = 0.05$.

$$e_{\boldsymbol{w}} = (\hat{\alpha} - \alpha)/\alpha.$$

We compute exceedance ratios and relative exceedance errors to levels $\alpha = 0.05$ and 0.01 for a set $\mathrm{W} = \{\boldsymbol{w}^*, \boldsymbol{w}_n;\ n = 1, \ldots, 100\}$ of portfolios, where each $\boldsymbol{w}_n = (w_{n,1}, \ldots, w_{n,6})^\top$ is a realization of a random vector uniformly distributed on $\mathcal{S} = \{(x_1, \ldots, x_6) \in \mathbb{R}^6 : \sum_{i=1}^6 x_i = 1, x_i \geq .1\}$, and $\boldsymbol{w}^* = 1/6\boldsymbol{I}_6$, with $\boldsymbol{I}_d$ denoting the $(d \times 1)$ vector of ones, is the equally weighted portfolio. The degree of diversification of a portfolio can be measured based on the majorization preordering on $\mathbf{S}$ (see Marshall and Olkin 1979). In other words, a portfolio $\boldsymbol{w}_a$ is more diversified than portfolio $\boldsymbol{w}_b$ if $\boldsymbol{w}_a \prec \boldsymbol{w}_b$. Under the majorization preordering the vector $\boldsymbol{w}^*$ satisfies $\boldsymbol{w}^* \preceq \boldsymbol{w}$ for all $\boldsymbol{w} \in \mathbf{S}$; therefore, the equally weighted portfolio is the most diversified portfolio from W, see Ibragimov and Walden (2007).

The average relative exceedance error over portfolios and the corresponding standard deviation

$$A_{\mathcal{W}} = \frac{1}{|\mathcal{W}|} \sum_{\boldsymbol{w} \in \mathcal{W}} e_{\boldsymbol{w}}$$

$$D_{\mathcal{W}} = \left\{ \frac{1}{|\mathcal{W}|} \sum_{\boldsymbol{w} \in \mathcal{W}} (e_{\boldsymbol{w}} - A_{\mathcal{W}})^2 \right\}^{\frac{1}{2}}$$

are used to evaluate the performances of the time-varying copulae and *RiskMetrics* methods in var estimation.

The dependence parameter estimated with LCP for stocks from groups 1 and 2 are shown in Figures 8 and 9. The different F8 F9 industry concentrations in each group are reflected in the higher parameter values obtained for group 1. The P&L and the var at level 0.05 estimated with LCP, moving windows, and

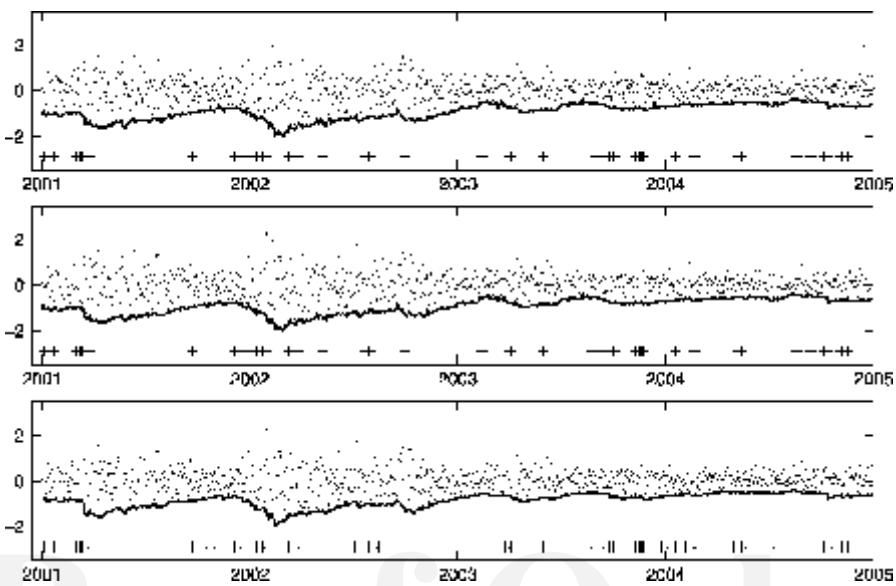


Figure 11. Estimated var across methods, group 2. P&L realizations l_t (dots), $\widehat{\text{var}}_t(\alpha)$ (line), and exceedance times (crosses). Estimated with LCP (top), moving windows (middle), and *RiskMetrics* (bottom) for equally weighted portfolio $\boldsymbol{w}^*$ at level $\alpha = 0.05$.

AU8 Table 4. Exceedance ratios and errors, group 1

	RiskMetrics		Moving windows		LCP	
	$\alpha = 5.00$	$\alpha = 1.00$	$\alpha = 5.00$	$\alpha = 1.00$	$\alpha = 5.00$	$\alpha = 1.00$
$\hat{\alpha}_{w^*}$	6.11	1.48	5.62	0.59	5.52	0.69
$\hat{\alpha}_{w_1}$	5.91	1.38	5.42	0.49	5.42	0.69
$\hat{\alpha}_{w_2}$	6.40	1.28	5.91	0.49	5.71	0.59
A_W	0.23	0.45	0.11	−0.49	0.11	−0.36
D_W	0.04	0.14	0.06	0.08	0.06	0.10

NOTE: Exceedance ratios for portfolios w^*, w_1, and w_2, and average and standard deviation from relative exceedance errors. Across levels and methods, ratios and levels are expressed as a percentage.

RiskMetrics methods for the equally weighted portfolio w^* are F10,11 in Figures 10 (group 1) and 11 (group 2). Exceedance ratios for portfolios w^*, w_1, and w_2; average relative exceedance errors; and corresponding standard deviations across methods and T4,T5 levels are shown in Tables 4 (group 1) and 5 (group 2).

Based on the exceedance errors, the LCP procedure outperforms the moving windows (second best) and *RiskMetrics* methods in var estimation in group 1. At level 0.05, the average error associated with copula methods is about half the error from *RiskMetrics* estimation for nearly the same standard deviation. At level 0.01, the LCP average error is the smallest in absolute value, and copula methods present less standard deviations. At this level, copula methods overestimate var, and *RiskMetrics* underestimates it. Although overestimation of var means that a financial institution would be requested to keep more capital aside than necessary to guarantee the desired confidence level, underestimation means that less capital is reserved and the desired level is not guaranteed. Therefore, from the regulatory point of view, overestimation is preferred to underestimation. In the less concentrated group 2, LCP outperforms moving windows and *RiskMetrics* at the level 0.05, presenting the smallest average error in magnitude for nearly the same value of D_W. At level 0.01, copula methods overestimate and *RiskMetrics* underestimates the var by about 60%.

It is interesting to note the effect of portfolio diversification on the exceedance errors for group 1 and level 0.01. The errors decrease with increasing portfolio diversification for copulae methods but become larger under the *RiskMetrics* estimation. For other groups and levels, the diversification effects are not clear. Refer to Ibragimov (2007) and Ibragimov and Walden (2007) for details on the effects of portfolio diversification under heavy-tailed distributions in risk management.

AU9 Table 5. Exceedance ratios and errors, group 2

	RiskMetrics		Moving windows		LCP	
	$\alpha = 5.00$	$\alpha = 1.00$	$\alpha = 5.00$	$\alpha = 1.00$	$\alpha = 5.00$	$\alpha = 1.00$
$\hat{\alpha}_{w^*}$	5.42	1.58	4.53	0.39	4.53	0.30
$\hat{\alpha}_{w_1}$	5.81	1.77	5.02	0.39	5.02	0.39
$\hat{\alpha}_{w_2}$	5.62	1.58	5.12	0.39	5.22	0.30
A_W	0.16	0.57	−0.10	−0.65	−0.09	−0.65
D_W	0.04	0.16	0.06	0.09	0.06	0.08

NOTE: Exceedance ratios for portfolios w^*, w_1, and w_2, and average and standard deviation from relative exceedance errors. Across levels and methods, ratios and levels are expressed as a percentage.

7. CONCLUSION

In this article we modeled the dependence structure from German equity returns using time-varying copulae with adaptively estimated parameters. In contrast to Patton (2006) and Rodriguez (2007), we neither specified the dynamics nor assumed regime switching models for the copula parameter. The parameter choice was performed under the local homogeneity assumption with homogeneity intervals recovered from the data through local change point analysis.

We used time-varying Clayton copulae, which are asymmetric and present lower tail dependence, to estimate the var from portfolios of two groups of German securities, presenting different levels of industry concentration. *RiskMetrics*, a widely used methodology based on multivariate normal distributions, was chosen as a benchmark for comparison. Based on backtesting, the adaptive copula achieved the best var estimation performance in both groups, with average exceedance errors mostly small in magnitude and corresponding to sufficient capital reserve for covering losses at the desired levels.

The better var estimates provided by Clayton copulae indicate that the dependence structure from German equities may contain nonlinearities and asymmetries, such as stronger dependence at lower tails than at upper tails, that cannot be captured by the multivariate normal distribution. This asymmetry translates into extremely negative returns being more correlated than extremely positive returns. Thus, our results for the German equities resemble those from Longin and Solnik (2001), Ang and Chen (2002) and Patton (2006) for international markets, U.S. equities, and Deutsch mark/Japanese yen exchange rates, where empirical evidence for asymmetric dependences with increasing correlations in market downturns were found.

Furthermore, in the non-Gaussian framework, with nonlinearities and asymmetries taken into consideration through the use of Clayton copulae, the adaptive estimation produces better var fits than the moving window estimation. The high sensitive adaptive procedure can capture local changes in the dependence parameter that are not detected by the estimation with a scrolling window of fixed size.

The main advantage of using time-varying copulae to model dependence dynamics is that the normality assumption is not needed. With the proposed adaptively estimated time-varying copulae, neither normality assumption nor specification for the dependence dynamics are necessary. Hence, the method provides more flexibility in modeling dependences between markets and economies over time.

ACKNOWLEDGMENTS

Financial support from the *Deutsche Forschungsgemeinschaft* via *SFB 649* "Ökonomisches Risiko," Humboldt-Universität zu Berlin is gratefully acknowledged. The authors also thank the editor, an associate editor, and two referees for their helpful comments.

[Received October 2006. Revised November 2007.]

REFERENCES

Andrews, D. W. K. (1993), "Tests for Parameter Instability and Structural Change With Unknown Change Point," *Econometrica,* 61, 821–856.

Andrews, D. W. K., and Ploberger, W. (1994), "Optimal Tests When a Nuisance Parameter Is Present Only Under the Alternative," *Econometrica,* 62, 1383–1414.

Ang, A., and Chen, J. (2002), "Asymmetric Correlations of Equity Portfolios," *Journal of Financial Economics,* 63, 443–494.

Belomestny, D., and Spokoiny, V. (2007), "Spatial Aggregation of Local Likelihood Estimates With Applications to Classification," *The Annals of Statistics,* 35, 2287–2311.

Chen, X., and Fan, Y. (2006), "Estimation and Model Selection of Semiparametric Copula-Based Multivariate Dynamic Models Under Copula Misspecification," *Journal of Econometrics,* 135, 125–154.

Chen, X., Fan, Y., and Tsyrennikov, V. (2006), "Efficient Estimation of Semiparametric Multivariate Copula Models," *Journal of the American Statistical Association,* 101, 1228–1240.

Cherubini, U., Luciano, E., and Vecchiato, W. (2004), *Copula Methods in Finance*, Chichester: Wiley.

Christoffersen, P., and Diebold, F. (2006), "Financial Asset Returns, Direction-of-Change Forecasting, and Volatility Dynamics," *Management Science,* 52, 1273–1287.

Cont, R. (2001), "Empirical Properties of Asset Returns: Stylized Facts and Statistical Issues," *Quantitative Finance,* 1, 223–236.

Embrechts, P., Hoeing, A., and Juri, A. (2003a), "Using Copulae to Bound the Value-at-Risk for Functions of Dependent Risks," *Finance and Stochastics,* 7, 145–167.

Embrechts, P., Lindskog, F., and McNeil, A. (2003b), "Modelling Dependence with Copulas and Applications to Risk Management," in *Handbook of Heavy Tailed Distributions in Finance*, ed. S. Rachev, Amsterdam: North-Holland, pp. 329–384.

Embrechts, P., McNeil, A., and Straumann, D. (2002), "Correlation and Dependence in Risk Management: Properties and Pitfalls," in *Risk Management: Value at Risk and Beyond*, ed. M. Dempster, Cambridge, UK: Cambridge University Press.

Fan, J., and Gu, J. (2003), "Semiparametric Estimation of Value-at-Risk," *The Econometrics Journal,* 6, 261–290.

Fleming, J., Kirby, C., and Ostdiek, B. (2001), "The Economic Value of Volatility Timing," *The Journal of Finance,* 56, 239–354.

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

Fréchet, M. (1951), "Sur les Tableaux de Correlation Dont les Marges Sont Données," *Annales de l'Université de Lyon, Sciences Mathématiques et Astronomie,* 14, 5–77.

Giacomini, E., and Härdle, W. (2005), "Value-at-Risk Calculations With Time Varying Copulae," in *Bulletin of the International Statistical Institute, Proceedings of the 55th Session*.

Granger, C. (2003), "Time Series Concept for Conditional Distributions," *Oxford Bulletin of Economics and Statistics,* 65, 689–701.

Hansen, B. E. (2001), "The New Econometrics of Structural Change: Dating Breaks in U.S. Labor Productivity," *The Journal of Economic Perspectives,* 15, 117–128.

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

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

Hoeffding, W. (1940), "Maßstabinvariante Korrelationstheorie," *Schriften des mathematischen Seminars und des Instituts für angewandte Mathematik der Universität Berlin,* 5, 181–233.

Hu, L. (2006), "Dependence Patterns Across Financial Markets: A Mixed Copula Approach," *Applied Financial Economics,* 16, 717–729.

Ibragimov, R. (2007), "Efficiency of Linear Estimators Under Heavy-Tailedness: Convolutions of α-Symmetric Distributions," *Econometric Theory,* 23, 501–517.

Ibragimov, R., and Walden, J. (2007), "The Limits of Diversification When Losses May be Large," *Journal of Banking and Finance,* 31, 2551–2569.

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

Jorion, P. (1995), "Predicting Volatility in the Foreign Exchange Market," *The Journal of Finance,* 50, 507–528.

Morgan, J. P. (1996), *RiskMetrics Technical Document*, New York: RiskMetrics Group.

Kim, J., Malz, A. M., and Mina, J. (1999), *Long Run Technical Document*, New York: RiskMetrics Group.

Longin, F., and Solnik, B. (2001), "Extreme Correlation on International Equity Markets," *The Journal of Finance,* 56, 649–676.

Mari, D., and Kotz, S. (2001), *Correlation and Dependence*, London: Imperial College Press.

Marshall, A., and Olkin, I. (1979), *Inequalities: Theory of Majorizations and Its Applications*, New York: Academic Press.

McNeil, A. J., Frey, R., and Embrechts, P. (2005), *Quantitative Risk Management: Concepts, Techniques and Tools*, Princeton, NJ: Princeton University Press.

Mercurio, D., and Spokoiny, V. (2004), "Estimation of Time Dependent Volatility via Local Change Point Analysis With Applications to Value-at-Risk," *Annals of Statistics,* 32, 577–602.

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

Patton, A. (2004), "On the Out-of-Sample Importance of Skewness and Asymmetric Dependence for Asset Allocation," *Journal of Financial Econometrics,* 2, 130–168.

——— (2006), "Modelling Asymmetric Exchange Rate Dependence," *International Economic Review,* 47, 527–556.

Perron, P. (1989), "The Great Crash, the Oil Price Shock and the Unit Root Hypothesis," *Econometrica,* 57, 1361–1401.

Polzehl, J., and Spokoiny, V. (2006), "Propagation–Separation Approach for Likelihood Estimation," *Probability Theory and Related Fields,* 135, 335–362.

Quintos, C., Fan, Z., and Philips, P. C. B. (2001), "Structural Change Tests in Tail Behaviour and the Asian Crisis," *The Review of Economic Studies,* 68, 633–663.

Rodriguez, J. C. (2007), "Measuring Financial Contagion: A Copula Approach," *Journal of Empirical Finance,* 14, 401–423.

Sklar, A. (1959), "Fonctions de Répartition à *n* Dimensions et Leurs Marges," *Publications de l'Institut de Statistique de l'Universite de Paris,* 8, 229–231.

Spokoiny, V. (2008), *Local Parametric Methods in Nonparametric Estimation*, Berlin, Heidelberg: Springer-Verlag.

Spokoiny, V., and Chen, Y. (2007), *Multiscale Local Change Point Detection with Applications to Value-at-Risk,* Preprint 904, Berlin: Weierstrass Institute Berlin.

Stock, J.H. (1994), "Unit Roots, Structural Breaks and Trends," in *Handbook of Econometrics,* Vol. 4, ed. R. F. Engle and D. McFadden, Amsterdam: North-Holland, pp. 2739–2841.

Zivot, E., and Andrews, D. W. K. (1992), "Further Evidence on the Great Crash, the Oil Price Shock and the Unit Root Hypothesis," *Journal of Business & Economic Statistics,* 10, 251–270.

Journal of Econometrics 150 (2009) 1–15

Contents lists available at ScienceDirect

Journal of Econometrics

journal homepage: www.elsevier.com/locate/jeconom

Dynamics of state price densities

Wolfgang Härdle [a], Zdeněk Hlávka [b,*]

[a] CASE—Center for Applied Statistics and Economics, Wirtschaftswissenschaftliche Fakultät, Humboldt-Universität zu Berlin, Spandauer Str. 1, 10178 Berlin, Germany

[b] Charles University in Prague, Department of Statistics, Sokolovská 83, 18675 Praha, Czech Republic

ARTICLE INFO

Article history:
Received 11 January 2009
Accepted 12 January 2009
Available online 15 January 2009

JEL classification:
C13
C14
G13

Keywords:
Option pricing
State price density
Nonlinear least squares
Constrained estimation

ABSTRACT

State price densities (SPDs) are an important element in applied quantitative finance. In a Black–Scholes world they are lognormal distributions, but in practice volatility changes and the distribution deviates from log-normality. In order to study the degree of this deviation, we estimate SPDs using EUREX option data on the DAX index via a nonparametric estimator of the second derivative of the (European) call pricing function. The estimator is constrained so as to satisfy no-arbitrage constraints and corrects for the intraday covariance structure in option prices. In contrast to existing methods, we do not use any parametric or smoothness assumptions.

1. Introduction

The dynamics of option prices carries information on changes in state price densities (SPDs). The SPD contains important information on the behavior and expectations of the market and is used for pricing and hedging. The most important application of an SPD is that it allows one to price options with complicated payoff functions simply by (numerical) integration of the payoff with respect to this density.

Prices $C_t(K, T)$ of European options with strike price K observed at time t and expiring at time T allow one to deduce the state price density $f(.)$ using the relationship (Breeden and Litzenberger, 1978)

$$f(K) = \exp\{r(T-t)\}\frac{\partial^2 C_t(K,T)}{\partial K^2}. \quad (1)$$

Eq. (1) can be used to estimate the SPD $f(K)$ from the observed option prices. An extensive overview of parametric and other estimation techniques can be found, for example, in Jackwerth (1999). An application to option pricing is given in Buehler (2006).

Kernel smoothers were in this framework proposed and successfully applied by, for example, Aït-Sahalia and Lo (1998), Aït-Sahalia and Lo (2000), Aït-Sahalia et al. (2000), or Huynh et al. (2002). Aït-Sahalia and Duarte (2003) proposed a method for nonparametric estimation of the SPD under constraints like positivity, convexity, and boundedness of the first derivative. Bondarenko (2003) calculates arbitrage-free SPD estimates using positive convolution approximation (PCA) methodology and demonstrates its properties in a Monte Carlo studied based on closing prices of the S&P 500 options. Another sophisticated approach based on smoothing splines allowing one to include these constraints is described and applied on simulated data in Yatchew and Härdle (2006). In the majority of these papers, the focus was more on the smoothing techniques rather than on a no-arbitrage argument, although a crucial element of local volatility models is the absence of arbitrage (Dupire, 1994). Highly numerically efficient pricing algorithms, for example, by Andersen and Brotherton-Ratcliffe (1997), rely heavily on no-arbitrage properties. Kahalé (2004) proposed a procedure that requires solving a set of nonlinear equations with no guarantee of a unique solution. Moreover, for that algorithm the data feed is already (unrealistically) expected to be arbitrage free (Fengler, 2005; Fengler et al., 2007). In addition, the covariance structure of the quoted option prices (Renault, 1997) is rarely incorporated into the estimation procedure.

In Table 1, we give an overview of selected properties of different estimation techniques. The parametric approach may be used to estimate parameters of a probability density lying in some preselected family. The parametric models may be further extended by considering more flexible probability densities or mixtures of distributions. Approaches based on nonparametric smoothing techniques are more flexible since the shape of a nonparametric SPD estimate is not fixed in advance and the method controls only the smoothness of the estimate. For example,

* Corresponding author. Tel.: +420 221 913 284; fax: +420 283 073 341.
E-mail addresses: haerdle@wiwi.hu-berlin.de (W. Härdle), hlavka@karlin.mff.cuni.cz (Z. Hlávka).

0304-4076/$ – see front matter © 2009 Elsevier B.V. All rights reserved.
doi:10.1016/j.jeconom.2009.01.005

Table 1
Summary of properties of parametric and nonparametric estimators.

	Methods			
	Parametric	Standard smoothing method	Nonparametric under constraints	This paper
Shape	Fixed	Flexible	Flexible	Flexible
Control	Choice of family	Smoothness	Smoothness	None
SPD support	Infinite	Restricted	Restricted	Restricted
Constraints	By design	Local	Yes	Yes

the smoothness of a kernel regression estimator depends mostly on the choice of the bandwidth parameter, the smoothness of the PCA estimator (Bondarenko, 2003) depends on the choice of the kernel, and the smoothness of the NNLS estimator (Yatchew and Härdle, 2006) is controlled by constraining the Sobolev norm of the SPD; using these nonparametric estimators, systematic bias may typically occur in the case of oversmoothing. Constraints on estimators are more easily implemented for globally valid parametric models than for local (nonparametric) models. The use of a standard smoothing technique which does not account for the constraints is not advisable. The value of the nonparametric estimate cannot be calculated in regions without any data and, therefore, the support of nonparametrically estimated SPDs is limited by the range of the observed strike prices even for nonparametric-under-constraints techniques.

Most of the commonly used estimation techniques do not specify explicitly the source of random error in the observed option prices; see Renault (1997) for an extensive review of this subject. A common approach in SPD estimation is to use either the closing option prices or to correct the intraday option prices by the current value of the underlying asset. Both approaches lack interpretation if the shape of the SPD changes rapidly. This can be made clear by a gedankenexperiment: if the shape of the SPD changes dramatically during the day, correcting the observed option prices by the value of the underlying asset and then estimating the SPD would lead to an estimate of some (nonexisting) daily average of the true SPDs. We try to circumvent this problem by introducing a simple model for the intraday covariance structure of option prices which allows us to estimate the value of the true SPD at an arbitrarily chosen fixed time; see also Hlávka and Svojík (2008). Most often, we are interested in the estimation of the current SPD.

We develop a simple estimation technique in order to construct constrained SPD estimates from the observed intraday option prices which are treated as repeated observations collected during a certain time period. The proposed technique involves constrained LS-estimation, it enables us to construct confidence intervals for the current value of the SPD and prediction intervals for its future development, and it does not depend on any tuning (smoothness) parameter. The construction of a simple approximation of the covariance structure of the observed option prices follows naturally from the derivation of our nonparametric constrained estimator. This covariance structure is interesting in itself; it separates two sources of random errors, and it is applicable to other SPD estimators.

We study the development of the estimated SPDs in Germany over 8 years. A no-arbitrage argument is imposed at each time point, leading (mathematically) to the above-mentioned no-arbitrage constraints. This, of course, is a vital feature for trading purposes where the derived (implied) volatility surfaces for different strikes and maturities are needed for proper judgment of risk and return.

The resulting SPDs and implied volatility surfaces are not smooth per se. In most applications, this is not a disadvantage though, since, first, we may smooth the resulting SPD estimates (Hlávka and Svojík, 2008) and, second, we are mostly interested in functionals of the estimated SPD like, for example, the expected payoff or the forward price. Another important feature that can be easily estimated from the nonsmooth SPDs are the quantiles; see Section 6.2 for an application.

In Section 2, we introduce the notation, discuss constraints that are necessary for estimating SPDs, and we construct a very simple unconstrained SPD estimator using simple linear regression. In Section 3, this estimator is modified so that it satisfies the shape constraints given in Section 2.1. We demonstrate that the covariance structure of the option prices exhibits correlations depending both on the strike price and time of the trade in Section 4. In Section 5, we apply our estimation technique on option prices observed in the year 1995, and we show that the proposed approximation of the covariance structure removes the dependency and heteroscedasticity of the residuals. The dynamics of the estimated SPDs in years 1995–2003 is studied in Section 6.

2. Construction of the estimate

The fair price of a European call option with payoff $(S_T - K)_+ = \max(S_T - K, 0)$, with S_T denoting the price of the stock at time T, t the current time, K the strike price, and r the risk-free interest rate, can be written as

$$C_t(K, T) = \exp\{-r(T-t)\} \int_0^\infty (S_T - K)_+ f(S_T) \mathrm{d}S_T, \tag{2}$$

i.e., as the discounted expected value of the payoff with respect to the SPD $f(.)$. For the sake of simplicity of the following presentation, we assume in the rest of the paper that the discount factor $\exp\{-r(T-t)\} = 1$. In applications, this is achieved by correcting the observed option prices by the known risk-free interest rate r and the time to maturity $(T-t)$ in (2). At the time of the trade, the current index price and volatility are common to all options and, hence, do not appear explicitly in Eq. (2).

Let us denote the i-th observation of the strike price by K_i and the corresponding option price, divided by the discount factor $\exp\{-r(T-t)\}$ from (2), by $C_i = C_{t,i}(K_i, T)$. In practice, on any given day t, one observes option prices repeatedly for a small number of distinct strike prices. Therefore, it is useful to adopt the following notation. Let $\mathcal{C} = (C_1, \ldots, C_n)^\top$ be the vector of the observed option prices on day t sorted by strike price. Then, the vector of strike prices has the following structure:

$$\mathcal{K} = \begin{pmatrix} K_1 \\ K_2 \\ \vdots \\ K_n \end{pmatrix} = \begin{pmatrix} k_1 \mathbf{1}_{n_1} \\ k_2 \mathbf{1}_{n_2} \\ \vdots \\ k_p \mathbf{1}_{n_p} \end{pmatrix},$$

where $k_1 < k_2 < \cdots < k_p$, $n_j = \sum_{i=1}^n \mathbf{I}(K_i = k_j)$, with $\mathbf{I}(.)$ denoting the indicator function and $\mathbf{1}_n$ a vector of ones of length n.

2.1. Assumptions and constraints

Let us now concentrate on options corresponding to a single maturity T observed at fixed time t. Let us assume that the i-th observed option price (corresponding to strike price K_i) follows the model

$$C_{t,i}(K_i, T) = \mu(K_i) + \varepsilon_i, \tag{3}$$

where ε_i are iid random variables with zero mean and variance σ^2. In practice, one might expect that the errors exhibit correlations depending on the strike price and time. Heteroscedasticity can

be incorporated in model (3) if we assume that the random errors ε_i have variance $\mathrm{Var}\,\varepsilon_i = \sigma^2_{K_i}$, leading to weighted least squares. The assumptions on the distribution of random errors will be investigated in more detail in Section 5.3. Following Renault (1997), we interpret the observed option price as the price given by a pricing formula plus an error term, and in Section 4 we suggest a covariance structure for the observed option prices taking into account the dependencies across strike prices and times of trade.

Harrison and Pliska (1981) characterized the absence of arbitrage by the existence of a unique risk neutral SPD $f(.)$. From formula (2) and the properties of a probability density it follows that, in a continuous setting, the function $\mu(.)$, defined on $\mathbb{R}^+$, has to satisfy the following no-arbitrage constraints:

1': it is positive,
2': it is decreasing in K,
3': it is convex,
4': its second derivative exists and it is a density (i.e., nonnegative and it integrates to one).

Let us now have a look at functions satisfying Constraints 1'–4'.

Lemma 1. *Suppose that $\mu : \mathbb{R}^+ \to \mathbb{R}^+$ satisfies Constraints 1'–4'. Then the first derivative, $\mu^{(1)}(.)$, is nondecreasing and such that $\lim_{x\to 0}\mu^{(1)}(x) = -1$ and $\lim_{x\to+\infty}\mu^{(1)}(x) = 0$.*

Proof. Constraint 4' implies that the first derivative, $\mu^{(1)}$, exists and that it is differentiable. $\lim_{x\to+\infty}\mu^{(1)}(x)$ exists since the function $\mu^{(1)}$ is nondecreasing (Constraint 3') and bounded (Constraint 2'). Next, $\lim_{x\to\infty}\mu^{(1)}(x) = 0$ since a negative limit would violate Constraint 1' for large x ($\mu^{(1)}(x)$ cannot be positive since $\mu(x)$ is decreasing). Finally, Constraint 4', $1 = \int_0^\infty \mu^{(2)}(x)\mathrm{d}x = \lim_{x\to+\infty}\mu^{(1)}(x) - \lim_{x\to 0}\mu^{(1)}(x)$, implies that $\lim_{x\to 0}\mu^{(1)}(x) = -1$. □

Remark 1. Lemma 1 allows us to restate Constraints 3' and 4' in terms of $\mu^{(1)}(.)$ by assuming that $\mu^{(1)}(.)$ is differentiable, nondecreasing, and such that $\lim_{x\to 0}\mu^{(1)}(x) = -1$ and $\lim_{x\to+\infty}\mu^{(1)}(x) = 0$.

In this section, we stated only constrains guaranteeing that the SPD estimate will be a probability density. Constraints for the expected value of the SPD estimate are discussed in Section 3.6.

2.2. Existence and uniqueness

In this subsection we address the issue of existence and uniqueness of a regression function, $\hat{C}(.)$, satisfying the required assumptions and constraints. In practice, we do not deal with a continuous function. Hence, we restate Constraints 1'–4' for discrete functions, defined only on a finite set of distinct points, say $k_1 < \cdots < k_p$, in terms of their function values, $C(k_i)$, and their scaled first differences, $C^{(1)}_{k_i,k_j} = \{C(k_i) - C(k_j)\}/\{k_i - k_j\}$.

1: $C(k_i) \geq 0, i = 1,\ldots,p$,
2: $k_i < k_j$ implies that $C(k_i) \geq C(k_j)$,
3: $k_i < k_j < k_l$ implies that $-1 \leq C^{(1)}_{k_i,k_j} \leq C^{(1)}_{k_j,k_l} \leq 0$.

It is easy to see that Constraints 1–2 are discrete versions of Constraints 1' and 2'. Constraint 3 is a discrete version of Constraints 3' and 4'; see Remark 1.

From now on, similarly as in Robertson et al. (1988), we think of the collection, $\mathcal{C}$, of functions satisfying Constraints 1–3 as a subset of a p-dimensional Euclidean space, where p is the number of distinct k_i's. The constrained regression, $\hat{C}$, is in this setting the closest point of $\mathcal{C}$ to the vector C of the observed option prices with distances measured by the usual Euclidean distance

$$d(f, C) = (f - C)^\top (f - C) = \sum_{i=1}^{n}\{f(K_i) - C(K_i)\}^2. \tag{4}$$

From this point of view, the regression function, $\hat{C}$, consists only of the values of the function in the points $k_1,\ldots,k_p$. The first and second differences are used to approximate the first and the second derivatives, respectively.

We claim that the set, $\mathcal{C}$, of functions satisfying Constraints 1–3 is closed in the topology induced by the metric given by Euclidean distance and it is convex, i.e., if $f, g \in \mathcal{C}$ and $0 \leq a \leq 1$, then $af + (1-a)g \in \mathcal{C}$.

Lemma 2. *If $\hat{C} \in \mathcal{C}$ is the regression of $C(K_i)$, $i = 1,\ldots,n$, on $k_1 < \cdots < k_p$ under Constraints 1–3 and if a and b are constants such that $a \leq C(K_i) \leq b$, $\forall i$, then $a \leq \hat{C}(k_i) \leq b + (k_p - k_1)$.*

Proof. It is not possible that $\hat{C}(k_i)$ lies above b for all k_i's (otherwise we would get a better fit only by shifting $\hat{C}(k_i)$). The upper bound now follows from Constraint 3.

The validity of the lower bound may be demonstrated similarly. Clearly, it is not possible that $\hat{C}(k_i)$ lie below a for all k_i's. Moreover, it is not possible that $\hat{C}(k_1) \geq \cdots \geq \hat{C}(k_i) \geq a > \hat{C}(k_{i+1}) \geq \cdots \geq \hat{C}(k_p)$ for any i, since in such a situation the fit could be trivially improved by increasing $\hat{C}(k_{i+1}),\ldots,\hat{C}(k_p)$ by some small amount, for example, by $a - \hat{C}(k_{i+1})$, without violating any of the Constraints 1–3. □

Theorem 1. *A regression, $\hat{C} = \arg\min_{f\in\mathcal{C}} d(f, C)$, satisfying Constraints 1–3, exists and it is unique.*

Proof. Lemma 2 implies that $\hat{C}$ belongs to a subset, $\mathcal{S}$, of $\mathcal{C}$ bounded below by a and above by $b + (k_p - k_1)$. Thinking of the functions as points in Euclidean space, it is clear that the continuous function $d(f, C)$ attains its minimum on the closed and bounded set $\mathcal{S}$. The uniqueness of $\hat{C}$ follows from the convexity of $\mathcal{S}$ using, for example, Robertson et al. (1988, Theorem 1.3.1). □

2.3. Linear model

With the given option data, Constraints 1–3 of Section 2.2 can be reformulated using linear regression models with constraints.

In the following, we fix the time t and the expiry date T and we omit these symbols from the notation. In Section 2.2 we have noted that the option prices are repeatedly observed for a small number p of distinct strike prices. Defining the expected values of the option prices for a given strike price, $\mu_j = \mu(k_j) = E\{C(k_j)\}$, we can write

$$\begin{aligned}
\mu_p &= \beta_0,\\
\mu_{p-1} &= \beta_0 + \beta_1,\\
\mu_{p-2} &= \beta_0 + 2\beta_1 + \beta_2,\\
\mu_{p-3} &= \beta_0 + 3\beta_1 + 2\beta_2 + \beta_3,\\
&\vdots\\
\mu_1 &= \beta_0 + (p-1)\beta_1 + (p-2)\beta_2 + \cdots + \beta_{p-1}.
\end{aligned}$$

Thus, we fit our data using coefficients β_j, $j = 1,\ldots,p$. The conditional means μ_i, $i = 1,\ldots,p$ are replaced by the same number of parameters β_j, $j = 0,\ldots,p-1$, which allow us to impose the shape constraints in a more natural way.

The interpretation of the coefficients β_j can be seen in Fig. 1, which shows a simple situation with only four distinct strike prices ($p = 4$). β_0 is the mean option price at point 4. Constraint 1', Section 2.1, implies that it has to be positive. β_1 is the difference between the mean option prices at point 4 and point 3; Constraint 2' implies that it has to be positive. The next coefficient, β_2, approximates the change in first derivative in point 3 and it can be interpreted as an approximation of the second derivative in point 3. Constraint 3' implies that β_2 has to be positive. Similarly, β_3 is an estimate of the (positive) second derivative in point 2. Constraint 4' can be rewritten as $\beta_2 + \beta_3 \leq 1$.

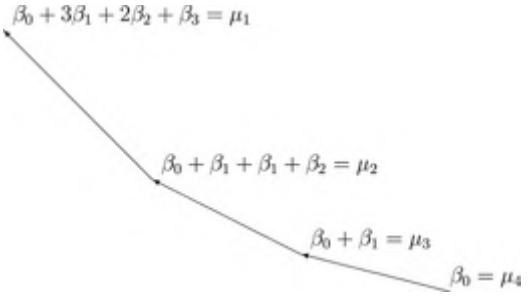


Fig. 1. Illustration of the dummy variables for call options.

In practice, we start with the construction of a design matrix which allows us to write the above model in the following linear form. For simplicity of presentation, we again set $p = 4$:

$$\begin{pmatrix}\mu_1\\ \mu_2\\ \mu_3\\ \mu_4\end{pmatrix} = \begin{pmatrix}1 & 3 & 2 & 1\\ 1 & 2 & 1 & 0\\ 1 & 1 & 0 & 0\\ 1 & 0 & 0 & 0\end{pmatrix}\begin{pmatrix}\beta_0\\ \beta_1\\ \beta_2\\ \beta_3\end{pmatrix}. \tag{5}$$

Ignoring the constraints on the coefficients would lead to a simple linear regression problem. Unfortunately, this approach does not have to lead, and usually does not, to interpretable and stable results.

Model (5) in the above form can be reasonably interpreted only if the observed strike prices are equidistant and if the distances between the neighboring observed strike prices are equal to one. If we want to keep the interpretation of the parameters β_j as the derivatives of the estimated function, we should use the design matrix

$$\Delta = \begin{pmatrix}1 & \Delta_p^1 & \Delta_{p-1}^1 & \Delta_{p-2}^1 & \cdots & \Delta_3^1 & \Delta_2^1\\ 1 & \Delta_p^2 & \Delta_{p-1}^2 & \Delta_{p-2}^2 & \cdots & \Delta_3^2 & 0\\ \vdots & & & & & & \vdots\\ 1 & \Delta_p^{p-2} & \Delta_{p-1}^{p-2} & 0 & \cdots & 0 & 0\\ 1 & \Delta_p^{p-1} & 0 & 0 & \cdots & 0 & 0\\ 1 & 0 & 0 & 0 & \cdots & 0 & 0\end{pmatrix}, \tag{6}$$

where $\Delta_j^i = \max(k_j - k_i, 0)$ denotes the positive part of the distance between k_i and k_j, the i-th and the j-th ($1 \le i \le j \le p$) sorted distinct observed values of the strike price.

The vector of conditional means μ can be written in terms of the parameters β as follows:

$$\begin{pmatrix}\mu_1\\ \mu_2\\ \vdots\\ \mu_p\end{pmatrix} = \mu = \Delta\beta = \Delta\begin{pmatrix}\beta_0\\ \beta_1\\ \vdots\\ \beta_{p-1}\end{pmatrix}. \tag{7}$$

The constraints on the conditional means μ_j can now be expressed as conditions on the parameters of the model (7). Namely, it suffices to request that $\beta_i > 0$, $i = 0, \ldots, p-1$ and that $\sum_{j=2}^{p-1}\beta_j \le 1$.

The model for the option prices can now be written as

$$C(\mathcal{K}) = \mathcal{X}_\Delta\beta + \varepsilon, \tag{8}$$

where $\mathcal{X}_\Delta$ is the design matrix obtained by repeating each row of matrix Δ n_i times, $i = 1, \ldots, p$.

3. Implementing the constraints

In order to impose Constraints 1–3 on parameters β_i, $i = 0, \ldots, p-1$, we propose the following reparameterization of the model in terms of parameters $\theta = (\theta_0, \ldots, \theta_{p-1})^\top$:

$$\begin{aligned}&\beta_0(\theta) = \exp(\theta_0),\\ &\beta_1(\theta) = \exp(\theta_1),\\ &\vdots\\ &\beta_{p-1}(\theta) = \exp(\theta_{p-1}),\end{aligned}$$

under the constraint that $\sum_{j=2}^{p-1}\exp(\theta_j) < 1$. Clearly, the parameters $\beta_i(\theta)$ satisfy the constraints

$$\begin{aligned}&\beta_i(\theta) > 0, \quad i = 0, \ldots, p-1,\\ &\sum_{j=2}^{p-1}\beta_j(\theta) < 1.\end{aligned}$$

This means that the parameters $\beta_2(\theta), \ldots, \beta_{p-1}(\theta)$ can be considered as point estimates of the state price density (the estimates have to be positive and integrate to less than one). Furthermore, in view of Lemma 1, it is worthwhile to note that the parameters also satisfy

$$-\sum_{j=1}^{k}\beta_j \in (-1, 0), \quad \text{for } k = 1, \ldots, p-1.$$

The model (8) rewritten in terms of parameters θ_i, $i = 0, \ldots, p$, is a nonlinear regression model which can be estimated using standard nonlinear least squares or maximum likelihood methods (Seber and Wild, 2003). The main advantage of these methods is that the asymptotic distribution is well known and that the asymptotic variance of the estimator can be approximated using numerical methods implemented in many statistical packages.

3.1. Reparameterization

The following reparameterization of the model in terms of parameters $\xi = (\xi_0, \ldots, \xi_p)^\top$ simplifies the calculation of the estimates because it guarantees that all constraints are automatically satisfied:

$$\begin{aligned}&\beta_0(\xi) = \exp(\xi_0),\\ &\beta_1(\xi) = \frac{\exp(\xi_1)}{\sum_{j=1}^{p}\exp(\xi_j)},\\ &\vdots\\ &\beta_{p-1}(\xi) = \frac{\exp(\xi_{p-1})}{\sum_{j=1}^{p}\exp(\xi_j)}.\end{aligned}$$

This property simplifies the numerical minimization algorithm needed for the calculation of the estimates.

The equality

$$\frac{1}{\sum_{j=1}^{p-1}\beta_j(\xi)} = 1 + \frac{\exp(\xi_p)}{\sum_{j=1}^{p-1}\exp(\xi_j)}$$

shows the meaning of the additional parameter ξ_p. Setting this parameter to $-\infty$ would be the same as requiring that $\sum_{j=1}^{p-1}\beta_j(\xi) = 1$. Large values of the parameter ξ_p indicate that the estimated coefficients sum to less than one or, in other words, the observed strike prices do not cover the support of the estimated SPD. Notice that, by setting $\xi_p = -\infty$, we could easily modify our procedure and impose the equality constraint $\sum_{j=1}^{p-1}\beta_j(\xi) = 1$.

3.2. Inverse transformation of model parameters

For the numerical algorithm, it is useful to know how to calculate ξ's from given β's. This is needed, for example, to obtain reasonable starting points for the iterative procedure maximizing the likelihood.

Lemma 3. *Given $\beta = (\beta_1, \ldots, \beta_p)^\top$, where $\beta_p = 1 - \sum_{i=1}^{p-1} \beta_i$, the parameters $\xi = (\xi_1, \ldots, \xi_p)^\top$ satisfy the system of equations*

$$\left(\beta \mathbf{1}_p^\top - \mathbf{I}_p\right) \exp \xi^\top = \mathcal{A} \exp \xi^\top = 0, \tag{9}$$

where $\mathbf{I}_p$ is the $(p \times p)$ identity matrix. Furthermore,

$$\operatorname{rank} \mathcal{A} = p - 1. \tag{10}$$

The system of Eq. (9) has infinitely many solutions, which can be expressed as

$$\exp(\xi) = \left(\mathcal{A}^- \mathcal{A} - \mathbf{I}_p\right) z, \tag{11}$$

where $\mathcal{A}^-$ denotes a generalized inverse of $\mathcal{A}$ and where z is an arbitrary vector in $\mathbb{R}^p$ such that the right-hand side of (11) is positive.

Proof. Parts (9) and (10) follow from the definition of $\beta(\xi)$ and from simple algebra (notice that the sum of rows of $\mathcal{A}$ is equal to zero). Part (11) follows, for example, from Anděl (1985, Theorem IV.18). □

It remains to choose the vector z in (11) so that the solution of the system of Eq. (9) is positive.

Proposition 1. *The rank of the matrix $\mathcal{A}^-\mathcal{A} - \mathbf{I}_p$ is 1. Hence, any solution of the system of Eq. (9) is a multiple of the first column of the matrix $\mathcal{A}^-\mathcal{A} - \mathbf{I}_p$. The vector z in (11) can be chosen, for example, as $z = \pm\mathbf{1}_p$, where the sign is chosen so that the resulting solution is positive.*

Proof. The definition of a generalized inverse is

$$\mathcal{A}\mathcal{A}^-\mathcal{A} - \mathcal{A} = \mathcal{A}(\mathcal{A}^-\mathcal{A} - \mathbf{I}_p) = 0. \tag{12}$$

Lemma 3 says that $\operatorname{rank}\mathcal{A} = p - 1$ and, hence, Eq. (12) implies that $\operatorname{rank}(\mathcal{A}^-\mathcal{A} - \mathbf{I}_p) \leq 1$. Noticing that $\mathcal{A}^-\mathcal{A} \neq \mathbf{I}_p$ means that $\operatorname{rank}(\mathcal{A}^-\mathcal{A} - \mathbf{I}_p) > 0$, and concludes the proof. □

3.3. The algorithm

The proposed algorithm consists of the following steps:

1: obtain a reasonable initial estimate $\hat{\beta}$, for example, by running the Pool-Adjacent-Violators algorithm (Robertson et al., 1988, Chapter 1) on the unconstrained least squares estimates of the first derivative of the curve,
2: transform the initial estimate $\hat{\beta}$ into the estimate $\hat{\xi}$ using the method described in Section 3.2,
3: estimate the parameters of the model (8) by minimizing the sum of squares $\{C(\mathcal{K}) - \mathcal{X}_\Delta \beta(\xi)\}^\top \{C(\mathcal{K}) - \mathcal{X}_\Delta \beta(\xi)\}$ in terms of ξ (see Section 3.1) using numerical methods.

An application of this simple algorithm on real data is given in Section 5.1.

3.4. Asymptotic confidence intervals

We construct confidence intervals based on the parameterization $\beta(\theta)$ introduced at the beginning of this section. The confidence limits for parameters θ_i are exponentiated in order to obtain valid pointwise confidence bounds for the true SPD. The main advantage of this approach is that such confidence bounds are always positive.

An alternative approach would be to construct confidence intervals based on the parameterizations in terms of β_i (Section 2.3) or ξ_i (Section 3.1). However, the limits of confidence intervals for β_i may be negative and confidence intervals for the SPD based on parameters ξ_i would have very complicated shapes in high-dimensional space and could not be easily calculated and interpreted.

Another approach to the construction of the asymptotic confidence intervals can be based on the maximum likelihood theory. Assuming normality, the log-likelihood for the model (8) can be written as

$$l(C, \mathcal{X}_\Delta, \theta, \sigma) = -n \log \sigma - \frac{1}{2\sigma^2} \{C - \mathcal{X}_\Delta \beta(\theta)\}^\top \times \{C - \mathcal{X}_\Delta \beta(\theta)\}, \tag{13}$$

where $\mathcal{X}_\Delta$ is the design matrix given in (8). This normality assumption is justified later by a residual analysis. The maximum likelihood estimator is defined as

$$\hat{\theta} = \arg\max_\theta l(C, \mathcal{X}_\Delta, \theta, \sigma), \tag{14}$$

and it has asymptotically a p-dimensional normal distribution with mean θ and the variance given by the inverse of the Fisher information matrix:

$$\mathcal{F}_n^{-1} = \left\{ -E\left(\frac{\partial^2}{\partial\theta\partial\theta^\top} l(C, \mathcal{X}_\Delta, \theta, \sigma) \right) \right\}^{-1}. \tag{15}$$

More precisely, $n^{1/2}(\hat{\theta} - \theta) \xrightarrow{\mathcal{L}} N_p(0, \mathcal{F}_n^{-1})$. In this framework, the Fisher information matrix can be estimated by using the numerically differentiated Hessian matrix of the log-likelihood. For details we refer, for example, to Serfling (1980, Chapter 4). The confidence intervals calculated for parameters θ may be transformed (exponentiated) to a confidence intervals for the SPD (β). We have not pursued the maximum likelihood approach since it was numerically less stable in this situation.

Note that, under the assumptions of normality, the maximum likelihood estimate is equal to the nonlinear least squares estimate (Seber and Wild, 2003, Section 2.2), and the asymptotic variance of $\hat{\theta} = \exp(\beta)$ may be approximated by $\widehat{\operatorname{Var}}\,\hat{\theta} = \{\operatorname{diag}(\exp\hat{\theta})\mathcal{X}_\Delta^\top \mathcal{X}_\Delta \operatorname{diag}(\exp\hat{\theta})\}^{-1}\hat{\sigma}^2$. Hence, asymptotic confidence intervals for θ_i may be calculated as $(\hat{\theta}_i \pm u_{1-\alpha/2}\hat{s}_{ii})$, where $u_{1-\alpha/2}$ is the $1 - \alpha/2$ quantile of the standard Normal distribution and $\hat{s}_{ii}$ denotes the i-th diagonal element of $\widehat{\operatorname{Var}}\,\hat{\theta}$. By exponentiating both limits of this confidence interval, we immediately obtain the $1 - \alpha$ confidence interval for $\beta_i = \exp\theta_i$.

The construction of the estimator guarantees that the matrix $\mathcal{X}_\Delta$ has full rank—this implies that $\mathcal{X}_\Delta^\top \mathcal{X}_\Delta$ is invertible and the asymptotic variance matrix $\widehat{\operatorname{Var}}\,\hat{\theta}$ always exists. If the number of observations is equal to the number of distinct strike prices (if there is only one option price for each strike price), it may happen that $\hat{\sigma}^2 = 0$ and the confidence intervals degenerate to a single point.

3.5. Put–Call parity

The prices of put options can be easily included in our estimation technique by applying the Put–Call parity of the option prices. Assuming that there are no dividends or costs connected with the ownership of the stock, each put option with price $P_t(K, T)$ corresponds to a call option with price

$$C_t(K, T) = P_t(K, T) + S_t - K\mathrm{e}^{-r(T-t)}.$$

In this way, the prices of the put options can be converted into the prices of call options and used in our model (Stoll, 1969). Statistically speaking, these additional observations will increase the precision of the SPD and will lead to more stable results.

In Germany, the Put–Call parity might be biased by an effect of the DAX index calculation which is based on the assumption that

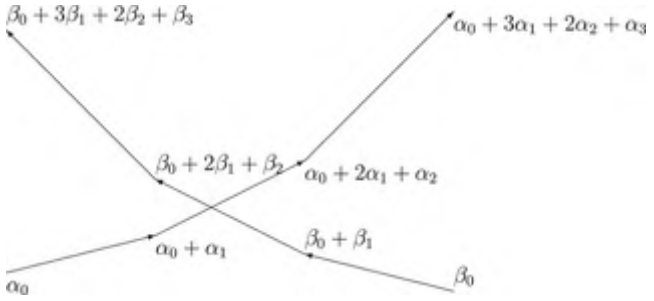


Fig. 2. Illustration of the dummy variables for both call (β) and put (α) options.

the dividends are reinvested after deduction of corporate income tax. As the income tax of some investors might be different, the value of the DAX has to be corrected before using Put–Call parity in subsequent analysis. For the exact description of this correction we refer to Hafner and Wallmeier (2000) who were analyzing the same data set.

The construction of our estimates allows us to include the put option prices in a more direct way by fitting the two curves separately using two sets of parameters. The situation is displayed in Fig. 2. Our assumption that the same SPD drives both the put and call option prices is naturally translated in terms of the coefficients α_i and β_i:

$$\alpha_i = \beta_{p-i+1}, \quad \text{for } i = 2, \ldots, p-1$$

$$\alpha_1 = 1 - \sum_{i=1}^{p-1} \beta_i.$$

The problem of estimating regression functions under such linear equality constraints is solved, for example, in Rao (1973). In Section 4.3, we will also investigate the covariance of the observed call and put option prices, and the suggested model will be presented in detail.

3.6. Expected value constraints

In Section 2.3, we have explained that the parameters $\beta_2, \ldots, \beta_{p-1}$ can be interpreted as estimates of the state price density in points $k_2, \ldots, k_{p-1}$. From the construction of the estimator, see also Fig. 1, it follows that parameter β_1 can be interpreted as the mass of the SPD lying to the right of k_{p-1}. Assuming that the observed strike prices entirely cover the support of the SPD, the mass β_1 could be attributed to the point k_p. Notice that the reparameterization introduced in Section 3 guarantees that $\sum_{i=1}^{p-1} \beta_i(\xi) < 1$, and it immediately follows that interpreting β_1 as the estimate of the SPD in point k_p does not violate any constraints described in Section 2.2.

Referring to Section 3.5, it is clear that the parameter $\beta_p \equiv \alpha_1 = 1 - \sum_{i=1}^{p-1} \beta_i$ can be interpreted as the estimator of the SPD in k_1. The parameterization of the problem now guarantees that $\sum_{i=1}^{p} \beta_i = 1$.

The expected value of the underlying stock under the risk-neutral measure can now be estimated as $\widehat{E^{\mathrm{SPD}}} = \sum_{i=1}^{p} k_i \beta_{p-i+1}$. From economic theory it follows that $\widehat{E^{\mathrm{SPD}}}$ has to be equal to the forward price of the stock. This constraint can be easily implemented by using the fact that β_1 and β_p estimate the mass of the SPD respectively to the right of k_{p-1} and to the left of k_2.

If $\widehat{E^{\mathrm{SPD}}}$ is smaller than the forward price $\exp\{r(T-t)\}S_t$ of the stock, it suffices to move the mass β_1 further to the right. If $\widehat{E^{\mathrm{SPD}}}$ is too large, we move the mass β_p to the left. More precisely, setting

$$\widetilde{k}_1 = k_1 - \mathbf{I}(\widehat{E^{\mathrm{SPD}}} > \exp\{r(T-t)\}S_t)(\widehat{E^{\mathrm{SPD}}} - \exp\{r(T-t)\}S_t)/\beta_p,$$

$$\widetilde{k}_p = k_p + \mathbf{I}(\widehat{E^{\mathrm{SPD}}} < \exp\{r(T-t)\}S_t)(\exp\{r(T-t)\}S_t - \widehat{E^{\mathrm{SPD}}})/\beta_1,$$

we get

$$\exp\{r(T-t)\}S_t = \widetilde{k}_1\beta_p + \sum_{i=2}^{p-1} k_i\beta_{p-i+1} + \widetilde{k}_p\beta_1.$$

This choice of $\widetilde{k}_1$ and $\widetilde{k}_p$ guarantees that the expected value corresponding to the estimator $\beta_1, \ldots, \beta_p$ is equal to the forward price S_t of the stock; see the beginning of Section 6 for an application of this technique.

In Sections 4 and 5, we will concentrate on the properties of $\beta_2, \ldots, \beta_{p-1}$ and further improvements in the estimation procedure.

4. Covariance structure

In this section, we use a model for the SPD development throughout the day to derive the covariance structure of the observed option prices depending on the strike prices and time of the trade. Considering the covariance structure in the estimation procedure solves the problems with heteroscedasticity and correlation of residuals that will be demonstrated in Section 5.3.

In this model, most recent option prices have the smallest variance and thus the largest weight in the estimation procedure. Similarly, the covariance of two option prices with the same strike price at approximately the same time is larger than the covariances of prices of some more dissimilar options.

We start by rewriting the model with iid error terms so that it can be more easily generalized. In Section 4.1, we present a model that accounts for heteroscedasticity and which is further developed in Sections 4.2 and 4.3, where an approximation of the covariance is calculated for any two options prices using only their strike prices and time of the trade. In Section 4.4, we suggest decomposing the error term into two parts, and we show how to estimate these additional parameters by the maximum likelihood method. The analysis of the resulting standardized residuals in Section 5.4 suggests that this covariance structure is applicable to our dataset.

Until now, we have assumed that the i-th option price (on a fixed day t) satisfies

$$C_i(k_j) = \Delta_j\widetilde{\beta} + \varepsilon_i \tag{16}$$

or

$$C_i(k_j) = \Delta_j\widetilde{\beta}_i + \varepsilon_i,$$

$$\widetilde{\beta}_i = \widetilde{\beta}_{i-1}, \tag{17}$$

where ε_i are iid random errors with zero mean and constant variance σ^2, $\widetilde{\beta} = \widetilde{\beta}_1 = \cdots = \widetilde{\beta}_i$ denotes the column vector of the unknown parameters, and Δ_j denotes the j-th row of the matrix Δ defined in (6), i.e.,

$$\Delta_j = (1, \Delta_p^j, \Delta_{p-1}^j, \ldots, \Delta_{j+1}^j, \underbrace{0, \ldots, 0}_{(j-1)}).$$

The residual analysis in Section 5.3 clearly demonstrates that the random errors ε_i are not independent and homoscedastic, and we have to consider some generalizations that lead to a better fit of the data set.

4.1. Heteroscedasticity

Assume that the i-th observation, corresponding to the j-th smallest exercise price k_j, can be written as

$$C_i(k_j) = \Delta_j\widetilde{\beta}_i, \tag{18}$$

$$\widetilde{\beta}_i = \widetilde{\beta} + \varepsilon_i, \tag{19}$$

i.e., there are iid random vectors ε_i having iid components with zero mean and variances σ^2 in the state price density $\widetilde{\beta}_i$. Clearly, the variance matrix of the vector of the observed option prices C is then

$$\mathrm{Var}\, C = \sigma^2 \mathrm{diag}(\mathcal{X}_\Delta \mathcal{X}_\Delta^\top), \tag{20}$$

where $\mathcal{X}_\Delta$ is the design matrix in which each row of the matrix Δ is repeated n_j times, $j = 1, \ldots, p$.

Remark 2. Assuming that the observed option prices have the covariance structure (20), the least squares estimates do not change, and

$$\operatorname{Var}\hat{\beta} = \sigma^2\{\mathcal{X}_\Delta^\top \operatorname{diag}(\mathcal{X}_\Delta\mathcal{X}_\Delta^\top)^{-1}\mathcal{X}_\Delta\}.$$

Another possible model for the heteroscedasticity would assume that the changes are multiplicative rather than additive.

$$C_i(k_j) = \Delta_j\widetilde{\beta}_i$$
$$\log\widetilde{\beta}_i = \log\widetilde{\beta} + \varepsilon_i.$$

This model leads to a variance of $C_i(k_j)$ that depends on the value of the SPD:

$$\operatorname{Var} C_i(k_j) = \sigma^2\{\beta_0^2 + (\Delta_p^j)^2\beta_1^2 + (\Delta_{p-1}^j)^2\beta_2^2 + (\Delta_{p-2}^j)^2\beta_3^2 + \cdots + (\Delta_{j+1}^j)^2\beta_j^2\}.$$

It is straightforward that Remark 2 also applies in this situation.

4.2. Covariance

Let us now assume that there are random changes in the state price density coefficients $\widetilde{\beta}_i$ over time so that we have

$$C_i(k_j) = \Delta_j\widetilde{\beta}_i,$$
$$\widetilde{\beta}_i = \widetilde{\beta}_{i-1} + \varepsilon_i, \tag{21}$$

where, for fixed i, $\widetilde{\beta}_i$ is the parameter vector and ε_k, $k = i, i-1, \ldots$, are iid random vectors having iid components with zero mean and variances σ^2. For nonequidistant time points, let δ_i denote the time between the i-th and $(i-1)$-th observation. The model is

$$C_i(k_j) = \Delta_j\widetilde{\beta}_i,$$
$$\widetilde{\beta}_i = \widetilde{\beta}_{i-1} + \delta_i^{1/2}\varepsilon_i, \tag{22}$$

and it leads to the covariance matrix with elements

$$\operatorname{Cov}\{C_{i-u}(k_j), C_{i-v}(k_i)\} = \operatorname{Cov}(\Delta_j\widetilde{\beta}_{i-u}, \Delta_i\widetilde{\beta}_{i-v}) = \sigma^2\Delta_j\Delta_i^\top\sum_{l=1}^{\min(u,v)}\delta_{i+1-l}. \tag{23}$$

When we observe the i-th observation, we are usually interested in the estimation of the current value of the vector of parameters $\widetilde{\beta}_i$.

4.3. Including put options

Similarly, we obtain the covariance for the price of the put options, $P_i(k_j)$. Using the relations between the α and β parameters, $\alpha_k = \beta_{p-k+1}$, for $k = 2, \ldots, p-1$, and after some simplifications, we can write the model for the price of the put options, $P_i(k_j)$, as

$$P_i(k_j) = \Delta_j\widetilde{\alpha}_i,$$
$$\widetilde{\alpha}_i = \widetilde{\alpha}_{i-1} + \delta_i^{1/2}\varepsilon_i, \tag{24}$$

where $\widetilde{\alpha} = (\alpha_0, \alpha_1, \beta_{p-1}, \beta_{p-2}, \ldots, \beta_2)^\top$ and Δ_j^P denotes the corresponding row of the design matrix, i.e.,

$$\Delta_j^P = (1, \Delta_j^1, \Delta_j^2, \ldots, \Delta_j^{j-1}, \underbrace{0, \ldots, 0}_{(p-j)}).$$

In this way, we obtain a joint estimation strategy for both the call and put option prices:

$$C_i(k_j) = \Delta_j\widetilde{\beta}_i,$$
$$P_i(k_j) = \Delta_j^P\widetilde{\alpha}_i,$$
$$\begin{pmatrix}\widetilde{\beta}_i\\ \widetilde{\alpha}_i\end{pmatrix} = \begin{pmatrix}\widetilde{\beta}_{i-1}\\ \widetilde{\alpha}_{i-1}\end{pmatrix} + \delta_i^{1/2}\varepsilon_i, \tag{25}$$

which directly leads to covariances

$$\operatorname{Cov}\{P_{i-u}(k_j), P_{i-v}(k_i)\} = \operatorname{Cov}(\Delta_j^P\widetilde{\alpha}_{i-u}, \Delta_i^P\widetilde{\alpha}_{i-v}) = \sigma^2\Delta_j^P(\Delta_i^P)^\top\sum_{l=1}^{\min(u,v)}\delta_{i+1-l} \tag{26}$$

and

$$\operatorname{Cov}\{C_{i-u}(k_j), P_{i-v}(k_i)\} = \operatorname{Cov}(\Delta_j\widetilde{\beta}_{i-u}, \Delta_i^P\widetilde{\alpha}_{i-v}) = \sigma^2\sum_{l=1}^{\min(u,v)}\delta_{i+1-l}\sum_{k=2}^{p-1}\Delta_{p+1-k}^j\Delta_i^{p+1-k}. \tag{27}$$

Together with (23), Eq. (26) and (27) allow us to calculate the covariance matrix of all observed option prices using only their strike prices and the times between the transactions.

4.4. Error term for option prices

Using the model (25) would mean that all changes observed in the option prices are due only to changes in the SPD. It seems natural to add another error term, η_i, as a description of the error in the option price:

$$C_i(k_j) = \Delta_j\widetilde{\beta}_i + \eta_i,$$
$$P_i(k_j) = \Delta_j^P\widetilde{\alpha}_i + \eta_i,$$
$$\begin{pmatrix}\widetilde{\beta}_i\\ \widetilde{\alpha}_i\end{pmatrix} = \begin{pmatrix}\widetilde{\beta}_{i-1}\\ \widetilde{\alpha}_{i-1}\end{pmatrix} + \delta_i^{1/2}\varepsilon_i, \tag{28}$$

where $\eta_i \sim N(0, \nu^2)$ are iid random variables independent of the random vectors ε_i. Here, normality assumptions are added both for η_i and ε_i so that the variance components parameters ν^2 and σ^2 may be estimated by the maximum likelihood method.

Next, in order to simplify the notation, let us fix the index i, and let Y denote the vector of observed call and put option prices, $\mathcal{X}_\Delta$ the corresponding design matrix consisting of the corresponding rows Δ_j and Δ_j^P, and $\widetilde{\gamma}$ the combined vector of unknown parameters. Denoting by Σ_i the matrix containing the covariances defined in (23), (26) and (27), we can rewrite model (25) as

$$Y = \mathcal{X}_\Delta\widetilde{\gamma} + \xi, \tag{29}$$

where $\operatorname{Var}\xi = \operatorname{Var} Y = \sigma^2\Sigma_i + \nu^2\mathcal{I}_n = \sigma^2(\Sigma_i + \psi^2\mathcal{I}_n) = \sigma^2 V$, where $\psi^2 = \nu^2/\sigma^2$. Differentiating the log-likelihood

$$l(\beta, \sigma^2, \psi^2) = -\frac{n}{2}\log(2\pi) - \frac{1}{2}\log|\sigma^2 V| - \frac{1}{2\sigma^2}(Y - \mathcal{X}_\Delta\widetilde{\gamma})^\top V^{-1}(Y - \mathcal{X}_\Delta\widetilde{\gamma}),$$

we obtain

$$\frac{\partial l(\beta, \sigma^2, \psi^2)}{\partial\psi^2} = -\frac{1}{2}\operatorname{tr}(V^{-1}) + \frac{1}{2\sigma^2}(Y - \mathcal{X}_\Delta\widetilde{\gamma})^\top V^{-2}(Y - \mathcal{X}_\Delta\widetilde{\gamma}). \tag{30}$$

For any fixed value of the parameter ψ^2, it is straightforward to calculate the optimal σ^2 and $\widetilde{\gamma}$. Hence, the numerical maximization of the log-likelihood can be based on a search for a root (zero) of the one-dimensional function (30).

Moreover, the variance components parameters σ^2 and $\nu^2 = \psi^2\sigma^2$ have a very natural econometric interpretation: σ^2 describes the speed of change of the SPD and ν^2 the error in observed option prices.

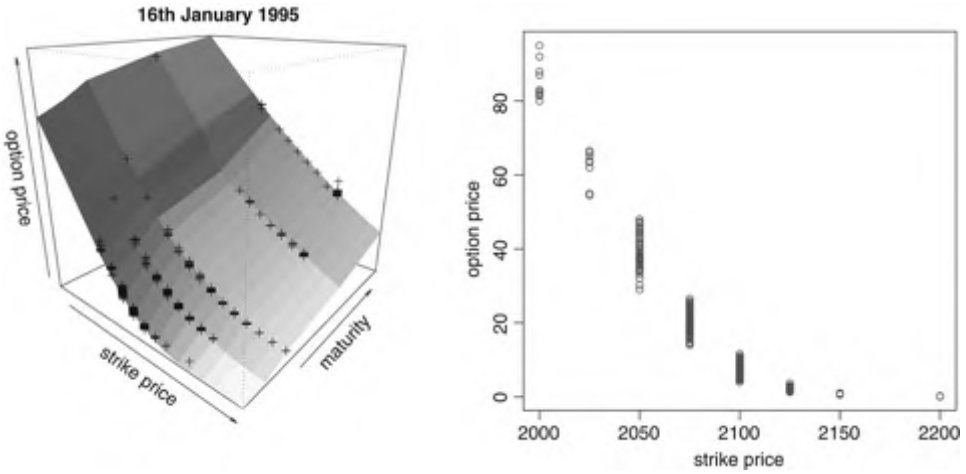


Fig. 3. Option prices plotted against strike price and time to maturity with a two-dimensional kernel regression surface (left) in January 1995 and the ensemble of the call option prices with shortest time to expiry against strike price (right) on 16 January 1995. SFB and CASE data base: sfb649.wiwi.hu-berlin.de.

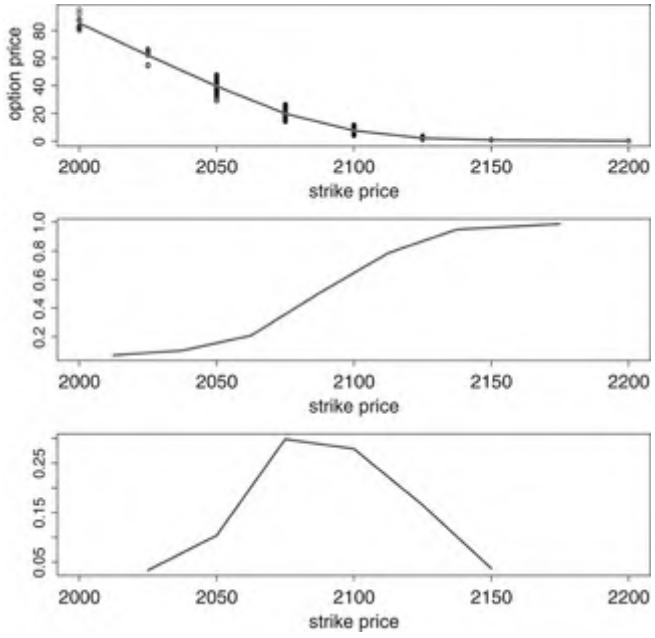


Fig. 4. On 16 January 1995, the unconstrained estimate satisfies the constraints. Hence, it is equal to the constrained estimate. The top panel shows the original data with the fitted call pricing functions. The second and the third panels show the estimates of the first and second derivatives, respectively.

5. Application to DAX data

We analyze a data set containing observed option prices for various strike prices and maturities. Other variables are the interest rate, date, and time. In 1995, one observed every day about 500 trades; in today's more liquid option markets this number has increased approximately 10 times. In our empirical study we will consider the time period from 1995 to 2003, thus also covering more recent liquid option market.

Fig. 3 displays the observed prices of European call options written on the DAX for the 16 January 1995. The left panel shows the ensemble of call option prices for different strikes and maturities as a free structure together with a smooth surface. The typical shape of dependency of the option price on the strike price can be observed in the right panel, containing the option prices only for the shortest time to expiry, $\tau = T - t = 4$ days.

In order to illustrate the method, we apply it to DAX option prices on two consecutive days. These days (16 and 17 January 1995) were selected since they provide a nice insight into the behavior of the presented methods.

5.1. Estimator with iid random errors

We start by a comparison of the unconstrained and constrained estimator described respectively in Sections 2.3 and 3.1.

For the European call option prices displayed in the right-hand plot in Fig. 3, we obtain the estimates plotted in Fig. 4. The top plot displays the original data, the second plot shows the estimate of the first derivative, and the third plot shows the estimate of the second derivative, i.e., the state price density. Actually, all plots contain two curves, both obtained using model (8). The thick line is calculated using the parameters β_i without constraints, whereas the thin line uses the reparameterization $\beta_i(\xi)$ given in Section 3.1. In Fig. 4, these two estimates coincide since the model maximizing the likelihood without constraints, by chance, fulfills the constraints $(\exists \xi : \beta_i = \beta_i(\xi), i = 0, \ldots, p-1)$, and hence it is clear that the same parameters also maximize the constrained likelihood.

The situation, in which the call pricing functions fitted with and without constraints differ, is displayed in Fig. 5. Notice that the difference between the two regression curves is small, whereas the difference between the estimates of the state price density (i.e., the second derivative of the curve) is surprisingly large. The unconstrained estimate shows very unstable behavior on the left-hand side of the plot. The constrained version behaves more reasonably. Very small differences between the fitted call pricing functions in the top plot in Fig. 5 lead to huge differences in the estimates of the second derivative.

We therefore conclude that a small error in the estimate of the call pricing function may lead to large scale error in the estimates of the first and second derivatives. The scale of this type of error seems to be limited by imposing the shape constraints given in Section 2.2.

5.2. Confidence intervals

In Figs. 6 and 7, we plot both estimates together with the 95% confidence intervals. Notice that, in the unconstrained model, the estimates of the values of the SPD are just the parameters of the linear regression model. Hence, the confidence intervals for the parameters are, at the same time, also confidence intervals for the SPD. These confidence intervals for 16 and 17 January 1995 are displayed in the upper plots in Figs. 6 and 7. The drawbacks of this method are clearly visible. In Fig. 6, the lower bounds of the confidence intervals only asymptotically satisfy the condition of

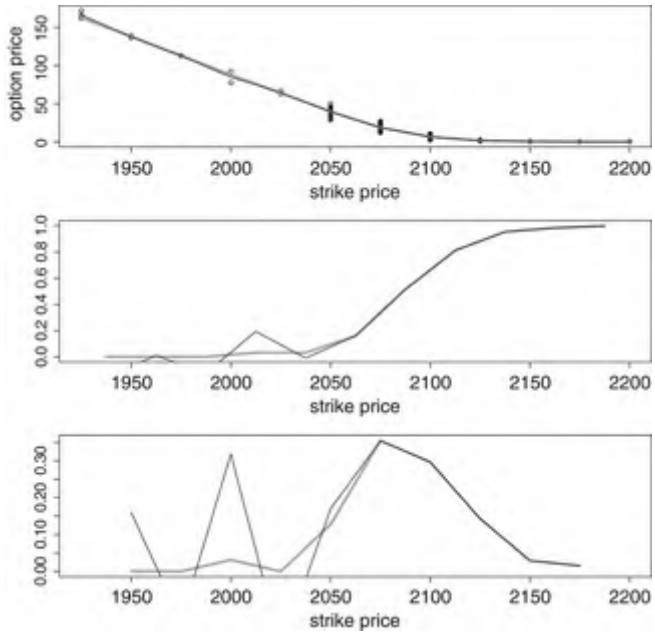


Fig. 5. On 17 January 1995, the unconstrained estimate, displayed using the thin line, does not satisfy the constraints. The top panel shows the original data with the two fitted call pricing functions. The estimates of the first derivative in the second panel look rather different. The constrained estimate of the second derivative in the bottom panel is clearly much more stable than the unconstrained estimate.

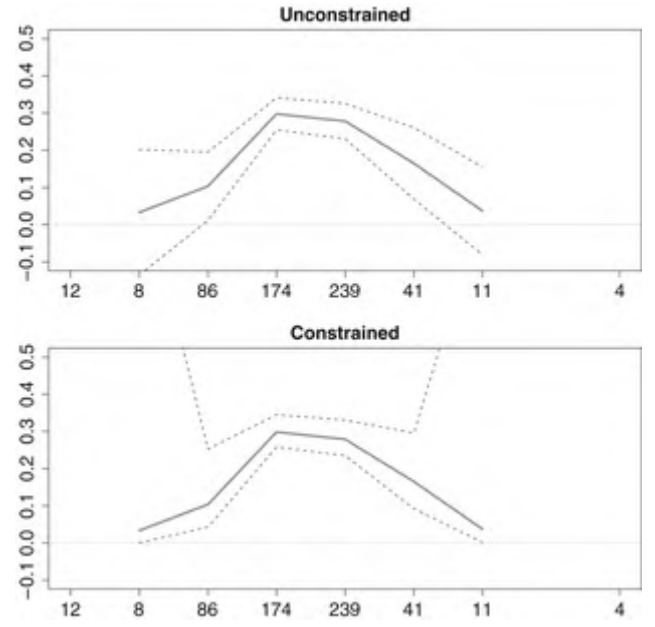


Fig. 6. The unconstrained and constrained confidence intervals for the SPD on 16 January 1995. The description on the *x*-axis shows the number of observations in each point.

positivity. In Fig. 7, we observe large variability on the left-hand side of the plot (the region with low number of observations). Again, some of the lower bounds are not positive. Clearly, the confidence intervals based on the unconstrained model make sense only if the constraints are, by chance, satisfied. Even if this is the case, there is no guarantee that the lower bounds will be positive. The lower panels in Figs. 6 and 7 display the nonnegative asymptotic confidence intervals calculated according to Section 3.4.

In Fig. 6, both types of confidence interval provide very similar results. The only difference is at the minimum and maximum value

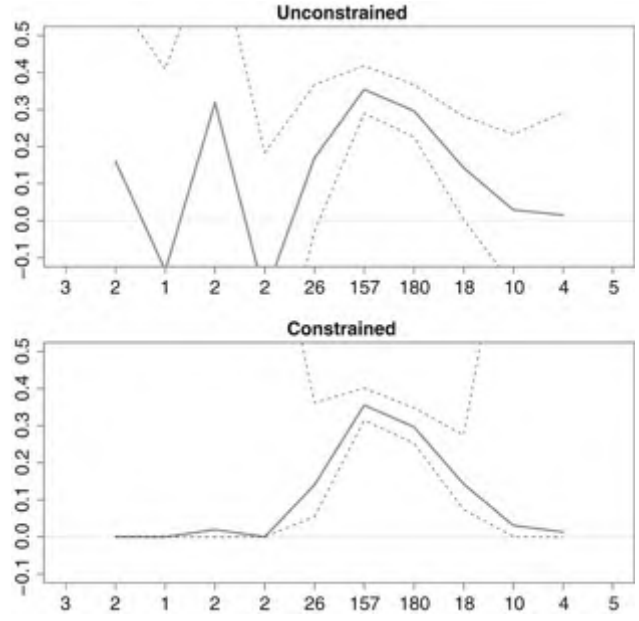


Fig. 7. Confidence intervals for SPD on 17 January 1995. The description on the *x*-axis shows the number of observations in each point.

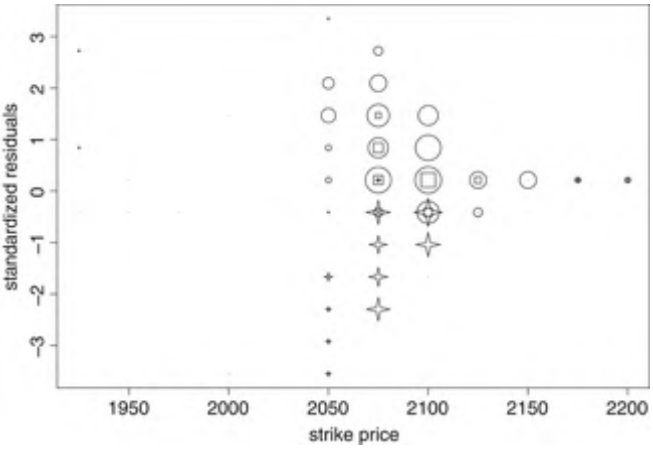


Fig. 8. The time dependency and the heteroscedasticity of the residuals during one day. The circle, square, and star denote the trades carried out in the morning, midday, and afternoon, respectively. The size of the symbols denotes the number of residuals.

of the independent variable (strike price), where the unconstrained method provides negative lower bounds and the conditional method leads to very large upper bounds of the confidence intervals.

In Fig. 7, we plot the confidence intervals for 17 January 1995. In the central region of the graphics, both types of confidence interval are quite similar. On the left-hand and right-hand sides, both methods tend to provide confidence intervals that seem to be overly wide. For the constrained method, we observe that the length of the confidence intervals explodes when the estimated value of the SPD is very close to zero and, at the same time, the number of observation in that region (see the description of the horizontal axis) is small.

5.3. Residual analysis

The residuals on 17 January 1995 are plotted in Fig. 8. The time of trade (in hours) is denoted by the plotting symbol. The circle, square, and star denote the trades carried out in the morning, midday, and afternoon, respectively. The size of the symbols

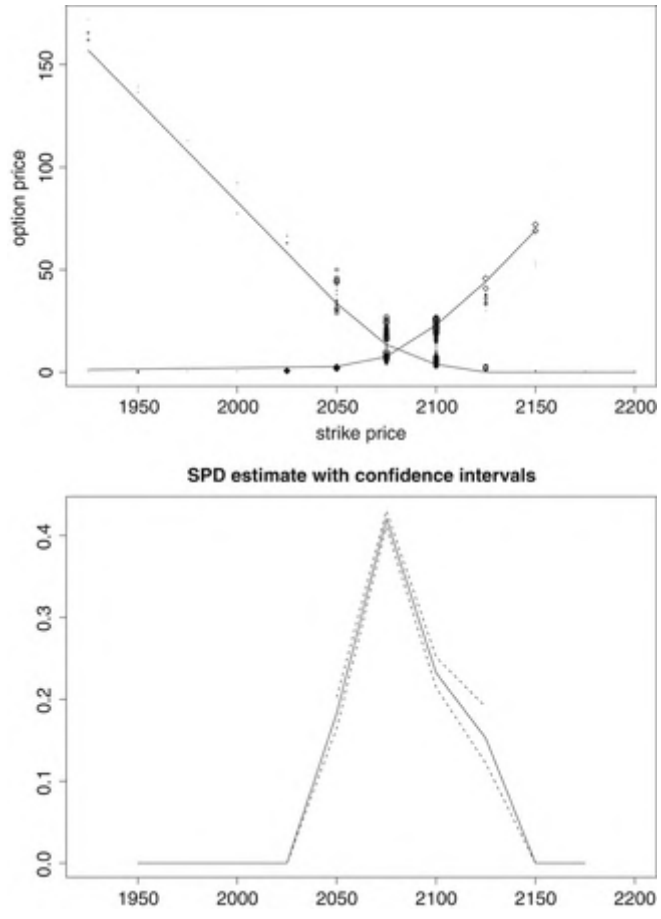


Fig. 9. Estimate using the covariance structure (28) on 17 January 1995. The upper plot shows the observed option prices and the constrained estimate. The size of the plotting symbols corresponds to the weight of the observations. The lower plot shows the estimated SPD with confidence intervals.

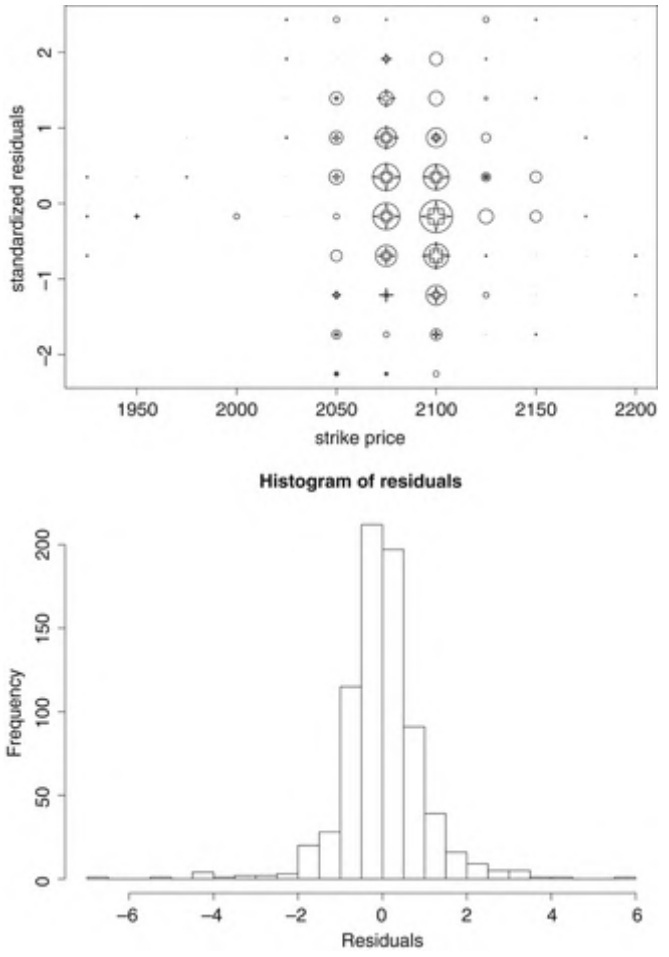


Fig. 10. The development of the standardized residuals resulting from the model with the covariance structure (28) on 17 January 1995 during the day, where circles, squares, and stars denote the residuals from morning, midday, and afternoon, and a histogram of the standardized residuals.

corresponds to the number of residuals lying in the respective areas.

The majority of the residuals correspond to the strike prices of 2075DEM and 2100DEM. The variance of the residuals is very low on the right-hand side of the plot and it rapidly increases when moving towards smaller strike prices. On the left-hand side of the plot, for strike prices smaller than 2000, we have only very few observations, and cannot judge the residual variability reliably.

Apart from the obvious heteroscedasticity we also observe a very strong systematic movement in the SPD throughout the day: the circles, corresponding to the first third of the day, are positive, and all stars, denoting the afternoon residuals, are negative. Similar patterns can be observed every day—residuals corresponding to the same time have the same sign.

We conclude that the assumption of iid random errors is obviously not fulfilled as the option prices tend to follow the changes of the market during the day.

5.4. Application of the covariance structure

In Fig. 9, we present the estimator combining both put and call option prices and using the covariance structure proposed in Section 4.4. In comparison with the results plotted in Fig. 7, we observe shorter length of the confidence intervals.

The estimates of the variance components parameters are $\hat{\psi}^2 = 17.77$, $\hat{\sigma}^2 = 0.0041$, and $\hat{\nu}^2 = 0.0722$. For interpretation, it is more natural to consider $\hat{\nu} = 0.2687$, suggesting that 95% of the option prices were on 17 January 1995 not further than 0.5DEM from the correct option price implied by the current (unobserved) SPD.

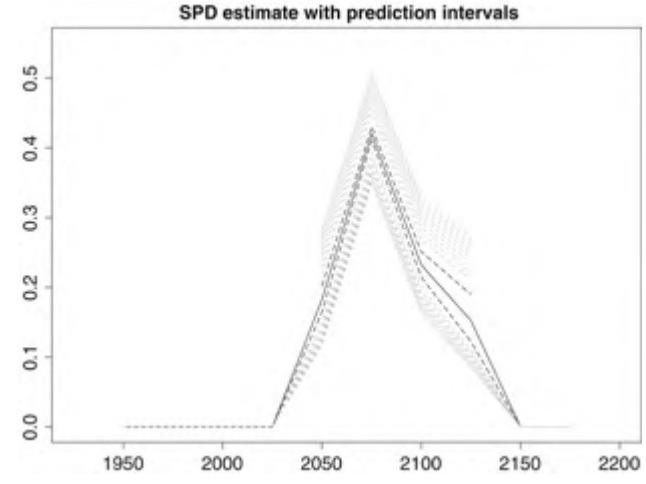


Fig. 11. SPD estimate on 17 January 1995 with prediction intervals for the next 5 h calculated for every 30 min.

The standardized residuals in the top panel of Fig. 10 were plotted using the same technique as the residuals in Fig. 8. Whereas the residuals for the iid model showed strong correlations and heteroscedasticity, the structure of the standardized residuals looks much better. It is natural that the residuals are larger in the central part since more than 90% of observations have strike price

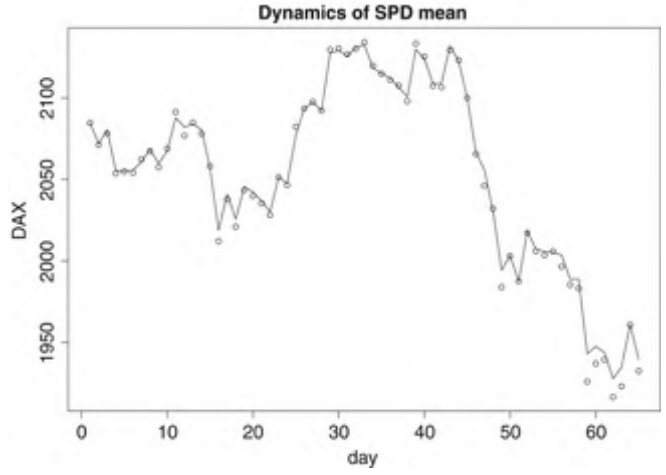


Fig. 12. Daily development of the expected value of the uncorrected SPD from January to March 1995. The circles denote the corresponding closing value of the DAX.

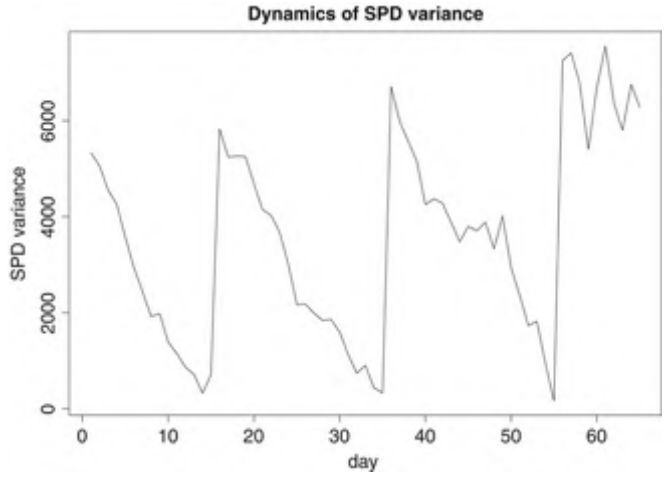


Fig. 13. Daily development of the SPD variance from January to March 1995.

between 2050 and 2100. The largest residuals were omitted in the residual plot so that the structure in the central part is more visible, but the lower panel of Fig. 10 displays the histogram of all residuals. The distribution of the residuals seems to be symmetric, and its shape is not too far from Normal distribution. However, the kurtosis of this distribution is too large, and formal tests reject normality.

In Fig. 11, we plot prediction intervals for the SPD obtained only by recalculating the covariance structure (28) with respect to some future time. More precisely, the prediction intervals are obtained from option prices observed until $\underset{\sim}{i}$. Then, using the notation of Section 4.4, we have, for the future $\widetilde{\beta}_{i+1}$ and $\widetilde{\alpha}_{i+1}$,

$$C_i(k_j) = \Delta_j \widetilde{\beta}_i + \eta_i,$$
$$P_i(k_j) = \Delta_j^p \widetilde{\alpha}_i + \eta_i,$$
$$\begin{pmatrix} \widetilde{\beta}_{i+1} \\ \widetilde{\alpha}_{i+1} \end{pmatrix} = \begin{pmatrix} \widetilde{\beta}_i \\ \widetilde{\alpha}_i \end{pmatrix} + \delta_{i+1}^{1/2} \varepsilon_{i+1}. \tag{31}$$

It is now easy to see that the only modification that has to be done for estimating $\widetilde{\beta}_{i+1}$ is to add the length of the forecasting horizon δ_{i+1} to the sum in (23), (26) and (27), and to recalculate the confidence regions using this variance matrix with the same estimates of the variance parameters σ^2 and ν^2. In Fig. 11, the 95% confidence intervals for the true SPD are denoted by the black dashed line. The grey dashed lines denote the prediction intervals calculated for each 30 min for the next 5 h. In this way, we can obtain a simple approximation for future short-term fluctuations of the SPD. In the long run, the prediction intervals become too wide to be informative.

6. Dynamics of the SPD

In order to study the dynamics of SPDs, we calculated the basic moment characteristics of the estimated SPDs. Note that the estimator does not allow one to estimate the SPD in the tails of the distribution. We can only estimate the probability mass lying to the left $(1 - \sum_{i=1}^{p-1} \beta_i)$ and to the right (β_1) of the available strike price range. Hence, the moments calculated in this section are only approximations which cannot be calculated more precisely without additional assumptions, for example, on the tail behavior or parametric shape of the SPD.

The estimated mean and variance in the first quarter of 1995 are plotted as lines in Figs. 12–13. Note that the SPDs in this period were always estimated using the options with shortest time to maturity. This means that the time to maturity is decreasing linearly in both plots, but it jumps up whenever the option with the shortest time to maturity expires. These jumps occurred at days 16, 36, and 56.

From no-arbitrage considerations, it follows that the mean of the SPD should correspond to the value of the DAX,

$$\widehat{E^{\mathrm{SPD}}} = \int S_T f(S_T) \mathrm{d}S_T = \exp\{r(T-t)\} S_t.$$

See also the discussion in Section 3.6. In Fig. 12, the observed values of the DAX multiplied by the factor $\exp\{r(T-t)\}$ are plotted as circles for the first 65 trading days in 1995, and we observe that the estimated means of the SPD estimates, displayed as the line, follow the theoretical value very closely. A small difference is mainly due to the fact that, in 1995, the observed strike prices do not entirely cover the support of the SPD. For example, on day 16, the difference between the SPD mean (2018.7) and the DAX multiplied by the discount factor (2012.1) is equal to 6.6. The fact that there are not any trades for strike prices smaller than 1925 means that we only know that the probability mass lying to the left from 1950 is equal to 0.25. In the calculation of the estimate of the SPD mean plotted in Fig. 12, this probability mass is assigned to the value 1925, as this is the leftmost observed strike price. Obviously, assigning this probability mass rather to the value $1925 - (6.6/0.25) = 1898.6$ leads a more realistic estimate of the SPD and to the equality of the SPD mean and the discounted DAX.

In Fig. 13, we see that the variance of the SPD decreases linearly as the option moves closer to its maturity. This observation suggests that SPD estimates calculated for neighboring maturities can be linearly interpolated in order to obtain an SPD estimate with arbitrary time to maturity. Such an estimate is important for making the SPD estimates comparable and for studying the development of the market expectations.

6.1. Estimate with the fixed time to expiry

The variances displayed in Fig. 13 suggest that the variance of the SPD estimates changes approximately linearly in time when moving closer to the date of expiry.

Hence, from the estimates $f_{\tau_1}(.)$ and $f_{\tau_2}(.)$ of centered SPDs corresponding to the times of expiry $\tau_1 < \tau_2$, we construct an estimate $f_\tau(.)$ for any time of expiry $\tau \in (\tau_1, \tau_2)$ as

$$f_\tau(.) = \frac{(\tau_2 - \tau) f_{\tau_1}(.) + (\tau - \tau_1) f_{\tau_2}(.)}{\tau_2 - \tau_1}. \tag{32}$$

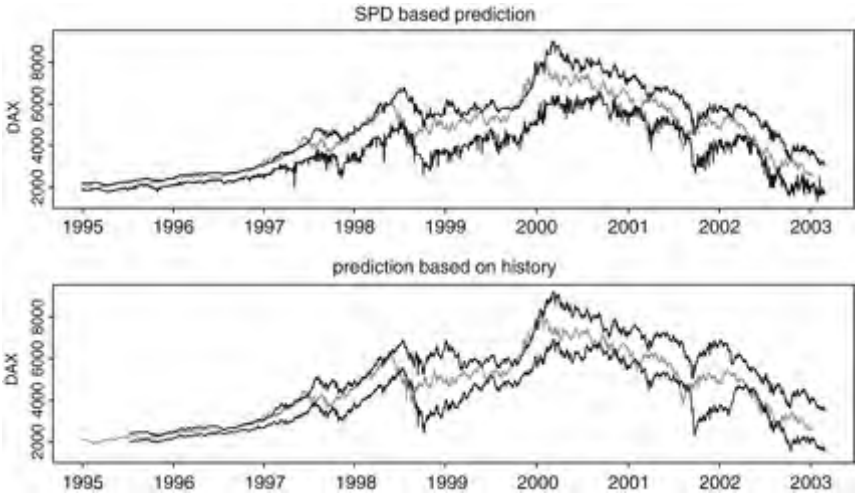


Fig. 14. Prediction intervals for the DAX based on SPDs and historical simulation from January 1995 to March 2003.

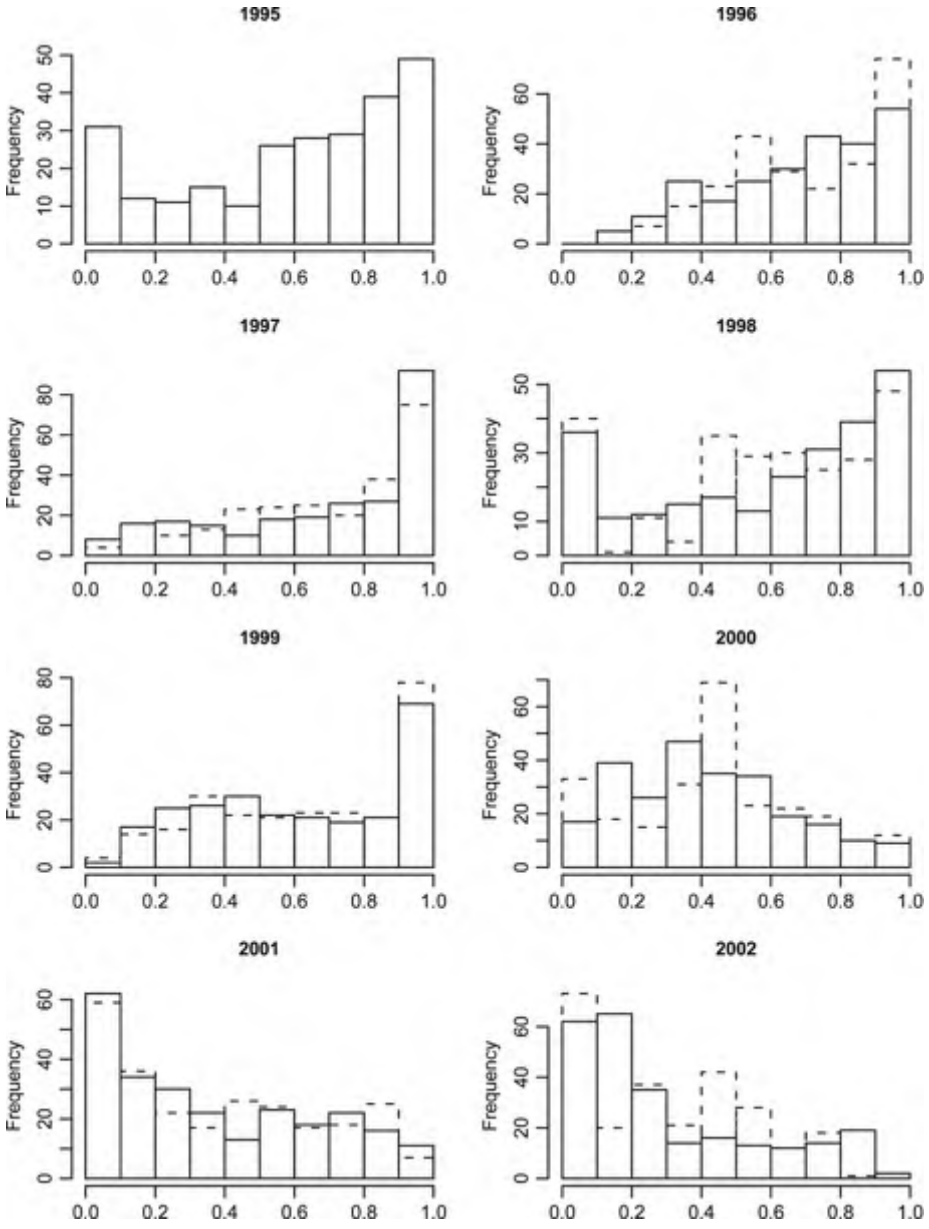


Fig. 15. Histograms for the SPDs (full line) and historical simulation (dashed line).

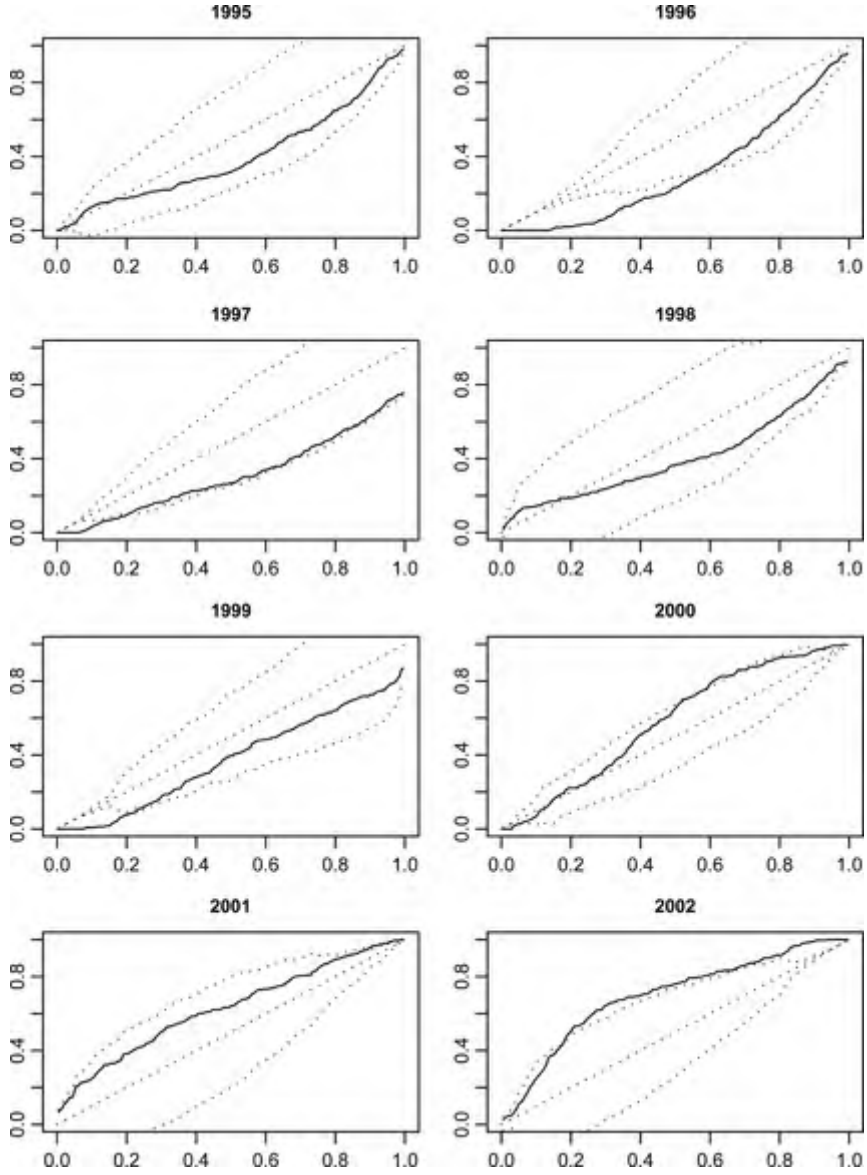


Fig. 16. Integral transformation for estimated SPDs.

In this way, the variance, V_τ, of the centered SPD with time to expiry equal to τ can be expressed as

$$\begin{aligned} V_\tau &= \int x^2 f_\tau(x)\mathrm{d}x \\ &= \int x^2 \frac{(\tau_2-\tau)f_{\tau_1}(x)+(\tau-\tau_1)f_{\tau_2}(x)}{\tau_2-\tau_1}\mathrm{d}x \\ &= \frac{(\tau_2-\tau)V_{\tau_1}+(\tau-\tau_1)V_{\tau_2}}{\tau_2-\tau_1}. \end{aligned}$$

We argue that such an estimate is reasonable since we observed in Fig. 13 that the SPD variances change linearly in time.

6.2. Verification of the market's expectations

Under the risk neutral (equivalent martingale) measure, the SPD reflects the market's expectation of the behavior of the value of the DAX in 45 days. Hence, it is interesting to use our data set to verify how these expectations compare with reality. In the left plot in Fig. 14, we plot intervals based on the SPD together with the true future value of the DAX: the black lines display the 2.5% and 97.5% quantiles of the estimated SPD; the future value of the DAX is displayed as a grey line. In the right plot, we show in the same way the 45-day ahead predictions based on the historical distribution of the 45-day absolute returns in the last 100 trading days; the 2.5% and 97.5% quantiles of this distribution are plotted as black lines.

Fig. 14 suggests that the method works well and that the DAX mostly stays well within the quantiles calculated from the estimated SPDs. The DAX was sometimes rising faster than the market expected from 1995 to mid-1998. After a fast decrease in the second half of 1998, the market increased again till the beginning of year 2000. Since then, the market has decreased. However, the changes stay mostly within or very close to the bounds predicted by our SPD estimates. The only exception is the large shock observed in September 2001, caused by the terrorist attack on the World Trade Center.

The upper quantiles, 97.5%, of the historical distribution of the 45-day absolute returns mostly agree with the upper quantiles of the SPD. The lower quantiles, 2.5%, of the SPDs seem to be much more variable than the same quantiles of the historical distribution. Both the lower and the upper quantiles of the historical distribution lie mostly above the corresponding quantiles of the estimated SPD, respectively in 69.44% and 81.75%.

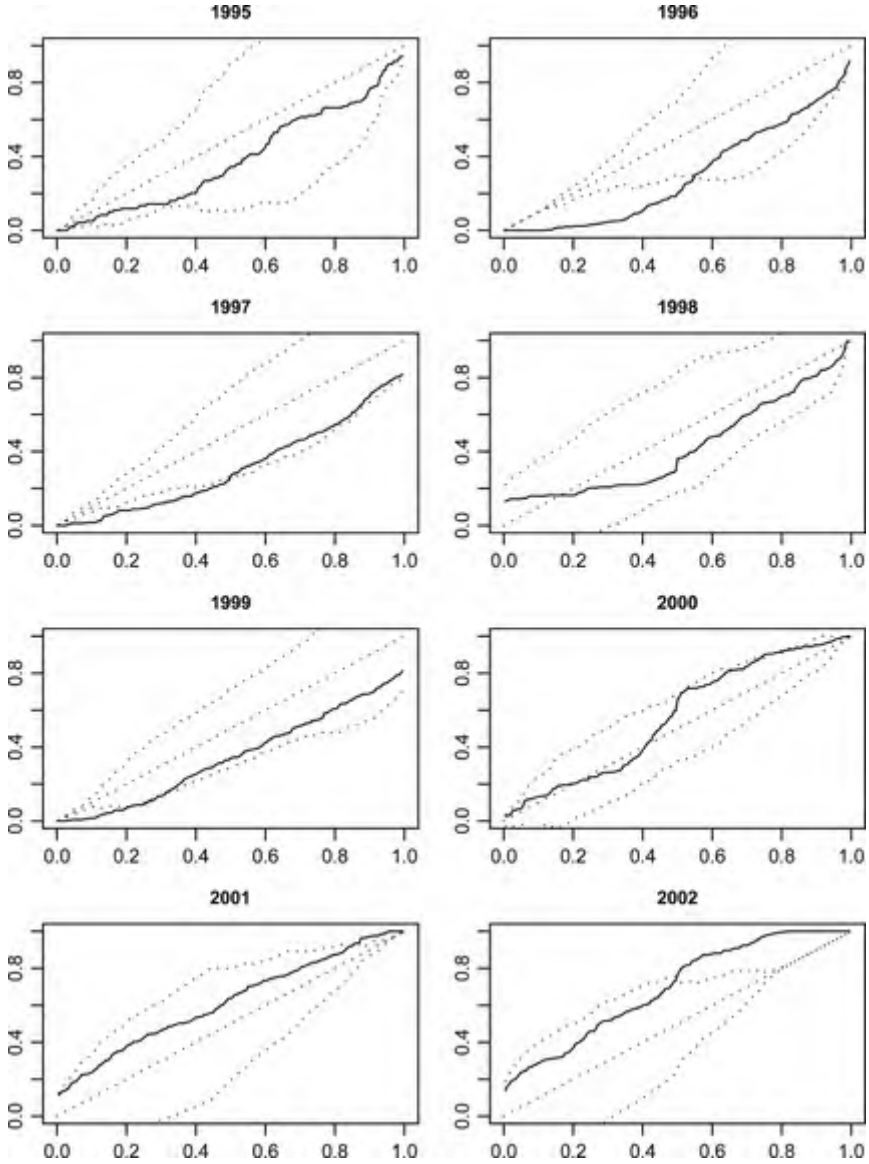


Fig. 17. Integral transformation for historical simulation.

Table 2
Fraction of the year that the DAX stays in the prediction corridor.

Year	1996	1997	1998	1999	2000	2001	2002
SPD (%)	84.40	66.13	75.30	74.60	97.22	85.66	94.84
Historical (%)	82.00	79.44	76.89	77.38	93.25	86.06	80.56

This observation just confirms the fact that the observed SPD includes effects of risk aversion.

In Table 2, we show the fraction of the year that the DAX stays in the prediction corridor. This suggests that the coverage is slightly better for the historical simulation if the DAX is increasing and better for the SPD based prediction if the DAX is decreasing (years 2000 and 2002).

6.3. Evaluation of the quality of the forecasts

The quality of the forecasts can be evaluated by comparing the true future observation with its predicted distribution (the SPD). Diebold et al. (1998) propose to evaluate density forecasts using the probability integral transformed observations $z_{h,t}$, where t denotes the time and h the forecasting horizon. More precisely, we define

$$z_{h,t} = \int_{-\infty}^{X_{t+h}} \widehat{f}_{h,t}(u)\mathrm{d}u,$$

where $\widehat{f}_{h,t}(.)$ denotes our estimate of the SPD h days ahead at time t and X_{t+h} is the future observation. In other words, $z_{h,t}$ is the probability value of X_{t+h} with respect to $\widehat{f}_{h,t}(.)$. Clearly, the $z_{h,t}$ should be uniformly $U(0, 1)$ distributed if the estimated SPD $\widehat{f}_{h,t}(.)$ is equal to the true density of X_{t+h}. In Fig. 15, we display the histograms of $z_{h,t}$'s for each year for the estimated SPDs and historical simulation using full and dashed histograms, respectively. Clearly, in the ideal case, the histograms should not be too far from a Uniform $U(0, 1)$ distribution. In our data, for the prediction horizon $h = 45$ days, we observe that the histograms look quite different from what we would expect. Especially in

years 1995–1999, the DAX was moving mainly in the upper quantiles of the predicted SPD. The forecasts based on the historical distribution of the 45-day returns behave similarly.

In order to account for the overlapping forecasting periods, we calculate the confidence limits for the empirical distribution function

$$\widehat{F}(u) = \frac{1}{T}\sum_{t=1}^{T} \mathbf{I}(z_{h,t} \leq u)$$

of $z_{h,t}$'s that take into account the autocorrelation structure.

$$\widehat{\mathrm{Var}}\{\widehat{F}(u)\} = \frac{1}{T}\left\{\widehat{\gamma}_u(0) + 2\sum_{j=1}^{h}\left(1 - \frac{j}{T}\right)\widehat{\gamma}_u(j)\right\}, \tag{33}$$

where $\gamma_u(j)$ is the sample autocovariance of order j:

$$\gamma_u(j) = \frac{1}{T}\sum_{t=j+1}^{T}\left\{\mathbf{I}(z_{h,t} \leq u) - \widehat{F}(u)\right\}\left\{\mathbf{I}(z_{h,t-j} \leq u) - \widehat{F}(u)\right\}.$$

The empirical distribution functions $\widehat{F}(.)$ are plotted separately for years 1995–2002 in Fig. 16. The distribution function of $U(0, 1)$ and the limits following from (33) are displayed as dotted lines. The year 2003 was not included since our dataset contains only two months of the year 2003, which did not leave enough observations to confirm the forecasts.

In 1996 and 1997, the market was growing much faster than the SPDs were indicating. In 1996, it never happened that the DAX fell below the 10% quantile of the SPD, and there were only a few days when this value was below 20%. The situation in 1998 and 1999 was less extreme even though the fast growth of the DAX continued. The distribution given by the SPD estimate $\widehat{f}_{t,h}(.)$ for the horizon $h = 45$ days does not differ significantly from the true distribution of X_{t+h} in 2000–2001, but in 2002 we again observe significant differences. Thus, the DAX was growing faster than the option market expected in 1996, 1997, and 1999 and it was falling faster in 2002.

Fig. 17 shows the same graphics for the forecast based on the historical distribution of the returns. The deviations are more clearly visible but the overall picture is very similar; the only difference arises in 2001 when the predictions did not stay between the limits.

7. Conclusion

We have proposed a simple nonparametric model for arbitrage-free estimation of the SPD. Our procedure takes care of the daily changing covariance structure and involves both types of European option. Moreover, the covariance structure allows us to calculate prediction intervals capturing future behavior of the SPD. We analyze the moment dynamics of the SPD from 1995–2003. An application to DAX EUREX data for the years 1995–2003 produces a corridor that is compared to the future DAX index value. The proposed technique enables us not only to price exotic options but also to measure the risk and volatility ahead of us.

Acknowledgments

We thank Volker Krätschmer for useful comments concerning the existence and uniqueness of the constrained regression function and the anonymous referee for many insightful comments leading to substantial improvements in both the presentation and the content of the paper. The research was supported by Deutsche Forschungsgemeinschaft, SFB 649 "Ökonomisches Risiko", by MSM0021620839, GAČR GA201/08/0486, and by MŠMT 1K04018.

References

Aït-Sahalia, Y., Duarte, J., 2003. Nonparametric option pricing under shape restrictions. Journal of Econometrics 116, 9–47.
Aït-Sahalia, Y., Lo, A.W., 1998. Nonparametric estimation of state-price densities implicit in financial asset prices. Journal of Finance 53, 499–547.
Aït-Sahalia, Y., Lo, A.W., 2000. Nonparametric risk management and implied risk aversion. Journal of Econometrics 94, 9–51.
Aït-Sahalia, Y., Wang, Y., Yared, F., 2000. Do option markets correctly price the probabilities of movement of the underlying asset? Journal of Econometrics 102, 67–110.
Anděl, J., 1985. Mathematical Statistics. SNTL/Alfa, Prague (in Czech).
Andersen, L.B.G., Brotherton-Ratcliffe, R., 1997. The equity option volatility smile: An implicit finite-difference approach. Journal of Computational Finance 1 (2), 5–37.
Bondarenko, O., 2003. Estimation of risk-neutral densities using positive convolution approximation. Journal of Econometrics 116, 85–112.
Breeden, D., Litzenberger, R., 1978. Prices of state-contingent claims implicit in option prices. Journal of Business 51, 621–651.
Buehler, H., 2006. Expensive martingales. Quantitative Finance 6 (3), 207–218.
Diebold, F.X., Gunther, T., Tay, A., 1998. Evaluating density forecasts, with applications to financial risk management. International Economic Review 39, 863–883.
Dupire, B., 1994. Pricing with a smile. RISK 7 (1), 18–20.
Fengler, M.R., 2005. Semiparametric Modeling of Implied Volatility. Springer, Heidelberg.
Fengler, M.R., Härdle, W., Mammen, E., 2007. A dynamic semiparametric factor model for implied volatility string dynamics. Journal of Financial Econometrics 5 (2), 189–218.
Hafner, R., Wallmeier, M., 2000. The Dynamics of DAX Implied Volatilities. University of Augsburg Working Paper. Available at SSRN: http://ssrn.com/abstract=234829 or doi:10.2139/ssrn.234829.
Harrison, J., Pliska, S., 1981. Martingale and stochastic integral in the theory of continuous trading. Stochastic Processes and their Applications 11, 215–260.
Hlávka, Z., Svojík, M., 2008. Application of extended Kalman filter to SPD estimation. In: Härdle, W., Hautsch, N., Overbeck, L. (Eds.), Applied Quantitative Finance. Springer, Berlin, pp. 233–247.
Huynh, K., Kervella, P., Zheng, J., 2002. Estimating state-price densities with nonparametric regression. In: Härdle, W., Kleinow, T., Stahl, G. (Eds.), Applied Quantitative Finance. Springer, Heidelberg, pp. 171–196.
Jackwerth, J.C., 1999. Option-implied risk-neutral distributions and implied binomial trees: A literature review. Journal of Derivatives 7, 66–82.
Kahalé, N., 2004. An arbitrage-free interpolation of volatilities. RISK 17 (5), 102–106.
Rao, C.R., 1973. Linear Statistical Inference and Its Applications. Wiley, New York.
Renault, E., 1997. Econometric models of option pricing errors. In: Kreps, D.M., Wallis, K.F. (Eds.), Advances in Economics and Econometrics: Theory and Applications, Seventh World Congress, vol. III. Cambridge University Press, Cambridge, pp. 223–278.
Robertson, T., Wright, F.T., Dykstra, R.L., 1988. Order Restricted Statistical Inference. Wiley, Chichester.
Seber, G.A.F., Wild, C.J., 2003. Nonlinear Regression. Wiley, Hoboken, New Jersey.
Serfling, R., 1980. Approximation Theorems of Mathematical Statistics. Wiley, New York.
Stoll, H.R., 1969. The relationship between put and call option prices. Journal of Finance 24, 801–824.
Yatchew, A., Härdle, W., 2006. Nonparametric state price density estimation using constrained least squares and the bootstrap. Journal of Econometrics 133 (2), 579–599.

Journal of Forecasting
J. Forecast. **28**, 512–534 (2009)
Published online 2 December 2008 in Wiley InterScience
(www.interscience.wiley.com) **DOI**: 10.1002/for.1109

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.

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 retransforming 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 mathematical properties such as strong convexity and desirable infinite 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

DOI: 10.1002/for

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

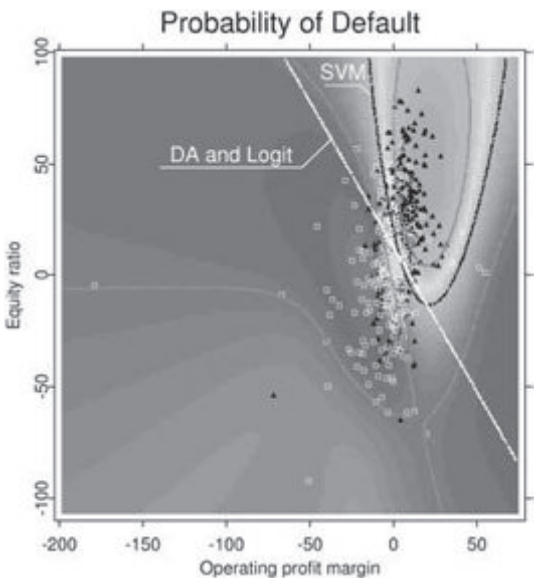


Figure 1. SVM-separating function (black) with margin in a two-dimensional space

DOI: 10.1002/for

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)

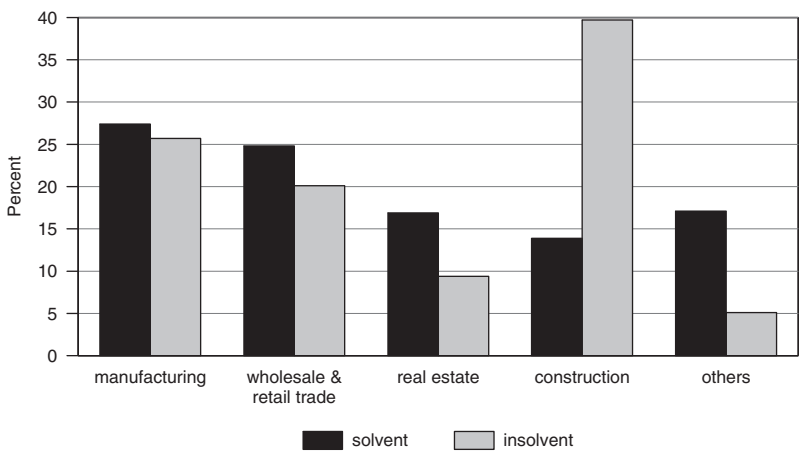


Figure 2. The distribution of solvent and insolvent companies across industries

DOI: 10.1002/for

- 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 (%);

DOI: 10.1002/for

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 Class	Positive	*True positive* (*TP*)	*False negative* (*FN*)
	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

DOI: 10.1002/for

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:

$$\begin{aligned} \min_{(w,b,\xi)\in\mathbb{R}^{d+1+n}} \; & C\sum\nolimits_{i=1}^{b}\xi_i + \frac{1}{2}\|w\|_2^2 \\ \text{s.t. } & y_i\left(w^\top \mathbf{x}_i + b\right) + \xi_i \geq 1 \\ & \xi_i \geq 0, \quad \text{for } i = 1, 2, \ldots, n \end{aligned} \tag{1}$$

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:

$$\begin{aligned} \min_{(w,b,\xi)\in\mathbb{R}^{d+1+n}} \; & \frac{C}{2}\sum\nolimits_{i=1}^{n}\xi_i^2 + \frac{1}{2}\left(\|w\|_2^2 + b^2\right) \\ \text{s.t. } & y_i\left(w^\top \mathbf{x}_i + b\right) + \xi_i \geq 1 \\ & \xi_i \geq 0, \quad \text{for } i = 1, 2, \ldots, n \end{aligned} \tag{2}$$

DOI: 10.1002/for

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:

$$\min_{(w,b)\in\mathbb{R}^{d+1}} \frac{C}{2}\sum_{i=1}^{n}\{1-y_i(w^\top \mathbf{x}_i + b)\}_+^2 + \frac{1}{2}\left(\|w\|_2^2 + b^2\right) \tag{3}$$

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:

$$\min_{(w,b)\in\mathbb{R}^{d+1}} \frac{C}{2}\sum_{i=1}^{n} p(\{1-y_i(w^\top \mathbf{x}_i + b)\}, \alpha)^2 + \frac{1}{2}\left(\|w\|_2^2 + b^2\right) \tag{4}$$

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:

$$\min_{(u,b)\in\mathbb{R}^{n+1}} \frac{C}{2}\sum_{i=1}^{n} p\left(\left[1-y_i\left\{\sum_{j=1}^{n} u_j K(\mathbf{x}_i, \mathbf{x}_j) + b\right\}\right], \alpha\right)^2 + \frac{1}{2}\left(\|u\|_2^2 + b^2\right) \tag{5}$$

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:

$$K(A, A^\top) \approx K(A, \tilde{A}^\top)K(\tilde{A}, \tilde{A}^\top)^{-1}K(\tilde{A}, A^\top) \tag{7}$$

DOI: 10.1002/for

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:

$$\begin{aligned} \min_{(w,b,\xi)\in\mathbb{R}^{d+1+n}} \; & C\sum_{i=1}^{n}\xi_i + \|w\|_1 \\ \text{s.t. } & y_i(w^\top \mathbf{x}_i + b) + \xi_i \geq 1 \\ & \xi_i \geq 0, \quad \text{for } i = 1, 2, \ldots, n. \end{aligned} \tag{9}$$

The objective function of (9) is a piecewise linear convex function. We can reformulate it as the following linear programming problem:

$$\begin{aligned} \min_{(w,s,b,\xi)\in\mathbb{R}^{d+d+1+n}} \; & C\sum_{i=1}^{n}\xi_i + \sum_{j=1}^{d}s_j \\ \text{s.t. } & y_i(w^\top \mathbf{x}_i + b) + \xi_i \geq 1 \\ & -s_j \leq w_j \leq s_j, \quad \text{for } j = 1, 2, \ldots, d, \\ & \xi_i \geq 0, \quad \text{for } i = 1, 2, \ldots, n \end{aligned} \tag{10}$$

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

DOI: 10.1002/for

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		

DOI: 10.1002/for

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

DOI: 10.1002/for

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 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 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

DOI: 10.1002/for

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

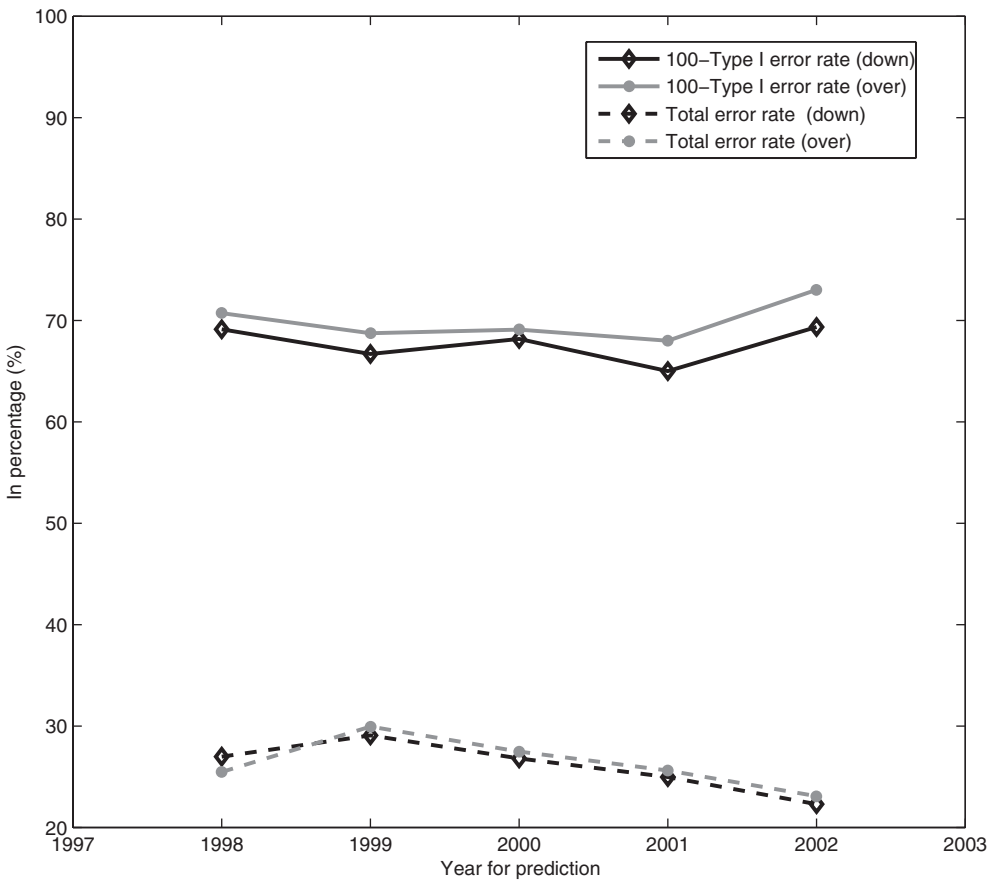


Figure 3. Learning curve for variables set V3

DOI: 10.1002/for

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 nonmonotonic. 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–test pair, we carry out oversampling for positive instances from 6 to 15 times. We show the trend and effect in Figure 4. It is

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*

DOI: 10.1002/for

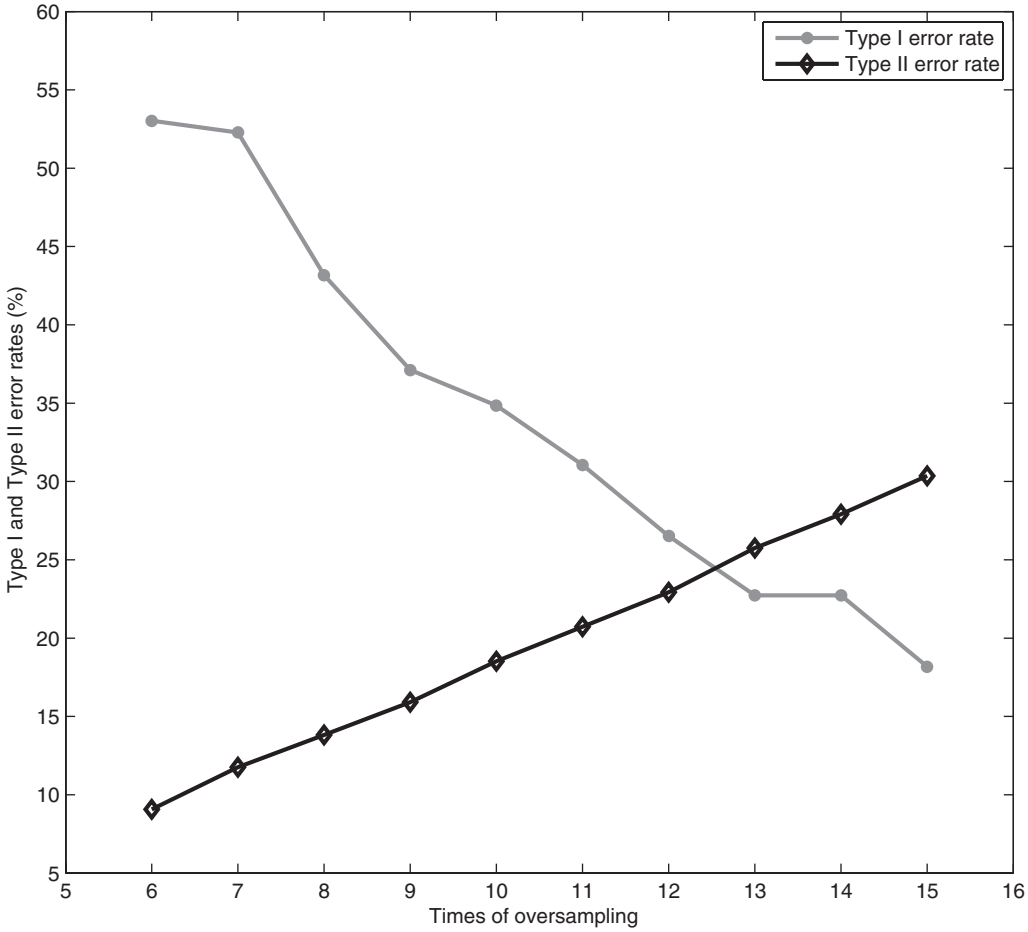


Figure 4. The effect of oversampling on Type I and Type II error rates for scenario S5 and variables set V3

easy to see that the Type I (II) error rate decreases (increases) as the oversampling times increase. 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.

DOI: 10.1002/for

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	LM	DA
		Mean	Mean	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	LM	DA
		Mean	Mean	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

DOI: 10.1002/for

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. LM				*	*
SSVM vs. DA					*
Downsampling					
SSVM vs. LM			*		*
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 insolvency (solvency). 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

DOI: 10.1002/for

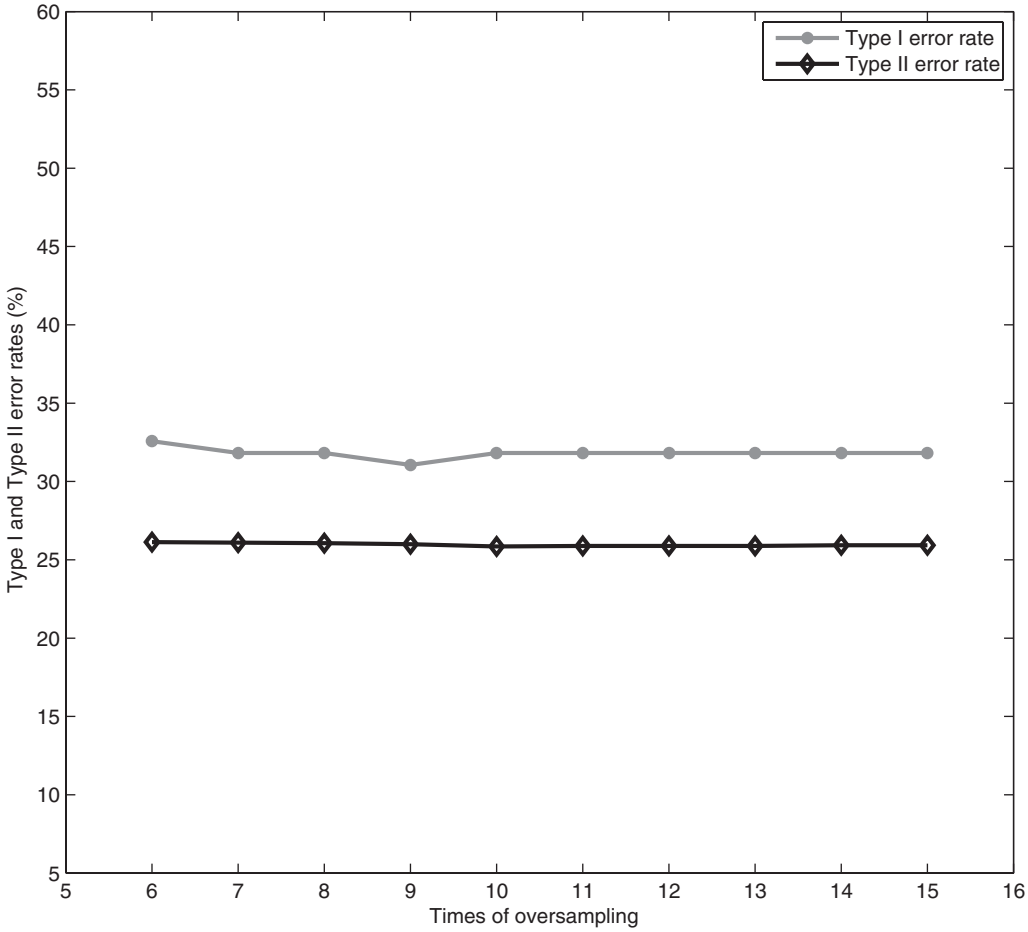


Figure 5. The effect of oversampling on Type I and Type II error rates for scenario S5 and variables set V3 in DA

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.

DOI: 10.1002/for

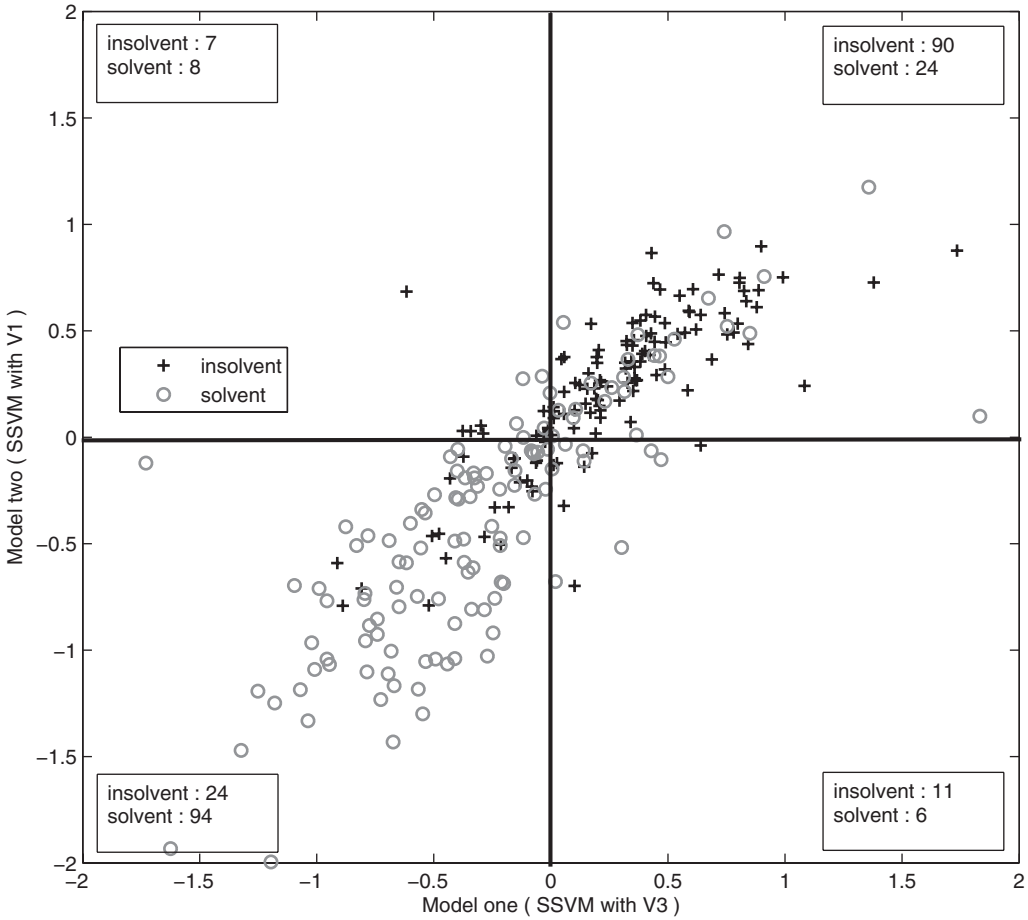


Figure 6. Data visualization via model one (generated by SSVM with V3) and model two (generated by SSVM with V1) in scenario S5

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

DOI: 10.1002/for

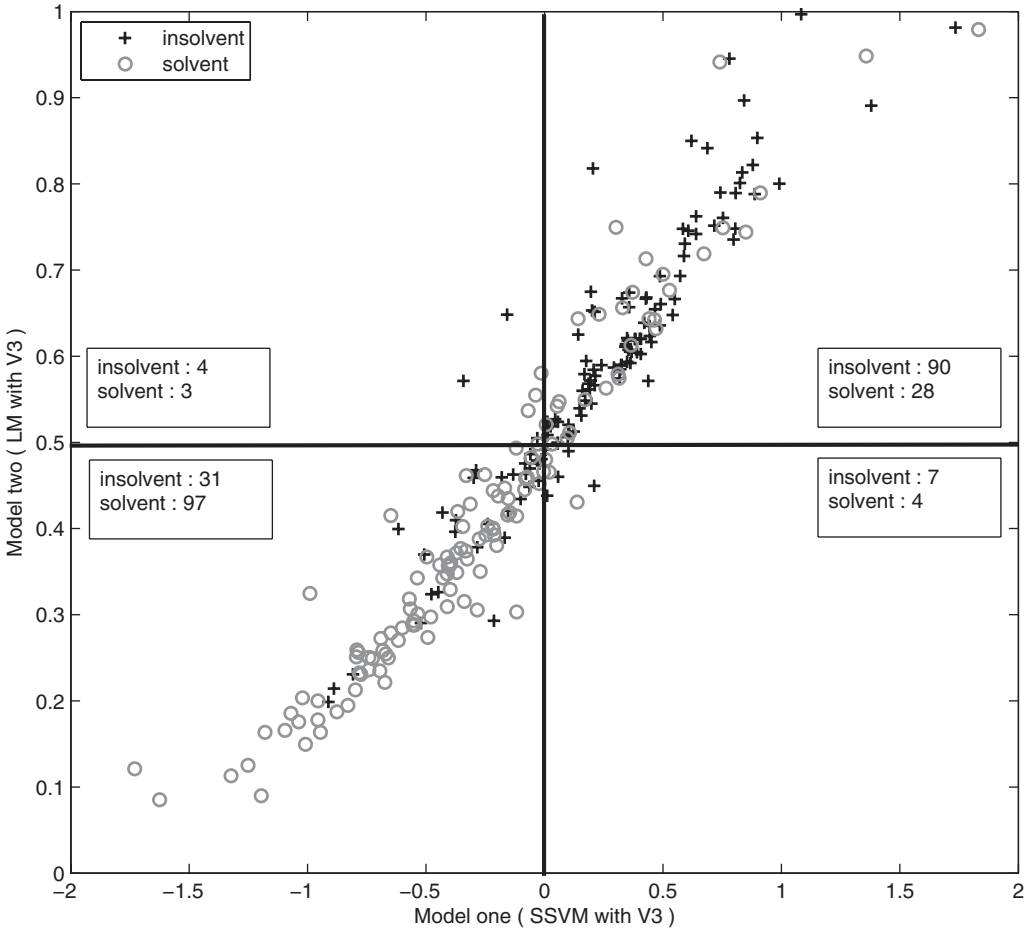


Figure 7. Data visualization via model one (generated by SSVM with V3) and model two (generated by LM with V3) in scenario S5

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.

DOI: 10.1002/for

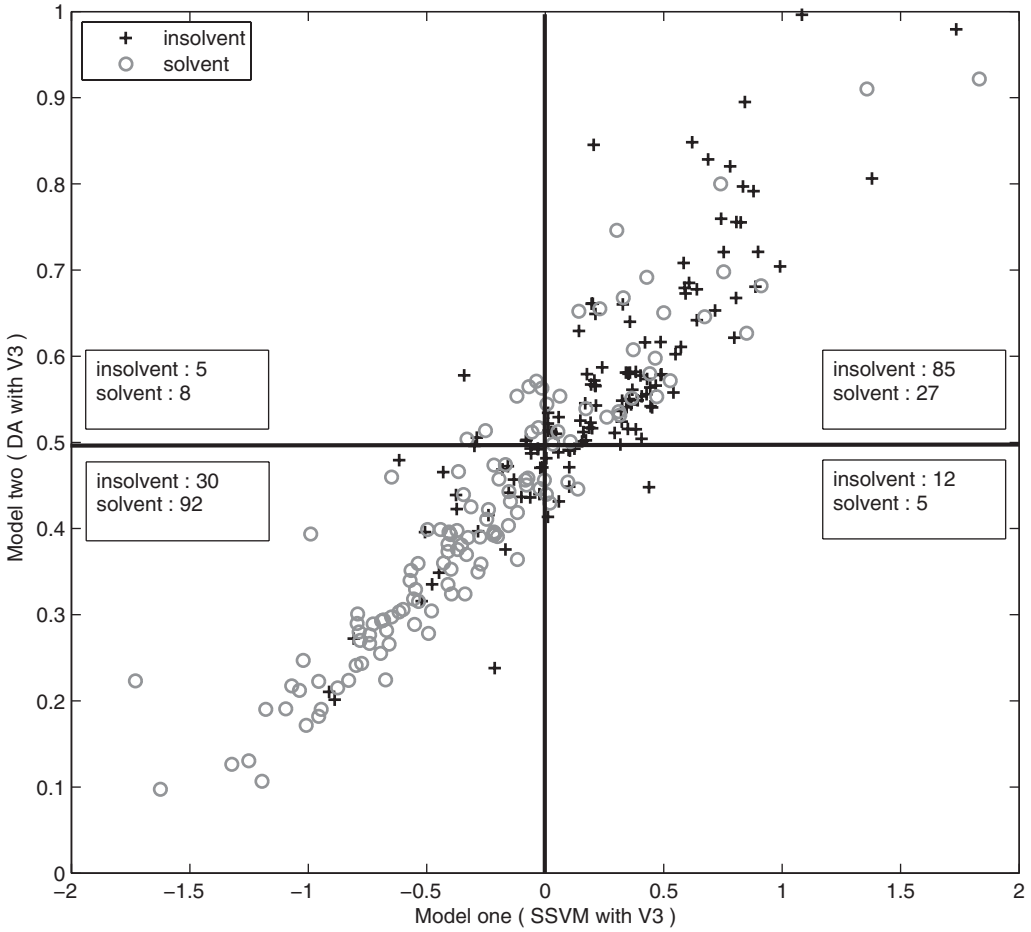


Figure 8, Data visualization via model one (generated by SSVM with V3) and model two (generated by DA with V3) in scenario S5

ACKNOWLEDGEMENTS

This research was supported by the 'Stiftung Geld und Währung' and by the Deutsche Forschungsgemeinschaft through the SFB 649 'Economic Risk'.

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.

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; 155–161.
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, Härdle W, Unwin A (eds). Springer: Heidelberg; 853–872.
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* **52**: 335–346. 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* **18**(5): 824–840.
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.

DOI: 10.1002/for

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* **16**: 49–56.

Authors' biographies:
Wolfgang Härdle did in 1982 his Dr. rer. nat. in Mathematics at Universität Heidelberg and in 1988 his Habilitation at Universität Bonn. He is currently chair professor of statistics at the Dept. of Economics and Business Administration, Humboldt-Universität zu Berlin. He is also director of CASE—Center for Applied Statistics & Economics and of the Collaborative Research Center 'Economic Risk'. His research focuses on dimension reduction techniques, computational statistics and quantitative finance. He has published 34 books and more than 200 papers in top statistical, econometrics and finance journals. He is one of the 'Highly cited Scientist' according to the Institute of Scientific Information.

Yuh-Jye Lee received his Master degree in Applied Mathematics from the National Tsing Hua University, Taiwan in 1992 and PhD degree in computer sciences from the University of Wisconsin-Madison in 2001. In 2002, Dr. Lee joined the Computer Science and Information Engineering Department, National Taiwan University of Science and Technology. He is an associate professor now. His research interests are in machine learning, data mining, optimization, information security and operations research. He developed new algorithms for large data mining problems such as classification problem, clustering, feature selection and dimension reduction. These algorithms have been used in intrusion detection systems (IDS), face detection, micro array gene expression analysis and breast cancer diagnosis and prognosis.

Dorothea Schäfer did in 1992 her Dr. rer. pol. in Economics and in the year 2000 her Habilitation at Freie Universität Berlin. She is currently coordinator of the research group Financial Markets and Financial Institutions and senior researcher at the German Institute for Economic Research (DIW) Berlin which she joined in 2002. She is managing editor of the Quarterly Journal of Economic Research (Vierteljahreshefte zur Wirtschaftsforschung) and adjunct lecturer at Freie Universität Berlin. Her research focuses on insolvency risk, financial management of firms and banks, and on behavioural finance.

Yi-Ren Yeh received the M.S. degree from the Department of Computer Science and Information Engineering, National Taiwan University of Science and Technology, Taipei, Taiwan, R.O.C., in 2006. He is currently working toward the PhD degree in the Department of Computer Science and Information Engineering, National Taiwan University of Science and Technology. His research interests include machine learning, data mining, and information security.

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.

DOI: 10.1002/for

Time Series Modelling With Semiparametric Factor Dynamics

Byeong U. PARK, Enno MAMMEN, Wolfgang HÄRDLE, and Szymon BORAK

High-dimensional regression problems, which reveal dynamic behavior, are typically analyzed by time propagation of a few number of factors. The inference on the whole system is then based on the low-dimensional time series analysis. Such high-dimensional problems occur frequently in many different fields of science. In this article we address the problem of inference when the factors and factor loadings are estimated by semiparametric methods. This more flexible modeling approach poses an important question: Is it justified, from an inferential point of view, to base statistical inference on the estimated times series factors? We show that the difference of the inference based on the estimated time series and "true" unobserved time series is asymptotically negligible. Our results justify fitting vector autoregressive processes to the estimated factors, which allows one to study the dynamics of the whole high-dimensional system with a low-dimensional representation. We illustrate the theory with a simulation study. Also, we apply the method to a study of the dynamic behavior of implied volatilities and to the analysis of functional magnetic resonance imaging (fMRI) data.

KEY WORDS: Asymptotic inference; Factor models; Implied volatility surface; Semiparametric models; Vector autoregressive process.

1 INTRODUCTION

Modeling for high-dimensional data is a challenging task in statistics especially when the data comes in a dynamic context and is observed at changing locations with different sample sizes. Such modeling challenges appear in many different fields. Examples are Stock and Watson (2005) in empirical macroeconomics, Lee and Carter (1992) in mortality analysis, Nelson and Siegel (1987) and Diebold and Li (2006) in bond portfolio risk management or derivative pricing, Martinussen and Scheike (2000) in biomedical research. Other examples include the studies of radiation treatment of prostate cancer by Kauermann (2000) and evoked potentials in Electroencephalogram (EEG) analysis by Gasser, Möcks, and Verleger (1983). In financial engineering, it is common to analyze the dynamics of implied volatility surface for risk management. For functional magnetic resonance imaging data (fMRI), one may be interested in analyzing the brain's response over time as well as identifying its activation area, see Worsley et al. (2002).

A successful modeling approach utilizes factor type models, which allow low-dimensional representation of the data. In an orthogonal L-factor model an observable J-dimensional random vector $Y_t = (Y_{t,1}, \ldots, Y_{t,J})^{\mathrm{T}}$ can be represented as

$$Y_{t,j} = m_{0,j} + Z_{t,1}m_{1,j} + \cdots + Z_{t,L}m_{L,j} + \varepsilon_{t,j}, \qquad (1)$$

where $Z_{t,l}$ are common factors, $\epsilon_{t,j}$ are errors or specific factors, and the coefficients $m_{l,j}$ are factor loadings. In most applications, the index $t = 1, \ldots, T$ reflects the time evolution of the whole system, and Y_t can be considered as a multidimensional time series. For a method to identify common factors in this model we refer to Peña and Box (1987). The study of high-dimensional Y_t is then simplified to the modeling of $Z_t = (Z_{t,1}, \ldots, Z_{t,L})^{\mathrm{T}}$, which is a more feasible task when $L \ll J$. The model (1) reduces to a special case of the generalized dynamic factor model considered by Forni, Hallin, Lippi, and Reichlin (2000), Forni and Lippi (2001) and Hallin and Liska (2007), when $Z_{t,l} = a_{l,1}(B)U_{t,1} + \cdots + a_{l,q}(B)U_{t,q}$ where the q-dimensional vector process $U_t = (U_{t,1}, \ldots, U_{t,q})^{\mathrm{T}}$ is an orthonormal white noise and B stands for the lag operator. In this case, the model (1) is expressed as $Y_{t,j} = m_{0,j} + \sum_{k=1}^{q} b_{k,j}(B)U_{t,k} + \varepsilon_{t,j}$, where $b_{k,j}(B) = \sum_{l=1}^{L} a_{l,k}(B)m_{l,j}$.

In a variety of applications, one has explanatory variables $X_{t,j} \in \mathbb{R}^d$ at hand that may influence the factor loadings m_l. An important refinement of the model (1) is to incorporate the existence of observable covariates $X_{t,j}$. The factor loadings are now generalized to functions of $X_{t,j}$, so that the model (1) is generalized to

$$Y_{t,j} = m_0(X_{t,j}) + \sum_{l=1}^{L} Z_{t,l}m_l(X_{t,j}) + \varepsilon_{t,j}, \ 1 \le j \le J_t. \qquad (2)$$

In this model, $Z_{t,l}$ for each l: $1 \le l \le L$ enters into all $Y_{t,j}$ for j such that $m_l(X_{t,j}) \neq 0$. Note that the probability of the event that $m_l(X_{t,j}) = 0$ for some $1 \le j \le J$ equals zero if $m_1(x) = 0$ at countably many points of x and the density f_t of $X_{t,j}$ is supported on an interval with nonempty interior, as we assume at (A2) in Section 5.

The model (2) can be interpreted as a discrete version of the following functional extension of the model (1):

$$Y_t(x) = m_0(x) + \sum_{l=1}^{L} Z_{t,l}m_l(x) + \varepsilon_t(x), \qquad (3)$$

where $\varepsilon_t(\cdot)$ is a mean zero stochastic process, and also regarded as a regression model with embedded time evolution. It is different from varying-coefficient models, such as in Fan, Yao, and Cai (2003) and Yang, Park, Xue, and Härdle (2006), because Z_t is unobservable. Our model also has some similarities to the one considered in Connor and Linton (2007) and Connor, Hagmann, and Linton (2007), which generalized the study of Fama and French (1992) on the common movements of stock price returns. There, the covariates, denoted by $X_{l,j}$, are

Byeong U. Park is Professor, Department of Statistics, Seoul National University Seoul 151-747, Korea (E-mail: *bupark@stats.snu.ac.kr*). Enno Mammen is Professor, Department of Economics, University of Mannheim, 68131 Mannheim, Germany (E-mail: *emammen@rumms.uni-mannheim.de*). Wolfgang Härdle is Professor, Institute for Statistics and Econometrics, Humboldt Universität zu Berlin, D-10178 Berlin, Germany (E-mail: *haerdle@wiwi.hu-berlin.de*). Szymon Borak is Ph.D. Student, Institute for Statistics and Econometrics, Humboldt Universität zu Berlin, D-10178 Berlin, Germany (E-mail: *szymon.borak@gmail.de*). The authors gratefully acknowledge financial support Deutsche Forschungsgemeinschaft and the Sonderforschungsbereich 649 "Ökonomisches Risiko." Byeong U. Park's research was supported by the Korea Research Foundation Grant funded by the Korean Government (MOEHRD) (KRF-2005-070-C00021). The authors thank the associate editor and referees for their helpful comments and suggestions.

Journal of the American Statistical Association
March 2009, Vol. 104, No. 485, Theory and Methods
DOI 10.1198/jasa.2009.0105

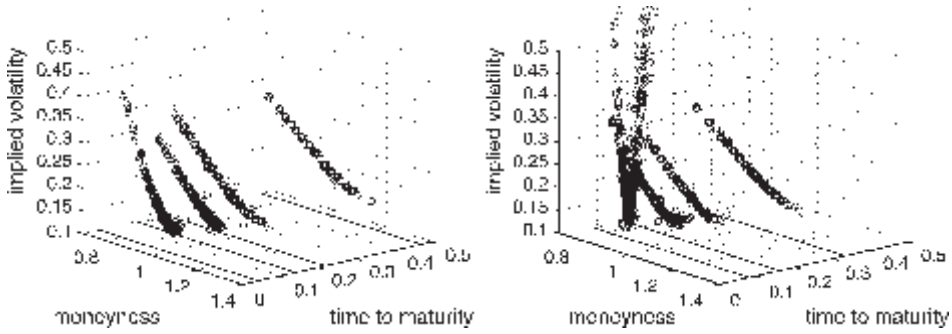


Figure 1. The typical IV data design on two different days. In the maturity direction observations appear in the discrete points for each particular day. Bottom solid lines indicate the observed maturities. Left panel: observations on 2004.07.08, $J_t = 5{,}606$. Right panel: observations on 2004.08.19, $J_t = 8{,}152$.

time-invariant and are different for different m_l, which allows a direct application of backfitting procedures and makes the problem quite different from our setting. Some linear models, which allow time-varying coefficients, as considered in Hansen, Nielsen, and Nielsen (2004) and Brumback and Rice (1998), may be recognized as a special case of (2).

In this article we consider the model (2) with unknown nonparametric functions m_l. We call this model a dynamic semiparametric factor model (DSFM). The evolution of complex high-dimensional objects may be described by (2), so that their analysis can be reduced to the study of a low-dimensional vector of factors Z_t. In the present article, we consider an efficient nonparametric method of fitting the model. We provide relevant theory for the method as well as illustrate its empirical aspects through a simulation and a real data application. Fengler, Härdle, and Mammen (2007) used a kernel smoothing approach for the same model, but it was focused on a particular data application without offering any discussion of numerical issues, statistical theory, and simulation analysis.

One of the main motivations for the model (2) comes from a special structure of the implied volatility (IV) data, as is observed in Figure 1. The IV is a volatility parameter that matches the observed plain vanilla option prices with the theoretical ones given by the formula of Black and Scholes (1973). Figure 1 shows the special "string" structure of the IV data obtained from the European option prices on the German stock index DAX (ODAX) for two different days. The volatility strings shift toward expiry, which is indicated by the bottom line in the figure. Moreover the shape of the IV strings is subject to stochastic deformation. Fengler et al. (2007) proposed to use the model (2) to describe the dynamics of the IV data, where $Y_{t,j}$ are the values of IV or those of its transformation on the day t, and $X_{t,j}$ are the two-dimensional vectors of the moneyness and time-to-maturity. For more details on the data design and econometric motivation, we refer to Fengler et al. (2007).

One may find another application of the model (2) in the analysis of functional magnetic resonance imaging (fMRI) data. The fMRI is a noninvasive technique of recording brain's signals on spatial area in every particular time period (usually 1–4 sec). One obtains a series of three-dimensional images of the blood-oxygen-level-dependent (BOLD) fMRI signals, whereas an exercised person is subject to certain stimuli. An example of the images in 15 different slices at one particular time point is presented in Figure 2. For the more detailed description on the fMRI methodology we refer to Logothetis and Wandell (2004). The main aims of the statistical methods in this field are identification of the brain's activation areas and analysis of its response over time. For this purpose the model (2) can be applied. DSFM may be applied to many other problems, such as modeling of yield curve evolution where the standard approach is to use the parametric factor model proposed by Nelson and Siegel (1987).

Our methods produce estimates of the true unobservable Z_t, say $\hat{Z}_t$, as well as estimates of the unknown functions m_l. In practice, one operates on these estimated values of Z_t for further statistical analysis of the data. In particular, for the IV application, one needs to fit an econometric model to the estimated factors $\hat{Z}_t$. For example, Hafner (2004) and Cont and da Fonseca (2002) fitted an AR(1) process to each factor, and Fengler et al. (2007) considered a multivariate VAR(2) model. The main question that arises from these applications is whether the inference based on $\hat{Z}_t$ is equivalent to the one based on Z_t. Attempting to give an answer to this question forms the core of this article.

It is worthwhile to note here that Z_t is not identifiable in the model (2). There are many versions of (Z_t, m), where $m = (m_0, \ldots, m_L)^{\mathrm{T}}$, that give the same distribution of Y_t. This means that estimates of Z_t and m_l are not uniquely defined. We show that for any version of $\{Z_t\}$ there exists a version of $\{\hat{Z}_t\}$ whose lagged covariances are asymptotically the same as those of $\{Z_t\}$. This justifies the inference based on $\{\hat{Z}_t\}$ when $\{Z_t\}$ is a VAR process, in particular. We confirm this theoretical result by a Monte Carlo simulation study. We also discuss fitting the model to the real ODAX IV and fMRI data.

The article is organized as follows. In the next section we propose a new method of fitting DSFM and an iterative algorithm that converges at a geometric rate. In Section 3 we present the results of a simulation study that illustrate the theoretical findings given in Section 5. In Section 4 we apply the model to the ODAX IV and fMRI data. Section 5 is devoted to the asymptotic analysis of the method. Technical details are provided in the Appendix.

2. METHODOLOGY

We observe $(X_{t,j}, Y_{t,j})$ for $j = 1, \ldots, J_t$ and $t = 1, \ldots, T$ such that

$$Y_{t,j} = \mathcal{Z}_t^{\top} m(X_{t,j}) + \varepsilon_{t,j}. \tag{4}$$

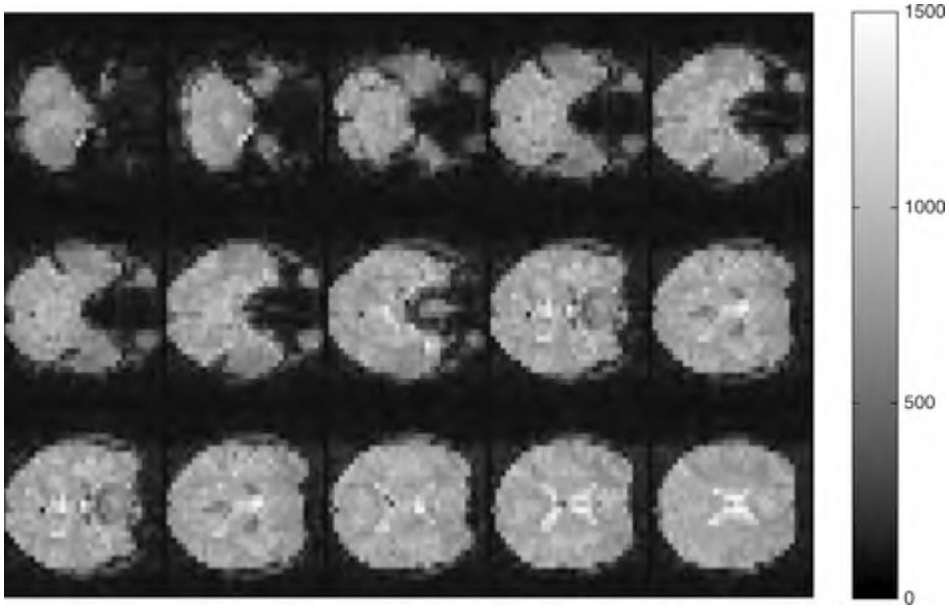


Figure 2. Typical fMRI data in one particular time point. The figure presents 15 parallel horizontal images. The brightness corresponds to the strength of the observed signals.

Here $\mathcal{Z}_t^\top = (1,\ Z_t^\top)$ and $Z_t = (Z_{t,1}, \ldots, Z_{t,L})^\top$ is an unobservable L-dimensional process. The function m is an $(L+1)$-tuple $(m_0, \ldots, m_L)$ of unknown real-valued functions m_l defined on a subset of $\mathbb{R}^d$. The variables $X_{1,1}, \ldots, X_{T,J_T}, \varepsilon_{1,1}, \ldots, \varepsilon_{T,J_T}$ are independent. The errors $\epsilon_{t,j}$ have zero means and finite second moments. For simplicity of notation, we will assume that the covariates $X_{t,j}$ have support $[0, 1]^d$, and also that $J_t \equiv J$ do not depend on t.

For the estimation of m, we use a series estimator. For an integer $K \geq 1$, we choose functions $\psi_1, \ldots, \psi_K$: $[0, 1]^d \to \mathbb{R}$, which are normalized so that $\int_{[0,1]^d} \psi_k^2(x)dx = 1$. For example, one may take $\{\psi_k\colon 1 \leq k \leq K\}$ to be a tensor B-spline basis (e.g., see de Boor 2001). Then, an $(L+1)$-tuple of functions $m = (m_0, \ldots, m_L)^\top$ may be approximated by $\mathcal{A}\psi$, where $\mathcal{A} = (\alpha_{l,k})$ is an $(L+1) \times K$ matrix and $\psi = (\psi_1, \ldots, \psi_K)^\top$. We define the least squares estimators $\widehat{\mathcal{Z}}_t = (\widehat{\mathcal{Z}}_{t,1}, \ldots, \widehat{\mathcal{Z}}_{t,L})^\top$ and $\widehat{\mathcal{A}} = \big(\hat{\alpha}_{l,k}\big)$:

$$S(\mathcal{A}, z) \equiv \sum_{t=1}^{T}\sum_{j=1}^{J}\big\{Y_{t,j} - (1, z_t^\top)\mathcal{A}\psi(X_{t,j})\big\}^2 = \min_{\mathcal{A},z}! \qquad (5)$$

where $z = \big(z_1^\top, \ldots, z_T^\top\big)^\top$ for L-dimensional vectors z_t. With $\widehat{\mathcal{A}}$ at hand, we estimate m by $\hat{m} = \widehat{\mathcal{A}}\psi$.

We note that, given z or $\mathcal{A}$, the function S in (5) is quadratic with respect to the other variables, and thus has an explicit unique minimizer. However, minimization of S with respect to $\mathcal{A}$ and z simultaneously is a fourth-order problem. The solution is neither unique nor explicit. It is unique only up to the values of $\widehat{\mathcal{Z}}_1^\top\widehat{\mathcal{A}}, \ldots, \widehat{\mathcal{Z}}_T^\top\widehat{\mathcal{A}}$, where $\widehat{\mathcal{Z}}_t^\top = (1, \widehat{Z}_t^\top)$. We will come back to this identifiability issue later in this section.

To find a solution $(\widehat{\mathcal{A}},\ \widehat{Z})$ of the minimization problem (5), one might adopt the following iterative algorithm: (i) Given an initial choice $Z^{(0)}$, minimize $S(\mathcal{A}, Z^{(0)})$ with respect to $\mathcal{A}$, which is an ordinary least squares problem and thus has an explicit unique solution. Call it $\mathcal{A}^{(1)}$. (ii) Minimize $S(\mathcal{A}^{(1)}, z)$ with respect to z now, which is also an ordinary least squares problem. (iii) Iterate (i) and (ii) until convergence. This is the approach taken by Fengler et al. (2007). However, the procedure is not guaranteed to converge to a solution of the original problem.

We propose to use a Newton-Raphson algorithm. Let $\alpha \equiv \alpha(\mathcal{A})$ denote the stack form of $\mathcal{A} = (\alpha_{l,k})$ [i.e., $\alpha = (\alpha_{0,1}, \ldots, \alpha_{L,1}, \alpha_{0,2}, \ldots, \alpha_{L,2}, \ldots, \alpha_{0,K}, \ldots, \alpha_{L,K})^\top$]. In a slight abuse of notation we write $S(\alpha, z)$ for $S(\mathcal{A}, z)$. Define

$$\begin{aligned}
F_{10}(\alpha, z) &= \frac{\partial}{\partial\alpha}S(\alpha, z), \quad F_{01}(\alpha, z) = \frac{\partial}{\partial z}S(\alpha, z),\\
F_{20}(\alpha, z) &= \frac{\partial^2}{\partial\alpha^2}S(\alpha, z), \quad F_{11}(\alpha, z) = \frac{\partial^2}{\partial\alpha\partial z}S(\alpha, z),\\
F_{02}(\alpha, z) &= \frac{\partial^2}{\partial z^2}S(\alpha, z).
\end{aligned}$$

Let $\Psi_t = [\psi(X_{t,1}), \ldots, \psi(X_{t,J})]$ be a $K \times J$ matrix. Define A to be the $L \times K$ matrix obtained by deleting the first row of $\mathcal{A}$. Writing $\zeta_t^\top = (1, z_t^\top)$, it can be shown that

$$F_{10}(\alpha, z) = 2\sum_{t=1}^{T}\big[(\Psi_t\Psi_t^\top) \otimes (\zeta_t\zeta_t^\top)\big]\ \alpha - 2\sum_{t=1}^{T}(\Psi_t Y_t) \otimes \zeta_t,$$

$$F_{20}(\alpha, z) = 2\sum_{t=1}^{T}\big[(\Psi_t\Psi_t^\top) \otimes (\zeta_t\zeta_t^\top)\big],$$

$F_{01}(\alpha, z)^\top = 2\big(\zeta_1^\top\mathcal{A}\Psi_1\Psi_1^\top A^\top - Y_1^\top\Psi_1^\top A^\top, \ldots, \zeta_T^\top\mathcal{A}\Psi_T\Psi_T^\top A^\top - Y_T^\top\Psi_T^\top A^\top\big)$, and $F_{02}(\alpha, z)$ equals a $(TL) \times (TL)$ matrix that consists of T diagonal blocks $A\Psi_t\Psi_t^\top A^\top$ for $t = 1, \ldots, T$. Here and later, $\otimes$ denotes the Kronecker product operator. Also, by some algebraic manipulations it can be shown that

$$\big[(\Psi_t\Psi_t^\top) \otimes (\zeta_t\zeta_t^\top)\big]\alpha\ =\ (\Psi_t\Psi_t^\top\mathcal{A}^\top\zeta_t) \otimes \zeta_t. \qquad (6)$$

Let $\mathcal{I}$ be an $(L+1) \times L$ matrix such that $\mathcal{I}^{\mathrm{T}} = (0, I_L)$ and I_L denote the identity matrix of dimension L. Define $F_{11,t}(\alpha, z) = (\Psi_t\Psi_t^\top A^\top) \otimes \zeta_t + (\Psi_t\Psi_t^\top\mathcal{A}^\top\zeta_t) \otimes \mathcal{I} - (\Psi_t Y_t) \otimes$

$\mathcal{I}$. Then, we get $F_{11}(\alpha, z) = 2\ (F_{11,1}(\alpha, z), F_{11,2}(\alpha, z), \ldots, F_{11,T}(\alpha, z))$. Let

$$F(\alpha, z) = \begin{pmatrix} F_{10}(\alpha, z) \\ F_{01}(\alpha, z) \end{pmatrix},\ F'(\alpha, z) = \begin{pmatrix} F_{20}(\alpha, z) & F_{11}(\alpha, z) \\ F_{11}(\alpha, z)^\top & F_{02}(\alpha, z) \end{pmatrix}.$$

We need to solve the equation $F(\alpha, z) = 0$ simultaneously for α and z. We note that the matrices $(\Psi_t \Psi_t^\top) \otimes (\zeta_t \zeta_t^\top) = (\Psi_t \otimes \zeta_t)(\Psi_t \otimes \zeta_t)^\top$ and $A\Psi_t \Psi_t^\top A^\top$ are nonnegative definite. Thus, by Miranda's existence theorem (for example, see Vrahatis 1989) the nonlinear system of equations $F(\alpha, z) = 0$ has a solution.

Given $(\alpha^{\mathrm{OLD}}, Z^{\mathrm{OLD}})$, the Newton-Raphson algorithm gives the updating equation for $(\alpha^{\mathrm{NEW}}, Z^{\mathrm{NEW}})$:

$$\begin{pmatrix} \alpha^{\mathrm{NEW}} \\ Z^{\mathrm{NEW}} \end{pmatrix} = \begin{pmatrix} \alpha^{\mathrm{OLD}} \\ Z^{\mathrm{OLD}} \end{pmatrix} - F_*'(\alpha^{\mathrm{OLD}}, Z^{\mathrm{OLD}})^{-1} F(\alpha^{\mathrm{OLD}}, Z^{\mathrm{OLD}}), \tag{7}$$

where $F_*'(\alpha, z)$ for each given (α, z) is the restriction to $\mathcal{F}_*$ of the linear map defined by the matrix $F'(\alpha, z)$ and $\mathcal{F}_*$ is the linear space of values of (α, z) with $\sum_{t=1}^T z_t = 0$ and $\sum_{t=1}^T Z_t^{(0)}\ (z_t - Z_t^{(0)})^\top = 0$. We denote the initial value of the algorithm by $(\alpha^{(0)}, Z^{(0)})$. We will argue later that under mild conditions, $(\hat{\alpha}, \widehat{Z})$ can be chosen as an element of $\mathcal{F}_*$.

The algorithm (7) is shown to converge to a solution of (5) at a geometric rate under some weak conditions on the initial choice $(\alpha^{(0)}, Z^{(0)})$, as is demonstrated by Theorem 1 later. We collect the conditions for the theorem.

(C1) It holds that $\sum_{t=1}^T Z_t^{(0)} = 0$. The matrix $\sum_{t=1}^T Z_t^{(0)} Z_t^{(0)\top}$ and the map $F_*'(\alpha^{(0)}, Z^{(0)})$ are invertible.
(C2) There exists a version $(\hat{\alpha}, \widehat{Z})$ with $\sum_{t=1}^T \widehat{Z}_t = 0$ such that $\sum_{t=1}^T \widehat{Z}_t Z_t^{(0)\top}$ is invertible. Also, $\hat{\alpha}_l = (\hat{\alpha}_{l1}, \ldots, \hat{\alpha}_{lK})^\top$ for $l = 0, \ldots, L$ are linearly independent.

Let $\alpha^{(k)}$ and $Z^{(k)}$ denote the kth updated vectors in the iteration with the algorithm (7). Also, we write $\mathcal{A}^{(k)}$ for the matrix that corresponds to $\alpha^{(k)}$, and $\mathcal{Z}_t^{(k)\top} = (1, Z_t^{(k)\top})$.

Theorem 1. Let T, J and K be held fixed. Suppose that the initial choice $(\alpha^{(0)}, Z^{(0)})$ satisfies (C1) and (C2). Then, for any constant $0 < \gamma < 1$ there exist $r > 0$ and $C > 0$, which are random variables depending on $\{(X_{t,j}, Y_{t,j})\}$, such that, if $\sum_{t=1}^T \| \mathcal{Z}_t^{(0)\top} \mathcal{A}^{(0)} - \widehat{\mathcal{Z}}_t^\top \widehat{\mathcal{A}} \|^2 \le r$, then

$$\sum_{t=1}^T \| \mathcal{Z}_t^{(k)\top} \mathcal{A}^{(k)} - \widehat{\mathcal{Z}}_t^\top \widehat{\mathcal{A}} \|^2 \le C 2^{-2(k-1)} \gamma^{2(2^k-1)}.$$

We now argue that under (C1) and (C2), $(\hat{\alpha}, \widehat{Z})$ can be chosen as an element of $\mathcal{F}_*$. Note first that one can always take $Z_t^{(0)}$ and $\widehat{Z}_t$ so that $\sum_{t=1}^T Z_t^{(0)} = 0$ and $\sum_{t=1}^T \widehat{Z}_t = 0$. This is because, for any version $(\hat{\alpha}, \widehat{Z})$, one has

$$\widehat{\mathcal{Z}}_t^\top \widehat{\mathcal{A}} = \hat{\alpha}_0^\top + \sum_{l=1}^L \widehat{Z}_{t,l} \hat{\alpha}_l^\top = \left(\hat{\alpha}_0^\top + \sum_{l=1}^L \overline{\widehat{Z}}_l \hat{\alpha}_l^\top \right) + \sum_{l=1}^L (\widehat{Z}_{t,l} - \overline{\widehat{Z}}_l) \hat{\alpha}_l^\top \overset{\mathrm{let}}{=} \hat{\alpha}_0^{*\top} + \sum_{l=1}^L \widehat{Z}_{t,l}^* \hat{\alpha}_l^\top = \widehat{\mathcal{Z}}_t^{*\top} \widehat{\mathcal{A}}^*,$$

where $\overline{\widehat{Z}}_l = T^{-1} \sum_{t=1}^T \widehat{Z}_{t,l}$, $\widehat{\mathcal{Z}}_t^{*\top} = (1, \widehat{Z}_t^{*\top})$ and $\widehat{\mathcal{A}}^*$ is the matrix obtained from $\widehat{\mathcal{A}}$ by replacing its first row by $\hat{\alpha}_0^{*\top}$. Furthermore, the minimization problem (5) has no unique solution. If $(\widehat{Z}_t, \widehat{\mathcal{A}})$ or $(\widehat{Z}_t, \hat{m} = \widehat{\mathcal{A}}\psi)$ is a minimizer, then also $(B^\top \widehat{Z}_t, \widetilde{B}^{-1} \hat{m})$ is a minimizer. Here

$$\widetilde{B} = \begin{pmatrix} 1 & 0 \\ 0 & B \end{pmatrix} \tag{8}$$

and B is an arbitrary invertible matrix. The special structure of $\widetilde{B}$ assures that the first component of $\widetilde{B}^\top \widehat{\mathcal{Z}}_t$ equals 1. In particular, with the choice $B = (\sum_{t=1}^T Z_t^{(0)} \widehat{Z}_t^\top)^{-1} \sum_{t=1}^T Z_t^{(0)} Z_t^{(0)\top}$ we get for $\widehat{Z}_t^* = B^\top \widehat{Z}_t$ that $\sum_{t=1}^T Z_t^{(0)} (\widehat{Z}_t^* - Z_t^{(0)})^\top = 0$.

In Section 5, we will show that, for any solution $\widehat{Z}_t$ and for any version of true Z_t, there exists a random matrix B such that $\widetilde{Z}_t = B^\top \widehat{Z}_t$ has asymptotically the same covariance structure as Z_t. This means that the difference of the inferences based on $\widetilde{Z}_t$ and Z_t is asymptotically negligible.

We also note that one can always choose $\hat{m} = \widehat{\mathcal{A}}\psi$ such that the components $\hat{m}_1, \ldots, \hat{m}_L$ are orthonormal in $L_2([0, 1]^d)$ or in other L_2 [e.g., in $L_2(T^{-1} \sum_{t=1}^T \widehat{f}_t)$ where $\widehat{f}_t$ is a kernel estimate of the density of $X_{t,j}$]. If one selects $\hat{m}$ in this way, then the matrix B should be an orthogonal matrix and the underlying time series Z_t is estimated up to such transformations.

In practice one needs to choose an initial estimate $(\alpha^{(0)}, Z^{(0)})$ to run the algorithm. One may generate normal random variates for $Z_{t,l}^{(0)}$, and then find the initial $\alpha^{(0)}$ by solving the equation $F_{10}(\alpha, Z^{(0)}) = 0$. This initial choice was found to work well in our numerical study presented in Sections 3 and 4.

As an alternative way of fitting the model (2), one may extend the idea of the principal component method that is used to fit the orthogonal factor model (1). In this way, the data $\{Y_{t,j}: 1 \le j \le J\}$ are viewed as the values of a functional datum $Y_t(\cdot)$ observed at $x = X_{t,j}$, $1 \le j \le J$, and the functional factor model given at (3) may be fitted with smooth approximations of Y_t obtained from the original dataset. If one assumes $EZ_t = 0$, $\mathrm{var}(Z_t) = I_L$, as is typically the case with the orthogonal factor model (1), then one can estimate m_l and Z_t by performing functional principal component analysis with the sample covariance function

$$\widehat{K}(x, x') = T^{-1} \sum_{t=1}^T \{Y_t(x) - \overline{Y}(x)\}\{Y_t(x') - \overline{Y}(x')\},$$

where $\overline{Y}(x) = T^{-1} \sum_{t=1}^T Y_t(x)$. There are some limitations for this approach. First, it requires initial fits to get smooth approximations of $Y_t(\cdot)$, which may be difficult when the design points $X_{t,j}$ are sparse as is the case with the IV application. Our method avoids the preliminary estimation and shifts the discrete representation directly to the functions m_l. Second, for the method to work one needs at least stationarity of Z_t and ε_t, whereas our theory does not rely on these assumptions.

3. SIMULATION STUDY

In Theorem 3 we will argue that the inference based on the covariances of the unobserved factors Z_t is asymptotically equivalent to the one based on $B^\top \widehat{Z}_t$ for some invertible B. In this section we illustrate the equivalence by a simulation study. We compare the covariances of Z_t and $\widetilde{Z}_t \equiv B^\top \widehat{Z}_t$, where

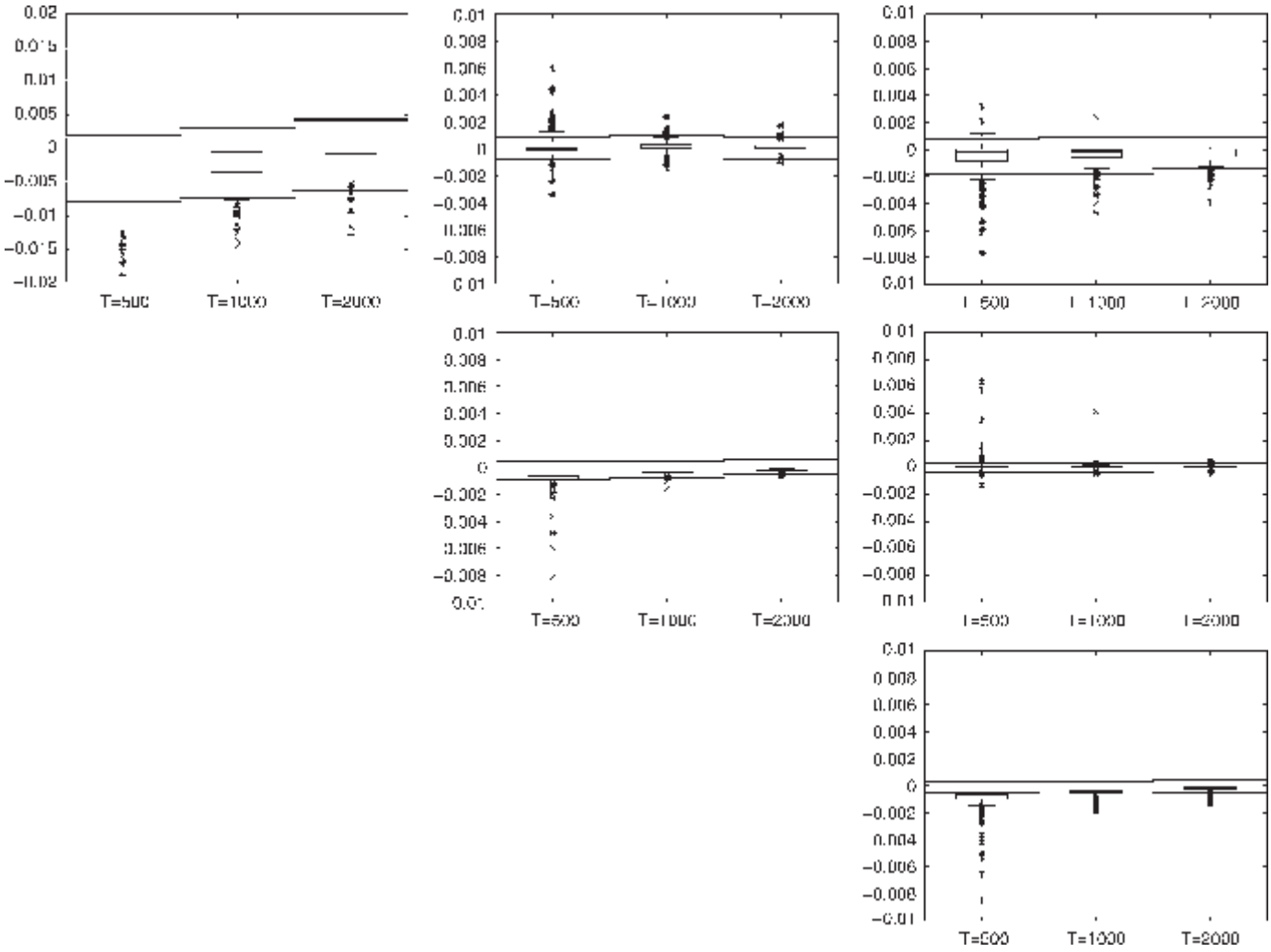


Figure 3. The boxplots based on 250 values of the entries of the scaled difference of the covariance matrices given at (10). The lengths of the series Z_t and $\tilde{Z}_t$ were 500, 1,000, 2,000. The thick lines represent the upper and lower quartiles of (11).

$$B = \left(T^{-1} \sum_{t=1}^{T} Z_{c,t} \widehat{Z}_{c,t}^{\top} \right)^{-1} T^{-1} \sum_{t=1}^{T} Z_{c,t} Z_{c,t}^{\top}, \tag{9}$$

$Z_{c,t} = Z_t - T^{-1} \sum_{s=1}^{T} Z_s$ and $\widehat{Z}_{c,t} = \widehat{Z}_t - T^{-1} \sum_{s=1}^{T} \widehat{Z}_s$. Note that B at (9) minimizes $\sum_{t=1}^{T} \| \widehat{Z}_{c,t} - (B^{\top})^{-1} Z_{c,t} \|^2$. In the Appendix we will prove that Theorem 3 holds with the choice at (9).

We took $T = 500, 1{,}000, 2{,}000$, $J = 100, 250, 1{,}000$ and $K = 36, 49, 64$. We considered $d = 2$, $L = 3$ and the following tuple of 2-dimensional functions:

$$\begin{aligned} &m_0(x_1, x_2) = 1, \quad m_1(x_1, x_2) = 3.46(x_1 - .5), \\ &m_2(x_1, x_2) = 9.45\left\{ (x_1 - .5)^2 + (x_2 - .5)^2 \right\} - 1.6, \\ &m_3(x_1, x_2) = 1.41 \sin(2\pi x_2). \end{aligned}$$

The coefficients in these functions were chosen so that m_1, m_2, m_3 are close to orthogonal. We generated Z_t from a centered VAR(1) process $Z_t = \mathcal{R} Z_{t-1} + U_t$, where U_t is $N_3(0, \Sigma_U)$ random vector, the rows of $\mathcal{R}$ from the top equal $(0.95, -0.2, 0)$, $(0, 0.8, 0.1)$, $(0.1, 0, 0.6)$, and $\Sigma_U = 10^{-4} I_3$. The design points $X_{t,j}$ were independently generated from a uniform distribution on the unit square, $\varepsilon_{t,j}$ were iid $N(0, \sigma^2)$ with $\sigma = 0.05$, and $Y_{t,j}$ were obtained according to the model (4). The simulation experiment was repeated 250 times for each combination of (T, J, K). For the estimation we employed, for ψ_j, the tensor products of linear B-splines. The one-dimensional linear B-splines $\widetilde{\psi}_k$ are defined on a consecutive equidistant knots x^k, x^{k+1}, x^{k+2} by $\widetilde{\psi}_k(x) = (x - x^k)/(x^{k+1} - x^k)$ for $x \in (x^k, x^{k+1}]$, $\widetilde{\psi}_k(x) = (x^{k+2} - x)/(x^{k+2} - x^{k+1})$ for $x \in (x^{k+1}, x^{k+2}]$, and $\widetilde{\psi}_k(x) = 0$ otherwise. We chose $K = 8 \times 8 = 64$.

We plotted in Figure 3 the entries of the scaled difference of the covariance matrices

$$\widetilde{D} = \frac{1}{\sqrt{T}} \left\{ \sum_{t=1}^{T} \left(\widetilde{Z}_t - \overline{\widetilde{Z}} \right) \left(\widetilde{Z}_t - \overline{\widetilde{Z}} \right)^{\top} - \sum_{t=1}^{T} \left(Z_t - \overline{Z} \right) \left(Z_t - \overline{Z} \right)^{\top} \right\}. \tag{10}$$

Each panel of Figure 3 corresponds to one entry of the matrix $\widetilde{D}$, and the three boxplots in each panel represent the distributions of the 250 values of the corresponding entry for $T = 500, 1{,}000, 2{,}000$. In the figure we also depicted, by thick lines, the upper and lower quartiles of

$$D = \frac{1}{\sqrt{T}} \left\{ \sum_{t=1}^{T} \left(Z_t - \overline{Z} \right) \left(Z_t - \overline{Z} \right)^{\top} - T\Gamma \right\}, \tag{11}$$

where Γ is the true covariance matrix of the simulated VAR process. We refer to Lütkepohl (1993) for a representation of Γ.

Our theory in Section 5 tells that the size of $\widetilde{D}$ is of smaller order than the normalized error D of the covariance estimator based on Z_t. It is known that the latter converges to a nondegenerate law as $T \to \infty$. This is well supported by the plots in Figure 3 showing that the distance between the two thick lines in each panel is almost invariant as T increases. The fact that the additional error incurred by using $\widetilde{Z}_t$ instead of Z_t is negligible for large T is also confirmed. In particular, the long stretches at tails of the distributions of $\widetilde{D}$ get shorter as T increases. Also, the upper and lower quartiles of each entry of $\widetilde{D}$, represented by the boxes, lie within those of the corresponding entry of D, represented by the thick lines, when $T =$ 1,000 and 2,000.

4. APPLICATIONS

This section presents an application of DSFM. We fit the model to the intraday IV based on ODAX prices and to fMRI data.

For our analysis we chose the data observed from July 1, 2004 to June 29, 2005. The one year period corresponds to the financial regulatory requirements. The data were taken from Financial and Economic Data Center of Humboldt-Universität zu Berlin. The IV data were regressed on the two-dimensional space of future moneyness and time-to-maturity, denoted by $(\kappa_t, \tau_t)^\top$. The future moneyness κ_t is a monotone function of the strike price K: $\kappa_t = K/(S_t e^{-r_t \tau_t})$, where S_t is the spot price at time t and r_t is the interest rate. We chose r_t as a daily Euro Interbank Offered Rate (EURIBOR) taken from the Ecowin Reuters database. The time-to-maturity of the options were measured in years. We took all trades with $10/365 < \tau < 0.5$. We limit also the moneyness range to $\kappa \in [0.7, 1.2]$.

The structure of the IV data, described already in Section 1, requires a careful treatment. Apart from the dynamic degeneration, one may also observe nonuniform frequency of the trades with significantly greater market activities for the options closer to expiry or at-the-money. Here, "at-the-money" means a condition in which the strike price of an option equals the spot price of the underlying security (i.e., $K = S_t$). To avoid the computational problems with the highly skewed empirical distribution of $X_t = (\kappa_t, \tau_t)$, we transformed the initial space $[0.7, 1.2] \times [0.03, 0.5]$ to $[0, 1]^2$ by using the marginal empirical distribution functions. We applied the estimation algorithm to the transformed space, and then transformed back the results to the original space.

Because the model is not nested, the number of the dynamic functions needs to be determined in advance. For this, we used

$$RV(L) = \frac{\sum_t^T \sum_j^{J_t} \left\{ Y_{t,j} - \hat{m}_0(X_{t,j}) - \sum_{l=1}^{L} \widehat{Z}_{t,l} \hat{m}_l(X_{t,j}) \right\}^2}{\sum_t^T \sum_j^{J_t} \left(Y_{t,j} - \overline{Y} \right)^2}, \tag{12}$$

although one may construct an Akaike information (AIC) or Bayesian information (BIC) type of criterion, where one penalizes the number of the dynamic functions in the model, or performs some type of cross-validation. The quantity $1 - RV(L)$ can be interpreted as a proportion of the variation explained by the model among the total variation. The computed values of $RV(L)$ are given in Table 1 for various L. Because the third, fourth, and fifth factor made only a small improvement in the fit, we chose $L = 2$.

Table 1. Proportion of the explained variation by the models with L = 1, . . ., 5 dynamic factors

No. factors	$L = 1$	$L = 2$	$L = 3$	$L = 4$	$L = 5$
$1 - RV(L)$	0.848	0.969	0.976	0.978	0.980

For the series estimators of $\hat{m}_l$ we used tensor B-splines that are cubic in the moneyness and quadratic in the maturity direction. In the transformed space we placed 10×5 knots, 10 in the moneyness and 5 in the maturity direction. We found that the results were not sensitive to the choice of the number of knots and the orders of splines. For several choices of knots in the range 5×5–15×10 and for the spline orders (2, 1), (2, 2), (3, 2), the values of $1 - RV(2)$ were between 0.949 and 0.974. Because the model is identifiable only up to the transformation (8), one has a freedom for the choice of factors. Here, we chose the approach taken by Fengler et al. (2007) with $L_2[0,1]^2$ norm. Specifically, we orthonormalized $\hat{m}_l$ and transformed $\widehat{Z}_t$ according to their Equation (19) with $\Gamma = \int \hat{m}(x)\hat{m}(x)^\top \, dx$, where $\hat{m} = (\hat{m}_1, \ldots, \hat{m}_L)^\top$. Call them $\hat{m}_l^*$ and $\widehat{Z}_t^*$, respectively. Then, we transformed them further by $\hat{m}_l^{**} = p_l^\top \hat{m}^*$ and $\widehat{Z}_{t,l}^{**} = p_l^\top \widehat{Z}_t^*$, where p_l were the orthonormal eigenvectors of the matrix $\sum_{t=1}^T \widehat{Z}_t^* \widehat{Z}_t^{*\top}$ that correspond to the eigenvalues $\lambda_1 > \lambda_2$. Note that $\widehat{Z}_t^{*\top} \hat{m}^* = \widehat{Z}_t^{**\top} \hat{m}^{**}$. In this way, $\{\widehat{Z}_{t,1}^{**} \hat{m}_1^{**}\}$ makes a larger contribution than $\{\widehat{Z}_{t,2}^{**} \hat{m}_2^{**}\}$ to the total variation $\sum_{t=1}^T \int (\widehat{Z}_t^{**\top} \hat{m}^{**})^2$ because $\sum_{t=1}^T \int (\widehat{Z}_{t,1}^{**} \hat{m}_1^{**})^2 = \lambda_1$ and $\sum_{t=1}^T \int (\widehat{Z}_t^{**\top} \hat{m}^{**})^2 = \lambda_1 + \lambda_2$. Later, we continue to write $\widehat{Z}_t$ and $\hat{m}$ for such $\widehat{Z}_t^{**}$ and $\hat{m}^{**}$, respectively.

The estimated functions $\hat{m}_1$ and $\hat{m}_2$ are plotted in Figure 4 in the transformed estimation space. The intercept function $\hat{m}_0$ was almost flat around zero, thus is not given. By construction, $\hat{m}_0 + \widehat{Z}_{t,1} \hat{m}_1$ explain the principal movements of the surface. It was observed by Cont and da Fonseca (2002) and Fengler et al. (2007) that most dominant innovations of the entire surface are parallel level shifts. Note that VDAX is an estimated at-the-money IV for an option with 45 days to maturity, and thus indicates up-and-down shifts. The left panel of Figure 5 shows the values of VDAX together with $\hat{m}_0(X_{t,0}) + \widehat{Z}_{t,1} \hat{m}_1(X_{t,0})$, where $X_{t,0}$ is the moneyness and maturity corresponding to an option at-the-money with 45 days to maturity. The right panel of Figure 5 depicts the factor $\widehat{Z}_t$, where one can find that $\widehat{Z}_t$ shows almost the same dynamic behavior as the index VDAX. This similarity supports that DSFM catches leading dynamic effects successfully. Obviously the model in its full setting explains other effects, such as skew or term structure changes, which are not explicitly stated here.

Statistical analysis on the evolution of a high-dimensional system ruling the option prices can be simplified to a low-dimensional analysis of the $\widehat{Z}_t$. In particular, as our theory in Section 5 and the simulation results in Section 3 assert, the inference based on the $\widehat{Z}_t$ is well justified in the VAR context. To select a VAR model we computed the Schwarz (SC), the Hannan-Quinn (HQ), and the Akaike criterion, as given in

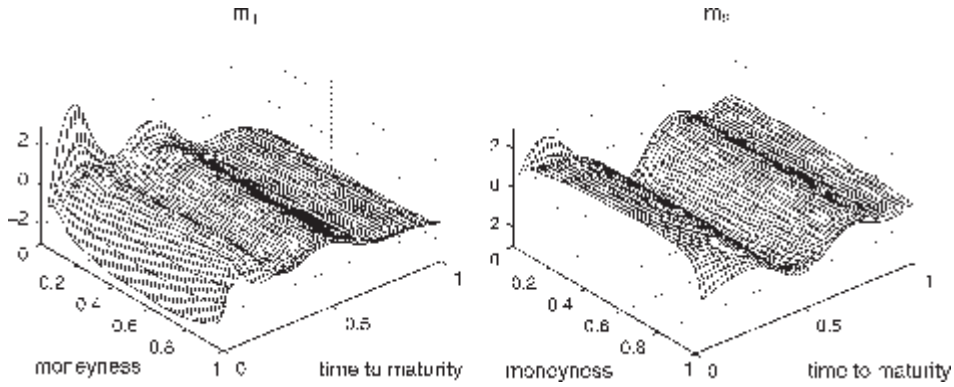


Figure 4. The estimated factor functions for the ODAX IV data in the period 20040701–20050629.

Table 2. One can find that SC and HQ suggest a VAR(1) process, whereas AIC selects VAR(2). The parameter estimates for each selected model are given in Table 3. The roots of the characteristic polynomial lie inside the unit circle, so the specified models satisfy the stationarity condition. For each of VAR(1) and VAR(2) models, we conducted a portmanteau test for the hypothesis that the autocorrelations of the error term at lags up to 12 are all zero, and also a series of LM tests, each of which tests whether the autocorrelation at a particular lag up to 5 equals zero. Some details on selection of lags for these tests can be found in Hosking (1980, 1981) and Brüggemann, Lütkepohl, and Saikkonen (2006). We found that in any test the null hypothesis was not rejected at 5% level. A closer inspection on the autocorrelations of the residuals, however, revealed that the autocorrelation of $\widehat{Z}_{t,2}$ residuals at lag one is slightly significant in the VAR(1) model, see Figure 6. But, this effect disappears in the VAR(2) case, see Figure 7. Similar analyses of characteristic polynomials, portmanteau and Lagrange multiplier (LM) tests supported VAR(2) as a successful model for $\widehat{Z}_t$.

As a second application of the model, we considered fitting an fMRI dataset. The data were obtained at Max-Planck Institut für Kognitions-und-Neurowissenschaften Leipzig by scanning a subject's brain using a standard head coil. The scanning was done every two seconds on the resolution of $3 \times 3 \times 2$ mm^3 with 1 mm gap between the slices. During the experiment, the subject was exposed to three types of objects (bench, phone and motorbike) and rotated around randomly changing axes for four seconds, followed by relaxation phase of six to ten seconds. Each stimulus was shown 16 times in pseudo-randomized order. As a result, a series of 290 images with $64 \times 64 \times 30$ voxels was obtained.

To apply the model (2) to the fMRI data, we took the voxel's index (i_1, i_2, i_3) as covariate $X_{t,j}$, and the BOLD signal as $Y_{t,j}$. For numerical tractability we reduced the original data to a series of $32 \times 32 \times 15$ voxels by taking every second slice in each direction. Thus, $J_t \equiv 32 \times 32 \times 15$ and $T = 290$. The voxels' indices (i_1, i_2, i_3) for $1 \le i_1, i_2 \le 32$;$1 \le i_3 \le 15$ are associated with $32 \times 32 \times 15$ equidistant points in $\mathbb{R}^3$. The function m_0 represents the ''average'' signal as a function of the three-dimensional location, and m_l for each $l \ge 1$ determines the effect of the lth common factor $Z_{t,l}$ on the brain's signal. In Figure 8, each estimated function $\hat{m}_l$ is represented by its sections on the 15 slices in the direction of i_3 [i.e., by those $\hat{m}_l(\cdot, \cdot, x_3)$ for which x_3 are fixed at the equidistant points corresponding to $i_3 = 1, \ldots, 15$]. We used quadratic tensor B-splines on equidistant knots. The number of knots in each direction was 8, 8, 4, respectively, so that $K = 9 \times 9 \times 5 = 405$. For the model identification we used the same method as in the IV application, but normalized $\widehat{Z}$ to have mean zero.

In contrast to the IV application, there was no significant difference between the values of $1 - RV(L)$ for different $L \ge 1$.

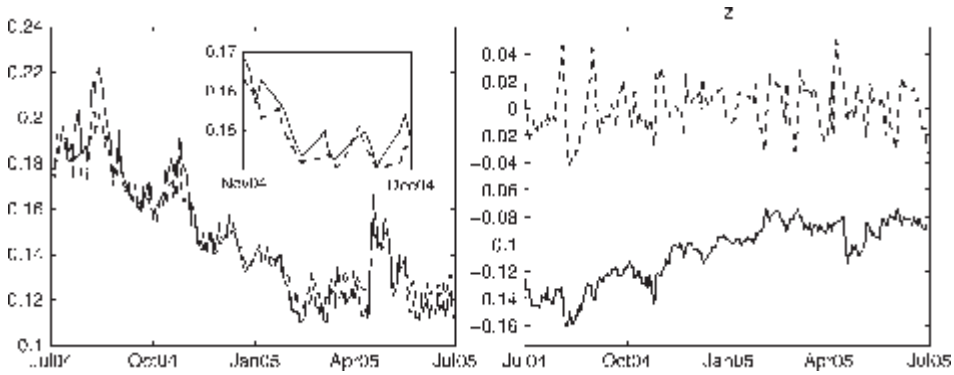


Figure 5. Left panel: VDAX in the period 20040701–20050629 (solid) and the dynamics of the corresponding IV given by the submodel $\hat{m}_0 + \widehat{Z}_{t,1}\hat{m}_1$ (dashed). Right panel: The obtained time series $\widehat{Z}_t$ on the ODAX IV data in the period 20040701–20050629. The solid line represents $\widehat{Z}_{t,1}$, the dashed line $\widehat{Z}_{t,2}$.

Table 2. The VAR model selection criteria. The smallest value for each criterion is marked by an asterisk

Order	AIC	SC	HQ
1	−14.06	−13.98*	−14.03*
2	−14.07*	−13.93	−14.02
3	−14.06	−13.86	−13.98
4	−14.06	−13.81	−13.96
5	−14.07	−13.76	−13.95

All the values for $L \geq 1$ were around 0.871. The fMRI signals $Y_{t,j}$ were explained mostly by $\hat{m}_0(X_{t,j}) + Z_{t,1}\hat{m}_1(X_{t,j})$, and the effects of the common factors $Z_{t,l}$ for $l \geq 2$ were relatively small. The slow increase in the value of $1 - RV(L)$ as $L \geq 1$ grows in the fMRI application, contrary to the case of the IV application, can be explained partly by the high complexity of human brain. Because the values of $1 - RV(L)$ were similar for $L \geq 1$, one might choose $L = 1$. However, we chose $L = 4$, which we think still allows relatively low complexity, to demonstrate some further analysis that might be possible with similar datasets. The estimated functions $\hat{m}_l$ for $0 \leq l \leq 4$ and the time series $\widehat{Z}_{t,l}$ for $1 \leq l \leq 4$ are plotted in Figures 8 and 9, respectively. The function $\hat{m}_0$ can be recognized as a smoothed version of the original signal. By construction the first factor and loadings incorporate the largest variation. One may see the strong positive trend in $\widehat{Z}_{t,1}$ and relatively flat patterns of $\widehat{Z}_{t,2}, \widehat{Z}_{t,3}, \widehat{Z}_{t,4}$. These effects could be typically explained by the mixture of several components, such as physiological pulsation, subtle head movement, machine noise, and so on. For a description of different artifacts, which significantly influence the fMRI signals, we refer to Biswal, Yetkin, Haughton, and Hyde (1995). The function estimates $\hat{m}_l$ for $1 \leq l \leq 4$ appear to have a clear peak, and $\widehat{Z}_{t,l}$ for $2 \leq l \leq 4$ show rather mild mean reverting behavior.

To see how the recovered signals interact with the given stimuli, we plotted $\widehat{Z}_{t+s,l} - \widehat{Z}_{s,l}$ against t in Figure 10, where s is the time when a stimulus appears. The mean changes of $\widehat{Z}_{t,1}$ and $\widehat{Z}_{t,3}$ show mild similarity, up to sign change, to the hemodynamic response (see Worsley et al. 2002). The case of $\widehat{Z}_{t,4}$ has a similar pattern as those of $\widehat{Z}_{t,1}$ and $\widehat{Z}_{t,3}$ but with larger amplitude, whereas the changes in $\widehat{Z}_{t,2}$ seem to be independent of the stimuli. In fitting the fMRI data, we did not use any external information on the signal. From the biological perspective it could be hardly expected that a pure statistical procedure gives full insight into understanding of the complex dynamics of MR images. For the latter one needs to incorporate into the procedure the shape of hemodynamic response, for example, or consider physiologically motivated identification of the factors. It goes however beyond the scope of this illustrative example.

5. ASYMPTOTIC ANALYSIS

In the simulation study and the real data application in Sections 3 and 4, we considered the case where Z_t is a VAR-process. Here, we only make some weak assumptions on the average behavior of the process. In our first theorem we allow that it is a deterministic sequence. In our second result we assume that it is a mixing sequence. For the asymptotic analysis, we let $K, J, T \to \infty$. This is a very natural assumption often also made in cross-sectional or panel data analysis. It is appropriate for data with many observations per data point that are available for many dates. It allows us to study how J and T have to grow with respect to each other for a good performance of a procedure. The distance between m and its best approximation $\mathcal{A}\psi$ does not tend to zero unless $K \to \infty$, see Assumption (A5) later. One needs to let $J \to \infty$ to get consistency of $\widehat{\mathcal{Z}}_t^\top \widehat{\mathcal{A}}$ and $\hat{m} = \widehat{\mathcal{A}}\psi$ as estimates of $\mathcal{Z}_t^\top \mathcal{A}^*$ and m, respectively, where $\mathcal{A}^*$ is defined at (A5). One should let $T \to \infty$ to describe the asymptotic equivalence between the lagged covariances of Z_t and those of $\widetilde{Z}_t$, see Theorem 3 below. In our analysis the dimension L is fixed. Clearly, one could also study our model with L growing to infinity. We treat the case where X_{it} are random. However, a theory for deterministic designs can be developed along the lines of our theory.

Our first result relies on the following assumptions.

(A1) The variables $X_{1,1}, \ldots, X_{T,J}, \varepsilon_{1,1}, \ldots, \varepsilon_{T,J}$, and $Z_1, \ldots, Z_T$ are independent. The process Z_t is allowed to be nonrandom.
(A2) For $t = 1, \ldots, T$ the variables $X_{t,1}, \ldots, X_{t,J}$ 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 $t = 1, \ldots, T$.
(A3) We assume that $E\varepsilon_{t,j} = 0$ for $1 \leq t \leq T$, $1 \leq j \leq J$, and for $c > 0$ small enough $\sup_{1\leq t\leq T,\ 1\leq j\leq J} E\exp(c\varepsilon_{t,j}^2) < \infty$.
(A4) The functions ψ_k may depend on the increasing indices T and J, but 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 vector of functions $m = (m_0, \ldots, m_L)^\top$ can be approximated by ψ_k, i.e.,

$$\delta_K \equiv \sup_{x\in[0,1]^d} \inf_{\mathcal{A}\in\mathbb{R}^{(L+1)\times K}} \|m(x) - \mathcal{A}\psi(x)\| \to 0$$

as $K \to \infty$. We denote $\mathcal{A}$ that fulfills $\sup_{x\in[0,1]^d} \|m(x) - \mathcal{A}\psi(x)\| \leq 2\delta_K$ by $\mathcal{A}^*$.
(A6) There exist constants $0 < C_L < C_U < \infty$ such that all eigenvalues of the matrix $T^{-1}\sum_{t=1}^T \mathcal{Z}_t\mathcal{Z}_t^\top$ lie in the interval $[C_L, C_U]$ with probability tending to one.

Table 3. The estimated parameters for VAR(1) and VAR(2) models. Those that are not significant at 5% level are marked by asterisk

	VAR(1)			VAR(2)				
	$\widehat{Z}_{t-1,1}$	$\widehat{Z}_{t-1,2}$	Const.	$\widehat{Z}_{t-1,1}$	$\widehat{Z}_{t-1,2}$	$\widehat{Z}_{t-2,1}$	$\widehat{Z}_{t-2,2}$	Const.
$\widehat{Z}_{t,1}$	0.984	−0.029*	−0.001	0.913	−0.025	0.071	−0.004	−0.001
$\widehat{Z}_{t,2}$	0.055	0.739	0.005	0.124	0.880	−0.065	−0.187*	0.006

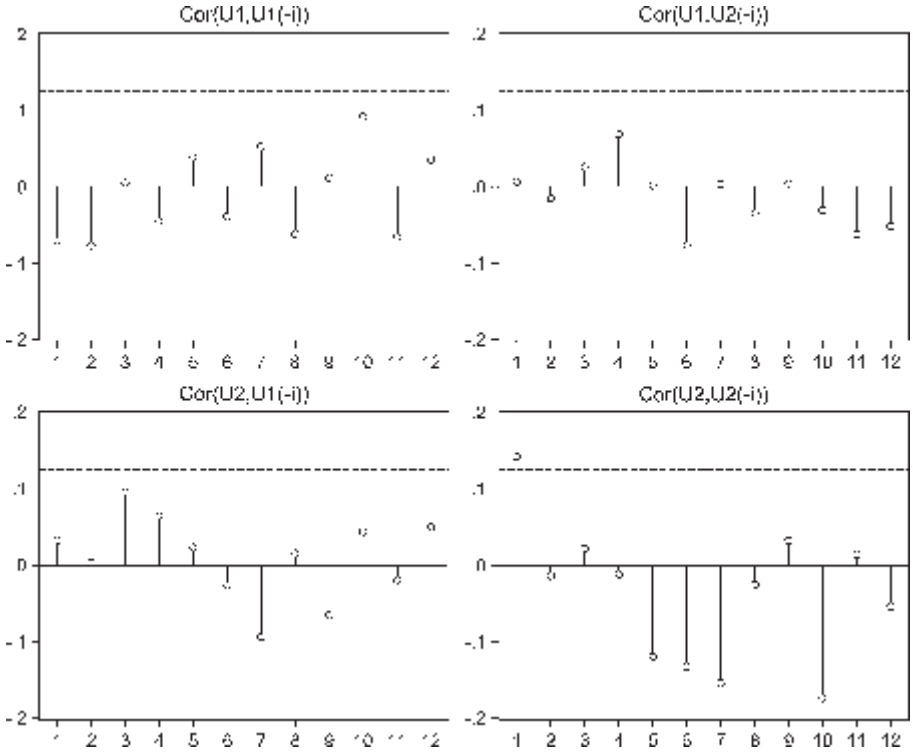


Figure 6. Cross-autocorrelogram for the VAR(1) residuals. The dashed line-bounds indicate $\pm 2 \times$ (standard deviations), which correspond to an approximate 95% confidence bound.

(A7) The minimization (5) runs over all values of $(\mathcal{A}, z)$ with

$$\sup_{x\in[0,1]} \max_{1\le t\le T} \| (1, z_t^\top)\mathcal{A}\psi(x) \| \le M_T,$$

where the constant M_T fulfils $\max_{1\le t\le T} \|Z_t\| \le M_T/C_m$ (with probability tending to one) for a constant C_m such that $\sup_{x\in[0,1]}\|m(x)\| < C_m$.

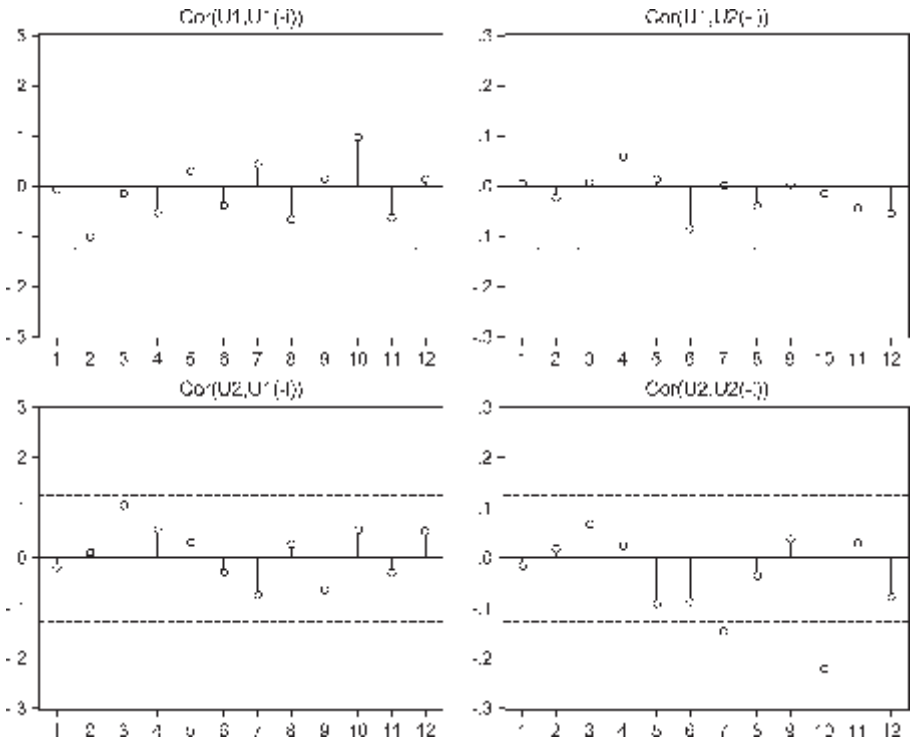


Figure 7. Cross-autocorrelogram for the VAR(2) residuals. The dashed line-bounds indicate $\pm 2 \times$ (standard deviations), which correspond to an approximate 95% confidence bound.

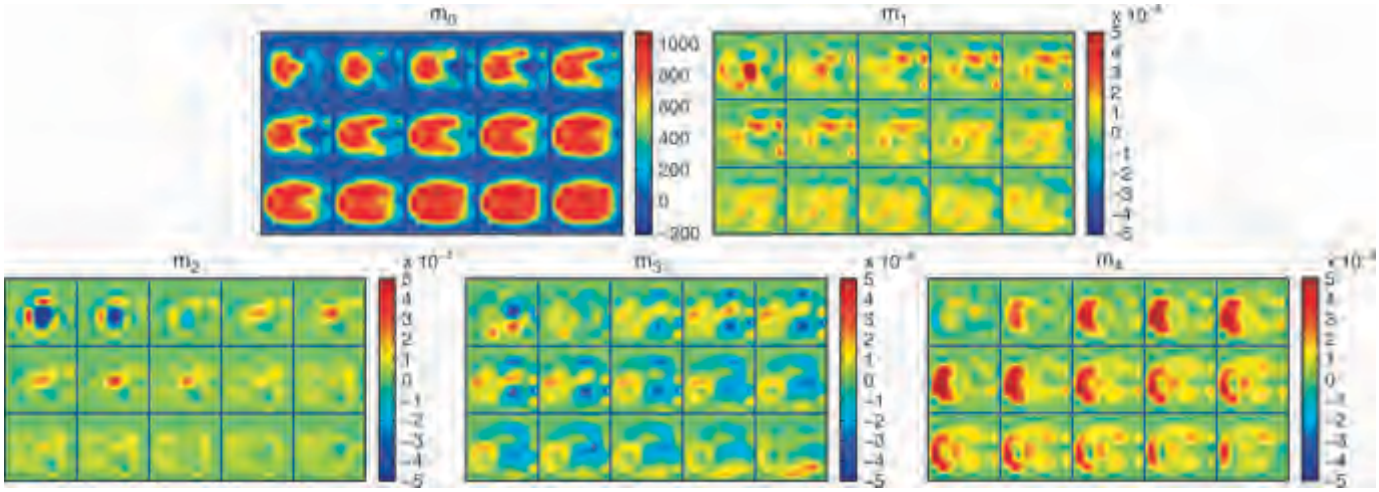


Figure 8. The estimated functions $\hat{m}_l$ for the fMRI signals.

(A8) It holds that $\rho^2 = (K+T)M_T^2 \log(JTM_T)/(JT) \to 0$. The dimension L is fixed.

Assumption (A7) and the additional bound M_T in the minimization is introduced for purely technical reasons. We conjecture that to some extent the asymptoic theory of this article could be developed under weaker conditions. The independence assumptions in (A1) and Assumption (A3) could be relaxed to assuming that the errors $\epsilon_{t,j}$ have a conditional mean zero and have a conditional distribution with subgaussian tails, given the past values $X_{s,i}$, Z_s $(1 \le i \le J, 1 \le s \le t)$. Such a theory would require an empirical process theory that is more explicitly designed for our model and it would also require a lot of more technical assumptions. We also expect that one could proceed with the assumption of subexponential instead of subgaussian tails, again at the cost of some additional conditions. Recall that the number of parameters to be estimated equals $TL + K(L+1)$. Because L is fixed, Assumption (A8) requires basically that, neglecting the factor $M_T{}^2 \log(JTM_T)$, the number of parameters grows slower than the number of observations, JT.

Our first result gives rates of convergence for the least squares estimators $\widehat{Z}_t$ and $\widehat{\mathcal{A}}$.

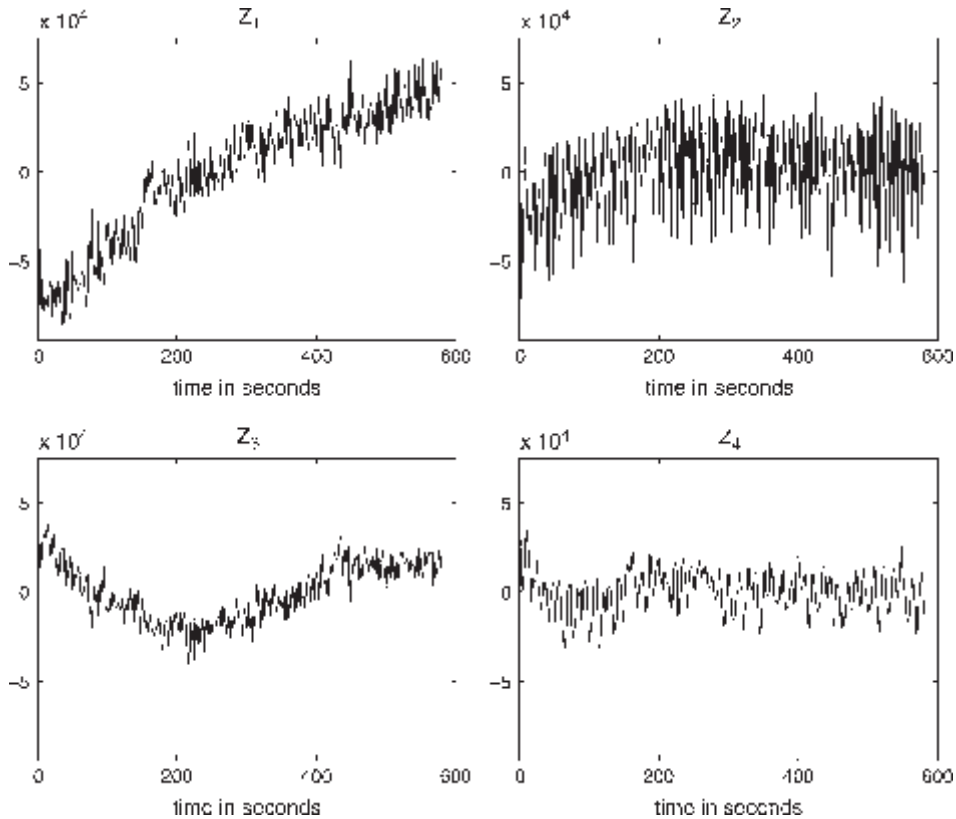


Figure 9. The estimated time series $\widehat{Z}_{t,l}$ for the fMRI signals.

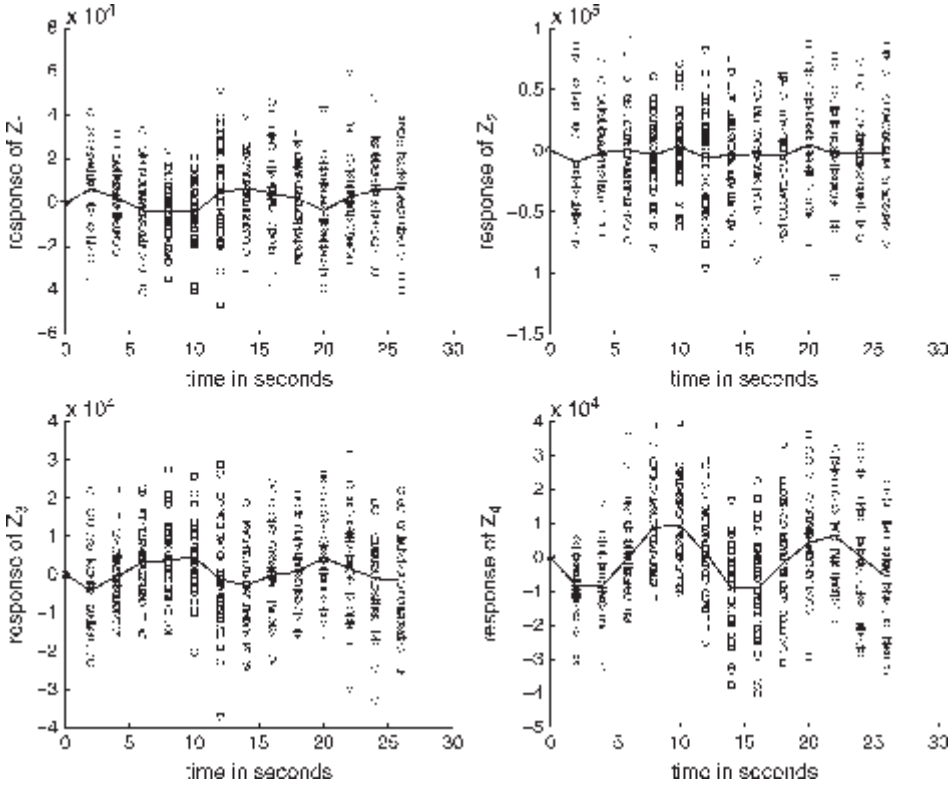


Figure 10. The responses of $\widehat{Z}_{t,l}$ to the stimuli.

Theorem 2. Suppose that model (4) holds and that $(\widehat{\mathcal{Z}}_t, \widehat{\mathcal{A}})$ is defined by the minimization problem (5). Make the Assumptions (A1)–(A8). Then it holds that

$$\frac{1}{T}\sum_{1\le t\le T}\left\|\widehat{\mathcal{Z}}_t^\top\widehat{\mathcal{A}} - \mathcal{Z}_t^\top\mathcal{A}^*\right\|^2 = \mathcal{O}_P(\rho^2+\delta_K^2). \qquad (13)$$

At this point we have made no assumptions on the sequence Z_t: $1 \le t \le T$, besides the bound in (A7). Up to now it is allowed to be a deterministic or a random sequence. We now assume that it is a random process. We discuss how a statistical analysis differs if inference on Z_t is based on $\widehat{Z}_t$ instead of using (the unobserved) process Z_t. We will show that the differences are asymptotically negligible (except an orthogonal transformation). This is the content of the following theorem, where we consider estimators of autocovariances and show that these estimators differ only by second order terms. This asymptotic equivalence carries over to classical estimation and testing procedures in the framework of fitting a vector autoregresssive model. For the statement of the theorem we need the following assumptions:

(A9) Z_t is a strictly stationary sequence with $E(Z_t) = 0$, $E(\|Z_t\|^\gamma) < \infty$ for some $\gamma > 2$. It is strongly mixing with $\sum_{i=1}^{\infty}\alpha(i)^{(\gamma-2)/\gamma} < \infty$. The matrix $EZ_tZ_t^{\mathrm{T}}$ has full rank. The process Z_t is independent of $X_{11}, \ldots, X_{TJ}, \epsilon_{11}, \ldots, \epsilon_{TJ}$.

(A10) The functions $m_0, \ldots, m_L$ are linearly independent. In particular, no function is equal to 0.

(A11) It holds that $[\log(KT)^2\{(KM_T/J)^{1/2} + T^{1/2}M_T^4J^{-2} + K^{3/2}J^{-1} + K^{4/3}J^{-2/3}T^{-1/6}\} + 1]T^{1/2}(\rho^2+\delta_K^2) = o(1)$.

Assumption (A11) poses very weak conditions on the growth of J, K, and T. Suppose, for example, that M_T is of logarithmic order and that K is of order $(TJ)^{1/5}$ so that the variance and the bias are balanced for twice differentiable functions. In this setting, (A11) only requires that T/J^2 times a logarithmic factor converges to zero. Define $\widetilde{Z}_t = B^\top\widehat{Z}_t$,

$$\widetilde{Z}_{c,t} = \widetilde{Z}_t - T^{-1}\sum_{s=1}^{T}\widetilde{Z}_s,$$
$$Z_{c,t} = Z_t - T^{-1}\sum_{s=1}^{T}Z_s,$$
$$\widetilde{Z}_{n,t} = (T^{-1}\textstyle\sum_{s=1}^{T}\widetilde{Z}_{c,s}\widetilde{Z}_{c,s}^\top)^{-1/2}\widetilde{Z}_{c,t},$$

and $Z_{n,t} = (T^{-1}\sum_{s=1}^{T}Z_{c,s}Z_{c,s}^\top)^{-1/2}Z_{c,t}$.

Theorem 3. Suppose that model (4) holds and that $(\widehat{Z}_t, \widehat{\mathcal{A}})$ is defined by the minimization problem (5). Make the Assumptions (A1)–(A11). Then there exists a random matrix B such that for $h \ne 0$

$$\frac{1}{T}\sum_{t=\max[1,-h+1]}^{\min[T,T-h]}\widetilde{Z}_{c,t}\big(\widetilde{Z}_{c,t+h} - \widetilde{Z}_{c,t}\big)^\top - \frac{1}{T}\sum_{t=\max[1,-h+1]}^{\min[T,T-h]}Z_{c,t}\big(Z_{c,t+h} - Z_{c,t}\big)^\top = o_P(T^{-1/2}),$$

$$\frac{1}{T}\sum_{t=\max[1,-h+1]}^{\min[T,T-h]}\widetilde{Z}_{n,t}\widetilde{Z}_{n,t+h}^\top - \frac{1}{T}\sum_{t=\max[1,-h+1]}^{\min[T,T-h]}Z_{n,t}Z_{n,t+h}^\top = o_P(T^{-1/2}).$$

To illustrate an implication of Theorem 3, suppose that the factor process Z_t in (4) is a stationary VAR(p) process in a mean adjusted form:

$$Z_t - \mu = \Theta_1(Z_{t-1} - \mu) + \ldots + \Theta_p(Z_{t-p} - \mu) + U_t, \quad (14)$$

where $\mu = E(Z_t)$, Θ_j is a $L \times L$ matrix of coefficients and U_t is a white noise with a nonsingular covariance matrix. Let Γ_h be the autocovariance matrix of the process Z_t with the lag $h \geq 0$, which is estimated by $\widehat{\Gamma}_h = T^{-1}\sum_{t=h+1}^{T}(Z_t - \overline{Z})(Z_{t-h} - \overline{Z})^\top$. Let $Y = (Z_{p+1} - \mu, \ldots, Z_T - \mu)$, $\Theta = (\Theta_1, \ldots, \Theta_p)$, and $U = (U_{p+1}, \ldots, U_T)$. Define $W_t = \left((Z_t - \mu)^\top, \ldots, (Z_{t-p+1} - \mu)^\top\right)^\top$ and $W = (W_p, \ldots, W_{T-1})$. Then, the model (14) can be rewritten as $Y = \Theta W + U$ and the least squares estimator of Θ is given by $\widehat{\Theta} = \widehat{Y}\widehat{W}^\top(\widehat{W}\widehat{W}^\top)^{-1}$, where $\widehat{Y}$ and $\widehat{W}$ are the same as Y and W, respectively, except that μ is replaced by $\overline{Z}$. Likewise, fitting a VAR(p) model with the estimated factor process $\widehat{Z}_t$ yields $\widetilde{\Theta} = \widetilde{Y}\widetilde{W}^\top(\widetilde{W}\widetilde{W}^\top)^{-1}$, where $\widetilde{Y}$ and $\widetilde{W}$ are defined as $\widehat{Y}$ and $\widehat{W}$ with Z_t being replaced by $\widehat{Z}_t$. Both $\widehat{Y}$ and $\widehat{W}$ are matrices composed of $\widehat{\Gamma}_h$ for various h. The matrices $\widetilde{Y}$ and $\widetilde{W}$ have the same forms as $\widehat{Y}$ and $\widehat{W}$, respectively, but with $\widehat{\Gamma}_h$ being replaced by $\widetilde{\Gamma}_h = T^{-1}\sum_{t=h+1}^{T}(\widehat{Z}_t - \overline{\widehat{Z}})(\widehat{Z}_{t-h} - \overline{\widehat{Z}})^\top$. It is well known that $\sqrt{T}(\widehat{\Theta} - \Theta) = \mathcal{O}_P(1)$, see Lütkepohl (1993). By Theorem 3, we have $\sqrt{T}(\widetilde{\Theta} - \widehat{\Theta}) = o_P(1)$.

APPENDIX: PROOFS OF THEOREMS

A.1 Proof of Theorem 1

We use the Newton-Kantorovich theorem to prove the theorem. The statement of the theorem may be found in Kantorovich and Akilov (1982), for example.

Suppose that $\sum_{t=1}^{T} \| \mathcal{Z}_t^{(0)\top}\mathcal{A}^{(0)} - \widehat{\mathcal{Z}}_t^\top\widehat{\mathcal{A}} \|^2 \leq r$ for some $r > 0$, which will be chosen later. With the Frobenius norm $\|M\|$ for a matrix M, we get

$$\begin{aligned}
\| \mathcal{A}^{(0)} - \widehat{\mathcal{A}} \|^2 \leq{}& \left\| \left(\sum_{t=1}^{T} \mathcal{Z}_t^{(0)} \mathcal{Z}_t^{(0)\top} \right)^{-1} \right\|^2 \cdot \\
& \left\| \sum_{t=1}^{T} \mathcal{Z}_t^{(0)} \mathcal{Z}_t^{(0)\top} (\mathcal{A}^{(0)} - \widehat{\mathcal{A}}) \right\|^2 \\
={}& \left\| \left(\sum_{t=1}^{T} \mathcal{Z}_t^{(0)} \mathcal{Z}_t^{(0)\top} \right)^{-1} \right\|^2 \cdot \left\| \sum_{t=1}^{T} \mathcal{Z}_t^{(0)} \mathcal{Z}_t^{(0)\top} \mathcal{A}^{(0)} \right. \\
& \left. - \sum_{t=1}^{T} \mathcal{Z}_t^{(0)} \widehat{\mathcal{Z}}_t^\top \widehat{\mathcal{A}} \right\|^2 \leq \left\| \left(\sum_{t=1}^{T} \mathcal{Z}_t^{(0)} \mathcal{Z}_t^{(0)\top} \right)^{-1} \right\|^2 \\
& \times \left(\sum_{t=1}^{T} \left\| \mathcal{Z}_t^{(0)} \mathcal{Z}_t^{(0)\top} \mathcal{A}^{(0)} - \mathcal{Z}_t^{(0)} \widehat{\mathcal{Z}}_t^\top \widehat{\mathcal{A}} \right\| \right)^2 \leq \\
& r \left\| \left(\sum_{t=1}^{T} \mathcal{Z}_t^{(0)} \mathcal{Z}_t^{(0)\top} \right)^{-1} \right\|^2 \left(\sum_{t=1}^{T} \| \mathcal{Z}_t^{(0)} \|^2 \right) \\
\equiv{}& r c_1^2.
\end{aligned} \quad (A.1)$$

For a matrix M, define $\|M\|_2 = \sup_{\|x\|=1} \|Mx\|$. It is known that $\|M\|_2 \leq \|M\|$. We get

$$\begin{aligned}
\| \widehat{\mathcal{A}}^\top(\mathcal{Z}_t^{(0)} - \widehat{\mathcal{Z}}_t) \| \geq{}& \| \widehat{\mathcal{A}} \|_2^{-1} \\
& \cdot \| (\widehat{\mathcal{A}}\widehat{\mathcal{A}}^\top)^{-1} \|^{-1} \cdot \| \mathcal{Z}_t^{(0)} - \widehat{\mathcal{Z}}_t \|,
\end{aligned} \quad (A.2)$$

$$\begin{aligned}
\| (\mathcal{Z}_t^{(0)} - \widehat{\mathcal{Z}}_t)^\top \widehat{\mathcal{A}} \| \leq{}& \| \mathcal{Z}_t^{(0)\top}(\widehat{\mathcal{A}} - \mathcal{A}^{(0)}) \| + \| \mathcal{Z}_t^{(0)\top}\mathcal{A}^{(0)} - \\
& \widehat{\mathcal{Z}}_t^\top \widehat{\mathcal{A}} \| \leq \| \mathcal{Z}_t^{(0)} \| \cdot \| \widehat{\mathcal{A}} - \mathcal{A}^{(0)} \| + \| \mathcal{Z}_t^{(0)\top}\mathcal{A}^{(0)} - \widehat{\mathcal{Z}}_t^\top \widehat{\mathcal{A}} \|.
\end{aligned} \quad (A.3)$$

The two inequalities (A.2) and (A.3) together with (A.1) give

$$\begin{aligned}
\| \mathcal{Z}^{(0)} - \widehat{\mathcal{Z}} \|^2 \leq{}& 2r \| \widehat{\mathcal{A}} \|_2^2 \cdot \| (\widehat{\mathcal{A}}\widehat{\mathcal{A}}^\top)^{-1} \|^2 \\
& \times \left(1 + c_1 \sum_{t=1}^{T} \| \mathcal{Z}_t^{(0)} \|^2 \right) \equiv r c_2^2.
\end{aligned} \quad (A.4)$$

Because $F'(\alpha, z)$ is quadratic in (α, z), there exists $0 < c_3 < \infty$ for any compact set D in $\mathbb{R}^{K(L+1)+TL}$ such that $\|F'(\alpha', z') - F'(\alpha, z)\|_2 \leq c_3 \|(\alpha'^\top, z'^\top)^\top - (\alpha^\top, z^\top)^\top\|$ for all $(\alpha^\top, z^\top)^\top$, $(\alpha'^\top, z'^\top)^\top \in D$. Let $c_4 = ||F'_*(\alpha^{(0)}, Z^{(0)})^{-1}||_2 < \infty$. Because F is continuous and $F(\hat{\alpha}, \widehat{Z}) = 0$, there exists $r' > 0$ such that, if $\| \alpha^{(0)} - \hat{\alpha} \| + \| Z^{(0)} - \widehat{Z} \| \leq r'$, then

$$\| F'_*(\alpha^{(0)}, Z^{(0)})^{-1} F(\alpha^{(0)}, Z^{(0)}) \| \leq \frac{\gamma}{2c_3c_4}.$$

By the Newton-Kantorovich theorem,

$$\| \alpha^{(k)} - \hat{\alpha} \| + \| Z^{(k)} - \widehat{Z} \| \leq C_1 2^{-(k-1)} \gamma^{2^k - 1} \quad (A.5)$$

for some $C_1 > 0$. This gives that if $\| \alpha^{(0)} - \hat{\alpha} \| + \| Z^{(0)} - \widehat{Z} \| \leq r'$, then

$$\begin{aligned}
\sum_{t=1}^{T} \| \mathcal{Z}_t^{(k)\top}\mathcal{A}^{(k)} - \widehat{\mathcal{Z}}_t^\top \widehat{\mathcal{A}} \|^2 \leq{}& C_2(\| \alpha^{(k)} - \hat{\alpha} \|^2 + \\
& \| Z^{(k)} - \widehat{Z} \|^2) \leq C 2^{-2(k-1)} \gamma^{2(2^k - 1)}
\end{aligned}$$

for some C, $C_2 > 0$. We take $r = (c_1 + c_2)^{-2} r'^2$. Then, by (A.1) and (A.4), $\| \alpha^{(0)} - \hat{\alpha} \| + \| Z^{(0)} - \widehat{Z} \| \leq r'$ if $\sum_{t=1}^{T} \| \mathcal{Z}_t^{(0)\top}\mathcal{A}^{(0)} - \widehat{\mathcal{Z}}_t^\top \widehat{\mathcal{A}} \|^2 \leq r$. This completes the proof of the theorem.

A.2 Proof of Theorem 2

For functions $g(t, x)$ we define the norms $\| g \|_1^2 = (1/TJ)\sum_{t=1}^{T}\sum_{j=1}^{J} g(t, X_{t,j})^2$, $\|g\|_2^2 = (1/T)\sum_{t=1}^{T} \int g(t,x)^2 f_t(x)\, dx$, and $\| g \|_3^2 = (1/T)\sum_{t=1}^{T} \int g(t,x)^2\, dx$. Note that because of Assumption (A2) the last two norms are equivalent. Thus, for the statement of the theorem we have to show for $\Delta(t,x) = (\widehat{\mathcal{Z}}_t^\top \widehat{\mathcal{A}} - \mathcal{Z}_t^\top \mathcal{A}^*)\psi(x)$ that

$$\| \Delta \|_2^2 = \mathcal{O}_P(\rho^2 + \delta_K^2). \quad (A.6)$$

We start by showing that

$$\| \Delta \|_1^2 = \mathcal{O}_P([(K+T)\log(JTM_T)]/(JT) + \delta_K^2). \quad (A.7)$$

For this aim we apply Theorem 10.11 in Van de Geer (2000) that treats rates of convergence for least squares estimators on sieves. In our case we have the following sieve: $\mathcal{G}_T^* = \{g\colon \{1, \ldots, T\} \times [0,1]^d \to \mathbb{R}, g(t,x) = (1, z_t^\top)\mathcal{A}\psi(x)$ for an $(L+1) \times K$ matrix $\mathcal{A}$ and $z_t \in \mathbb{R}^L$ with the following properties: $|(1, z_t^\top)\mathcal{A}\psi(x)| \leq M_T$ for $1 \leq t \leq T$ and $x \in [0,1]^d\}$. With a constant C the δ-entropy $H_T(\delta, \mathcal{G}_T^*)$ of $\mathcal{G}_T^*$ with respect to the empirical norm $\|g\|_1$ is bounded by

$$H_T(\delta, \mathcal{G}_T^*) \leq CT\log(M_T/\delta) + CK \log(KM_T/\delta). \quad (A.8)$$

For the proof of (A.8) note first that each element $g(t,x) = (1, z_t^\top)\mathcal{A}\psi(x)$ of $\mathcal{G}_T^*$ can be chosen such that $T^{-1}\sum_{t=1}^T z_t z_t^\top$ is equal to the $L \times L$ identity matrix I_L. Then the bound $|(1, z_t^\top)\mathcal{A}\psi(x)| \le M_T$ implies that $\| \mathcal{A}\psi(x) \| \le M_T$. For the proof of (A.8) we use that the (δ/M_T)-entropy of a unit ball in $\mathbb{R}^T$ is of order $\mathcal{O}(T \log(M_T/\delta))$ and that the δ-entropy with respect to the sup-norm for functions $\mathcal{A}\psi(x)$ with $\| \mathcal{A}\psi(x) \| \le M_T$ is of order $\mathcal{O}(K \log(KM_T/\delta))$. In the last entropy bound we used that for each x it holds that $||\psi(x)|| \le K^{1/2}$. These two entropy bounds imply (A.8). Application of Theorem 10.11 in Van de Geer (2000) gives (A.7).

We now show that (A.7) implies (A.6). For this aim note first that by Bernstein's inequality for $a,\ d > 0,\ g \in \mathcal{G}_T^*$ with $||g||_2^2 \le d$

$$P(|\| g \|_1^2 - \| g \|_2^2 | \ge a) \le 2 \exp\left(-\frac{a^2 JT}{2(a+d)M_T^2}\right). \tag{A.9}$$

Furthermore, for $g, h \in \mathcal{G}_T^*$ it holds with constants C, C' that

$$\begin{aligned}&|\| g \|_1^2 - \| h \|_1^2 | \le CK\left(T^{-1}\sum_{t=1}^T \| e_t - f_t \|^2\right)^{1/2} \\ &\left(T^{-1}\sum_{t=1}^T \| e_t + f_t \|^2\right)^{1/2} \le C'K \| g - h \|_2 (\| g \|_2 + \| h \|_2),\end{aligned} \tag{A.10}$$

where e_t and f_t are chosen such that $g(x,t) = e_t^\top \psi(x)$ and $h(x,t) = f_t^\top \psi(x)$. From (A.9) and (A.10) we get with a constant $C > 0$ that for $d = 1, 2, \ldots$

$$\begin{aligned}&P\Bigg(\sup_{g \in \mathcal{G}_T^*, d\rho^2 \le \|g\|_2^2 \le (d+1)\rho^2} |\| g \|_1^2 - \| g \|_2^2 | \ge d\rho^2/2\Bigg) \\ &\le C \exp((C + K + T) \log(dKM_T) - d\rho^2 JT/[20M_T^2]).\end{aligned}$$

By summing these inequalities over $d \ge 1$ we get $||\Delta||_2^2 \le \rho^2$ or

$$||\Delta||_2^2 \le |||\Delta||_1^2 - ||\Delta||_2^2| + ||\Delta||_1^2 \le ||\Delta||_2^2/2 + ||\Delta||_1^2$$

with probability tending to one. This shows Equation (A.6) and concludes the proof of Theorem 2.

A.3 Proof of Theorem 3

We will prove the first equation of the theorem for $h \ne 0$. The second equation follows from the first equation. We first prove that the matrix $T^{-1}\sum_{t=1}^T \mathcal{Z}_{c,t}\widehat{\mathcal{Z}}_{c,t}^\top$ is invertible, where $\mathcal{Z}_{c,t}^\top = (1, Z_{c,t}^\top)$, $\widehat{\mathcal{Z}}_{c,t}^\top = (1, \widehat{Z}_{c,t}^\top)$, and $\widehat{Z}_{c,t} = \widehat{Z}_t - T^{-1}\sum_{s=1}^T \widehat{Z}_s$. This implies that $T^{-1}\sum_{t=1}^T Z_{c,t}\widehat{Z}_{c,t}^\top$ is invertible. Suppose that the assertion is not true. We can choose a random vector e such that $\|e\| = 1$ and $e^\top \sum_{t=1}^T \mathcal{Z}_{c,t}\widehat{\mathcal{Z}}_{c,t}^\top = 0$. Let $\widehat{A}$ and A^* be the $L \times K$ matrices that are obtained by deleting the first rows of $\widehat{\mathcal{A}}$ and $\mathcal{A}^*$, respectively. Let $\widehat{\mathcal{A}}_c$ and $\mathcal{A}_c^*$ be the matrices obtained from $\widehat{\mathcal{A}}$ and $\mathcal{A}^*$ by replacing their first rows by $\widehat{\alpha}_0^\top + (T^{-1}\sum_{t=1}^T \widehat{Z}_t)^\top \widehat{A}$ and $\alpha_0^{*\top} + (T^{-1}\sum_{t=1}^T Z_t)^\top A^*$, respectively. By definition, it follows that

$$\widehat{\mathcal{Z}}_t^\top \widehat{\mathcal{A}} = \widehat{\mathcal{Z}}_{c,t}^\top \widehat{\mathcal{A}}_c, \quad \mathcal{Z}_t^\top \mathcal{A}^* = \mathcal{Z}_{c,t}^\top \mathcal{A}_c^*. \tag{A.11}$$

Note that

$$\begin{aligned}&\left\| T^{-1}\sum_{t=1}^T \mathcal{Z}_{c,t}\widehat{\mathcal{Z}}_{c,t}^\top \widehat{\mathcal{A}}_c - T^{-1}\sum_{t=1}^T \mathcal{Z}_{c,t}\mathcal{Z}_{c,t}^\top \mathcal{A}_c^* \right\| \\ &\le T^{-1}\sum_{t=1}^T \left\| \mathcal{Z}_{c,t}\left(\widehat{\mathcal{Z}}_{c,t}^\top \widehat{\mathcal{A}}_c - \mathcal{Z}_{c,t}^\top \mathcal{A}_c^*\right)\right\| \\ &\le \left(T^{-1}\sum_{t=1}^T \left\| \mathcal{Z}_{c,t}\right\|^2\right)^{1/2} \left(T^{-1}\sum_{t=1}^T \left\| \widehat{\mathcal{Z}}_t^\top \widehat{\mathcal{A}} - \mathcal{Z}_t^\top \mathcal{A}^*\right\|^2\right)^{1/2} \\ &= \mathcal{O}_P(\rho + \delta_K),\end{aligned} \tag{A.12}$$

because of Assumption (A6) and Theorem 2. Thus with $f = T^{-1}\sum_{t=1}^T \mathcal{Z}_{c,t}\mathcal{Z}_{c,t}^\top e$, we obtain

$$\begin{aligned}\| f^\top m \| &= \| f^\top (\mathcal{A}_c^* \psi) \| + \mathcal{O}_P(T^{-1/2} + \delta_K) \\ &= \left\| e^\top T^{-1}\sum_{t=1}^T \mathcal{Z}_{c,t}\widehat{\mathcal{Z}}_{c,t}^\top \widehat{\mathcal{A}}_c \psi \right\| + \mathcal{O}_P(T^{-1/2} + \rho + \delta_K) \\ &= \mathcal{O}_P(T^{-1/2} + \rho + \delta_K).\end{aligned}$$

This implies that $m_0, \ldots, m_d$ are linearly dependent, contradicting to Assumption (A10).

Let $\widetilde{B}$ be the matrix given at (8) with B defined as in (9). Define $\widetilde{\mathcal{Z}}_{c,t} = \widetilde{B}^\top \widehat{\mathcal{Z}}_{c,t}$ and $\widetilde{\mathcal{A}}_c = \widetilde{B}^{-1}\widehat{\mathcal{A}}_c$. Then $\widetilde{\mathcal{Z}}_{c,t}^\top \widetilde{\mathcal{A}}_c = \widehat{\mathcal{Z}}_{c,t}^\top \widehat{\mathcal{A}}_c$ and $T^{-1}\sum_{t=1}^T \mathcal{Z}_{c,t}\widetilde{\mathcal{Z}}_{c,t}^\top = T^{-1}\sum_{t=1}^T \mathcal{Z}_{c,t}\mathcal{Z}_{c,t}^\top$. This gives with (A.12)

$$\begin{aligned}\left\|\widetilde{\mathcal{A}}_c - \mathcal{A}_c^*\right\| &= \left\| T^{-1}\sum_{t=1}^T \mathcal{Z}_{c,t}\mathcal{Z}_{c,t}^\top (\widetilde{\mathcal{A}}_c - \mathcal{A}_c^*)\right\| \mathcal{O}_P(1) \\ &= \left\| T^{-1}\sum_{t=1}^T \mathcal{Z}_{c,t}\widetilde{\mathcal{Z}}_{c,t}^\top \widetilde{\mathcal{A}}_c - T^{-1}\sum_{t=1}^T \mathcal{Z}_{c,t}\mathcal{Z}_{c,t}^\top \mathcal{A}_c^*\right\| \mathcal{O}_P(1) \\ &= \mathcal{O}_P(\rho + \delta_K).\end{aligned} \tag{A.13}$$

Because of Theorem 2 this implies

$$\left\|\widetilde{\mathcal{A}} - \mathcal{A}^*\right\| = \mathcal{O}_P(\rho + \delta_K). \tag{A.14}$$

Define $\widetilde{\mathcal{Z}}_{c,t}$ by $\widetilde{\mathcal{Z}}_{c,t}^\top = (1, \widetilde{Z}_{c,t}^\top)$. Note that $\widetilde{\mathcal{Z}}_{c,t} = B^\top \widehat{\mathcal{Z}}_{c,t}$. Also, define $\widetilde{A} = B^{-1}\widehat{A}$, which equals $\widetilde{\mathcal{A}}_c$ without the first row. From (A10), (A5), (A.14), and Theorem 2, we get

$$\begin{aligned}&T^{-1}\sum_{t=1}^T \left\|\widetilde{Z}_t - Z_t\right\|^2 = T^{-1}\sum_{t=1}^T \left\|\widetilde{\mathcal{Z}}_t - \mathcal{Z}_t\right\|^2 \\ &= T^{-1}\sum_{t=1}^T \left\|\widetilde{\mathcal{Z}}_t^\top (m_0, \ldots, m_L)^\top - \mathcal{Z}_t^\top (m_0, \ldots, m_L)^\top\right\|^2 \mathcal{O}_P(1) \\ &= T^{-1}\sum_{t=1}^T \left\|\widetilde{Z}_t^\top A^* - \widetilde{Z}_t^\top \widetilde{A}\right\|^2 \mathcal{O}_P(1) + T^{-1}\sum_{t=1}^T \left\|\widetilde{Z}_t^\top \widetilde{A} - Z_t^\top A^*\right\|^2 \\ &\quad \times \mathcal{O}_P(1) + \mathcal{O}_P(\delta_K^2) \\ &\le T^{-1}\sum_{t=1}^T \| \widetilde{Z}_t - Z_t\|^2 \| \widetilde{A} - A^*\|^2 \mathcal{O}_P(1) + T^{-1}\sum_{t=1}^T \| Z_t\|^2 \\ &\quad \times \| \widetilde{A} - A^* \|^2 \mathcal{O}_P(1) \\ &+ T^{-1}\sum_{t=1}^T \| \widetilde{Z}_t^\top \widetilde{A} - Z_t^\top A^* \|^2 \mathcal{O}_P(1) + \mathcal{O}_P(\rho^2 + \delta_K^2) \\ &= \mathcal{O}_P(\rho^2 + \delta_K^2).\end{aligned} \tag{A.15}$$

From Equation (A.15) one gets

$$T^{-1}\sum_{t=1}^{T}\left\|\widetilde{Z}_{c,t}-Z_{c,t}\right\|^2=\mathcal{O}_P(\rho^2+\delta_K^2). \tag{A.16}$$

We will show that for $h\neq 0$

$$T^{-1}\sum_{t=h+1}^{T}\{(\widetilde{Z}_{c,t+h}-Z_{c,t+h})-(\widetilde{Z}_{c,t}-Z_{c,t})\}Z_{c,t}^{\top}=o_P(T^{-1/2}). \tag{A.17}$$

This implies the first statement of Theorem 3, because by (A.16)

$$T^{-1}\sum_{t=-h+1}^{T}(\widetilde{Z}_{c,t}-Z_{c,t})(\widetilde{Z}_{c,t+h}^{\top}-Z_{c,t+h}^{\top})=\mathcal{O}_P(\rho^2+\delta_K^2)=\mathcal{O}_P(T^{-1/2}).$$

For the proof of (A.17), let $\widetilde{\alpha}_c$ be the stack form of $\widetilde{\mathcal{A}}_c$ and $\widetilde{\alpha}_{c,0}^{\top}$ be its first row. Using the representation (6) and the first identity of (A.11), it can be verified that

$$\widetilde{Z}_{c,t}=\widetilde{S}_{t,Z}^{-1}J^{-1}\sum_{j=1}^{J}\{Y_{t,j}\widetilde{A}\psi(X_{t,j})-\widetilde{A}\psi(X_{t,j})\psi(X_{t,j})^{\top}\widetilde{\alpha}_{c,0}\}, \tag{A.18}$$

$$\widetilde{\alpha}_c=\widetilde{\mathcal{S}}_{\alpha}^{-1}T^{-1}J^{-1}\sum_{t=1}^{T}\sum_{j=1}^{J}\{\psi(X_{t,j})\otimes\widetilde{\mathcal{Z}}_{c,t}\}Y_{t,j}, \tag{A.19}$$

where $\widetilde{S}_{t,Z}=J^{-1}\sum_{j=1}^{J}\widetilde{A}\psi(X_{t,j})\psi(X_{t,j})^{\top}\widetilde{A}^{\top}$ and $\widetilde{\mathcal{S}}_{\alpha}=T^{-1}J^{-1}\sum_{t=1}^{T}\sum_{j=1}^{J}\{\psi(X_{t,j})\otimes\widetilde{\mathcal{Z}}_{c,t}\}\{\psi(X_{t,j})\otimes\widetilde{\mathcal{Z}}_{c,t}\}^{\top}$. Define $\widetilde{\mathcal{S}}_{t,Z}$ as $\widetilde{S}_{t,Z}$ with $\widetilde{\mathcal{A}}_c$ replacing $\widetilde{A}$. Also, define $\mathcal{S}_{t,Z}=\mathcal{A}_c^{*}E\{\psi(X_{t,j})\psi(X_{t,j})^{\top}\}\mathcal{A}_c^{*\top}$, $S_{t,Z}=A^{*}E\left\{\psi(X_{t,j})\psi(X_{t,j})^{\top}\right\}A^{*\top}$ and

$$\mathcal{S}_{\alpha}=T^{-1}\sum_{t=1}^{T}E[\{\psi(X_{t,j})\otimes\mathcal{Z}_{c,t}\}\{\psi(X_{t,j})\otimes\mathcal{Z}_{c,t}\}^{\top}|Z_t].$$

Let $\gamma=T^{-1/2}(\rho+\delta_K)^{-1}$. We argue that

$$\sup_{1\le t\le T}\|\widetilde{\mathcal{S}}_{t,Z}-\mathcal{S}_{t,Z}\|=o_P(\gamma),\quad \|\widetilde{\mathcal{S}}_{\alpha}-\mathcal{S}_{\alpha}\|=o_P(\gamma). \tag{A.20}$$

We show the first part of (A.20). The second part can be shown similarly. To prove the first part it suffices to show that, uniformly for $1\le t\le T$,

$$J^{-1}\sum_{j=1}^{J}\mathcal{A}_c^{*}[\psi(X_{t,j})\psi(X_{t,j})^{\top}-E\{\psi(X_{t,j})\psi(X_{t,j})^{\top}\}](\widetilde{\mathcal{A}}_c-\mathcal{A}_c^{*})^{\top}=o_P(\gamma), \tag{A.21}$$

$$J^{-1}\sum_{j=1}^{J}(\widetilde{\mathcal{A}}_c-\mathcal{A}_c^{*})[\psi(X_{t,j})\psi(X_{t,j})^{\top}-E\{\psi(X_{t,j})\psi(X_{t,j})^{\top}\}](\widetilde{\mathcal{A}}_c-\mathcal{A}_c^{*})^{\top}=o_P(\gamma), \tag{A.22}$$

$$J^{-1}\sum_{j=1}^{J}\mathcal{A}_c^{*}[\psi(X_{t,j})\psi(X_{t,j})^{\top}-E\{\psi(X_{t,j})\psi(X_{t,j})^{\top}\}]\mathcal{A}_c^{*\top}=o_P(\gamma), \tag{A.23}$$

$$J^{-1}\sum_{j=1}^{J}\mathcal{A}_c^{*}E\{\psi(X_{t,j})\psi(X_{t,j})^{\top}\}(\widetilde{\mathcal{A}}_c-\mathcal{A}_c^{*})^{\top}=o_P(\gamma), \tag{A.24}$$

$$J^{-1}\sum_{j=1}^{J}(\widetilde{\mathcal{A}}_c-\mathcal{A}_c^{*})E\{\psi(X_{t,j})\psi(X_{t,j})^{T}\}(\widetilde{\mathcal{A}}_c-\mathcal{A}_c^{*})^{\top}=o_P(\gamma). \tag{A.25}$$

The proof of (A.23)–(A.25) follows by simple arguments. We now show (A.21). Claim (A.22) can be shown similarly. For the proof of (A.21) we use Bernstein's inequality for the following sum:

$$P\left(|\sum_{j=1}^{J}W_j|>x\right)\le 2\exp\left(-\frac{1}{2}\frac{x^2}{V+Mx/3}\right). \tag{A.26}$$

Here for a value of t with $1\le t\le T$, the random variable W_j is an element of the $(L+1)\times 1$-matrix $S=J^{-1}\mathcal{A}_c^{*}[\psi(X_{t,j})\psi(X_{t,j})^{\top}e-E\{\psi(X_{tj})\psi(X_{tj})^{\top}e\}]$ where $e\in\mathbb{R}^K$ with $\|e\|=1$. In (A.26), V is an upper bound for the variance of $\sum_{j=1}^{J}W_j$ and M is a bound for the absolute values of W_j (i.e. $|W_j|\le M$ for $1\le j\le J$, a.s.). With some constants C_1 and C_2 that do not depend on t and the row number we get $V\le C_1J^{-1}$ and $M\le C_2K^{1/2}J^{-1}$. Application of Bernstein's inequality gives that, uniformly for $1\le t\le T$ and $e\in\mathbb{R}^K$ with $\|e\|=1$, all $(L+1)$ elements of S are of order $o_P(\gamma)$. This shows claim (A.21).

From (A.13), (A.15), (A.18), (A.19), and (A.20) it follows that uniformly for $1\le t\le T$,

$$\begin{aligned}\widetilde{Z}_{c,t}-Z_{c,t}&=S_{t,Z}^{-1}J^{-1}\sum_{j=1}^{J}\varepsilon_{t,j}A^{*}\psi(X_{t,j})+S_{t,Z}^{-1}J^{-1}\sum_{j=1}^{J}\varepsilon_{t,j}\\&\quad\times(\widetilde{A}-A^{*})\psi(X_{t,j})\\&+S_{t,Z}^{-1}J^{-1}\sum_{j=1}^{J}(\widetilde{A}-A^{*})\psi(X_{t,j})\psi(X_{t,j})^{\top}\mathcal{A}_c^{*\top}\mathcal{Z}_{c,t}+o_P(T^{-1/2})\\&\equiv\Delta_{t,1,Z}+\Delta_{t,2,Z}+\Delta_{t,3,Z}+o_P(T^{-1/2}).\end{aligned} \tag{A.27}$$

For the proof of the theorem it remains to show that for $1\le j\le 3$

$$T^{-1}\sum_{t=-h+1}^{T}(\Delta_{t+h,j,Z}-\Delta_{t,j,Z})Z_{c,t}^{\top}=o_P(T^{-1/2}). \tag{A.28}$$

This can be easily checked for $j=1$. For $j=2$ it follows from $\|\widetilde{A}-A^{*}\|=\mathcal{O}(\rho+\delta_k)$ and

$$E\left\{\|T^{-1}J^{-1}\sum_{t=1}^{T}\sum_{j=1}^{J}\varepsilon_{t,j}S_{t,Z}^{-1}\mathcal{M}\psi(X_{t,j})\|^2\right\}=\mathcal{O}(KJ^{-1}T^{-1})$$

for any $L\times K$ matrix $\mathcal{M}$ with $\|\mathcal{M}\|=1$. For the proof of (A.28) for $j=3$, it suffices to show that

$$T^{-1}\sum_{t=1}^{T+h}\Delta_{t,j,Z}(Z_{c,t-h}-Z_{c,t})^{\top}=o_P(T^{-1/2}). \tag{A.29}$$

We note first that for $1\le l\le L$

$$T^{-1}\sum_{t=1}^{T+h}\Delta_{t,3,Z}(Z_{c,t-h,l}-Z_{c,t,l})=T^{-1}J^{-1}\sum_{t=1}^{T+h}\sum_{j=1}^{J}\left\{\left(V_{h,t}^{\top}\mathcal{A}_c^{*}\psi(X_{t,j})\psi(X_{t,j})^{\top}\right)\otimes S_{t,Z}^{-1}\right\}(\widetilde{\alpha}-\alpha^{*}),$$

where $V_{h,t} = (\underset{\sim}{Z}_{c,t-h,l} - Z_{c,t,l})\mathcal{Z}_{c,t}$, and $\widetilde{\alpha}$ and α^* denote the stack forms of $\widetilde{A}$ and A^*, respectively. For the proof of (A.29) it suffices to show

$$T^{-1}J^{-1}\sum_{t=1}^{T+h}\sum_{j=1}^{J}\{(E[V_{h,t}]^\top \mathcal{A}_c^*\psi(X_{t,j})\psi(X_{t,j})^\top)\otimes S_{t,Z}^{-1}\} \times(\widetilde{\alpha}-\alpha^*) = o_P(T^{-1/2}), \tag{A.30}$$

$$\left\| T^{-1}J^{-1}\sum_{t=1}^{T+h}\sum_{j=1}^{J}\{(\{V_{h,t}-E[V_{h,t}]\}^\top \mathcal{A}_c^*\psi(X_{t,j})\psi(X_{t,j})^\top) \otimes S_{t,Z}^{-1}\}\right\|^2 = \mathcal{O}_P(KJ^{-1}T^{-1}). \tag{A.31}$$

Claim (A.31) can be easily shown by calculating the expectation of the left hand side of (A.31) and by using the mixing condition at Assumption (A9). For a proof of (A.30) we remark first that by construction

$$0 = T^{-1}\sum_{t=1}^{T}(\widetilde{Z}_{c,t} - Z_{c,t})Z_{c,t}^T.$$

Using (A.27) and similar arguments as in the proof of (A.28) for $j = 1, 2$ we get that

$$T^{-1}\sum_{t=1}^{T}\Delta_{t,3,Z}Z_{c,t}^T = T^{-1}J^{-1}\sum_{t=1}^{T}\sum_{j=1}^{J}\left\{(Z_{c,t}\mathcal{Z}_{c,t}^\top\mathcal{A}_c^*\psi(X_{t,j})\psi(X_{t,j})^\top)\otimes S_{t,Z}^{-1}\right\}(\widetilde{\alpha}-\alpha^*) = o_P(T^{-1/2}).$$

As in the proof of (A.31) one can show that

$$\left\| T^{-1}J^{-1}\sum_{t=1}^{T+h}\sum_{j=1}^{J}\{(\{Z_{c,t}\mathcal{Z}_{c,t}^\top - E[Z_{c,t}\mathcal{Z}_{c,t}^\top]\}\mathcal{A}_c^*\psi(X_{t,j})\psi(X_{t,j})^\top \otimes S_{t,Z}^{-1}\}\right\|^2 = \mathcal{O}_P(KJ^{-1}T^{-1}).$$

The last two equalities imply that

$$T^{-1}J^{-1}\sum_{t=1}^{T}\sum_{j=1}^{J}\{(E[Z_{c,t}\mathcal{Z}_{c,t}^\top]\mathcal{A}_c^*\psi(X_{t,j})\psi(X_{t,j})^\top)\otimes S_{t,Z}^{-1}\} \times(\widetilde{\alpha}-\alpha^*) = o_P(T^{-1/2}).$$

Because of Assumption (A9) this implies claim (A.29) and concludes the proof of Theorem 3.

[Received June 2007. Revised August 2008.]

REFERENCES

Biswal, B., Yetkin, F., Haughton, V., and Hyde, J. (1995), "Functional Connectivity in the Motor Cortex of Resting Human Brain Using Echo-Planar MRI," *Magnetic Resonance in Medicine,* 34, 537–541.

Black, F., and Scholes, M. (1973), "The Pricing of Options and Corporate Liabilities," *The Journal of Political Economy,* 81, 637–654.

Brüggemann, R., Lütkepohl, H., and Saikkonen, P. (2006), "Residual Autocorrelation Testing for Vector Error Correction Models," *Journal of Econometrics,* 134, 579–604.

Brumback, B., and Rice, J. A. (1998), "Smooting Spline Models for the Analysis of Nested and Crossed Samples of Curves," *Journal of the American Statistical Association,* 93, 961–994.

Connor, G., Hagmann, M., and Linton, O. (2007). *Efficient Semiparametric Estimation of the Fama-French Model and Extensions*, Preprint.

Connor, G., and Linton, O. (2007), "Semiparametric Estimation of a Characteristic-based Factor Model of Stock Returns," *Journal of Empirical Finance,* 14, 694–717.

Cont, R., and da Fonseca, J. (2002), "The Dynamics of Implied Volatility Surfaces," *Quantitative Finance,* 2, 45–60.

de Boor, C. (2001). *A Practical Guide to Splines*, Berlin, Heidelberg: Springer-Verlag.

Diebold, F. X., and Li, C. (2006), "Forecasting the Term Structure of Government Bond Yields," *Journal of Econometrics,* 130, 337–364.

Fama, E. F., and French, K. R. (1992), "The Cross-Section of Expected Stock Returns," *The Journal of Finance,* 47, 427–465.

Fan, J., Yao, Q., and Cai, Z. (2003), "Adaptive Varying-Coefficient Linear Models," *Journal of the Royal Statistical Society:* Series B, 65, 57–80.

Fengler, M. R., Härdle, W., and Mammen, E. (2007), "A Semiparametric Factor Model for Implied Volatility Surface Dynamics," *Journal of Financial Econometrics,* 5, 189–218.

Forni, M., Hallin, M., Lippi, M., and Reichlin, L. (2000), "The Generalized Dynamic Factor Model: Identification and Estimation," *The Review of Economics and Statistics,* 82, 540–554.

Forni, M., and Lippi, M. (2001), "The Generalized Factor Model: Representation Theory," *Econometric Theory,* 17, 1113–1141.

Gasser, T., Möcks, R., and Verleger, R. (1983), "Selavco: A Method to Deal With Trial-to-Trial Variability of Evoked Potential," *Electroencephalography and Clinical Neurophysiology,* 55, 717–723.

Hafner, R. (2004). *Stochastic Implied Volatility*, Berlin: Springer.

Hallin, M., and Liska, R. (2007), "Determining the Number of Factors in the Generalized Dynamic Factor Model," *Journal of the American Statistical Association,* 102, 603–617.

Hansen, L.H., Nielsen, B., and Nielsen, J.P. (2004). "Two Sided Analysis of the Variance With a Latent Time Series," Nuffield College Economic Working Paper 2004-W25, University of Oxford.

Hosking, J. R. M. (1980), "The Multivariate Portmanteau Statistic," *Journal of the American Statistical Association,* 75, 602–608.

Hosking, J. R. M. (1981), "Lagrange-Multiplier Tests of Multivariate Time-Series Models," *Journal of the Royal Statistical Society,* Series B, 43, 219–230.

Kantorovich, L. V., and Akilov, G. P. (1982). *Functional Analysis* (2nd ed.), Oxford, U.K: Pergamon Press.

Kauermann, G. (2000), "Modeling Longitudinal Data With Ordinal Response by Varying Coefficients," *Biometrics,* 56, 1692–1698.

Lee, R. D., and Carter, L. (1992), "Modeling and Forecasting the Time Series of U.S. Mortality," *Journal of the American Statistical Association,* 87, 659–671.

Logothetis, N., and Wandell, B. (2004), "Interpreting the Bold Signal," *Annual Review of Physiology,* 66, 735–769.

Lütkepohl, H. (1993). *Intorduction to Multiple Time Series Analysis*, Berlin, Heidelberg: Springer-Verlag.

Martinussen, T., and Scheike, T. (2000), "A Nonparametric Dynamic Additive Regression Model for Longitudinal Data," *Annals of Statistics,* 28, 1000–1025.

Nelson, C. R., and Siegel, A. F. (1987), "Parsimonious Modeling of Yield Curves," *Journal of Business,* 60, 473–489.

Peña, D., and Box, E. P. (1987), "Identifying a Simplifying Structure in Time Series," *Journal of the American Statistical Association,* 82, 836–843.

Stock, J. H., and Watson, M.W. (2005). "Implications of Dynamic Factor Models for VAR Analysis," NBER Working Papers 11467, National Bureau of Economic Research, Inc., available at http://ideas.repec.org/p/nbr/nberwo/11467.html.

Van de Geer, S. (2000). *Empirical Processes in M-Estimation*, Cambridge, U.K.: Cambridge University Press.

Vrahatis, M. N. (1989), "A Short Proof and a Generalization of Miranda's Existence Theorem," *Proceedings of the American Mathematical Society,* 107, 701–703.

Worsley, K., Liao, C., Aston, J., Petre, V., Duncan, G., Morales, F., and Evans, A. (2002), "A General Statistical Analysis for fMRI Data," *NeuroImange,* 15, 1–15.

Yang, L., Park, B. U., Xue, L., and Härdle, W. (2006), "Estimation and Testing for Varying Coefficients in Additive Models With Marginal Integration," *Journal of the American Statistical Association,* 101, 1212–1227.

A Generalized ARFIMA Process with Markov-Switching Fractional Differencing Parameter

Wen-Jen Tsay [1] and Wolfgang Karl Härdle [2] *

April 24, 2007

[1] The Institute of Economics, Academia Sinica, Taiwan

[2] CASE – Center for Applied Statistics and Economics

Humboldt-Universität zu Berlin,

Spandauer Straße 1, 10178 Berlin, Germany

Abstract

We propose a general class of Markov-switching-ARFIMA processes in order to combine strands of long memory and Markov-switching literature. Although the coverage of this class of models is broad, we show that these models can be easily estimated with the DLV algorithm proposed. This algorithm combines the Durbin-Levinson and Viterbi procedures. A Monte Carlo experiment reveals that the finite sample performance of the proposed algorithm for a simple mixture model of Markov-switching mean and ARFIMA$(1, d, 1)$ process is satisfactory. We apply the Markov-switching-ARFIMA models to the U.S. real interest rates, the Nile river level, and the U.S. unemployment rates, respectively. The results are all highly consistent with the conjectures made or empirical results found in the literature. Particularly, we confirm the conjecture in Beran and Terrin (1996) that the observations 1 to about 100 of the Nile river data seem to be more independent than the subsequent observations, and the value of differencing parameter is lower for the first 100 observations than for the subsequent data.

Key words: Markov chain; ARFIMA process; Viterbi algorithm; Long memory

JEL classification: C14, C22, C32, C52, C53, G12

*This research was supported by the Deutsche Forschungsgemeinschaft through the SFB 649 'Economic Risk'. Correspondence to Wen-Jen Tsay. The Institute of Economics, Academia Sinica, Taipei, Taiwan, R.O.C. Tel: (886-2) 2782-2791 ext. 296. Fax: (886-2) 2785-3946. E-Mail: wtsay@ieas.econ.sinica.edu.tw

1 Introduction

It is well known that many time series data exhibit long memory, or long-range dependence, including the Nile river level, *ex post* real interest rate, forward premium, and the dynamics of aggregate partisanship and macroideology. Among the many other examples that Beran (1994) gives the Nile river data has been known for its long memory behavior since ancient times, and this is one of the time series that led to the discovery of the Hurst effect (Hurst, 1951) and motivated Mandelbrot and his co-workers (Mandelbrot and van Ness, 1968; Mandelbrot and Wallis, 1969) to introduce fractional Gaussian noise to model long memory phenomenon.

Long range dependence also has been observed in financial data. As demonstrated by Ding et al. (1993), de Lima and Crato (1993) and Bollerslev and Mikkelsen (1996) that the volatility of most financial time series exhibits strong persistency and can be well described as a long memory process. Evidence of financial market volatility's strong persistency inspired Breidt et al. (1998) to propose a class of long memory stochastic volatility (LMSV) models. Deo et al. (2006) also show that the LMSV model is useful for forecasting realized volatility (RV) which is an important quantity in finance.

Figure 1 displays the yearly Nile river minima based on measurements at the Roda gauge near Cairo during the years 622-1284. Beran (1994, p.33) documents that "When one only looks at short time periods, then there seem to be cycles or local trend. However, looking at the whole series, there is no apparent persisting cycle." The changing pattern of the Nile river data leads Bhattacharya et al. (1983) to argue that the so-called Hurst effect can also be explained as if the observations are composed as the sum of a weakly dependent stationary process and a deterministic function. As a consequence it is important to distinguish between a long memory time series and a weakly dependent time series with change-points in the mean. This question has been intensively considered in the literature, including Künsch (1986) and Heyde and Dai (1996). Berkes et al. (2006) presents an overview about this strand of literature. Similarly, Diebold and Inoue (2001) shows that long memory also may be easily confused with a Markov-switching mean. Thus, most of the existing literature considers long memory as a competing modeling framework against the structural change and Markov-switching models.

The Nile river level time series is far more complicated than a pure long memory or

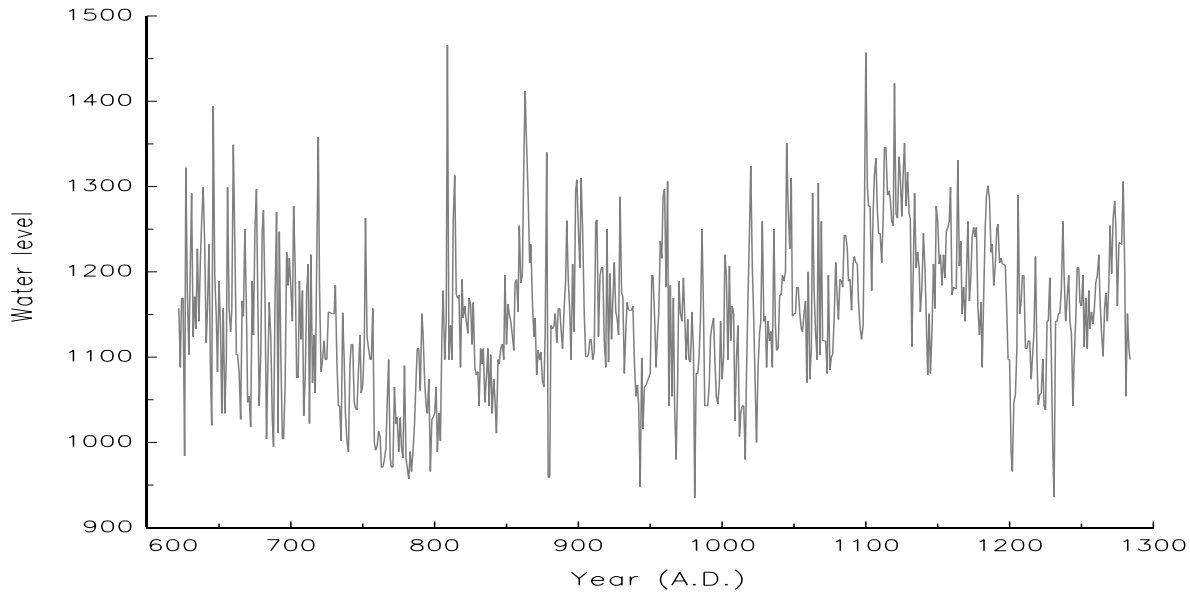


Figure 1: Yearly Nile river minima based on measurements at the Roda gauge near Cairo.

a weakly dependent time series with change-points in the mean to describe. Beran and Terrin (BT) (1996) suggest therefore that the Hurst parameter characterizing the yearly Nile river might change over time. When estimating the Nile river data with the autoregressive fractionally-integrated moving-average (ARFIMA) or $I(d)$ process introduced by Granger (1980), Granger and Joyeux (1980) and Hosking (1981), Beran and Terrin (1996, p.629) show that the data can be well fitted with an ARFIMA$(0, d, 0)$ model with $d = 0.4$, where the fractional differencing parameter d of ARFIMA process acts like the Hurst parameter H of fractional Gaussian noise in characterizing the hyperbolic decay of the autocovariance function of a long memory process. BT further claim that the observations 1 to about 100 seem to be more independent than the subsequent observations, and the value of the fractional differencing parameter might be *lower* for the first 100 observations than for the subsequent data. If this claim is right, then there should be a structural change in the long range persistence of the Nile river data around the year 720, and the Nile river data neither can be described with a pure long memory nor a weakly dependent time series with change-points in the mean.

The possible change of the differencing parameter stimulate BT to propose a statistic for testing the stability of the fractional differencing parameter. This testing statistic has been further discussed and extended in Horváth and Shao (1999) and Horváth (2001). However,

their methods can not identify the change points of the fractional differencing parameter. A Bayesian random persistent-shift (RPS) method for detecting structural change in the differencing parameter and the process level has been considered in Ray and Tsay (2002). Nevertheless, the RPS method is not built on the Markov-switching framework, thus may not fully characterize the cycling behavior of the data series, i.e., "seven years of great abundance" and "seven years of famine" — the Joseph effect named by Mandelbrot and van Ness (1968) and Mandelbrot and Wallis (1969).

The above considerations lead us to combine the long memory and Markov-switching literature into a unified framework. We introduce a Markov-switching-ARFIMA (MS-ARFIMA) process by extending the hidden Markov model. Given that the hidden Markov model has become extremely popular in speech recognition as shown in Juang and Rabiner (1991) and Qian and Titterington (1991), and in econometrics, finance, genetics, and neurophysiology as outlined in Robert et al. (2000), the MS-ARFIMA model provides a flexible modeling framework for many applications to these fields. Moreover, the research conducted in this paper also solve the puzzle raised by Diebold and Inoue (2001) by estimating the differencing parameter allowing for the parameters of interest are Markov-switching.

The remaining parts of this paper are arranged as follows: Section 2 presents the MS-ARFIMA process and the algorithms for estimating the parameters of interest. In Section 3 we consider the finite sample performance of the proposed algorithm under the simple mixture of a Markov-switching mean and an ARFIMA$(1, d, 1)$ process. We then apply the proposed methodology to the U.S. real interest rates, the Nile river data, and the U.S. unemployment rates in Section 4. Section 5 provides a conclusion.

2 Models and Main Results

The objective of this paper is to propose a general class of Markov-switching-ARFIMA processes in order to combine strands of long memory and Markov-switching literature. This class of models offers a rich dynamic mixture of a Markov chain and an $I(d)$ process.

Let $\{s_t\}_{t=1}^{T}$ be the latent sample path of an N-state Markov chain. At each time s_t can

assume only an integer value of $1, 2, \cdots, N$, and its transition probability matrix is

$$\mathcal{P} \equiv \begin{bmatrix} p_{11} & p_{21} & \cdots & p_{N1} \\ p_{12} & p_{22} & \cdots & p_{N2} \\ \vdots & \vdots & \ddots & \vdots \\ p_{1N} & p_{2N} & \cdots & p_{NN} \end{bmatrix},$$

where $p_{ij} = P(s_t = j \mid s_{t-1} = i)$ and $\sum_{j=1}^{N} p_{ij} = 1$ for all i.

An $I(d)$ process, x_t, is defined as:

$$(1 - L)^d x_t = h_t,$$

where L is the lag operator ($Lk_t = k_{t-1}$) and h_t is a short memory process. When $d > 0$, the $I(d)$ process is often called the long memory process, because its autocovariance function is not summable so as to capture the long range dependence of a time series. In addition, the $I(d)$ process is nonstationary when $d \geq \frac{1}{2}$, otherwise, it is covariance stationary.

Combining the defining feature of a Markov chain and that of an $I(d)$ process, we propose the following MS-ARFIMA(p, d, q) process:

$$w_t = \mu_{s_t} I\{t \geq 1\} + (1 - L)^{-d_{s_t}} \sigma_{s_t} z_t I\{t \geq 1\} = \mu_{s_t} I\{t \geq 1\} + y_{s_t}, \tag{1}$$

where $I\{.\}$ is the indicator function and z_t is stationary process with mean zero and bounded positive spectral density $f_u(\lambda) \sim G_0$ as $\lambda \to 0$ at each possible regime, thus including stationary and invertible ARMA process as its special case. The most distinguished feature of the process is that the fractional differencing parameter d_{s_t} well known in the long memory literature is allowed to be a Markov chain satisfying the following Assumption A:

Assumption A. *s_t is independent of z_τ for all t and τ.*

The model in (1) subsumes many interesting models in the literature. When $N = 1$, w_t reduces to the specification in (7) of Shimotsu and Phillips (2005):

$$w_t = \mu_0 + (1 - L)^{-d_0} \sigma_0 z_t I\{t \geq 1\} \tag{2}$$

which also can be represented as:

$$w_t = \mu_0 + \sum_{k=0}^{t-1} \frac{(d_0)_k}{k} \sigma_0 z_{t-k}, \tag{3}$$

where

$$(d_0)_k = \frac{\Gamma(d_0 + k)}{\Gamma(d_0)} = (d_0)(d_0 + 1)\dots(d_0 + k - 1) \tag{4}$$

is Pochhammer's symbol for the forward factorial and $\Gamma(.)$ is the gamma function. Moreover, under the model in (1) and $d_{s_t} = 0$, w_t still includes the Markov-switching AR model considered in Hamilton (1989) as one of its special cases. We will show that the estimation of the model in (1) can be easily implemented with the algorithm proposed in this paper, even though the parameter estimation from a noisy version of realizations of Markov models is extremely difficult in all but very simple examples as well documented in Qian and Titterington (1991).

Let the total sample size be T, and denote $\mathcal{W}_t \equiv (w_1, w_2, \cdots, w_t)^\top$ the column vector containing the observations from time 1 to time t, while $\mathcal{S}_t = (s_1, s_2, \cdots, s_t)^\top$ represents the corresponding states, and $\mathcal{Y}_t = (y_1, y_2, \cdots, y_t)^\top$ in (1) is similarly defined. The column vector $\alpha = (\mu_1, \dots, \mu_N, \sigma_1, \dots, \sigma_N, \phi_{11}, \dots, \phi_{1p}, \phi_{21}, \dots, \phi_{Np},\ d_1, \dots, d_N, \theta_{11}, \cdots, \theta_{Nq})^\top$ and p_{ij} (transition probabilities) consist of the parameters characterizing the conditional density function (cdf) of w_t. After stacking the parameter vector α and the transition probabilities p_{ij} into one column vector ξ, we can represent the cdf of w_t as $f(w_t \mid \mathcal{S}_t, \mathcal{W}_{t-1}; \xi)$, clearly showing that the cdf of w_t depends on the entire past routes of states (in general). Indeed, there are N^T possible paths of states running throughout the observations $\mathcal{W}_T$.

To illustrate the proposed algorithm for the model in (1), we first consider the simplest case where w_t in (1) is generated as:

$$w_t = \mu_{s_t} I\{t \geq 1\} + (1 - L)^{-d_0} \sigma_0 \varepsilon_t I\{t \geq 1\} = \mu_{s_t} I\{t \geq 1\} + y_t, \tag{5}$$

where $d < \frac{1}{2}$ and ε_t is a zero mean normally, independently and identically distributed white noise (i.i.d.) with $E(\varepsilon_t^2) = 1$. That is, w_t in (5) is a special type of MS-ARFIMA$(0, d, 0)$ process whose differencing parameter is fixed across different regimes. Under Assumption A and $\varepsilon_t \sim N(0, 1)$ i.i.d. process, the likelihood function of $\mathcal{W}_T$, $L(\mathcal{S}_T, \mathcal{W}_T; \xi)$ hereafter, for the hidden Markov model in (5) equals

$$L(\mathcal{S}_T, \mathcal{W}_T; \xi) = (2\pi)^{-T/2} |\Lambda|^{-1/2} \exp\left(-\frac{1}{2} \mathcal{Y}_T^\top \Lambda^{-1} \mathcal{Y}_T\right) \prod_{t=1}^{T} \Pr(s_t \mid s_{t-1}), \tag{6}$$

where $\Lambda = E(\mathcal{Y}_T \mathcal{Y}_T^\top)$, and $\Pr(s_1 \mid s_0)$ is evaluated with the unconditional probability that the process will be in regime s_1. Given that y_t in (5) is a simple ARFIMA$(0, d, 0)$ process,

we can use the Durbin-Levinson algorithm to derive

$$(2\pi)^{-T/2}|\Lambda|^{-1/2}\exp\left(-\frac{1}{2}\mathcal{Y}_T^{\top}\Lambda^{-1}\mathcal{Y}_T\right)=\prod_{t=1}^{T}(2\pi)^{-1/2}v_{t-1}^{-1/2}\exp\left\{-\frac{(y_t-\widehat{y}_t)^2}{2v_{t-1}}\right\}, \tag{7}$$

where $\widehat{y}_t$ denotes the one-step ahead predictor of y_t with the observation $\mathcal{Y}_{t-1}$ as $j \geq 2$, and v_{t-1} is the corresponding one-step ahead prediction variance. Deriche and Tewfik (1993) also have employed the Durbin-Levinson algorithm to estimate a univariate ARFIMA$(0, d, 0)$ processes without Markov-switching characteristic. Note that as $t = 1$, $\widehat{y}_1 = 0$, and $v_0 = \gamma_0$ corresponds to the variance of y_t. As a result, the likelihood function in (6) can be rewritten as:

$$L(\mathcal{S}_T, \mathcal{W}_T; \xi)=\prod_{t=1}^{T}(2\pi)^{-1/2}v_{t-1}^{-1/2}\exp\left\{-\frac{(y_t-\widehat{y}_t)^2}{2v_{t-1}}\right\}\Pr(s_t \mid s_{t-1}), \tag{8}$$

indicating that the *unconditional* likelihood function of the mixture model in (5) can be exactly and recursively evaluated provided that we can identify the true path of s_t, $\mathcal{S}_T^*$.

We do not know in reality the value of $\mathcal{S}_T^*$. However, the recursive structure shown in (8) is especially suitable for implementing the Viterbi (1967) algorithm in the digital communication literature to identify the most likely path of states among the N^T possible routes within $\mathcal{W}_T$. We thus combine the Durbin-Levinson algorithm and the Viterbi algorithm to suggest a *Durbin-Levinson-Viterbi* (DLV) algorithm for the model in (5). When compared to the original Viterbi algorithm designed for solving the problem of maximum a posteriori probability estimate of the state sequence of a finite-state discrete-time Markov process observed in white noise, the DLV algorithm proposed in this paper is concerned with the hidden Markov process observed in a much more general ARFIMA noise. Since the DLV algorithm can estimate the differencing parameter of a time series allowing for the presence of a Markov-switching mean, the puzzle raised by Diebold and Inoue (2001) that long memory can be easily confused with a Markov-switching mean is thus resolved by using this DLV algorithm.

To locate the most likely path running through the data $\mathcal{W}_T$ with the idea of Viterbi (1967), we note first that, for each time t, there are N possible states ending at time t, i.e., $(s_t = i)$, $i = 1, \ldots, N$. For a particular node of these N end points at time t, say $(s_t = j)$, there exists a corresponding most likely path:

$$(\mathcal{S}_{t-1}(s_t = j), s_t = j) = (s_1(s_t = j), s_2(s_t = j), \cdots, s_{t-1}(s_t = j), s_t = j)\,, \tag{9}$$

which ends at this particular node $(s_t = j)$. We refer to the path $(\mathcal{S}_{t-1}(s_t = j), s_t = j)$ in (9) as the *survivor* associated with the node $(s_t = j)$. Note that, with little loss of clarity, we do not explicitly specify that the path depends on the parameter ξ and the observations $\mathcal{W}_t$ in order to simplify the notation. The likelihood function generated from this survivor $(\mathcal{S}_{t-1}(s_t = j), s_t = j)$ and the formula in (8) is recorded as $L(\mathcal{S}_{t-1}(s_t = j), s_t = j, \mathcal{W}_t; \xi)$ and is crucial for locating the most likely path running from time 1 to time T. In short, for each node $(s_t = j)$ at time t, there exists a most likely path, survivor $(\mathcal{S}_{t-1}(s_t = j), s_t = j)$, and its associated likelihood function $L(\mathcal{S}_{t-1}(s_t = j), s_t = j, \mathcal{W}_t; \xi)$. Most importantly, the number of survivors at each time t is always equal to N.

Given the N survivors at time t and in order to locate the survivor $(\mathcal{S}_t(s_{t+1} = i), s_{t+1} = i)$ for a particular node $(s_{t+1} = i)$ at time $t + 1$, among the N segments connecting the node $(s_{t+i} = i)$ and the N time-t survivors $(\mathcal{S}_{t-1}(s_t = j), s_t = j)$ recorded at time t, we select the one producing the largest likelihood function $L(\mathcal{S}_t(s_{t+1} = i), s_{t+1} = i, \mathcal{W}_{t+1}; \xi)$ among these N possible candidates, and name it as the survivor $(\mathcal{S}_t(s_{t+1} = i), s_{t+1} = i)$ for this particular node $(s_{t+1} = i)$. The computation of the aforementioned likelihoods is simple, because we record the likelihood functions of the N time-t survivors at each time t.

This recursive updating process proceeds from time 1 to time T and results in N time-T survivors $(\mathcal{S}_{T-1}(s_T = i), s_T = i)$ and their associated likelihood function $L(\mathcal{S}_{T-1}(s_T = i), s_T = i, \mathcal{W}_T; \xi)$, for each $i = 1, \ldots, N$. From these N time-T survivors we select the one producing the largest likelihood function, say $L(\mathcal{S}_{T-1}(s_T = g), s_T = g, \mathcal{W}_T; \xi)$, as the most likely path running from time 1 to time T. Combining a numerical optimization procedure and this chosen likelihood function $L(\mathcal{S}_{T-1}(s_T = g), s_T = g, \mathcal{W}_T; \xi)$ generated from the Viterbi algorithm and the Durbin-Levinson algorithm displayed in (7), we can estimate the parameters ξ and identify the states $\mathcal{S}_T$ hidden in the observations $\mathcal{W}_T$.

We now consider another special type of MS-ARFIMA(p, d, q) process:

$$w_t = \mu_{s_t} I\{t \geq 1\} + y_t = \mu_{s_t} I\{t \geq 1\} + (1 - L)^{-d_0} \sigma_0 z_t I\{t \geq 1\}, \quad \phi(L) z_t = \theta(L) \varepsilon_t, \tag{10}$$

where

$$\phi(L) = 1 - \phi_1 L - \ldots - \phi_p L^p, \quad \theta(L) = 1 + \theta_1 L + \ldots + \theta_q L^q, \tag{11}$$

and the roots of the polynomial $\phi(L)$ and those of $\theta(L)$ in (11) are all outside the unit circle and share no common roots. The model in (10) is much more general than that in (5), but

still can be estimated with the preceding Viterbi algorithm after some modifications. Please note that the value of fractional differencing parameter is unchanged across different regimes as that imposed in (5).

Note that the term y_t in (10) can be rearranged as

$$y_t = (1-L)^{-d_0}\sigma_0\phi(L)^{-1}\theta(L)\varepsilon_t, \qquad t = 1, 2, \ldots. \tag{12}$$

We then have

$$\phi(L)y_t = (1-L)^{-d_0}\sigma_0\theta(L)\varepsilon_t = \sigma_0\theta(L)(1-L)^{-d_0}\varepsilon_t = \sigma_0\theta(L)\widetilde{y}_t, \qquad t = 1, 2, \ldots, \tag{13}$$

where $\widetilde{y}_t = (1-L)^{-d_0}\varepsilon_t$ is an ARFIMA$(0, d, 0)$ process. Dueker and Serletis (2000) use the same transformation method for estimating an ARFIMA(p, d, q) process. Conditional on a set of $\phi(L)$ and $\theta(L)$ and a suitable starting value, the *conditional* likelihood function of y_t in (12) can still be evaluated exactly with the transformed ARFIMA$(0, d, 0)$ $\widetilde{y}_t$ in (13) and the Durbin-Levinson algorithm defined in (7). For example, conditional on y_0 being equal to 0, we can extract an ARFIMA$(0, d, 0)$ process from an ARFIMA$(1, d, 1)$ process as follows:

$$\sigma_0\widetilde{y}_t = y_t - \phi_1 y_{t-1} - \sigma_0\theta_1\widetilde{y}_{t-1}, \qquad t = 1, \ldots, T. \tag{14}$$

Conditional on a set of $\phi(L)$ and $\theta(L)$ and a suitable starting value for the parameter ξ, we can recursively and exactly evaluate the conditional likelihood function of the hidden Markov model using the DLV algorithm proposed previously.

The same idea also applies to the class of MS-ARFIMA(p, d, q) processes in (1) where d can be Markov-switching. However, we cannot use the Durbin-Levinson algorithm when the fractional differencing parameter is allowed to be Markov-switching. Nevertheless, the Viterbi algorithm is still powerful enough to locate the most likely path under this circumstance. That is, conditional on a suitable starting value for the parameter ξ, we employ the recursive structure inherent in Viterbi algorithms to identify the most likely path running through the data set.

3 Monte Carlo Experiment

In this section we consider a Monte Carlo experiment to demonstrate the finite sample performance of the proposed DLV algorithm on a special version of the model in (1):

$$w_t = \mu_{s_t} I\{t \geq 1\} + (1-L)^{-d_0}\sigma_0(1-\phi_1 L)^{-1}(1+\theta_1 L)\varepsilon_t I\{t \geq 1\}. \tag{15}$$

We employ three different values of the fractional differencing parameter:

$$d_0 = \{0.2, 0.3, 0.4\}, \tag{16}$$

along with the following parameters:

$$\mu_1 = 4, \;\; \mu_2 = 1, \;\; \phi_1 = 0.5, \;\; \theta_1 = 0.5, \;\; p_{11} = p_{22} = 0.95, \tag{17}$$

and σ_0 is chosen to ensure that the variance of the ARFIMA$(1, d, 1)$ noise in (15) is equal to 1 across different configurations. Note that the positive values of d_0 in (16) are chosen to reflect the variations used in the long memory literature.

All the computations are performed with GAUSS. Two hundred replications are conducted for each specification at 3 different sample sizes ($T = 100, 200, 400$) usually encountered in the empirical applications. For each sample size T, 200 additional values are generated in order to obtain random starting values. The optimization algorithm used to implement the DLV algorithm is the quasi-Newton algorithm of Broyden, Fletcher, Goldfarb, and Shanno (BFGS) contained in the GAUSS MAXLIK library. The maximum number of iterations for each replication is 100.

Table 1 contains the simulation results when the true value of parameters are used as the initial values for estimation procedure. The results reveal that the bias performance from the DLV algorithm is satisfactory (especially when the sample size is larger) for all configurations considered. Moreover, the associated root-mean-squared error (RMSE) almost always decreases with the increasing sample size. We find only two cases where the pattern of RMSE change is not what we expect, i.e., when $d_0 = 0.4$, the RMSE of estimating the parameters μ_1 and μ_2 as $T = 400$ is found to be a little higher than that of estimating the parameters μ_1 and μ_2 as $T = 200$. These two observations demonstrate the ability of the DLV algorithm to deal with the mixture model considered in this section. The performance of DLV algorithm for estimating the fractional differencing parameter is particularly displayed with the box-plots in Figure 2. The above-mentioned observations are clearly borne out in this figure.

We also check the robustness of the preceding simulation results by changing the choice of initial values for estimation. The simulations in Table 1 are replicated by setting the initial values for parameters at the true values except that of d_0 is set at zero. The results

Table 1. Finite sample performance of the DLV algorithm: Initial values of parameters are set at the true values of parameters

Parameter		μ_1	μ_2	p_{11}	p_{22}	σ_0	d_0	ϕ_1	θ_1
				$d_0 = \mathbf{0.4}$					
$T = 100$	Bias	-0.010	-0.106	0.010	0.016	0.008	0.183	-0.128	-0.040
	RMSE	1.008	0.991	0.039	0.060	0.022	0.294	0.233	0.138
$T = 200$	Bias	-0.094	-0.098	0.006	0.006	0.003	0.135	-0.101	-0.026
	RMSE	0.978	0.978	0.028	0.025	0.015	0.233	0.191	0.086
$T = 400$	Bias	-0.074	-0.076	0.004	0.004	0.001	0.096	-0.073	-0.013
	RMSE	0.990	0.990	0.019	0.017	0.010	0.192	0.163	0.060
				$d_0 = \mathbf{0.3}$					
$T = 100$	Bias	-0.057	-0.070	0.009	0.017	0.012	0.175	-0.109	-0.041
	RMSE	1.042	1.024	0.037	0.060	0.030	0.319	0.245	0.131
$T = 200$	Bias	-0.058	-0.055	0.006	0.006	0.005	0.122	-0.079	-0.030
	RMSE	0.947	0.944	0.027	0.025	0.020	0.260	0.212	0.086
$T = 400$	Bias	-0.038	-0.043	0.004	0.004	0.001	0.090	-0.061	-0.016
	RMSE	0.885	0.883	0.019	0.017	0.015	0.217	0.185	0.061
				$d_0 = \mathbf{0.2}$					
$T = 100$	Bias	-0.017	-0.041	0.009	0.017	0.014	0.201	-0.115	-0.044
	RMSE	0.874	0.853	0.037	0.060	0.037	0.341	0.258	0.128
$T = 200$	Bias	-0.042	-0.047	0.006	0.006	0.006	0.167	-0.106	-0.037
	RMSE	0.795	0.792	0.028	0.025	0.024	0.297	0.239	0.088
$T = 400$	Bias	-0.038	-0.046	0.004	0.004	0.002	0.122	-0.085	-0.019
	RMSE	0.670	0.669	0.019	0.017	0.018	0.239	0.203	0.061

Notes: Simulations are based on 200 replications. The data is generated from the mixture model defined in (15), (16) and (17). DLV algorithm is the Durbin-Levinson-Viterbi algorithm proposed in this paper. Bias is computed as the true parameter minus the corresponding average estimated values.

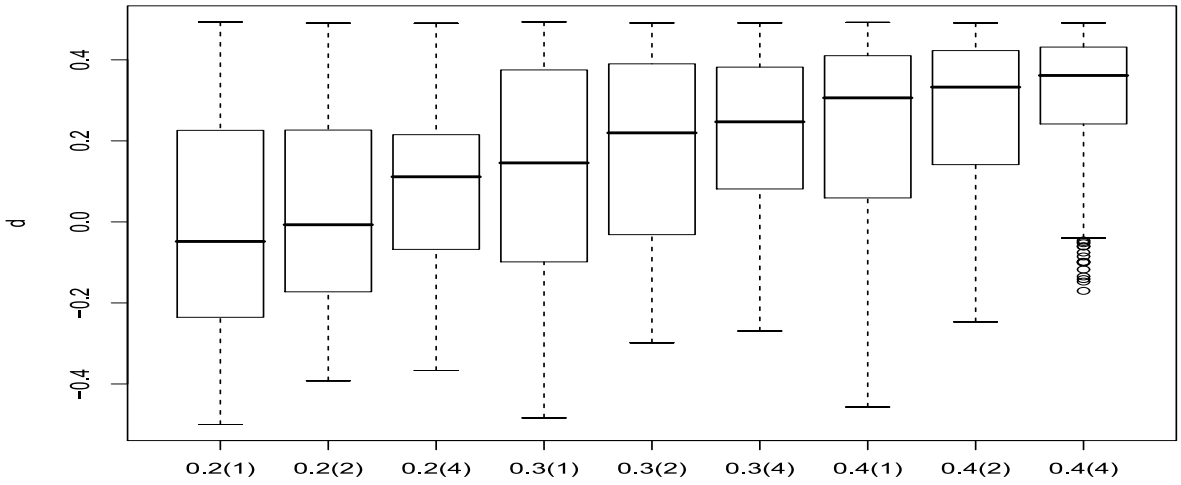


Figure 2: Box-plots of the estimated d from the model defined in (15), (16) and (17) with 200 realizations. The initial values of parameters are set at the true values of parameters. The value $f(g)$ denotes the model specification where $d = f$ and $T = 100 \times g$.

contained Table 2 and Figure 3 indicate that the finite sample performance of our procedure is not sensitive to the initial values used for estimation.

4 Empirical Applications

The methodology developed in this paper is motivated by the dynamic pattern of long memory behavior. Evidence has been given by many methods for such a changing covariance behavior of the Nile river. The applications of the proposed MS-ARFIMA model to actual data are far reaching. For that reason, we consider three data set. The first one is the U.S. real interest rates, the second one is the Nile river data, and the third one is the U.S. unemployment rates.

4.1 Example with real interest rates

In this subsection we first consider the U.S. ex post monthly real interest rate constructed from monthly inflation and Treasury bill rates from January 1953 to December 1990 in Mishkin (1990). The reason we use the original dataset of Mishkin (1990) is to employ it as a benchmark for a clear comparison between the results from the MS-ARFIMA model and

Table 2. Finite sample performance of the DLV algorithm: Initial values of parameters are set at the true values of parameters except that of d_0 is set at zero

Parameter		μ_1	μ_2	p_{11}	p_{22}	σ_0	d_0	ϕ_1	θ_1
					$d_0 = \mathbf{0.4}$				
$T = 100$	Bias	-0.116	-0.122	0.010	0.017	0.009	0.188	-0.130	-0.041
	RMSE	1.030	1.017	0.039	0.060	0.021	0.298	0.235	0.137
$T = 200$	Bias	-0.093	-0.096	0.006	0.006	0.003	0.138	-0.103	-0.027
	RMSE	0.979	0.979	0.028	0.025	0.015	0.238	0.193	0.087
$T = 400$	Bias	-0.074	-0.076	0.004	0.004	0.001	0.096	-0.073	-0.013
	RMSE	0.990	0.990	0.019	0.017	0.010	0.192	0.163	0.060
					$d_0 = \mathbf{0.3}$				
$T = 100$	Bias	-0.021	-0.034	0.010	0.017	0.012	0.186	-0.115	-0.040
	RMSE	0.972	0.949	0.039	0.060	0.030	0.325	0.241	0.127
$T = 200$	Bias	-0.046	-0.049	0.006	0.006	0.005	0.126	-0.081	-0.030
	RMSE	0.936	0.937	0.028	0.025	0.021	0.261	0.212	0.086
$T = 400$	Bias	-0.040	-0.044	0.004	0.004	0.002	0.088	-0.059	-0.016
	RMSE	0.912	0.912	0.019	0.017	0.015	0.217	0.184	0.060
					$d_0 = \mathbf{0.2}$				
$T = 100$	Bias	-0.018	-0.038	0.009	0.016	0.014	0.195	-0.110	-0.045
	RMSE	0.892	0.864	0.037	0.060	0.037	0.340	0.260	0.130
$T = 200$	Bias	-0.036	-0.040	0.006	0.006	0.006	0.160	-0.100	-0.037
	RMSE	0.804	0.801	0.028	0.025	0.024	0.294	0.238	0.087
$T = 400$	Bias	-0.044	-0.051	0.004	0.004	0.002	0.117	-0.082	-0.018
	RMSE	0.674	0.673	0.019	0.017	0.018	0.235	0.200	0.060

Notes: Simulations are based on 200 replications. The data is generated from the mixture model defined in (15), (16) and (17). DLV algorithm is the Durbin-Levinson-Viterbi algorithm proposed in this paper. Bias is computed as the true parameter minus the corresponding average estimated values.

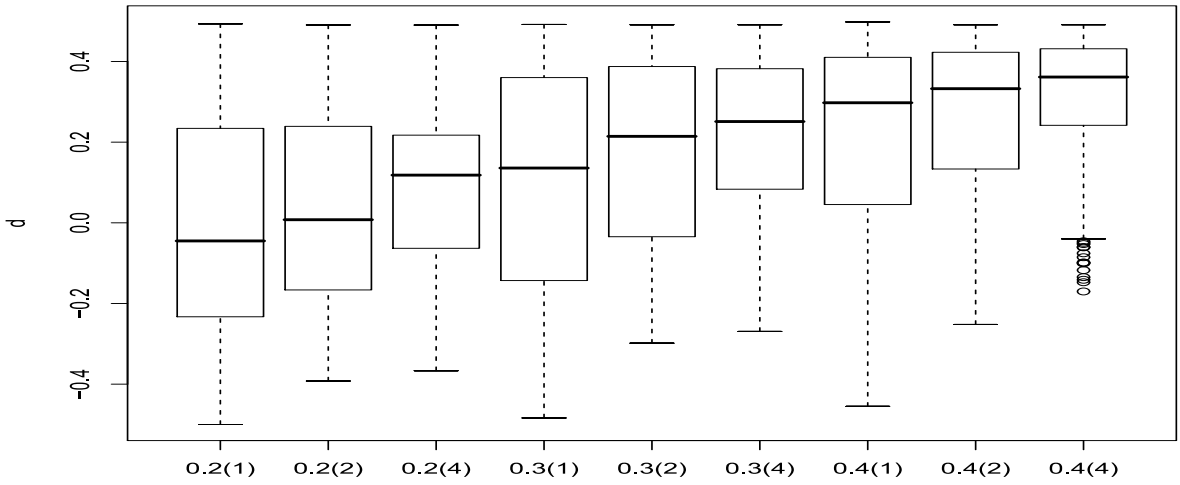


Figure 3: Box-plots of the estimated d from the model defined in (15), (16) and (17) with 200 realizations. The initial values of parameters are set at the true values except that of d_0 is set at zero. The value $f(g)$ denotes the model specification where $d = f$ and $T = 100 \times g$.

those generated from the methodology employed in earlier papers.

The main feature of the real interest rate is that the whole dataset can be split into three subperiods, January 1953-October 1979, November 1979-October 1982, and November 1982-December 1990, because the operating procedure of the monetary authority changed in October 1979 and October 1982 as argued in Mishkin (1990). Another interesting feature of the real interest rate is that the data of these three subperiods can be well described with the ARFIMA models as shown in Tsay (2000). The simultaneous presence of structural break and long memory within the real interest rate allows itself to be an ideal subject to be investigated with the MS-ARFIMA model.

Allowing the break points to be endogenously determined, Table 3 contains the parameter estimates from the following mixture model with a 2-state Markov chain and an ARFIMA$(1, d, 1)$ noise:

$$w_t = \mu_{s_t} I\{t \geq 1\} + (1 - L)^{-d_0} \sigma_0 z_t I\{t \geq 1\}, \quad (1 - \phi_1 L) z_t = (1 + \theta_1 L) \varepsilon_t, \tag{18}$$

where ϕ_1 or θ_1 is assumed to be zero depending on the noise specification. Following Hamilton (1989), asymptotic standard errors are calculated numerically.

Table 3 shows that the estimates of μ_1, μ_2, p_{11}, p_{22}, σ_0, and d_0 from the DLV algorithm

Table 3. Estimates of Parameters Based on Data for U.S. Monthly Real Interest Rate and the DLV Algorithm

	ARFIMA$(0,d,0)$		ARFIMA$(0,d,1)$		ARFIMA$(1,d,0)$		ARFIMA$(1,d,1)$	
	Estimate	S.E.	Estimate	S.E.	Estimate	S.E.	Estimate	S.E.
μ_1	5.3455	0.7494	5.3168	0.7162	5.3116	0.7124	5.3626	0.7706
μ_2	0.7226	0.4814	0.7194	0.4383	0.7184	0.4322	0.7352	0.4958
p_{11}	0.9833	0.0150	0.9833	0.0150	0.9833	0.0150	0.9833	0.0150
p_{22}	0.9977	0.0023	0.9977	0.0023	0.9977	0.0023	0.9977	0.0023
σ_0	2.5094	0.0831	2.5091	0.0831	2.5091	0.0831	2.4979	0.0827
d_0	0.2225	0.0367	0.2062	0.0520	0.2034	0.0653	0.2337	0.0376
ϕ_1	-	-	-	-	0.0324	0.0946	-0.9847	0.0155
θ_1	-	-	0.0279	0.0663	-	-	0.9675	0.0200
L^*	1079.0875		1079.0009		1078.9918		1077.0173	

Notes: The results are based on the MS-ARFIMA model defined in (18). S.E. stands for the standard error of the estimate. L^* represents the negative of the log-likelihood function of the switching model. DLV algorithm is the Durbin-Levinson-Viterbi algorithm proposed in this paper.

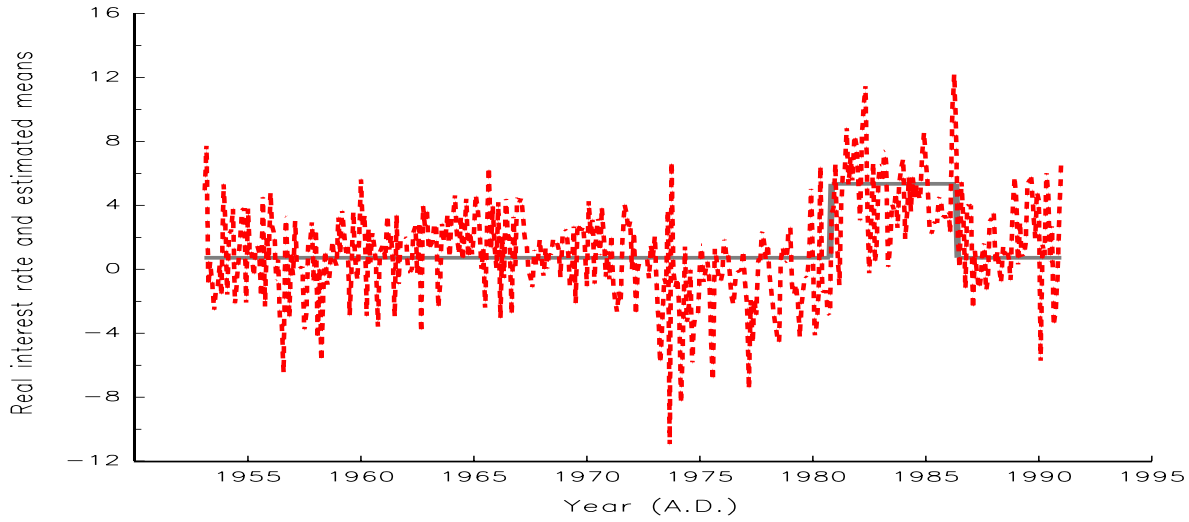


Figure 4: US monthly ex post real interest rates, January 1953-December 1990. Solid line denotes the path of estimated switching means from the specification ARFIMA$(0, d, 0)$ in Table 3, while dotted line denotes the observed monthly ex post real interest rates.

are quite robust across all 4 different configurations. More importantly, two identical break points are identified with these four models, thus divide the whole data into three subperiods as suggested in Mishkin (1990). The endogenous break points identified are November 1980 and May 1986, respectively.

Figure 4 displays the U.S. monthly ex post real interest rates and the path of estimated switching means generated from the DLV algorithm. Without loss of generality, only the path of the estimated switching means from the specification ARFIMA$(0, d, 0)$ in Table 3 is reported. Figure 4 shows that the model in (18) provides a satisfactory fitting of the U.S. monthly real interest rates. Although the endogenously identified break points are later than the well-known monetary operating procedure change points (October 1979 and October 1982), this finding is quite reasonable, because it takes some time for the ex post real interest rate to adjust its path after new information arrives. This argument is buttressed with the findings in Figure 4 that the endogenously identified break points are more closely connected to the observed path of the U.S. monthly ex post real interest rates than the monetary operating procedure change points are.

Table 3 also shows that a long memory phenomenon is found in the real interest rate as has been documented in Tsay (2000). Nevertheless, the estimate of the fractional differencing parameter in Table 3 is much lower than that of 0.666 in Table 3 of Tsay (2000) where the

change points are exogenenously determined, and it is more in line with the estimates of 0.204, 0.275, and 0.193 from the individual subperiod data presented in Table 3 of Tsay (2000). This implies that the persistence of long memory in the real interest rate is much more mitigated, once we take the potentially switching mean of the data into account, thus confirming the arguments of Diebold and Inoue (2001) that the presence of Markov-switching level might increase the persistence of the data under investigation.

4.2 Example with Nile river data

In this subsection we apply the Viterbi algorithm to the Nile river data with the following model:

$$w_t = \mu_{s_t} I\{t \geq 1\} + (1-L)^{-d_{s_t}} \sigma_{s_t} \varepsilon_t I\{t \geq 1\}, \tag{19}$$

where N is assumed to be 2. For the purpose of comparison, we estimate a fixed regime ARFIMA$(0, d, 0)$ model for the Nile river data, i.e., $N = 1$ is imposed on this model. The estimated value of d from such a fixed regime ARFIMA$(0, d, 0)$ model is 0.3986 and is almost identical to the finding in Beran and Terrin (1996).

When estimating the model in (19) with the Viterbi algorithm, we find that the value of the differencing parameter in Table 4 is 0.5770 (nonstationary) for one state, and is 0.2143 (stationary) for the other one. In addition, we identify 5 transitions within the Nile river data in the year 720, 805, 815, 878, and 1070. The estimated path of d_{s_t} from the MS-ARFIMA$(0, d, 0)$ model in Table 4 is graphed in Figure 5.

Most impressively, the first transition data occurs in the year of 720, and the associated estimated value of d_{s_t} within the period 622 to 719 is 0.2143 which is lower than the 0.5770 observed in the other regime. These two findings correspond closely to the conjectures in Beran and Terrin (1996) that the observations 1 to about 100 seem to be more independent than the subsequent observations and the value of differencing parameter might be lower for the first 100 observations than for the subsequent data.

In Figures 6 and 7 we present the observations and the fitted values generated from the estimated parameters displayed in Table 4. It is clear that the fitted value from the MS-ARFIMA$(0, d, 0)$ model is much closer to the real data than that generated from the model whose differencing parameter is not Markov switching. Combining the findings of the likelihood values in Table 4, we find that the MS-ARFIMA$(0, d, 0)$ model is a promising

Table 4. Estimates of MS-ARFIMA$(0, d, 0)$ Model based on the Nile River Data

	MS-ARFIMA$(0, d, 0)$		ARFIMA$(0, d, 0)$	
	Estimate	S.E.	Estimate	S.E.
μ_1	10.8593	0.6903	11.4847	0.2607
μ_2	11.4939	0.0917	-	-
p_{11}	0.9930	0.0042	-	-
p_{22}	0.9918	0.0050	-	-
σ_1	0.5430	0.0202	0.6995	0.0192
σ_2	0.8143	0.0332	-	-
d_1	0.5770	0.0430	0.3986	0.0309
d_2	0.2143	0.0510	-	-
L^*	687.5642		703.8541	

Notes: The MS-ARFIMA$(0, d, 0)$ model is defined in (19). S.E. stands for the standard error of the estimate based on numerical derivative. L^* represents the negative of the log-likelihood function of the estimated model.

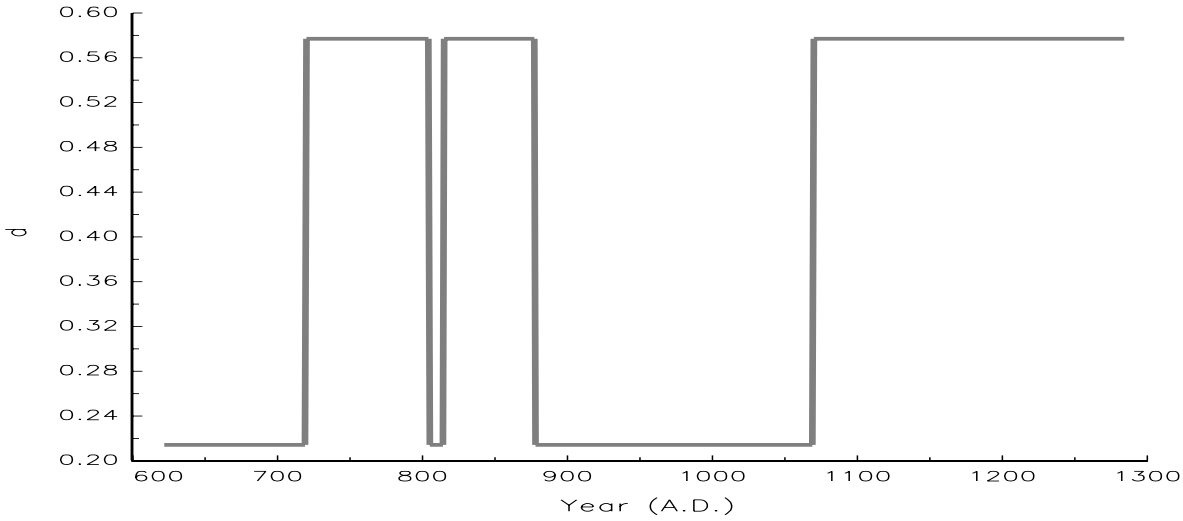


Figure 5: Estimated d_{s_t} from the MS-ARFIMA$(0, d, 0)$ model in Table 4.

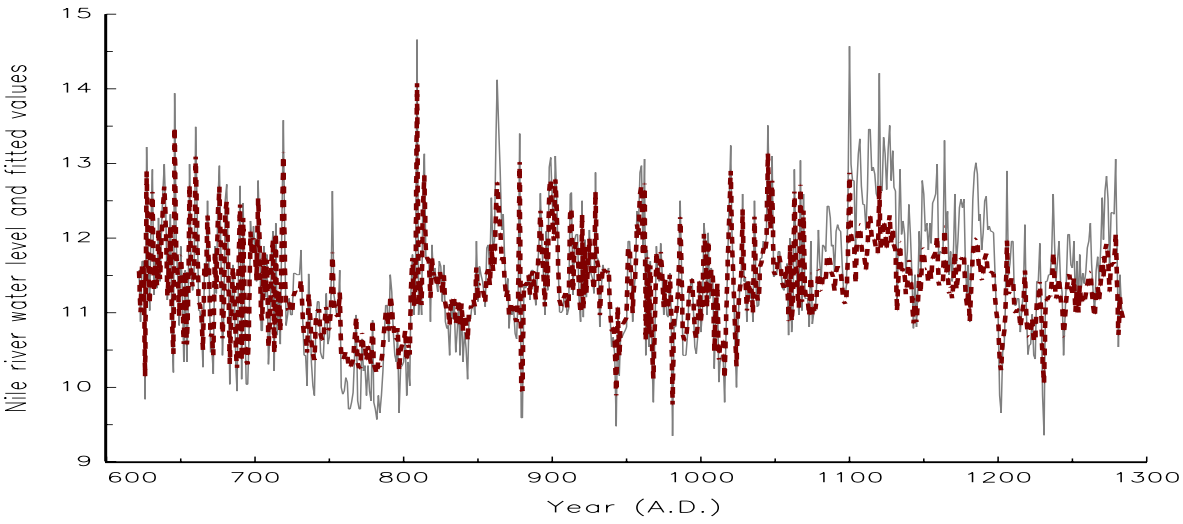


Figure 6: Solid line denotes the Nile river water level divided by 100, while dotted line denotes the corresponding fitted values from the MS-ARFIMA$(0, d, 0)$ model in Table 4.

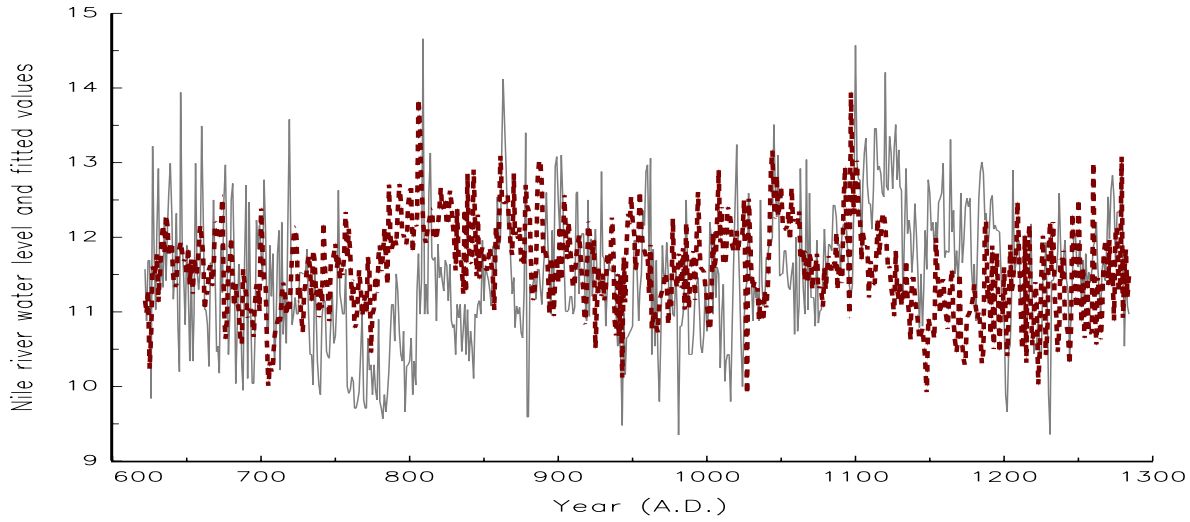


Figure 7: Solid line denotes the Nile river water level divided by 100, while dotted line denotes the corresponding fitted values from the ARFIMA$(0, d, 0)$ model in Table 4.

alternative to describe the Nile river data.

4.3 Example with unemployment rates

In this subsection we apply the Viterbi algorithm to the U.S. quarterly unemployment rates rates from 1948 to 2006. This data is based on the monthly unemployment rates contained in *Bureau of Labour Statistics* as those employed in van Dijk et al. (2002) for estimating a fractionally integrated smooth transition autoregressive (FI-STAR) model. However, van Dijk et al (2002) employ the original monthly unemployment rates ranging from July 1986 to December 1999, while we use all the data contained in *Bureau of Labour Statistics*, but focusing on the quarterly frequency usually considered in the business cycle related studies.

As clearly argued in van Dijk et al. (2002) and shown in Figure 8, there are two important empirical features of U.S. unemployment rates, i.e., the shocks to the series is quite persistent and the series seem to rise faster during recessions than it falls during expansions. van Dijk et al. (2002) find that the estimated d is 0.43 from a FI-STAR model presented in their Table 1. This implies that a time series model describing long memory and nonlinearity simultaneously may be useful for modeling U.S. unemployment rates and many other applications.

The aforementioned two features contained in U.S. unemployment also provide another good opportunity to test the applicability of the MS-ARFIMA model. As a consequence we estimate the U.S. quarterly unemployment rates with the following MS-ARFIMA$(p, d, 0)$

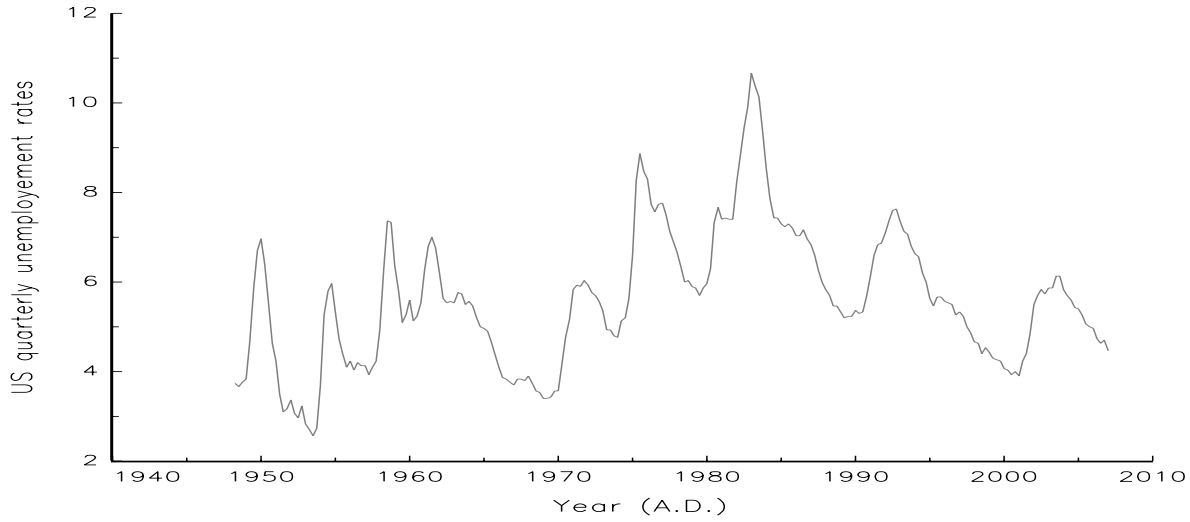


Figure 8: U.S. quarterly seasonally adjusted unemployment rates, 1948-2006.

model:

$$w_t = \mu_{s_t} I\{t \geq 1\} + (1-L)^{-d_{s_t}} \sigma_{s_t} \phi(B)^{-1} \varepsilon_t I\{t \geq 1\}, \tag{20}$$

where N is assumed to be 2, and $p = \{3, 4\}$. The choice of $p = 4$ is adopted by following the model specification in (30) of van Dijk et al (2002), while $p = 3$ is chosen to check the robustness of the estimation results from the specification $p = 4$. The major objective of this subsection is to investigate whether the long memory observed in van Dijk et al. (2002) can also be retained from the MS-ARFIMA methodology.

When estimating the model in (20) with the Viterbi algorithm, we find that the values of the estimated fractional differencing parameter from both MS-ARFIMA$(3, d, 0)$ and MS-ARFIMA$(4, d, 0)$ models in Table 5 are very close to that found in van Dijk et al. (2002), thus confirming that long memory phenomenon seems to be present in the U.S. unemployment rates. For clarity of exposition, the estimated path of d_{s_t} from the MS-ARFIMA$(3, d, 0)$ model and that of d_{s_t} from the MS-ARFIMA$(4, d, 0)$ one are graphed in Figure 9 and Figure 10, respectively. These figures clearly show that d_{s_t} are around 0.4-0.5 for both regimes estimated in each MS-ARFIMA$(p, d, 0)$ model in Table 5.

We also check to what extent the fitted values generated from the models in Table 5 can capture the feature of U.S. unemployment rates. This task is not taken in van Dijk et al. (2002) when estimating their FI-STAR model for the U.S. monthly unemployment rates. It is interesting to find in Figure 11 and Figure 12 that the MS-ARFIMA$(p, d, 0)$ model in (20)

Table 5. Estimates of MS-ARFIMA$(p, d, 0)$ Model based on the U.S. quarterly unemployment rates

	MS-ARFIMA$(3, d, 0)$		MS-ARFIMA$(4, d, 0)$	
	Estimate	S.E.	Estimate	S.E.
μ_1	3.8080	0.1552	3.4572	0.3711
μ_2	5.1358	0.4403	3.8254	0.3334
p_{11}	0.9939	0.0067	0.9877	0.0093
p_{22}	0.9896	0.0083	0.9867	0.0101
σ_1	0.1973	0.0135	0.1535	0.0101
σ_2	0.3380	0.0206	0.3921	0.0274
d_1	0.4919	0.1215	0.4429	0.0987
d_2	0.4143	0.1337	0.4342	0.1058
ϕ_1	1.2570	0.1415	1.1325	0.1215
ϕ_2	-0.3822	0.1510	-0.2301	0.1239
ϕ_3	-0.0666	0.0788	-0.0141	0.1053
ϕ_4	-	-	-0.0495	0.0712
L^*	36.1377		14.7662	

Notes: The results are based on the MS-ARFIMA$(p, d, 0)$ model defined in (20). S.E. stands for the standard error of the estimate based on numerical derivative. L^* represents the negative of the log-likelihood function of the estimated model.

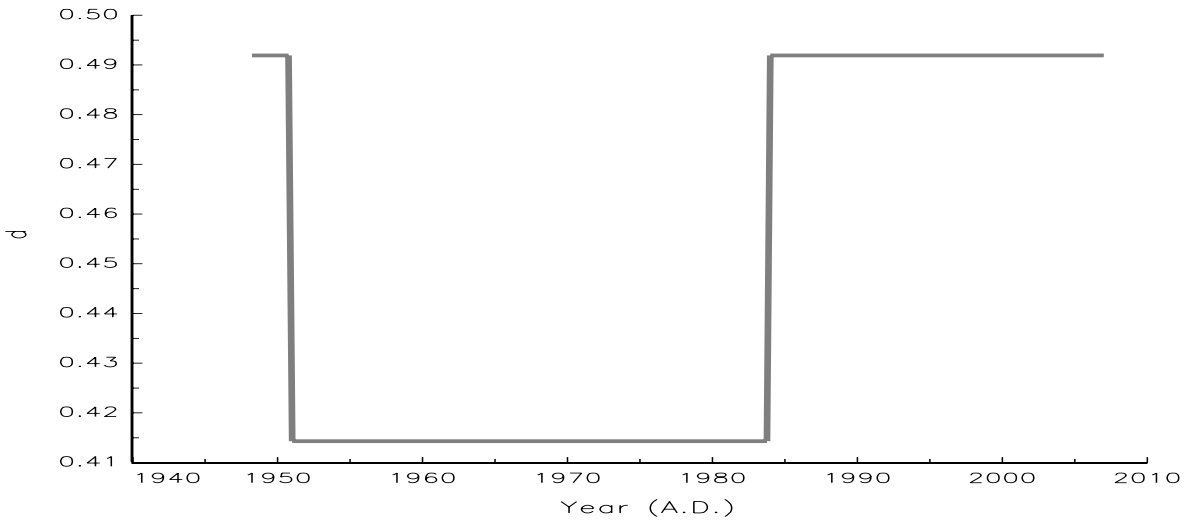


Figure 9: Estimated d_{s_t} from the MS-ARFIMA$(3, d, 0)$ model in Table 5.

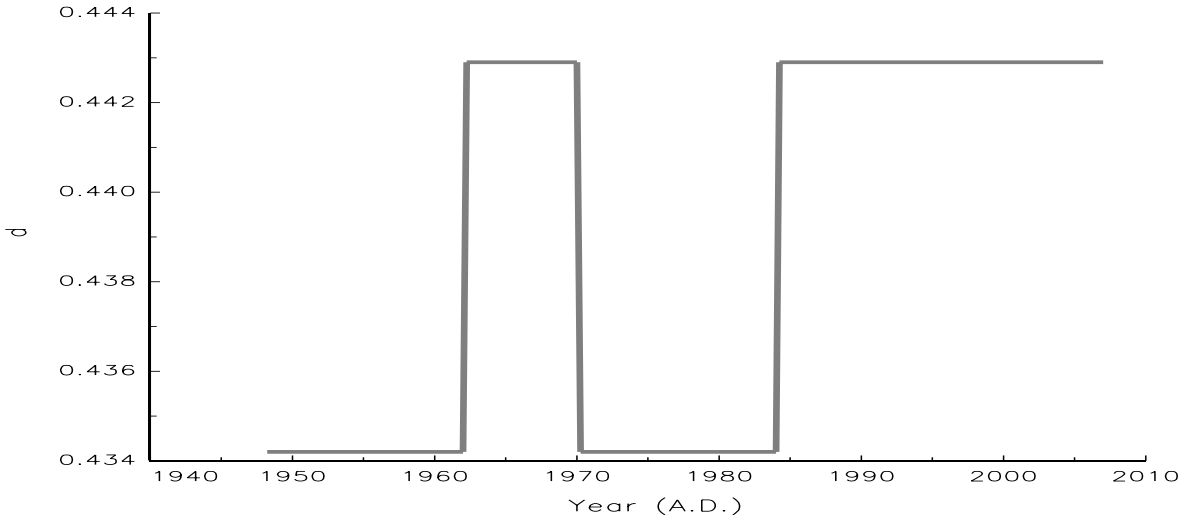


Figure 10: Estimated d_{s_t} from the MS-ARFIMA$(4, d, 0)$ model in Table 5.

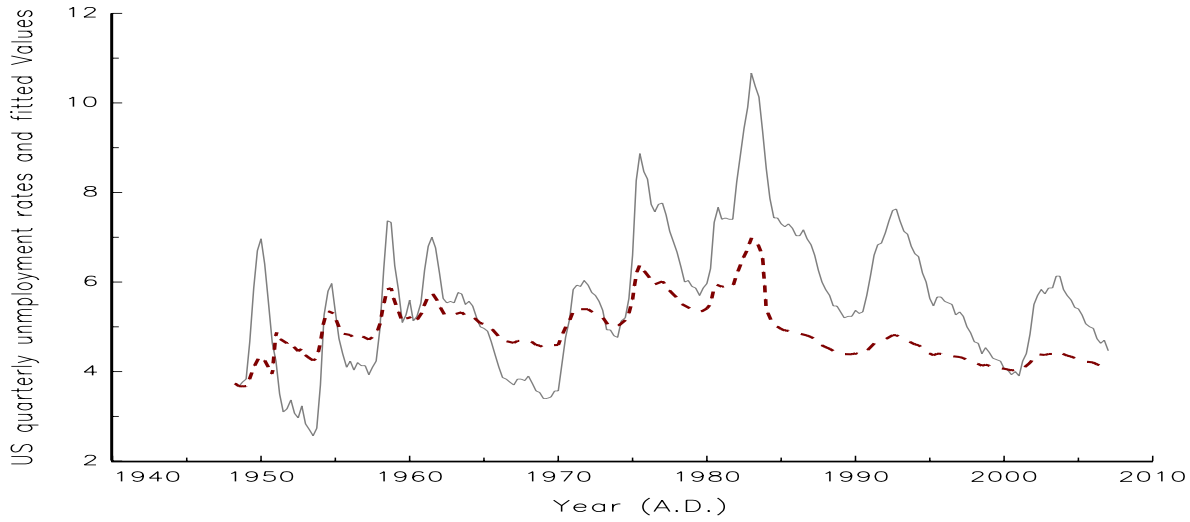


Figure 11: Solid line denotes the U.S. quarterly seasonally adjusted unemployment rates (1948-2006), while dotted line denotes the corresponding fitted values from the MS-ARFIMA$(3, d, 0)$ model in Table 5.

provides a reasonable fit to the data, even though we do not include some seasonal control variables, like seasonal difference operator, as van Dijk et al. (2002) have done for their empirical studies.

5 Conclusions

A general class of MS-ARFIMA processes is suggested to combine long memory and Markov-switching models into one unified framework. The coverage of this class of MS-ARFIMA models is far-reaching, but we show that they still can be easily estimated with the original Viterbi algorithm or the DLV algorithm proposed in this paper. In addition, the simulation reveals that the finite sample performance of the DLV algorithm for a simple mixture model of Markov-switching mean and ARFIMA$(1, d, 1)$ process is satisfactory. When applying the MS-ARFIMA models to the U.S. real interest rates, the Nile river level, and the U.S. unemployment rates, the estimation results are both highly compatible with the conjectures made in the literature. Accordingly, the MS-ARFIMA model considered in this paper not only can be used for solving the puzzle raised by Diebold and Inoue (2001), but can also find many potential applications in several scientific research fields.

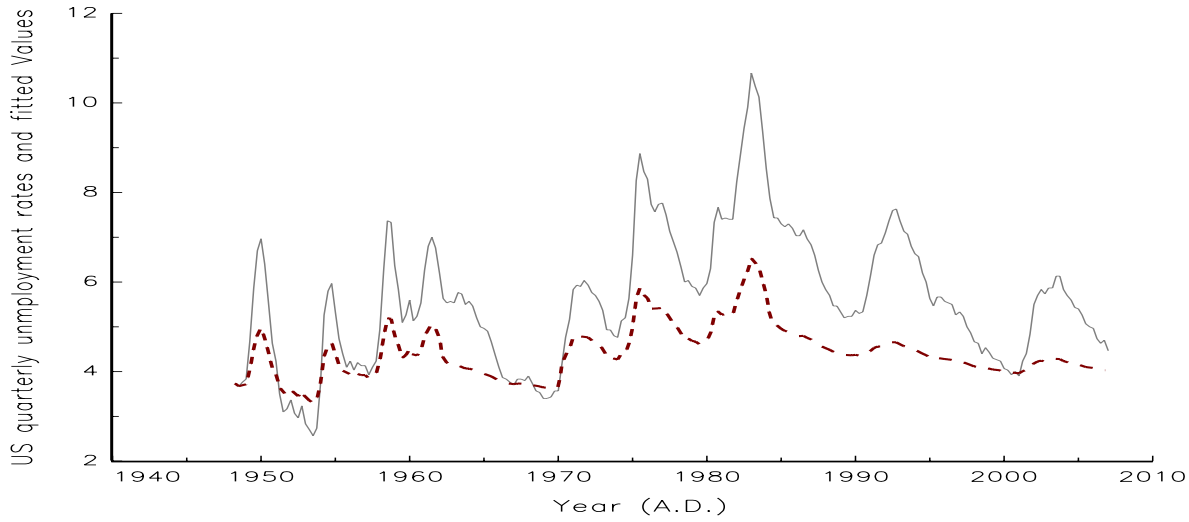


Figure 12: Solid line denotes the U.S. quarterly seasonally adjusted unemployment rates (1948-2006), while dotted line denotes the corresponding fitted values from the MS-ARFIMA$(4, d, 0)$ model in Table 5.

References

Beran, J. (1994), *Statistics for Long-Memory Processes*. New York: Chapman and Hall.

Beran, J. and Terrin, N. (1996), "Testing for a Change of the Long-Memory Parameter", *Biometrika*, 83, 627-638.

Berkes, I., Horváth, L., Kokoszka, P. and Shao, Q.M. (2006), "On Discriminating between Long-Range Dependence and Changes in Mean", *The Annals of Statistics*, 34, 1140-1165.

Bhattacharya, R.N., Gupta, V.K. and Waymire, E. (1983), "The Hurst Effect under Trends", *Journal of Applied Probability*, 20, 649-662.

Bollerslev, T. and Mikkelsen, H.O.A. (1996), "Modeling and Pricing Long-Memory in Stock Market Volatility", *Journal of Econometrics*, 73, 151-184.

Breidt, F.J., Crato, N. and de Lima, P. (1998), "The Detection and Estimation of Long Memory in Stochastic Volatility", *Journal of Econometrics*, 83, 325-348.

de Lima, P. and Crato, N. (1993), " Long-Range Dependence in the Conditional Variance of Stock Returns", *Proceedings of the Business and Economic Statistics Section*, August 1993 Joint Statistical Meetings, San Francisco.

Deo, R., Hurvich, C. and Lu, Y. (2006), "Forecasting Realized Volatility Using a Long-Memory Stochastic Volatility Model: Estimation, Prediction and Seasonal Adjustment", *Journal of Econometrics* 131, 29-58.

Deriche, J.A. and Tewfik, A.H. (1993), "Maximum Likelihood Estimation of the Parameters of Discrete Fractionally Differenced Gaussian Noise Process," *IEEE Transactions on Signal Processing*, 41, 2977-2989.

Diebold, F.X. and Inoue, A. (2001), "Long Memory and Regime Switching", *Journal of Econometrics*, 105, 131-159.

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.

Dueker, M. and Serletis, A. (2000), "Do Real Exchange Rates Have Autoregressive Unit Roots? A Test under the Alternative of Long Memory and Breaks", Working Paper 2000-016A, Federal Reserve Bank of St. Louis.

Granger, C.W.J. (1980), "Long Memory Relationships and the Aggregation of Dynamic Models", *Journal of Econometrics*, 14, 227-238.

Granger, C.W.J. and Joyeux, R. (1980), "An Introduction to Long-Memory Time Series Models and Fractional Differencing", *Journal of Time Series Analysis*, 1, 15-29.

Hamilton, J.D. (1989), "A New Approach to the Economic Analysis of Nonstationary Time Series and the Business Cycle", *Econometrica*, 57, 357-384.

Heyde, C.C. and Dai, W. (1996), "On the Robustness to Small Trends of Estimation based on the Smoothed Periodogram", *Journal of Time Series Analysis*, 17, 141-150.

Horváth, L. (2001), "Change-Point Detection in Long-Memory Processes. *Journal of Multivariate Analysis*, 78, 218-234.

Horváth, L. and Shao, Q.M. (1999), "Limit Theorems for the Union-Intersection Test", *Journal of Statistical Planning and Inference*, 44, 133-148.

Hosking, J.R.M. (1981), "Fractional Differencing", *Biometrika*, 68, 165-176.

Hurst, H.E. (1951), "Long-Term Storage Capacity of Reservoirs", *Transactions of the American Society of Civil Engineers*, 116, 770-799.

Juang, B.H. and Rabiner, L.R. (1991), "Hidden Marker Models for Speech Recognition", *Technometrics*, 33, 251-272.

Künsch, H. (1986), "Discrimination between Monotonic Trends and Long-Range Dependence", *Journal of Applied Probability*, 23, 1025-1030.

Mandelbrot, B.B. and van Ness, J.W. (1968), "Fractional Brownian Motions, Fractional Noises and Applications", *SIAM Review*, 10, 422-437.

Mandelbrot, B.B. and Wallis, J.R. (1969), "Some Long-Run Properties of Geophysical Records", *Water Resources Research*, 5, 321-340.

Mishkin, F.S. (1990), "What does the Term Structure of Interest Rate Tell Us about Future Inflation? *Journal of Monetary Economics*, 25, 77-95.

Qian, W. and Titterington, D.M. (1991), "Estimation of Parameters in Hidden Markov Models", *Philosophical Transactions: Physical Sciences and Engineering*, 337, 407-428.

Ray, B.K. and Tsay, R.S. (2002), "Bayesian Methods for Change-Point Detection in Long-Range Dependent Process", *Journal of Time Series Analysis*, 23, 687-705.

Robert, C.P., Rydén, R. and Titterington, D.M. (2000), "Bayesian Inference in Hidden Markov Models through the Reversible Jump Markov Chain Monte Carlo Method", *Journal of the Royal Statistical Society B*, 62, 57-75.

Shimotsu, K. and Phillips, P.C.B. (2005), "Exact Local Whittle Estimation of Fractional Integration", *The Annals of Statistics*, 33, 1890-1933.

Tsay, W.J. (2000), "Long Memory Story of the Real Interest Rate", *Economic Letters*, 67, 325-330.

van Dijk, D., Franses, P.H. and Raap, R. (2002), "A Nonlinear Long Memory Model, with an Application to US Unemployment", *Journal of Econometrics*, 110, 135-165.

Viterbi, A.J. (1967), "Error Bounds for Convolutional Codes and an Asymptotic Optimum Decoding Algorithm", *IEEE Transactions on Signal Processing*, IT-13, 260-269.

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.

SFB 649, Spandauer Straße 1, D-10178 Berlin
http://sfb649.wiwi.hu-berlin.de

This research was s
Forschungsgemeinschaft thr

SFB 649, Spandauer Straße 1, D-10178
http://sfb649.wiwi.hu-berlin.de

This research was supported by the Deutsche Forschungsgemeinschaft through the SFB 649 "Economic Risk".

The Annals of Statistics
2009, Vol. 37, No. 1, 1–34
DOI: 10.1214/07-AOS516

COMMON FUNCTIONAL PRINCIPAL COMPONENTS[1]

By Michal Benko, Wolfgang Härdle and Alois Kneip

Humboldt-Universität, Humboldt-Universität and Bonn Universität

Functional principal component analysis (FPCA) based on the Karhunen–Loève decomposition has been successfully applied in many applications, mainly for one sample problems. In this paper we consider common functional principal components for two sample problems. Our research is motivated not only by the theoretical challenge of this data situation, but also by the actual question of dynamics of implied volatility (IV) functions. For different maturities the log-returns of IVs are samples of (smooth) random functions and the methods proposed here study the similarities of their stochastic behavior. First we present a new method for estimation of functional principal components from discrete noisy data. Next we present the two sample inference for FPCA and develop the two sample theory. We propose bootstrap tests for testing the equality of eigenvalues, eigenfunctions, and mean functions of two functional samples, illustrate the test-properties by simulation study and apply the method to the IV analysis.

1. Introduction. In many applications in biometrics, chemometrics, econometrics, etc., the data come from the observation of continuous phenomenons of time or space and can be assumed to represent a sample of i.i.d. smooth random functions $X_1(t), \ldots, X_n(t) \in L^2[0,1]$. Functional data analysis has received considerable attention in the statistical literature during the last decade. In this context functional principal component analysis (FPCA) has proved to be a key technique. An early reference is Rao (1958), and important methodological contributions have been given by various authors. Case studies and references, as well as methodological and algorithmical details, can be found in the books by Ramsay and Silverman (2002, 2005) or Ferraty and Vieu (2006).

Received January 2006; revised February 2007.
[1]Supported by the Deutsche Forschungsgemeinschaft and the Sonderforschungsbereich 649 "Ökonomisches Risiko."
AMS 2000 subject classifications. Primary 62H25, 62G08; secondary 62P05.
Key words and phrases. Functional principal components, nonparametric regression, bootstrap, two sample problem.

The well-known Karhunen–Loève (KL) expansion provides a basic tool to describe the distribution of the random functions X_i and can be seen as the theoretical basis of FPCA. For $v, w \in L^2[0,1]$, let $\langle v, w\rangle = \int_0^1 v(t)w(t)\,dt$, and let $\|\cdot\| = \langle\cdot,\cdot\rangle^{1/2}$ denote the usual L^2-norm. With $\lambda_1 \geq \lambda_2 \geq \cdots$ and $\gamma_1, \gamma_2, \ldots$ denoting eigenvalues and corresponding orthonormal eigenfunctions of the covariance operator Γ of X_i, we obtain $X_i = \mu + \sum_{r=1}^{\infty} \beta_{ri}\gamma_r, i = 1, \ldots, n$, where $\mu = \mathrm{E}(X_i)$ is the mean function and $\beta_{ri} = \langle X_i - \mu, \gamma_r\rangle$ are (scalar) factor loadings with $\mathrm{E}(\beta_{ri}^2) = \lambda_r$. Structure and dynamics of the random functions can be assessed by analyzing the "functional principal components" γ_r, as well as the distribution of the factor loadings. For a given functional sample, the unknown characteristics λ_r, γ_r are estimated by the eigenvalues and eigenfunctions of the empirical covariance operator $\hat{\Gamma}_n$ of $X_1, \ldots, X_n$. Note that an eigenfunction γ_r is identified (up to sign) only if the corresponding eigenvalue λ_r has multiplicity one. This therefore establishes a necessary regularity condition for any inference based on an estimated functional principal component $\hat{\gamma}_r$ in FPCA. Signs are arbitrary (γ_r and β_{ri} can be replaced by $-\gamma_r$ and $-\beta_{ri}$) and may be fixed by a suitable standardization. More detailed discussion on this topic and precise assumptions can be found in Section 2.

In many important applications a small number of functional principal components will suffice to approximate the functions X_i with a high degree of accuracy. Indeed, FPCA plays a much more central role in functional data analysis than its well-known analogue in multivariate analysis. There are two major reasons. First, distributions on function spaces are complex objects, and the Karhunen–Loève expansion seems to be the only practically feasible way to access their structure. Second, in multivariate analysis a substantial interpretation of principal components is often difficult and has to be based on vague arguments concerning the correlation of principal components with original variables. Such a problem does not at all exists in the functional context, where $\gamma_1(t), \gamma_2(t), \ldots$ are functions representing the major modes of variation of $X_i(t)$ over t.

In this paper we consider inference and tests of hypotheses on the structure of functional principal components. Motivated by an application to implied volatility analysis, we will concentrate on the two sample case. A central point is the use of bootstrap procedures. We will show that the bootstrap methodology can also be applied to functional data.

In Section 2 we start by discussing one-sample inference for FPCA. Basic results on asymptotic distributions have already been derived by Dauxois, Pousse and Romain (1982) in situations where the functions are directly observable. Hall and Hosseini-Nasab (2006) develop asymptotic Taylor expansions of estimated eigenfunctions in terms of the difference $\hat{\Gamma}_n - \Gamma$.

Without deriving rigorous theoretical results, they also provide some qualitative arguments as well as simulation results motivating the use of bootstrap in order to construct confidence regions for principal components.

In practice, the functions of interest are often not directly observed, but are regression curves which have to be reconstructed from discrete, noisy data. In this context the standard approach is to first estimate individual functions nonparametrically (e.g., by B-splines) and then to determine principal components of the resulting estimated empirical covariance operator—see Besse and Ramsay (1986), Ramsay and Dalzell (1991), among others. Approaches incorporating a smoothing step into the eigenanalysis have been proposed by Rice and Silverman (1991), Pezzulli and Silverman (1993) or Silverman (1996). Robust estimation of principal components has been considered by Lacontore et al. (1999). Yao, Müller and Wang (2005) and Hall, Müller and Wang (2006) propose techniques based on nonparametric estimation of the covariance function $\mathrm{E}[\{X_i(t)-\mu(t)\}\{X_i(s)-\mu(s)\}]$ which can also be applied if there are only a few scattered observations per curve.

Section 2.1 presents a new method for estimation of functional principal components. It consists in an adaptation of a technique introduced by Kneip and Utikal (2001) for the case of density functions. The key-idea is to represent the components of the Karhunen–Loève expansion in terms of an (L^2) scalar-product matrix of the sample. We investigate the asymptotic properties of the proposed method. It is shown that under mild conditions the additional error caused by estimation from discrete, noisy data is first-order asymptotically negligible, and inference may proceed "as if" the functions were directly observed. Generalizing the results of Dauxois, Pousse and Romain (1982), we then present a theorem on the asymptotic distributions of the empirical eigenvalues and eigenfunctions. The structure of the asymptotic expansion derived in the theorem provides a basis to show consistency of bootstrap procedures.

Section 3 deals with two-sample inference. We consider two independent samples of functions $\{X_i^{(1)}\}_{i=1}^{n_1}$ and $\{X_i^{(2)}\}_{i=1}^{n_2}$. The problem of interest is to test in how far the distributions of these random functions coincide. The structure of the different distributions in function space can be accessed by means of the respective Karhunen–Loève expansions

$$X_i^{(p)} = \mu^{(p)} + \sum_{r=1}^{\infty} \beta_{ri}^{(p)} \gamma_r^{(p)}, \qquad p = 1, 2.$$

Differences in the distribution of these random functions will correspond to differences in the components of the respective KL expansions above. Without restriction, one may require that signs are such that $\langle \gamma_r^{(1)}, \gamma_r^{(2)} \rangle \geq 0$. Two sample inference for FPCA in general has not been considered in the literature so far. In Section 3 we define bootstrap procedures for testing

the equality of mean functions, eigenvalues, eigenfunctions and eigenspaces. Consistency of the bootstrap is derived in Section 3.1, while Section 3.2 contains a simulation study providing insight into the finite sample performance of our tests.

It is of particular interest to compare the functional components characterizing the two samples. If these factors are "common," this means $\gamma_r := \gamma_r^{(1)} = \gamma_r^{(2)}$, then only the factor loadings $\beta_{ri}^{(p)}$ may vary across samples. This situation may be seen as a functional generalization of the concept of "common principal components" as introduced by Flury (1988) in multivariate analysis. A weaker hypothesis may only require equality of the eigenspaces spanned by the first $L \in \mathbb{N}$ functional principal components. [$\mathbb{N}$ denotes the set of all natural numbers $1, 2, \ldots$ $(0 \notin \mathbb{N})$]. If for both samples the common L-dimensional eigenspaces suffice to approximate the functions with high accuracy, then the distributions in function space are well represented by a low-dimensional factor model, and subsequent analysis may rely on comparing the multivariate distributions of the random vectors $(\beta_{r1}^{(p)}, \ldots, \beta_{rL}^{(p)})^\top$.

The idea of "common functional principal components" is of considerable importance in implied volatility (IV) dynamics. This application is discussed in detail in Section 4. Implied volatility is obtained from the pricing model proposed by Black and Scholes (1973) and is a key parameter for quoting options prices. Our aim is to construct low-dimensional factor models for the log-returns of the IV functions of options with different maturities. In our application the first group of functional observations—$\{X_i^{(1)}\}_{i=1}^{n_1}$, are log-returns on the maturity "1 month" (1M group) and second group—$\{X_i^{(2)}\}_{i=1}^{n_2}$, are log-returns on the maturity "3 months" (3M group).

The first three eigenfunctions (ordered with respect to the corresponding eigenvalues), estimated by the method described in Section 2.1, are plotted in Figure 1. The estimated eigenfunctions for both groups are of similar structure, which motivates a common FPCA approach. Based on discretized vectors of functional values, a (multivariate) common principal components analysis of implied volatilities has already been considered by Fengler, Härdle and Villa (2003). They rely on the methodology introduced by Flury (1988) which is based on maximum likelihood estimation under the assumption of multivariate normality. Our analysis overcomes the limitations of this approach by providing specific hypothesis tests in a fully functional setup. It will be shown in Section 4 that for both groups $L = 3$ components suffice to explain 98.2% of the variability of the sample functions. An application of the tests developed in Section 3 does not reject the equality of the corresponding eigenspaces.

2. Functional principal components and one sample inference. In this section we will focus on one sample of i.i.d. smooth random functions $X_1, \ldots,$

$X_n \in L^2[0,1]$. We will assume a well-defined mean function $\mu = \mathrm{E}(X_i)$, as well as the existence of a continuous covariance function $\sigma(t,s) = \mathrm{E}[\{X_i(t) - \mu(t)\}\{X_i(s) - \mu(s)\}]$. Then $\mathrm{E}(\|X_i - \mu\|^2) = \int \sigma(t,t)\,dt < \infty$, and the covariance operator Γ of X_i is given by

$$(\Gamma v)(t) = \int \sigma(t,s)v(s)\,ds, \qquad v \in L^2[0,1].$$

The Karhunen–Loève decomposition provides a basic tool to describe the distribution of the random functions X_i. With $\lambda_1 \geq \lambda_2 \geq \cdots$ and $\gamma_1, \gamma_2, \ldots$ denoting eigenvalues and a corresponding complete orthonormal basis of eigenfunctions of Γ, we obtain

$$X_i = \mu + \sum_{r=1}^{\infty} \beta_{ri}\gamma_r, \qquad i = 1, \ldots, n, \tag{1}$$

where $\beta_{ri} = \langle X_i - \mu, \gamma_r \rangle$ are uncorrelated (scalar) factor loadings with $\mathrm{E}(\beta_{ri}) = 0$, $\mathrm{E}(\beta_{ri}^2) = \lambda_r$ and $\mathrm{E}(\beta_{ri}\beta_{ki}) = 0$ for $r \neq k$. Structure and dynamics of the random functions can be assessed by analyzing the "functional principal components" γ_r, as well as the distribution of the factor loadings.

A discussion of basic properties of (1) can, for example, be found in Gihman and Skorohod (1973). Under our assumptions, the infinite sums in (1) converge with probability 1, and $\sum_{r=1}^{\infty} \lambda_r = \mathrm{E}(\|X_i - \mu\|^2) < \infty$. Smoothness of X_i carries over to a corresponding degree of smoothness of $\sigma(t,s)$ and γ_r. If, with probability 1, $X_i(t)$ is twice continuously differentiable, then σ as well as γ_r are also twice continuously differentiable. The particular case of a Gaussian random function X_i implies that the β_{ri} are independent $N(0, \lambda_r)$-distributed random variables.

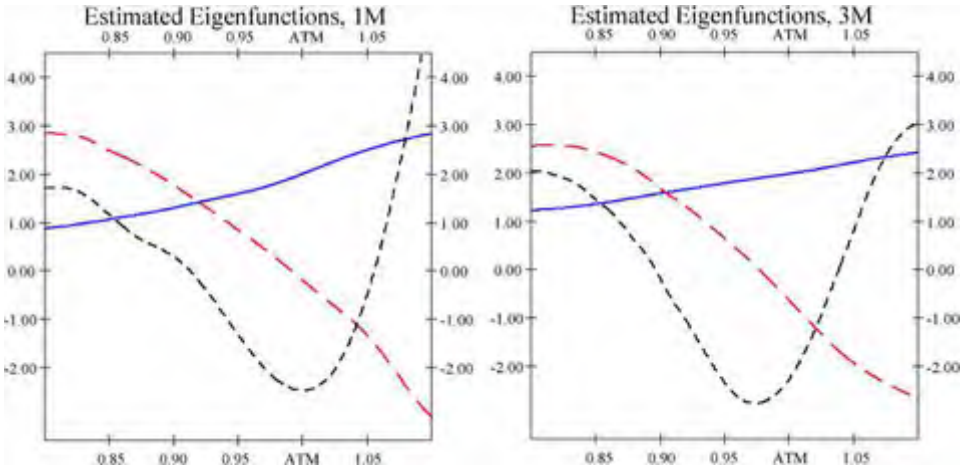


FIG. 1. *Estimated eigenfunctions for 1M group in the left plot and 3M group in the right plot: solid—first function, dashed—second function, finely dashed—third function.*

An important property of (1) consists in the known fact that the first L principal components provide a "best basis" for approximating the sample functions in terms of the integrated square error; see Ramsay and Silverman (2005), Section 6.2.3, among others. For any choice of L orthonormal basis functions $v_1, \ldots, v_L$, the mean integrated square error

$$\rho(v_1, \ldots, v_L) = \mathrm{E}\left(\left\| X_i - \mu - \sum_{r=1}^{L} \langle X_i - \mu, v_r \rangle v_r \right\|^2 \right) \tag{2}$$

is minimized by $v_r = \gamma_r$.

2.1. *Estimation of functional principal components.* For a given sample an empirical analog of (1) can be constructed by using eigenvalues $\hat{\lambda}_1 \geq \hat{\lambda}_2 \geq \cdots$ and orthonormal eigenfunctions $\hat{\gamma}_1, \hat{\gamma}_2, \ldots$ of the empirical covariance operator $\hat{\Gamma}_n$, where

$$(\hat{\Gamma}_n v)(t) = \int \hat{\sigma}(t,s) v(s)\, ds,$$

with $\bar{X} = n^{-1} \sum_{i=1}^{n} X_i$ and $\hat{\sigma}(t,s) = n^{-1} \sum_{i=1}^{n} \{X_i(t) - \bar{X}(t)\}\{X_i(s) - \bar{X}(s)\}$ denoting sample mean and covariance function. Then

$$X_i = \bar{X} + \sum_{r=1}^{n} \hat{\beta}_{ri} \hat{\gamma}_r, \qquad i = 1, \ldots, n, \tag{3}$$

where $\hat{\beta}_{ri} = \langle \hat{\gamma}_r, X_i - \bar{X} \rangle$. We necessarily obtain $n^{-1} \sum_i \hat{\beta}_{ri} = 0$, $n^{-1} \sum_i \hat{\beta}_{ri} \hat{\beta}_{si} = 0$ for $r \neq s$, and $n^{-1} \sum_i \hat{\beta}_{ri}^2 = \hat{\lambda}_r$.

Analysis will have to concentrate on the leading principal components explaining the major part of the variance. In the following we will assume that $\lambda_1 > \lambda_2 > \cdots > \lambda_{r_0} > \lambda_{r_0+1}$, where r_0 denotes the maximal number of components to be considered. For all $r = 1, \ldots, r_0$, the corresponding eigenfunction γ_r is then uniquely defined up to sign. Signs are arbitrary, decompositions (1) or (3) may just as well be written in terms of $-\gamma_r, -\beta_{ri}$ or $-\hat{\gamma}_r, -\hat{\beta}_{ri}$, and any suitable standardization may be applied by the statistician. In order to ensure that $\hat{\gamma}_r$ may be viewed as an estimator of γ_r rather than of $-\gamma_r$, we will in the following only assume that signs are such that $\langle \gamma_r, \hat{\gamma}_r \rangle \geq 0$. More generally, any subsequent statement concerning differences of two eigenfunctions will be based on the condition of a nonnegative inner product. This does not impose any restriction and will go without saying.

The results of Dauxois, Pousse and Romain (1982) imply that, under regularity conditions, $\|\hat{\gamma}_r - \gamma_r\| = \mathcal{O}_p(n^{-1/2})$, $|\hat{\lambda}_r - \lambda_r| = \mathcal{O}_p(n^{-1/2})$, as well as $|\hat{\beta}_{ri} - \beta_{ri}| = \mathcal{O}_p(n^{-1/2})$ for all $r \leq r_0$.

However, in practice, the sample functions X_i are often not directly observed, but have to be reconstructed from noisy observations Y_{ij} at discrete

design points t_{ik}:

$$Y_{ik} = X_i(t_{ik}) + \varepsilon_{ik}, \qquad k = 1, \dots, T_i, \tag{4}$$

where ε_{ik} are independent noise terms with $\mathrm{E}(\varepsilon_{ik}) = 0$, $\mathrm{Var}(\varepsilon_{ik}) = \sigma_i^2$.

Our approach for estimating principal components is motivated by the well-known duality relation between row and column spaces of a data matrix; see Härdle and Simar (2003), Chapter 8, among others. In a first step this approach relies on estimating the elements of the matrix:

$$M_{lk} = \langle X_l - \bar{X}, X_k - \bar{X} \rangle, \qquad l, k = 1, \dots, n. \tag{5}$$

Some simple linear algebra shows that all nonzero eigenvalues $\hat{\lambda}_1 \geq \hat{\lambda}_2 \cdots$ of $\hat{\Gamma}_n$ and $l_1 \geq l_2 \cdots$ of M are related by $\hat{\lambda}_r = l_r/n$, $r = 1, 2, \dots$. When using the corresponding orthonormal eigenvectors $p_1, p_2, \dots$ of M, the empirical scores $\hat{\beta}_{ri}$, as well as the empirical eigenfunctions $\hat{\gamma}_r$, are obtained by $\hat{\beta}_{ri} = \sqrt{l_r} p_{ir}$ and

$$\hat{\gamma}_r = \frac{1}{\sqrt{l_r}} \sum_{i=1}^{n} p_{ir}(X_i - \bar{X}) = \frac{1}{\sqrt{l_r}} \sum_{i=1}^{n} p_{ir} X_i. \tag{6}$$

The elements of M are functionals which can be estimated with asymptotically negligible bias and a parametric rate of convergence $T_i^{-1/2}$. If the data in (4) is generated from a balanced, equidistant design, then it is easily seen that for $i \neq j$ this rate of convergence is achieved by the estimator

$$\widehat{M}_{ij} = T^{-1} \sum_{k=1}^{T} (Y_{ik} - \bar{Y}_{\cdot k})(Y_{jk} - \bar{Y}_{\cdot k}), \qquad i \neq j,$$

and

$$\widehat{M}_{ii} = T^{-1} \sum_{k=1}^{T} (Y_{ik} - \bar{Y}_{\cdot k})^2 - \hat{\sigma}_i^2,$$

where $\hat{\sigma}_i^2$ denotes some nonparametric estimator of variance and $\bar{Y}_{\cdot k} = n^{-1} \times \sum_{j=1}^{n} Y_{jk}$.

In the case of a random design some adjustment is necessary: Define the ordered sample $t_{i(1)} \leq t_{i(2)} \leq \cdots \leq t_{i(T_i)}$ of design points, and for $j = 1, \dots, T_i$, let $Y_{i(j)}$ denote the observation belonging to $t_{i(j)}$. With $t_{i(0)} = -t_{i(1)}$ and $t_{i(T_i+1)} = 2 - t_{i(T_i)}$, set

$$\chi_i(t) = \sum_{j=1}^{T_i} Y_{i(j)} I\left(t \in \left[\frac{t_{i(j-1)} + t_{i(j)}}{2}, \frac{t_{i(j)} + t_{i(j+1)}}{2} \right) \right), \qquad t \in [0, 1],$$

where $I(\cdot)$ denotes the indicator function, and for $i \neq j$, define the estimate of M_{ij} by

$$\widehat{M}_{ij} = \int_0^1 \{\chi_i(t) - \bar{\chi}(t)\}\{\chi_j(t) - \bar{\chi}(t)\}\, dt,$$

where $\bar{\chi}(t) = n^{-1}\sum_{i=1}^{n}\chi_i(t)$. Finally, by redefining $t_{i(1)} = -t_{i(2)}$ and $t_{i(T_i+1)} = 2 - t_{i(T_i)}$, set $\chi_i^*(t) = \sum_{j=2}^{T_i} Y_{i(j-1)} I(t \in [\frac{t_{i(j-1)}+t_{i(j)}}{2}, \frac{t_{i(j)}+t_{i(j+1)}}{2}))$, $t \in [0,1]$. Then construct estimators of the diagonal terms M_{ii} by

$$\widehat{M}_{ii} = \int_0^1 \{\chi_i(t) - \bar{\chi}(t)\}\{\chi_i^*(t) - \bar{\chi}(t)\}\,dt. \tag{7}$$

The aim of using the estimator (7) for the diagonal terms is to avoid the additional bias implied by $\mathrm{E}_\varepsilon(Y_{ik}^2) = X_i(t_{ij})^2 + \sigma_i^2$. Here E_ε denotes conditional expectation given t_{ij}, X_i. Alternatively, we can construct a bias corrected estimator using some nonparametric estimation of variance σ_i^2, for example, the difference based model-free variance estimators studied in Hall, Kay and Titterington (1990) can be employed.

The eigenvalues $\hat{l}_1 \geq \hat{l}_2 \cdots$ and eigenvectors $\hat{p}_1, \hat{p}_2, \ldots$ of the resulting matrix $\widehat{M}$ then provide estimates $\hat{\lambda}_{r;T} = \hat{l}_r/n$ and $\hat{\beta}_{ri;T} = \sqrt{\hat{l}_r}\hat{p}_{ir}$ of $\hat{\lambda}_r$ and $\hat{\beta}_{ri}$. Estimates $\hat{\gamma}_{r;T}$ of the empirical functional principal component $\hat{\gamma}_r$ can be determined from (6) when replacing the unknown true functions X_i by nonparametric estimates $\hat{X}_i$ (as, for example, local polynomial estimates) with smoothing parameter (bandwidth) b:

$$\hat{\gamma}_{r;T} = \frac{1}{\sqrt{\hat{l}_r}}\sum_{i=1}^{n}\hat{p}_{ir}\hat{X}_i. \tag{8}$$

When considering (8), it is important to note that $\hat{\gamma}_{r;T}$ is defined as *a weighted average* of all estimated sample functions. Averaging reduces variance, and efficient estimation of $\hat{\gamma}_r$ therefore requires *undersmoothing* of individual function estimates $\hat{X}_i$. Theoretical results are given in Theorem 1 below. Indeed, if, for example, n and $T = \min_i T_i$ are of the same order of magnitude, then under suitable additional regularity conditions it will be shown that for an optimal choice of a smoothing parameter $b \sim (nT)^{-1/5}$ and twice continuously differentiable X_i, we obtain the rate of convergence $\|\hat{\gamma}_r - \hat{\gamma}_{r;T}\| = \mathcal{O}_p\{(nT)^{-2/5}\}$. Note, however, that the bias corrected estimator (7) may yield negative eigenvalues. In practice, these values will be small and will have to be interpreted as zero. Furthermore, the eigenfunctions determined by (8) may not be exactly orthogonal. Again, when using reasonable bandwidths, this effect will be small, but of course (8) may by followed by a suitable orthogonalization procedure.

It is of interest to compare our procedure to more standard methods for estimating $\hat{\lambda}_r$ and $\hat{\gamma}_r$ as mentioned above. When evaluating eigenvalues and eigenfunctions of the empirical covariance operator of nonparametrically estimated curves $\hat{X}_i$, then for fixed $r \leq r_0$ the above rate of convergence for the estimated eigenfunctions may well be achieved for a suitable choice of smoothing parameters (e.g., number of basis functions). But as will be seen

from Theorem 1, our approach also implies that $|\hat{\lambda}_r - \frac{\hat{l}_r}{n}| = \mathcal{O}_p(T^{-1} + n^{-1})$. When using standard methods it does not seem to be possible to obtain a corresponding rate of convergence, since any smoothing bias $|\mathrm{E}[\hat{X}_i(t)] - X_i(t)|$ will invariably affect the quality of the corresponding estimate of $\hat{\lambda}_r$.

We want to emphasize that any finite sample interpretation will require that T is sufficiently large such that our nonparametric reconstructions of individual curves can be assumed to possess a fairly small bias. The above arguments do not apply to extremely sparse designs with very few observations per curve [see Hall, Müller and Wang (2006) for an FPCA methodology focusing on sparse data].

Note that, in addition to (8), our final estimate of the empirical mean function $\hat{\mu} = \bar{X}$ will be given by $\hat{\mu}_T = n^{-1} \sum_i \hat{X}_i$. A straightforward approach to determine a suitable bandwidth b consists in a "leave-one-individual-out" cross-validation. For the maximal number r_0 of components to be considered, let $\hat{\mu}_{T,-i}$ and $\hat{\gamma}_{r;T,-i}$, $r = 1, \ldots, r_0$, denote the estimates of $\hat{\mu}$ and $\hat{\gamma}_r$ obtained from the data (Y_{lj}, t_{lj}), $l = 1, \ldots, i-1, i+1, \ldots, n$, $j = 1, \ldots, T_k$. By (8), these estimates depend on b, and one may approximate an optimal smoothing parameter by minimizing

$$\sum_i \sum_j \left\{ Y_{ij} - \hat{\mu}_{T,-i}(t_{ij}) - \sum_{r=1}^{r_0} \hat{\vartheta}_{ri} \hat{\gamma}_{r;T,-i}(t_{ij}) \right\}^2$$

over b, where $\hat{\vartheta}_{ri}$ denote ordinary least squares estimates of $\hat{\beta}_{ri}$. A more sophisticated version of this method may even allow to select different bandwidths b_r when estimating different functional principal components by (8). Although, under certain regularity conditions, the same qualitative rates of convergence hold for any arbitrary *fixed* $r \leq r_0$, the quality of estimates decreases when r becomes large. Due to $\langle \gamma_s, \gamma_r \rangle = 0$ for $s < r$, the number of zero crossings, peaks and valleys of γ_r has to increase with r. Hence, in tendency γ_r will be less and less smooth as r increases. At the same time, $\lambda_r \to 0$, which means that for large r the rth eigenfunctions will only possess a very small influence on the structure of X_i. This in turn means that the relative importance of the error terms ε_{ik} in (4) on the structure of $\hat{\gamma}_{r;T}$ will increase with r.

2.2. *One sample inference.* Clearly, in the framework described by (1)–(4) we are faced with two sources of variability of estimated functional principal components. Due to sampling variation, $\hat{\gamma}_r$ will differ from the true component γ_r, and due to (4), there will exist an additional estimation error when approximating $\hat{\gamma}_r$ by $\hat{\gamma}_{r;T}$.

The following theorems quantify the order of magnitude of these different types of error. Our theoretical results are based on the following assumptions on the structure of the random functions X_i.

ASSUMPTION 1. $X_1, \dots, X_n \in L^2[0,1]$ is an i.i.d. sample of random functions with mean μ and continuous covariance function $\sigma(t,s)$, and (1) holds for a system of eigenfunctions satisfying $\sup_{s\in\mathbb{N}} \sup_{t\in[0,1]} |\gamma_s(t)| < \infty$. Furthermore, $\sum_{r=1}^{\infty}\sum_{s=1}^{\infty} \mathrm{E}[\beta_{ri}^2\beta_{si}^2] < \infty$ and $\sum_{q=1}^{\infty}\sum_{s=1}^{\infty} \mathrm{E}[\beta_{ri}^2\beta_{qi}\beta_{si}] < \infty$ for all $r \in \mathbb{N}$.

Recall that $\mathrm{E}[\beta_{ri}] = 0$ and $\mathrm{E}[\beta_{ri}\beta_{si}] = 0$ for $r \neq s$. Note that the assumption on the factor loadings is necessarily fulfilled if X_i are Gaussian random functions. Then β_{ri} and β_{si} are independent for $r \neq s$, all moments of β_{ri} are finite, and hence $\mathrm{E}[\beta_{ri}^2\beta_{qi}\beta_{si}] = 0$ for $q \neq s$, as well as $\mathrm{E}[\beta_{ri}^2\beta_{si}^2] = \lambda_r\lambda_s$ for $r \neq s$; see Gihman and Skorohod (1973).

We need some further assumptions concerning smoothness of X_i and the structure of the discrete model (4).

ASSUMPTION 2. (a) X_i is a.s. twice continuously differentiable. There exists a constant $D_1 < \infty$ such that the derivatives are bounded by $\sup_t \mathrm{E}[X_i'(t)^4] \le D_1$, as well as $\sup_t \mathrm{E}[X_i''(t)^4] \le D_1$.

(b) The design points t_{ik}, $i = 1, \dots, n$, $k = 1, \dots, T_i$, are i.i.d. random variables which are independent of X_i and ε_{ik}. The corresponding design density f is continuous on $[0,1]$ and satisfies $\inf_{t\in[0,1]} f(t) > 0$.

(c) For any i, the error terms ε_{ik} are i.i.d. zero mean random variables with $\mathrm{Var}(\varepsilon_{ik}) = \sigma_i^2$. Furthermore, ε_{ik} is independent of X_i, and there exists a constant D_2 such that $\mathrm{E}(\varepsilon_{ik}^8) < D_2$ for all i, k.

(d) The estimates $\hat{X}_i$ used in (8) are determined by either a local linear or a Nadaraya–Watson kernel estimator with smoothing parameter b and kernel function K. K is a continuous probability density which is symmetric at 0.

The following theorems provide asymptotic results as $n, T \to \infty$, where $T = \min_{i=1}^{n}\{T_i\}$.

THEOREM 1. *In addition to Assumptions 1 and 2, assume that* $\inf_{s\neq r} |\lambda_r - \lambda_s| > 0$ *holds for some* $r = 1, 2, \dots$. *Then we have the following:*

(i) $n^{-1}\sum_{i=1}^{n}(\hat{\beta}_{ri} - \hat{\beta}_{ri;T})^2 = \mathcal{O}_p(T^{-1})$ *and*

$$\left|\hat{\lambda}_r - \frac{\hat{l}_r}{n}\right| = \mathcal{O}_p(T^{-1} + n^{-1}). \tag{9}$$

(ii) *If additionally* $b \to 0$ *and* $(Tb)^{-1} \to 0$ *as* $n, T \to \infty$, *then for all* $t \in [0,1]$,

$$|\hat{\gamma}_r(t) - \hat{\gamma}_{r;T}(t)| = \mathcal{O}_p\{b^2 + (nTb)^{-1/2} + (Tb^{1/2})^{-1} + n^{-1}\}. \tag{10}$$

A proof is given in the Appendix.

THEOREM 2. *Under Assumption 1 we obtain the following:*

(i) *For all* $t \in [0,1]$,

$$\sqrt{n}\{\bar{X}(t) - \mu(t)\} = \sum_r \left\{ \frac{1}{\sqrt{n}} \sum_{i=1}^n \beta_{ri} \right\} \gamma_r(t) \xrightarrow{\mathcal{L}} N\left(0, \sum_r \lambda_r \gamma_r(t)^2\right).$$

If, furthermore, $\lambda_{r-1} > \lambda_r > \lambda_{r+1}$ *holds for some fixed* $r \in \{1,2,\ldots\}$, *then*

(ii)

$$\sqrt{n}(\hat{\lambda}_r - \lambda_r) = \frac{1}{\sqrt{n}} \sum_{i=1}^n (\beta_{ri}^2 - \lambda_r) + \mathcal{O}_p(n^{-1/2}) \xrightarrow{\mathcal{L}} N(0, \Lambda_r), \tag{11}$$

where $\Lambda_r = \mathrm{E}[(\beta_{ri}^2 - \lambda_r)^2]$,

(iii) *and for all* $t \in [0,1]$

$$\hat{\gamma}_r(t) - \gamma_r(t) = \sum_{s \neq r} \left\{ \frac{1}{n(\lambda_r - \lambda_s)} \sum_{i=1}^n \beta_{si}\beta_{ri} \right\} \gamma_s(t) + R_r(t), \tag{12}$$

where $\|R_r\| = \mathcal{O}_p(n^{-1})$.

Moreover,

$$\sqrt{n} \sum_{s \neq r} \left\{ \frac{1}{n(\lambda_r - \lambda_s)} \sum_{i=1}^n \beta_{si}\beta_{ri} \right\} \gamma_s(t) \xrightarrow{\mathcal{L}} N\left(0, \sum_{q \neq r} \sum_{s \neq r} \frac{\mathrm{E}[\beta_{ri}^2 \beta_{qi} \beta_{si}]}{(\lambda_q - \lambda_r)(\lambda_s - \lambda_r)} \gamma_q(t)\gamma_s(t)\right).$$

A proof can be found in the Appendix. The theorem provides a generalization of the results of Dauxois, Pousse and Romain (1982) who derive explicit asymptotic distributions by assuming Gaussian random functions X_i. Note that in this case $\Lambda_r = 2\lambda_r^2$ and $\sum_{q \neq r} \sum_{s \neq r} \frac{\mathrm{E}[\beta_{ri}^2 \beta_{qi} \beta_{si}]}{(\lambda_q - \lambda_r)(\lambda_s - \lambda_r)} \gamma_q(t)\gamma_s(t) = \sum_{s \neq r} \frac{\lambda_r \lambda_s}{(\lambda_s - \lambda_r)^2} \gamma_s(t)^2$.

When evaluating the bandwidth-dependent terms in (10), best rates of convergence $|\hat{\gamma}_r(t) - \hat{\gamma}_{r;T}(t)| = \mathcal{O}_p\{(nT)^{-2/5} + T^{-4/5} + n^{-1}\}$ are achieved when choosing an undersmoothing bandwidth $b \sim \max\{(nT)^{-1/5}, T^{-2/5}\}$. Theoretical work in functional data analysis is usually based on the implicit assumption that the additional error due to (4) is negligible, and that one can proceed "as if" the functions X_i were directly observed. In view of Theorems 1 and 2, this approach is justified in the following situations:

(1) T is much larger than n, that is, $n/T^{4/5} \to 0$, and the smoothing parameter b in (8) is of order $T^{-1/5}$ (optimal smoothing of individual functions).

(2) An undersmoothing bandwidth $b \sim \max\{(nT)^{-1/5}, T^{-2/5}\}$ is used and $n/T^{8/5} \to 0$. This means that T may be smaller than n, but T must be at least of order of magnitude larger than $n^{5/8}$.

In both cases (1) and (2) the above theorems imply that $|\hat{\lambda}_r - \frac{\hat{l}_r}{n}| = \mathcal{O}_p(|\hat{\lambda}_r - \lambda_r|)$, as well as $\|\hat{\gamma}_r - \hat{\gamma}_{r;T}\| = \mathcal{O}_p(\|\hat{\gamma}_r - \gamma_r\|)$. Inference about functional principal components will then be first-order equivalent to an inference based on known functions X_i.

In such situations Theorem 2 suggests bootstrap procedures as tools for one sample inference. For example, the distribution of $\|\hat{\gamma}_r - \gamma_r\|$ may by approximated by the bootstrap distribution of $\|\hat{\gamma}_r^* - \hat{\gamma}_r\|$, where $\hat{\gamma}_r^*$ are estimates to be obtained from i.i.d. bootstrap resamples $X_1^*, X_2^*, \ldots, X_n^*$ of $\{X_1, X_2, \ldots, X_n\}$. This means that $X_1^* = X_{i_1}, \ldots, X_n^* = X_{i_n}$ for some indices $i_1, \ldots, i_n$ drawn independently and with replacement from $\{1, \ldots, n\}$ and, in practice, $\hat{\gamma}_r^*$ may thus be approximated from corresponding discrete data $(Y_{i_1 j}, t_{i_1 j})_{j=1,\ldots,T_{i_1}}, \ldots, (Y_{i_n j}, t_{i_n j})_{j=1,\ldots,T_{i_n}}$. The additional error is negligible if either (1) or (2) is satisfied.

One may wonder about the validity of such a bootstrap. Functions are complex objects and there is no established result in bootstrap theory which readily generalizes to samples of random functions. But by (1), i.i.d. bootstrap resamples $\{X_i^*\}_{i=1,\ldots,n}$ may be equivalently represented by corresponding, i.i.d. resamples $\{\beta_{1i}^*, \beta_{2i}^*, \ldots\}_{i=1,\ldots,n}$ of factor loadings. Standard multivariate bootstrap theorems imply that for any $q \in \mathbb{N}$ the distribution of moments of the random vectors $(\beta_{1i}, \ldots, \beta_{qi})$ may be consistently approximated by the bootstrap distribution of corresponding moments of $(\beta_{1i}^*, \ldots, \beta_{qi}^*)$. Together with some straightforward limit arguments as $q \to \infty$, the structure of the first-order terms in the asymptotic expansions (11) and (12) then allows to establish consistency of the functional bootstrap. These arguments will be made precise in the proof of Theorem 3 below, which concerns related bootstrap statistics in two sample problems.

REMARK. Theorem 2(iii) implies that the variance of $\hat{\gamma}_r$ is large if one of the differences $\lambda_{r-1} - \lambda_r$ or $\lambda_r - \lambda_{r+1}$ is small. In the limit case of eigenvalues of multiplicity $m > 1$ our theory does not apply. Note that then only the m-dimensional eigenspace is identified, but not a particular basis (eigenfunctions). In multivariate PCA Tyler (1981) provides some inference results on corresponding projection matrices assuming that $\lambda_r > \lambda_{r+1} \geq \cdots \geq \lambda_{r+m} > \lambda_{r+m+1}$ for known values of r and m.

Although the existence of eigenvalues λ_r, $r \leq r_0$, with multiplicity $m > 1$ may be considered as a degenerate case, it is immediately seen that $\lambda_r \to 0$ and, hence, $\lambda_r - \lambda_{r+1} \to 0$ as r increases. Even in the case of fully observed

functions X_i, estimates of eigenfunctions corresponding to very small eigenvalues will thus be poor. The problem of determining a sensible upper limit of the number r_0 of principal components to be analyzed is addressed in Hall and Hosseini-Nasab (2006).

3. Two sample inference. The comparison of functional components across groups leads naturally to two sample problems. Thus, let

$$X_1^{(1)}, X_2^{(1)}, \ldots, X_{n_1}^{(1)} \quad \text{and} \quad X_1^{(2)}, X_2^{(2)}, \ldots, X_{n_2}^{(2)}$$

denote two independent samples of smooth functions. The problem of interest is to test in how far the distributions of these random functions coincide. The structure of the different distributions in function space can be accessed by means of the respective Karhunen–Loève decompositions. The problem to be considered then translates into testing equality of the different components of these decompositions given by

$$X_i^{(p)} = \mu^{(p)} + \sum_{r=1}^{\infty} \beta_{ri}^{(p)} \gamma_r^{(p)}, \qquad p = 1, 2, \tag{13}$$

where again $\gamma_r^{(p)}$ are the eigenfunctions of the respective covariance operator $\Gamma^{(p)}$ corresponding to the eigenvalues $\lambda_1^{(p)} = \mathrm{E}\{(\beta_{1i}^{(p)})^2\} \geq \lambda_2^{(p)} = \mathrm{E}\{(\beta_{2i}^{(p)})^2\} \geq \cdots$. We will again suppose that $\lambda_{r-1}^{(p)} > \lambda_r^{(p)} > \lambda_{r+1}^{(p)}$, $p = 1, 2$, for all $r \leq r_0$ components to be considered. Without restriction, we will additionally assume that signs are such that $\langle \gamma_r^{(1)}, \gamma_r^{(2)} \rangle \geq 0$, as well as $\langle \hat{\gamma}_r^{(1)}, \hat{\gamma}_r^{(2)} \rangle \geq 0$.

It is of great interest to detect possible variations in the functional components characterizing the two samples in (13). Significant difference may give rise to substantial interpretation. Important hypotheses to be considered thus are as follows:

$$H_{0_1} : \mu^{(1)} = \mu^{(2)} \quad \text{and} \quad H_{0_{2,r}} : \gamma_r^{(1)} = \gamma_r^{(2)}, \qquad r \leq r_0.$$

Hypothesis $H_{0_{2,r}}$ is of particular importance. Then $\gamma_r^{(1)} = \gamma_r^{(2)}$ and only the factor loadings β_{ri} may vary across samples. If, for example, $H_{0_{2,r}}$ is accepted, one may additionally want to test hypotheses about the distributions of $\beta_{ri}^{(p)}$, $p = 1, 2$. Recall that necessarily $\mathrm{E}\{\beta_{ri}^{(p)}\} = 0$, $\mathrm{E}\{\beta_{ri}^{(p)}\}^2 = \lambda_r^{(p)}$, and $\beta_{si}^{(p)}$ is uncorrelated with $\beta_{ri}^{(p)}$ if $r \neq s$. If the $X_i^{(p)}$ are Gaussian random variables, the $\beta_{ri}^{(p)}$ are independent $N(0, \lambda_r)$ random variables. A natural hypothesis to be tested then refers to the equality of variances:

$$H_{0_{3,r}} : \lambda_r^{(1)} = \lambda_r^{(2)}, \qquad r = 1, 2, \ldots.$$

Let $\hat{\mu}^{(p)}(t) = \frac{1}{n_p} \sum_i X_i^{(p)}(t)$, and let $\hat{\lambda}_1^{(p)} \geq \hat{\lambda}_2^{(p)} \geq \cdots$ and $\hat{\gamma}_1^{(p)}, \hat{\gamma}_2^{(p)}, \ldots$ denote eigenvalues and corresponding eigenfunctions of the empirical covariance operator $\hat{\Gamma}_{n_p}^{(p)}$ of $X_1^{(p)}, X_2^{(p)}(t), \ldots, X_{n_p}^{(p)}$. The following test statistics are

defined in terms of $\hat{\mu}^{(p)}$, $\hat{\lambda}_r^{(p)}$ and $\hat{\gamma}_r^{(p)}$. As discussed in the proceeding section, all curves in both samples are usually not directly observed, but have to be reconstructed from noisy observations according to (4). In this situation, the "true" empirical eigenvalues and eigenfunctions have to be replaced by their discrete sample estimates. Bootstrap estimates are obtained by resampling the observations corresponding to the unknown curves $X_i^{(p)}$. As discussed in Section 2.2, the validity of our test procedures is then based on the assumption that T is sufficiently large such that the additional estimation error is asymptotically negligible.

Our tests of the hypotheses $H_{0_1}, H_{0_{2,r}}$ and $H_{0_{3,r}}$ rely on the statistics

$$D_1 \stackrel{\text{def}}{=} \|\hat{\mu}^{(1)} - \hat{\mu}^{(2)}\|^2,$$
$$D_{2,r} \stackrel{\text{def}}{=} \|\hat{\gamma}_r^{(1)} - \hat{\gamma}_r^{(2)}\|^2,$$
$$D_{3,r} \stackrel{\text{def}}{=} |\hat{\lambda}_r^{(1)} - \hat{\lambda}_r^{(2)}|^2.$$

The respective null-hypothesis has to be rejected if $D_1 \geq \Delta_{1;1-\alpha}$, $D_{2,r} \geq \Delta_{2,r;1-\alpha}$ or $D_{3,r} \geq \Delta_{3,r;1-\alpha}$, where $\Delta_{1;1-\alpha}$, $\Delta_{2,r;1-\alpha}$ and $\Delta_{3,r;1-\alpha}$ denote the critical values of the distributions of

$$\Delta_1 \stackrel{\text{def}}{=} \|\hat{\mu}^{(1)} - \mu^{(1)} - (\hat{\mu}^{(2)} - \mu^{(2)})\|^2,$$
$$\Delta_{2,r} \stackrel{\text{def}}{=} \|\hat{\gamma}_r^{(1)} - \gamma_r^{(1)} - (\hat{\gamma}_r^{(2)} - \gamma_r^{(2)})\|^2,$$
$$\Delta_{3,r} \stackrel{\text{def}}{=} |\hat{\lambda}_r^{(1)} - \lambda_r^{(1)} - (\hat{\lambda}_r^{(2)} - \lambda_r^{(2)})|^2.$$

Of course, the distributions of the different Δ's cannot be accessed directly, since they depend on the unknown true population mean, eigenvalues and eigenfunctions. However, it will be shown below that these distributions and, hence, their critical values are approximated by the bootstrap distribution of

$$\Delta_1^* \stackrel{\text{def}}{=} \|\hat{\mu}^{(1)*} - \hat{\mu}^{(1)} - (\hat{\mu}^{(2)*} - \hat{\mu}^{(2)})\|^2,$$
$$\Delta_{2,r}^* \stackrel{\text{def}}{=} \|\hat{\gamma}_r^{(1)*} - \hat{\gamma}_r^{(1)} - (\hat{\gamma}_r^{(2)*} - \hat{\gamma}_r^{(2)})\|^2,$$
$$\Delta_{3,r}^* \stackrel{\text{def}}{=} |\hat{\lambda}_r^{(1)*} - \hat{\lambda}_r^{(1)} - (\hat{\lambda}_r^{(2)*} - \hat{\lambda}_r^{(2)})|^2,$$

where $\hat{\mu}^{(1)*}$, $\hat{\gamma}_r^{(1)*}$, $\hat{\lambda}_r^{(1)*}$, as well as $\hat{\mu}^{(2)*}$, $\hat{\gamma}_r^{(2)*}$, $\hat{\lambda}_r^{(2)*}$, are estimates to be obtained from independent bootstrap samples $X_1^{1*}(t), X_2^{1*}(t), \ldots, X_{n_1}^{1*}(t)$, as well as $X_1^{2*}(t), X_2^{2*}(t), \ldots, X_{n_2}^{2*}(t)$.

This test procedure is motivated by the following insights:

(1) Under each of our null-hypotheses the respective test statistics D is equal to the corresponding Δ. The test will thus asymptotically possess the correct level: $P(D > \Delta_{1-\alpha}) \approx \alpha$.

(2) If the null hypothesis is false, then $D \neq \Delta$. Compared to the distribution of Δ, the distribution of D is shifted by the difference in the true means, eigenfunctions or eigenvalues. In tendency D will be larger than $\Delta_{1-\alpha}$.

Let $1 < L \leq r_0$. Even if for $r \leq L$ the equality of eigenfunctions is rejected, we may be interested in the question of whether at least the L-dimensional eigenspaces generated by the first L eigenfunctions are identical. Therefore, let $\mathcal{E}_L^{(1)}$, as well as $\mathcal{E}_L^{(2)}$, denote the L-dimensional linear function spaces generated by the eigenfunctions $\gamma_1^{(1)}, \ldots, \gamma_L^{(1)}$ and $\gamma_1^{(2)}, \ldots, \gamma_L^{(2)}$, respectively. We then aim to test the null hypothesis:

$$H_{0_{4,L}} : \mathcal{E}_L^{(1)} = \mathcal{E}_L^{(2)}.$$

Of course, $H_{0_{4,L}}$ corresponds to the hypothesis that the operators projecting into $\mathcal{E}_L^{(1)}$ and $\mathcal{E}_L^{(2)}$ are identical. This in turn translates into the condition that

$$\sum_{r=1}^{L} \gamma_r^{(1)}(t)\gamma_r^{(1)}(s) = \sum_{r=1}^{L} \gamma_r^{(2)}(t)\gamma_r^{(2)}(s) \qquad \text{for all } t, s \in [0,1].$$

Similar to above, a suitable test statistic is given by

$$D_{4,L} \stackrel{\text{def}}{=} \int\int \left\{ \sum_{r=1}^{L} \hat{\gamma}_r^{(1)}(t)\hat{\gamma}_r^{(1)}(s) - \sum_{r=1}^{L} \hat{\gamma}_r^{(2)}(t)\hat{\gamma}_r^{(2)}(s) \right\}^2 dt\, ds$$

and the null hypothesis is rejected if $D_{4,L} \geq \Delta_{4,L;1-\alpha}$, where $\Delta_{4,L;1-\alpha}$ denotes the critical value of the distribution of

$$\Delta_{4,L} \stackrel{\text{def}}{=} \int\int \left[\sum_{r=1}^{L} \{\hat{\gamma}_r^{(1)}(t)\hat{\gamma}_r^{(1)}(s) - \gamma_r^{(1)}(t)\gamma_r^{(1)}(s)\} - \sum_{r=1}^{L} \{\hat{\gamma}_r^{(2)}(t)\hat{\gamma}_r^{(2)}(s) - \gamma_r^{(2)}(t)\gamma_r^{(2)}(s)\} \right]^2 dt\, ds.$$

The distribution of $\Delta_{4,L}$ and, hence, its critical values are approximated by the bootstrap distribution of

$$\Delta_{4,L}^* \stackrel{\text{def}}{=} \int\int \left[\sum_{r=1}^{L} \{\hat{\gamma}_r^{(1)*}(t)\hat{\gamma}_r^{(1)*}(s) - \hat{\gamma}_r^{(1)}(t)\hat{\gamma}_r^{(1)}(s)\} - \sum_{r=1}^{L} \{\hat{\gamma}_r^{(2)*}(t)\hat{\gamma}_r^{(2)*}(s) - \hat{\gamma}_r^{(2)}(t)\hat{\gamma}_r^{(2)}(s)\} \right]^2 dt\, ds.$$

It will be shown in Theorem 3 below that under the null hypothesis, as well as under the alternative, the distributions of $n\Delta_1, n\Delta_{2,r}, n\Delta_{3,r}, n\Delta_{4,L}$ converge to continuous limit distributions which can be consistently approximated by the bootstrap distributions of $n\Delta_1^*, n\Delta_{2,r}^*, n\Delta_{3,r}^*, n\Delta_{4,L}^*$.

3.1. *Theoretical results.* Let $n = (n_1+n_2)/2$. We will assume that asymptotically $n_1 = n \cdot q_1$ and $n_2 = n \cdot q_2$ for some fixed proportions q_1 and q_2. We will then study the asymptotic behavior of our statistics as $n \to \infty$.

We will use $\mathcal{X}_1 = \{X_1^{(1)}, \ldots, X_{n_1}^{(1)}\}$ and $\mathcal{X}_2 = \{X_1^{(2)}, \ldots, X_{n_2}^{(2)}\}$ to denote the observed samples of random functions.

THEOREM 3. *Assume that* $\{X_1^{(1)}, \ldots, X_{n_1}^{(1)}\}$ *and* $\{X_1^{(2)}, \ldots, X_{n_2}^{(2)}\}$ *are two independent samples of random functions, each of which satisfies Assumption 1. As* $n \to \infty$ *we then obtain the following:*

(i) *There exists a nondegenerated, continuous probability distribution* F_1 *such that* $n\Delta_1 \xrightarrow{\mathcal{L}} F_1$, *and for any* $\delta > 0$,

$$|P(n\Delta_1 \geq \delta) - P(n\Delta_1^* \geq \delta | \mathcal{X}_1, \mathcal{X}_2)| = \mathcal{O}_p(1).$$

(ii) *If, furthermore,* $\lambda_{r-1}^{(1)} > \lambda_r^{(1)} > \lambda_{r+1}^{(1)}$ *and* $\lambda_{r-1}^{(2)} > \lambda_r^{(2)} > \lambda_{r+1}^{(2)}$ *hold for some fixed* $r = 1, 2, \ldots$, *there exist a nondegenerated, continuous probability distributions* $F_{k,r}$ *such that* $n\Delta_{k,r} \xrightarrow{\mathcal{L}} F_{k,r}$, $k = 2, 3$, *and for any* $\delta > 0$,

$$|P(n\Delta_{k,r} \geq \delta) - P(n\Delta_{k,r}^* \geq \delta | \mathcal{X}_1, \mathcal{X}_2)| = \mathcal{O}_p(1), \qquad k = 2, 3.$$

(iii) *If* $\lambda_r^{(1)} > \lambda_{r+1}^{(1)} > 0$ *and* $\lambda_r^{(2)} > \lambda_{r+1}^{(2)} > 0$ *hold for all* $r = 1, \ldots, L$, *there exists a nondegenerated, continuous probability distribution* $F_{4,L}$ *such that* $n\Delta_{4,L} \xrightarrow{\mathcal{L}} F_{4,L}$, *and for any* $\delta > 0$,

$$|P(n\Delta_{4,L} \geq \delta) - P(n\Delta_{4,L}^* \geq \delta | \mathcal{X}_1, \mathcal{X}_2)| = \mathcal{O}_p(1).$$

The structures of the distributions F_1, $F_{2,r}$, $F_{3,r}$, $F_{4,L}$ are derived in the proof of the theorem which can be found in the Appendix. They are obtained as limits of distributions of quadratic forms.

3.2. *Simulation study.* In this paragraph we illustrate the finite behavior of the proposed test. The basic simulation-setup (setup "a") is established as follows: the first sample is generated by the random combination of orthonormalized sine and cosine functions (Fourier functions) and the second sample is generated by the random combination of the same but shifted factor functions:

$$X_i^{(1)}(t_{ik}) = \beta_{1i}^{(1)}\sqrt{2}\sin(2\pi t_{ik}) + \beta_{2i}^{(1)}\sqrt{2}\cos(2\pi t_{ik}),$$
$$X_i^{(2)}(t_{ik}) = \beta_{1i}^{(2)}\sqrt{2}\sin\{2\pi(t_{ik}+\delta)\} + \beta_{2i}^{(2)}\sqrt{2}\cos\{2\pi(t_{ik}+\delta)\}.$$

The factor loadings are i.i.d. random variables with $\beta_{1i}^{(p)} \sim N(0, \lambda_1^{(p)})$ and $\beta_{2i}^{(p)} \sim N(0, \lambda_2^{(p)})$. The functions are generated on the equidistant grid $t_{ik} = t_k = k/T$, $k = 1, \ldots T = 100$, $i = 1, \ldots, n = 70$. The simulation setup is based

TABLE 1
The results of the simulations for $\alpha = 0.1$, $n = 70$, $T = 100$, *number of simulations* 250

Setup/shift	**0**	**0.05**	**0.1**	**0.15**	**0.2**	**0.25**
(a) 10, 5, 8, 4	0.13	0.41	0.85	0.96	1	1
(a) 4, 2, 2, 1	0.12	0.48	0.87	0.96	1	1
(a) 2, 1, 1.5, 2	0.14	0.372	0.704	0.872	0.92	0.9
(b) 10, 5, 8, 4 D_1	0.10	0.44	0.86	0.95	1	1
(b) 10, 5, 8, 4 D_2	1	1	1	1	1	1

on the fact that the error of the estimation of the eigenfunctions simulated by sine and cosine functions is, in particular, manifested by some shift of the estimated eigenfunctions. The focus of this simulation study is the test of common eigenfunctions.

For the presentation of results in Table 1, we use the following notation: "(a) $\lambda_1^{(1)}, \lambda_2^{(1)}, \lambda_2^{(2)}, \lambda_2^{(2)}$." The shift parameter δ is changing from 0 to 0.25 with the step 0.05. It should be mentioned that the shift $\delta = 0$ yields the simulation of level and setup with shift $\delta = 0.25$ yields the simulation of the alternative, where the two factor functions are exchanged.

In the second setup (setup "b") the first factor functions are the same and the second factor functions differ:

$$X_i^{(1)}(t_{ik}) = \beta_{1i}^{(1)}\sqrt{2}\sin(2\pi t_{ik}) + \beta_{2i}^{(1)}\sqrt{2}\cos(2\pi t_{ik}),$$
$$X_i^{(2)}(t_{ik}) = \beta_{1i}^{(2)}\sqrt{2}\sin\{2\pi(t_{ik}+\delta)\} + \beta_{2i}^{(2)}\sqrt{2}\sin\{4\pi(t_{ik}+\delta)\}.$$

In Table 1 we use the notation "(b) $\lambda_1^{(1)}, \lambda_2^{(1)}, \lambda_2^{(2)}, \lambda_2^{(2)}, D_r$." D_r means the test for the equality of the rth eigenfunction. In the bootstrap tests we used 500 bootstrap replications. The critical level in this simulation is $\alpha = 0.1$. The number of simulations is 250.

We can interpret Table 1 in the following way: In power simulations ($\delta \neq 0$) test behaves as expected: less powerful if the functions are "hardly distinguishable" (small shift, small difference in eigenvalues). The level approximation seems to be less precise if the difference in the eingenvalues $(\lambda_1^{(p)} - \lambda_2^{(p)})$ becomes smaller. This can be explained by relative small sample-size n, small number of bootstrap-replications and increasing estimation-error as argued in Theorem 2, assertion (iii).

In comparison to our general setup (4), we used an equidistant and common design for all functions. This simplification is necessary, it simplifies and speeds-up the simulations, in particular, using general random and observation-specific design makes the simulation computationally untractable.

Second, we omitted the additional observation error, this corresponds to the standard assumptions in the functional principal components theory. As

TABLE 2
The results of the simulation for $\alpha = 0.1$, $n = 70$, $T = 100$ with additional error in observation

Setup/shift	**0**	**0.05**	**0.1**	**0.15**	**0.2**	**0.25**
(a) 10, 5, 8, 4	0.09	0.35	0.64	0.92	0.94	0.97

argued in Section 2.2, the inference based on the directly observed functions and estimated functions X_i is first-order equivalent under mild conditions implied by Theorems 1 and 2. In order to illustrate this theoretical result in the simulation, we used the following setup:

$$X_i^{(1)}(t_{ik}) = \beta_{1i}^{(1)}\sqrt{2}\sin(2\pi t_{ik}) + \beta_{2i}^{(1)}\sqrt{2}\cos(2\pi t_{ik}) + \varepsilon_{ik}^{(1)},$$
$$X_i^{(2)}(t_{ik}) = \beta_{1i}^{(2)}\sqrt{2}\sin\{2\pi(t_{ik}+\delta)\} + \beta_{2i}^{(2)}\sqrt{2}\cos\{2\pi(t_{ik}+\delta)\} + \varepsilon_{ik}^{(2)},$$

where $\varepsilon_{ik}^{(p)} \sim N(0, 0.25)$, $p = 1, 2$, all other parameters remain the same as in the simulation setup "a." Using this setup, we recalculate the simulation presented in the second "row" of Table 1, for estimation of the functions $X_i^{(p)}, p = 1, 2$, we used the Nadaraya–Watson estimation with Epanechnikov kernel and bandwidth $b = 0.05$. We run the simulations with various bandwidths, the choice of the bandwidth does not have a strong influence on results except by oversmoothing (large bandwidths). The results are printed in Table 2. As we can see, the difference of the simulation results using estimated functions is not significant in comparison to the results printed in the second line of Table 1—directly observed functional values.

The last limitation of this simulation study is the choice of a particular alternative. A more general setup of this simulation study might be based on the following model: $X_i^{(1)}(t) = \beta_{1i}^{(1)}\gamma_1^{(1)}(t) + \beta_{2i}^{(1)}\gamma_2^{(1)}(t)$, $X_i^{(2)}(t) = \beta_{1i}^{(2)}\gamma_1^{(2)}(t) + \beta_{2i}^{(2)}\gamma_2^{(2)}(t)$, where $\gamma_1^{(1)}, \gamma_1^{(2)}, \gamma_2^{(1)}$ and g are mutually orthogonal functions on $L^2[0,1]$ and $\gamma_2^{(2)} = (1+\upsilon^2)^{-1/2}\{\gamma_2^{(1)} + \upsilon g\}$. Basically we create the alternative by the contamination of one of the "eigenfunctions" (in our case the second one) in the direction g and ensure $\|\gamma_2^{(2)}\| = 1$. The amount of the contamination is controlled by the parameter υ. Note that the exact squared integral difference $\|\gamma_2^{(1)} - \gamma_2^{(2)}\|^2$ does not depend on function g. Thus, in the "functional sense" particular "direction of the alternative hypothesis" represented by the function g has no impact on the power of the test. However, since we are using a nonparametric estimation technique, we might expect that rough (highly fluctuating) functions g will yield higher error of estimation and, hence, decrease the precision (and power) of the test. Finally, a higher number of factor functions (L) in simulation may cause less precise approximation of critical values and more bootstrap replications and

larger sample-size may be needed. This can also be expected from Theorem 2 in Section 2.2—the variance of the estimated eigenfunctions depends on all eigenfunctions corresponding to nonzero eingenvalues.

4. Implied volatility analysis. In this section we present an application of the method discussed in previous sections to the implied volatilities of European options on the German stock index (ODAX). Implied volatilities are derived from the Black–Scholes (BS) pricing formula for European options; see Black and Scholes (1973). European call and put options are derivatives written on an underlying asset with price process S_i, which yield the pay-off $\max(S_I - K, 0)$ and $\max(K - S_I, 0)$, respectively. Here i denotes the current day, I the expiration day and K the strike price. Time to maturity is defined as $\tau = I - i$. The BS pricing formula for a Call option is

$$C_i(S_i, K, \tau, r, \sigma) = S_i \Phi(d_1) - K e^{-r\tau} \Phi(d_2), \tag{14}$$

where $d_1 = \frac{\ln(S_i/K) + (r + \sigma^2/2)\tau}{\sigma\sqrt{\tau}}$, $d_2 = d_1 - \sigma\sqrt{\tau}$, r is the risk-free interest rate, σ is the (unknown and constant) volatility parameter, and Φ denotes the c.d.f. of a standard normal distributed random variable. In (14) we assume the zero-dividend case. The Put option price P_i can be obtained from the put–call parity $P_i = C_i - S_i + e^{-\tau r} K$.

The implied volatility $\tilde{\sigma}$ is defined as the volatility σ, for which the BS price C_i in (14) equals the price $\tilde{C}_i$ observed on the market. For a single asset, we obtain at each time point (day i) and for each maturity τ a IV function $\tilde{\sigma}_i^{\tau}(K)$. Practitioners often rescale the strike dimension by plotting this surface in terms of (futures) moneyness $\kappa = K/F_i(\tau)$, where $F_i(\tau) = S_i e^{r\tau}$.

Clearly, for given parameters S_i, r, K, τ the mapping from prices to IVs is a one-to-one mapping. The IV is often used for quoting the European options in financial practice, since it reflects the "uncertainty" of the financial market better than the option prices. It is also known that if the stock price drops, the IV raises (so-called leverage effect), motivates hedging strategies based on IVs. Consequently, for the purpose of this application, we will regard the BS–IV as an individual financial variable. The practical relevance of such an approach is justified by the volatility based financial products such as VDAX, which are commonly traded on the option markets.

The goal of this analysis is to study the dynamics of the IV functions for different maturities. More specifically, our aim is to construct low dimensional factor models based on the truncated Karhunen–Loève expansions (1) for the log-returns of the IV functions of options with different maturities and compare these factor models using the methodology presented in the previous sections. Analysis of IVs based on a low-dimensional factor model gives directly a descriptive insight into the structure of distribution

of the log-IV-returns—structure of the factors and empirical distribution of the factor loadings may be a good starting point for further pricing models. In practice, such a factor model can also be used in Monte Carlo based pricing methods and for risk-management (hedging) purposes. For comprehensive monographs on IV and IV-factor models, see Hafner (2004) or Fengler (2005b).

The idea of constructing and analyzing the factor models for log-IV-returns for different maturities was originally proposed by Fengler, Härdle and Villa (2003), who studied the dynamics of the IV via PCA on discretized IV functions for different maturity groups and tested the Common Principal Components (CPC) hypotheses (equality of eigenvectors and eigenspaces for different groups). Fengler, Härdle and Villa (2003) proposed a PCA-based factor model for log-IV-returns on (short) maturities 1, 2 and 3 months and grid of moneyness $[0.85, 0.9, 0.95, 1, 1.05, 1.1]$. They showed that the factor functions do not significantly differ and only the factor loadings differ across maturity groups. Their method relies on the CPC methodology introduced by Flury (1988) which is based on maximum likelihood estimation under the assumption of multivariate normality. The log-IV-returns are extracted by the two-dimensional Nadaraya–Watson estimate.

The main aim of this application is to reconsider their results in a functional sense. Doing so, we overcome two basic weaknesses of their approach. First, the factor model proposed by Fengler, Härdle and Villa (2003) is performed only on a sparse design of moneyness. However, in practice (e.g., in Monte Carlo pricing methods), evaluation of the model on a fine grid is needed. Using the functional PCA approach, we may overcome this difficulty and evaluate the factor model on an arbitrary fine grid. The second difficulty of the procedure proposed by Fengler, Härdle and Villa (2003) stems from the data design—on the exchange we cannot observe options with desired maturity on each day and we need to estimate them from the IV-functions with maturities observed on the particular day. Consequently, the two-dimensional Nadaraya–Watson estimator proposed by Fengler, Härdle and Villa (2003) results essentially in the (weighted) average of the IVs (with closest maturities) observed on a particular day, which may affect the test of the common eigenfunction hypothesis. We use the linear interpolation scheme in the *total variance* $\sigma^2_{\mathrm{TOT},i}(\kappa,\tau) \stackrel{\mathrm{def}}{=} (\sigma_i^\tau(\kappa))^2\tau$, in order to recover the IV functions with fixed maturity (on day i). This interpolation scheme is based on the arbitrage arguments originally proposed by Kahalé (2004) for zero-dividend and zero-interest rate case and generalized for deterministic interest rate by Fengler (2005a). More precisely, having IVs with maturities observed on a particular day i: $\tilde\sigma_i^{\tau_{j_i}}(\kappa)$, $j_i = 1, \ldots, p_{\tau_i}$, we calculate the corresponding total variance $\tilde\sigma_{\mathrm{TOT},i}(\kappa,\tau_{j_i})$. From these total variances

we linearly interpolate the total variance with the desired maturity from the nearest maturities observed on day i. The total variance can be easily transformed to corresponding IV $\tilde{\sigma}_i^\tau(\kappa)$. As the last step, we calculate the log-returns $\Delta \log \tilde{\sigma}_i^\tau(\kappa) \stackrel{\text{def}}{=} \log \tilde{\sigma}_{i+1}^\tau(\kappa) - \log \tilde{\sigma}_i^\tau(\kappa)$. The log-IV-returns are observed for each maturity τ on a discrete grid κ_{ik}^τ. We assume that observed log-IV-return $\Delta \log \tilde{\sigma}_i^\tau(\kappa_{ik}^\tau)$ consists of true log-return of the IV function denoted by $\Delta \log \sigma_i^\tau(\kappa_{ik}^\tau)$ and possibly of some additional error ε_{ik}^τ. By setting $Y_{ik}^\tau := \Delta \log \tilde{\sigma}_i^\tau(\kappa_{ik}^\tau)$, $X_i^\tau(\kappa) := \Delta \log \sigma_i^\tau(\kappa)$, we obtain an analogue of the model (4) with the argument κ:

$$Y_{ik}^\tau = X_i^\tau(\kappa_{ik}) + \varepsilon_{ik}^\tau, \qquad i = 1, \ldots, n_\tau. \tag{15}$$

In order to simplify the notation and make the connection with the theoretical part clear, we will use the notation of (15).

For our analysis we use a recent data set containing daily data from January 2004 to June 2004 from the German–Swiss exchange (EUREX). Violations of the arbitrage-free assumptions ("obvious" errors in data) were corrected using the procedure proposed by Fengler (2005a). Similarly to Fengler, Härdle and Villa (2003), we excluded options with maturity smaller then 10 days, since these option-prices are known to be very noisy, partially because of a special and arbitrary setup in the pricing systems of the dealers. Using the interpolation scheme described above, we calculate the log-IV-returns for two maturity groups: "1M" group with maturity $\tau = 0.12$ (measured in years) and "3M" group with maturity $\tau = 0.36$. The observed log-IV-returns are denoted by Y_{ik}^{1M}, $k = 1, \ldots, K_i^{1M}$, Y_{ik}^{3M}, $k = 1, \ldots, K_i^{3M}$. Since we ensured that for no i, the interpolation procedure uses data with the same maturity for both groups, this procedure has no impact on the independence of both samples.

The underlying models based on the truncated version of (3) are as follows:

$$X_i^{1M}(\kappa) = \bar{X}^{1M}(\kappa) + \sum_{r=1}^{L_{1M}} \hat{\beta}_{ri}^{1M} \widehat{\gamma}_r^{1M}(\kappa), \qquad i = 1, \ldots, n_{1M}, \tag{16}$$

$$X_i^{3M}(\kappa) = \bar{X}^{3M}(\kappa) + \sum_{r=1}^{L_{3M}} \hat{\beta}_{ri}^{3M} \widehat{\gamma}_r^{3M}(\kappa), \qquad i = 1, \ldots, n_{3M}. \tag{17}$$

Models (16) and (17) can serve, for example, in a Monte Carlo pricing tool in the risk management for pricing exotic options where the whole path of implied volatilities is needed to determine the price. Estimating the factor functions in (16) and (17) by eigenfunctions displayed in Figure 1, we only need to fit the (estimated) factor loadings $\hat{\beta}_{ji}^{1M}$ and $\hat{\beta}_{ji}^{3M}$. The pillar of the model is the dimension reduction. Keeping the factor function fixed for a certain time period, we need to analyze (two) multivariate random processes

of the factor loadings. For the purposes of this paper we will focus on the comparison of factors from models (16) and (17) and the technical details of the factor loading analysis will not be discussed here, since in this respect we refer to Fengler, Härdle and Villa (2003), who proposed to fit the factor loadings by centered normal distributions with diagonal variance matrix containing the corresponding eigenvalues. For a deeper discussion of the fitting of factor loadings using a more sophisticated approach, basically based on (possibly multivariate) GARCH models; see Fengler (2005b).

From our data set we obtained 88 functional observations for the 1M group (n_{1M}) and 125 observations for the 3M group (n_{3M}). We will estimate the model on the interval for futures moneyness $\kappa \in [0.8, 1.1]$. In comparison to Fengler, Härdle and Villa (2003), we may estimate models (16) and (17) on an arbitrary fine grid (we used an equidistant grid of 500 points on the interval $[0.8, 1.1]$). For illustration, the Nadaraya–Watson (NW) estimator of resulting log-returns is plotted in Figure 2. The smoothing parameters have been chosen in accordance with the requirements in Section 2.2. As argued in Section 2.2, we should use small smoothing parameters in order to avoid a possible bias in the estimated eigenfunctions. Thus, we use for each i essentially the smallest bandwidth b_i that guarantees that estimator $\hat{X}_i$ is defined on the entire support $[0.8, 1.1]$.

Using the procedures described in Section 2.1, we first estimate the eigenfunctions of both maturity groups. The estimated eigenfunctions are plotted in Figure 1. The structure of the eigenfunctions is in accordance with other empirical studies on IV-surfaces. For a deeper discussion and economical interpretation, see, for example, Fengler, Härdle and Mammen (2007) or Fengler, Härdle and Villa (2003).

Clearly, the ratio of the variance explained by the kth factor function is given by the quantity $\hat{\nu}_k^{1M} = \hat{\lambda}_k^{1M} / \sum_{j=1}^{n_{1M}} \hat{\lambda}_j^{1M}$ for the 1M group and, correspondingly, by $\hat{\nu}_k^{3M}$ for the 3M group. In Table 3 we list the contributions of the factor functions. Looking at Table 3, we can see that 4th factor functions explain less than 1% of the variation. This number was the "threshold" for the choice of L_{1M} and L_{2M}.

We can observe (see Figure 1) that the factor functions for both groups are similar. Thus, in the next step we use the bootstrap test for testing the

TABLE 3
Variance explained by the eigenfunctions

	Var. explained 1M	**Var. explained 3M**
$\hat{\nu}_1^\tau$	89.9%	93.0%
$\hat{\nu}_2^\tau$	7.7%	4.2%
$\hat{\nu}_3^\tau$	1.7%	1.0%
$\hat{\nu}_4^\tau$	0.6%	0.4%

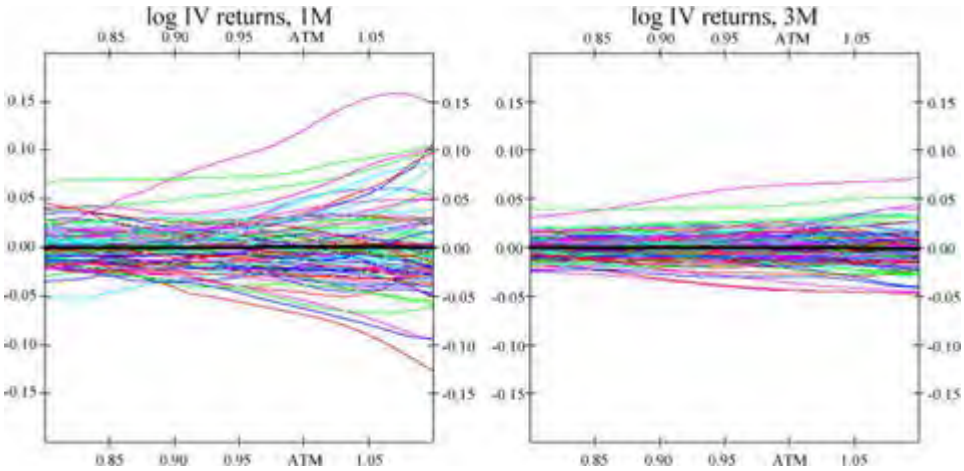


FIG. 2. *Nadaraya–Watson estimate of the log-IV-returns for maturity 1M (left figure) and 3M (right figure). The bold line is the sample mean of the corresponding group.*

equality of the factor functions. We use 2000 bootstrap replications. The test of equality of the eigenfunctions was rejected for the first eigenfunction for the analyzed time period (January 2004–June 2004) at a significance level $\alpha = 0.05$ (P-value 0.01). We may conclude that the (first) factor functions are not identical in the factor model for both maturity groups. However, from a practical point of view, we are more interested in checking the appropriateness of the entire models for a fixed number of factors: $L = 2$ or $L = 3$ in (16) and (17). This requirement translates into the testing of the equality of eigenspaces. Thus, in the next step we use the same setup (2000 bootstrap replications) to test the hypotheses that the first two and first three eigenfunctions span the same eigenspaces $\mathcal{E}_L^{1M}$ and $\mathcal{E}_L^{3M}$. None of the hypotheses for $L = 2$ and $L = 3$ is rejected at significance level $\alpha = 0.05$ (P-value is 0.61 for $L = 2$ and 0.09 for $L = 3$). Summarizing, even in the functional sense we have no significant reason to reject the hypothesis of common eigenspaces for these two maturity groups. Using this hypothesis, the factors governing the movement of the returns of IV surface are invariant to time to maturity, only their relative importance can vary. This leads to the common factor model: $X_i^\tau(\kappa) = \bar{X}^\tau(\kappa) + \sum_{r=1}^{L_\tau} \hat{\beta}_{ri}^\tau \widehat{\gamma_r}(\kappa), i = 1, \ldots, n_\tau,\ \tau = 1M, 3M$, where $\gamma_r := \gamma_r^{1M} = \gamma_r^{3M}$. Beside contributing to the understanding of the structure of the IV function dynamics, the common factor model helps us to reduce the number of functional factors by half compared to models (16) and (17). Furthermore, from the technical point of view, we also obtain an additional dimension reduction and higher estimation precision, since under this hypothesis we may estimate the eigenfunctions from the (individually centered) pooled sample $X_i(\kappa)^{1M}, i = 1, \ldots, n_{1M}$, $X_i^{3M}(\kappa), i =$

$1, \ldots, n_{3M}$. The main improvement compared to the multivariate study by Fengler, Härdle and Villa (2003) is that our test is performed in the functional sense – it does not depend on particular discretization and our factor model can be evaluated on an arbitrary fine grid.

APPENDIX: MATHEMATICAL PROOFS

In the following, $\|v\| = (\int_0^1 v(t)^2\,dt)^{1/2}$ will denote the L^2-norm for any square integrable function v. At the same time, $\|a\| = (\frac{1}{k}\sum_{i=1}^k a_i^2)^{1/2}$ will indicate the Euclidean norm, whenever $a \in \mathbb{R}^k$ is a k-vector for some $k \in \mathbb{N}$.

In the proof of Theorem 1, E_ε and Var_ε denote expectation and variance with respect to ε only (i.e., conditional on t_{ij} and X_i).

PROOF OF THEOREM 1. Recall the definition of the $\chi_i(t)$ and note that $\chi_i(t) = \chi_i^X(t) + \chi_i^\varepsilon(t)$, where

$$\chi_i^\varepsilon(t) = \sum_{j=1}^{T_i} \varepsilon_{i(j)} I\left(t \in \left[\frac{t_{i(j-1)} + t_{i(j)}}{2}, \frac{t_{i(j)} + t_{i(j+1)}}{2}\right)\right),$$

as well as

$$\chi_i^X(t) = \sum_{j=1}^{T_i} X_i(t_{i(j)}) I\left(t \in \left[\frac{t_{i(j-1)} + t_{i(j)}}{2}, \frac{t_{i(j)} + t_{i(j+1)}}{2}\right)\right)$$

for $t \in [0,1]$, $t_{i(0)} = -t_{i(1)}$ and $t_{i(T_i+1)} = 2 - t_{i(T_i)}$. Similarly, $\chi_i^*(t) = \chi_i^{X_*}(t) + \chi_i^{\varepsilon *}(t)$.

By Assumption 2, $\mathrm{E}(|t_{i(j)} - t_{i(j-1)}|^s) = \mathcal{O}(T^{-s})$ for $s = 1, \ldots, 4$, and the convergence is uniform in $j < n$. Our assumptions on the structure of X_i together with some straightforward Taylor expansions then lead to

$$\langle \chi_i, \chi_j \rangle = \langle X_i, X_j \rangle + \mathcal{O}_p(1/T)$$

and

$$\langle \chi_i, \chi_i^* \rangle = \|X_i\|^2 + \mathcal{O}_p(1/T).$$

Moreover,

$$\mathrm{E}_\varepsilon(\langle \chi_i^\varepsilon, \chi_j^X \rangle) = 0, \qquad \mathrm{E}_\varepsilon(\|\chi_i^\varepsilon\|^2) = \sigma_i^2,$$

$$\mathrm{E}_\varepsilon(\langle \chi_i^\varepsilon, \chi_i^{\varepsilon *} \rangle) = 0, \qquad \mathrm{E}_\varepsilon(\langle \chi_i^\varepsilon, \chi_i^{\varepsilon *} \rangle^2) = \mathcal{O}_p(1/T),$$

$$\mathrm{E}_\varepsilon(\langle \chi_i^\varepsilon, \chi_j^X \rangle^2) = \mathcal{O}_p(1/T), \qquad \mathrm{E}_\varepsilon(\langle \chi_i^\varepsilon, \chi_j^X \rangle \langle \chi_k^\varepsilon, \chi_l^X \rangle) = 0 \quad \text{for } i \neq k,$$

$$\mathrm{E}_\varepsilon(\langle \chi_i^\varepsilon, \chi_j^\varepsilon \rangle \langle \chi_i^\varepsilon, \chi_k^\varepsilon \rangle) = 0 \quad \text{for } j \neq k \text{ and } \mathrm{E}_\varepsilon(\|\chi_i^\varepsilon\|^4) = \mathcal{O}_p(1)$$

hold (uniformly) for all $i, j = 1, \ldots, n$.

Consequently, $\mathrm{E}_\varepsilon(\|\bar{\chi}\|^2 - \|\bar{X}\|^2) = \mathcal{O}_p(T^{-1} + n^{-1})$.

When using these relations, it is easily seen that for all $i, j = 1, \dots, n$

$$\begin{aligned} \widehat{M}_{ij} - M_{ij} &= \mathcal{O}_p(T^{-1/2} + n^{-1}) \quad \text{and} \\ \operatorname{tr}\{(\widehat{M} - M)^2\}^{1/2} &= \mathcal{O}_p(1 + nT^{-1/2}). \end{aligned} \tag{18}$$

Since the orthonormal eigenvectors p_q of M satisfy $\|p_q\| = 1$, we furthermore obtain for any $i = 1, \dots, n$ and all $q = 1, 2, \dots$

$$\sum_{j=1}^{n} p_{jq} \left\{ \widehat{M}_{ij} - M_{ij} - \int_0^1 \chi_i^{\varepsilon}(t) \chi_j^X(t)\, dt \right\} = \mathcal{O}_p(T^{-1/2} + n^{-1/2}), \tag{19}$$

as well as

$$\sum_{j=1}^{n} p_{jq} \int_0^1 \chi_i^{\varepsilon}(t) \chi_j^X(t)\, dt = \mathcal{O}_p\left(\frac{n^{1/2}}{T^{1/2}} \right) \tag{20}$$

and

$$\sum_{i=1}^{n} a_i \sum_{j=1}^{n} p_{jq} \int_0^1 \chi_i^{\varepsilon}(t) \chi_j^X(t)\, dt = \mathcal{O}_p\left(\frac{n^{1/2}}{T^{1/2}} \right) \tag{21}$$

for any further vector a with $\|a\| = 1$.

Recall that the jth largest eigenvalue l_j satisfies $n\hat{\lambda}_j = l_j$. Since by assumption $\inf_{s \neq r} |\lambda_r - \lambda_s| > 0$, the results of Dauxois, Pousse and Romain (1982) imply that $\hat{\lambda}_r$ converges to λ_r as $n \to \infty$, and $\sup_{s \neq r} \frac{1}{|\hat{\lambda}_r - \hat{\lambda}_s|} = \mathcal{O}_p(1)$, which leads to $\sup_{s \neq r} \frac{1}{|l_r - l_s|} = \mathcal{O}_p(1/n)$. Assertion (a) of Lemma A of Kneip and Utikal (2001) together with (18)–(21) then implies that

$$\begin{aligned} \left| \hat{\lambda}_r - \frac{\hat{l}_r}{n} \right| &= n^{-1} |l_r - \hat{l}_r| = n^{-1} |p_r^{\top} (\widehat{M} - M) p_r| + \mathcal{O}_p(T^{-1} + n^{-1}) \\ &= \mathcal{O}_p\{(nT)^{-1/2} + T^{-1} + n^{-1}\}. \end{aligned} \tag{22}$$

When analyzing the difference between the estimated and true eigenvectors $\hat{p}_r$ and p_r, assertion (b) of Lemma A of Kneip and Utikal (2001) together with (18) lead to

$$\hat{p}_r - p_r = -\mathcal{S}_r(\widehat{M} - M) p_r + \mathcal{R}_r, \qquad \text{with } \|\mathcal{R}_r\| = \mathcal{O}_p(T^{-1} + n^{-1}) \tag{23}$$

and $\mathcal{S}_r = \sum_{s \neq r} \frac{1}{l_s - l_r} p_s p_s^{\top}$. Since $\sup_{\|a\|=1} a^{\top} \mathcal{S}_r a \leq \sup_{s \neq r} \frac{1}{|l_r - l_s|} = \mathcal{O}_p(1/n)$, we can conclude that

$$\|\hat{p}_r - p_r\| = \mathcal{O}_p(T^{-1/2} + n^{-1}), \tag{24}$$

and our assertion on the sequence $n^{-1} \sum_i (\hat{\beta}_{ri} - \hat{\beta}_{ri;T})^2$ is an immediate consequence.

Let us now consider assertion (ii). The well-known properties of local linear estimators imply that $|\mathrm{E}_\varepsilon\{\hat{X}_i(t) - X_i(t)\}| = \mathcal{O}_p(b^2)$, as well as $\mathrm{Var}_\varepsilon\{\hat{X}_i(t)\} = \mathcal{O}_p\{Tb\}$, and the convergence is uniform for all i, n. Furthermore, due to the independence of the error term ε_{ij}, $\mathrm{Cov}_\varepsilon\{\hat{X}_i(t), \hat{X}_j(t)\} = 0$ for $i \neq j$. Therefore,

$$\left| \hat{\gamma}_r(t) - \frac{1}{\sqrt{l_r}} \sum_{i=1}^{n} p_{ir} \hat{X}_i(t) \right| = \mathcal{O}_p\left(b^2 + \frac{1}{\sqrt{nTb}} \right).$$

On the other hand, (18)–(24) imply that with $\hat{X}(t) = (\hat{X}_1(t), \ldots, \hat{X}_n(t))^\top$

$$\begin{aligned}
&\left| \hat{\gamma}_{r;T}(t) - \frac{1}{\sqrt{l_r}} \sum_{i=1}^{n} p_{ir} \hat{X}_i(t) \right| \\
&\quad = \left| \frac{1}{\sqrt{l_r}} \sum_{i=1}^{n} (\hat{p}_{ir} - p_{ir}) X_i(t) + \frac{1}{\sqrt{l_r}} \sum_{i=1}^{n} (\hat{p}_{ir} - p_{ir}) \{\hat{X}_i(t) - X_i(t)\} \right| \\
&\qquad + \mathcal{O}_p(T^{-1} + n^{-1}) \\
&\quad = \frac{\|\mathcal{S}_r X(t)\|}{\sqrt{l_r}} \left| p_r^\top (\hat{M} - M) \mathcal{S}_r \frac{X(t)}{\|\mathcal{S}_r X(t)\|} \right| \\
&\qquad + \mathcal{O}_p(b^2 T^{-1/2} + T^{-1} b^{-1/2} + n^{-1}) \\
&\quad = \mathcal{O}_p(n^{-1/2} T^{-1/2} + b^2 T^{-1/2} + T^{-1} b^{-1/2} + n^{-1}).
\end{aligned}$$

This proves the theorem. □

Proof of Theorem 2. First consider assertion (i). By definition,

$$\bar{X}(t) - \mu(t) = n^{-1} \sum_{i=1}^{n} \{X_i(t) - \mu(t)\} = \sum_r \left(n^{-1} \sum_{i=1}^{n} \beta_{ri} \right) \gamma_r(t).$$

Recall that, by assumption, β_{ri} are independent, zero mean random variables with variance λ_r, and that the above series converges with probability 1. When defining the truncated series

$$V(q) = \sum_{r=1}^{q} \left(n^{-1} \sum_{i=1}^{n} \beta_{ri} \right) \gamma_r(t),$$

standard central limit theorems therefore imply that $\sqrt{n} V(q)$ is asymptotically $N(0, \sum_{r=1}^{q} \lambda_r \gamma_r(t)^2)$ distributed for any possible $q \in \mathbb{N}$.

The assertion of a $N(0, \sum_{r=1}^{\infty} \lambda_r \gamma_r(t)^2)$ limiting distribution now is a consequence of the fact that for all $\delta_1, \delta_2 > 0$ there exists a q_δ such that $P\{|\sqrt{n} V(q) - \sqrt{n} \sum_r (n^{-1} \sum_{i=1}^{n} \beta_{ri}) \gamma_r(t)| > \delta_1\} < \delta_2$ for all $q \geq q_\delta$ and all n sufficiently large.

In order to prove assertions (i) and (ii), consider some fixed $r \in \{1, 2, \ldots\}$ with $\lambda_{r-1} > \lambda_r > \lambda_{r+1}$. Note that Γ as well as $\hat{\Gamma}_n$ are nuclear, self-adjoint and non-negative linear operators with $\Gamma v = \int \sigma(t,s)v(s)\,ds$ and $\hat{\Gamma}_n v = \int \hat{\sigma}(t,s)v(s)\,ds$, $v \in L^2[0,1]$. For $m \in \mathbb{N}$, let Π_m denote the orthogonal projector from $L^2[0,1]$ into the m-dimensional linear space spanned by $\{\gamma_1, \ldots, \gamma_m\}$, that is, $\Pi_m v = \sum_{j=1}^m \langle v, \gamma_j\rangle \gamma_j$, $v \in L^2[0,1]$. Now consider the operator $\Pi_m \hat{\Gamma}_n \Pi_m$, as well as its eigenvalues and corresponding eigenfunctions denoted by $\hat{\lambda}_{1,m} \geq \hat{\lambda}_{2,m} \geq \cdots$ and $\hat{\gamma}_{1,m}, \hat{\gamma}_{2,m}, \ldots$, respectively. It follows from well-known results in the Hilbert space theory that $\Pi_m \hat{\Gamma}_n \Pi_m$ converges strongly to $\hat{\Gamma}_n$ as $m \to \infty$. Furthermore, we obtain (Rayleigh–Ritz theorem)

$$\lim_{m\to\infty} \hat{\lambda}_{r,m} = \lambda_r \quad \text{and} \quad \lim_{m\to\infty} \|\hat{\gamma}_r - \hat{\gamma}_{r,m}\| = 0 \qquad \text{if } \hat{\lambda}_{r-1} > \hat{\lambda}_r > \hat{\lambda}_{r+1}. \tag{25}$$

Note that under the above condition $\hat{\gamma}_r$ is uniquely determined up to sign, and recall that we always assume that the right "versions" (with respect to sign) are used so that $\langle \hat{\gamma}_r, \hat{\gamma}_{r,m}\rangle \geq 0$. By definition, $\beta_{ji} = \int \gamma_j(t)\{X_i(t) - \mu(t)\}\,dt$, and therefore, $\int \gamma_j(t)\{X_i(t) - \bar{X}(t)\}\,dt = \beta_{ji} - \bar{\beta}_j$, as well as $X_i - \bar{X} = \sum_j (\beta_{ji} - \bar{\beta}_j)\gamma_j$, where $\bar{\beta}_j = \frac{1}{n}\sum_{i=1}^n \beta_{ji}$. When analyzing the structure of $\Pi_m \hat{\Gamma}_n \Pi_m$ more deeply, we can verify that $\Pi_m \hat{\Gamma}_n \Pi_m v = \int \hat{\sigma}_m(t,s)v(s)\,ds$, $v \in L^2[0,1]$, with

$$\hat{\sigma}_m(t,s) = g_m(t)^\top \hat{\Sigma}_m g_m(s),$$

where $g_m(t) = (\gamma_1(t), \ldots, \gamma_m(t))^\top$, and where $\hat{\Sigma}_m$ is the $m \times m$ matrix with elements $\{\frac{1}{n}\sum_{i=1}^n (\beta_{ji} - \bar{\beta}_j)(\beta_{ki} - \bar{\beta}_k)\}_{j,k=1,\ldots,m}$. Let $\lambda_1(\hat{\Sigma}_m) \geq \lambda_2(\hat{\Sigma}_m) \geq \cdots \geq \lambda_m(\hat{\Sigma}_m)$ and $\hat{\zeta}_{1,m}, \ldots, \hat{\zeta}_{m,m}$ denote eigenvalues and corresponding eigenvectors of $\hat{\Sigma}_m$. Some straightforward algebra then shows that

$$\hat{\lambda}_{r,m} = \lambda_r(\hat{\Sigma}_m), \qquad \hat{\gamma}_{r,m} = g_m(t)^\top \hat{\zeta}_{r,m}. \tag{26}$$

We will use Σ_m to represent the $m \times m$ diagonal matrix with diagonal entries $\lambda_1 \geq \cdots \geq \lambda_m$. Obviously, the corresponding eigenvectors are given by the m-dimensional unit vectors denoted by $e_{1,m}, \ldots, e_{m,m}$. Lemma A of Kneip and Utikal (2001) now implies that the differences between eigenvalues and eigenvectors of Σ_m and $\hat{\Sigma}_m$ can be bounded by

$$\begin{aligned} \hat{\lambda}_{r,m} - \lambda_r &= \operatorname{tr}\{e_{r,m}e_{r,m}^\top(\hat{\Sigma}_m - \Sigma_m)\} + \tilde{R}_{r,m}, \\ &\text{with } \tilde{R}_{r,m} \leq \frac{6\sup_{\|a\|=1} a^\top(\hat{\Sigma}_m - \Sigma_m)^2 a}{\min_s |\lambda_s - \lambda_r|}, \end{aligned} \tag{27}$$

$$\begin{aligned} \hat{\zeta}_{r,m} - e_{r,m} &= -S_{r,m}(\hat{\Sigma}_m - \Sigma_m)e_{r,m} + R^*_{r,m}, \\ &\text{with } \|R^*_{r,m}\| \leq \frac{6\sup_{\|a\|=1} a^\top(\hat{\Sigma}_m - \Sigma_m)^2 a}{\min_s |\lambda_s - \lambda_r|^2}, \end{aligned} \tag{28}$$

where $S_{r,m} = \sum_{s \neq r} \frac{1}{\lambda_s - \lambda_r} e_{s,m} e_{s,m}^\top$.

Assumption 1 implies $\mathrm{E}(\bar{\beta}_r) = 0$, $\mathrm{Var}(\bar{\beta}_r) = \frac{\lambda_r}{n}$, and with $\delta_{ii} = 1$, as well as $\delta_{ij} = 0$ for $i \neq j$, we obtain

$$
\begin{aligned}
&\mathrm{E}\Big\{ \sup_{\|a\|=1} a^\top (\hat{\Sigma}_m - \Sigma_m)^2 a \Big\} \\
&\quad \leq \mathrm{E}\{\mathrm{tr}[(\hat{\Sigma}_m - \Sigma_m)^2]\} \\
&\quad = \mathrm{E}\Bigg\{ \sum_{j,k=1}^{m} \Bigg[\frac{1}{n} \sum_{i=1}^{n} (\beta_{ji} - \bar{\beta}_j)(\beta_{ki} - \bar{\beta}_k) - \delta_{jk} \lambda_j \Bigg]^2 \Bigg\} \\
&\quad \leq \mathrm{E}\Bigg\{ \sum_{j,k=1}^{\infty} \Bigg[\frac{1}{n} \sum_{i=1}^{n} (\beta_{ji} - \bar{\beta}_j)(\beta_{ki} - \bar{\beta}_k) - \delta_{jk} \lambda_j \Bigg]^2 \Bigg\} \\
&\quad = \frac{1}{n} \Bigg(\sum_j \sum_k \mathrm{E}\{\beta_{ji}^2 \beta_{ki}^2\} \Bigg) + \mathcal{O}(n^{-1}) = \mathcal{O}(n^{-1}),
\end{aligned}
\tag{29}
$$

for all m. Since $\mathrm{tr}\{e_{r,m} e_{r,m}^\top (\hat{\Sigma}_m - \Sigma_m)\} = \frac{1}{n} \sum_{i=1}^{n} (\beta_{ri} - \bar{\beta}_r)^2 - \lambda_r$, (25), (26), (27) and (29) together with standard central limit theorems imply that

$$
\begin{aligned}
\sqrt{n}(\hat{\lambda}_r - \lambda_r) &= \frac{1}{\sqrt{n}} \sum_{i=1}^{n} (\beta_{ri} - \bar{\beta}_r)^2 - \lambda_r + \mathcal{O}_p(n^{-1/2}) \\
&= \frac{1}{\sqrt{n}} \sum_{i=1}^{n} [(\beta_{ri})^2 - \mathrm{E}\{(\beta_{ri})^2\}] + \mathcal{O}_p(n^{-1/2}) \\
&\xrightarrow{\mathcal{L}} N(0, \Lambda_r).
\end{aligned}
\tag{30}
$$

It remains to prove assertion (iii). Relations (26) and (28) lead to

$$
\begin{aligned}
\hat{\gamma}_{r,m}(t) - \gamma_r(t) &= g_m(t)^\top (\hat{\zeta}_{r,m} - e_{r,m}) \\
&= -\sum_{s \neq r}^{m} \Bigg\{ \frac{1}{n(\lambda_s - \lambda_r)} \sum_{i=1}^{n} (\beta_{si} - \bar{\beta}_s)(\beta_{ri} - \bar{\beta}_r) \Bigg\} \gamma_s(t) \\
&\quad + g_m(t)^\top R_{r,m}^*,
\end{aligned}
\tag{31}
$$

where due to (29) the function $g_m(t)^\top R_{r,m}^*$ satisfies

$$
\begin{aligned}
\mathrm{E}(\|g_m^\top R_{r,m}^*\|) &= \mathrm{E}(\|R_{r,m}^*\|) \\
&\leq \frac{6}{n \min_s |\lambda_s - \lambda_r|^2} \Bigg(\sum_j \sum_k \mathrm{E}\{\beta_{ji}^2 \beta_{ki}^2\} \Bigg) + \mathcal{O}(n^{-1}),
\end{aligned}
$$

for all m. By Assumption 1, the series in (31) converge with probability 1 as $m \to \infty$.

Obviously, the event $\hat{\lambda}_{r-1} > \hat{\lambda}_r > \hat{\lambda}_{r+1}$ occurs with probability 1. Since m is arbitrary, we can therefore conclude from (25) and (31) that

$$\begin{aligned}
&\hat{\gamma}_r(t) - \gamma_r(t) \\
&\quad = -\sum_{s \neq r} \left\{ \frac{1}{n(\lambda_s - \lambda_r)} \sum_{i=1}^{n} (\beta_{si} - \bar{\beta}_s)(\beta_{ri} - \bar{\beta}_r) \right\} \gamma_s(t) \;+\; R_r^*(t) \\
&\quad = -\sum_{s \neq r} \left\{ \frac{1}{n(\lambda_s - \lambda_r)} \sum_{i=1}^{n} \beta_{si}\beta_{ri} \right\} \gamma_s(t) \;+\; R_r(t),
\end{aligned} \tag{32}$$

where $\|R_r^*\| = \mathcal{O}_p(n^{-1})$, as well as $\|R_r\| = \mathcal{O}_p(n^{-1})$. Moreover, $\sqrt{n} \times \sum_{s \neq r} \{ \frac{1}{n(\lambda_s - \lambda_r)} \sum_{i=1}^{n} \beta_{si}\beta_{ri} \} \gamma_s(t)$ is a zero mean random variable with variance $\sum_{q \neq r} \sum_{s \neq r} \frac{\mathrm{E}[\beta_{ri}^2 \beta_{qi} \beta_{si}]}{(\lambda_q - \lambda_r)(\lambda_s - \lambda_r)} \gamma_q(t)\gamma_s(t) < \infty$. By Assumption 1, it follows from standard central limit arguments that for any $q \in \mathbb{N}$ the truncated series $\sqrt{n} W(q) \stackrel{\mathrm{def}}{=} \sqrt{n} \sum_{s=1, s \neq r}^{q} [\frac{1}{n(\lambda_s - \lambda_r)} \sum_{i=1}^{n} \beta_{si}\beta_{ri}] \gamma_s(t)$ is asymptotically normal distributed. The asserted asymptotic normality of the complete series then follows from an argument similar to the one used in the proof of assertion (i). $\square$

Proof of Theorem 3. The results of Theorem 2 imply that

$$\begin{aligned}
n\Delta_1 = \int \Bigg(&\sum_r \frac{1}{\sqrt{q_1 n_1}} \sum_{i=1}^{n_1} \beta_{ri}^{(1)} \gamma_r^{(1)}(t) \\
&- \sum_r \frac{1}{\sqrt{q_2 n_2}} \sum_{i=1}^{n_2} \beta_{ri}^{(2)} \gamma_r^{(2)}(t) \Bigg)^2 dt.
\end{aligned} \tag{33}$$

Furthermore, independence of $X_i^{(1)}$ and $X_i^{(2)}$ together with (30) imply that

$$\begin{aligned}
&\sqrt{n}[\hat{\lambda}_r^{(1)} - \lambda_r^{(1)} - \{\hat{\lambda}_r^{(2)} - \lambda_r^{(2)}\}] \xrightarrow{\mathcal{L}} N\left(0, \frac{\Lambda_r^{(1)}}{q_1} + \frac{\Lambda_r^{(2)}}{q_2}\right) \quad \text{and} \\
&\frac{n}{\Lambda_r^{(1)}/q_1 + \Lambda_r^{(2)}/q_2} \Delta_{3,r} \xrightarrow{\mathcal{L}} \chi_1^2.
\end{aligned} \tag{34}$$

Furthermore, (32) leads to

$$\begin{aligned}
n\Delta_{2,r} = \Bigg\| &\sum_{s \neq r} \left\{ \frac{1}{\sqrt{q_1 n_1}(\lambda_s^{(1)} - \lambda_r^{(1)})} \sum_{i=1}^{n_1} \beta_{si}^{(1)} \beta_{ri}^{(1)} \right\} \gamma_s^{(1)} \\
&- \sum_{s \neq r} \left\{ \frac{1}{\sqrt{q_2 n_2}(\lambda_s^{(2)} - \lambda_r^{(2)})} \sum_{i=1}^{n_2} \beta_{si}^{(2)} \beta_{ri}^{(2)} \right\} \gamma_s^{(2)} \Bigg\|^2 + \mathcal{O}_p(n^{-1/2})
\end{aligned} \tag{35}$$

and

$$
\begin{aligned}
n\Delta_{4,L} = n \iint \Bigg[& \sum_{r=1}^{L} \gamma_r^{(1)}(t)\{\hat{\gamma}_r^{(1)}(u) - \gamma_r^{(1)}(u)\} \\
& + \gamma_r^{(1)}(u)\{\hat{\gamma}_r^{(1)}(t) - \gamma_r^{(1)}(t)\} \\
& - \sum_{r=1}^{L} \gamma_r^{(2)}(t)\{\hat{\gamma}_r^{(2)}(u) - \gamma_r^{(2)}(u)\} \\
& + \gamma_r^{(2)}(u)\{\hat{\gamma}_r^{(2)}(t) - \gamma_r^{(2)}(t)\} \Bigg]^2 dt\, du + \mathcal{O}_p(n^{-1/2})
\end{aligned}
$$

$$
\begin{aligned}
= \iint \Bigg[& \sum_{r=1}^{L} \sum_{s>L} \left\{ \frac{1}{\sqrt{q_1 n_1}(\lambda_s^{(1)} - \lambda_r^{(1)})} \sum_{i=1}^{n_1} \beta_{si}^{(1)} \beta_{ri}^{(1)} \right\} \\
& \times \{\gamma_r^{(1)}(t)\gamma_s^{(1)}(u) + \gamma_r^{(1)}(u)\gamma_s^{(1)}(t)\} \\
& - \sum_{r=1}^{L} \sum_{s>L} \left\{ \frac{1}{\sqrt{q_2 n_2}(\lambda_s^{(2)} - \lambda_r^{(2)})} \sum_{i=1}^{n_2} \beta_{si}^{(2)} \beta_{ri}^{(2)} \right\} \\
& \times \{\gamma_r^{(2)}(t)\gamma_s^{(2)}(u) + \gamma_r^{(2)}(u)\gamma_s^{(2)}(t)\} \Bigg]^2 dt\, du \\
& + \mathcal{O}_p(n^{-1/2}).
\end{aligned} \tag{36}
$$

In order to verify (36), note that $\sum_{r=1}^{L} \sum_{s=1, s\neq r}^{L} \frac{1}{(\lambda_s^{(p)} - \lambda_r^{(p)})} a_r a_s = 0$ for $p = 1, 2$ and all possible sequences $a_1, \ldots, a_L$. It is clear from our assumptions that all sums involved converge with probability 1. Recall that $\mathrm{E}(\beta_{ri}^{(p)} \beta_{si}^{(p)}) = 0$, $p = 1, 2$ for $r \neq s$.

It follows that $\tilde{X}_r^{(p)} := \frac{1}{\sqrt{q_p n_p}} \sum_{s\neq r} \sum_{i=1}^{n_p} \frac{\beta_{si}^{(p)} \beta_{ri}^{(p)}}{\lambda_s^{(p)} - \lambda_r^{(p)}} \gamma_s^{(p)}$, $p = 1, 2$, is a continuous, zero mean random function on $L^2[0,1]$, and, by assumption, $\mathrm{E}(\|\tilde{X}_r^{(p)}\|^2) < \infty$. By Hilbert space central limit theorems [see, e.g., Araujo and Giné (1980)], $\tilde{X}_r^{(p)}$ thus converges in distribution to a Gaussian random function $\xi_r^{(p)}$ as $n \to \infty$. Obviously, $\xi_r^{(1)}$ is independent of $\xi_r^{(2)}$. We can conclude that $n\Delta_{4,L}$ possesses a continuous limit distribution $F_{4,L}$ defined by the distribution of $\iint [\sum_{r=1}^{L} \{\xi_r^{(1)}(t)\gamma_r^{(1)}(u) + \xi_r^{(1)}(u)\gamma_r^{(1)}(t)\} - \sum_{r=1}^{L} \{\xi_r^{(2)}(t)\gamma_r^{(2)}(u) + \xi_r^{(2)}(u) \times \gamma_r^{(2)}(t)\}]^2\, dt\, du$. Similar arguments show the existence of continuous limit distributions F_1 and $F_{2,r}$ of $n\Delta_1$ and $n\Delta_{2,r}$.

For given $q \in \mathbb{N}$, define vectors $b_{i1}^{(p)} = (\beta_{1i}^{(p)}, \ldots, \beta_{qi}^{(p)},)^\top \in \mathbb{R}^q$, $b_{i2}^{(p)} = (\beta_{1i}^{(p)} \beta_{ri}^{(p)}, \ldots, \beta_{r-1,i}^{(p)} \beta_{ri}^{(p)}, \beta_{r+1,i}^{(p)} \beta_{ri}^{(p)}, \ldots, \beta_{qi}^{(p)} \beta_{ri}^{(p)})^\top \in \mathbb{R}^{q-1}$ and $b_{i3} = (\beta_{1i}^{(p)} \beta_{2i}^{(p)},$

$\ldots, \beta_{qi}^{(p)} \beta_{Li}^{(p)})^\top \in \mathbb{R}^{(q-1)L}$. When the infinite sums over r in (33), respectively $s \neq r$ in (35) and (36), are restricted to $q \in \mathbb{N}$ components (i.e., $\sum_r$ and $\sum_{s>L}$ are replaced by $\sum_{r \leq q}$ and $\sum_{L<s \leq q}$), then the above relations can generally be presented as limits $n\Delta = \lim_{q\to\infty} n\Delta(q)$ of quadratic forms

$$
\begin{aligned}
n\Delta_1(q) &= \begin{pmatrix} \frac{1}{\sqrt{n_1}} \sum_{i=1}^{n_1} b_{i1}^{(1)} \\ \frac{1}{\sqrt{n_2}} \sum_{i=1}^{n_2} b_{i1}^{(2)} \end{pmatrix}^\top Q_1^q \begin{pmatrix} \frac{1}{\sqrt{n_1}} \sum_{i=1}^{n_1} b_{i1}^{(1)} \\ \frac{1}{\sqrt{n_2}} \sum_{i=1}^{n_2} b_{i1}^{(2)} \end{pmatrix}, \\
n\Delta_{2,r}(q) &= \begin{pmatrix} \frac{1}{\sqrt{n_1}} \sum_{i=1}^{n_1} b_{i2}^{(1)} \\ \frac{1}{\sqrt{n_2}} \sum_{i=1}^{n_2} b_{i2}^{(2)} \end{pmatrix}^\top Q_2^q \begin{pmatrix} \frac{1}{\sqrt{n_1}} \sum_{i=1}^{n_1} b_{i2}^{(1)} \\ \frac{1}{\sqrt{n_2}} \sum_{i=1}^{n_2} b_{i2}^{(2)} \end{pmatrix}, \\
n\Delta_{4,L}(q) &= \begin{pmatrix} \frac{1}{\sqrt{n_1}} \sum_{i=1}^{n_1} b_{i3}^{(1)} \\ \frac{1}{\sqrt{n_2}} \sum_{i=1}^{n_2} b_{i3}^{(2)} \end{pmatrix}^\top Q_3^q \begin{pmatrix} \frac{1}{\sqrt{n_1}} \sum_{i=1}^{n_1} b_{i3}^{(1)} \\ \frac{1}{\sqrt{n_2}} \sum_{i=1}^{n_2} b_{i3}^{(2)} \end{pmatrix},
\end{aligned} \tag{37}
$$

where the elements of the $2q \times 2q$, $2(q-1) \times 2(q-1)$ and $2L(q-1) \times 2L(q-1)$ matrices Q_1^q, Q_2^q and Q_3^q can be computed from the respective (q-element) version of (33)–(36). Assumption 1 implies that all series converge with probability 1 as $q \to \infty$, and by (33)–(36), it is easily seen that for all $\epsilon, \delta > 0$ there exist some $q(\epsilon, \delta), n(\epsilon, \delta) \in \mathbb{N}$ such that

$$
\begin{gathered}
P(|n\Delta_1 - n\Delta_1(q)| > \epsilon) < \delta, \qquad P(|n\Delta_{2,r} - n\Delta_{2,r}(q)| > \epsilon) < \delta, \\
P(|n\Delta_{4,L} - n\Delta_{4,L}(q)| > \epsilon) < \delta
\end{gathered} \tag{38}
$$

hold for all $q \geq q(\epsilon, \delta)$ and all $n \geq n(\epsilon, \delta)$. For any given q, we have $\mathrm{E}(b_{i1}) = \mathrm{E}(b_{i2}) = \mathrm{E}(b_{i3}) = 0$, and it follows from Assumption 1 that the respective covariance structures can be represented by finite covariance matrices $\Omega_{1,q}$, $\Omega_{2,q}$ and $\Omega_{3,q}$. It therefore follows from our assumptions together with standard multivariate central limit theorems that the vectors $\{\frac{1}{\sqrt{n_1}} \sum_{i=1}^{n_1} (b_{ik}^{(1)})^\top$, $\frac{1}{\sqrt{n_2}} \sum_{i=1}^{n_2} (b_{ik}^{(2)})^\top\}^\top$, $k = 1, 2, 3$, are asymptotically normal with zero means and covariance matrices $\Omega_{1,q}$, $\Omega_{2,q}$ and $\Omega_{3,q}$. One can thus conclude that, as $n \to \infty$,

$$
n\Delta_1(q) \xrightarrow{\mathcal{L}} F_{1,q}, \qquad n\Delta_{2,r}(q) \xrightarrow{\mathcal{L}} F_{2,r,q}, \qquad n\Delta_{4,L}(q) \xrightarrow{\mathcal{L}} F_{4,L,q}, \tag{39}
$$

where $F_{1,q}, F_{2,r,q}, F_{4,L,q}$ denote the continuous distributions of the quadratic forms $z_1^\top Q_1^q z_1$, $z_2^\top Q_2^q z_2$, $z_3^\top Q_3^q z_3$ with $z_1 \sim N(0, \Omega_{1,q})$, $z_2 \sim N(0, \Omega_{2,q})$, $z_3 \sim$

$N(0, \Omega_{3,q})$. Since ϵ, δ are arbitrary, (38) implies

$$\lim_{q\to\infty} F_{1,q} = F_1, \qquad \lim_{q\to\infty} F_{2,r,q} = F_{2,r}, \qquad \lim_{q\to\infty} F_{4,L,q} = F_{4,L}. \tag{40}$$

We now have to consider the asymptotic properties of bootstrapped eigenvalues and eigenfunctions. Let $\bar{X}^{(p)*} = \frac{1}{n_p}\sum_{i=1}^{n_p} X_i^{(p)*}$, $\beta_{ri}^{(p)*} = \int \gamma_r^{(p)}(t)\{X_i^{(p)*}(t) - \mu(t)\}$, $\bar{\beta}_r^{(p)*} = \frac{1}{n_p}\sum_{i=1}^{n_p}\beta_{ri}^{(p)*}$, and note that $\int \gamma_r^{(p)}(t)\{X_i^{(p)*}(t) - \bar{X}^{(p)*}(t)\} = \beta_{ri}^{(p)*} - \bar{\beta}_r^{(p)*}$. When considering unconditional expectations, our assumptions imply that for $p = 1, 2$

$$\mathrm{E}[\beta_{ri}^{(p)*}] = 0, \qquad \mathrm{E}[(\beta_{ri}^{(p)*})^2] = \lambda_r^{(p)},$$

$$\mathrm{E}[(\bar{\beta}_r^{(p)*})^2] = \frac{\lambda_r^{(p)}}{n_p}, \qquad \mathrm{E}\{[(\beta_{ri}^{(p)*})^2 - \lambda_r^{(p)}]^2\} = \Lambda_r^{(p)},$$

$$\mathrm{E}\left\{\sum_{l,k=1}^{\infty}\left[\frac{1}{n_p}\sum_{i=1}^{n_p}(\beta_{li}^{(p)*} - \bar{\beta}_l^{(p)*})(\beta_{ki}^{(p)*} - \bar{\beta}_k^{(p)*}) - \delta_{lk}\lambda_l^{(p)}\right]^2\right\} \tag{41}$$

$$= \frac{1}{n_p}\left(\sum_l \Lambda_l^{(p)} + \sum_{l\neq k}\lambda_l^{(p)}\lambda_k^{(p)}\right) + o(n_p^{-1}).$$

One can infer from (41) that the arguments used to prove Theorem 1 can be generalized to approximate the difference between the bootstrap eigenvalues and eigenfunctions $\hat{\lambda}_r^{(p)*}$, $\hat{\gamma}_r^{(p)*}$ and the true eigenvalues $\lambda_r^{(p)}$, $\gamma_r^{(p)}$. All infinite sums involved converge with probability 1. Relation (30) then generalizes to

$$\begin{aligned}
&\sqrt{n_p}(\hat{\lambda}_r^{(p)*} - \hat{\lambda}_r^{(p)}) \\
&\quad = \sqrt{n_p}(\hat{\lambda}_r^{(p)*} - \lambda_r^{(p)}) - \sqrt{n_p}(\hat{\lambda}_r^{(p)} - \lambda_r^{(p)}) \\
&\quad = \frac{1}{\sqrt{n_p}}\sum_{i=1}^{n_p}(\beta_{ri}^{(p)*} - \bar{\beta}_r^{(p)*})^2 \\
&\qquad - \frac{1}{\sqrt{n_p}}\sum_{i=1}^{n_p}(\beta_{ri}^{(p)} - \bar{\beta}_r^{(p)})^2 + \mathcal{O}_p(n_p^{-1/2}) \\
&\quad = \frac{1}{\sqrt{n_p}}\sum_{i=1}^{n_p}\left\{(\beta_{ri}^{(p)*})^2 - \frac{1}{n_p}\sum_{k=1}^{n_p}(\beta_{rk}^{(p)})^2\right\} + \mathcal{O}_p(n_p^{-1/2}).
\end{aligned} \tag{42}$$

Similarly, (32) becomes

$$\begin{aligned}
&\hat{\gamma}_r^{(p)*} - \hat{\gamma}_r^{(p)} \\
&\quad = \hat{\gamma}_r^{(p)*} - \gamma_r^{(p)} - (\hat{\gamma}_r^{(p)} - \gamma_r^{(p)})
\end{aligned} \tag{43}$$

$$
\begin{aligned}
&= -\sum_{s\neq r}\Bigg\{\frac{1}{\lambda_s^{(p)}-\lambda_r^{(p)}}\frac{1}{n_p}\sum_{i=1}^{n_p}(\beta_{si}^{(p)*}-\bar\beta_s^{(p)*})(\beta_{ri}^{(p)*}-\bar\beta_r^{(p)*})\\
&\qquad\qquad -\frac{1}{\lambda_s^{(p)}-\lambda_r^{(p)}}\frac{1}{n_p}\sum_{i=1}^{n_p}(\beta_{si}^{(p)}-\bar\beta_s^{(p)})(\beta_{ri}^{(p)}-\bar\beta_r^{(p)})\Bigg\}\gamma_s^{(p)}(t)\\
&\quad + R_r^{(p)*}(t)\\
&= -\sum_{s\neq r}\Bigg\{\frac{1}{\lambda_s^{(p)}-\lambda_r^{(p)}}\frac{1}{n_p}\sum_{i=1}^{n_p}\Bigg(\beta_{si}^{(p)*}\beta_{ri}^{(p)*}-\frac{1}{n_p}\sum_{k=1}^{n_p}\beta_{sk}^{(p)}\beta_{rk}^{(p)}\Bigg)\Bigg\}\gamma_s^{(p)}(t)\\
&\quad + \tilde R_r^{(p)*}(t),
\end{aligned}
$$

where due to (28), (29) and (41), the remainder term satisfies $\|R_r^{(p)*}\| = \mathcal{O}_p(n_p^{-1})$.

We are now ready to analyze the bootstrap versions Δ^* of the different Δ. First consider $\Delta_{3,r}^*$ and note that $\{(\beta_{ri}^{(p)*})^2\}$ are i.i.d. bootstrap resamples from $\{(\beta_{ri}^{(p)})^2\}$. It therefore follows from basic bootstrap results that the conditional distribution of $\frac{1}{\sqrt{n_p}}\sum_{i=1}^{n_p}[(\beta_{ri}^{(p)*})^2-\frac{1}{n_p}\sum_{k=1}^{n_p}(\beta_{rk}^{(p)})^2]$ given $\mathcal{X}_p$ converges to the same $N(0,\Lambda_r^{(p)})$ limit distribution as $\frac{1}{\sqrt{n_p}}\sum_{i=1}^{n_p}[(\beta_{ri}^{(p)})^2-\mathrm{E}\{(\beta_{ri}^{(p)})^2\}]$. Together with the independence of $(\beta_{ri}^{(1)*})^2$ and $(\beta_{ri}^{(2)*})^2$, the assertion of the theorem is an immediate consequence.

Let us turn to Δ_1^*, $\Delta_{2,r}^*$ and $\Delta_{4,L}^*$. Using (41)–(43), it is then easily seen that $n\Delta_1^*$, $n\Delta_{2,r}^*$ and $n\Delta_{4,L}^*$ admit expansions similar to (33), (35) and (36), when replacing there $\frac{1}{\sqrt{n_p}}\sum_{i=1}^{n_p}\beta_{ri}^{(p)}$ by $\frac{1}{\sqrt{n_p}}\sum_{i=1}^{n_p}(\beta_{ri}^{(p)*}-\frac{1}{n_p}\sum_{k=1}^{n_p}\beta_{rk}^{(p)})$, as well as $\frac{1}{\sqrt{n_p}}\sum_{i=1}^{n_p}\beta_{si}^{(p)}\beta_{ri}^{(p)}$ by $\frac{1}{\sqrt{n_p}}\sum_{i=1}^{n_p}(\beta_{si}^{(p)*}\beta_{ri}^{(p)*}-\frac{1}{n_p}\sum_{k=1}^{n_p}\beta_{sk}^{(p)}\beta_{rk}^{(p)})$.

Replacing $\beta_{ri}^{(p)}$, $\beta_{si}^{(p)}$ by $\beta_{ri}^{(p)*}$, $\beta_{si}^{(p)*}$ leads to bootstrap analogs $b_{ik}^{(p)*}$ of the vectors $b_{ik}^{(p)}$, $k=1,2,3$. For any $q\in\mathbb{N}$, define bootstrap versions $n\Delta_1^*(q)$, $n\Delta_{2,r}^*(q)$ and $n\Delta_{4,L}^*(q)$ of $n\Delta_1(q)$, $n\Delta_{2,r}(q)$ and $n\Delta_{4,L}(q)$ by using $(\frac{1}{\sqrt{n_1}}\sum_{i=1}^{n_1}(b_{ik}^{(1)*}-\frac{1}{n_1}\sum_{k=1}^{n_1}b_{ik}^{(1)})^\top, \frac{1}{\sqrt{n_2}}\sum_{i=1}^{n_2}(b_{ik}^{(2)*}-\frac{1}{n_2}\sum_{k=1}^{n_2}b_{ik}^{(2)})^\top)$ instead of $(\frac{1}{\sqrt{n_1}}\sum_{i=1}^{n_1}(b_{ik}^{(1)})^\top, \frac{1}{\sqrt{n_2}}\sum_{i=1}^{n_2}(b_{ik}^{(2)})^\top)$, $k=1,2,3$, in (37). Applying again (41)–(43), one can conclude that for any $\epsilon>0$ there exists some $q(\epsilon)$ such that, as $n\to\infty$,

$$
\begin{aligned}
P(|n\Delta_1^* - n\Delta_1^*(q)|<\epsilon)&\to 1,\\
P(|n\Delta_{2,r}^* - n\Delta_{2,r}^*(q)|<\epsilon)&\to 1,\\
P(|n\Delta_{4,L}^* - n\Delta_{4,L}^*(q)|<\epsilon)&\to 1
\end{aligned}
\tag{44}
$$

hold for all $q \geq q(\epsilon)$. Of course, (44) generalizes to the conditional probabilities given $\mathcal{X}_1$, $\mathcal{X}_2$.

In order to prove the theorem, it thus only remains to show that for *any* given q and all δ

$$|\mathrm{P}(n\Delta(q) \geq \delta) - \mathrm{P}(n\Delta^*(q) \geq \delta |\ \mathcal{X}_1, \mathcal{X}_2)| = \mathcal{O}_p(1) \tag{45}$$

hold for either $\Delta(q) = \Delta_1(q)$ and $\Delta^*(q) = \Delta_1^*(q)$, $\Delta(q) = \Delta_{2,r}(q)$ and $\Delta^*(q) = \Delta_{2,r}^*(q)$, or $\Delta(q) = \Delta_{4,L}(q)$ and $\Delta^*(q) = \Delta_{4,L}^*(q)$. But note that for $k = 1, 2, 3, \mathrm{E}(b_{ik}) = 0$, $\{b_{ik}^{(j)*}\}$ are i.i.d. bootstrap resamples from $\{b_{ik}^{(p)}\}$, and $\mathrm{E}(b_{ik}^{(p)*}|\mathcal{X}_1, \mathcal{X}_2) = \frac{1}{n_p}\sum_{k=1}^{n_p} b_{ik}^{(p)}$ are the corresponding conditional means. It therefore follows from basic bootstrap results that as $n \to \infty$ the conditional distribution of $(\frac{1}{\sqrt{n_1}}\sum_{i=1}^{n_1}(b_{ik}^{(1)*} - \frac{1}{n_1}\sum_{k=1}^{n_1} b_{ik}^{(1)})^\top, \frac{1}{\sqrt{n_2}}\sum_{i=1}^{n_2}(b_{ik}^{(2)*} - \frac{1}{n_2}\sum_{k=1}^{n_2} b_{ik}^{(2)})^\top)$ given $\mathcal{X}_1$, $\mathcal{X}_2$ converges to the same $N(0, \Omega_{k,q})$ limit distribution as $(\frac{1}{\sqrt{n_1}}\sum_{i=1}^{n_1}(b_{ik}^{(1)})^\top, \frac{1}{\sqrt{n_2}}\sum_{i=1}^{n_2}, (b_{ik}^{(2)})^\top)$. This obviously holds for all $q \in \mathbb{N}$, and (45) is an immediate consequence. The theorem then follows from (38), (39), (40), (44) and (45). □

REFERENCES

Araujo, A. and Giné, E. (1980). *The Central Limit Theorem for Real and Banach Valued Random Variables*. Wiley, New York. MR0576407

Besse, P. and Ramsay, J. (1986). Principal components of sampled functions. *Psychometrika* **51** 285–311. MR0848110

Black, F. and Scholes, M. (1973). The pricing of options and corporate liabilities. *J. 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. *J. Multivariate Anal.* **12** 136–154. MR0650934

Fengler, M. (2005a). Arbitrage-free smoothing of the implied volatility surface. SFB 649 Discussion Paper No. 2005–019, SFB 649, Humboldt-Universität zu Berlin.

Fengler, M. (2005b). *Semiparametric Modeling of Implied Volatility*. Springer, Berlin. MR2183565

Fengler, M., Härdle, W. and Villa, P. (2003). The dynamics of implied volatilities: A common principle components approach. *Rev. Derivative Research* **6** 179–202.

Fengler, M., Härdle, W. and Mammen, E. (2007). A dynamic semiparametric factor model for implied volatility string dynamics. *Financial Econometrics* **5** 189–218.

Ferraty, F. and Vieu, P. (2006). *Nonparametric Functional Data Analysis*. Springer, New York. MR2229687

Flury, B. (1988). *Common Principal Components and Related Models*. Wiley, New York. MR0986245

Gihman, I. I. and Skorohod, A. V. (1973). *The Theory of Stochastic Processes. II*. Springer, New York. MR0375463

Hall, P. and Hosseini-Nasab, M. (2006). On properties of functional principal components analysis. *J. Roy. Statist. Soc. Ser. B* **68** 109–126. MR2212577

Hall, P., Müller, H. G. and Wang, J. L. (2006). Properties of principal components methods for functional and longitudinal data analysis. *Ann. Statist.* **34** 1493–1517. MR2278365

HALL, P., KAY, J. W. and TITTERINGTON, D. M. (1990). Asymptotically optimal difference-based estimation of variance in nonparametric regression. *Biometrika* **77** 520–528. MR1087842

HAFNER, R. (2004). *Stochastic Implied Volatility*. Springer, Berlin. MR2090447

HÄRDLE, W. and SIMAR, L. (2003). *Applied Multivariate Statistical Analysis*. Springer, Berlin. MR2061627

KAHALÉ, N. (2004). An arbitrage-free interpolation of volatilities. *Risk* **17** 102–106.

KNEIP, A. and UTIKAL, K. (2001). Inference for density families using functional principal components analysis. *J. Amer. Statist. Assoc.* **96** 519–531. MR1946423

LACANTORE, N., MARRON, J. S., SIMPSON, D. G., TRIPOLI, N., ZHANG, J. T. and COHEN, K. L. (1999). Robust principal component analysis for functional data. *Test* **8** 1–73. MR1707596

PEZZULLI, S. D. and SILVERMAN, B. (1993). Some properties of smoothed principal components analysis for functional data. *Comput. Statist.* **8** 1–16. MR1220336

RAMSAY, J. O. and DALZELL, C. J. (1991). Some tools for functional data analysis (with discussion). *J. Roy. Statist. Soc. Ser. B* **53** 539–572. MR1125714

RAMSAY, J. and SILVERMAN, B. (2002). *Applied Functional Data Analysis*. Springer, New York. MR1910407

RAMSAY, J. and SILVERMAN, B. (2005). *Functional Data Analysis*. Springer, New York. MR2168993

RAO, C. (1958). Some statistical methods for comparison of growth curves. *Biometrics* **14** 1–17.

RICE, J. and SILVERMAN, B. (1991). Estimating the mean and covariance structure nonparametrically when the data are curves. *J. Roy. Statist. Soc. Ser. B* **53** 233–243. MR1094283

SILVERMAN, B. (1996). Smoothed functional principal components analysis by choice of norm. *Ann. Statist.* **24** 1–24. MR1389877

TYLER, D. E. (1981). Asymptotic inference for eigenvectors. *Ann. Statist.* **9** 725–736. MR0619278

YAO, F., MÜLLER, H. G. and WANG, J. L. (2005). Functional data analysis for sparse longitudinal data. *J. Amer. Statist. Assoc.* **100** 577–590. MR2160561

M. BENKO
W. HÄRDLE
CASE—CENTER FOR APPLIED STATISTICS AND ECONOMICS
HUMBOLDT-UNIVERSITÄT ZU BERLIN
SPANDAUERSTR 1
D-10178 BERLIN
GERMANY
E-MAIL: benko@wiwi.hu-berlin.de
haerdle@wiwi.hu-berlin.de
URL: http://www.case.hu-berlin.de/

A. KNEIP
STATISTISCHE ABTEILUNG
DEPARTMENT OF ECONOMICS
UNIVERSITÄT BONN
ADENAUERALLEE 24-26
D-53113 BONN
GERMANY
E-MAIL: akneip@uni-bonn.de

GHICA - Risk Analysis with GH Distributions and Independent Components

Ying Chen[1,2], **Wolfgang Härdle**[1] **and Vladimir Spokoiny**[1,2]

[1] CASE - Center for Applied Statistics and Economics
Humboldt-Universität zu Berlin
Wirtschaftswissenschaftliche Fakultät
Spandauerstrasse 1, 10178 Berlin, Germany
[2] Weierstraß - Institute für Angewandte Analysis und Stochastik
Mohrenstrasse 39, 10117 Berlin, Germany

Abstract

Over recent years, study on risk management has been prompted by the Basel committee for regular banking supervisory. There are however limitations of some widely-used risk management methods that either calculate risk measures under the Gaussian distributional assumption or involve numerical difficulty. The primary aim of this paper is to present a realistic and fast method, **GHICA**, which overcomes the limitations in multivariate risk analysis. The idea is to first retrieve independent components (ICs) out of the observed high-dimensional time series and then individually and adaptively fit the resulting ICs in the generalized hyperbolic (GH) distributional framework. For the volatility estimation of each IC, the local exponential smoothing technique is used to achieve the best possible accuracy of estimation. Finally, the fast Fourier transformation technique is used to approximate the density of the portfolio returns.

The proposed GHICA method is applicable to covariance estimation as well. It is compared with the dynamic conditional correlation (DCC) method based on the simulated data with $d = 50$ GH distributed components. We further implement the GHICA method to calculate risk measures given 20-dimensional German DAX portfolios and a dynamic exchange rate portfolio. Several alternative methods are considered as well to compare the accuracy of calculation with the GHICA one.

Keywords: multivariate risk management, independent component analysis, generalized hyperbolic distribution, local exponential estimation, value at risk, expected shortfall

JEL Codes: C14, C16, C32, C61, G20

Acknowledgement: This research was supported by the Deutsche Forschungsgemeinschaft through the SFB 649 "Economic Risk". In the numerical analysis, the Matlab DCC function developed by Kevin Sheppard and the Matlab FastICA function developed by Aapo Hyvärinen are used.

1 Introduction

Over recent years, study on risk management has been prompted by the Basel committee for regular banking supervisory. Given a d-dimensional portfolio, the conditionally heteroscedastic model is widely used to describe the movement of the underlying series:

$$x(t) = \Sigma_x^{1/2}(t)\varepsilon_x(t), \tag{1}$$

where $x(t) \in \mathbb{R}^d$ are risk factors of the portfolio, e.g. (log) returns of the financial instruments. The covariance Σ_x is assumed to be predictable with respect to (w.r.t.) the past information and $\varepsilon_x(t) \in \mathbb{R}^d$ is a sequence of standardized innovations with $\mathsf{E}[\varepsilon_x(t)|\mathcal{F}_{t-1}] = 0$ and $\mathsf{E}[\varepsilon_x^2(t)|\mathcal{F}_{t-1}] = I_d$. There is a sizeable literature on risk management methods. Among others, we refer to Jorion (2001) for a systematic description.

In this paper, we focus on the calculation of two risk measures, value at risk (VaR) and expected shortfall (ES). These two risk measures are inherently related to the joint density of $x(t)$. The VaR is in fact the distributional quantile of loss, i.e. $-x(t)$, at a prescribed level over a target time horizon and the ES measures the size of loss once the loss exceeds the VaR value. Indicated by formula (1), the joint density estimation depends on the covariance estimation and the distributional assumption of the innovations.

The largest challenge of risk management is due to the high-dimensionality of real portfolios. Above all, the covariance estimation is really computationally demanding as high dimensional series, e.g. a dimension $d > 10$, is considered, see Härdle, Herwartz and Spokoiny (2003). For example, the dynamic conditional correlation (DCC) model proposed by Engle (2002), Engle and Sheppard (2001), which is one multivariate GARCH model, is recommended due to the good performance of its univariate version. In the estimation, the covariance matrix is approximated by the product of a diagonal matrix and a correlation matrix, which reduces the number of unknown parameters much relative to the BEKK specification proposed by Engle and Kroner (1995). In spite of the appealing dimensional reduction, the mentioned estimation method is time consuming and numerically difficult to handle given high-dimensional data.

Moreover, many widely-used risk management methods rely on the unrealistic Gaussian distributional assumption, e.g. the RiskMetrics product introduced by JP Morgan in 1994. In the Gaussian framework with an estimate $\hat{\Sigma}_x(t)$ of $\Sigma_x(t)$, the standardized returns $\hat{\varepsilon}_x(t) = \hat{\Sigma}_x^{(-1/2)}(t)x(t)$ are asymptotically independent and the joint distributional behavior can be easily measured by the marginal distributions. However the Gaussian distributional assumption is merely used for computational and numerical purposes and not for statistical reasons. The conditional Gaussian marginal distributions and the resulting joint Gaussian distribution are at odds with empirical facts, i.e. financial series are heavy tailed distributed.

The heavy tails are typically reduced but not eliminated as the series are standardized by the estimated volatility, see Anderson, Bollerslev, Diebold and Labys (2001).

We illustrate this effect based on two real data sets, the Allianz stock and a DAX portfolio from 1988/01/04 to 1996/12/30. The DAX is the leading index of Frankfurt stock exchange and a 20-dimensional hypothetic portfolio with a static trading strategy $b(t) = (1/20, \cdots, 1/20)^\top$ is considered. The portfolio returns $r(t) = b(t)^\top x(t)$ are analyzed in the univariate version of (1). This simplified calculation is used in practice, but it often suffers from low accuracy of calculation. Suppose now that the two return processes have been properly standardized, by using a local volatility estimation technique discussed later. The standardized returns are empirically heavy-tailed distributed, indicated by the sample kurtoses 12.07 for the Allianz and 22.38 for the portfolio respectively.

Figure 1 displays the estimated logarithmic density curves under several distributional assumptions. Among them, the estimate using the nonparametric kernel estimation is considered as benchmark. The comparison w.r.t. the Allianz stock shows that the GH estimate is most close to the benchmark among others. The Gaussian estimate presents lighter tails. To alleviate the limitation, the Student-$t(6)$ distribution with degrees of freedom of 6 has been recommended in practice. However this distribution is found to over-fit the heavy tails, namely the $t(6)$ estimate displays heavier tails relative to the benchmark. The similar result is observed w.r.t. the DAX portfolio. It is rational to surmise that the risk management methods under the Gaussian and $t(6)$ distributional assumptions generate low accurate results.

To overcome these limitations, Chen, Härdle and Spokoiny (2006) present a simple VaR calculation approach that achieves much better accuracy than the alternative RiskMetrics method. In their study, univariate approaches that involve more realistic but complex procedures can be easily extended for multivariate risk measurement. To be more specific, financial risk factors are first converted to independent components (ICs) using a linear filtering and the univariate method is applied to identify the distributional behavior of each IC. We name here two univariate approaches which measure the risk exposure in the realistic distributional framework. One is the univariate VaR calculation proposed by Chen, Härdle and Jeong (2005), which implements local constant model to estimate volatility and fit the standardized returns under the GH distributional assumption. The other is proposed by Chen and Spokoiny (2006), who apply the local exponential smoothing method to estimate volatility and calculate the risk measure in the GH distributional framework. The standardization of the Allianz and DAX returns in Figure 1 is in fact based on the local exponential smoothing technique.

The primary aim of this paper is to present an realistic and fast multivariate risk management method, **GHICA**, by implementing the IC analysis (ICA) to the high dimensional series and adaptively fitting the ICs in the GH distributional framework. The GHICA

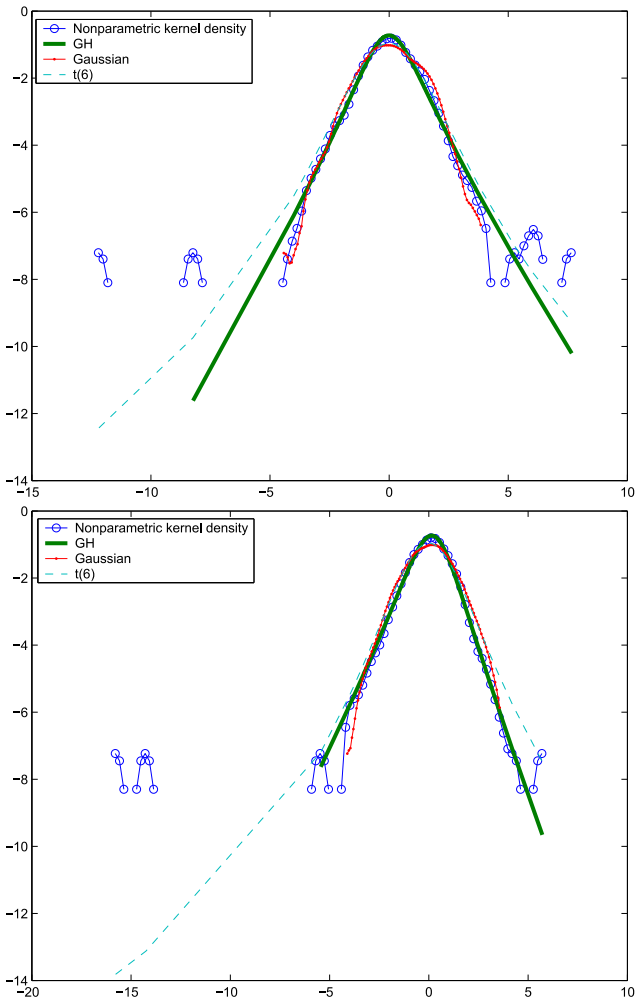


Fig. 1: Density comparisons of the standardized returns in log scale based on the Allianz stock (top) and the DAX portfolio (bottom) with static weights $b(t) = \text{unit}(1/20)$. Time interval: 1988/01/04 - 1996/12/30. The nonparametric kernel density is considered as benchmark. The GH distributional parameters are respectively $\text{GH}(-0.5, 1.01, 0.05, 1.11, -0.03)$ for the Allianz and $\text{GH}(-0.5, 1.21, -0.21, 1.21, 0.24)$ for the DAX portfolio. Data source: FEDC (http://sfb649.wiwi.hu-berlin.de).

method improves the work of Chen et al. (2006) from two aspects. The volatility estimation is driven by the local exponential smoothing technique to achieve the best possible accuracy of estimation. The fast Fourier transformation (FFT) technique is used to approximate the density of the portfolio returns. Compared to the Monte Carlo simulation technique used in the former study, it significantly speeds up the calculation.

In addition, the proposed GHICA method is easily applicable for covariance estimation. Relative to the widely used DCC setup, the GHICA method is fast and delivers sensitive estimates. We demonstrate the comparison based on simulated data. Furthermore, the

GHICA method is implemented to risk management on the base of DAX stocks and foreign exchange rates. Several hypothetic portfolios are constructed by assigning static and dynamic trading strategies to the data sets. The results are compared with those calculated using alternative methods, i.e. the RiskMetrics method, the method using the exponential smoothing to estimate volatility and assuming the Student-$t(6)$ distribution, and the method using the DCC to estimate covariance in the Gaussian distributional framework. All the results are analyzed from the viewpoints of regulatory, investors and internal supervisory. The GHICA method, in general, produces better results than the others.

The paper is organized as follows. The GHICA method is described in Section 2, by which the ICA method, the local exponential smoothing technique and the FFT technique are detailed. Section 3 compares the covariance estimation using the GHICA and DCC methods based on the simulated data with $d = 50$ GH components. The real data analysis in Section 4 demonstrates the implementation of the GHICA method in risk management based on the 20-dimensional German DAX portfolios and a dynamic exchange rate portfolio. Several alternative methods are considered as well to compare the accuracy of calculation with the GHICA one.

2 GHICA Methodology

Given multidimensional time series, for example prices of financial assets, $s(t) \in \mathbb{R}^d$, the (log) returns are calculated as $x(t) = \log\{s(t)/s(t-1)\}$. Without loss of generality, the drift of the returns is set to be 0. Given the time homogeneous model, $x(t) = \Sigma_x^{1/2}\varepsilon_x(t)$ with standardized innovations $\varepsilon_x(t)$, the maximum Gaussian likelihood estimate of the time independent covariance Σ_x is the sample covariance based on the whole past information. Since the covariance is in fact time dependent, one considers the conditional heteroscedastic model:

$$x(t) = \Sigma_x^{1/2}(t)\varepsilon_x(t).$$

Many techniques have been used to approximate the local covariance by specifying a "local homogeneous" interval (e.g. one year or 250 trading days). Inside the homogeneous interval, the unknown covariance should be time-invariant and can be identified using the ML estimation. Among many others, the multivariate GARCH setup such as the DCC is successful in characterizing the clustering feature of covariance under the Gaussian distributional assumption. As the dimension d increases, it however needs to estimate many parameters and becomes numerically difficult. Moreover, the standardized returns $\hat{\varepsilon}_x(t) = \hat{\Sigma}_x^{-1/2}(t)x(t)$ are empirically not Gaussian distributed. Under a realistic distributional assumption, on the other hand, by which the distributional behaviors such as asymmetry and heavy tails are well matched, it is hard to identify the unknown distributional parameters due to complex density form.

The GHICA method proposes a solution to balance the numerical tractability and the realistic distributional assumption on the risk factors. It first converts the return series using a linear transformation and filters out ICs: $y(t) = Wx(t)$. The transformation matrix W is assumed to be time constant and nonsingular and $y(t)$ is the independent vector. The heteroscedastic model is now reformulated as:

$$x(t) = W^{-1}y(t) = W^{-1}\Sigma_y^{1/2}(t)\varepsilon_y(t) = W^{-1}D_y^{1/2}(t)\varepsilon_y(t).$$

Due to the statistical property of independence, the covariance of the ICs $\Sigma_y(t)$ is a diagonal matrix and is denoted as $D_y(t)$ to emphasize this feature. Its diagonal elements are the time varying variances of the ICs. The stochastic innovations $\varepsilon_y(t) = \{\varepsilon_{y_1}(t), \cdots, \varepsilon_{y_d}(t)\}^\top$ are cross independent and can be individually identified in the realistic and univariate distributional framework. By doing so, the GHICA method converts the high dimensional analysis to univariate study and significantly speeds up the calculation.

In this section, the building blocks of the GHICA method are detailed: The FastICA procedure is used to estimate the transformation matrix W; The resulting ICs are individually analyzed, by which the univariate volatility process is estimated using the local exponential smoothing technique and the innovations are assumed to be GH distributed; The quantile of the portfolio return is approximated using the FFT technique.

The GHICA algorithm is summarized as follows:

1. Do ICA to the given risk factors to get ICs.
2. Implement local exponential smoothing to estimate the variance of each IC
3. Identify the distribution of every IC's innovation in the GH distributional framework
4. Estimate the density of the portfolio return using the FFT technique
5. Calculate risk measures

In addition, the GHICA method can be used to estimate the covariance matrix $\Sigma_x(t)$. Given the matrix estimate $\hat{W}$ in the ICA and the variance estimates of the ICs, the covariance of the observed time series are: $\hat{\Sigma}_x(t) = \hat{W}^{-1}\hat{D}_y(t)\hat{W}^{-1\top}$. An alternative covariance estimation approach, the DCC, is briefly described as well. We will compare the GHICA-based covariance estimation with the DCC estimation in the later simulation study.

2.1 Independent component analysis (ICA) and FastICA approach

The aim of ICA is to retrieve, out of high dimensional time series, stochastically ICs through a linear transformation: $y(t) = Wx(t)$, where the transformation matrix $W = (w_1, \cdots, w_d)^\top$

is nonsingular. It is essential to use high order moments in the ICA. In the Gaussian framework, high order moments are however fixed such as skewness with value of 0 and kurtosis with value of 3. Therefore the ICs are assumed to be nongaussian distributed. Furthermore, the ICA transformation has scale identification problem, i.e. the equation holds true by simultaneously multiplying the same constants to the unknown terms $y(t)$ and W: $\{cy(t)\} = \{cW\}x(t)$. To avoid this problem, it is natural to standardize the dependent series and assume that every IC has unit variance $\mathsf{E}(y_j) = 1$ with $j = 1, \cdots, d$. The Mahalanobis transformation $\tilde{x}(t) = \tilde{\Sigma}_x^{-1/2} x(t)$ helps to standardize the return series and the resulting series are considered:

$$y(t) = \tilde{W}\tilde{x}(t),$$

where $\tilde{\Sigma}_x$ is the sample covariance based on the available data. It is easy to show that after the standardization the transformation matrix $\tilde{W}$ turns to be an orthogonal matrix with unit norm. The corresponding matrix w.r.t. the return series is $W = \tilde{W}\tilde{\Sigma}_x^{-1/2}$. For notational simplification, we eliminate the mark $\tilde{\cdot}$ in the following text in this section.

Various ideas have been proposed to estimate the transformation matrix W. Among others, one intuitive ICA estimation is motivated by the definition of mutual information. The mutual information is a natural measure of independence. It is defined as the difference of the sum of marginal entropy and the mutual entropy:

$$\begin{aligned} I(y) &= \sum_{j=1}^{d} H(y_j) - H(y) \\ \text{where } H(y_j) &= -\int f_{y_j}(u) \log f_{y_j}(u) du \end{aligned} \tag{2}$$

The mutual information is nonnegative and goes to 0 if the vector y is cross independent, see Cover and Thomas (1991). Hence for a candidate transformation W, one can minimize the mutual information to achieve independence. Based on the linear transformation of the ICA, the mutual information in (2) can be reformulated as:

$$I(W, y) = \sum_{j=1}^{d} H(y_j) - H(x) - \log|\det(W)|.$$

Notice that the entropy of the return series $H(x)$ is a fixed value and does not depend on the ICs, and the last term in the equation is 0 due to the orthogonality of the transformation matrix W. The optimization problem is: $\min_W \sum_{j=1}^{d} H(y_j)$ and can be further simplified to d optimization problems according to the inequality:

$$\min_W \sum_{j=1}^{d} H(y_j) \quad \geq \quad \sum_{j=1}^{d} \min_{w_j} H(y_j)$$

This simplification leads to some loss in the W estimation but it extensively speeds up the estimation procedure by merely considering d elements of W every time. Equivalently, one can formulate the optimization problem concerning negentropy $J(y_j) = H(y_0) - H(y_j)$ since the entropy and the negentropy are in one-to-one correspondence, where $y_0 \sim \mathrm{N}(0,1)$ is a standard Gaussian vector and $H(y_0)$ is merely a constant. The negentropy is always nonnegative since the Gaussian random variable has the largest entropy given the same variance, see Hyvärinen (1998).

$$\hat{w}_j = \mathrm{argmin} H(y_j) = \mathrm{argmax} J(w_j, y_j).$$

In the estimation, the approximation of negentropy is used to construct the optimization object function w.r.t. the j-th row of the transformation matrix W:

$$\begin{aligned} \hat{w}_j &= \mathrm{argmin} H(y_j) = \mathrm{argmax} J(y_j) \\ J(y_j) &\approx const.\{\mathsf{E}[G(y)] - \mathsf{E}[G(y_0)]\}^2 \\ &= const.\{\mathsf{E}[G(w_j^\top x)] - \mathsf{E}[G(y_0)]\}^2 \\ G(y_j) &= \log\cosh(y_j) \end{aligned} \tag{3}$$

This optimization problem is solved by using the symmetric FastICA algorithm, see Hyvärinen, Karhunen and Oja (2001):

1. Initialization: Choose initial vectors $\hat{w}_j^{(1)}$ for $W = \{w_1, \cdots, w_d\}^\top$ with $j = 1, \cdots, d$, each has a unit norm.

2. Loop:

   - At step n, Calculate $\hat{w}_j^{(n)} = \mathsf{E}\left[x^\top(t) g\left\{\hat{w}_j^{(n-1)\top} x(t)\right\}\right] - \mathsf{E}\left[g'\left\{\hat{w}_j^{(n-1)\top} x(t)\right\}\right] \hat{w}_j^{(n-1)}$, where g is the first derivative of $G(y)$ in form (3) and g' is the second derivative. The expectation $\mathsf{E}[\cdot]$ is approximated by the sample mean.
   - Do a symmetric orthogonalization of the estimated transformation matrix $\hat{W}^{(n)}$:

   $$\hat{W}^{(n)} = \{\hat{W}^{(n)} \hat{W}^{(n)\top}\}^{-1/2} \hat{W}^{(n)}$$

   - If not converged, i.e. $\det\{\hat{W}^{(n)} - \hat{W}^{(n-1)}\} \neq 0$, go back to 2. Otherwise, the algorithm stops.

3. Final result: the last (converged) estimate is the final estimate $\hat{W}$.

2.2 Local exponential smoothing and dynamically conditional correlation

Suppose that the ICs and the transformation matrix W are given. The covariance matrices of the ICs and the original return series are respectively:

$$\begin{aligned} D_y(t) &= \text{diag}\{\sigma^2_{y_1}(t), \cdots, \sigma^2_{y_d}(t)\} \\ \Sigma_x(t) &= W^{-1} D_y(t) W^{-1\top} \end{aligned} \tag{4}$$

where $\sigma_{y_j}(t)$ is the heteroscedastic volatility of the j-th IC with $j = 1, \cdots, d$. Recall that (4) has a similar decomposition structure as the often-used principal component analysis (PCA), by which the covariance is decomposed as: $\Sigma_x = \Gamma\Lambda\Gamma^\top$ with the eigenvector matrix Γ and the diagonal eigenvalue matrix Λ, see Flury (1998). Among other distinctions, the PCA method orders the resulting PCs whereas the ICs have equal importance. In the estimation of the unknown variance, the local exponential smoothing method is used.

Local exponential smoothing: Given the univariate conditional heteroscedastic model: $y_j(t) = \sigma_{y_j}(t)\varepsilon_{y_j}(t)$ with $\mathsf{E}[\varepsilon_{y_j}(t)|\mathcal{F}_{t-1}] = 0$ and $\mathsf{E}[\varepsilon^2_{y_j}(t)|\mathcal{F}_{t-1}] = 1$, we now focus on the adaptive estimation of the volatility σ_{y_j} for $j = 1, \cdots, d$. For notational simplification, the subscripts y_j in σ_{y_j} and j in y_j are eliminated here.

Suppose that a finite set $\{\eta_k, k = 1, \cdots, K\}$ of values of smoothing parameter is given. Every value η_k leads to a localizing weighting scheme $\{\eta_k^{t-s}\}$ for $s \le t$ to the local Gaussian MLE $\tilde{\sigma}^{(k)}(t)$

$$\tilde{\sigma}^{(k)}(t) = \left[\{ \sum_{m=0}^{\infty} \eta_k^m y^2(t-m-1) \} / \{ \sum_{m=0}^{\infty} \eta_k^m \} \right]^{1/2}$$

In practice, one truncates the smoothing window at M_k such that $\eta_k^{M_k+1} \le c \to 0$:

$$\tilde{\sigma}^{(k)}(t) = \left[\{ \sum_{m=0}^{M_k} \eta_k^m y^2(t-m-1) \} / \{ \sum_{m=0}^{M_k} \eta_k^m \} \right]^{1/2}$$

where the Gaussian log-likelihood function given η_k is:

$$\begin{aligned} L(\eta_k, \tilde{\sigma}^{(k)}(t)) &= -\frac{N_k}{2} \log\left(2\pi\{\sigma^{(k)}(t)\}^2\right) - \frac{1}{2\{\sigma^{(k)}(t)\}^2} \sum_{m=0}^{M_k} \eta_k^m y^2(t-m-1) \\ \text{where } N_k &= \sum_{m=0}^{M_k} \eta_k^m \end{aligned} \tag{5}$$

The fitted log-likelihood ratio $L\left(\eta_k, \tilde{\sigma}^{(k)}(t), \sigma(t)\right)$ reads as:

$$L\left(\eta_k, \tilde{\sigma}^{(k)}(t), \sigma(t)\right) \quad = \quad L\left(\eta_k, \tilde{\sigma}^{(k)}(t)\right) - L(\eta_k, \sigma(t))$$

The idea of local exponential smoothing is to aggregate all the local likelihood estimate to achieve the best possible accuracy of estimation. In this sense, the local MLEs $\tilde{\sigma}^{(k)}(t)$ are referred as "weak" estimates.

In our study, we concern the heavy-tailedness of financial time series and assume the normal inverse Gaussian (NIG) distribution, one subclass of the GH distribution, see Section 2.3 for more details. Since the NIG distributional parameters of the innovations are unknown at this stage, we use the quasi ML estimation instead of estimating the variance based on the NIG density form. The quasi ML estimation is applicable if the exponential moment of the squared innovations $\mathsf{E}[\exp\{\rho\varepsilon^2(t)\}]$ exists. A power transformation guarantees that:

$$\begin{aligned} y_p(t) &= \text{sign}\{y(t)\}|y(t)|^p \\ \theta(t) &= Var\{y_p(t)|\mathcal{F}_{t-1}\} = \mathsf{E}\{y_p^2(t)|\mathcal{F}_{t-1}\} = \mathsf{E}\{|y(t)|^{2p}|\mathcal{F}_{t-1}\} \\ &= \sigma^{2p}(t)\,\mathsf{E}\,|\varepsilon(t)|^{2p} = \sigma^{2p}(t)C_p \end{aligned} \tag{6}$$

where $C_p = \mathsf{E}(|\varepsilon(t)|^{2p}|\mathcal{F}_{t-1})$ is a constant and only relies on $0 \leq p < 1/2$. Notice that the power transformed variable $\theta(t)$ is one-to-one correspondence to the variance $\sigma^2(t)$ and can be estimated on the base of the transformed observations $|y(t)|^{2p}$:

$$\tilde{\theta}^{(k)}(t) \quad = \quad \{\sum_{m=0}^{M_k} \eta_k^m |y(t-m-1)|^{2p}\}/N_k$$

Here the smoothing parameter η_k is designed to run over a wide range from values close to zero to one, so that the variability of the unknown process $\theta(t)$ reduces and at least one of the resulting MLEs is good in the sense of small estimation bias. Polzehl and Spokoiny (2006) show that the inverse of N_k in (5) is positively related to the variation of the MLEs. This result is used to construct the sequence of the smoothing parameter $\{\eta_k\}$:

$$\frac{N_{k+1}}{N_k} \approx \frac{1-\eta_k}{1-\eta_{k+1}} = a > 1, \tag{7}$$

where the coefficient a controls the decreasing speed of the variations.

The procedure is sequential and starts with the estimate $\tilde{\theta}^{(1)}(t)$ that has the largest variability but small bias, i.e. we set $\hat{\theta}^{(1)}(t) = \tilde{\theta}^{(1)}(t)$. At every step $k \geq 2$, the new estimate $\hat{\theta}^{(k)}(t)$ is constructed by aggregating the next "weak" estimate $\tilde{\theta}^{(k)}(t)$ and the previously constructed estimate $\hat{\theta}^{(k-1)}(t)$. Following to Belomestny and Spokoiny (2006), the aggregation is done in terms of the parameter $v = -1/(2\theta)$ so that the variable $y(t)$

belongs to the exponential distributional family with a density form: $p(y,v) = p(y)\exp\{yv - d(v)\}$:

$$\begin{aligned} \hat{v}^{(k)}(t) &= \gamma_k \tilde{v}^{(k)}(t) + (1-\gamma_k)\hat{v}^{(k-1)}(t) \\ \text{or equivalently, } \hat{\theta}^{(k)}(t) &= \left(\frac{\gamma_k}{\tilde{\theta}^{(k)}(t)} + \frac{1-\gamma_k}{\hat{\theta}^{(k-1)}(t)}\right)^{-1} \end{aligned}$$

The mixing weights $\{\gamma_k\}$ are computed on the base of the fitted log-likelihood ratio by checking that the previously accepted estimate $\hat{\theta}^{(k-1)}(t)$ is in agreement with the next "weak" estimate $\tilde{\theta}^{(k)}(t)$, i.e. the difference between these two estimates is bounded by critical values ϑ_k:

$$\gamma_k = K_{ag}\left\{L\left(\eta_k, \tilde{\theta}^{(k)}(t), \hat{\theta}^{(k-1)}(t)\right)/\vartheta_k\right\}$$

The aggregation kernel K_{ag} guarantees that the mixing coefficient γ_k is one if there is no essential difference between $\tilde{\theta}^{(k)}(t)$ and $\hat{\theta}^{(k-1)}(t)$, and zero if the difference is significant. The significance level is measured by the critical value ζ_k. In the intermediate case, the mixing coefficient γ_k is between zero and one. The procedure terminates after step k if $\gamma_k = 0$ and we define in this case $\hat{\theta}^{(m)}(t) = \hat{\theta}^{(k-1)}(t)$ for all $m \geq k$.

The critical values $\{\zeta_k\}$ are calculated by using Monte Carlo simulation. We briefly summarize the procedure here. Since the NIG distributional parameters of the innovations are unknown and the transformed variable is close to Gaussian variable, we start from the Gaussian assumption. To be more specific, we generate $y(t) = \sigma^* \varepsilon(t)$ with $\varepsilon(t) \sim \mathrm{N}(0,1)$ and $\sigma^* \stackrel{\text{def}}{=} 1$. The "weak" estimates are calculated given the sequence of $\{\eta_k\}$. For $k = 2, \ldots, K$ with $\zeta_1, \infty, \cdots, \infty$, the value ζ_1 is selected as the minimal one to fulfill

$$\mathsf{E}_{\theta^*}|L\left(\eta_k, \tilde{\theta}^{(k)}(t), \hat{\theta}^{(k)}_{\zeta_1}(t)\right)|^r \leq \frac{\alpha\tau_r}{K-1}, \tag{8}$$

where $\tau_r = 2r\int_{\zeta\geq 0}\zeta^{r-1}e^{-\zeta}d\zeta = 2r\Gamma(r)$, and $r = 0.5$ and $\alpha = 1$ have been suggested in Chen and Spokoiny (2006). Consequently for $l = k+1, \ldots, K$ with the parameters $\zeta_1, \ldots, \zeta_k, \infty, \ldots, \infty$, we select ζ_k as the minimal value which fulfills

$$\mathsf{E}_{\theta^*}|L\left(\eta_l, \tilde{\theta}^{(l)}(t), \hat{\theta}^{(l)}_{\zeta_1,\ldots,\zeta_k}(t)\right)|^r \leq \frac{k\alpha\tau_r}{K-1}. \tag{9}$$

As said before, the transformed variable is close to Gaussian variable, we use the generated critical values under the Gaussian assumption to estimate the volatility. The constant C_p is calculated based on the estimates $\hat{\theta}(t)$ such that the innovation is standardized, i.e. $Var\{\hat{\varepsilon}(t)\} = Var\left[y(t)\{\hat{C}_p/\hat{\theta}(t)\}^{\frac{1}{2p}}\right] = 1$. One then estimates the NIG distributional parameters of $\hat{\varepsilon}(t) = y(t)/\hat{\sigma}(t)$ where $\hat{\sigma}(t) = \{\hat{\theta}(t)/\hat{C}_p\}^{\frac{1}{2p}}$. To get more accurate results, one

generates NIG innovations with the estimated distributional parameters and recalculates the critical values as in the Gaussian case.

The local exponential smoothing algorithm is described as follows:

1. Initialization: $\hat{\theta}^{(1)}(t) = \tilde{\theta}^{(1)}(t)$.
2. Loop: for $k \geq 2$,
$$\hat{\theta}^{(k)}(t) = (\frac{\gamma_k}{\tilde{\theta}^{(k)}(t)} + \frac{1-\gamma_k}{\hat{\theta}^{(k-1)}(t)})^{-1}$$
where the aggregating parameter γ_k is computed as:
$$\gamma_k = K_{\text{ag}}(L(\eta_k, \tilde{\theta}^{(k)}(t), \hat{\theta}^{(k-1)}(t))/\zeta_{k-1}) \tag{10}$$
If $\gamma_k = 0$ then terminate by letting $\hat{\theta}^{(k)}(t) = \ldots = \hat{\theta}^{(K)}(t) = \hat{\theta}^{(k-1)}(t)$.
3. Aggregation estimate: $\hat{\theta}(t) = \hat{\theta}^{(K)}(t)$.
4. Final estimate: $\hat{\sigma}(t) = \{\hat{\theta}(t)/C_p\}^{\frac{1}{2p}}$, where the constant C_p is computed such that the residuals $\hat{\varepsilon}(t) = y(t)/\hat{\sigma}(t)$ have a unit variance as assumed in the heteroscedastic model.

Consequently, the covariance matrices $D_y(t)$ and $\Sigma_x(t)$ are calculated.

Dynamic conditional correlation (DCC) model: Alternatively, the covariance of the return series can be estimated by the DCC model:

$$\Sigma_x(t) = D_x(t) R_x(t) D_x(t)^\top.$$

This technique first identifies the elements of the diagonal matrix $D_x(t)$ in the GARCH(1,1) setup and adaptively specifies the correlation matrix as:

$$R_x(t) = \tilde{R}_x(1-\theta_1-\theta_2) + \theta_1\{\varepsilon_x(t-1)\varepsilon_x(t-1)^\top\} + \theta_2 R_x(t-1),$$

where $\tilde{R}_x$ is the sample correlation of the risk factors, $\varepsilon_x \in \mathbb{R}^d$ are the standardized returns, i.e. risk factors divided by the univariate GARCH(1,1) volatilities, or equivalently by the squared diagonal elements in $D_x(t)$. The standardized returns are assumed to be Gaussian distributed. The parameters θ_1 and θ_2 are identified by the ML estimation.

2.3 Normal inverse Gaussian (NIG) distribution and fast Fourier transformation (FFT)

The estimated ICs are assumed to be NIG distributed. The NIG is a subclass of the GH distribution with a fixed value of $\lambda = -1/2$, see Eberlein and Prause (2002). With 4

distributional parameters, the NIG distribution is flexible to well match the behavior of real data. Compared to many other subclasses of GH distribution, the NIG distribution has a desirable property, saying that the scaled NIG variable belongs to the NIG distribution as well. The density of NIG random variable has a form of:

$$f_{\text{NIG}}(y;\alpha,\beta,\delta,\mu) = \frac{\alpha\delta}{\pi}\frac{K_1\left\{\alpha\sqrt{\delta^2+(y-\mu)^2}\right\}}{\sqrt{\delta^2+(y-\mu)^2}}\exp\{\delta\sqrt{\alpha^2-\beta^2}+\beta(y-\mu)\},$$

where the distributional parameters fulfill $\mu \in \mathbb{R}$, $\delta > 0$ and $|\beta| \leq \alpha$. The modified Bessel function of the third kind $K_\lambda(\cdot)$ with an index $\lambda = 1$ has a form of:

$$K_\lambda(y) = \frac{1}{2}\int_0^\infty y^{\lambda-1}\exp\{-\frac{y}{2}(y+y^{-1})\}\,dy$$

The characteristic function of the NIG variable is:

$$\varphi_y(z) = \exp\left[\mathbf{i}z\mu + \delta\{\sqrt{\alpha^2-\beta^2}-\sqrt{\alpha^2-(\beta+\mathbf{i}z)^2}\}\right]$$

Proof: The characteristic function of the GH random variable has a form of:

$$\varphi_y(z) = \exp(\mathbf{i}z\mu)\left\{\frac{\alpha^2-\beta^2}{\alpha^2-(\beta+\mathbf{i}z)^2}\right\}^{\lambda/2}\frac{K_\lambda\{\delta\sqrt{\alpha^2-(\beta+\mathbf{i}z)^2}\}}{K_\lambda(\delta\sqrt{\alpha^2-\beta^2})}$$

Using the representation of the modified Bessel function with a fixed index $\lambda = -1/2$ derived in Barndorff-Nielsen and Blæsild (1981):

$$K_\lambda(y) = \sqrt{\frac{2}{\pi}}y^{-1/2}e^{-y},$$

it is straightforwardly to show that the assertion holds. □

One desirable feature of the NIG distribution is its explicit scaling transformation. Multiplying the random variable by c, the resulting variable $y' = cy$ belongs to the NIG distribution as well:

$$f_{\text{NIG}}(y';\alpha',\beta',\delta',\mu') = f_{\text{NIG}}(cy;\alpha/|c|,\beta/c,|c|\delta,c\mu). \tag{11}$$

Proof: It is easy to show the result by using the Jacobian transformation, see Härdle and Simar (2003). Given the density of y and let $\alpha' = \alpha/|c|$, $\beta' = \beta/c$, $\delta' = |c|\delta$ and $\mu' = c\mu$, the density of $y' = cy$ has a form of:

$$\begin{aligned} f(y') &= \frac{1}{|c|}f_y(\frac{y}{c}) = \frac{\alpha'\delta'}{\pi}\frac{K_1\left\{\alpha'\sqrt{\delta'^2+(y'-\mu')^2}\right\}}{\sqrt{\delta'^2+(y'-\mu')^2}}\exp\{\delta'\sqrt{\alpha'^2-\beta'^2}+\beta'(y'-\mu')\} \\ &= f_{\text{NIG}}(y';\alpha',\beta',\delta',\mu'). \end{aligned}$$

□

To calculate risk measures, it requires the identification of the portfolio returns' density. Based on the GHICA model, the portfolio returns are calculated as:

$$r(t) = b(t)^\top W^{-1} D_y(t)^{1/2} \varepsilon_y(t)$$

where $b(t)$ is the trading strategy. Notice that the linear transformation of the NIG variable is not necessarily NIG distributed. In other words, the density of the return is unknown although the marginal densities are clear. On the meanwhile its characteristic function is explicitly writable. This is the same case as approximating the α-stable distribution in Menn and Rachev (2004), by which the Fourier transformation is used to approximate the density of the variable based on its characteristic function. This motivates us to use the technique to approximate the density of the return in the GHICA procedure.

Set $a = (a_1, \cdots, a_d) = b(t)^\top W^{-1} D_y(t)^{1/2}$, the variable $\zeta_j = a_j \varepsilon_j$ is NIG distributed with $j = 1, \cdots, d$, according to (11):

$$\zeta_j \sim \text{NIG}(\zeta_j, \breve{\alpha}_j, \breve{\beta}_j, \breve{\delta}_j, \breve{\mu}_j) = \text{NIG}(\zeta_j, \alpha_j/|a_j|, \beta_j/a_j, |a_j|\delta_j, a_j\mu_j).$$

The characteristic function of the return $r = \sum_{j=1}^d \zeta_j$ at time t is:

$$\varphi_r(z) = \prod_{j=1}^d \varphi_{\zeta_j}(z) = \exp\left[\mathbf{i}z \sum_{j=1}^d \breve{\mu}_j + \sum_{j=1}^d \breve{\delta}_j \{\sqrt{\breve{\alpha}_j^2 - \breve{\beta}_j^2} - \sqrt{\breve{\alpha}_j^2 - (\breve{\beta}_j + \mathbf{i}z)^2}\}\right]$$

The density function is approximated by the Fourier transformation:

$$f(r) = \frac{1}{2\pi} \int_{-\infty}^{+\infty} \exp(-\mathbf{i}tr)\psi(z)dt \approx \frac{1}{2\pi} \int_{-s}^{s} \exp(-\mathbf{i}tr)\psi(z)dt$$

The procedure of quantile estimation is summarized as follows:

- Implement the discrete fast Fourier transformation (DFT) to approximate the density of r at every time point t:
  1. Let $N = 2^m$ with $m \in \mathbb{N}$ and define an equidistance grid over the integral interval $[-s, s]$ by setting $h = \frac{2s}{N}$ and the grid points $z_j = -s + j * h$ with $j = 0, \cdots, N$.
  2. Calculate the input of the DFT: $y_j = (-1)^j \psi(z_j^*)$ with $z_j^* = 0.5(z_j + z_{j+1})$ are the middle points. Notice that the characteristic function is time dependent.
  3. The density $f(r) = \frac{1}{2\pi} C_k \text{DFT}(y)_k$ with $C_k = \frac{2s}{N}(-1)^k \exp(-\frac{\mathbf{i}k\pi}{N})\mathbf{i}$ with $k = 0, \cdots, N-1$. We refer to Borak, Detlefsen and Härdle (2005) and Menn and Rachev (2004) for more details. The corresponding values of $r = -\frac{N\pi}{2a} + \frac{\pi k}{a}$.
- The cumulative density function and the quantile are then approximated based on

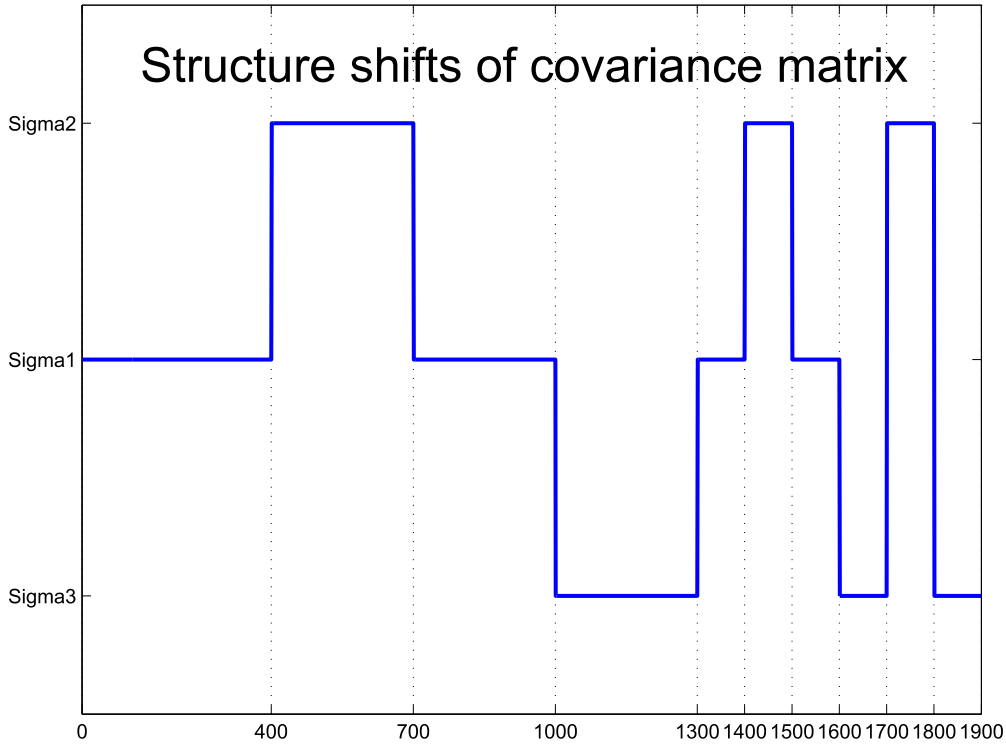


Fig. 2: Structure shifts of the generated covariance through time. Notice that there are shifts among matrices not up-and-down movements.

the resulting density.

3 Covariance estimation with simulated data

In this section, the GHICA versus the DCC, are implemented to estimate covariance of simulated data. The dimension is set to be $d = 50$. The simulation study is designed to include structure shifts of covariance. To be more specific, the designed covariance changes among three matrices over time, one is an identity matrix denoted as Σ_1, meaning uncorrelatedness, and two symmetric and semi-positive defined matrices Σ_2 and Σ_3. (Here we first generate $d * d$ matrix U_1 whose elements are uniform random variables for Σ_2 and standard Gaussian variables for Σ_3, then calculate a new matrix $U_2 = U_1 * U_1'$ to guarantee the semi-positiveness. The elements $\Sigma(i,j)$ of the target matrix are calculated as $\Sigma(i,j) = U_2(i,j)/\sqrt{U_2(i,i)U_2(j,j)}$.) The eigenvalues of these two matrices are distributed in $[5.92e-004, 3.779]$ (Σ_2) and $[0.002, 3.573]$ (Σ_3) respectively. The off-diagonal values span over $[-0.433, 0.468]$ in the first self-correlated matrix (Σ_2) and $[-0.447, 0.464]$ in the second one (Σ_3). Temporal stationarity is assumed to be long for 400 time units and short for 100 units. The structure shifts of the generated covariance are illustrated in Figure 2. The level of the shifts is either small with a shift from one self-correlated matrix (Σ_2 or Σ_3) to the identity matrix or contrariwise, e.g. at the point 700, or large with a shift between the two

self-correlated matrices, e.g. at the point 1800.

Furthermore, two distributional parameters μ and β of the standardized NIG innovations $\varepsilon_x(t)$ are set to be 0, meaning that the innovations are centered around 0 and symmetric distributed, see Barndorff-Nielsen and Blæsild (1981). By doing so, the mean and variance of the NIG innovations only depend on α and δ:

$$\begin{aligned} \mathsf{E}(\varepsilon_x) &= \mu + \frac{\beta\delta}{\sqrt{\alpha^2-\beta^2}} = 0 \\ \mathrm{Var}(\varepsilon_x) &= \frac{\delta}{\sqrt{\alpha^2-\beta^2}} + \frac{\beta^2}{\delta^3\sqrt{\alpha^2-\beta^2}} = \frac{\delta}{\alpha} = 1 \end{aligned}$$

This result is used to generate the standardized innovations, by which $\alpha \sim U[1,2]$ is suggested by our experience on real data analysis and $\delta = \alpha$.

In the Monte Carlo simulation, we generate $d = 50$ NIG variables with the designed covariance and distributional parameters:

$$x(t) = \Sigma_x^{1/2}(t)\varepsilon_x(t).$$

The sample size is $T = 1900$ and the scenarios are repeated $N = 100$ times. The covariance matrix is estimated using the GHICA procedure and the DCC method respectively.

The GHICA method first converts the underlying series to ICs by a linear transformation:

$$x(t) = W^{-1}y(t) = W^{-1}D_y^{1/2}(t)\varepsilon_y(t),$$

by which the elements of $D_y(t)$ on the diagonal are estimated using the local exponential smoothing method. In the local exponential smoothing estimation, we set the involved parameters $c = 0.01$, $a = 1.25$ and $p = 0.25$. The sequence of the smoothing parameters $\{\eta_k\}$ are $0.600, \cdots, 0.982$ with $K = 15$, based on the condition $(1-\eta_k)/(1-\eta_{k+1}) = a$ in (7). The first 300 observations are reserved as training set for the very beginning estimations, since the largest smoothing parameter used in this study corresponds to a window with 259 observations.

The covariance of $x(t)$ is calculated by the basic statistical property:

$$\Sigma_x(t) = W^{-1}D_y(t)W^{-1\top}$$

The DCC method assumes that the underlying series are Gaussian distributed. It decomposes the covariance matrix to a product of diagonal variance matrix and correlation matrix:

$$\Sigma_x(t) = D_x(t)R_x(t)D_x(t)^\top.$$

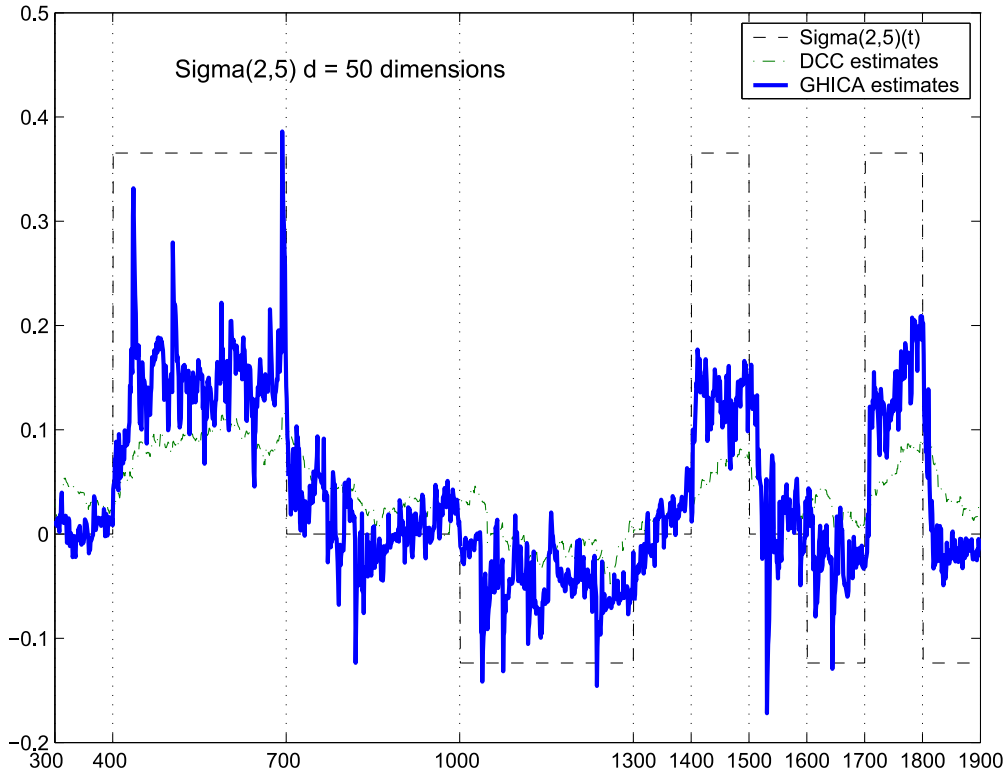


Fig. 3: Realized estimates of $\Sigma(2,5)$ based on the GHICA and DCC methods. The generated data consists of 50 NIG distributed components.

where $D_x(t)$ consists of the variances of $x(t)$ on the diagonal that are estimated in the GARCH(1,1) setup.

Figure 3 displays one realization of $\Sigma(2,5)$, i.e. the covariance of the second and fifth risk factors $x_2(t)$ and $x_5(t)$, based on one simulation data. The true values are 0.365 in Σ_2 and -0.124 in Σ_3. As expected, the GHICA estimates are sensitive to structure shifts through time. The DCC estimates, on the contrary, are over-smooth and slowly follow the shifts. Given more often shifts around the last hundreds of time points, the DCC estimates deliver less information on the movements. Recall that 100 points correspond to 4 months observations of daily returns. It is rational to surmise that structure shifts happen so often in the active financial markets, see Merton (1973). The similar estimation results are observed in the other elements of the covariance, which are eliminated here.

To measure the accuracy of estimation, ratio of absolute estimation error (RAE) of the estimates w.r.t. the true covariance are calculated pointwise.

$$\mathrm{RAE}(i,j) \quad = \quad \frac{\sum_{t=301}^{T} |\hat{\Sigma}^{\mathrm{GHICA}}_{(i,j)}(t) - \Sigma_{(i,j)}(t)|}{\sum_{t=301}^{T} |\hat{\Sigma}^{\mathrm{DCC}}_{(i,j)}(t) - \Sigma_{(i,j)}(t)|}$$

If $\mathrm{RAE}(i,j) \leq 1$, it means that the GHICA method reaches higher accuracy in the estima-

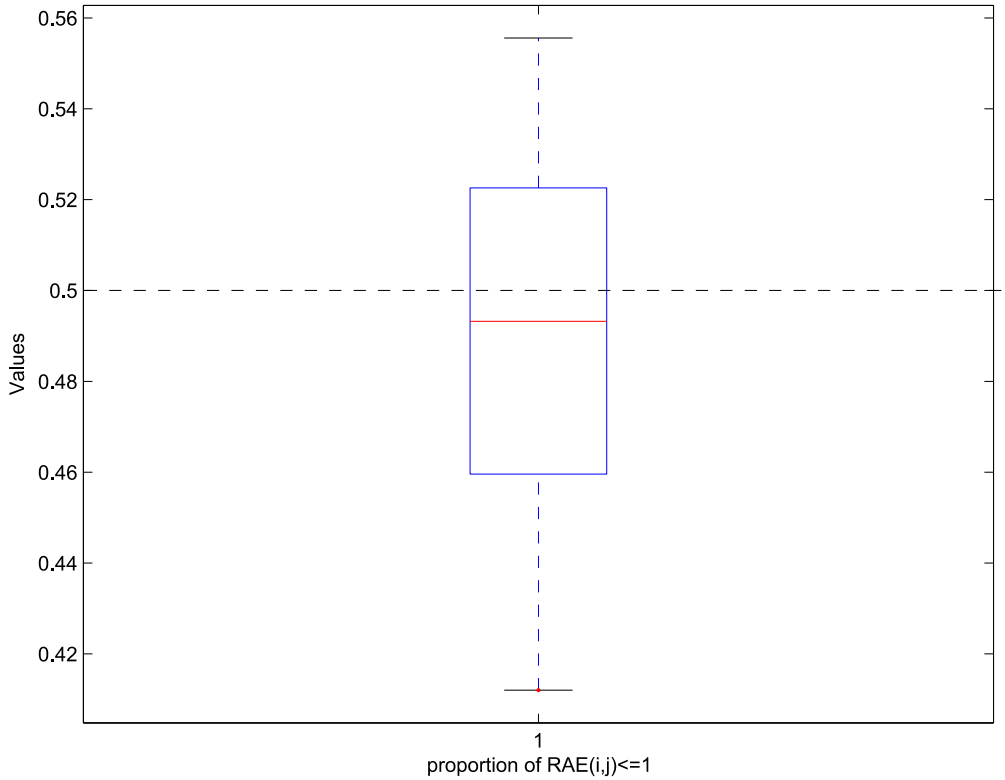


Fig. 4: Boxplot of the proportion $\frac{\sum_i \sum_j \mathbf{1}(\text{RAE}_{(i,j)\leq 1})}{d\times d}$ for $i, j = 1, \cdots, d$. Here $d = 50$ and the proportions on the base of 100 simulations are considered.

tion of $\Sigma(i, j)$ than the DCC. To compare the general performance of these two methods in covariance estimation, we check the proportion of the RAEs among the 2500 $(d*d)$ elements that are smaller or equal to one, i.e. $\frac{\sum_i \sum_j \mathbf{1}(\text{RAE}_{(i,j)\leq 1})}{d\times d}$ for $i, j = 1, \cdots, d$. Notice that the proportion with value of 0.5 indicates that half elements are better estimated by using the GHICA and the other half are better done by the DCC. In other words, the considered methods have a comparable accuracy of estimation. Figure 4 displays the boxplot of the 100 proportions. The mean of the proportion is 0.4904 among the 100 simulations. It states that the DCC method performs a little bit better than the GHICA in the sense of accuracy. On the meanwhile, the GHICA method is much fast and sensitive to structure shifts.

4 Risk management with real data

In this section, we implement the proposed GHICA method to calculate risk measures using real data sets: 20-dimensional German DAX portfolio and 7-dimensional exchange rate portfolio. The results are compared with those based on alternative risk management models. The data sets have been kindly provided by the financial and economic data center (FEDC) of the Collaborative Research Center 649 on Economic Risk of the Humboldt-

Universität zu Berlin (http://sfb649.wiwi.hu-berlin.de). Before giving detailed description of the data sets, we analyze the risk measures from the viewpoints of regulatory, investors and internal supervisory.

Regulatory requirement: Financial institutions generally face market risk that arises from the uncertainty due to changes in market prices and rates such as share prices, foreign exchange rates and interest rates, the correlations among them and their levels of volatility, see Jorion (2001). The market risk is the main risk source and has a great negative influence on the development of economic. The famous example is the stock crashes in the autumn 1929 and 1987 which caused a violent depression in the United States and some other countries, with the collapse of financial markets and the contraction of production and employment. To alleviate the down influence of market risks, regulation on banking and other financial institutions has been strengthened since the mid-1990s. The goals of the regulation are to restrict the happening of extremely large losses and require banks to reserve adequate capital. In 1998 the Basel accord officially allowed financial institutions to use their internal models to measure market risks. Among others, Value at Risk (VaR) has been considered as industry standard risk measure:

$$\mathrm{VaR}_{t,\mathrm{pr}} = -\mathrm{quantile}_{\mathrm{pr}}\{r(t)\}.$$

where pr is the $h = 1$-day or $h = 5$-day forecasted probability of the portfolio returns. Internal models for risk management are verified in accordance with the "traffic light" rule that counts the number of exceptions over VaR at 1% probability spanning the last 250 days and identifies the multiplicative factor M_f in the market risk charge calculation, see Franke, Härdle and Hafner (2004):

$$\text{Risk charge}_t = \max\left(M_f \frac{1}{60}\sum_{i=1}^{60} \mathrm{VaR}_{t-i,1\%}, \mathrm{VaR}_{t,1\%}\right)$$

The multiplicative factor M_f has a floor value 3. It increases corresponding to the number of exceptions, see Table 1. For example, if an internal model generates 7 exceptions at 1% probability over the last 250 days, the model is in the yellow zone and its multiplicative factor is $M_f = 3.65$. Financial institutions whose internal model is located in the yellow or red zone, with a very high probability, are required to reserve more risk capital than their internal-model-based VaRs. Notice that the increase of risk charge will reduce the ratio of profit since the reserved capital can not be invested. On the meanwhile, an internal model is automatically accepted if the number of exceptions does not exceed 4. This regulatory rule in fact suggests banks to control VaR at 1.6% (i.e. 4/250) instead of 1% probability. It is clear that 1.6%-VaR is smaller than 1%-VaR. Therefore an internal model is particularly desirable by financial institutions if its empirical probability is smaller or equal to 1.6%, and simultaneously requires risk charge as small as possible. Here a simplified calculation

No. exceptions	Increase of M_f	Zone
0 bis 4	0	**green**
5	0.4	**yellow**
6	0.5	**yellow**
7	0.65	**yellow**
8	0.75	**yellow**
9	0.85	**yellow**
More than 9	1	**red**

Tab. 1: Traffic light as a factor of the exceeding amount, cited from Franke, Härdle and Hafner (2004).

on the average value of VaRs is used as risk charge for comparison:

$$\text{Risk charge (RC)} = \text{mean}\left(\text{VaR}_{t,\text{pr}}\right)$$

Investor: It is known that VaR is inappropriate for the measurement of capital adequacy, since it controls only the probability of default, i.e. the frequency of losses, but not the size of losses in the case of default. For this reason, investors concern expected shortfall (ES) more than VaR to measure and control their risks.

$$\text{ES} = \mathsf{E}\{-r(t)| - r(t) > \text{VaR}_{t,\text{pr}}\}$$

Investors suffer loss once bankruptcy happens. Even in the "best" situation, their loss equals to the difference between the total loss and the reserved risk capital, i.e. the value of ES. Generally risk-averse investors care the amount of loss and thus prefer an internal model with small value of ES. Risk-seeking investors, on the other hand, care profit and hence the small value of risk charge favors their requirement.

Internal supervisory: It is important for internal supervisory to exactly measure the market risk exposures before risk controlling. For this reason, internal supervisory prefers the model delivering accurate probability prediction, i.e. the empirical probability $\hat{\text{pr}}$ is as close to the expected values as possible:

$$\hat{\text{pr}} = \frac{\text{No. exceptions}}{\text{No. total observations}}$$

Given two models with the same empirical probability, the model has a smaller value of ES is considered better than the other. Here two extreme probabilities are considered, i.e. $\text{pr} = 1\%$ for regulatory reason and $\text{pr} = 0.5\%$ used by financial institutions with AAA rating.

4.1 Data analysis 1: DAX portfolio

The primary target of the real data analysis is to compare the forecasting ability of the GHICA method with two alternatives, the RiskMetrics method under the Gaussian distributional assumption and a modification with the Student-$t(6)$ distributional assumption (abbreviated as $t(6)$ method) in the market. The comparison is demonstrated based on 20 DAX stocks over a long time period, starting on 1974/01/02 and ending on 1996/12/30 (5748 observations). The return series are all centered around 0 and have heavy tails (kurtosis$>$ 3), the smallest correlation coefficient is 0.3654. Hypothetical German DAX portfolios are constructed with two static trading strategies $b(t) = b^{(1)} = (1/d, \cdots, 1/d)^\top$ and $b(t) = b^{(2)} \sim U[0,1]^d$. Such a simple portfolio construction eliminates the influence of strategy adjustments on the calculation. The portfolio returns are analyzed using the RiskMetrics or the $t(6)$ method. Here the unknown volatility process of the portfolio is estimated using the exponential smoothing method with $\eta = 0.94$:

$$\begin{aligned} r(t) &= b^\top x(t) = \sigma_r(t)\varepsilon_r(t) \\ \sigma_r^2(t) &= \{\sum_{m=0}^{M} \eta^m r^2(t-m-1)\}/(\sum_{m=0}^{M} \eta^m) \end{aligned}$$

where the truncated value M fulfills the condition $\eta^{(M+1)} \leq 0.01$. Notice that given a dynamic trading strategy, this simplification needs to repeatedly estimate the density of the time varying hypothetical portfolio returns, and it often suffers from a low accuracy of estimation.

Figure 5 depicts the one day log-returns of the DAX portfolio with the static trading strategy $b(t) = b^{(1)}$. The VaRs from 1975/03/17 to 1996/12/30 at pr $= 0.5\%$ are displayed w.r.t. three methods, the GHICA, the RiskMetrics and the $t(6)$. The most volatile time period over $t \in [3300, 4300]$ is detailed in the bottom diagram. Recall that on the Monday, 19 October 1987, the worldwide downward jump of stocks happened. Dow Jones Industrial Average for example dropped by over 500 points. At this market quiver around $t = 3446$, the GHICA method exactly achieves the locations of extreme losses whereas the RiskMetrics and $t(6)$ methods over-react to them. Such over reactions induce large risk charges unnecessarily. On the other hand, it is observed that these two alternative methods give close forecasts to some extreme losses, e.g. around time points 4000 and 4500. As a result, the associating values of ES are small and satisfy the requirement of risk-averse investors.

Table 2 reports the risk measures based on the three methods. In general, the RiskMetrics is successful in fulfilling the minimal requirement of regulatory. The $t(6)$ method is preferred by investors who consider risk happened with 1% probability. The GHICA method performs better than the other two for internal supervisory and requirement of

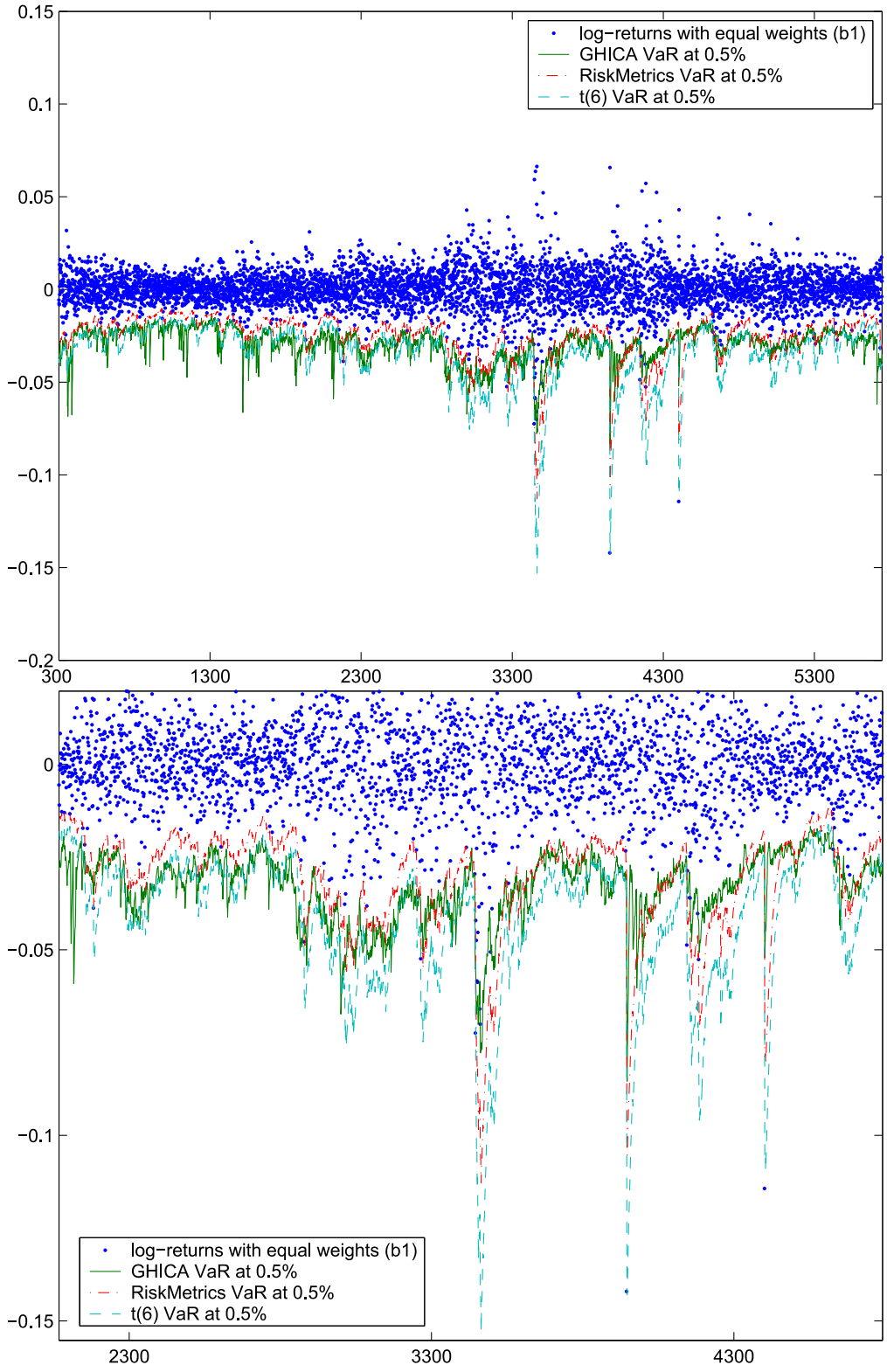


Fig. 5: One day log-returns of the DAX portfolio with the static trading strategy $b(t) = b^{(1)}$. The VaRs are from 1975/03/17 to 1996/12/30 at pr = 0.5% w.r.t. three methods, the GHICA, the RiskMetrics and the $t(6)$. Part of the VaR time plot is enlarged and displayed on the bottom.

			GHICA			RiskMetrics N(μ, σ^2)			Exponential smoothing $t(6)$		
h	$b(t)$	pr	p̂r	RC	ES	p̂r	RC	ES	p̂r	RC	ES
1	$b^{(1)}$	1%	0.55%	0.0264	0.0456	1.18%^{s}	**0.0229**^{r}	0.0279	0.40%	0.0292	**0.0269**^{i}
	$b^{(1)}$	0.5%	0.44%^{s}	0.0297	**0.0472**^{i}	0.75%	0.0254	0.0317	0.23%	0.0345	0.0506
	$b^{(2)}$	1%	0.59%	0.0265	0.0448	1.03%^{s}	**0.0231**^{r}	0.0288	0.38%	0.0294	**0.0406**^{i}
	$b^{(2)}$	0.5%	0.42%^{s}	0.0298	**0.0476**^{i}	0.71%	0.0256	0.0315	0.21%	0.0347	0.0514
5	$b^{(1)}$	1%	0.83%	0.0550	0.0841	1.15%^{s}	**0.0481**^{r}	0.0602	0.19%	0.0665	**0.0833**^{i}
	$b^{(1)}$	0.5%	0.51%^{s}	0.0612	**0.0939**^{i}	0.64%	0.0536	0.0683	0.09%	0.0784	0.1067
	$b^{(2)}$	1%	0.83%^{s}	0.0554	**0.0828**^{i}	1.18%	**0.0488**^{r}	0.0613	0.16%	0.0673	0.0852
	$b^{(2)}$	0.5%	0.50%^{s}	0.0617	**0.0943**^{i}	0.63%	0.0543	0.0676	0.07%	0.0794	0.1218

Tab. 2: Risk analysis of the DAX portfolios with two static trading strategies. The concerned forecasting interval is $h = 1$ or $h = 5$ days. The best results to fulfill the regulatory requirement are marked by ^{r}. The method preferred by investor is marked by ^{i}. For the internal supervisory, the method marked by ^{s} is recommended.

risk-averse investors who care the extreme risk happened with 0.5% probability.

4.2 Data analysis 2: Foreign exchange rate portfolio

In financial markets, traders adjust trading strategy according to information obtained. The GHICA is easily applicable to dynamic portfolios. We consider here 7 actively traded exchange rates, Euro (EUR), the US dollar (USD), the British pounds (GBP), the Japanese yen (JPY) and the Singapore dollar (SGD) from 1997/01/02 to 2006/01/05 (2332 observations). The foreign exchange rate (FX) market is the most active and liquid financial market in the world. It is realistic to analyze a dynamic portfolio with daily time varying trading strategy $b^{(3)}(t)$. The strategy at time point t relies on the realized returns at $t-1$, the proportions of which w.r.t the sum of returns:

$$b^{(3)}(t) = \frac{x(t-1)}{\sum_{j=1}^{d} x_j(t-1)}$$

where $x(t) = \{x_1(t), \cdots, x_d(t)\}^\top$. Among these data sets, the returns of the EUR/SGD and USD/JPY rates are least correlated with the correlation coefficient 0.0071 whereas the returns of the EUR/USD and EUR/SGD rates are most correlated with the coefficient 0.6745. The resulting portfolio returns span over $[-0.7962, 0.7074]$.

The GHICA method is compared with an alternative method, abbreviated as DCCN, that applies the DCC covariance estimation under the Gaussian distributional assumption.

$$r(t) \quad = \quad b(t)^\top x(t) = b(t)^\top \Sigma_x^{(1/2)}(t)\varepsilon_x(t)$$

where $\varepsilon_x \sim \mathrm{N}(\mu, \Sigma_\varepsilon)$ with the diagonal covariance matrix Σ_ε. Notice that the quantile

			GHICA			DCCN		
h	$b(t)$	pr	p̂r	RC	ES	p̂r	RC	ES
1	$b^{(3)}(t)$	1%	1.28%^{s}	**0.0453**^{r}	0.0778	1.59%	0.0494	**0.0254**^{i}
	$b^{(3)}(t)$	0.5%	0.59%^{s}	0.0493	**0.1944**^{i}	0.94%	0.0547	0.0289
5	$b^{(3)}(t)$	1%	1.53%^{s}	**0.0806**^{r}	**0.2630**^{i}	4.17%	0.0993	0.1735
	$b^{(3)}(t)$	0.5%	0.79%^{s}	0.1092	**0.2801**^{i}	3.44%	0.1100	0.1389

Tab. 3: Risk analysis of the dynamic exchange rate portfolio. The best results to fulfill the regulatory requirement are marked by ^{r}. The recommended method to the investor is marked by ^{i}. For the internal supervisory, we recommend the method marked by ^{s}.

vector with pr-quantiles of individual innovations does not necessarily correspond to the pr-quantile of the portfolio return. Under the Gaussian distributional assumption, the standardized DCCN returns are theoretically cross independent and the Gaussian quantiles of the portfolio can be easily calculated. The dynamic mean, variance of the portfolio's returns have values of:

$$\begin{aligned} \mathsf{E}\{r(t)\} &= b(t)^{\top}\Sigma_x^{(1/2)}(t)\,\mathsf{E}\{\varepsilon_x(t)\} \\ Var\{r(t)\} &= b(t)^{\top}\Sigma_x^{(1/2)}(t)\; Var\{\varepsilon_x(t)\}\Sigma_x^{(1/2)\top}(t)b(t) \end{aligned}$$

The GHICA method in general presents better results than the DCCN. Except the value of ES at 1% level, the GHICA fulfills the requirements of regulatory, internal supervisory and investors, see Table 3. For $h = 1$ day forecasts, the DCCN gives although a closer VaR value to 1.6%, i.e. the ideal probability for regulatory, its risk charge with a value of 0.0494 is larger than that based on the GHICA, 0.0453. Therefore the GHICA is more favored in fulfilling the minimal regulatory requirement.

The two real data studies show that the GHICA method fulfills the minimal regulatory requirement by controlling the risk inside 1.6% level and requiring small risk charge, in particular satisfies the internal supervisory requirement by precisely measuring risk level as expected and favors the investors' requirement by delivering small size of loss. In summary, the GHICA method is not only a realistic and fast procedure given either static or dynamic portfolios but also produces better results than several alternative risk management methods.

References

Anderson, T., Bollerslev, T., Diebold, F. and Labys, P. (2001). The distribution of realized exchange rate volatility, *Journal of the American Statistical Association* pp. 42–55.

Barndorff-Nielsen, O. and Blæsild, P. (1981). Hyperbolic distribution and ramifications: Contributions to theory and applications, *in* C. Taillie, P. Patil and A. Baldessari (eds), *Statistical Distributions in Scientific Work*, Vol. 4, D. Reidel, pp. 19–44.

Belomestny, D. and Spokoiny, V. (2006). Spatial aggregation of local likelihood estimates with applications to classification, *WIAS Preprint*.

Borak, S., Detlefsen, K. and Härdle, W. (2005). FFT-based option pricing, *in* P. Cizek, W. Härdle and R. Weron (eds), *Statistical Tools for Finance and Insurance*, Springer Verlag.

Chen, Y. and Spokoiny, V. (2006). Local exponential smoothing with applications to volatility estimation and risk management, *working paper*.

Chen, Y., Härdle, W. and Jeong, S. (2005). Nonparametric risk management with generalized hyperbolic distributions, *SFB 649, Discussion paper 2005-001, http://sfb649.wiwi.hu-berlin.de*.

Chen, Y., Härdle, W. and Spokoiny, V. (2006). Portfolio value at risk based on independent components analysis, *Journal of Computational and Applied Mathematics, forthcoming*.

Cover, T. and Thomas, J. (1991). *Elements of information theory*, Wiley.

Eberlein, E. and Prause, K. (2002). The generalized hyperbolic model: financial derivatives and risk measures, *in* H. Geman, D. Madan, S. Pliska and T. Vorst (eds), *Mathematical Finance-Bachelier Congress 2000*, Springer Verlag.

Engle, R. (2002). Dynamic conditional correlation - a simple class of multivariate garch models, *Journal of Business and Economic Statistics, 20(3)* pp. 339–350.

Engle, R. and Kroner, F. (1995). Multivariate simultaneous generalized arch, *Econometric Theory 11* pp. 122–150.

Engle, R. and Sheppard, K. (2001). Theoretical and empirical properties of dynamic conditional correlation multivariate garch, *NBER Working Paper 8554*.

Flury, B. (1998). *Common Principal Components and Related Multivariate Models*, John Wiley & Sons, Inc.

Franke, J., Härdle, W. and Hafner, C. (2004). *Statistics of Financial Markets*, Springer-Verlag Berlin Heidelberg New York.

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

Härdle, W., Herwartz, H. and Spokoiny, V. (2003). Time inhomogeneous multiple volatility modelling, *Journal of Financial Econometrics* **1**: 55–95.

Hyvärinen, A. (1998). *New Approximations of Differential Entropy for Independent Component Analysis and Projection Pursuit*, MIT Press, pp. 273–279.

Hyvärinen, A., Karhunen, J. and Oja, E. (2001). *Independent Component Analysis*, John Wiley & Sons, Inc.

Jorion, P. (2001). *Value at Risk*, McGraw-Hill.

Menn, C. and Rachev, S. (2004). Calibrated FFT-based density approximations for α-stable distributions.

Merton, R. (1973). Theory of rational option pricing, *The Bell Journal of Economics and Management Science* **4**: 141–183.

Polzehl, J. and Spokoiny, V. (2006). Propagation-separation approach for local likelihood estimation, *Probability Theory and Related Fields* pp. 335–362.

Empirical Pricing Kernels and Investor Preferences

K. Detlefsen[1], W. K. Härdle[2], R. A. Moro[3],

[1]CASE – Center for Applied Statistics and Economics, Humboldt-Universität zu Berlin, Spandauer Straße 1, 10178 Berlin, Germany; e-mail: detlefsen@wiwi.hu-berlin.de; phone: +49(0)30 2093-5807

[2]CASE – Center for Applied Statistics and Economics, Humboldt-Universität zu Berlin, Spandauer Straße 1, 10178 Berlin, Germany; e-mail: haerdle@wiwi.hu-berlin.de; phone: +49(0)30 2093-5630

[3]German Institute for Economic Research, Königin-Luise-Straße 5, 14195 Berlin, Germany; e-mail: rmoro@diw.de; phone: +49(0)30 8978-9262 and CASE – Center for Applied Statistics and Economics, Humboldt-Universität zu Berlin, Spandauer Straße 1, 10178 Berlin

Abstract

This paper analyzes empirical market utility functions and pricing kernels derived from the DAX and DAX option data for three market regimes. A consistent parametric framework of stochastic volatility is used. All empirical market utility functions show a region of risk proclivity that is reproduced by adopting the hypothesis of heterogeneous individual investors whose utility functions have a switching point between bullish and bearish attitudes. The inverse problem of finding the distribution of individual switching points is formulated in the space of stock returns by discretization as a quadratic optimization problem. The resulting distributions vary over time and correspond to different market regimes.

JEL classification: G12, G13, C50

Keywords: Utility function, pricing kernel, behavioral finance, risk aversion, risk proclivity, Heston model

1 Introduction

Numerous attempts have been undertaken to describe basic principles on which the behaviour of individuals are based. Expected utility theory was originally proposed by J. Bernoulli in 1738. In his work J. Bernoulli used such terms as risk aversion and risk premium and proposed a concave (logarithmic) utility function, see Bernoulli (1956). The utilitarianism theory that emerged in the 18th century considered utility maximization as a principle for the organisation of society. Later the expected utility idea was applied to game theory and formalized by von Neumann and Morgenstern (1944). A utility function relates some observable variable, in most cases consumption, and an unobservable utility level that this consumption delivers. It was suggested that individuals' preferences are based on this unobservable utility: such bundles of goods are preferred that are associated with higher utility levels. It was claimed that three types of utility functions – concave, convex and linear – correspond to three types of individuals – risk averse, risk neutral and risk seeking. A typical economic agent was considered to be risk averse and this was quantified by coefficients of relative or absolute risk aversion. Another important step in the development of utility theory was the prospect theory of Kahneman and Tversky (1979). By behavioural experiments they found that people act risk averse above a certain reference point and risk seeking below it. This implies a concave form of the utility function above the reference point and a convex form below it.

Besides these individual utility functions, market utility functions have recently been analyzed in empirical studies by Jackwerth (2000), Rosenberg and Engle (2002) and others. Across different markets, the authors observed a common pattern in market utility functions: There is a reference point near the initial wealth and in a region around this reference point the market utility functions are convex. But for big losses or gains they show a concave form – risk aversion. Such utility functions disagree with the classical utility functions of von Neumann and Morgenstern (1944) and also with the findings of Kahneman and Tversky (1979). They are however in concordance with the utility function form proposed by Friedman and Savage (1948).

In this paper, we analyze how these market utility functions can be explained by aggregating individual investors' attitudes. To this end, we first determine empirical pricing kernels from DAX data. Our estimation procedure is based on historical and risk neutral densities and these distributions are derived with stochastic volatility models that are widely used in industry. From these pricing kernels we construct the corresponding market utility functions. Then we describe our method of aggregating individual utility functions to a market utility function. This leads to an inverse problem for

the density function that describes how many investors have the utility function of each type. We solve this problem by discrete approximation. In this way, we derive utility functions and their distribution among investors that allow to recover the market utility function. Hence, we explain how (and what) individual utility functions can be used to form the behaviour of the whole market.

The paper is organized as follows: In section 2, we describe the theoretical connection between utility functions and pricing kernels. In section 3, we present a consistent stochastic volatility framework for the estimation of both the historical and the risk neutral density. Moreover, we discuss the empirical pricing kernel implied by the DAX in 2000, 2002 and 2004. In section 4, we explain the utility aggregation method that relates the market utility function and the utility functions of individual investors. This aggregation mechanism leads to an inverse problem that is analyzed and solved in this section. In section 5, we conclude and discuss related approaches.

2 Pricing kernels and utility functions

In this section, we derive the fundamental relationship between utility functions and pricing kernels. It describes how a representative utility function can be derived from historical and risk-neutral distributions of assets. In the following sections, we estimate the empirical pricing kernel and observe in this way the market utility function.

First, we derive the price of a security in an equilibrium model: we consider an investor with a utility function U who has as initial endowment one share of stock. He can invest into the stock and a bond up to a final time when he can consume. His problem is to choose a strategy that maximizes the expected utility of his initial and terminal wealth. In continuous time, this leads to a well known optimization problem introduced by Merton (1973) for stock prices modelled by diffusions. In discrete time, it is a basic optimization problem, see Cochrane (2001).

From this result, we can derive the asset pricing equation

$$P_0 = \mathrm{E}^P \left[\psi(S_T) M_T\right]$$

for a security on the stock (S_t) with payoff function ψ at maturity T. Here, P_0 denotes the price of the security at time 0 and E^P is the expectation with respect to the real/historical measure P. The stochastic discount factor M_T is given by

$$M_T = \beta U'(S_T)/U'(S_0) \tag{1}$$

where β is a fixed discount factor. This stochastic discount factor is actually the projection of the general stochastic discount factor on the traded asset (S_t). The stochastic discount factor can depend on more variables in general. But as discussed in Cochrane (2001) this projection has the same interpretation for pricing as the general stochastic discount factor.

Besides this equilibrium based approach, Black and Scholes (1973) derived the price of a security relative to the underlying by constructing a perfect hedge. The resulting continuous delta hedging strategy is equivalent to pricing under a risk neutral measure Q under which the discounted price process of the underlying becomes a martingale. Hence, the price of a security is given by an expected value with respect to a risk neutral measure Q:

$$P_0 = \mathrm{E}^Q \left[\exp(-rT)\psi(S_T)\right]$$

If p denotes the historical density of S_T (i.e. $P(S_T \leq s) = \int_{-\infty}^{s} p(x)\ dx$) and q the risk neutral density of S_T (i.e. $Q(S_T \leq s) = \int_{-\infty}^{s} q(x)\ dx$) then we get

$$\begin{aligned} P_0 &= \exp(-rT) \int \psi(x) q(x) dx \\ &= \exp(-rT) \int \psi(x) \frac{q(x)}{p(x)} p(x) dx \\ &= \mathrm{E}^P \left[\exp(-rT)\psi(S_T)\frac{q(S_T)}{p(S_T)}\right] \end{aligned} \tag{2}$$

Combining equations (1) and (2) we see

$$\beta \frac{U'(s)}{U'(S_0)} = \exp(-rT)\frac{q(s)}{p(s)}.$$

Defining the pricing kernel by $K = q/p$ we conclude that the form of the market utility function can be derived from the empirical pricing kernel by integration:

$$\begin{aligned} U(s) &= U(S_0) + \int_{S_0}^{s} U'(S_0) \frac{\exp(-rT)}{\beta} \frac{q(x)}{p(x)} dx \\ &= U(S_0) + \int_{S_0}^{s} U'(S_0) \frac{\exp(-rT)}{\beta} K(x) dx \end{aligned}$$

because S_0 is known.

As an example, we consider the model of Black and Scholes (1973) where the stock follows a geometric Brownian motion

$$dS_t/S_t = \mu dt + \sigma dW_t \tag{3}$$

Here the historical density p of S_t is log-normal, i.e.

$$p(x) = \frac{1}{x}\frac{1}{\sqrt{2\pi\tilde{\sigma}^2}} \exp\left\{-\frac{1}{2}\left(\frac{\log x - \tilde{\mu}}{\tilde{\sigma}}\right)^2\right\},\ x > 0$$

where $\tilde{\mu} = (\mu - \sigma^2/2)t + \log S_0$ and $\tilde{\sigma} = \sigma\sqrt{t}$. Under the risk neutral measure Q the drift μ is replaced by the riskless interest rate r, see e.g. Harrison and Pliska (1981). Thus, also the risk neutral density q is log-normal. In this way, we can derive the pricing kernel

$$K(x) = \left(\frac{x}{S_0}\right)^{-\frac{\mu-r}{\sigma^2}} \exp\{(\mu - r)(\mu + r - \sigma^2)T/(2\sigma^2)\}.$$

This pricing kernel has the form of a derivative of a power utility

$$K(x) = \lambda\left(\frac{x}{S_0}\right)^{-\gamma}$$

where the constants are given by $\lambda = e^{\frac{(\mu-r)(\mu+r-\sigma^2)T}{2\sigma^2}}$ and $\gamma = \frac{\mu-r}{\sigma^2}$. This gives a utility function corresponding to the underlying (3)

$$U(S_T) = (1 - \frac{\mu - r}{\sigma^2})^{-1}\ S_T^{(1-\frac{\mu-r}{\sigma^2})}$$

where we ignored additive and multiplicative constants. In this power utility function the risk aversion is not given by the market price of risk $(\mu - r)/\sigma$. Instead investors take the volatility more into account. The expected return $\mu - r$ that is adjusted by the riskfree return is related to the variance. This results in a higher relative risk aversion than the market price of risk.

A utility function corresponding to the Black-Scholes model is shown in the upper panel of figure 1 as a function of returns. In order to make different market situations comparable we consider utility functions as functions of (half year) returns $R = S_{0.5}/S_0$. We chose the time horizon of half a year ahead for our analysis. Shorter time horizons are interesting economically and moreover the historical density converges to the Dirac measure so that results become trivial (in the end). Longer time horizons are economically

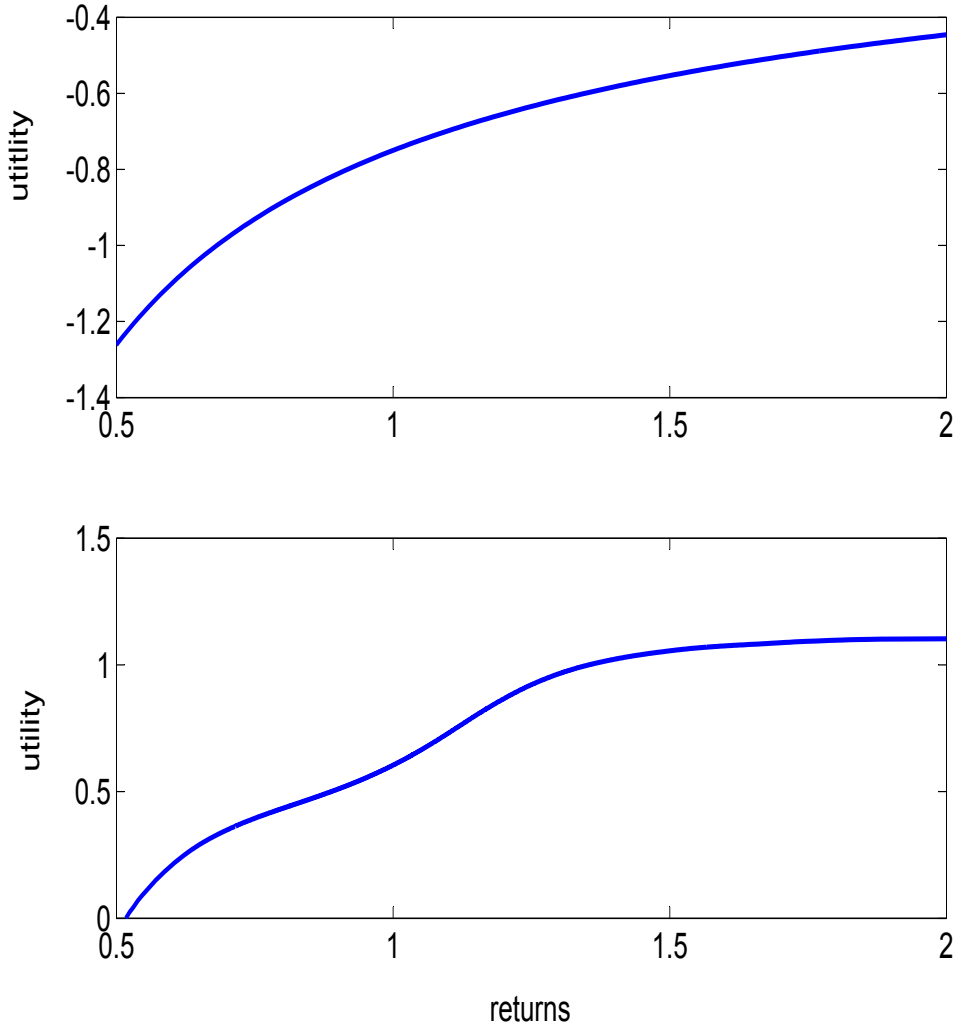


Figure 1: up: Utility function in the Black Scholes model for $T = 0.5$ years ahead and drift $\mu = 0.1$, volatility $\sigma = 0.2$ and interest rate $r = 0.03$. down: Market utility function on 06/30/2000 for $T = 0.5$ years ahead.

more interesting but it is hardly possible to estimate the historical density for a long time ahead. It neither seems realistic to assume that investors have clear ideas where the DAX will be in e.g. 10 years. For these reasons we use half a year as future horizon. Utility functions $\tilde{U}$ of returns are defined by:

$$\tilde{U}(R) := U(RS_0),\ R > 0$$

where S_0 denotes the value of the DAX on the day of estimation. Because of $U' = cK$ for a constant c we have $\tilde{U}'(R) = cK(RS_0)S_0$ and we see that also utility functions of returns are given as integrals of the pricing kernel. The change to returns allows us to compare different market regimes independently of the initial wealth. In the following we denote the utility functions of returns by the original notation U. Hence, we suppress in the notation the dependence of the utility function U on the day of estimation t.

The utility function corresponding to the model of Black and Scholes (1973) is a power utility, monotonically increasing and concave. But such classical utility functions are not observed on the market. Parametric and nonparametric models that replicate the option prices all lead to utility functions with a hump around the initial wealth level. This is described in detail later but is shown already in figure 1. The upper panel presents the utility function corresponding to Black-Scholes model with a volatility of 20% and an expected return of 10%. The function is concave and implies a constant relative risk aversion. The utility function estimated on the bullish market in summer 2000 is presented in the lower panel. Here, the hump around the money is clearly visible. The function is no more concave but has a region where investors are risk seeking. This risk proclivity around the money is reflected in a negative relative risk aversion.

3 Estimation

In this section, we start by reviewing some recent approaches for estimating the pricing kernel. Then we describe our method that is based on estimates of the risk neutral and the historical density. The risk neutral density is derived from option prices that are given by an implied volatility surface and the historical density is estimated from the independent data set of historical returns. Finally, we present the empirical pricing kernels and the inferred utility and relative risk aversion functions.

3.1 Estimation approaches for the pricing kernel

There exist several ways and methods to estimate the pricing kernel. Some of these methods assume parametric models while others use nonparametric techniques. Moreover, some methods estimate first the risk neutral and subjective density to infer the pricing kernel. Other approaches estimate directly the pricing kernel.

Ait-Sahalia and Lo (1998) derive a nonparametric estimator of the risk neutral density based on option prices. In Ait-Sahalia and Lo (2000), they consider the empirical pricing kernel and the corresponding risk aversion using this estimator. Moreover, they derive asymptotic properties of the estimator that allow e.g. the construction of confidence bands. The estimation procedure consists of two steps: First, the option price function is determined by nonparametric kernel regression and then the risk neutral density is computed by the formula of Breeden and Litzenberger (1978). Advantages of this approach are the known asymptotic properties of the estimator and the few assumptions necessary.

Jackwerth (2000) analyses risk aversion by computing the risk neutral density from option prices and the subjective density from historical data of the underlying. For the risk neutral distribution, he applies a variation of the estimation procedure described in Jackwerth and Rubinstein (1996): A smooth volatility function derived from observed option prices gives the risk neutral density by differentiating it twice. The subjective density is approximated by a kernel density computed from historical data. In this method bandwidths have to be chosen as in the method of Ait-Sahalia and Lo (1998).

Rosenberg and Engle (2002) use a different approach and estimate the subjective density and directly (the projection of) the pricing kernel. This gives the same information as the estimation of the two densities because the risk neutral density is the product of the pricing kernel and the subjective density. For the pricing kernel, they consider two parametric specifications as power functions and as exponentials of polynomials. The evolution of the underlying is modelled by GARCH processes. As the parametric pricing kernels lead to different results according to the parametric form used this parametric approach appears a bit problematic.

Chernov (2003) also estimates the pricing kernel without computing the risk neutral and subjective density explicitly. Instead of assuming directly a parametric form of the kernel he starts with a (multi dimensional) modified model of Heston (1993) and derives an analytic expression for the pricing kernel by the Girsanov theorem, see Chernov (2000) for details. The ker-

nel is estimated by a simulated method of moments technique from equity, fixed income and commodities data and by reprojection. An advantage of this approach is that the pricing kernel is estimated without assuming an equity index to approximate the whole market portfolio. But the estimation procedure is rather complex and model dependent.

In a recent paper, Barone-Adesi et al. (2004) price options in a GARCH framework allowing the volatility to differ between historical and risk neutral distribution. This approach leads to acceptable calibration errors between the observed option prices and the model prices. They estimate the historical density as a GARCH process and consider the pricing kernel only on one day. This kernel is decreasing which coincides with standard economic theory. But the general approach of changing explicitly the volatility between the historical and risk neutral distribution is not supported by the standard economic theory.

We estimate the pricing kernel in this paper by estimating the risk neutral and the subjective density and then deriving the pricing kernel. This approach does not impose a strict structure on the kernel. Moreover, we use accepted parametric models because nonparametric techniques for the estimation of second derivatives depend a lot on the bandwidth selection although they yield the same pricing kernel behaviour over a wide range of bandwidths. For the risk neutral density we use a stochastic volatility model that is popular both in academia and in industry. The historical density is more difficult to estimate because the drift is not fixed. Hence, the estimation depends more on the model and the length of the historical time series. In order to get robust results we consider different (discrete) models and different lengths. In particular, we use a GARCH model that is the discrete version of the continuous model for the risk neutral density. In the following, we describe these models, their estimation and the empirical results.

3.2 Estimation of the risk neutral density

Stochastic volatility models are popular in industry because they replicate the observed smile in the implied volatility surfaces (IVS) rather well and moreover imply rather realistic dynamics of the surfaces. Nonparametric approaches like the local volatility model of Dupire (1994) allow a perfect fit to observed price surfaces but their dynamics are in general contrary to the market. As Bergomi (2005) points out the dynamics are more important for modern products than a perfect fit. Hence, stochastic volatility models are popular.

We consider the model of Heston (1993) for the risk neutral density be-

cause it can be interpreted as the limit of GARCH models. The Heston model has been refined further in order to improve the fit, e.g. by jumps in the stock price or by a time varying mean variance level. We use the original Heston model in order to maintain a direct connection to GARCH processes. Although it is possible to estimate the historical density also with the Heston model e.g. by Kalman filter methods we prefer more direct approaches in order to reduce the dependence of the results on the model and the estimation technique.

The stochastic volatility model of Heston (1993) is given by the two stochastic differential equations:

$$\frac{dS_t}{S_t} = rdt + \sqrt{V_t}dW_t^1$$

where the variance process is modelled by a square-root process:

$$dV_t = \xi(\eta - V_t)dt + \theta\sqrt{V_t}dW_t^2$$

and W^1 and W^2 are Wiener processes with correlation ρ and r is the risk free interest rate. The first equation models the stock returns by normal innovations with stochastic variance. The second equation models the stochastic variance process as a square-root diffusion.

The parameters of the model all have economic interpretations: η is called the long variance because the process always returns to this level. If the variance V_t is e.g. below the long variance then $\eta - V_t$ is positive and the drift drives the variance in the direction of the long variance. ξ controls the speed at which the variance is driven to the long variance. In calibrations, this parameter changes a lot and makes also the other parameters instable. To avoid this problem, the reversion speed is kept fixed in general. We follow this approach and choose $\xi = 2$ as Bergomi (2005) does. The volatility of variance θ controls mainly the kurtosis of the distribution of the variance. Moreover, there are the initial variance V_0 of the variance process and the correlation ρ between the Brownian motions. This correlation models the leverage effect: When the stock goes down then the variance goes up and vice versa. The parameters also control different aspects of the implied volatility surface. The short (long) variance determines the level of implied volatility for short (long) maturities. The correlation creates the skew effect and the volatility of variance controls the smile.

The variance process remains positive if the volatility of variance θ is small enough with respect to the product of the mean reversion speed ξ and

the long variance level η (i.e. $2\xi\eta > \theta^2$). As this constraint leads often to significantly worse fits to implied volatility surfaces it is in general not taken into account and we follow this approach.

The popularity of this model can probably be attributed to the semiclosed form of the prices of plain vanilla options. Carr and Madan (1999) showed that the price $C(K,T)$ of a European call option with strike K and maturity T is given by

$$C(K,T) = \frac{\exp\{-\alpha \ln(K)\}}{\pi} \int_0^{+\infty} \exp\{-\mathbf{i}v \ln(K)\}\psi_T(v)dv$$

for a (suitable) damping factor $\alpha > 0$. The function ψ_T is given by

$$\psi_T(v) = \frac{\exp(-rT)\phi_T\{v-(\alpha+1)\mathbf{i}\}}{\alpha^2+\alpha-v^2+\mathbf{i}(2\alpha+1)v}$$

where ϕ_T is the characteristic function of $\log(S_T)$. This characteristic function is given by

$$\begin{aligned}\phi_T(z) &= \exp\{\frac{-(z^2+\mathbf{i}z)V_0}{\gamma(z)\coth\frac{\gamma(z)T}{2}+\xi-\mathbf{i}\rho\theta z}\} \\ &\quad \times \frac{\exp\{\frac{\xi\eta T(\xi-\mathbf{i}\rho\theta z)}{\theta^2}+\mathbf{i}zTr+\mathbf{i}z\log(S_0)\}}{(\cosh\frac{\gamma(z)T}{2}+\frac{\xi-\mathbf{i}\rho\theta z}{\gamma(z)}\sinh\frac{\gamma(z)T}{2})^{\frac{2\xi\eta}{\theta^2}}} \end{aligned} \quad (4)$$

where $\gamma(z) \stackrel{\text{def}}{=} \sqrt{\theta^2(z^2+\mathbf{i}z)+(\xi-\mathbf{i}\rho\theta z)^2}$, see e.g. Cizek et al. (2005).

For the calibration we minimize the absolute error of implied volatilities based on the root mean square error:

$$\text{ASE}_t \stackrel{\text{def}}{=} \sqrt{\sum_{i=1}^{n} n^{-1}\{IV_i^{mod}(t) - IV_i^{mar}(t)\}^2}$$

where *mod* refers to a model quantity, *mar* to a quantity observed on the market and $IV(t)$ to an implied volatility on day t. The index i runs over all n observations of the surface on day t.

It is essential for the error functional ASE_t which observed prices are used for the calibration. As we investigate the pricing kernel for half a year to maturity we use only the prices of options that expire in less than 1.5 years. In order to exclude liquidity problems occurring at expiry we consider for the

calibration only options with more than 1 month time to maturity. In the moneyness direction we restrict ourselves to strikes 50% above or below the spot for liquidity reasons.

The risk neutral density is derived by estimation of the model parameters by a least squares approach. This amounts to the minimization of the error functional ASE_t. Cont and Tankov (2004) provided evidence that such error functionals may have local minima. In order to circumvent this problem we apply a stochastic optimization routine that does not get trapped in a local minimum. To this end, we use the method of differential evolution developed by Storn and Price (1997).

Having estimated the model parameters we know the distribution of $X_T = \log S_T$ in form of the characteristic function ϕ_T, see (4). Then the corresponding density f of X_T can be recovered by Fourier inversion:

$$f(x) = \frac{1}{2\pi} \int_{-\infty}^{\infty} e^{\mathbf{i}tx} \phi_T(t) dt,$$

see e.g. Billingsley (1995). This integral can be computed numerically.

Finally, the risk neutral density q of $S_T = \exp(X_T)$ is given as a transformed density:

$$q(x) = \frac{1}{x} f\{\log(x)\}.$$

This density q is risk neutral because it is derived from option prices and options are priced under the risk neutral measure. This measure is applied because banks replicate the payoff of options so that no arbitrage conditions determine the option price, see e.g. Rubinstein (1994). An estimated risk neutral density is presented in figure 2. It is estimated from the implied volatility shown in figure 3 for the day 24/03/2000. The distribution is right skewed and its mean is fixed by the martingale property. This implies that the density is low for high profits and high for high losses. Moreover, the distribution is not symmetrical around the neutral point where there are neither profits nor losses. For this and all the following estimations 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 have to consider dividend payments explicitly.

3.3 Estimation of the historical density

While the risk neutral density is derived from option prices observed on the day of estimation we derive the subjective density from the historical time

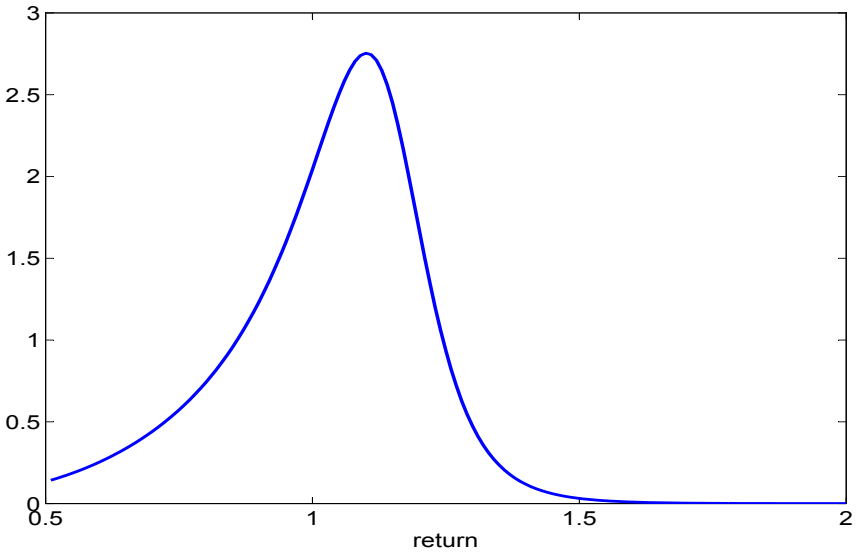


Figure 2: Risk neutral density on 24/03/2000 half a year ahead.

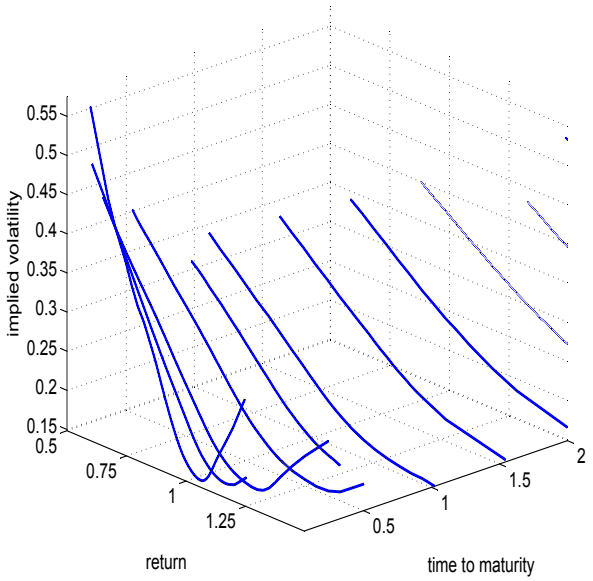


Figure 3: Implied volatility surface on 24/03/00.

model	time period
GARCH in mean	2.0y
discrete Heston	2.0y
observed returns	1.0y

Table 1: Models and the time periods used for their estimation.

series of the index. Hence, the two data sets are independent in the sense that the option prices reflect the future movements and the historical time series the past.

The estimation of the historical density seems more difficult than the estimation of the risk neutral density because the drift is not fixed and it depends in general on the length of the time series. Because of these difficulties we use different models and time horizons for the historical density: First, we estimate a GARCH in mean model for the returns. Returns are generally assumed to be stationary and we confirmed this at least in the time intervals we consider. The mean component in the GARCH model is important to reflect different market regimes. We estimate the GARCH model from the time series of the returns of the last two year because GARCH models require quite long time series for the estimation in order to make the standard error reasonably small. We do not choose longer time period for the estimation because we want to consider special market regimes. Besides this popular model choice we apply a GARCH model that converges in the limit to the Heston model that we used for the risk neutral density. As this model is also hard to estimate we use again the returns of the last 2 years for this model. Moreover, we consider directly the observed returns of the last year. The models and their time period for the estimation are presented in table 1. All these models give by simulation and smoothing the historical density for half a year ahead.

The GARCH estimations are based on the daily log-returns

$$R_i = \log(S_{t_i}) - \log(S_{t_{i-1}})$$

where (S_t) denotes the price process of the underlying and $t_i,\ i = 1, 2, \ldots$ denote the settlement times of the trading days. Returns of financial assets have been analyzed in numerous studies, see e.g. Cont (2001). A model that has often been successfully applied to financial returns and their stylized facts

is the GARCH(1,1) model. This model with a mean is given by

$$R_i = \mu + \sigma_i Z_i$$
$$\sigma_i^2 = \omega + \alpha R_{i-1}^2 + \beta \sigma_{i-1}^2$$

where (Z_i) are independent identically distributed innovations with a standard normal distribution, see e.g. Franke et al. (2004). On day t_j the model parameters μ, ω, α and β are estimated by quasi maximum likelihood from the observations of the last two years, i.e. $R_{j-504}, \ldots, R_j$ assuming 252 trading days per year.

After the model parameters have been estimated on day t_j from historical data the process of logarithmic returns (R_i) is simulated half a year ahead, i.e. until time $t_j + 0.5$. In such a simulation μ, ω, α and β are given and the time series (σ_i) and (R_i) are unknown. The values of the DAX corresponding to the simulated returns are then given by inverting the definition of the log returns:

$$S_{t_i} = S_{t_{i-1}} \exp(R_i)$$

where we start with the observed DAX value on day t_j. Repeating the simulation N times we obtain N samples of the distribution of $S_{t_j+0.5}$. We use $N = 2000$ simulations because tests have shown that the results become robust around this number of simulations.

From these samples we estimate the probability density function of $S_{t_j+0.5}$ (given $(S_{t_{j-126}}, \ldots, S_{t_j})$) by kernel density estimation. We apply the Gaussian kernel and choose the bandwidth by Silverman's rule of thumb, see e.g. Silverman (1986). This rule provides a trade-off between oversmoothing – resulting in a high bias – and undersmoothing – leading to big variations of the density. We have moreover checked the robustness of the estimate relative to this bandwidth choice. The estimation results of a historical density are presented in figure 4 for the day 24/03/2000. This density that represents a bullish market is has most of its weight in the profit region and its tail for the losses is relatively light.

As we use the Heston model for the estimation of the risk neutral density we consider in addition to the described GARCH model a GARCH model that is a discrete version of the Heston model. Heston and Nandi (2000) show that the discrete version of the square-root process is given by

$$V_i = \omega + \beta V_{i-1} + \alpha(Z_{i-1} - \gamma\sqrt{V_{i-1}})$$

and the returns are modelled by

$$R_i = \mu - \frac{1}{2}V_i + \sqrt{V_i} Z_i$$

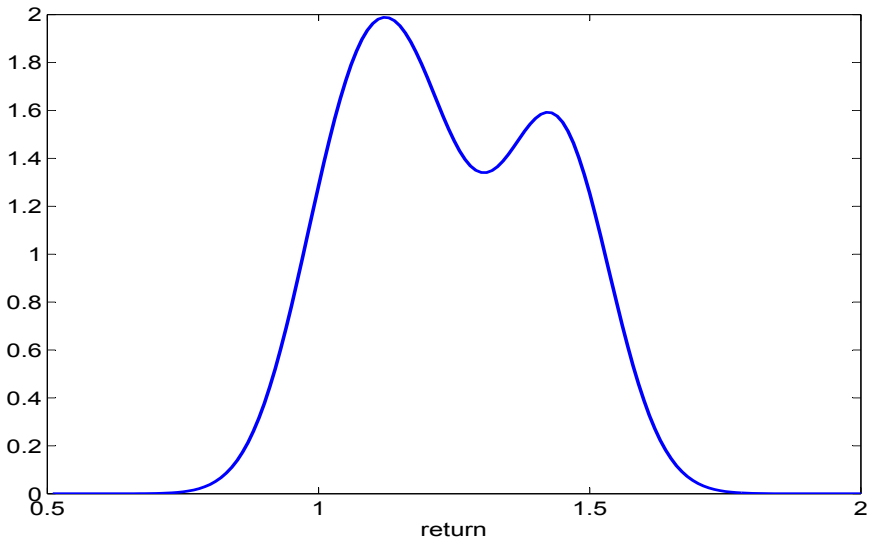


Figure 4: Historical density on 24/03/2000 half a year ahead.

where (Z_i) are independent identically distributed innovations with a standard normal distribution. Having estimated this model by maximum likelihood on day t_j we simulate it half a year ahead and then smooth the samples of $S_{t_j+0.5}$ in the same way as in the other GARCH model.

In addition to these parametric models, we consider directly the observed returns over half a year

$$\tilde{R}_i = S_{t_i}/S_{t_{i-126}}.$$

In this way, we interpret these half year returns as samples from the distribution of the returns for half a year ahead. Smoothing these historical samples of returns gives an estimate of the density of returns and in this way also an estimate of the historical density of $S_{t_j+0.5}$.

3.4 Empirical pricing kernels

In contrast to many other studies that concentrate on the S&P500 index we analyze the German economy by focusing on the DAX, the German stock index. This broad index serves as an approximation to the German economy. We use two data sets: A daily time series of the DAX for the estimation of the subjective density and prices of European options on the DAX for the estimation of the risk neutral density.

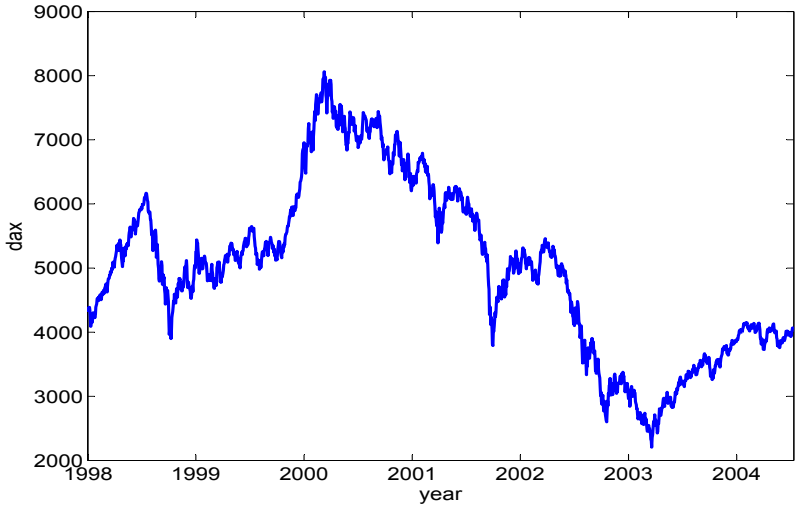


Figure 5: DAX, 1998 - 2004.

	1.0y	2.0y
03/2000	1.63	1.57
07/2002	0.66	0.54
06/2004	1.11	0.98

Table 2: Market regimes in 2000, 2002 and 2004 described by the return $S_0/S_{0-\Delta}$ for periods $\Delta = 1.0y, 2.0y$.

In figure 5, we present the DAX in the years 1998 to 2004. This figure shows that the index reached its peak in 2000 when all the internet firms were making huge profits. But in the same year this bubble burst and the index fell afterwards for a long time. The historical density is estimated from the returns of this time series. We analyze the market utility functions in March 2000, July 2002 and June 2004 in order to consider different market regimes. We interpret 2000 as a bullish, 2002 as a bearish and 2004 as a unsettled market. These interpretations are based on table 2 that describes the changes of the DAX over the preceding 1 or 2 years. (In June 2004 the market went up by 11% in the last 10 months.)

A utility function derived from the market data is a market utility function. It is estimated as an aggregate for all investors as if the representative investor existed. A representative investor is however just a convenient con-

struction because the existence of the market itself implies that the asset is bought and sold, i.e. at least two counterparties are required for each transaction.

In section 2 we identified the market utility function (up to linear transformations) as

$$U(R) = \int_{R_0}^{R} K(x)dx$$

where K is the pricing kernel for returns. It is defined by

$$K(x) = q(x)/p(x)$$

in terms of the historical and risk neutral densities p and q of returns. Any utility function (both cardinal and ordinal) can be defined up to a linear transformation, therefore we have identified the utility functions sufficiently. In section 3.3 we proposed different models for estimating the historical density. In figure 6 we show the pricing kernels resulting from the different estimation approaches for the historical density. The figure shows that all three kernels are quite similar: They have the same form, the same characteristic features like e.g. the hump and differ in absolute terms only a little. This demonstrates the economic equivalence of the three estimation methods on this day and this equivalence holds also for the other days. In the following we work with historical densities that are estimated by the observed returns.

Besides the pricing kernel and the utility function we consider also the risk attitudes in the markets. Such risk attitudes are often described in terms of relative risk aversion that is defined by

$$RRA(R) = -R\frac{U''(R)}{U'(R)}.$$

Because of $U' = cK = cq/p$ for a constant c the relative risk aversion is also given by

$$RRA(R) = -R\frac{q'(R)p(R) - q(R)p'(R)}{p^2(R)} \Big/ \frac{q(R)}{p(R)} = R\left(\frac{p'(R)}{p(R)} - \frac{q'(R)}{q(R)}\right).$$

Hence, we can estimate the relative risk aversion from the estimated historical and risk neutral densities.

In figure 7 we present the empirical pricing kernels in March 2000, July 2002 and June 2004. The dates represent a bullish, a bearish and an unsettled markets, see table 2. All pricing kernels have a proclaimed hump located

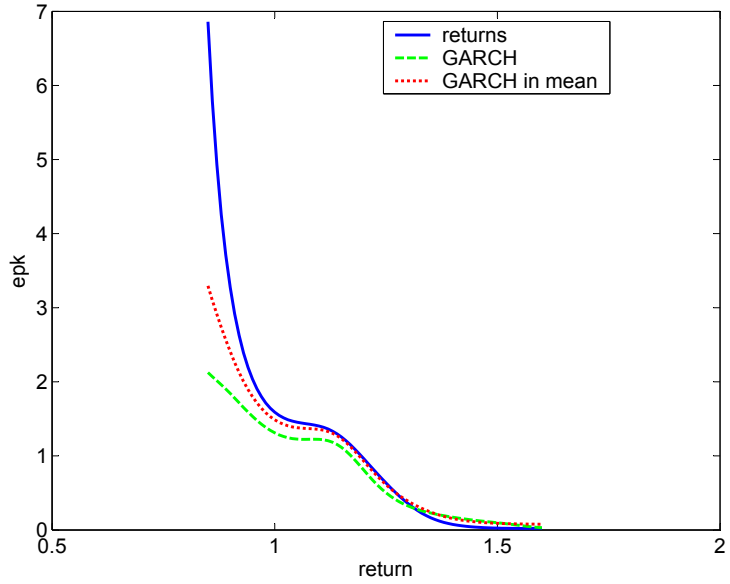


Figure 6: Empirical pricing kernel on 24/03/2000 (bullish market).

at small profits. Hence, the market utility functions do not correspond to standard specification of utility functions. We present the pricing kernels only in regions around the initial DAX (corresponding to a return of 1) value because the kernels explode outside these regions. This explosive behaviour reflects the typical pricing kernel form for losses. The explosion of the kernel for large profits is due to numerical problems in the estimation of the very low densities in this region. But we can see that in the unsettled market the kernel is concentrated on a small region while the bullish and bearish markets have wider pricing kernels. The hump of the unsettled market is also narrower than in the other two regimes. The bullish and bearish regimes have kernels of similar width but the bearish kernel is shifted to the loss region and the bullish kernel is located mainly in the profit area. Moreover, the figures show that the kernel is steeper in the unsettled markets than in the other markets. But this steepness cannot be interpreted clearly because pricing kernels are only defined up to a multiplicative constant.

The pricing kernels are the link between the relative risk aversion and the utility functions that are presented in figure 8. These utility functions are only defined up to linear transformations, see section 2. All the utility functions are increasing but only the utility function of the bullish market is concave. This concavity can be seen from the monotonicity of the kernel, see figure 7. Actually, this non convexity can be attributed to the quite special

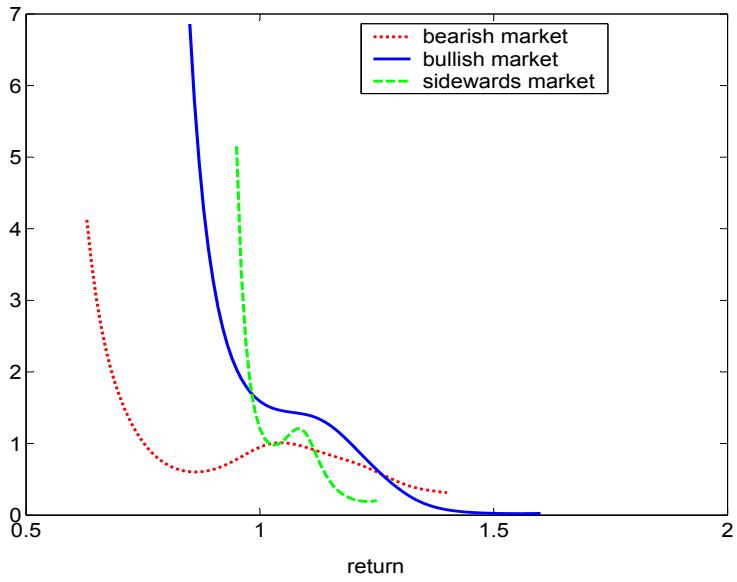


Figure 7: Empirical pricing kernel on 24/03/2000 (bullish), 30/07/2002 (bearish) and 30/06/2004 (unsettled or sidewards market).

form of the historical density which has two modes on this date, see figure 4. Hence, we presume that also this utility function has in general a region of convexity. The other two utility functions are convex in a region of small profits where the bullish utility is almost convex. The derivatives of the utility functions cannot be compared directly because utility functions are identified only up to multiplicative constants. But we can compare the ratio of the derivatives in the loss and profit regions for the three dates because the constants cancel in these ratios. We see that the derivatives in the loss region are highest in the bullish and lowest in the bearish market and vice versa in the profit region. Economically these observations can be interpreted in such a way that in the bullish market a loss (of 1 unit) reduces the utility stronger than in the bearish market. On the other hand, a gain (of 1 unit) increases the utility less than in the bearish market. The unsettled market shows a behaviour between these extreme markets. Hence, investors fear in a good market situation losses more than in a bad situation and they appreciate profits in a good situation less than in a bad situation.

Finally, we consider the relative risk aversions in the three market regimes. These risk aversions are presented in figure 9, they do not depend on any constants but are completely identified. We see that the risk aversion is smallest in all markets for a small profit that roughly corresponds to the

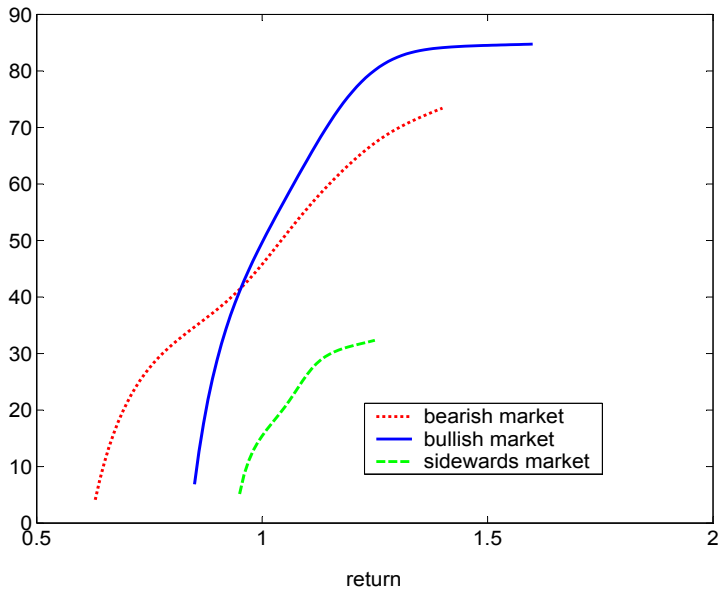


Figure 8: Market utility functions on 24/03/2000 (bullish), 30/07/2002 (bearish) and 30/06/2004 (unsettled or sidewards market).

initial value plus a riskless interest on it. In the unsettled regime the market is risk seeking in a small region around this minimal risk aversion. But then the risk aversion increases quite fast. Hence, the representative agent in this market is willing to take small risks but is sensitive to large losses or profits. In the bullish and bearish regimes the representative agent is less sensitive to large losses or profits than in the unsettled market. In the bearish situation the representative agent is willing to take more risks than in the bullish regime. In the bearish regime the investors are risk seeking in a wider region than in the unsettled regime. In this sense they are more risk seeking in the bearish market. In the bullish market – on the other hand – the investors are never risk seeking so that they are less risk seeking than in the unsettled market.

The estimated utility functions most closely follow the specification proposed by Friedman & Savage (1948). The utility function proposed by Kahneman & Tversky (1979) consists of one concave and one convex segment and is less suitable for describing the observed behaviour, see figure 10. Both utility functions were proposed to account for two opposite types of behaviour with respect to risk attitudes: buying insurance and gambling. Any utility function that is strictly concave fails to describe both risk attitudes. Most notable examples are the quadratic utility function with the linear pricing

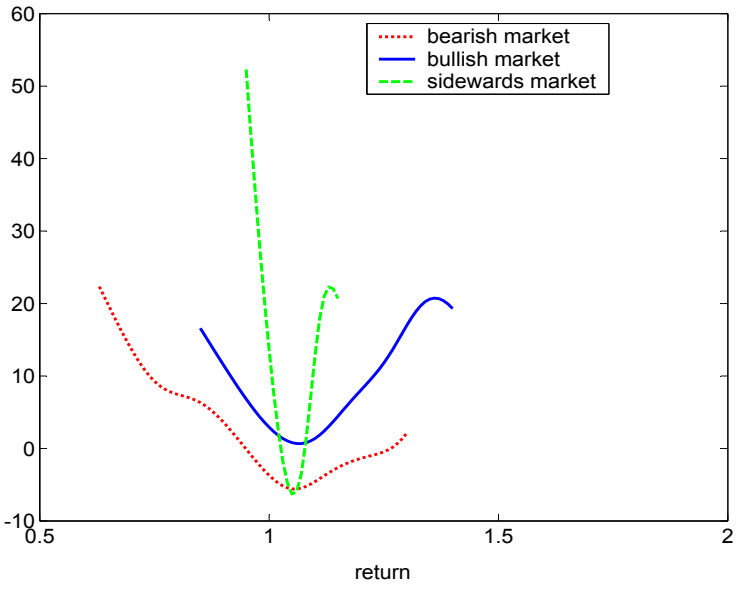


Figure 9: Relative risk aversions on 24/03/2000 (bullish), 30/07/2002 (bearish) and 30/06/2004 (unsettled or sidewards market).

kernel as in the CAPM model and the CRRA utility function. These functions are presented in figure 10. Comparing this theoretical figure with the empirical results in figure 7 we see clearly the shortcoming of the standard specifications of utility functions to capture the characteristic hump of the pricing kernels.

4 Individual investors and their utility functions

In this section, we introduce a type of utility function that has two regions of different risk aversion. Then we describe how individual investors can be aggregated to a representative agent that has the market utility function. Finally, we solve the resulting estimation problem by discretization and estimate the distribution of individual investors.

4.1 Individual Utility Function

We learn from figures 10 and 7 that the market utility differs significantly from the standard specification of utility functions. Moreover, we can observe

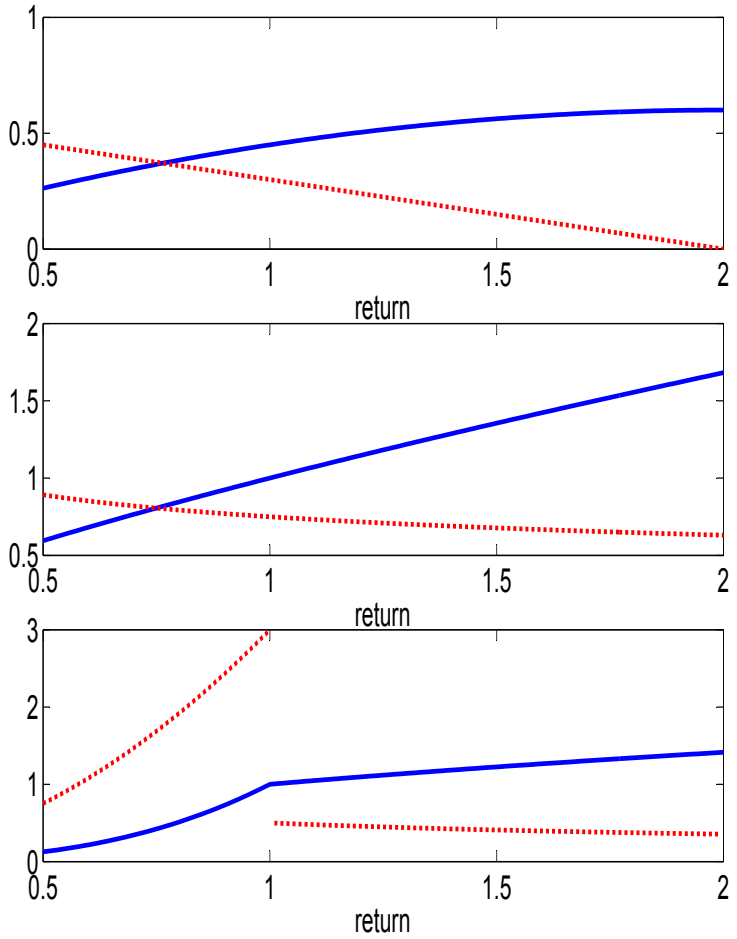


Figure 10: Common utility functions (solid) and their pricing kernels (dotted) (upper: quadratic, middle: power, lower panel: Kahneman and Tversky utility function).

from the estimated utility functions 8 that the loss part and the profit part of the utility functions can be quite well approximated with hyperbolic absolute risk aversion (HARA) functions, $k = 1, 2$:

$$U^{(k)}(R) = a_k(R - c_k)^{\gamma_k} + b_k,$$

where the shift parameter is c_k. These power utility functions become infinitely negative for $R = c_k$ and can be extended by $U^{(k)}(R) = -\infty$ for $R \leq c_k$, i.e. investors will avoid by all means the situation when $R \leq c_k$. The CRRA utility function has $c_k = 0$.

We try to reconstruct the market utility of the representative investor by individual utility functions and hence assume that there are many investors on the market. Investor i will be attributed with a utility function that consists of two HARA functions:

$$U_i(R) = \begin{cases} \max\left\{U(R, \theta_1, c_1); U(R, \theta_2, c_{2,i})\right\}, & \text{if} \quad R > c_1 \\ -\infty, & \text{if} \quad R \leq c_1 \end{cases}$$

where $U(R, \theta, c) = a(R - c)^{\gamma} + b$, $\theta = (a, b, \gamma)^{\top}$, $c_{2,i} > c_1$. If $a_1 = a_2 = 1$, $b_1 = b_2 = 0$ and $c_1 = c_2 = 0$, we get the standard CRRA utility function.

The parameters θ_1 and θ_2 and c_1 are the same for all investors who differ only with the shift parameter c_2. θ_1 and c_1 are estimated from the lower part of the utility market function, where all investors probably agree that the market is "bad". θ_2 is estimated from the upper part of the utility function where all investors agree that the state of the world is "good". The distribution of c_2 uniquely defines the distribution of switching points and is computed in section 4.3. In this way a bear part $U_{bear}(R) = U(R, \theta_1, c_1)$ and a bull part $U_{bull}(R) = U(R, \theta_1, c_2)$ can be estimated by least squares.

The individual utility function can then be denoted conveniently as:

$$U_i(R) = \begin{cases} \max\left\{U_{bear}(R); U_{bull}(R, c_i)\right\}, & \text{if} \quad \mathrm{R} > \mathrm{c}_1; \\ -\infty, & \text{if} \quad \mathrm{R} \leq \mathrm{c}_1. \end{cases} \tag{5}$$

Switching between U_{bear} and U_{bull} happens at the *switching point* z, whereas $U_{bear}(z) = U_{bull}(z, c_i)$. The switching point is uniquely determined by $c_i \equiv c_{2,i}$. The notations *bear* and *bull* have been chosen because U_{bear} is activated when returns are low and U_{bull} when returns are high.

Each investor is characterised by a switching point z. The smoothness of the market utility function is the result of the aggregation of different attitudes. U_{bear} characterizes more cautious attitudes when returns are low and U_{bull} describes the attitudes when the market is booming. Both U_{bear}

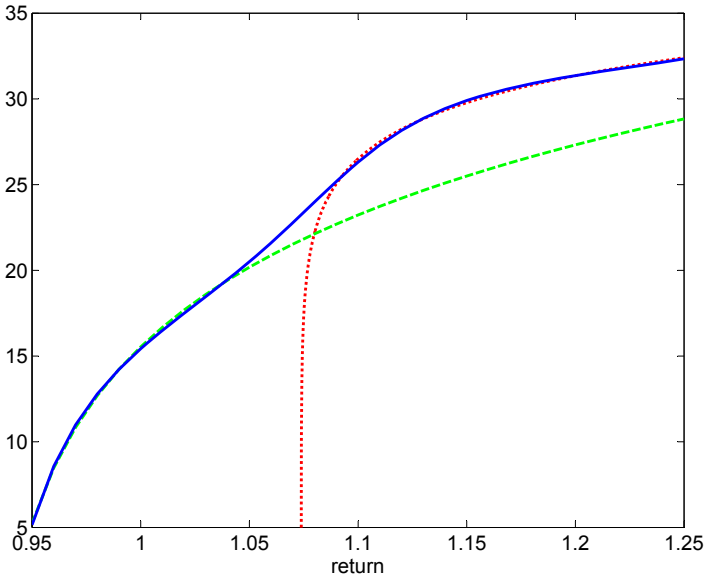


Figure 11: Market utility function (solid) with bearish (dashed) and bullish (dotted) part of an individual utility function 5 estimated in the unsettled market of 30/06/2004.

and U_{bull} are concave. However, due to switching the total utility function can be locally convex.

These utility functions are illustrated in figure 11 that shows the results for the unsettled market. We observe/estimate the market utility function that does not correspond to standard utility approaches because of the convex region. We propose to reconstruct this phenomenon by individual utility functions that consist of a bearish part and a bullish part. While the bearish part is fixed for all investors the bullish part starts at the switching point that characterizes an individual investor. By aggregating investors with different switching points we reconstruct the market utility function. We describe the aggregation in section 4.2 and estimate the distribution of switching points in section 4.3. In this way we explain the special form of the observed market utility functions.

4.2 Market Aggregation Mechanism

We consider the problem of aggregating individual utility functions to a representative market utility function. A simple approach to this problem is to identify the market utility function with an average of the individual utility functions. To this end one needs to specify the *observable* states of the world

in the future by returns R and then find a weighted average of the utility functions for each state. If the importance of the investors is the same, then the weights are equal:

$$U(R) = \frac{1}{N}\sum_{i=1}^{N} U_i(R),$$

where N is the number of investors. The problem that arises in this case is that utility functions of different investors can not be summed up since they are incomparable.

Therefore, we propose an alternative aggregation technique. First we specify the *subjective* states of the world given by utility levels u and then aggregate the outlooks concerning the returns in the future R for each perceived state. For a *subjective* state described with the utility level U, such that

$$u = U_1(R_1) = U_2(R_2) = \ldots = U_N(R_N)$$

the aggregate estimate of the resulting returns is

$$R_A(u) = \frac{1}{N}\sum_{i=1}^{N} U_i^{-1}(u) \tag{6}$$

if all investors have the same market power. The market utility function U_M resulting from this aggregation is given by the inverse R_A^{-1}.

In contrast to the naive approach described at the beginning of this section, this aggregation mechanism is consistent under transformations: if all individual utility functions are changed by the same transformation then the resulting market utility is also given by the transformation of the original aggregated utility. We consider the individual utility functions U_i and the resulting aggregate U_M. In addition, we consider the transformed individual utility functions $U_i^{\phi}(x) = \phi\{U_i(x)\}$ and the corresponding aggregate U_M^{ϕ} where ϕ is a transformation. Then the aggregation is consistent in the sense that $U_M^{\phi} = \phi(U_M)$. This property can be seen from

$$\begin{aligned}(U_M^{\phi})^{-1}(u) &= \frac{1}{N}\sum_{i=1}^{N}(U_i^{\phi})^{-1}(u)\\ &= \frac{1}{N}\sum_{i=1}^{N} U_i^{-1}\{\phi^{-1}(u)\}\\ &= U_M^{-1}\{\phi^{-1}(u)\}\end{aligned}$$

The naive aggregation is not consistent in the above sense as the following example shows: We consider the two individual utility functions $U_1(x) = \sqrt{x}$

and $U_2(x) = \sqrt{x}/2$ under the logarithmic transformation $\phi = \log$. Then the naively aggregated utility is given by $U_M(x) = 3\sqrt{x}/4$. Hence, the transformed aggregated utility is $\phi\{U_M(x)\} = \log(3/4) + \log(x)/2$. But the aggregate of the transformed individual utility functions is

$$\begin{aligned} U_M^{\phi}(x) &= \frac{1}{2}\left\{\log(\sqrt{x}) + \log(\sqrt{x}/2)\right\} \\ &= \frac{1}{2}\log\left(\frac{1}{2}\right) + \log(x)/2. \end{aligned}$$

This implies that $U_M^{\phi} \neq \phi(U_M)$ in general.

This described aggregation approach can be generalized in two ways: If the individual investors have different market power then we use the corresponding weights w_i in the aggregation (6) instead of the uniform weights. As the number of market participants is in general big and unknown it is better to use a continuous density f instead of the discrete distributions given by the weights w_i. These generalizations lead to the following aggregation

$$R_A(u) = \int U^{-1}(\cdot, z)(u) f(z) dz$$

where $U(\cdot, z)$ is the utility function of investor z. We assume in the following that the investors have utility function of the form described in section 4.1. In the next section we estimate the distribution of the investors who are parametrized by z.

4.3 The Estimation of the Distribution of Switching Points

Using the described aggregation procedure, we consider now the problem of replicating the market utility by aggregating individual utility functions. To this end, we choose the parametric utility functions $U(\cdot, z)$ described in 4.1 and try to recover with them the market utility U_M. We do not consider directly the utility functions but minimize instead the distance between the inverse functions:

$$\min_f \| \int U^{-1}(\cdot, z) f(z) dz - U_M^{-1} \|_{L^2(\tilde{P})} \tag{7}$$

where $\tilde{P}$ is image measure of the historical measure P on the returns under the transformation U_M. As the historical measure has the density p the

transformation theorem for densities implies that $\tilde{P}$ has the density

$$\tilde{p}(u) = p\{U_M^{-1}(u)\}/U_M'\{U_M^{-1}(u)\}.$$

With this density the functional to be minimized in problem (7) can be stated as

$$\begin{aligned}
&\int \left(\int U^{-1}(u,z) f(z) dz - U_M^{-1}(u) \right)^2 \; \tilde{p}(u) \; du \\
&= \int \left(\int U^{-1}(u,z) f(z) dz - U_M^{-1}(u) \right)^2 \; p\{U_M^{-1}(u)\}/U_M'\{U_M^{-1}(u)\} \; du \\
&= \int \left(\int U^{-1}(u,z) f(z) dz - U_M^{-1}(u) \right)^2 \; p\{U_M^{-1}(u)\}(U_M^{-1})'(u) \; du
\end{aligned}$$

because the derivative of the inverse is given by $(g^{-1})'(y) = 1/g'\{g^{-1}(y)\}$. Moreover, we can apply integration by substitution to simplify this expression further

$$\begin{aligned}
&\int \left(\int U^{-1}(u,z) f(z) dz - U_M^{-1}(u) \right)^2 \; p\{U_M^{-1}(u)\}(U_M^{-1})'(u) \; du \\
&= \int \left(\int U^{-1}\{U_M(x), z\} f(z) dz - x \right)^2 \; p(x) \; dx.
\end{aligned}$$

For replicating the market utility by minimizing (7) we observe first that we have samples of the historical distribution with density p. Hence, we can replace the outer integral by the empirical expectation and the minimization problem can be restated as

$$\min_f \frac{1}{n} \sum_{i=1}^{n} \left(\int g\{U_M(x_i), z\} f(z) dz - x_i \right)^2$$

where $x_1 \ldots, x_n$ are the samples from the historical distribution and $g = U^{-1}$.

Replacing the density f by a histogram $f(z) = \sum_{j=1}^{J} \theta_j I_{B_j}(z)$ with bins B_j, $h_j = |B_j|$, the problem is transformed into

$$\min_{\theta_j} \frac{1}{n} \sum_{i=1}^{n} \left\{ \sum_{j=1}^{J} \tilde{g}(i,j) \theta_j - x_i \right\}^2$$

where $\tilde{g}(i,j) = \int_{B_j} g\{U_M(x_i), z\} dz$.

Hence, the distribution of switching points can be estimated by solving the quadratic optimization problem

$$\min_{\theta_j} \frac{1}{n} \sum_{i=1}^{n} \left\{ \sum_{j=1}^{J} \tilde{g}(i,j)\theta_j - x_i \right\}^2,$$
$$\text{s.t.} \qquad \theta_j \geq 0,$$
$$\sum_{j=1}^{J} \theta_j h_j = 1.$$

Such quadratic optimization problems are well known and their solutions can be obtained using standard techniques, see e.g. Mehrotra (1992) or Wright (1998).

We present in figures 12–14 the estimated distribution of switching points in the bullish (24/03/2000), bearish (30/07/2002) and unsettled (30/06/2004) markets. The distribution density f was computed for 100 bins but we checked the broad range of binwidths. The width of the distribution varies greatly depending on the regularisation scheme, for example as represented by the number of bins. The location of the distribution maximum, however, remains constant and independent from the computational method.

The maximum and the median of the distribution, i.e. the returns at which half of investors have bearish and bullish attitudes, depend on the year. For example, in the bullish market (Figure 12) the peak of the switching point distribution is located in the area of high returns around $R = 1.07$ for half a year. On the contrary, in the bearish market (Figure 13) the peak of switching points is around $R = 0.93$. This means that when the market is booming, such as in year 1999–2000 prior to the dot-com crash, investors get used to high returns and switch to the bullish attitude only for comparatively high R's. An overall high level of returns serves in this respect as a reference level and investors form their judgements about the market relative to it. Since different investors have different initial wealth, personal habits, attitudes and other factors that our model does not take into account, we have a distribution of switching points. In the bearish market the average level of returns is low and investors switch to bullish attitudes already at much lower R's.

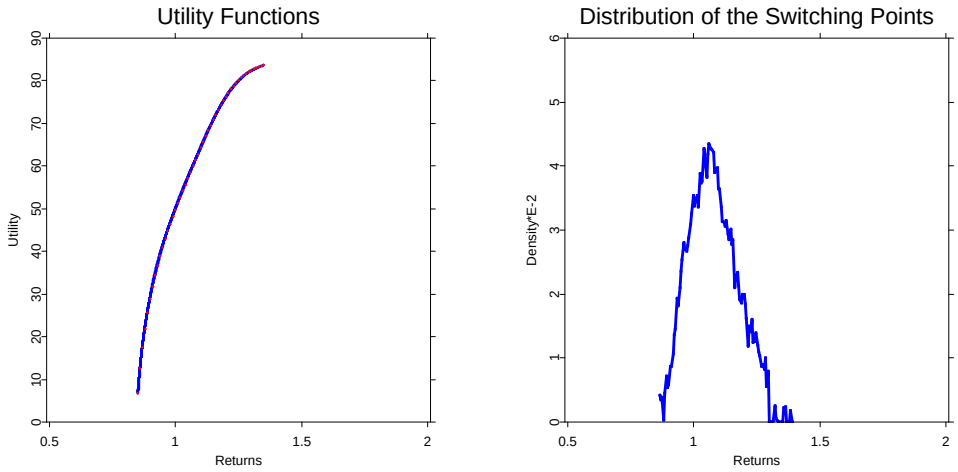


Figure 12: Left panel: the market utility function (red) and the fitted utility function (blue). Right panel: the distribution of the reference points. 24 March 2000, a bullish market.

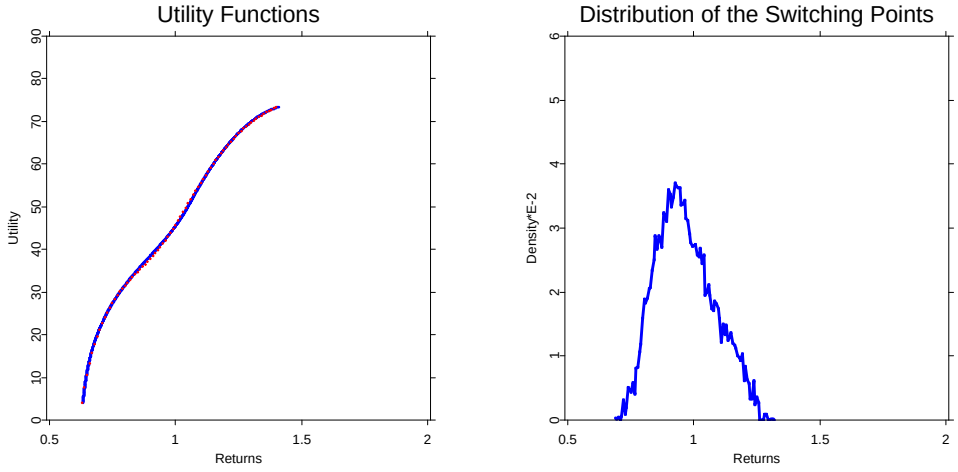


Figure 13: Left panel: the market utility function (red) and the fitted utility function (blue). Right panel: the distribution of the reference points. 30 July 2002, a bearish market.

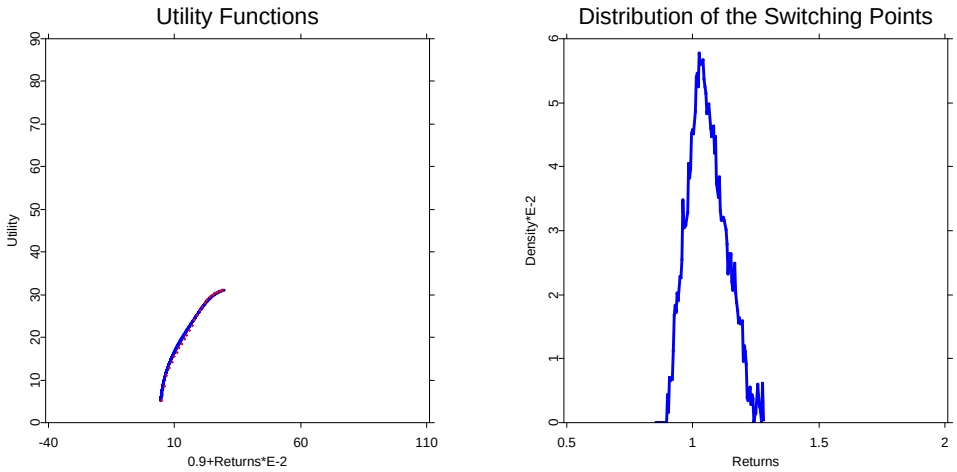


Figure 14: Left panel: the market utility function (red) and the fitted utility function (blue). Right panel: the distribution of the reference points. 30 June 2004, an unsettled market.

5 Conclusion

We have analyzed in this paper empirical pricing kernels in three market regimes using data on the German stock index and options on this index. In the bullish, bearish and unsettled market regime we estimate the pricing kernel and derive the corresponding utility functions and relative risk aversions.

In the unsettled market of June 2004, the market investor is risk seeking in a small region around the riskless return but risk aversion increases fast for high absolute returns. In the bullish market of March 2000, the investor is on the other hand never risk seeking while he becomes more risk seeking in the bearish market of July 2002. Before the stock market crash in 1987 European options did not show the smile and the Black-Scholes model captured the data quite well. Hence, utility functions could be estimated at that times by power utility functions with a constant positive risk aversion. Our analysis shows that this simple structure does not hold anymore and discusses different structures corresponding to different market regimes.

The empirical pricing kernels of all market regimes demonstrate that the corresponding utility functions do not correspond to standard specifications of utility functions including Kahneman and Tversky (1979). The observed utility functions are closest to the general utility functions of Friedman and Savage (1948). We propose a parametric specification of these functions,

estimate it and explain the observed market utility function by aggregating individual utility functions. In this way, we can estimate a distribution of individual investors.

The proposed aggregation mechanism is based on homogeneous investors in the sense that they differ only with switching points. Future research can reveal how nonlinear aggregation procedures could be applied to heterogeneous investors.

6 Acknowledgements

The research work of R. A. Moro was supported by the German Academic Exchange Service (DAAD). K. Detlefsen was supported by Bankhaus Sal. Oppenheim. This research was supported by Deutsche Forschungsgemeinschaft through the SFB 649 "Economic Risk".

References

Ait-Sahalia, Y. and A. Lo, 1998: Nonparametric estimation of state-price densitites implicit in financial asset prices. *Journal of Finance*, **53**(2).

Ait-Sahalia, Y. and A. Lo, 2000: Nonparametric risk-management and implied risk aversion. *Journal of Econometrics*, **94**(9).

Barone-Adesi, G., R. Engle, and L. Mancini, 2004: Garch options in incomplete markets. working paper, University of Lugano.

Bergomi, L., 2005: Smile dynamics 2. *Risk*, **18**(10).

Bernoulli, D., 1956: Exposition of a new theory on the measurement of risk. *Econometrica*, **22**, 23–36.

Billingsley, P., 1995: *Probability and Measure*. Wiley-Interscience.

Black, F. and M. Scholes, 1973: The pricing of options and corporate liabilities. *Journal of Political Economy*, **81**, 637–659.

Breeden, D. and R. Litzenberger, 1978: Prices of state-contingent claims implicit in option prices. *Journal of business*, **51**, 621–651.

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

Chernov, M., 2000: Essays in financial econometrics. Phd thesis, Pennsylvania State University.

Chernov, M., 2003: Empirical reverse engineering of the pricing kernel. *Journal of Econometrics*, **116**, 329–364.

Cizek, P., W. Härdle, and R. Weron, 2005: *Statistical Tools in Finance and Insurance*. Springer, Berlin.

Cochrane, J., 2001: *Asset Pricing*. Princeton University Press.

Cont, R., 2001: Empirical properties of asset returns: stylized facts and statistical issues. 223-349.

Cont, R. and P. Tankov, 2004: Nonparametric calibration of jump-diffusion option pricing models. *Journal of Computational Finance*, **7**(3), 1–49.

Dupire, B., 1994: Pricing with a smile. *Risk*, **7**, 327–343.

Franke, J., W. Härdle, and C. Hafner, 2004: *Statistics of Financial Markets*. Springer Verlag, Berlin.

Friedman, M. and L. P. Savage, 1948: The utility analysis of choices involving risk. *Journal of Political Economy*, **56**, 279–304.

Harrison, M. and S. Pliska, 1981: Martingales and stochastic integrals in the theory of continuous trading. *Stochastic Processes and their Applications*, **11**, 215–260.

Heston, S., 1993: A closed-form solution for options with stochastic volatility with applications to bond and currency options. *Review of Financial Studies*, **6**(2), 327–343.

Heston, S. and S. Nandi, 2000: A clsed form garch option pricing model. *Review of Financial Studies*, **13**, 585–625.

Jackwerth, J., 2000: Recovering risk aversion from option prices and realized returns. *Review of Financial Studies*, **13**(2), 433–451.

Jackwerth, J. and M. Rubinstein, 1996: Recovering probability distributions from option prices. *Journal of Finance*, **51**(5), 1611–1631.

Kahneman, D. and A. Tversky, 1979: Prospect theory: An analysis of decision under risk. *Econometrica*, **47**, 263–291.

Mehrotra, S., 1992: On the implementation of a primal-dual interior point method. *SIAM Journal on Optimization*, **2**(4), 575–601.

Merton, R. C., 1973: An intertemporal capital asset pricing model. *Econometrica*, **41**(5), 867–887.

Rosenberg, J. and R. Engle, 2002: Empirical pricing kernels. *Journal of Financial Economics*, **64**(7), 341–372.

Rubinstein, M., 1994: Implied binomial trees. *Journal of Finance*, **69**, 771–818.

Silverman, B., 1986: *Density Estimation*. Chapman and Hall, London.

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.

von Neumann, J. and O. Morgenstern, 1944: *The Theory of Games and Economic Behavior*. Princeton University Press.

Wright, S., 1998: Primal-dual interior-point methods. *Mathematics of Computation*, **67**(222), 867–870.

De copulis non est disputandum*

Copulae: An Overview

Wolfgang Karl Härdle†, Ostap Okhrin‡

May 27, 2009

Abstract: Normal distribution of the residuals is the traditional assumption in the classical multivariate time series models. Nevertheless it is not very often consistent with the real data. Copulae allows for an extension of the classical time series models to nonelliptically distributed residuals. In this paper we apply different copulae to the calculation of the static and dynamic Value-at-Risk of portfolio returns and Profit-and-Loss function. In our findings copula based multivariate model provide better results than those based on the normal distribution.

Keywords: copula; multivariate distribution; value-at-risk; multivariate dependence.
JEL Classification: C13, C14, C50.

1 Introduction

Understanding the joint distribution of high dimensional data is fundamental in applied statistics. The conventional procedure to model joint distributions is to approximate them with *multivariate normal distributions*.

That implies, however, that the dependence structures is reduced to a fixed type. Predetermining a multivariate normal distribution means that the tails of the distribution are not too heavy, the distribution is symmetric and that the dependence between variables is linear.

Empirical evidence for these assumptions are barely verified and an alternative model is needed, with more flexible dependence structure and arbitrary marginal distributions. These are exactly the characteristics of *copulae*.

Copulae are very useful for modelling and estimating multivariate distributions. The flexibilty of copulae basically follows from *Sklar's Theorem*, which says that each joint

*The financial support from the Deutsche Forschungsgemeinschaft via SFB 649 "Okonomisches Risiko", Humboldt-Universität zu Berlin is gratefully acknowledged.

†CASE - Center for Applied Statistics and Economics, Institute for Statistics and Econometrics of Humboldt-Universität zu Berlin, Spandauer Straße 1, 10178 Berlin, Germany. Email: haerdle@wiwi.hu-berlin.de.

‡CASE - Center for Applied Statistics and Economics, Institute for Statistics and Econometrics of Humboldt-Universität zu Berlin, Spandauer Straße 1, 10178 Berlin, Germany. Email: ostap.okhrin@wiwi.hu-berlin.de

distribution can be "decomposed" into its marginal distributions and a copula C "responsible" for the dependence structure:

$$F(x_1 \ldots, x_d) \;=\; C\{F_1(x_1), \ldots, F_d(x_d)\}.$$

Two important factors for practical applications rely on this theorem:

1. The construction of multivariate distributions may be done in two independent steps: the specification of marginal distributions - not necessarily identical - and the specification of a dependence structure. Copulae "couple together" the marginal distributions into a multivariate distribution with the desired dependence structure.
2. Joint distributions can be separately estimated from a sample of observations: the marginal distributions are estimated first, the dependence structure later.

The copula approach gives us more freedom than the normality assumptions, marginal distributions with asymmetric heavy tails (typical for financial returns) can be combined with different dependence structures, resulting in multivariate distributions (far different from the multivariate normal) that better describe the empirical characteristics of financial returns distribution.

Moreover, copulae allow for dynamical modelling and adaption to portfolios, different copulae with distinct properties can be associated to different portfolios according to their specific dependence structures. Furthermore, copulae may change as time evolves, reflecting the evolution of the dependence between financial assets.

The structure of this paper is as follows. In the next section we give a short review of the copula theory. In the Section 3 we deals with different copula classes used in the calculation. The simulation and estimation techniques are provided in Sections 4 and 5 respectively. The first static problem on the calculation of the Value-at-Risk for the portfolio return has been discussed in Sections 6 and in the beginning of Section 7. Subsections 7.1 and 7.2 deals with the dynamic estimation of the Value-at-Risk for the Profit and Loss function. The paper is finished with summary.

2 Copulae

The description of copulae for measuring and modelling dependence with its main properties is the subject of this section. The term copula goes back to the works of Sklar (1959) were it was first mentioned. There are a lot of different equivalent definitions that could define the copula, but the most general is the following one.

Definition 1 (Copula) *A d-dimensional copula is a d-dimensional distribution with all uniform marginal distributions.*

Note that by considering random variables $X_1, \ldots, X_d$ with univariate distribution functions $F_{X_1}, \ldots, F_{X_d}$ and the random variables $U_i = F_{X_i}(X_i)$, $i = 1, \ldots, d$ uniformly distributed in $[0, 1]$, a copula may be interpreted as *the joint distribution of the marginal distributions.*

Copulae gained popularity through Sklar's (1959) work where the term was first coined. However, many results had already been proved by Hoeffding (1940) and Hoeffding (1941), who could have been the founder of a copula theory, if he had considered the stochastically more intuitive dependency over the unit cube $[0,1]^2$ rather than over $[-1/2,1/2]^2$ as he had done. Copulae allow marginal distributions to be separated from the dependency structure. Sklar's theorem connects copulae with distribution functions such that from the one side every distribution function can be "decomposed" into its marginal distribution and (at least) one copula and from the other side a (unique) copula is obtained from "decoupling" every (continuous) multivariate distribution function from its marginal distributions.

Theorem 1 (Sklar's theorem) *Let F be a multivariate distribution function with margins $F_1,\dots,F_d$, then a copula C exists such that*

$$F(x_1,\dots,x_d) = C\{F_1(x_1),\dots,F_k(x_d)\}, \quad x_1,\dots,x_d \in \overline{\mathbb{R}}.$$

If F_i are continuous for $i=1,\dots,d$ then C is unique. Otherwise C is uniquely determined on $F_1(\overline{\mathbb{R}}) \times \cdots \times F_d(\overline{\mathbb{R}})$.

Conversely, if C is a copula and $F_1,\dots,F_d$ are univariate distribution functions, then the function F defined above is a multivariate distribution function with margins $F_1,\dots,F_d$.

The representation in Sklar's Theorem can be used to construct new multivariate distributions by changing either the copula function or marginal distributions. For an arbitrary continuous multivariate distribution we can determine its copula from the transformation

$$C(u_1,\dots,u_d) = F\{F_1^{-1}(u_1),\dots,F_d^{-1}(u_d)\}, \quad u_1,\dots,u_d \in [0,1], \tag{1}$$

where F_i^{-1} are inverse marginal distribution functions.

Since the copula function is a multivariate distribution with uniform margins, it follows that the copula density can be determined in the usual way

$$c(u_1,\dots,u_d) = \frac{\partial^d C(u_1,\dots,u_d)}{\partial u_1 \dots \partial u_d}, \quad u_1,\dots,u_d \in [0,1],$$

Being armed with Theorem 1 and (**??**) we can write the density function $f(\cdot)$ of the d-variate distribution F in terms of copula as follows

$$f(x_1,\dots,x_d) = c\{F_1(x_1),\dots,F_d(x_d)\} \prod_{i=1}^{d} f_i(x_i), \quad x_1,\dots,x_d \in \overline{\mathbb{R}}.$$

A detailed discussion with proofs and deep mathematical treatment can be found in Joe (1997) and Nelsen (2006). A practical introduction is given in Deutsch and Eller (1999). Embrechts, McNeil and Straumann (1999b) discuss restrictions of the copula technique and their relation to the classical correlation analysis.

3 Copula Classes

Since there are plenty of functions satisfying the assumption of Theorem 1 they should be classified by construction and properties. Here we consider several main classes, like *simplest*, *elliptical*, *Archimedean copulae* and *hierarchical Archimedean copulae*.

3.1 Simplest Copulae

Special cases, like independence and perfect positive or negative dependence can be represented by copulae. If d random variables $X_1, \ldots, X_d$ are stochastically independent from Theorem 1, then the structure of such a relationship is given by the product copula

$$\Pi(u_1, \ldots, u_d) = \prod_{j=1}^{d} u_j. \tag{2}$$

Copulae are bounded, this means that for all $u = (u_1, \ldots, u_d)^\top \in [0,1]^d$:

$$W(u_1, \ldots, u_d) \leq C(u_1, \ldots, u_d) \leq M(u_1, \ldots, u_d)$$

where

$$M(u_1, \ldots, u_d) = \min(u_1, \ldots, u_d)$$

is called the *Fréchet-Hoeffding lower bound* and

$$W(u_1, \ldots, u_d) = \max\left(\sum_{i=1}^{d} u_i - d + 1, 0\right)$$

is the *Fréchet-Hoeffding upper bound*. While M is not a copula for $d > 2$, W is a copula for all d. Both structures represent the perfect negative and perfect positive dependence. From this observation we may conclude that an arbitrary copula C reflects dependence which lies between the perfect negative and positive one.

3.2 Elliptical Copulae

The elliptical copulae are derived from the elliptical distributions using Theorem 1. In the bivariate case one has that a bivariate copula is elliptical if, and only if, it is equal to its associated copula

$$\begin{aligned} C(u_1, u_2, \theta) &= \overline{C}(u_1, u_2, \theta) \\ &= u_1 + u_2 - 1 + C(1 - u_1, 1 - u_2, \theta), \quad u_1, u_2 \in [0,1]. \end{aligned}$$

The most prominent examples of elliptical copulae are Gaussian and t-copula.

Gaussian Copula

The Gaussian copula represents the *dependence structure* of the multivariate normal distribution, that means that *normal* marginal distributions are combined with a Gaussian copula to form multivariate normal distributions. The combination of *non-normal* marginal distributions with a Gaussian copula results in *meta-Gaussian* distributions, i.e., distributions where *only* the dependence structure is Gaussian.

To obtain the Gaussian copula, let $X = (X, \ldots, X_d)^\top \sim \mathrm{N}_d(\mu, \Sigma)$ with $X_j \sim \mathrm{N}(\mu_j, \sigma_j)$ for $j = 1, \ldots, d$. A copula C exists:

$$F(x_1, \ldots, x_d) = C\{F_1(x_1), \ldots, F_d(x_d)\},$$

where F_j is the distribution function of X_j and F the distribution function of X. Let $Y_j = T_j(X_j)$, $T_j(x) = (x - \mu_j)/\sigma_j$. Then $Y_j \sim \mathrm{N}(0, 1)$ and $Y = (Y_1, \ldots, Y_d)^\top \sim \mathrm{N}_d(0, \Psi)$ where Ψ is the correlation matrix associated with Σ. A copula C_Ψ^{Ga}, called *Gaussian copula* exists as follows:

$$F_Y(y_1, \ldots, y_d) = C_\Psi^{Ga}\{\Phi(y_1), \ldots, \Phi(y_d)\}. \tag{3}$$

An explicit expression for the Gaussian copula is obtained by rewriting (3) with $u_j = \Phi(y_j)$:

$$\begin{aligned} C_\Psi^{Ga}(u_1, \ldots, u_d) &= F_Y\{\Phi^{-1}(u_1), \ldots, \Phi^{-1}(u_d)\} \\ &= \int_{-\infty}^{\Phi^{-1}(u_1)} \ldots \int_{-\infty}^{\Phi^{-1}(u_d)} (2\pi)^{-\frac{d}{2}} \mid \Psi \mid^{-\frac{1}{2}} \exp\left(-\frac{1}{2} r^\top \Psi^{-1} r\right) dr_1 \ldots dr_d. \end{aligned}$$

The density of the Gaussian copula is given by

$$c_\Psi^{Ga}(u_1, \ldots, u_d) = \mid \Psi \mid^{-\frac{1}{2}} \exp\left\{-\frac{1}{2} \zeta^\top (\Psi^{-1} - I_d)\zeta\right\}. \tag{4}$$

Student's t-Copula

The t-copula, containing the dependence structure from the multivariate t-distribution, may be obtained in a similar way.

Let $X = (X_1, \ldots, X_d)^\top \sim t_d(\nu, \mu, \Sigma)$ and $Y = (Y_1, \ldots, Y_d)^\top \sim t_d(\nu, 0, \Psi)$ where Ψ is the correlation matrix associated with Σ. The unique copula from Y is the *Student's t-copula* $C_{\nu,\Psi}^t$. For $u = (u_1, \ldots, u_d)^\top \in [0, 1]^d$, the *Student's t-copula* is given by

$$C_{\nu,\Psi}^t(u_1, \ldots, u_d) = t_{\nu,\Psi}\{t_\nu^{-1}(u_1), \ldots, t_\nu^{-1}(u_d)\}$$

where t_ν^{-1} is the quantile function from the univariate t-distribution and $t_{\nu,\Psi}$ the distribution function of Y.

The *density of the t-copula* is given by

$$\begin{aligned} c_{\nu,\Psi}^t(u_1, \ldots, u_d) &= \frac{t_{\nu,\Psi}\{t_\nu^{-1}(u_1), \ldots, t_\nu^{-1}(u_d)\}}{\prod_{j=1}^d t_{\nu,\Psi}\{t_\nu^{-1}(u_j)\}}. \\ &= \mid \Psi \mid^{-\frac{1}{2}} \frac{\Gamma(\frac{\nu+d}{2}) \left\{\Gamma(\frac{\nu}{2})\right\}^{d-1} \left(1 + \frac{1}{\nu}\zeta^\top \Psi^{-1} \zeta\right)^{-\frac{\nu+d}{2}}}{\left\{\Gamma(\frac{\nu+1}{2})\right\}^d \prod_{j=1}^d \left(1 + \frac{1}{\nu}\zeta_j^2\right)^{-\frac{\nu+1}{2}}}. \end{aligned}$$

3.3 Archimedean Copulae

As opposed to elliptical copulae, Archimedean copulae are not constructed using Theorem 1, but are related to Laplace transforms of univariate distribution functions. Let $\mathbb{L}$ denote the class of Laplace transforms which consists of strictly decreasing differentiable functions Joe (1997), i.e.

$$\mathbb{L} = \{\phi : [0;\infty) \to [0,1] \,|\, \phi(0) = 1,\ \phi(\infty) = 0;\ (-1)^j \phi^{(j)} \geq 0;\ j = 1, \ldots, \infty\}.$$

The function $C : [0,1]^d \to [0,1]$ defined as

$$C(u_1, \ldots, u_d) = \phi\{\phi^{-1}(u_1) + \cdots + \phi^{-1}(u_d)\}, \quad u_1, \ldots, u_d \in [0,1]$$

is a d-dimensional Archimedean copula, where $\phi \in \mathbb{L}$ and is called the *generator of the copula*. It is straightforward to show that $C(u_1, \ldots, u_d)$ satisfies the conditions of Definition 1.

Some d-dimensional Archimedean copulae are presented below.

Frank (1979) copula, $0 \leq \theta < \infty$.

The first popular Archimedean copula is the so called Frank copula, which is the only elliptical Archimedean copula. Its generator and copula functions are

$$\begin{aligned} \phi(x, \theta) &= \theta^{-1} \log\{1 - (1 - e^{-\theta})e^{-x}\}, \quad 0 \leq \theta < \infty,\ x \in [0, \infty). \\ C_\theta(u_1, \ldots, u_d) &= -\frac{1}{\theta} \log \left[1 + \frac{\prod_{j=1}^{d} \{\exp(-\theta u_j) - 1\}}{\{\exp(-\theta) - 1\}^{d-1}} \right]. \end{aligned}$$

The dependence becomes maximal when θ tends to infinity and independence is achieved when $\theta = 0$.

Gumbel (1960) copula, $1 \leq \theta < \infty$.

The Gumbel copula is frequently used in financial applications. Its generator and copula functions are

$$\begin{aligned} \phi(x, \theta) &= \exp\{-x^{1/\theta}\}, \quad 1 \leq \theta < \infty,\ x \in [0, \infty) \\ C_\theta(u_1, \ldots, u_d) &= \exp\left[-\left\{ \sum_{j=1}^{d} (-\log u_j)^\theta \right\}^{\theta^{-1}} \right]. \end{aligned}$$

Consider a bivariate distribution based on the Gumbel copula with univariate extreme value marginal distributions. Genest and Rivest (1989) showed that this distribution is

the only bivariate extreme value distribution based on an Archimedean copula. Moreover, all distributions based on Archimedean copulae belong to its domain of attraction under common regularity conditions. In contrary to the elliptical copulae, the Gumbel copula leads to asymmetric contour diagrams. The Gumbel copula shows stronger linkage between positive values, however, it also shows more variability and more mass in the negative tail.

For $\theta > 1$ this copula allows for the generation of dependence in the upper tail. For $\theta \to 1$, the Gumbel copula reduces to the product copula and for $\theta \to \infty$ we obtain the Fréchet-Hoeffding upper bound.

Clayton (1978) copula, $-1 \leq \theta < \infty,\ \theta \neq 0$.

The Clayton copula which, in contrast to the Gumbel copula, has more mass on the lower tail, and less on the upper. The generator and copula function are

$$\begin{aligned} \phi(x,\theta) &= (\theta x+1)^{-\frac{1}{\theta}}, \quad -1 \leq \theta < \infty,\ \theta \neq 0,\ x \in [0,\infty), \\ C_\theta(u_1,\ldots,u_d) &= \left\{ \left(\sum_{j=1}^{d} u_j^{-\theta} \right) - d + 1 \right\}^{-\theta^{-1}}. \end{aligned}$$

The Clayton copula is one of few copulae that has a simple explicit form of density for any dimension

$$c_\theta(u_1,\ldots,u_d) = \prod_{j=1}^{d} \{1+(j-1)\theta\} u_j^{-(\theta+1)} \left(\sum_{j=1}^{d} u_j^{-\theta} - d + 1 \right)^{-(\theta^{-1}+d)}.$$

As the parameter θ tends to infinity, dependence becomes maximal and as θ tends to zero, we have independence. As $\theta \to -1$, the distribution tends to the lower Fréchet bound.

3.4 Hierarchical Archimedean Copulae

A recently developed flexible method is provided by hierarchical Archimedean copulae (HAC). The special, so called fully nested case of the copula function is:

$$\begin{aligned} C(u_1,\ldots,u_d) &= \phi_{d-1}\big\{\phi_{d-1}^{-1} \circ \phi_{d-2}\big(\ldots [\phi_2^{-1} \circ \phi_1\{\phi_1^{-1}(u_1) + \phi_1^{-1}(u_2)\} \\ &+ \phi_2^{-1}(u_3)] + \cdots + \phi_{d-2}^{-1}(u_{d-1})\big) + \phi_{d-1}^{-1}(u_d)\big\} \\ &= \phi_{d-1}[\phi_{d-1}^{-1} \circ C(\{\phi_1,\ldots,\phi_{d-2}\})(u_1,\ldots,u_{d-1}) + \phi_{d-1}^{-1}(u_d)] \end{aligned}$$

for $\phi_{d-i}^{-1} \circ \phi_{d-j} \in \mathbb{L}^*$, $i < j$, where

$$\begin{aligned} \mathbb{L}^* = \{&\omega : [0;\infty) \to [0,\infty) \,|\, \omega(0) = 0, \\ &\omega(\infty) = \infty;\ (-1)^{j-1}\omega^{(j)} \geq 0;\ j = 1,\ldots,\infty\}. \end{aligned}$$

In contrast to the Archimedean copula, the HAC defines the whole dependency structure in a recursive way. At the lowest level the dependency between the first two variables is

modelled by a copula function with the generator ϕ_1, i.e. $z_1 = C(u_1, u_2) = \phi_1\{\phi_1^{-1}(u_1) + \phi_1^{-1}(u_2)\}$. At the second level an another copula function is used to model the dependency between z_1 and u_3, etc. Note that the generators ϕ_i can come from the same family and they differ only through the parameter or, to introduce more flexibility, they come from different generator families. As an alternative to the fully nested model, we can consider copula functions, with arbitrary chosen combinations at each copula level. Okhrin, Okhrin and Schmid (2009a) provide several methodologies in determining the structure of the HAC from the data. The case of $d = 3$ which we use further in applications is quite a simple one. If τ_{12}, τ_{13} and τ_{23} are Kendall's τ, pairwise rank correlation coefficients, we join together those X_i and X_j such that $\max_{i,j\in\{1,2,3\},\ i\neq j} = \tau_{ij}$. Next we introduce $z = \widehat{C}\{\hat{F}_i(X_i), \hat{F}_i(X_j)\}$. Estimation techniques will be considered later. Variable X_{i^*}, $i^* \in \{1, 2, 3\}/\{i, j\}$ is joined afterwards with the z.

Whelan (2004) provides tools for generating samples from Archimedean copulae, Savu and Trede (2006) derived the density of such copulae and Joe (1997) proves their positive quadrant dependence (see Theorem 4.4). Okhrin et al. (2009a) and Okhrin, Okhrin and Schmid (2009b) considered methods for determining the optimal structure of the HAC, provided asymptotic theory for the estimated parameters and derive theoretical properties of this copula family.

4 Monte Carlo Simulation

The Monte-Carlo simulation is often a single reliable solution to many financial problems. Within the simulation study the random variables are generated from some prescribed distributions. There are numerous methods of simulating from copula-based distributions, see Frees and Valdez (1998), Whelan (2004), Marshall and Olkin (1988),McNeil (2008), Embrechts, McNeil and Straumann (1999), Frey and McNeil (2003), Devroye (1986), etc. Here we focus on two of them, on the conditional inversion method and on the method proposed by Marshall and Olkin (1988) for Archimedean copulae with generalizations to hierarchical Archimedean copulae by McNeil (2008).

4.1 Conditional Inverse Method

The simulation from d pseudo random variables with joint distribution defined by a copula C and d marginal distributions F_j, $j = 1, \ldots, d$, may follow different techniques.

Defining the copula j-dimensional marginal distribution C_j for $j = 2, \ldots, d-1$ as $C_j(u_1, \ldots, u_j) = C(u_1, \ldots, u_j, 1, \ldots, 1)$ and the derivative of C_j with respect to the first $j-1$ arguments as

$$c_{j-1}^j(u_1, \ldots, u_j) = \frac{\partial^{j-1} C_j(u_1, \ldots, u_j)}{\partial u_1, \ldots, \partial u_{j-1}}$$

the probability $P(U_j \leq u_j, U_1 = u_1, \ldots, U_{j-1} = u_{j-1})$ can be written as

$$\lim_{\Delta u_1, \ldots, \Delta u_{j-1} \to 0} \frac{C_j(u_1 + \Delta u_1, \ldots, u_{j-1} + \Delta u_{j-1}, u_j) - C_j(u_1, \ldots, u_j)}{\Delta u_1, \ldots, \Delta u_{j-1}}$$
$$= c_{j-1}^j(u_1, \ldots, u_j).$$

Thus, the conditional distribution $\Lambda(u_j)$ (given fixed $u_1, \ldots, u_{j-1}$) is a function of the ratio of derivatives:

$$\begin{aligned} \Lambda(u_j) &= P(U_j \leq u_j \mid U_1 = u_1, \ldots, U_{j-1} = u_{j-1}) \\ &= \frac{c_{j-1}^{j}(u_1, \ldots, u_j)}{c_{j-1}^{j-1}(u_1, \ldots, u_{j-1})}. \end{aligned}$$

The generation of d pseudo random numbers with given marginal distributions F_j, $j = 1, \ldots, d$ and dependence structure given by the copula C follows the steps:

1. generate iid $v_1, \ldots, v_d \sim U[0, 1]$.
2. for $j = 1, \ldots, d$ calculate $u_j = \Lambda^{-1}(v_j)$.
3. set $x_j = F_j^{-1}(u_j)$.

4.2 Marshal-Olkin Method

The Marshal-Olkin method is developed for the simulations only from Archimedean copulae. The idea this approach is based on the fact that the Archimedean copulae are derived from Laplace transforms. Let M be a univariate cdf of a positive random variable (so that $M(0) = 0$) and ϕ be the Laplace transform of M, i.e.

$$\phi(s) = \int_0^\infty \exp\{-sw\}\, dM(w),\ s \geq 0.$$

For any univariate distribution function F, a unique distribution G exists:

$$F(x) = \int_0^\infty G^\alpha(x)\, dM(\alpha) = \phi\{-\log G(x)\}.$$

Considering d different univariate distributions $F_1, \ldots, F_d$, we obtain

$$C(u_1, \ldots, u_d) = \int_0^\infty \prod_{i=1}^{d} G_i^\alpha\, dM(\alpha) = \phi\left[\sum_{i=1}^{d} \phi^{-1}\{F_i(u_i)\}\right]$$

which is a multivariate distribution function. By replacing the product of univariate distributions G_i for $i = 1, \ldots, d$ with an arbitrary copula function R we get:

$$C(u_1, \ldots, u_d) = \int_0^\infty \ldots \int_0^\infty R(G_1^\alpha, \ldots, G_d^\alpha)\, dM(\alpha).$$

Note that for the classical Archimedean copula R is equal to a product copula.

One proceeds with the following three steps to make a draw from a distribution described by an Archimedean copula:

1. generate an observation u from M;
2. generate an observations $(v_1, \ldots, v_d)$ from R;

3. the generated vector is computed by $x_j = G_j^{-1}(v_j^{1/u})$.

This method works faster than the conditional inverse technique. The drawback is that the distribution M can be determined explicitly only for a few generator functions ϕ like, for example for the Frank, Gumbel and Clayton families. The same problem arises in the case of hierarchical copulae, where $\phi_i \circ \phi_{i+1}^{-1}$ should satisfy the properties of generator functions.

5 Copula Estimation

The estimation of a copula based multivariate distribution involves both the estimation of the copula parameters θ and the estimation of the margins F_j, $j = 1, \ldots, d$, however all the parameters from the copula and from the margins could be also estimated in one step. The properties and goodness of the estimator of θ heavily depend on the estimators of F_j, $j = 1, \ldots, d$. We distinguish between a parametric and a nonparametric specification of the margins. If we are interested only in the dependency structure, the estimator of $\{\delta_1, \ldots, \delta_d, \theta\}$ should be independent of any parametric models for the margins. In practical applications, however, we are interested in a complete distribution model and, therefore, parametric models for margins are preferred.

For nonparametrically estimated margins, one may show the consistency and asymptotic normality of maximum-likelihood (ML) estimators and derive the moments of the asymptotic distribution. The ML estimation can be performed simultaneously for the parameters of the margins and of the copula function. Alternatively, a two-stage procedure can be applied, where we estimate the parameters of margins at the first stage and the copula parameters at the second stage.

Let X be a d-dimensional random variable with parametric univariate marginal distributions $F_j(x_j; \delta_j)$, $j = 1, \ldots, d$. Further let a copula belong to a parametric family $\mathcal{C} = \{C_\theta, \theta \in \Theta\}$. The distribution of X can be expressed as

$$F(x_1, \ldots, x_d) = C\{F_1(x_1; \delta_1), \ldots, F_d(x_d; \delta_d); \theta\}$$

and its density as

$$f(x_1, \ldots, x_d; \delta_1, \ldots, \delta_d, \theta) = c\{F_1(x_1; \delta_1), \ldots, F_d(x_d; \delta_d); \theta\} \prod_{j=1}^{d} f_j(x_j; \delta_j)$$

where $c(\cdot)$ is the copula density (**??**). For a sample of observations $\{x_t\}_{t=1}^T$, $x_t = (x_{1,t}, \ldots, x_{d,t})^\top$ and a vector of parameters $\alpha = (\delta_1, \ldots, \delta_d, \theta)^\top \in \mathbb{R}^{d+1}$ the likelihood function is given by

$$L(\alpha; x_1, \ldots, x_T) = \prod_{t=1}^{T} f(x_{1,t}, \ldots, x_{d,t}; \delta_1, \ldots, \delta_d, \theta)$$

and the log-likelihood function by

$$\begin{aligned} \ell(\alpha; x_1, \ldots, x_T) &= \sum_{t=1}^{T} \log c\{F_1(x_{1,t}; \delta_1), \ldots, F_d(x_{d,t}; \delta_d); \theta\} \\ &+ \sum_{t=1}^{T} \sum_{j=1}^{d} \log f_j(x_{j,t}; \delta_j). \end{aligned}$$

The vector of parameters $\alpha = (\delta_1, \ldots, \delta_d, \theta)^\top$ contains d parameters δ_j from the marginals and the copula parameter θ. All these parameters can be estimated *in one step*. For practical applications, however, a two step estimation procedure is more efficient.

5.1 FML – Full Maximum Likelihood Estimation

In the Maximum Likelihood estimation method (also called *full maximum likelihood*), the vector of parameters α is estimated in one single step through

$$\tilde{\alpha}_{FML} = \arg\max_{\alpha} \ell(\alpha)$$

The estimates $\tilde{\alpha}_{FML} = (\tilde{\delta}_1, \ldots, \tilde{\delta}_d, \tilde{\theta})^\top$ solve

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

Following the standard theory on ML estimation it is efficient and asymptotically normal. However, it is often computationally demanding to solve the system simultaneously.

5.2 IFM – Inference for Margins

In the IFM (*inference for margins*) method, the parameters δ_j from the marginal distributions are estimated in the first step and used to estimate the dependece parameter θ in the second step:

1. for $j = 1, \ldots, d$ the log-likelihood function for each of the marginal distributions are

   $$\ell_j(\delta_j) = \sum_{t=1}^{T} \log f_j(x_{j,t}; \delta_j)$$

   and the estimated parameters

   $$\hat{\delta}_j = \arg\max_{\delta} \ell_j(\delta_j)$$

2. the *pseudo log-likelihood* function

   $$\ell(\theta, \hat{\delta}_1, \ldots, \hat{\delta}_d) = \sum_{t=1}^{T} \log c\{F_1(x_{1,t}; \hat{\delta}_1), \ldots, F_d(x_{d,t}; \hat{\delta}_d); \theta\}$$

   is maximised over θ to get the dependence parameter estimate $\hat{\theta}$.

The estimates $\hat{\alpha}_{IFM} = (\hat{\delta}_1, \ldots, \hat{\delta}_d, \hat{\theta})^\top$ solve

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

Detailed discussion on this method could be found in Joe and Xu (1996) Note, that this procedure does not lead to efficient estimators, however, as argued by Joe (1997) the loss in the efficiency is modest. The advantage of the inference for margins procedure lies in the dramatic reduction of the numerical complexity. Detailed discussion on the inference for margins procedure can be found in Joe and Xu (1996). Note, that this method does not lead to efficient estimators, however, as argued by Joe (1997) the loss in the efficiency is modest.

5.3 CML – Canonical Maximum Likelihood

In the CML (*canonical maximum likelihood*) method, the univariate marginal distributions are estimated through the edf $\hat{F}$. The asymptotic properties of the multistage estimators of θ do not depend explicitly on the type of the nonparametric estimator, but on its convergence properties. For $j = 1, \ldots, d$

$$\hat{F}_j(x) = \frac{1}{T+1} \sum_{t=1}^{T} \mathbf{I}(x_{j,t} \leq x).$$

The *pseudo log-likelihood* function is

$$\ell(\theta) = \sum_{t=1}^{T} \log c\{\hat{F}_1(x_{1,t}), \ldots, \hat{F}_d(x_{d,t}); \theta\}$$

and the copula parameter estimator $\hat{\theta}_{CML}$ is given by

$$\hat{\theta}_{CML} = \arg\max_{\theta} \ell(\theta).$$

Notice that the first step of the IMF and CML methods estimates the marginal distributions. After marginals are estimated, a *pseudo sample* $\{u_t\}$ of observations transformed in the unit d-cube is obtained and used in the *copula* estimation. As in the IFM, the semiparametric estimator $\hat{\theta}$ is asymptotically normal under suitable regularity conditions.

6 Asset Allocation

We illustrate the extension of the classical asset allocation problem to copula-based models. We consider an investor with a CRRA utility function $U(x) = (1-\gamma)^{-1}x^{1-\gamma}$ willing to allocate his wealth to d risky assets. We denote the d-dimensional vector of d asset prices by $S_t = (S_{1,t}, \ldots, S_{d,t})^\top$ and their continuously compounded asset returns at time $t+1$ by $X_{t+1} = (X_{1,t+1}, \ldots, X_{d,t+1})^\top$ where $X_{t+1} = \log S_{t+1} - \log S_t$. The vector of portfolio weights by $w = (w_1, \ldots, w_d)^\top$. Let F_{t+1} be the d-dimensional distribution function of X_{t+1} with the mean μ_{t+1} and covariance matrix Σ_{t+1}. The aim is to forecast F_{t+1} for the time period $t+1$ using the data up to time t. The estimator is denoted by $\hat{F}_{t+1}$ with the mean $\hat{\mu}_{t+1}$, the covariance matrix $\hat{\Sigma}_{t+1}$ and the density $\hat{f}_{t+1}$. The objective of the investor is to maximise the expected utility at the time point $t+1$. This leads to the optimisation problem

$$\max_{w \in \mathcal{W}} \mathsf{E}_{\hat{F}_{t+1}} U(1 + w^\top X_{t+1}). \tag{5}$$

In the case of no short sales constraint we set $\mathcal{W} = \{w \in [0,1]^d : w^\top 1 = 1\}$ else we set $\mathcal{W} = \{w \in \mathbb{R}^d : w^\top 1 = 1\}$. The conditional expectation in (5) implies that we integrate the utility with respect to the forecasted distribution $\hat{F}_{t+1}$. This reduces the problem (5) to the problem

$$\max_{w \in \mathcal{W}} \int \cdots \int U(1 + w^\top X_{t+1}) \hat{f}_{t+1}(X_{t+1}) dX_{t+1}.$$

There are several alternative parametric approaches to modelling F_{t+1}. Let $\Sigma_{d,t+1}$ denote the diagonal matrix containing only the main diagonal of Σ_{t+1}. Then $\Sigma_{t+1} = \Sigma_{d,t+1}^{1/2} R_{t+1} \Sigma_{d,t+1}^{1/2}$, where R_{t+1} denotes the correlation matrix. A standard approach is to define the model of the asset returns in the form

$$\Sigma_{d,t}^{-1/2}(X_t - \mu_t) \sim \mathrm{N}_d(0, R_t), \tag{6}$$

where the conditional moments μ_t and Σ_t are modelled by a GARCH type process.

To introduce a copula-based distribution into the asset allocation we deviate from the normality assumption and assume that $F = C(F_1, \ldots, F_d)$. Thus (7) is replaced by:

$$\Sigma_{d,t}^{-1/2}(X_t - \mu_t) \sim C(F_1, \ldots, F_d) \tag{7}$$

with some given functional forms of the copula and the marginal distributions. Similarly as above, the parameters of the conditional moments of the copula and of the marginal distributions are estimated using the ML method.

In Patton (2004) the investor allocates his wealth between small cap and large cap stocks (i.e. $d = 2$). The conditional mean is defined as linear function of the lagged asset returns and additional explanatory variables. The conditional variance is stated in the TARCH(1,1) form. The rotated Gumbel copula with skewed t margins are used to construct the bivariate distribution of the residuals. This model reveals the highest likelihood function and the lowest AIC and BIC criterion. It is concluded that unconstrained portfolios derived from the normality assumption performed worse in 9 of 10 different trading strategies compared to the Gumbel model.

7 Value-at-Risk of the Portfolio Returns

If the return of the stock i at time point t is denoted as X_{it} then the portfolio value V at time t is defined recursively as

$$V_t = V_{t-1}\left(1 + \sum_{i=1}^{d} w_i X_{it}\right),$$

where w_i for $i = 1, \ldots, d$ are the corresponding portfolio weights. Ruled with this notation the portfolio return is then given by

$$R_{tp} = \frac{V_t}{V_{t-1}} - 1 = \sum_{i=1}^{d} X_{it} w_i.$$

In our study we consider the case of equally weighted portfolio, i.e. $w_i = \frac{1}{d}$ for $i = 1, \ldots, d$. The portfolio return is the random variable and its distribution strongly depends on the underlying distribution of the indices.

The distribution function of R_p, dropping the time index, is given by

$$F_{R_p}(\xi) = \mathrm{P}(R_p \leq \xi). \tag{8}$$

One of the main advantages of copulae is the fact that they allow flexible modelling of the tail behaviour of multivariate distributions. Since the tail behaviour explains the

simultaneous outliers of asset returns, it is of special interest in risk management. The *Value-at-Risk* of a portfolio at level α is defined as the lower α-quantile of the distribution of the portfolio return, i.e.

$$\mathrm{VaR}(\alpha) = F_{R_p}^{-1}(\alpha). \tag{9}$$

The VaR is a reasonable measure of risk if we assume that the returns are elliptically distributed. Moreover, the assumption of ellipticity implies that minimising the variance in the Markowitz problem also minimises the VaR, the expected shortfall and any other coherent measure of risk. However, this statement is false in the non-elliptical case. Moreover, regarding the effect of diversification the variance is the smallest (highest) for perfect negative (positive) correlation of the assets. This also holds for the VaR in the elliptical case, however, not for the non-elliptical distributions. This implies that for copula based distribution the VaR should be used with caution and its computation should be awarded more attention. Detailed description of the VaR estimation procedure at prescribed level α can be found in Giacomini and Härdle (2005).

Our aim is to determine such ξ that $\mathrm{P}(R_p \leq \xi) = \alpha$. Note that

$$R_p = w^\top X = \sum_{i=1}^{d} w_i X_i = \sum_{i=1}^{d} w_i F_i^{-1}(u_i),$$

where F_i denotes the marginal distributions of individual asset returns, $u_i = F_i(X_i) \sim U[0,1]$ for all $i = 1, \ldots, d$ and $u_1, \ldots, u_d \sim C$. The copula C defines the dependency structure between the asset returns. This implies that

$$F_{R_p}(\xi) = \mathrm{P}(R_p \leq \xi) = \int_{\mathcal{U}} c(u_1, \ldots, u_d) du_1 \ldots du_d, \tag{10}$$

with

$$\mathcal{U} = \{[0,1]^{d-1} \times [0, u_d(\xi)]\}, \quad u_d(\xi) \;=\; F_d\Big\{\xi/w_d - \sum_{i=1}^{d-1} w_i F_i^{-1}(u_i)/w_d\Big\}. \tag{11}$$

For fixed α, the VaR is determined by solving (10) numerically for ξ. Direct multidimensional numerical integration is a tedious task which can be substantially simplified by using the Monte-Carlo integration. For this purpose we have to generate random samples from C using the methods described in Section 4.

In the empirical study we consider four countries Canada, Germany, U.S. and U.K. from the MCSI index and eleven models of the joint multivariate distribution of indices, which include t-copula, Gaussian copula, simple exchangeable Archimedean copula, binary HAC and aggregated binary HAC, with normally and t-distributed margins. As a benchmark we use the empirical VaR, based purely on the real data.

In the cases where margins are t-distributed, we consider t-distribution with three degrees of freedom, while estimated t-distributions for this data are $t_{3.163}$, $t_{3.420}$, $t_{3.023}$, $t_{2.879}$. Multivariate t-copula in this case has eight degrees of freedom. Let us consider the simulation procedure, where on the first stage we estimate the covariance matrix $\widehat{\Sigma} = \{\widehat{\Sigma}_{ij}\}_{i,j=1,\ldots,d}$, mean vector $\widehat{\mu} = \{\widehat{\mu}_i\}_{i=1,\ldots,d}$ from the real data set and assume, or estimate, the marginal distributions $\widehat{F}_i(\cdot)$ (in our case they are normally or t-distributed), for $i =$

$1, \ldots, d$. Next we show how to sample $u_1, \ldots, u_d \in \mathcal{U}$ from (11). First we simulate the vector u of a dimension $d-1$

$$u_1, \ldots, u_{d-1} \sim U(0,1).$$

Based on u we consider $x = \{x_i\}_{i=1,\ldots,d-1}$ which for normal margins is equal to

$$x_i = \Phi^{-1}(u_i)\sqrt{\widehat{\Sigma}_{ii}} + \widehat{\mu}_i, \quad i = 1, \ldots, d-1,$$

and for t margins is

$$x_i = t^{-1}(u_i)\sqrt{\frac{\nu_i - 2}{\nu_i}\widehat{\Sigma}_{ii}} + \widehat{\mu}_i, \quad i = 1, \ldots, d-1,$$

where ν_i, $i = 1, \ldots, d$ are degrees of freedom for marginal distributions. This transformation returns a normally or t-distributed vector x with the same parameters as the real data set.

Theoretically, in further steps we have to find bounds for the last stock (or index) to gain the portfolio ξ which is the α quantile. Thus, we separate our maximally reachable portfolio return ξ into two parts

$$\xi = \sum_{i=1}^{d-1} \frac{1}{d} X_i + \frac{1}{d} X_d,$$

then the return of the last index given the return of the portfolio is

$$X_d = d\xi - \sum_{i=1}^{d-1} X_i,$$

where the upper bound for our last value in vector u is then

$$u_d^* = \widehat{F}_d \left(d\xi - \sum_{i=1}^{d-1} x_i \right).$$

Value u_d^* is uniformly distributed on $[0,1]$ and we simulate the last element of the vector $u_d \sim U(0, u_d^*)$.

As mentioned above, the goal is to compute (10) which for this setting is

$$F_{R_p}(\xi) = \int\limits_{[0,1]^{d-1} \times [0, u_d^*]} \cdots \int c(u_1, \ldots, u_d) du_1 \ldots du_d.$$

Then by solving $F_{R_p}(\xi) = \alpha$ we find $R_\alpha = \mathrm{VaR}(\alpha)$. In our study we solve the equations numerically using the golden section method. The integration is performed using the Monte-Carlo technique

$$\widehat{P(R_p \leq \xi)} = \frac{1}{n_s} \sum_{i=1}^{n_s} c(u_{1i}, \ldots, u_{di})$$

where n_s is equal to 10^8, α is set to be 1% and the values $u_{1i}, \ldots, u_{di}$ for $i = 1, \ldots, n_s$ are simulated using the method described above. The precision of R is set at 0.00015.

Table 1: VaR for the 4-dimensional data set

	N	t_3
N	-0.0194	-0.0210
t_8	**-0.0199**	**-0.0213**
AC	*-0.0174*	*-0.0154*
HAC_{binary}	-0.0187	-0.0194
$HAC_{binary\ aggr.}$	-0.0188	-0.0194
Empirical	-0.0235	

The final results for all methods are given in Table 1. In the left-hand column we provide the models with normal margins and in the right-hand column with t margins. From top to bottom we have five different copula functions like Gaussian, t, simple Archimedean copula, binary HAC and binary aggregated HAC. The empirical VaR which is at the bottom of the table is derived from the empirical quantile. Bold fonts in the table emphasize those results which are closest in absolute value to the empirical one in each column, and italic fonts the worst cases in absolute value.

As can be seen from Table 1, the results which are the best in absolute value are those returned by the model with t-copula and t margins. The model based on the simple Archimedean copula is the worst one. This is quite natural, since this copula needs exchangeability between variables, which is not observable here (see previous section). HAC with binary as well as aggregated binary structures, unfortunately, give us results that are not much worse compared to t-copula and Gaussian copula. For VaR(0.01) the t-copula with t margins provided the best result.

7.1 VaR of the P&L

This sub-section introduces the main assumptions and steps necessary to estimate the VaR from a Profit and Loss of a linear portfolio using copulae. Static and time-varying methods and their VaR performance evaluation through backtesting are described below.

In this section w is the portfolio, which is represented by the number of assets for a specified stock in the portfolio, $w = \{w_1, \ldots, w_d\},\ w_i \in \mathbb{Z}$. The value V_t of the portfolio w is given non-recursively by

$$V_t = \sum_{j=1}^{d} w_j S_{j,t} \tag{12}$$

and the random variable

$$\begin{aligned} L_{t+1} &= (V_{t+1} - V_t) \\ &= \sum_{j=1}^{d} w_j S_{j,t} \left\{ \exp(X_{j,t+1}) - 1 \right\}. \end{aligned}$$

also called *profit and loss (P&L) function*, expresses the absolute change in the portfolio value in one period.

Similarly to the previous case, the distribution function of L, dropping the time index, is given by

$$F_L(x) = \mathrm{P}(L \leq x). \tag{13}$$

As usual the *Value-at-Risk* at level α from a portfolio w is defined as the α-quantile from F_L:

$$\mathrm{VaR}(\alpha) = F_L^{-1}(\alpha). \tag{14}$$

It follows from (13) that F_L depends on the d-dimensional distribution of log-returns F_X. In general, the *loss distribution* F_L depends on a random process representing the *risk factors* influencing the P&L from a portfolio. In the present case log-returns are a suitable risk factor choice. Thus, modelling their distribution is essential to obtain the quantiles from F_L.

Contrary to the previous section, here log-returns are assumed to be time-dependent, thus a log-returns process $\{X_t\}$ can be modelled as

$$X_{j,t} = \mu_{j,t} + \sigma_{j,t}\varepsilon_{j,t}$$

where $\varepsilon_t = (\varepsilon_{1,t}, \ldots, \varepsilon_{d,t})^\top$ are standardised *i.i.d.* innovations with $\mathsf{E}[\varepsilon_{j,t}] = 0$ and $\mathsf{E}[\varepsilon_{j,t}^2] = 1$ for $j = 1, \ldots, d$; $\mathcal{F}_t$ is the available information at time t:

$$\mu_{j,t} = E[X_{j,t} \mid \mathcal{F}_{t-1}]$$

is the conditional mean given $\mathcal{F}_{t-1}$ and

$$\sigma_{j,t}^2 = E[(X_{j,t} - \mu_{j,t})^2 \mid \mathcal{F}_{t-1}]$$

is the conditional variance given $\mathcal{F}_{t-1}$. The innovations $\varepsilon = (\varepsilon_1, \ldots, \varepsilon_d)^\top$ have joint distribution

$$F_\varepsilon(\varepsilon_1, \ldots, \varepsilon_d) = C_\theta\{F_1(\varepsilon_1), \ldots, F_d(\varepsilon_d)\}, \tag{15}$$

where C_θ is a copula belonging to a parametric family $\mathcal{C} = \{C_\theta, \theta \in \Theta\}$, and F_j, $j = 1, \ldots, d$ are continuous marginal distributions of ε_j. To obtain the Value-at-Risk in this set up, the dependence parameter and distribution function from residuals are estimated from a sample of log-returns and used to generate P&L Monte Carlo samples. Their quantiles at different levels are the estimators for the Value-at-Risk.

For a portfolio w on d assets and a sample $\{x_{j,t}\}_{t=1}^T$, $j = 1, \ldots, d$ of log-returns, the Value-at-Risk at level α is estimated according to the following steps:

1. Estimation of residuals $\hat{\varepsilon}_t$ from the prespecified time-series model;
2. Specification and estimation of marginal distributions $F_j(\hat{\varepsilon}_j)$;
3. Specification of a parametric copula family $\mathcal{C}$ and estimation of dependence parameter θ;
4. Generation of Monte Carlo sample of innovations ε and losses L, for the forecast on the one day;

5. Estimation of $\widehat{VaR}(\alpha)$, the empirical α-quantile from the forecasted L.

The application of the (*static*) procedure described above on sliding windows of a time series $\{x_{j,t}\}_{t=1}^{T}$ delivers a sequence of parameters for a copula family. Hence the denomination *time-varying copulae*.

Using moving windows of size r in time t

$$\{x_t\}_{t=s-w+1}^{s}$$

for $s = r, \ldots, T$, the procedure described in the section above generates the time series $\{\widehat{VaR}_t\}_{t=r}^{T}$ of Value-at-Risk and $\{\hat{\theta}_t\}_{t=r}^{T}$ dependence parameters estimates.

Afterwards *Backtesting* is used to evaluate the performance of the specified copula family $\mathcal{C}$. The estimated values for the VaR are compared with the true realisations $\{l_t\}$ of the P&L function, an *exceedance* occuring for each l_t smaller than $\widehat{VaR}_t(\alpha)$. The ratio of the number of exceedances to the number of observations gives the *exceedances ratio* $\hat{\alpha}$:

$$\hat{\alpha} = \frac{1}{T-r} \sum_{t=r}^{T} \mathbf{I}\{l_t < \widehat{VaR}_t(\alpha)\}.$$

The estimation methods described before are used on two portfolio, the first composed of 2 positions, the second of 3 positions. Different copulae are used in static and dynamic setups and their VaR performance is compared based on backtesting.

In this section, the Value-at-Risk of portfolios for two companies (Tyssenkrupp (TKA) and Volkswagen (VOW) from 01.12.1997 to 03.07.2007) is computed using different copulae.

Assuming the log-returns $\{X_{j,t}\}$ follow a GARCH(1,1) process we have

$$X_{j,t} = \mu_{j,t} + \sigma_{j,t}\varepsilon_{j,t}$$

where

$$\sigma_{j,t}^2 = \omega_j + \alpha_j \sigma_{j,t-1}^2 + \beta_j (X_{j,t-1} - \mu_{j,t-1})^2$$

and $\omega > 0$, $\alpha_j \geq 0$, $\beta_j \geq 0$, $\alpha_j + \beta_j < 1$.

The fit of a GARCH(1,1) model to the sample of log returns $\{x_t\}_{t=1}^{T}$, $X_t = (X_{1,t}, X_{2,t})^\top$, $T = 2500$, gives the estimates $\hat{\omega}_j$, $\hat{\alpha}_j$ and $\hat{\beta}_j$, as in Table 2, and empirical residuals $\{\hat{\varepsilon}_t\}_{t=1}^{T}$, where $\hat{\varepsilon}_t = (\hat{\varepsilon}_{1,t}, \hat{\varepsilon}_{2,t})^\top$. The marginal distributions are specified as normal, i.e., $\hat{\varepsilon}_j \sim \mathrm{N}(\hat{\mu}_j, \hat{\sigma}_j)$ with parameters $\hat{\delta}_j = (\hat{\mu}_j, \hat{\sigma}_j)$ estimated from the data.

Figure 1 displays the Kernel density estimator of the residuals and of the normal density, estimated with an Quartic kernel. The dependence parameters are estimated for different copula families (Gaussian, Clayton and Gumbel). Residuals $\hat{\varepsilon}$ and fitted copulae (Gaussian, Clayton and Gumbel) are plotted in Figure 2.

In the dynamic approach, the empirical residuals are sampled in moving windows with a fixed size $r = 250$, $\{\hat{\varepsilon}_t\}_{t=s-r+1}^{s}$, for $s = r, \ldots, T$. The time series from estimated dependence parameters for each copula family are in Figure 3.

The same portfolio compositions as in the static case are used to generate P&L samples. The series of estimated Value-at-Risk and the P&L function for selected portfolios are plotted in Figure 4, 5 and 6.

	$\hat{\mu}_j$	$\hat{\omega}_j$	$\hat{\alpha}_j$	$\hat{\beta}_j$	BL	KS
MRK	7.392e-04	4.588e-06	3.333e-02	9.572e-01	0.1285	1.255e-11
	(3.672e-04)	(1.557e-06)	(6.225e-03)	(8.568e-03)		
TKA	7.845e-04	3.549e-06	7.087e-02	9.252e-01	0.1360	4.189e-05
	(3.308e-04)	(1.149e-06)	(9.837e-03)	(9.915e-03)		
VOW	9.720e-04	1.239e-05	9.303e-02	8.830e-01	1.927e-05	3.422e-06
	(3.480e-04)	(2.699e-06)	(1.301e-02)	(1.566e-02)		

Table 2: Fitting of univariate GARCH(1,1) to asset returns. The standard deviation of the parameters are given in parentheses. The last two columns provide the p-values of the Box-Ljung test (BL) for autocorrelations and Kolmogorov-Smirnov test (KS) for normality applied to the residuals

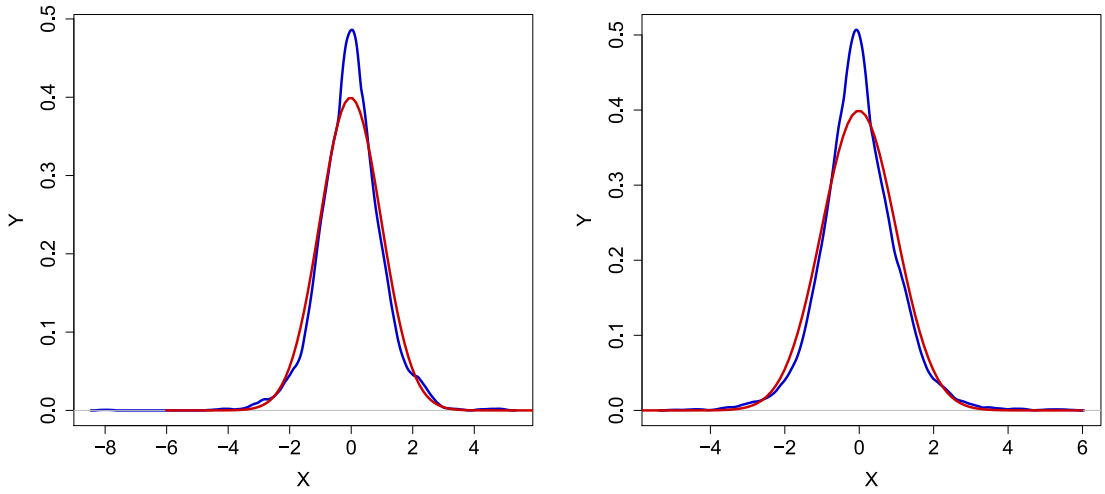


Fig. 1: Kernel density estimator of the residuals and of the normal density from TKA (left) and VOW (right). Quartic kernel, $\hat{h} = 2.78\hat{\sigma}n^{-0.2}$.

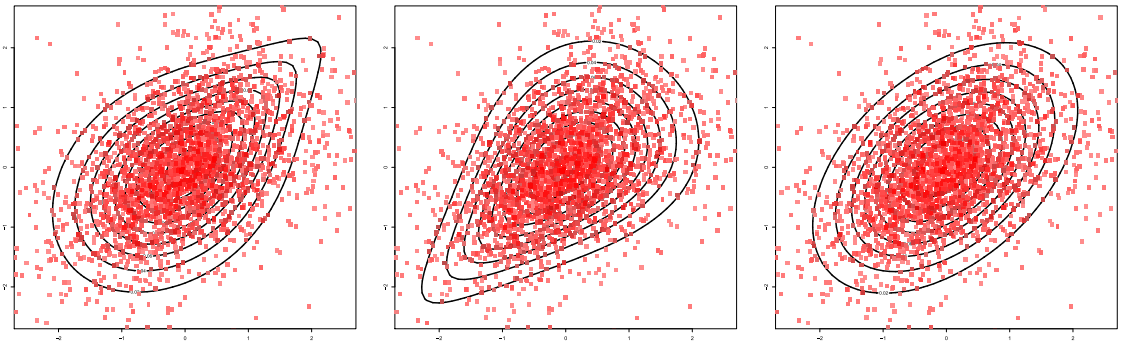


Fig. 2: Residuals $\hat{\varepsilon}$ and fitted copulae: Gaussian ($\hat{\rho} = 0.462$), Clayton ($\hat{\theta} = 0.880$), Gumbel ($\hat{\theta} = 1.439$).

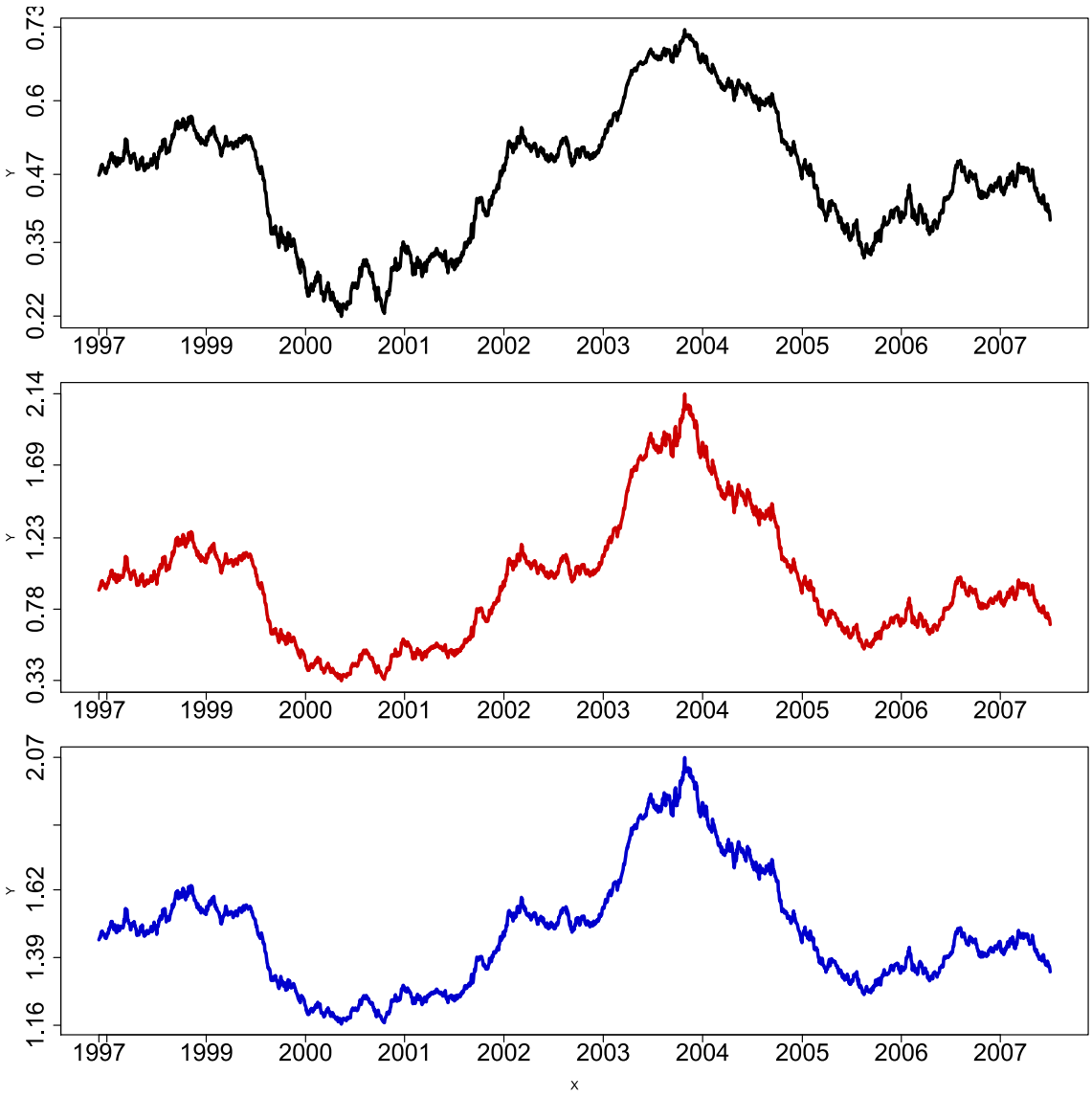


Fig. 3: Dependence parameter $\hat{\theta}$, estimated using the IFM method, Gaussian (upper panel), Gumbel (middle panel) and Clayton (lower panel) copulae, moving window ($w = 250$).

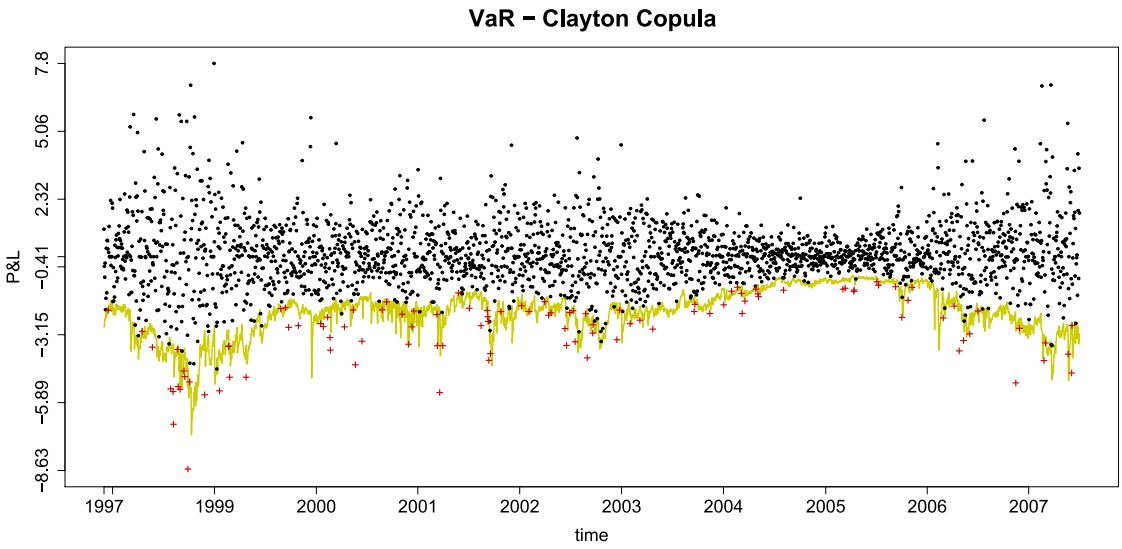


Fig. 4: $\widehat{VaR}(\alpha)$ (solid line), P&L (dots) and exceedances (crosses), $\alpha = 0.05$, $\hat{\alpha} = 0.0424$. P&L samples generated with Clayton copula.

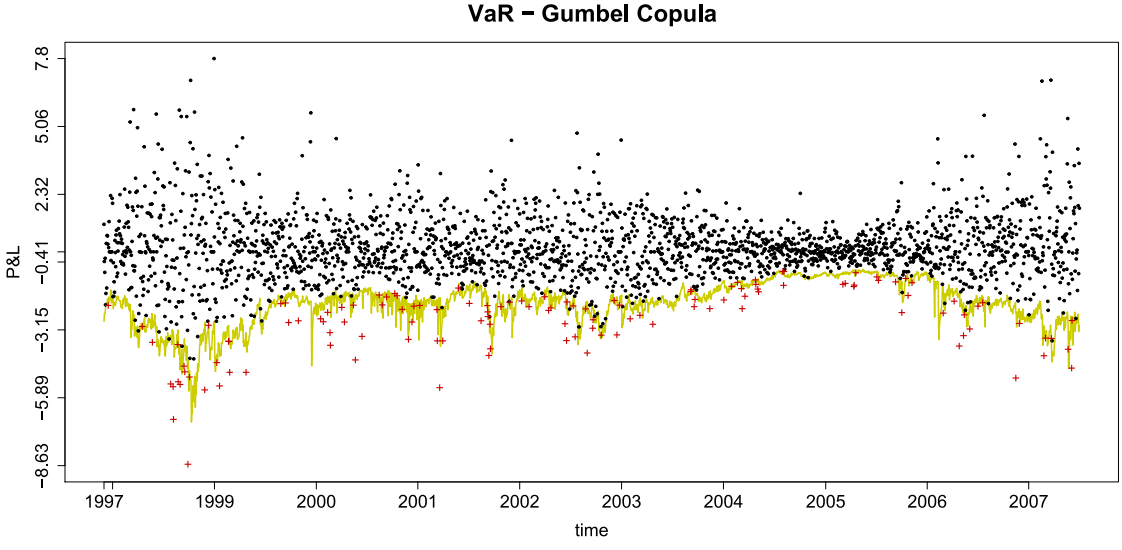


Fig. 5: $\widehat{VaR}(\alpha)$ (solid line), P&L (dots) and exceedances (crosses), $\alpha = 0.05$, $\hat{\alpha} = 0.0508$. P&L samples generated with Gumbel copula.

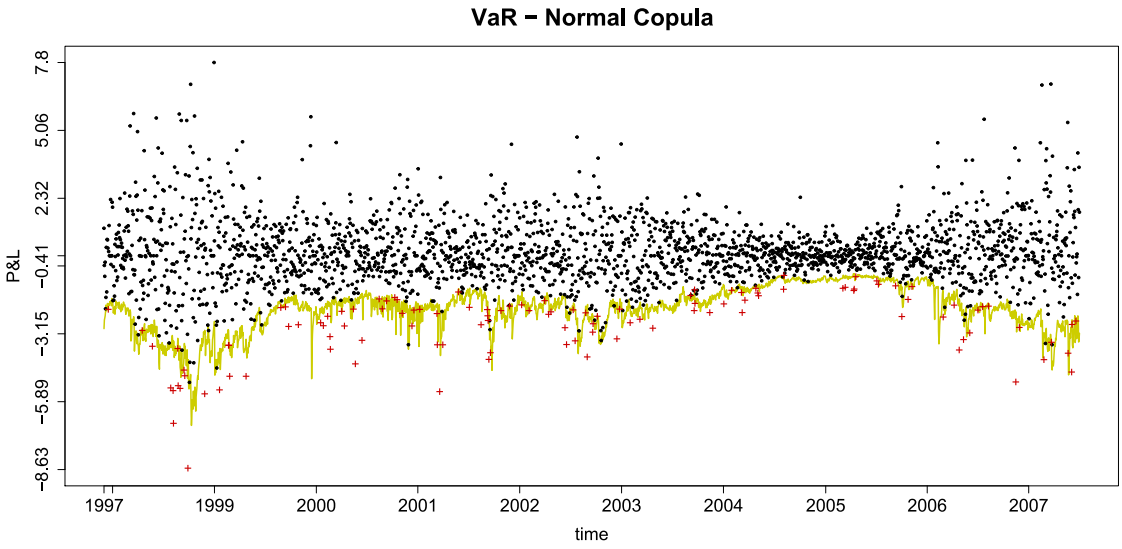


Fig. 6: $\widehat{VaR}(\alpha)$ (solid line), P&L (dots) and exceedances (crosses), $\alpha = 0.05$, $\hat{\alpha} = 0.0464$. P&L samples generated with Gaussian copula.

7.2 3-dimensional Portfolio

In this section, the Value-at-Risk of portfolios composed of 3 positions (Merck (MRK), Tyssenkrupp (TKA) and Volkswagen (VOW) from 01.12.1997 to 03.07.2007) is computed using a time-varying simple Gumbel copula and time-varying hierarchical Archimedean copula with generators from the Gumbel family.

The estimation of the parameters of the 3-dimensional copula was done by the IFM method. Concerning the HAC, we determine the structure under each window and re-estimate the parameters.

The fit of a GARCH(1,1) model to the sample of log returns $\{X_t\}_{t=1}^{T}$, $X_t = (X_{1,t}, X_{2,t}, X_{3,t})^\top$, $T = 2500$, gives the estimates $\hat{\omega}_j$, $\hat{\alpha}_j$ and $\hat{\beta}_j$, as in Table 2, and empirical residuals $\{\hat{\varepsilon}_t\}_{t=1}^{T}$, where $\hat{\varepsilon}_t = (\hat{\varepsilon}_{1,t}, \hat{\varepsilon}_{2,t}, \hat{\varepsilon}_{3,t})^\top$, as in upper right part of Figure 8. The marginal distributions are specified as normal, $\hat{\varepsilon}_j \sim \mathrm{N}(\hat{\mu}_j, \hat{\sigma}_j)$ with the estimated parameters $\hat{\delta}_j = (\hat{\mu}_j, \hat{\sigma}_j)$.

The estimated Value-at-Risk at level α together with the P&L function are plotted in Figure 9 for the simple Archimedean Copula (AC) and on 10 for the HAC. As can be seen from the backtesting results for different VaR levels, HAC outperforms the simple AC in all levels. This implies the necessity of dependence flexibility in modelling of log-returns.

8 Summary

To conclude, a summary of the main findings of this paper. We calculated the Value-at-Risk for the static and dynamic portfolio constructed by different methods. Three different copulae - Gumbel, Clayton and Gaussian - were used to estimate the Value-at-Risk from the two- (MRK and TKA) and three- (MRK, TKA and VOW) dimensional portfolios. From the time series of estimated dependence parameters, we can verify that the dependence structure is represented in a similar form with all copula families, as in Figure 3.

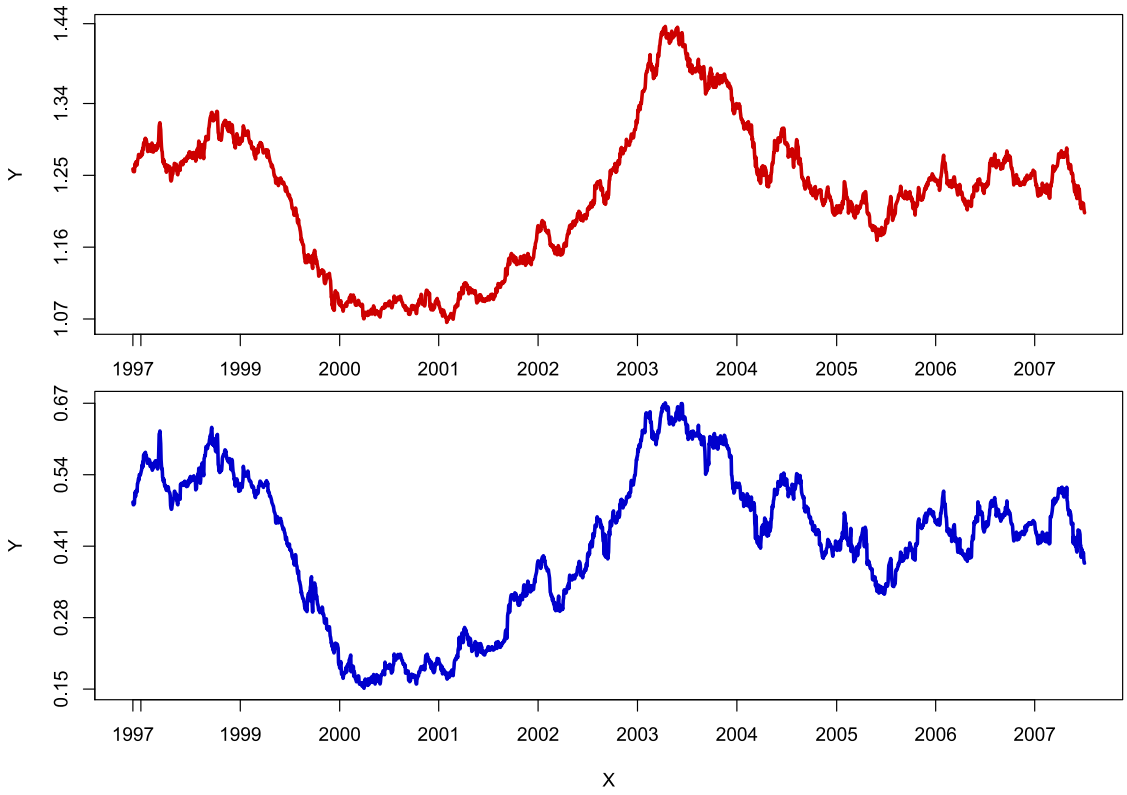


Fig. 7: Dependence parameter $\hat{\theta}$, estimated using the IFM method, Clayton (upper panel) and Gumbel (lower panel) copulae, moving window ($w = 250$).

Using backtesting results to compare the performance in the VaR estimation, we remark that on average the Clayton and Gaussian copulae *overestimate* the VaR. In terms of capital requirement, a financial institution computing VaR with those copulae would be requested to keep *more* capital aside than necessary to guarantee the desired confidence level.

The estimation with Gumbel copula, on another side, produced results close to the desired level. Gumbel copulae seems to represent specific data dependence structures (like lower tail dependencies, relevant to explain simultaneous losses) better than Gaussian and Clayton copulae.

References

Clayton, D. G. (1978). A model for association in bivariate life tables and its application in epidemiological studies of familial tendency in chronic disease incidence, *Biometrika* **65**: 141–151.

Deutsch, H. and Eller, R. (1999). *Derivatives and Internal Models*, Macmillan Press.

Devroye, L. (1986). *Non-uniform Random Variate Generation*, Springer Verlag, New York.

Embrechts, P., McNeil, A. J. and Straumann, D. (1999). Correlation and dependence in risk management: Properties and pitfalls, *RISK* pp. 69–71.

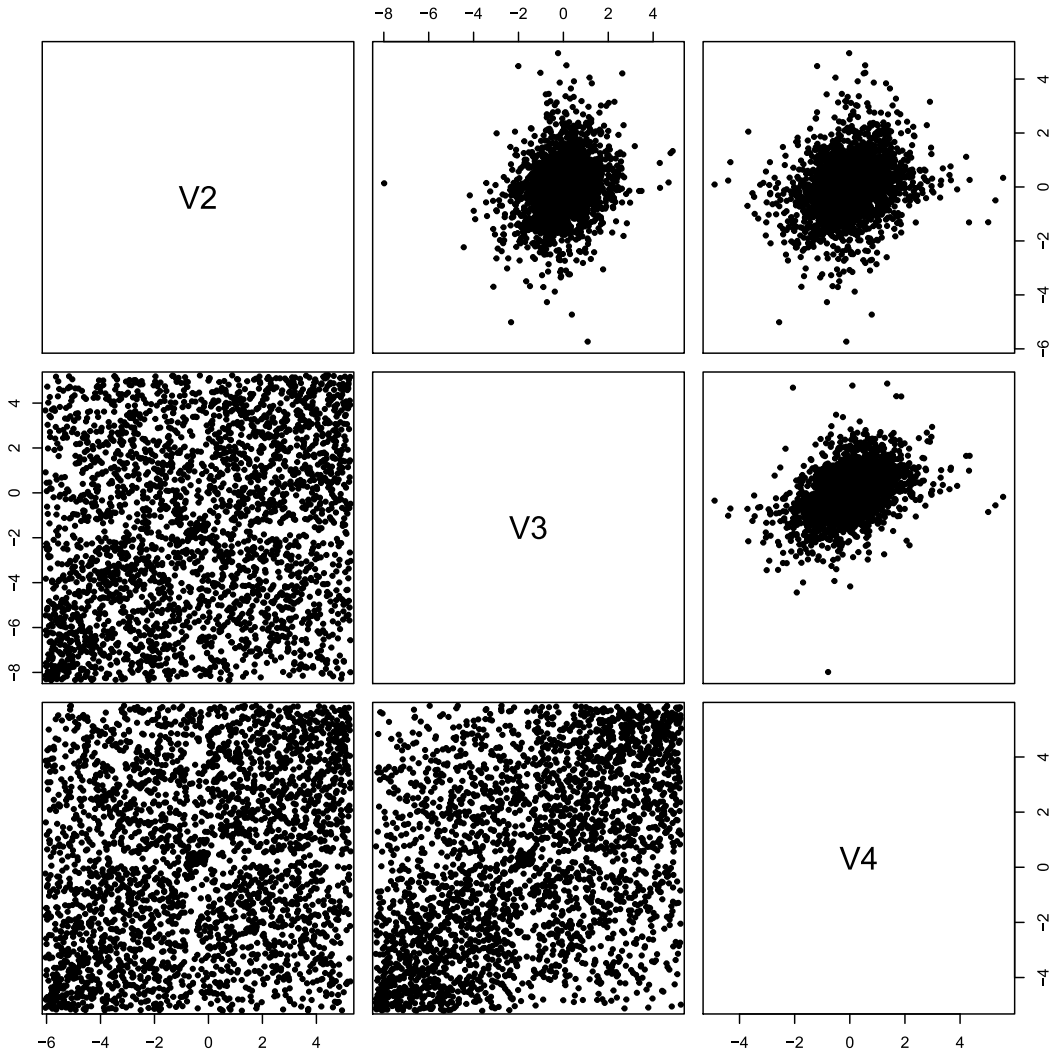


Fig. 8: Scatterplots from GARCH residulas (upper triangular) and from residuals mapped on unit square by the cdf (lower triangular).

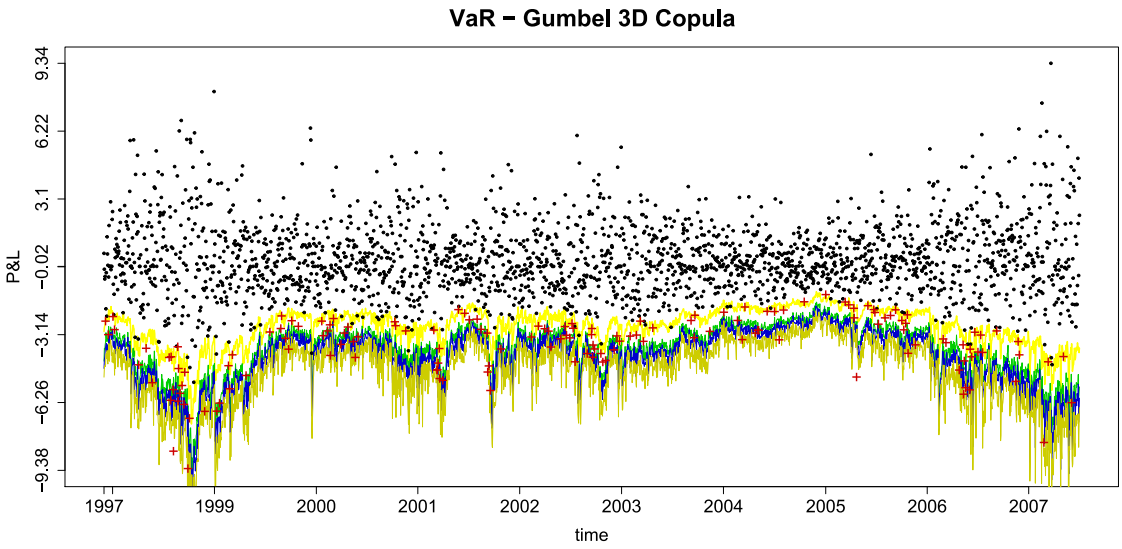


Fig. 9: $\widehat{VaR}(\alpha)$ and P&L (dots), estimated with 3-dimensional simple Gumbel copula, $\alpha_1 = 0.05$ ($\hat{\alpha}_1 = 0.0612$), $\alpha_2 = 0.01$ ($\hat{\alpha}_2 = 0.0232$), $\alpha_3 = 0.005$ ($\hat{\alpha}_3 = 0.016$) and $\alpha_4 = 0.001$ ($\hat{\alpha}_4 = 0.006$).

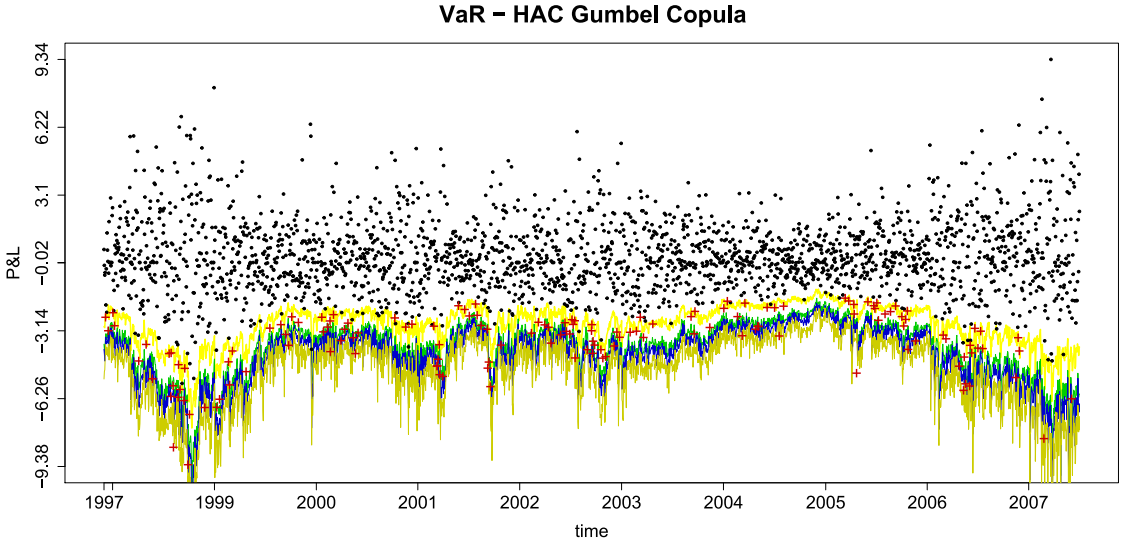


Fig. 10: $\widehat{VaR}(\alpha)$ and P&L (dots), estimated with 3-dimensional HAC with Gumbel generators, $\alpha_1 = 0.05$ ($\hat{\alpha}_1 = 0.0592$), $\alpha_2 = 0.01$ ($\hat{\alpha}_2 = 0.0208$), $\alpha_3 = 0.005$ ($\hat{\alpha}_3 = 0.014$) and $\alpha_4 = 0.001$ ($\hat{\alpha}_4 = 0.004$).

Embrechts, P., McNeil, A. and Straumann, D. (1999b). Correlation: Pitfalls and alternatives, *RISK* **May**: 69–71.

Frank, M. J. (1979). On the simultaneous associativity of $f(x, y)$ and $x + y - f(x, y)$, *Aequationes Mathematicae* **19**: 194–226.

Frees, E. and Valdez, E. (1998). Understanding relationships using copulas, *North American Actuarial Journal* **2**: 1–125.

Frey, R. and McNeil, A. J. (2003). Dependent defaults in models of portfolio credit risk, *Journal of Risk* **6**(1): 59–92.

Genest, C. and Rivest, L.-P. (1989). A characterization of Gumbel family of extreme value distributions, *Statistics and Probability Letters* **8**: 207–211.

Giacomini, E. and Härdle, W. (2005). Value-at-risk calculations with time varying copulae, *Proceedings 55th International Statistical Institute, Sydney 2005* .

Gumbel, E. J. (1960). Distributions des valeurs extrêmes en plusieurs dimensions, *Publ. Inst. Statist. Univ. Paris* **9**: 171–173.

Hoeffding, W. (1940). Masstabinvariante Korrelationstheorie, *Schriften des Mathematischen Instituts und des Instituts für Angewandte Mathematik der Universität Berlin* **5**(3): 179–233.

Hoeffding, W. (1941). Masstabinvariante Korrelationsmasse für diskontinuierliche Verteilungen, *Archiv für die mathematische Wirtschafts- und Sozialforschung* **7**: 49–70.

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

Joe, H. and Xu, J. J. (1996). The estimation method of inference functions for margins for multivariate models, *Technical Report 166*, Department of Statistics, University of British Columbia.

Marshall, A. W. and Olkin, J. (1988). Families of multivariate distributions, *Journal of the American Statistical Association* **83**: 834–841.

McNeil, A. J. (2008). Sampling nested Archimedean copulas, *Journal Statistical Computation and Simulation* . forthcoming.

Nelsen, R. B. (2006). *An Introduction to Copulas*, Springer Verlag, New York.

Okhrin, O., Okhrin, Y. and Schmid, W. (2009a). On the structure and estimation of hierarchical Archimedean copulas. under revision in Journal of Econometrics.

Okhrin, O., Okhrin, Y. and Schmid, W. (2009b). Properties of Hierarchical Archimedean Copulas, *SFB 649 Discussion Paper 2009-014*, Sonderforschungsbereich 649, Humboldt-Universität zu Berlin, Germany. available at http://sfb649.wiwi.hu-berlin.de/papers/pdf/SFB649DP2009-014.pdf.

Patton, A. J. (2004). On the out-of-sample importance of skewness and asymmetric dependence for asset allocation, *Journal of Financial Econometrics* **2**: 130–168.

Savu, C. and Trede, M. (2006). Hierarchical Archimedean copulas, *Discussion paper*, University of Muenster.

Sklar, A. (1959). Fonctions de répartition à n dimensions et leurs marges, **8**: 229–231.

Whelan, N. (2004). Sampling from Archimedean copulas, *Quantitative Finance* **4**: 339–352.

Econometric Theory, **26**, 2010, 1180–1200.
doi:10.1017/S0266466609990491

CONFIDENCE BANDS IN QUANTILE REGRESSION

WOLFGANG K. HÄRDLE AND SONG SONG
Humboldt-Universität zu Berlin

Let $(X_1, Y_1), \ldots, (X_n, Y_n)$ be independent and identically distributed random variables and let $l(x)$ be the unknown p-quantile regression curve of Y conditional on X. A quantile smoother $l_n(x)$ is a localized, nonlinear estimator of $l(x)$. The strong uniform consistency rate is established under general conditions. In many applications it is necessary to know the stochastic fluctuation of the process $\{l_n(x) - l(x)\}$. Using strong approximations of the empirical process and extreme value theory, we consider the asymptotic maximal deviation $\sup_{0 \leqslant x \leqslant 1} |l_n(x) - l(x)|$. The derived result helps in the construction of a uniform confidence band for the quantile curve $l(x)$. This confidence band can be applied as a econometric model check. An economic application considers the relation between age and earnings in the labor market by means of parametric model specification tests, which presents a new framework to describe trends in the entire wage distribution in a parsimonious way.

1. INTRODUCTION

In standard regression function estimation, most investigations are concerned with the conditional mean regression. However, new insights about the underlying structures can be gained by considering other aspects of the conditional distribution. The quantile curves are key aspects of inference in various economic problems and are of great interest in practice. These describe the conditional behavior of a response variable (e.g., wage of workers) given the value of an explanatory variable (e.g., education level, experience, occupation of workers) and investigate changes in both tails of the distribution, other than just the mean.

When examining labor markets, economists are concerned with whether discrimination exists, e.g., for different genders, nationalities, union status, etc. To study this question, we need to separate out other effects first, e.g., age, education, etc. The crucial relation between age and earnings or salaries belongs to the most carefully studied subjects in labor economics. The fundamental work in mean regression can be found in Murphy and Welch (1990). Quantile regression estimates could provide more accurate measures. Koenker and Hallock (2001) present a group of important economic applications, including quantile

Financial support from the Deutsche Forschungsgemeinschaft via SFB 649 "Ökonomisches Risiko," Humboldt-Universität zu Berlin, is gratefully acknowledged. We thank the editor and two referees for concrete suggestions on improving the manuscript and restructuring the paper. Their valuable comments and suggestions are gratefully acknowledged. Address correspondence to Song Song, Institute for Statistics and Econometrics, Humboldt-Universität zu Berlin, Spandauer Straße 1, 10178 Berlin, Germany; e-mail: songsong@cms.hu-berlin.de.

  0266-4666/10 $15.00

Engel curves, and claim that "quantile regression is gradually developing into a comprehensive strategy for completing the regression prediction." Besides this, it is also well known that a quantile regression model (e.g., the conditional median curve) is more robust to outliers, especially for fat-tailed distributions. For symmetric conditional distributions the quantile regression generates the nonparametric mean regression analysis because the $p = 0.5$ (median) quantile curve coincides with the mean regression.

As first introduced by Koenker and Bassett (1978), one may assume a parametric model for the p-quantile curve and estimate parameters by the interior point method discussed by Koenker and Park (1996) and Portnoy and Koenker (1997). Similarly, we can also adopt nonparametric methods to estimate conditional quantiles. The first one, a more direct approach using a check function such as a robustified local linear smoother, is provided by Fan, Hu, and Troung (1994) and further extended by Yu and Jones (1997, 1998). An alternative procedure is first to estimate the conditional distribution function using the double-kernel local linear technique of Fan, Yao, and Tong (1996) and then to invert the conditional distribution estimator to produce an estimator of a conditional quantile by Yu and Jones (1997, 1998). Beside these, Hall, Wolff, and Yao (1999) proposed a weighted version of the Nadaraya–Watson estimator, which was further studied by Cai (2002). Recently Jeong and Härdle (2008) have developed the conditional quantile causality test. More generally, for an M-regression function that involves quantile regression as a special case, the uniform Bahadur representation and application to the additive model are studied by Kong, Linton, and Xia (2010). An interesting question for parametric fitting, especially from labor economists, would be how well these models fit the data, when compared with the nonparametric estimation method.

Let $(X_1, Y_1), (X_2, Y_2), \ldots, (X_n, Y_n)$ be a sequence of independent and identically distributed (i.i.d.) bivariate random variables with joint probability density function (pdf) $f(x, y)$, joint cumulative distribution function (cdf) $F(x, y)$, conditional pdf $f(y|x)$, $f(x|y)$, conditional cdf $F(y|x)$, $F(x|y)$ for Y given X and X given Y, respectively, and marginal pdf $f_X(x)$ for X, $f_Y(y)$ for Y where $x \in J$ and J is a possibly infinite interval in $\mathbb{R}^d$ and $y \in \mathbb{R}$. In general, X may be a multivariate covariate, although here we restrict attention to the univariate case and $J = [0, 1]$ for convenience. Let $l(x)$ denote the p-quantile curve, i.e., $l(x) = F^{-1}_{Y|x}(p)$.

Under a "check function," the quantile regression curve $l(x)$ can be viewed as the minimizer of $L(\theta) \stackrel{\text{def}}{=} \mathsf{E}\{\rho_p(y - \theta)|X = x\}$ (with respect to θ) with $\rho_p(u) = pu\mathbf{1}\{u \in (0, \infty)\} - (1 - p)u\mathbf{1}\{u \in (-\infty, 0)\}$, which was originally motivated by an exercise in Ferguson (1967, p. 51) in the literature.

A kernel-based p-quantile curve estimator $l_n(x)$ can naturally be constructed by minimizing:

$$L_n(\theta) = n^{-1} \sum_{i=1}^{n} \rho_p(Y_i - \theta) K_h(x - X_i) \tag{1}$$

with respect to $\theta \in I$ where I is a possibly infinite, or possibly degenerate, interval in $\mathbb{R}$ and $K_h(u) = h^{-1}K(u/h)$ is a kernel with bandwidth h. The numerical solution of (1) may be found iteratively as in Lejeune and Sarda (1988) and Yu, Lu, and Stander (2003).

In light of the concepts of M-estimation as in Huber (1981), if we define $\psi(u)$ as

$$\begin{aligned}\psi_p(u) &= p\mathbf{1}\{u \in (0,\infty)\} - (1-p)\mathbf{1}\{u \in (-\infty, 0)\} \\ &= p - \mathbf{1}\{u \in (-\infty, 0)\},\end{aligned}$$

$l_n(x)$ and $l(x)$ can be treated as a zero (with respect to θ) of the function

$$\widetilde{H}_n(\theta, x) \stackrel{\text{def}}{=} n^{-1} \sum_{i=1}^{n} K_h(x - X_i)\psi(Y_i - \theta), \tag{2}$$

$$\widetilde{H}(\theta, x) \stackrel{\text{def}}{=} \int_{\mathbb{R}} f(x, y)\psi(y - \theta)\, dy, \tag{3}$$

correspondingly.

To show the uniform consistency of the quantile smoother, we shall reduce the problem of strong convergence of $l_n(x) - l(x)$, uniformly in x, to an application of the strong convergence of $\widetilde{H}_n(\theta, x)$ to $\widetilde{H}(\theta, x)$, uniformly in x and θ, as given by Theorem 2.2 in Härdle, Janssen, and Serfling (1988). It is shown that under general conditions almost surely (a.s.)

$$\sup_{x \in J} |l_n(x) - l(x)| \leqslant B^* \max\left\{(nh/(\log n))^{-1/2}, h^{\tilde{\alpha}}\right\}, \quad \text{as } n \to \infty,$$

where B^* and $\tilde{\alpha}$ are parameters defined more precisely in Section 2.

Note that without assuming K has compact support (as we do here) under similar assumptions Franke and Mwita (2003) obtain

$$l_n(x) = \hat{F}^{-1}_{Y|x}(p),$$

$$\hat{F}(y|x) = \frac{\sum_{i=1}^{n} K_h(x - X_i)\mathbf{1}(Y_i < y)}{\sum_{i=1}^{n} K_h(x - X_i)},$$

$$\sup_{x \in J} |l_n(x) - l(x)| \leqslant B^{**}\left\{(nh/(s_n \log n))^{-1/2} + h^2\right\}, \quad \text{as } n \to \infty$$

for α-mixing data where B^{**} is some constant and $s_n, n \geqslant 1$ is an increasing sequence of positive integers satisfying $1 \leqslant s_n \leqslant n/2$ and some other criteria. Thus $\{nh/(\log n)\}^{-1/2} \leqslant \{nh/(s_n \log n)\}^{-1/2}$.

By employing similar methods to those developed in Härdle (1989) it is shown in this paper that

$$\begin{aligned}&\mathrm{P}\left((2\delta \log n)^{1/2}\left[\sup_{x \in J} r(x)|\{l_n(x) - l(x)\}|/\lambda(K)^{1/2} - d_n\right] < z\right) \\ &\quad \to \exp\{-2\exp(-z)\}, \quad \text{as } n \to \infty\end{aligned} \tag{4}$$

from the asymptotic Gumbel distribution where $r(x)$, δ, $\lambda(K)$, d_n are suitable scaling parameters. The asymptotic result (4) therefore allows the construction of (asymptotic) uniform confidence bands for $l(x)$ based on specifications of the stochastic fluctuation of $l_n(x)$. The strong approximation with Brownian bridge techniques that we use in this paper is available only for the approximation of the two-dimensional empirical process. The extension to the multivariate covariable can be done by partial linear modeling, which deserves further research.

The plan of the paper is as follows. In Section 2, the stochastic fluctuation of the process $\{l_n(x) - l(x)\}$ and the uniform confidence band are presented through the equivalence of several stochastic processes, with a strong uniform consistency rate of $\{l_n(x) - l(x)\}$ also shown. In Section 3, in a small Monte Carlo study we investigate the behavior of $l_n(x)$ when the data are generated by fat-tailed conditional distributions of $(Y|X = x)$. In Section 4, an application considers a wage-earning relation in the labor market. All proofs are sketched in the Appendix.

2. RESULTS

The following assumptions will be convenient. To make x and X clearly distinguishable, we replace x by t sometimes, but they are essentially the same.

(A1) The kernel $K(\cdot)$ is positive and symmetric, has compact support $[-A, A]$, and is Lipschitz continuously differentiable with bounded derivatives.

(A2) $(nh)^{-1/2}(\log n)^{3/2} \to 0$, $(n \log n)^{1/2} h^{5/2} \to 0$, $(nh^3)^{-1}(\log n)^2 \leqslant M$, where M is a constant.

(A3) $h^{-3}(\log n) \int_{|y|>a_n} f_Y(y)dy = \mathcal{O}(1)$, where $f_Y(y)$ is the marginal density of Y and $\{a_n\}_{n=1}^{\infty}$ is a sequence of constants tending to infinity as $n \to \infty$.

(A4) $\inf_{t \in J} |q(t)| \geqslant q_0 > 0$, where $q(t) = \partial \mathsf{E}\{\psi(Y - \theta)|t\}/\partial\theta|_{\theta=l(t)} \cdot f_X(t) = f\{l(t)|t\} f_X(t)$.

(A5) The quantile function $l(t)$ is Lipschitz twice continuously differentiable for all $t \in J$.

(A6) $0 < m_1 \leqslant f_X(t) \leqslant M_1 < \infty$, $t \in J$; the conditional densities $f(\cdot|y)$, $y \in \mathbb{R}$, are uniform local Lipschitz continuous of order $\tilde{\alpha}$ (ulL-$\tilde{\alpha}$) on J, uniformly in $y \in \mathbb{R}$, with $0 < \tilde{\alpha} \leqslant 1$.

Define also

$$\sigma^2(t) = \mathsf{E}[\psi^2\{Y - l(t)\}|t] = p(1-p),$$

$$H_n(t) = (nh)^{-1} \sum_{i=1}^{n} K\{(t - X_i)/h\} \psi\{Y_i - l(t)\},$$

$$D_n(t) = \partial(nh)^{-1} \sum_{i=1}^{n} K\{(t - X_i)/h\} \psi\{Y_i - \theta\}/\partial\theta|_{\theta=l(t)}$$

and assume that $\sigma^2(t)$ and $f_X(t)$ are differentiable.

Assumption (A1) on the compact support of the kernel could possibly be relaxed by introducing a cutoff technique as in Csörgö and Hall (1982) for density estimators. Assumption (A2) has purely technical reasons: to keep the bias at a lower rate than the variance and to ensure the vanishing of some nonlinear remainder terms. Assumption (A3) appears in a somewhat modified form also in Johnston (1982). Assumptions (A5) and (A6) are common assumptions in robust estimation as in Huber (1981) and Härdle et al. (1988) that are satisfied by exponential and generalized hyperbolic distributions.

For the uniform strong consistency rate of $l_n(x) - l(x)$, we apply the result of Härdle et al. (1988) by taking $\beta(y) = \psi(y-\theta)$, $y \in \mathbb{R}$, for $\theta \in I = \mathbb{R}$, $q_1 = q_2 = -1$, $\gamma_1(y) = \max\{0, -\psi(y-\theta)\}$, $\gamma_2(y) = \min\{0, -\psi(y-\theta)\}$, and $\lambda = \infty$ to satisfy the representations for the parameters there. Thus from Härdle et al.'s Theorem 2.2 and Remark 2.3(v), we immediately have the following lemma.

LEMMA 2.1. *Let $\widetilde{H}_n(\theta, x)$ and $\widetilde{H}(\theta, x)$ be given by (2) and (3). Under Assumption (A6) and $(nh/\log n)^{-1/2} \to \infty$ through Assumption (A2), for some constant A^* not depending on n, we have a.s. as $n \to \infty$*

$$\sup_{\theta \in I} \sup_{x \in J} \left| \widetilde{H}_n(\theta, x) - \widetilde{H}(\theta, x) \right| \leq A^* \max \left\{ (nh/\log n)^{-1/2}, h^{\tilde{\alpha}} \right\}. \tag{5}$$

For our result on $l_n(\cdot)$, we shall also require

$$\inf_{x \in J} \left| \int \psi\{y - l(x) + \varepsilon\} \, dF(y|x) \right| \geqslant \tilde{q}|\varepsilon|, \quad \text{for } |\varepsilon| \leqslant \delta_1, \tag{6}$$

where δ_1 and $\tilde{q}$ are some positive constants; see also Härdle and Luckhaus (1984). This assumption is satisfied if there exists a constant $\tilde{q}$ such that $f(l(x)|x) > \tilde{q}/p$, $x \in J$.

THEOREM 2.1. *Under the conditions of Lemma 2.1 and also assuming (6), we have a.s. as $n \to \infty$*

$$\sup_{x \in J} \left| l_n(x) - l(x) \right| \leq B^* \max \left\{ (nh/\log n)^{-1/2}, h^{\tilde{\alpha}} \right\} \tag{7}$$

with $B^* = A^*/m_1\tilde{q}$ not depending on n and m_1 a lower bound of $f_X(t)$. If additionally $\tilde{\alpha} \geqslant \{\log(\sqrt{\log n}) - \log(\sqrt{nh})\}/\log h$, it can be further simplified to

$$\sup_{x \in J} |l_n(x) - l(x)| \leq B^* \left\{ (nh/\log n)^{-1/2} \right\}.$$

THEOREM 2.2. *Let $h = n^{-\delta}$, $\frac{1}{5} < \delta < \frac{1}{3}$, $\lambda(K) = \int_{-A}^{A} K^2(u)\,du$, and*

$$d_n = (2\delta \log n)^{1/2} + (2\delta \log n)^{-1/2} \left[\log \left\{ c_1(K)/\pi^{1/2} \right\} + \frac{1}{2} \left\{ \log \delta + \log\log n \right\} \right],$$

if $c_1(K) = \{K^2(A) + K^2(-A)\}/\{2\lambda(K)\} > 0$;

$d_n = (2\delta \log n)^{1/2} + (2\delta \log n)^{-1/2} \log\{c_2(K)/2\pi\}$

otherwise with $c_2(K) = \int_{-A}^{A} \{K'(u)\}^2 \, du/\{2\lambda(K)\}$. *Then (4) holds with*

$r(x) = (nh)^{1/2} f\{l(x)|x\}\{f_X(x)/p(1-p)\}^{1/2}$.

This theorem can be used to construct uniform confidence intervals for the regression function as stated in the following corollary.

COROLLARY 2.1. *Under the assumptions of Theorem 2.2, an approximate* $(1-\alpha) \times 100\%$ *confidence band over* $[0, 1]$ *is*

$$l_n(t) \pm (nh)^{-1/2} \left\{ p(1-p)\lambda(K)/\hat{f}_X(t) \right\}^{1/2} \hat{f}^{-1}\{l(t)|t\} \left\{ d_n + c(\alpha)(2\delta \log n)^{-1/2} \right\},$$

where $c(\alpha) = \log 2 - \log|\log(1-\alpha)|$ *and* $\hat{f}_X(t)$, $\hat{f}\{l(t)|t\}$ *are consistent estimates for* $f_X(t)$, $f\{l(t)|t\}$.

In the literature, according to Fan et al. (1994, 1996), Yu and Jones (1997, 1998), Hall et al. (1999), Cai (2002), and others, asymptotic normality at interior points for various nonparametric smoothers, e.g., local constant, local linear, reweighted Nadaraya–Watson methods, etc., has been shown:

$$\sqrt{nh}\{l_n(t) - l(t)\} \sim \mathrm{N}\big(0, \tau^2(t)\big)$$

with $\tau^2(t) = \lambda(K)p(1-p)/[f_X(t)f^2\{l(t)|t\}]$. Note that the bias term vanishes here as we adjust h. With $\tau(t)$ introduced, we can further write Corollary 2.1 as

$$l_n(t) \pm (nh)^{-1/2} \left\{ d_n + c(\alpha)(2\delta \log n)^{-1/2} \right\} \hat{\tau}(t).$$

Through minimizing the approximation of asymptotic mean square error, the optimal bandwidth h_p can be computed. In practice, the rule of thumb for h_p is given by Yu and Jones (1998):

1. Use ready-made and sophisticated methods to select optimal bandwidth h_{mean} from conditional mean regression, e.g., Ruppert, Sheather, and Wand (1995);
2. $h_p = [p(1-p)/\varphi^2\{\Phi^{-1}(p)\}]^{1/5} \cdot h_{\text{mean}}$ with φ, Φ as the pdf and cdf of a standard normal distribution

Obviously the further p lies from 0.5, the more smoothing is necessary.

The proof is essentially based on a linearization argument after a Taylor series expansion. The leading linear term will then be approximated in a similar way as in Johnston (1982) and Bickel and Rosenblatt (1973). The main idea behind the proof is a strong approximation of the empirical process of $\{(X_i, Y_i)_{i=1}^n\}$ by a sequence of Brownian bridges as proved by Tusnady (1977).

As $l_n(t)$ is the zero (with respect to θ) of $\widetilde{H}_n(\theta, t)$, it follows by applying second-order Taylor expansions to $\widetilde{H}_n(\theta, t)$ around $l(t)$ that

$$l_n(t) - l(t) = \{H_n(t) - \mathsf{E}\, H_n(t)\}/q(t) + R_n(t), \tag{8}$$

where $\{H_n(t) - \mathsf{E}\, H_n(t)\}/q(t)$ is the leading linear term and

$$R_n(t) = H_n(t)\{q(t) - D_n(t)\}/\{D_n(t) \cdot q(t)\} + \mathsf{E}\, H_n(t)/q(t)$$

$$+ \frac{1}{2}\{l_n(t) - l(t)\}^2 \cdot \{D_n(t)\}^{-1} \tag{9}$$

$$\cdot (nh)^{-1} \sum_{i=1}^{n} K\{(x - X_i)/h\}\psi''\{Y_i - l(t) + r_n(t)\}, \tag{10}$$

$|r_n(t)| < |l_n(t) - l(t)|$

is the remainder term. In the Appendix it is shown (Lemma A.1) that $\|R_n\| = \sup_{t \in J} |R_n(t)| = \mathcal{O}_p\left\{(nh \log n)^{-1/2}\right\}$.

Furthermore, the rescaled linear part

$$Y_n(t) = (nh)^{1/2}\left\{\sigma^2(t) f_X(t)\right\}^{-1/2}\{H_n(t) - \mathsf{E}\, H_n(t)\}$$

is approximated by a sequence of Gaussian processes, leading finally to the Gaussian process

$$Y_{5,n}(t) = h^{-1/2} \int K\{(t - x)/h\}\, dW(x). \tag{11}$$

Drawing upon the result of Bickel and Rosenblatt (1973), we finally obtain asymptotically the Gumbel distribution.

We also need the Rosenblatt (1952) transformation,

$$T(x, y) = \{F_{X|y}(x|y), F_Y(y)\},$$

which transforms (X_i, Y_i) into $T(X_i, Y_i) = (X_i', Y_i')$ mutually independent uniform random variables. In the event that x is a d-dimensional covariate, the transformation becomes

$$T(x_1, x_2, \ldots, x_d, y) = \{F_{X_1|y}(x_1|y), F_{X_2|y}(x_2|x_1, y), \ldots, F_{X_k|x_{d-1},\ldots,x_1,y}$$

$$(x_k|x_{d-1}, \ldots, x_1, y), F_Y(y)\}. \tag{12}$$

With the aid of this transformation, Theorem 1 of Tusnady (1977) may be applied to obtain the following lemma.

LEMMA 2.2. *On a suitable probability space a sequence of Brownian bridges B_n exists such that*

$$\sup_{x \in J, y \in \mathbb{R}} |Z_n(x, y) - B_n\{T(x, y)\}| = \mathcal{O}\left\{n^{-1/2}(\log n)^2\right\} \quad a.s.,$$

where $Z_n(x,y) = n^{1/2}\{F_n(x,y) - F(x,y)\}$ *denotes the empirical process of* $\{(X_i, Y_i)\}_{i=1}^n$.

For $d > 2$, it is still an open problem that deserves further research.

Before we define the different approximating processes, let us first rewrite (11) as a stochastic integral with respect to the empirical process $Z_n(x,y)$:

$$Y_n(t) = \{hg'(t)\}^{-1/2} \iint K\{(t-x)/h\}\psi\{y - l(t)\}\, dZ_n(x,y),$$

$$g'(t) = \sigma^2(t) f_X(t).$$

The approximating processes are now

$$Y_{0,n}(t) = \{hg(t)\}^{-1/2} \iint_{\Gamma_n} K\{(t-x)/h\}\psi\{y - l(t)\}\, dZ_n(x,y), \tag{13}$$

where $\Gamma_n = \{|y| \leqslant a_n\}$, $g(t) = \mathsf{E}[\psi^2\{y - l(t)\} \cdot \mathbf{1}(|y| \leqslant a_n)|X = t] \cdot f_X(t)$

$$Y_{1,n}(t) = \{hg(t)\}^{-1/2} \iint_{\Gamma_n} K\{(t-x)/h\}\psi\{y - l(t)\}\, dB_n\{T(x,y)\}, \tag{14}$$

$\{B_n\}$ being the sequence of Brownian bridges from Lemma 2.2.

$$Y_{2,n}(t) = \{hg(t)\}^{-1/2} \iint_{\Gamma_n} K\{(t-x)/h\}\psi\{y - l(t)\}\, dW_n\{T(x,y)\}, \tag{15}$$

$\{W_n\}$ being the sequence of Wiener processes satisfying

$$B_n(x', y') = W_n(x', y') - x'y'W_n(1,1),$$

$$Y_{3,n}(t) = \{hg(t)\}^{-1/2} \iint_{\Gamma_n} K\{(t-x)/h\}\psi\{y - l(x)\}\, dW_n\{T(x,y)\}, \tag{16}$$

$$Y_{4,n}(t) = \{hg(t)\}^{-1/2} \int g(x)^{1/2} K\{(t-x)/h\}\, dW(x), \tag{17}$$

$$Y_{5,n}(t) = h^{-1/2} \int K\{(t-x)/h\}\, dW(x), \tag{18}$$

$\{W(\cdot)\}$ being the Wiener process.

Lemmas A.2–A.7 in the Appendix ensure that all these processes have the same limit distributions. The result then follows from the next lemma.

LEMMA 2.3 (Theorem 3.1 in Bickel and Rosenblatt, 1973). *Let* d_n, $\lambda(K)$, δ *as in Theorem 2.2. Let*

$$Y_{5,n}(t) = h^{-1/2} \int K\{(t-x)/h\}\, dW(x).$$

Then, as $n \to \infty$, the supremum of $Y_{5,n}(t)$ has a Gumbel distribution:

$$\mathrm{P}\left\{(2\delta \log n)^{1/2}\left[\sup_{t\in J}|Y_{5,n}(t)|/\{\lambda(K)\}^{1/2} - d_n\right] < z\right\} \to \exp\{-2\exp(-z)\}.$$

3. A MONTE CARLO STUDY

We generate bivariate data $\{(X_i, Y_i)\}_{i=1}^n, n = 500$ with joint pdf:

$$f(x, y) = g\left(y - \sqrt{x + 2.5}\right)\mathbf{1}(x \in [-2.5, 2.5]), \tag{19}$$

$$g(u) = \frac{9}{10}\varphi(u) + \frac{1}{90}\varphi(u/9).$$

The p-quantile curve $l(x)$ can be obtained from a zero (with respect to θ) of

$$9\Phi(\theta) + \Phi(\theta/9) = 10p,$$

with Φ as the cdf of a standard normal distribution. Solving it numerically gives the 0.5-quantile curve $l(x) = \sqrt{x + 2.5}$ and the 0.9-quantile curve $l(x) = 1.5296 + \sqrt{x + 2.5}$. We use the quartic kernel:

$$K(u) = \frac{15}{16}(1 - u^2)^2, \qquad |u| \leqslant 1,$$

$$= 0, \qquad |u| > 1.$$

In Figure 1 the raw data, together with the 0.5-quantile curve, are displayed. The random variables generated with probability $\frac{1}{10}$ from the fat-tailed pdf $\frac{1}{9}\varphi(u/9)$ (see eqn. (19)) are marked as squares whereas the standard normal random variables are shown as stars. We then compute both the Nadaraya–Watson estimator $m_n^*(x)$ and the 0.5-quantile smoother $l_n(x)$. The bandwidth is set to 1.25, which is equivalent to 0.25 after rescaling x to $[0, 1]$ and fulfills the requirements of Theorem 2.2.

In Figure 1 $l(x)$, $m_n^*(x)$, and $l_n(x)$ are shown as a dotted line, dashed-dot line, and solid line, respectively. At first sight $m_n^*(x)$ has clearly more variation and has the expected sensitivity to the fat tails of $f(x, y)$. A closer look reveals that $m_n^*(x)$ for $x \approx 0$ apparently even leaves the 0.5-quantile curve. It may be surprising that this happens at $x \approx 0$ where no outlier is placed, but a closer look at Figure 1 shows that the large negative data values at both $x \approx -0.1$ and $x \approx 0.25$ cause the problem. This data value is inside the window ($h = 1.10$) and therefore distorts $m_n^*(x)$ for $x \approx 0$. The quantile smoother $l_n(x)$ (solid line) is unaffected and stays fairly close to the 0.5-quantile curve. Similar results can be obtained in Figure 2 corresponding to the 0.9 quantile ($h = 1.25$) with the 95% confidence band.

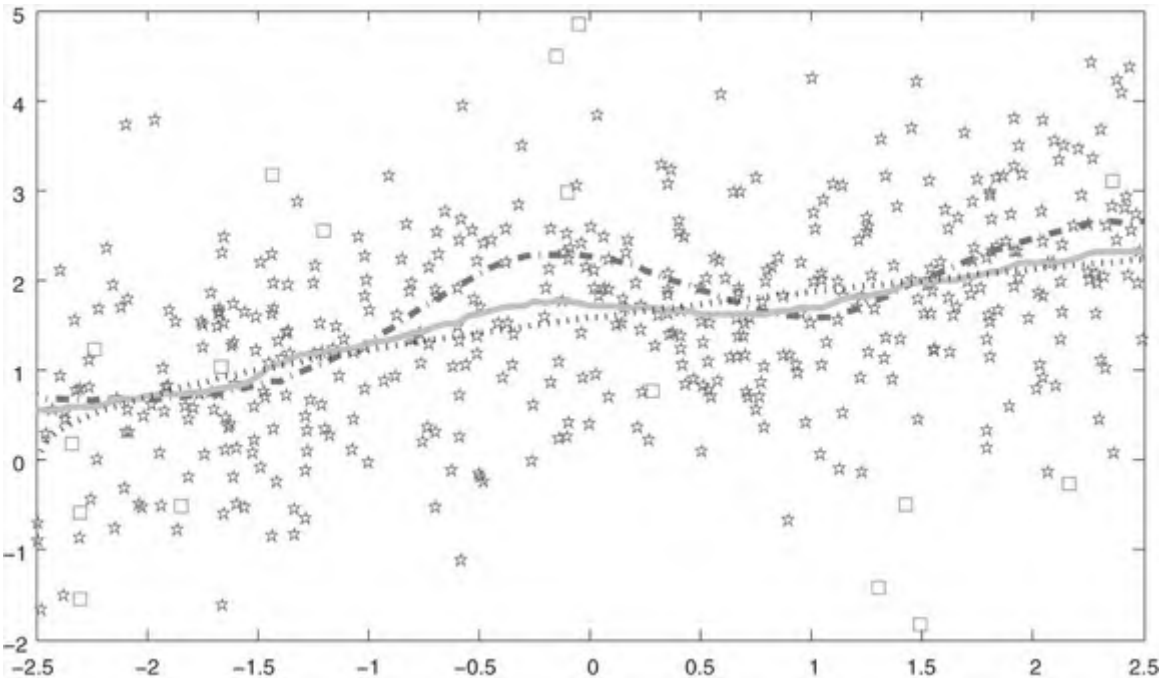


FIGURE 1. The 0.5-quantile curve, the Nadaraya–Watson estimator $m_n^*(x)$, and the 0.5-quantile smoother $l_n(x)$.

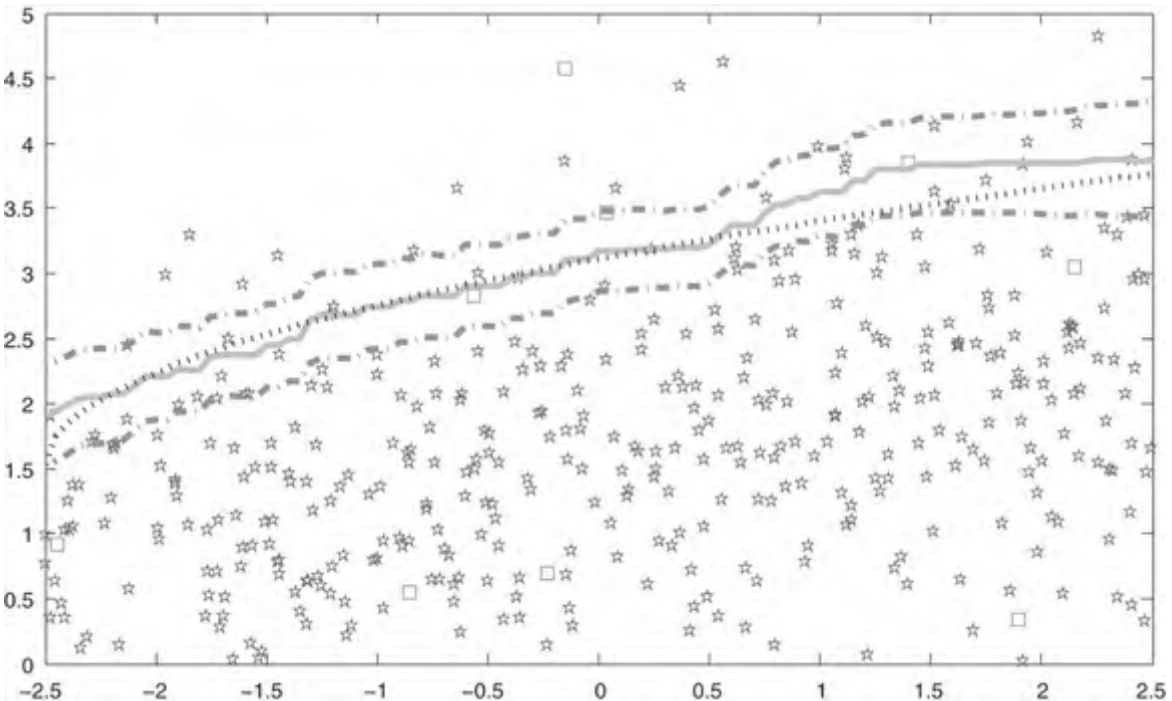


FIGURE 2. The 0.9-quantile curve, the 0.9-quantile smoother, and 95% confidence band.

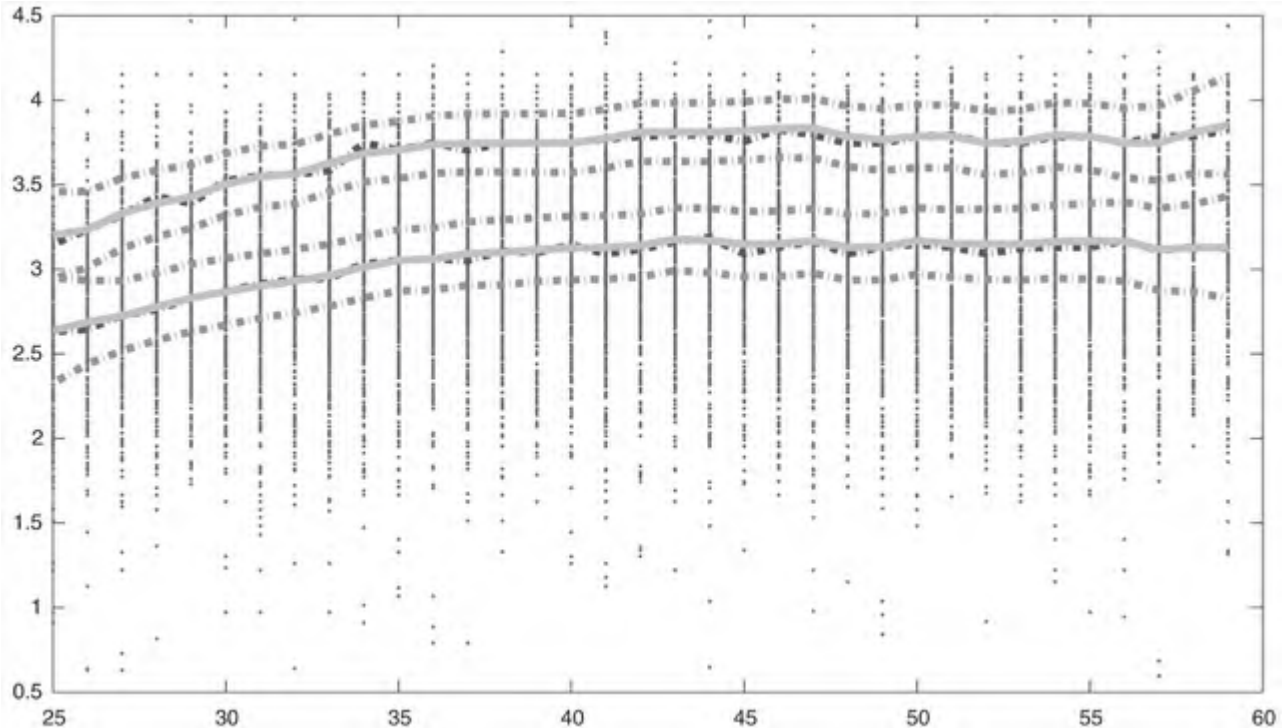


FIGURE 3. The original observations, local quantiles, 0.5- and 0.9-quantile smoothers, and corresponding 95% confidence bands.

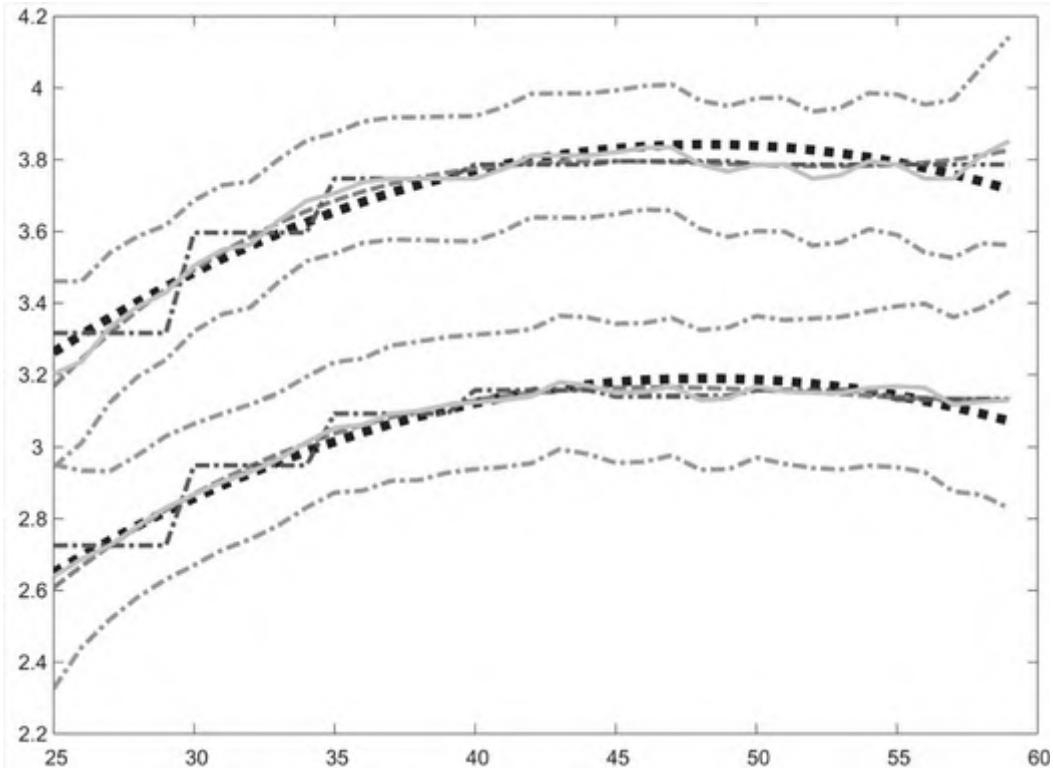


FIGURE 4. Quadratic, quartic, set of dummies (for age groups) estimates, 0.5- and 0.9-quantile smoothers, and their corresponding 95% confidence bands.

4. APPLICATION

Recently there has been great interest in finding out how the financial returns of a job depend on the age of the employee. We use the Current Population Survey (CPS) data from 2005 for the following group: male aged 25–59, full-time employed, and college graduate containing 16,731 observations, for the age-earning estimation. As is usual for wage data, a log transformation to hourly real wages (unit: U.S. dollar) is carried out first. In the CPS all ages (25–59) are reported as integers. We rescaled them into $[0, 1]$ by dividing 40 by bandwidth 0.059 for nonparametric quantile smoothers. This is equivalent to setting bandwidth 2 for the original age data.

In Figure 3 the original observations are displayed as small stars. The local 0.5 and 0.9 quantiles at the integer points of age are shown as dashed lines, whereas the corresponding nonparametric quantile smoothers are displayed as solid lines with corresponding 95% uniform confidence bands shown as dashed-dot lines. A closer look reveals a quadratic relation between age and logged hourly real wages. We use several popular parametric methods to estimate the 0.5 and 0.9 conditional quantiles, e.g., quadratic, quartic, and set of dummies (a dummy variable for each 5-year age group) models; the results are displayed in Figure 4. With the help of the 95% uniform confidence bands, we can conduct the parametric model specification test. At the 5% significance level, we could not reject any model. However, when the confidence level further decreases and the uniform confidence bands get narrower, the "set of dummies" parametric model will be the first one to be rejected. At the 10% significance level, the set of dummies (for age groups) model is rejected whereas the other two are not. As the quadratic model performs quite similarly to the quartic one, for simplicity it is suggested in practice to measure the log(wage)-earning relation in mean regression, which coincides with the approach of Murphy and Welch (1990).

REFERENCES

Bickel, P. & M. Rosenblatt (1973) On some global measures of the deviation of density function estimatiors. *Annals of Statistics* 1, 1071–1095.

Cai, Z.W. (2002) Regression quantiles for time series. *Econometric Theory* 18, 169–192.

Csörgö, S. & P. Hall (1982) Upper and lower classes for triangular arrays. *Zeitschrift für Wahrscheinlichkeitstheorie und verwandte Gebiete* 61, 207–222.

Fan, J., T.C. Hu, & Y.K. Troung (1994) Robust nonparametric function estimation. *Scandinavian Journal of Statistics* 21, 433–446.

Fan, J., Q. Yao, & H. Tong (1996) Estimation of conditional densities and sensitivity measures in nonlinear dynamical systems. *Biometrika* 83, 189–206.

Ferguson, T.S. (1967) *Mathematical Statistics: A Decision Theoretic Approach.* Academic Press.

Franke, J. & P. Mwita (2003) Nonparametric Estimates for Conditional Quantiles of Time Series. Report in Wirtschaftsmathematik 87, University of Kaiserslautern.

Hall, P., R. Wolff, & Q. Yao (1999) Methods for estimating a conditional distribution function. *Journal of the American Statistical Association* 94, 154–163.

Härdle, W. (1989) Asymptotic maximal deviation of M-smoothers. *Journal of Multivariate Analysis* 29, 163–179.

Härdle, W., P. Janssen & R. Serfling (1988) Strong uniform consistency rates for estimators of conditional functionals. *Annals of Statistics* 16, 1428–1429.

Härdle, W. & S. Luckhaus (1984) Uniform consistency of a class of regression function estimators. *Annals of Statistics* 12, 612–623.

Huber, P. (1981) *Robust Statistics*. Wiley.

Jeong, K. & W. Härdle. (2008) A Consistent Nonparametric Test for Causality in Quantile. SFB 649 Discussion Paper.

Johnston, G. (1982) Probabilities of maximal deviations of nonparametric regression function estimates. *Journal of Multivariate Analysis* 12, 402–414.

Koenker, R. & G.W. Bassett (1978) Regression quantiles. *Econometrica* 46, 33–50.

Koenker, R. & K.F. Hallock (2001) Quantile regression. *Journal of Econometric Perspectives* 15, 143–156.

Koenker, R. & B.J. Park (1996) An interior point algorithm for nonlinear quantile regression. *Journal of Econometrics* 71, 265–283.

Kong, E., O. Linton, & Y. Xia (2010) Uniform Bahadur representation for local polynomial estimates of M-regression and its application to the additive model. *Econometric Theory*, forthcoming.

Lejeune, M.G. & P. Sarda (1988) Quantile regression: A nonparametric approach. *Computational Statistics and Data Analysis* 6, 229–239.

Murphy, K. & F. Welch (1990) Empirical age-earnings profiles. *Journal of Labor Economics* 8, 202–229.

Parzen, M. (1962) On estimation of a probability density function and mode. *Annals of Mathematical Statistics* 32, 1065–1076.

Portnoy, S. & R. Koenker (1997) The Gaussian hare and the Laplacian tortoise: Computability of squared-error versus absolute-error estimators (with discussion). *Statistical Sciences* 12, 279–300.

Rosenblatt, M. (1952) Remarks on a multivariate transformation. *Annals of Mathematical Statistics* 23, 470–472.

Ruppert, D., S.J. Sheather, & M.P. Wand (1995) An effective bandwidth selector for local least squares regression. *Journal of the American Statistical Association* 90, 1257–1270.

Tusnady, G. (1977) A remark on the approximation of the sample distribution function in the multidimensional case. *Periodica Mathematica Hungarica* 8, 53–55.

Yu, K. & M.C. Jones (1997) A comparison of local constant and local linear regression quantile estimation. *Computational Statistics and Data Analysis* 25, 159–166.

Yu, K. & M.C. Jones (1998) Local linear quantile regression. *Journal of the American Statistical Association* 93, 228–237.

Yu, K., Z. Lu, & J. Stander (2003) Quantile regression: Applications and current research areas. *Journal of the Royal Statistical Society, Series D* 52, 331–350.

APPENDIX

Proof of Theorem 2.1 . By the definition of $l_n(x)$ as a zero of (2), we have, for $\varepsilon > 0$,

$$\text{if} \quad l_n(x) > l(x) + \varepsilon, \quad \text{then} \quad \widetilde{H}_n\{l(x) + \varepsilon, x\} > 0. \tag{A.1}$$

Now

$$\widetilde{H}_n\{l(x) + \varepsilon, x\} \leqslant \widetilde{H}\{l(x) + \varepsilon, x\} + \sup_{\theta \in I} \left| \widetilde{H}_n(\theta, x) - \widetilde{H}(\theta, x) \right|. \tag{A.2}$$

Also, by the identity $\widetilde{H}\{l(x), x\} = 0$, the function $\widetilde{H}\{l(x)+\varepsilon, x\}$ is not positive and has a magnitude $\geqslant m_1\tilde{q}\varepsilon$ by Assumption (A6) and (6), for $0 < \varepsilon < \delta_1$. That is, for $0 < \varepsilon < \delta_1$,

$$\widetilde{H}\{l(x)+\varepsilon, x\} \leqslant -m_1\tilde{q}\varepsilon. \tag{A.3}$$

Combining (A.1)–(A.3), we have, for $0 < \varepsilon < \delta_1$,

$$\text{if} \quad l_n(x) > l(x)+\varepsilon, \quad \text{then} \quad \sup_{\theta\in I}\sup_{x\in J}\left|\widetilde{H}_n(\theta, x) - \widetilde{H}(\theta, x)\right| > m_1\tilde{q}\varepsilon.$$

With a similar inequality proved for the case $l_n(x) < l(x)+\varepsilon$, we obtain, for $0 < \varepsilon < \delta_1$,

$$\text{if} \quad \sup_{x\in J}|l_n(x) - l(x)| > \varepsilon, \quad \text{then} \quad \sup_{\theta\in I}\sup_{x\in J}\left|\widetilde{H}_n(\theta, x) - \widetilde{H}(\theta, x)\right| > m_1\tilde{q}\varepsilon. \tag{A.4}$$

It readily follows that (A.4) and (5) imply (7). ■

Subsequently we first show that $\|R_n\|_\infty = \sup_{t\in J}|R_n(t)|$ vanishes asymptotically faster than the rate $(nh\log n)^{-1/2}$; for simplicity we will just use $\|\cdot\|$ to indicate the sup-norm.

LEMMA A.1. *For the remainder term $R_n(t)$ defined in (9) we have*

$$\|R_n\| = \mathcal{O}_p\{(nh\log n)^{-1/2}\}. \tag{A.5}$$

Proof. First we have by the positivity of the kernel K,

$$\|R_n\| \leqslant \Big[\inf_{0\leqslant t\leqslant 1}\{|D_n(t)|\cdot q(t)\}\Big]^{-1}\{\|H_n\|\cdot\|q - D_n\| + \|D_n\|\cdot\|\mathsf{E}\, H_n\|\}$$

$$+ C_1\cdot\|l_n - l\|^2\cdot\Big\{\inf_{0\leqslant t\leqslant 1}|D_n(t)|\Big\}^{-1}\cdot\|f_n\|_\infty,$$

where $f_n(x) = (nh)^{-1}\sum_{i=1}^n K\{(x - X_i)/h\}$.

The desired result, Lemma A.1, will then follow if we prove

$$\|H_n\| = \mathcal{O}_p\Big\{(nh)^{-1/2}(\log n)^{1/2}\Big\}, \tag{A.6}$$

$$\|q - D_n\| = \mathcal{O}_p\Big\{(nh)^{-1/4}(\log n)^{-1/2}\Big\}, \tag{A.7}$$

$$\|\mathsf{E}\, H_n\| = \mathcal{O}\big(h^2\big), \tag{A.8}$$

$$\|l_n - l\|^2 = \mathcal{O}_p\Big\{(nh)^{-1/2}(\log n)^{-1/2}\Big\}. \tag{A.9}$$

Because (A.8) follows from the well-known bias calculation

$$\mathsf{E}\, H_n(t) = h^{-1}\int K\{(t-u)/h\}\,\mathsf{E}[\psi\{y - l(t)\}|X = u]f_X(u)\,du = \mathcal{O}(h^2),$$

where $\mathcal{O}(h^2)$ is independent of t in Parzen (1962), we have from Assumption (A2) that $\|\mathsf{E}\, H_n\| = \mathcal{O}_p\{(nh)^{-1/2}(\log n)^{-1/2}\}$.

According to Lemma A.3 in Franke and Mwita (2003),

$$\sup_{t\in J}|H_n(t) - \mathsf{E}\, H_n(t)| = \mathcal{O}\Big\{(nh)^{-1/2}(\log n)^{1/2}\Big\}$$

and the following inequality

$$\begin{aligned}\|H_n\| &\leqslant \|H_n - \mathsf{E}\, H_n\| + \|\mathsf{E}\, H_n\| \\ &= \mathcal{O}\left\{(nh)^{-1/2}(\log n)^{1/2}\right\} + \mathcal{O}_p\left\{(nh)^{-1/2}(\log n)^{-1/2}\right\} \\ &= \mathcal{O}\left\{(nh)^{-1/2}(\log n)^{1/2}\right\},\end{aligned}$$

statement (A.6) thus is obtained.

Statement (A.7) follows in the same way as (A.6) using Assumption (A2) and the Lipschitz continuity properties of K, ψ', l.

According to the uniform consistency of $l_n(t) - l(t)$ shown before, we have

$$\|l_n - l\| = \mathcal{O}_p\{(nh)^{-1/2}(\log n)^{1/2}\},$$

which implies (A.9).

Now the assertion of the lemma follows, because by tightness of $D_n(t)$, $\inf_{0\leqslant t\leqslant 1} |D_n(t)| \geqslant q_0$ a.s. and thus

$$\|R_n\| = \mathcal{O}_p\{(nh\log n)^{-1/2}\}(1 + \|f_n\|).$$

Finally, by Theorem 3.1 of Bickel and Rosenblatt (1973), $\|f_n\| = \mathcal{O}_p(1)$; thus the desired result $\|R_n\| = \mathcal{O}_p\{(nh\log n)^{-1/2}\}$ follows. ■

We now begin with the subsequent approximations of the processes $Y_{0,n}$–$Y_{5,n}$.

LEMMA A.2.

$$\|Y_{0,n} - Y_{1,n}\| = \mathcal{O}\left\{(nh)^{-1/2}(\log n)^2\right\} \quad a.s.$$

Proof. Let t be fixed and put $L(y) = \psi\{y - l(t)\}$ still depending on t. Using integration by parts, we obtain

$$\begin{aligned}&\iint_{\Gamma_n} L(y)K\{(t-x)/h\}\,dZ_n(x,y) \\ &\quad= \int_{u=-A}^{A}\int_{y=-a_n}^{a_n} L(y)K(u)\,dZ_n(t - h\cdot u, y) \\ &\quad= -\int_{-A}^{A}\int_{-a_n}^{a_n} Z_n(t - h\cdot u, y)\,d\{L(y)K(u)\} \\ &\qquad + L(a_n)(a_n)\int_{-A}^{A} Z_n(t - h\cdot u, a_n)\,dK(u) \\ &\qquad - L(-a_n)(-a_n)\int_{-A}^{A} Z_n(t - h\cdot u, -a_n)\,dK(u) \\ &\qquad + K(A)\left\{\int_{-a_n}^{a_n} Z_n(t - h\cdot A, y)\,dL(y)\right. \\ &\qquad\qquad \left. + L(a_n)(a_n)Z_{n_a}(t - h\cdot A, a_n) - L(-a_n)(-a_n)Z_n(t - h\cdot A, -a_n)\right\}\end{aligned}$$

$$- K(-A)\left\{\int_{-a_n}^{a_n} Z_n(t+h\cdot A, y)\,dL(y) + L(a_n)(a_n)Z_n(t+h\cdot A, a_n)\right.$$
$$\left. - L(-a_n)(-a_n)Z_n(t+h\cdot A, -a_n)\right\}.$$

If we apply the same operation to $Y_{1,n}$ with $B_n\{T(x,y)\}$ instead of $Z_n(x,y)$ and use Lemma 2.2, we finally obtain

$$\sup_{0\leqslant t\leqslant 1} h^{1/2} g(t)^{1/2} |Y_{0,n}(t) - Y_{1,n}(t)| = \mathcal{O}\left\{n^{-1/2}(\log n)^2\right\} \quad \text{a.s.} \qquad \blacksquare$$

LEMMA A.3. $\|Y_{1,n} - Y_{2,n}\| = \mathcal{O}_p(h^{1/2})$.

Proof. Note that the Jacobian of $T(x,y)$ is $f(x,y)$. Hence

$$Y_{1,n}(t) - Y_{2,n}(t) = \left|\{g(t)h\}^{-1/2} \iint_{\Gamma_n} \psi\{y - l(t)\} K\{(t-x)/h\} f(x,y)\,dx\,dy\right| \cdot |W_n(1,1)|.$$

It follows that

$$h^{-1/2}\|Y_{1,n} - Y_{2,n}\| \leqslant |W_n(1,1)| \cdot \left\|g^{-1/2}\right\|$$
$$\cdot \sup_{0\leqslant t\leqslant 1} h^{-1} \iint_{\Gamma_n} |\psi\{y - l(t)\} K\{(t-x)/h\}|\, f(x,y)\,dx\,dy.$$

Because $\|g^{-1/2}\|$ is bounded by assumption, we have

$$h^{-1/2}\|Y_{1,n} - Y_{2,n}\| \leqslant |W_n(1,1)| \cdot C_4 \cdot h^{-1} \int K\{(t-x)/h\}\,dx = \mathcal{O}_p(1). \qquad \blacksquare$$

LEMMA A.4. $\|Y_{2,n} - Y_{3,n}\| = \mathcal{O}_p(h^{1/2})$.

Proof. The difference $|Y_{2,n}(t) - Y_{3,n}(t)|$ may be written as

$$\left|\{g(t)h\}^{-1/2} \iint_{\Gamma_n} [\psi\{y - l(t)\} - \psi\{y - l(x)\}] K\{(t-x)/h\}\,dW_n\{T(x,y)\}\right|.$$

If we use the fact that l is uniformly continuous, this is smaller than

$$h^{-1/2}|g(t)|^{-1/2} \cdot \mathcal{O}_p(h),$$

and the lemma thus follows. $\blacksquare$

LEMMA A.5. $\|Y_{4,n} - Y_{5,n}\| = \mathcal{O}_p(h^{1/2})$.

Proof.

$$|Y_{4,n}(t) - Y_{5,n}(t)| = h^{-1/2}\left|\int \left[\left\{\frac{g(x)}{g(t)}\right\}^{1/2} - 1\right] K\{(t-x)/h\}\,dW(x)\right|$$
$$\leqslant h^{-1/2}\left|\int_{-A}^{A} W(t-hu)\frac{\partial}{\partial u}\left[\left\{\frac{g(t-hu)}{g(t)}\right\}^{1/2} - 1\right] K(u)\,du\right|$$

$$+h^{-1/2}\left|K(A)W(t-hA)\left[\left\{\frac{g(t-Ah)}{g(t)}\right\}^{1/2}-1\right]\right|$$

$$+h^{-1/2}\left|K(-A)W(t+hA)\left[\left\{\frac{g(t+Ah)}{g(t)}\right\}^{1/2}-1\right]\right|$$

$$S_{1,n}(t)+S_{2,n}(t)+S_{3,n}(t), \quad \text{say.}$$

The second term can be estimated by

$$h^{-1/2}\|S_{2,n}\| \leqslant K(A)\cdot \sup_{0\leqslant t\leqslant 1}|W(t-Ah)|\cdot \sup_{0\leqslant t\leqslant 1} h^{-1}\left|\left[\left\{\frac{g(t-Ah)}{g(t)}\right\}^{1/2}-1\right]\right|.$$

By the mean value theorem it follows that

$$h^{-1/2}\|S_{2,n}\| = \mathcal{O}_p(1).$$

The first term $S_{1,n}$ is estimated as

$$h^{-1/2}S_{1,n}(t) = \left|h^{-1}\int_{-A}^{A} W(t-uh)K'(u)\left[\left\{\frac{g(t-uh)}{g(t)}\right\}^{1/2}-1\right]du\right.$$

$$\left.\cdot\frac{1}{2}\int_{-A}^{A} W(t-uh)K(u)\left\{\frac{g(t-uh)}{g(t)}\right\}^{1/2}\left\{\frac{g'(t-uh)}{g(t)}\right\}du\right|$$

$$= |T_{1,n}(t)-T_{2,n}(t)|, \quad \text{say;}$$

$\|T_{2,n}\| \leqslant C_5\cdot\int_{-A}^{A}|W(t-hu)|du = \mathcal{O}_p(1)$ by assumption on $g(t)=\sigma^2(t)\cdot f_X(t)$. To estimate $T_{1,n}$ we again use the mean value theorem to conclude that

$$\sup_{0\leqslant t\leqslant 1} h^{-1}\left|\left\{\frac{g(t-uh)}{g(t)}\right\}^{1/2}-1\right| < C_6\cdot|u|;$$

hence

$$\|T_{1,n}\| \leqslant C_6\cdot\sup_{0\leqslant t\leqslant 1}\int_{-A}^{A}|W(t-hu)|K'(u)u/du = \mathcal{O}_p(1).$$

Because $S_{3,n}(t)$ is estimated as $S_{2,n}(t)$, we finally obtain the desired result. ■

The next lemma shows that the truncation introduced through $\{a_n\}$ does not affect the limiting distribution.

LEMMA A.6. $\|Y_n - Y_{0,n}\| = \mathcal{O}_p\{(\log n)^{-1/2}\}$.

Proof. We shall only show that $g'(t)^{-1/2}h^{-1/2}\iint_{\mathbb{R}-\Gamma_n}\psi\{y-l(t)\}K\{(t-x)/h\}dZ_n(x,y)$ fulfills the lemma. The replacement of $g'(t)$ by $g(t)$ may be proved as in Lemma A.4 of Johnston (1982). The preceding quantity is less than $h^{-1/2}\|g^{-1/2}\|\cdot\|\iint_{\{|y|>a_n\}}\psi\{y-$

$l(\cdot)\}K\{(\cdot - x)/h\}dZ(x, y)\|$. It remains to be shown that the last factor tends to zero at a rate $\mathcal{O}_p\{(\log n)^{-1/2}\}$. We show first that

$$V_n(t) = (\log n)^{1/2} h^{-1/2} \iint_{\{|y|>a_n\}} \psi\{y - l(t)\}K\{(t-x)/h\}\,dZ_n(x, y)$$

$$\xrightarrow{p} 0 \quad \text{for all } t,$$

and then we show tightness of $V_n(t)$. The result then follows:

$$V_n(t) = (\log n)^{1/2}(nh)^{-1/2} \sum_{i=1}^{n} [\psi\{Y_i - l(t)\}\mathbf{1}(|Y_i| > a_n)K\{(t - X_i)/h\}$$

$$- \mathsf{E}\,\psi\{Y_i - l(t)\}\mathbf{1}(|Y_i| > a_n)K\{(t - X_i)/h\}]$$

$$= \sum_{i=1}^{n} X_{n,t}(t),$$

where $\{X_{n,t}(t)\}_{i=1}^{n}$ are i.i.d. for each n with $\mathsf{E}\,X_{n,t}(t) = 0$ for all $t \in [0, 1]$. We then have

$$\mathsf{E}\,X_{n,t}^2(t) \leqslant (\log n)(nh)^{-1}\,\mathsf{E}\,\psi^2\{Y_i - l(t)\}\mathbf{1}(|Y_i| > a_n)K^2\{(t - X_i)/h\}$$

$$\leqslant \sup_{-A\leqslant u\leqslant A} K^2(u) \cdot (\log n)(nh)^{-1}\,\mathsf{E}\,\psi^2\{Y_i - l(t)\}\mathbf{1}(|Y_i| > a_n).$$

Hence

$$\mathrm{Var}\{V_n(t)\} = \mathsf{E}\left\{\sum_{i=1}^{n} X_{n,t}(t)\right\}^2 = n \cdot \mathsf{E}\,X_{n,t}^2(t)$$

$$\leqslant \sup_{-A\leqslant u\leqslant A} K^2(u) h^{-1}(\log n) \int_{\{|y|>a_n\}} f_y(y)\,dy \cdot M_\psi,$$

where M_ψ denotes an upper bound for ψ^2. This term tends to zero by Assumption (A3). Thus by Markov's inequality we conclude that

$$V_n(t) \xrightarrow{p} 0 \quad \text{for all } t \in [0, 1].$$

To prove tightness of $\{V_n(t)\}$ we refer again to the following moment condition as stated in Lemma A.1:

$$\mathsf{E}\{|V_n(t) - V_n(t_1)| \cdot |V_n(t_2) - V_n(t)|\} \leqslant C' \cdot (t_2 - t_1)^2$$

C' denoting a constant, $\quad t \in [t_1, t_2]$.

We again estimate the left-hand side by Schwarz's inequality and estimate each factor separately:

$$\mathsf{E}\{V_n(t) - V_n(t_1)\}^2 = (\log n)(nh)^{-1}\,\mathsf{E}\left[\sum_{i=1}^{n} \Psi_n(t, t_1, X_i, Y_i) \cdot \mathbf{1}(|Y_i| > a_n)\right.$$

$$\left. - \mathsf{E}\{\Psi_n(t, t_1, X_i, Y_i) \cdot \mathbf{1}(|Y_i| > a_n)\}\right]^2,$$

where $\Psi_n(t, t_1, X_i, Y_i) = \psi\{Y_i - l(t)\}K\{(t - X_i)/h\} - \psi\{Y_i - l(t_1)\}K\{(t_1 - X_1)/h\}$. Because ψ, K are Lipschitz continuous except at one point and the expectation is taken afterward, it follows that

$$[\mathsf{E}\{V_n(t) - V_n(t_1)\}^2]^{1/2}$$

$$\leqslant C_7 \cdot (\log n)^{1/2} h^{-3/2} |t - t_1| \cdot \left\{ \int_{\{|y| > a_n\}} f_y(y)\, dy \right\}^{1/2}.$$

If we apply the same estimation to $V_n(t_2) - V_n(t_1)$ we finally have

$$\mathsf{E}\{|V_n(t) - V_n(t_1)| \cdot |V_n(t_2) - V_n(t)|\}$$

$$\leqslant C_7^2 (\log n) h^{-3} |t - t_1||t_2 - t| \times \int_{\{|y| > a_n\}} f_y(y)\, dy$$

$$\leqslant C' \cdot |t_2 - t_1|^2 \quad \text{because } t \in [t_1, t_2] \quad \text{by Assumption (A3).}$$ ■

LEMMA A.7. *Let $\lambda(K) = \int K^2(u)\, du$ and let $\{d_n\}$ be as in Theorem 2.2. Then*

$$(2\delta \log n)^{1/2} [\|Y_{3,n}\| / \{\lambda(K)\}^{1/2} - d_n]$$

has the same asymptotic distribution as

$$(2\delta \log n)^{1/2} [\|Y_{4,n}\| / \{\lambda(K)\}^{1/2} - d_n].$$

Proof. $Y_{3,n}(t)$ is a Gaussian process with

$$\mathsf{E}\, Y_{3,n}(t) = 0$$

and covariance function

$$r_3(t_1, t_2) = \mathsf{E}\, Y_{3,n}(t_1) Y_{3,n}(t_2)$$

$$= \{g(t_1) g(t_2)\}^{-1/2} h^{-1} \iint_{\Gamma_n} \psi^2\{y - l(x)\} K\{(t_1 - x)/h\}$$

$$\times K\{(t_2 - x)/h\} f(x, y)\, dx\, dy$$

$$= \{g(t_1) g(t_2)\}^{-1/2} h^{-1} \iint_{\Gamma_n} \psi^2\{y - l(x)\} f(y|x)\, dy K\{(t_1 - x)/h\}$$

$$\times K\{(t_2 - x)/h\} f_X(x)\, dx$$

$$= \{g(t_1) g(t_2)\}^{-1/2} h^{-1} \int g(x) K\{(t_1 - x)/h\} K\{(t_2 - x)/h\}\, dx$$

$$= r_4(t_1, t_2),$$

where $r_4(t_1, t_2)$ is the covariance function of the Gaussian process $Y_{4,n}(t)$, which proves the lemma. ■

Statistica Sinica **20** (2010), 771-785

INVESTORS' PREFERENCE: ESTIMATING AND DEMIXING OF THE WEIGHT FUNCTION IN SEMIPARAMETRIC MODELS FOR BIASED SAMPLES

Ya'acov Ritov and Wolfgang K. Härdle

The Hebrew University of Jerusalem and Humboldt-Universität zu Berlin

Abstract: We consider a semiparametric model for the weight function in a biased sample model. The object of our interest parametrizes the weight function, and it is non-Euclidean. The model discussed is motivated by the estimation of the mixing distribution of individual utility functions in the DAX market. We discuss the estimation rate of different functionals of the weight functions.

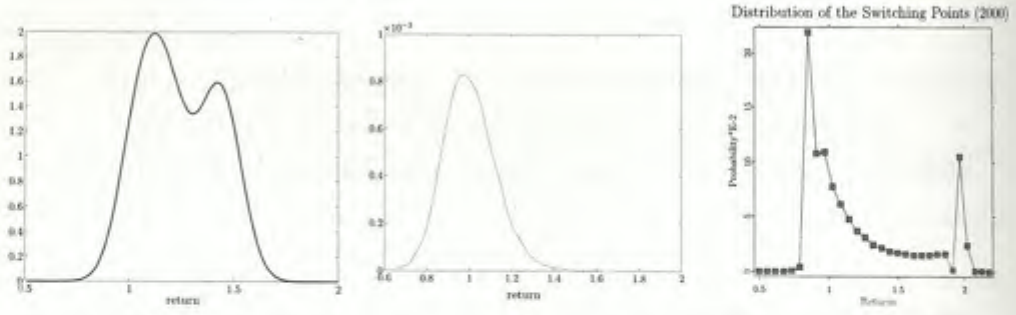
Key words and phrases: Empirical pricing kernel, exponential mixture, inverse problem, mixture distribution, risk aversion.

1. Introduction

A sample $X_1, \ldots, X_n$ is considered biased if it is sampled from a density p which is represented as

$$p(x) = \frac{q(x)w(x)}{\int q(u)w(u)du}. \tag{1.1}$$

Here q is some 'natural' pdf (probability density function) for the problem, representing the 'true' underlying distribution, while w is a given weight function that biases the sample. In a standard example, X represents the severity of the disease, and q is the density of X among patients at admission to the hospital. However, it may be more convenient to take a random sample from the population of patients who are in the hospital at a given time. If the time of hospitalization is proportional to the severity of the case, then the sample is taken from the density p, which is equal to q 'length biased' with $w(x) \equiv x$. Vardi (1985) was the first to systematically analyze these models; asymptotic theory was developed in Gill, Vardi and Wellner (1988); Gilbert, Lele and Vardi (1999) extended the model to the situation where the weight function depends on some parameter, $w(x) = w(x; f)$; the large sample properties were discussed in Gilbert (2000). Equation (1.1) has some similarities to the classical choice-based sample problem, Manski and Lerman (1977), or retrospective case-control studies, Mantel (1973). In fact one can consider the situation as if one has an infinite



(a) (b) (c)

Figure 1. The DAX data, 24/03/2000 half a year look ahead: (a) p, the historical density; (b) q, the risk neutral density; (c) The estimate of f, the mixing density. Figures are taken from DHM.

sample from the control group, and hence q is known, and a finite sample from the control, the biased sample. The likelihood ratio between the two is the given $w(x; f)$. The main difficulty we face in this paper is the particular form of $w(x; f)$ we have.

Technically speaking, our paper is about estimating f, the parameter of the weight function, $w(x) = w(x; f)$. In the model we consider, q is taken as known, while the weight function is parametrized by a non-Euclidean parameter. This brings us to an inverse problem of estimating and demixing the weight function.

In subject matter, our model is motivated by the research on risk aversion and proclivity, and more precisely on the empirical pricing kernel (EPK), see Detlefsen, Härdle and Moro (2007) (hereafter DHM). The EPK describes the apparent utility behavior as function of the individual investors utility function. In this model q is the risk neutral density of asset pricing, and is derived from theoretical considerations. The density p on the other hand is the density of the empirical (historical) prices. See parts (a) and (b) of Figure 1 for an example. In asset pricing the EPK links a risk neutral investor's behavior to individual utilities, which gives in our notation a semiparametric modeling of the weight function w. The integral function of the pricing kernel q/p is the utility function used by a representing individual. Knowing p and q yields the exact form of the utility function, cf. Ait-Sahalia and Lo (2000), and Rosenberg and Engle (2002). The risk neutral (state price) density (SPD) q can be calculated from market data on European options. There are more than 5,000 observations each day for maturity from one week to two years. The SPD can therefore be estimated very precisely. Much empirical research work has demonstrated the so called EPK paradox: the resulting utility function is partially concave and partially convex, more precisely of the Friedman and Savage type, Friedman and Savage (1948).

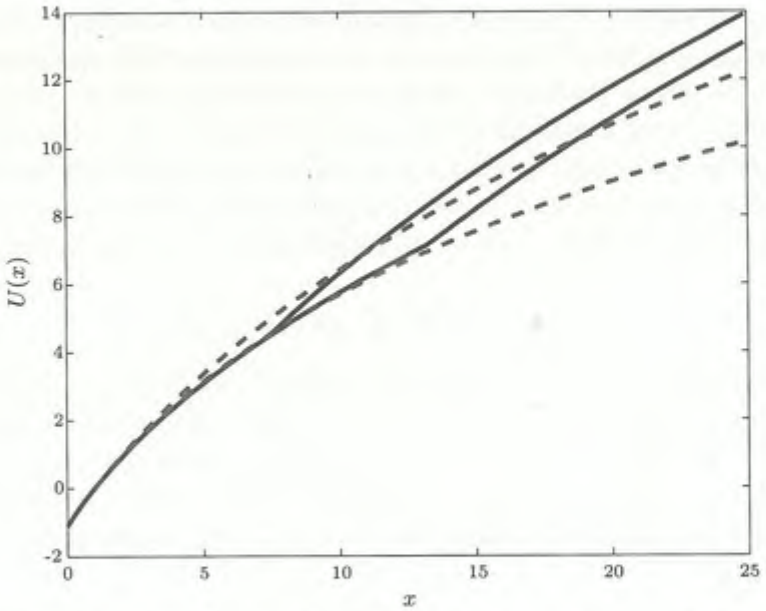


Figure 2. The utility function $U(\cdot;\xi)$ of (3.5) ($\alpha_1 = 2$, $\alpha_2 = 2.25$, $c = 2$) for two different values of ξ (solid lines), and of (3.8) for two values broken lines.

This so called risk aversion puzzle has also been recently discussed in Chabi-Yo, Garcia and Renault (2008); a recursive utility approach to dynamic pricing kernel estimation is published in Gallant and Hong (2007); a fundamental reference on asset pricing theory is the book by Cochrane (2005).

It is assumed in DHM that the observed density of the DAX value has density of the form $p(x) = cq(x)w(x;f)$, where $q \in \{q_\nu, \nu \in N \subseteq \mathbb{R}^d\}$ is the theoretical derived risk neutral density, assumed to follow a given parametric function, and c is a normalization factor, that is, of the type (1.1). The weight function is theoretically derived as

$$w(x;f) = \frac{1}{U'}(x), \tag{1.2}$$

where U is the market utility function, and prime denotes derivative. The market utility is estimated for option data and available historical data, and it also showed the risk aversion puzzle for the DAX stock market. In DHM an aggregation mechanism was proposed that similarly to Chabi-Yo, Garcia and Renault (2008) uses a switching point ξ. This point characterizes the investors switch from a bearish (low return) to a bullish (high return) risk aversion pattern. A graph of two different utility functions $u(\cdot;\xi)$ with switching points $\xi_1 < \xi_2$ is presented in Figure 2.

Simply averaging the utilities is not possible since utilities for different investors are incomparable. One therefore specifies first a utility level u and aggregates the outlooks on the returns R_i with $u = U(R_i; \xi_i)$, $i = 1, 2, \ldots$. The aggregate estimator of the switching return equals average$\{U^{-1}(u, \xi_i), i = 1, 2, \ldots\}$ if all investors have the same market power. Denoting the investors inverse utility function by g and assuming a distribution of switching points, the market utility function U_f is itself assumed to be a function of the mixture of the individual investors:

$$x = U_f^{-1}(u) = \int_{\Xi} g(u; \xi) f(\xi) d\xi. \tag{1.3}$$

Here $\xi \in \Xi$ denotes an investor type, f is the density of the investors' distribution, and $\{g(\cdot; \xi) : \xi \in \Xi\}$ is the (known) class of possible inverse utility functions of the different investors. A subject of type ξ has the inverse utility function $g(\cdot; \xi)$ or, equivalently, he has the utility function $u(\cdot; \xi)$ satisfying $g\{u(x; \xi); \xi\} \equiv x$. The problem we consider is finding the density f. We obtain from (1.1)–(1.3) the representation:

$$p(x) = cq(x) \int \frac{\partial}{\partial u} g(u; \xi) f(\xi) d\xi,$$

where u solves

$$x = \int g(u; \xi) f(\xi) d\xi. \tag{1.4}$$

See Figure 1 for an example taken from DHM of estimates of p, q, and f. See also Figure 2 for an example of $g^{-1}(\cdot; \xi)$.

Aggregation problem (1.3) is a way of aggregating preferences that is not based on the equilibrium theory usually associated with Walras (1874). The situation considered here is of a different type and is hypothetical when applied to real markets. The DAX market data were mentioned as suitable for testing the disaggregation techniques described in the paper.

Aggregation procedure (1.3) relates to the situation where the price of an asset is obtained as the result of a survey of investors (or experts) before they made trades. Thus, this price should be considered as a forecast for the next period, not a reflection of the struggle for limited resources in the market between investors with different preferences and endowments.

The survey proceeds as following. Each market participant is asked what the price will be if the conditions in the market are, for example, extremely good. Extremely good corresponds to some utility level $\tilde{u}_1$ in the minds of investors. In this way all investors agree that they are discussing an economic situation with the same utility level. As the next step, each investor forms his forecast about how high the prices would be in such a situation. Those forecasted prices are recorded and averaged to produce an aggregate opinion of all market participants

(or experts). If the investors have equal market power, their individual opinions will be averaged with equal weights. The forecast for different economic situations corresponding to other utility levels is formed in a similar way.

To sum up, (1.3) describes a mechanism for forming a forecast about future prices. It gives an idea of which opinions prevailed in a group of investors or experts that was able to predict prices correctly before trading, for example if they were more optimistic or pessimistic investors (experts), and to what degree.

In this paper we investigate the estimation of the non-Euclidean parameter f of a few utility functions. The result is typical for inverse problems, in that slightly different assumption yield completely different results. In fact, we present three similar models, similar to those investigated in DHM, that exhibit these behaviors:

(i) there is no consistent estimator of f;

(ii) f can be estimated at a regular nonparametric rate of $n^{-\alpha}$;

(iii) f can be estimated, but at a very slow rate.

Interestingly, there is a a sort of uncertainty principle: the better we can estimate the function $U^{-1}(u)$, the worse we can demix it and estimate f. This is not unexpected. We cannot estimate f well when large differences in f have only minor impact on $\int g(\cdot;\xi)f(\xi)d\xi$.

The structure of the rest of the paper is as follows. In Section 2, we suggest an algorithm for calculating the generalized maximum-likelihood estimator (GMLE) for the semiparametric weight function of the model suggested by DHM. Rates of convergence of the demixing estimator for the DHM's model are discussed in Section 3, as well as of estimates of the mixture itself.

2. EPK: Model and an EM estimator

We consider the EPK problem. We start from (1.4) and we assume that q is known. In practice, it is assumed only to belong to some parametric family $\{q_\nu\}$. However, we deal in the following with rates that are much slower than the parametric $\sqrt{n}$ rate, and the estimate of ν is based on a much larger sample than the estimates of the rest of the parameters. Therefore, the assumption that ν is known considerably simplifies the discussion without impacting the results.

Rewrite (1.4) as

$$p\left\{\int g(u;\xi)f(\xi)d\mu(\xi)\right\}\int \frac{\partial}{\partial u}g(u;\xi)f(\xi)d\mu(\xi)$$
$$= cq\left\{\int g(u;\xi)f(\xi)d\mu(\xi)\right\}\left\{\int \frac{\partial}{\partial u}g(u;\xi)f(\xi)d\mu(\xi)\right\}^2, \qquad (2.1)$$

where μ is some dominating measure (e.g., Lebesgue or the counting measure). Noting that the LHS of (2.1) integrates to 1, c can be found to yield

$$p\left\{\int g(u;\xi)f(\xi)d\mu(\xi)\right\} = \frac{q\{\int g(u;\xi)f(\xi)d\mu(\xi)\}\int \frac{\partial}{\partial u}g(u;\xi)f(\xi)d\mu(\xi)}{\int q\{\int g(v;\xi)f(\xi)d\mu(\xi)\}\{\int \frac{\partial}{\partial u}g(v;\xi)f(\xi)d\mu(\xi)\}^2 dv}.$$

The market utility $U(x) = U(x;f)$ is given by

$$x \equiv \int g\Big\{U(x;f);\xi\Big\}f(\xi)d\mu(\xi) \equiv \psi_f\Big\{U(x;f)\Big\}.$$

We obtain

$$p(x) = \frac{q(x)\int \frac{\partial}{\partial u}g(U(x;f);\xi)f(\xi)d\mu(\xi)}{\int q(y)\int \frac{\partial}{\partial u}g(U(y;f);\xi)f(\xi)d\mu(\xi)dy} = \frac{q(x)\psi_f'\{\psi_f^{-1}(x)\}}{\int q(y)\psi_f'\{\psi_f^{-1}(y)\}dy}. \tag{2.2}$$

The statistical model assumed by DHM is that we obtain a simple random sample from p, where p is parametrized in (2.2) by the non-Euclidean parameter f. A natural approach is to estimate f by the MLE or a variant of it, which we develop now. Note that $\nabla_f\psi_f(u) = g(u;\cdot)$, and by taking the gradient of $x \equiv \int g\{\psi_f^{-1}(x);\xi\}f(\xi)d\mu(\xi)$ we obtain

$$0 = g\{\psi_f^{-1}(x);\cdot\} + \psi_f'\{\psi_f^{-1}(x)\}\nabla_f\psi_f^{-1}(x).$$

The derivative of the log-likelihood is given therefore by

$$\begin{aligned}\dot{\ell}_f(\xi) &= \sum_{i=1}^n \frac{1}{\psi_f'\{\psi_f^{-1}(X_i)\}}\left[\frac{\partial}{\partial u}g\{\psi_f^{-1}(X_i);\xi\} - \frac{\psi_f''}{\psi_f'}\{\psi_f^{-1}(X_i)\}g\{\psi_f^{-1}(X_i);\xi\}\right] \\ &\quad - nA_f(\xi), \\ &= \sum_{i=1}^n \frac{1}{\psi_f'\{U_i\}}\left\{\frac{\partial}{\partial u}g\{U_i;\xi\} - \frac{\psi_f''}{\psi_f'}(U_i)g(U_i;\xi)\right\} - nA_f(\xi),\end{aligned}$$

with $U_i = \psi_f^{-1}(X_i)$, and for all $\xi \in \text{suppf}$, where $A_f(\xi)$ is the mean of the first term under f. Since the density of U_i is given by

$$r_f(u) = p\{\psi_f(u)\}\psi_f'(u) = \frac{q\{\psi_f(u)\}\{\psi_f'(u)\}^2}{\int q\{\psi_f(v)\}\{\psi_f'(v)\}^2 dv},$$

we obtain that

$$A_f(\xi) = \frac{\int q\{\psi_f(u)\}\{\psi_f'(u)\frac{\partial}{\partial u}g(u;\xi) - \psi_f''(u)g(u;\xi)\}du}{\int q\{\psi_f(v)\}\{\psi_f'(v)\}^2 dv}.$$

We discusse now how a GMLE can be constructed, and suggest a pseudo-EM algorithm, that is justified as being the limiting result of proper EM algorithms

applied in approximate models. To be clear, the approximation introduced in the following is needed only as a justification for an algorithm applied to the formal model. The algorithm itself is "exact" and maximizes the exact likelihood. The technical problem we want to circumvent is the exact functional dependency of X_i and U_i which affects the EM. As an intermediate step we weaken the functional dependency into a proper statistical dependency.

The model of a random sample from the density p can be well-approximated as $\sigma \to 0$ by a $X_i = \psi_f(U_i) + \varepsilon_i$, $i = 1, \ldots, n$, where $\varepsilon_1, \ldots, \varepsilon_n$ is a random sample from $N(0, \sigma^2)$ independent from the random sample $U_1, \ldots, U_n$ taken from the density r_f. Now, the log-likelihood of the joint density is given by

$$\ell_f = \sum_{i=1}^{n} \Big[\log q\{\psi_f(U_i)\} + 2 \log\{\psi_f'(U_i)\} \Big] - nC_f - \frac{1}{2\sigma^2} \sum_{i=1}^{n} (X_i - \psi_f(U_i))^2,$$

where $C_f = \log \int q l\{\psi_f(v)\}\{\psi_f'(v)\}^2 dv$. By a well-known formula for the Bayes estimator in the Gaussian measurement error model, here the distribution of $\psi_f(U_i) - X_i$, given X_i, is normal with mean $\sigma^2 f_X'(X_i)/f_X(X_i)$ and second moment $\sigma^4 f_X''(X_i)/f_X(X_i) + \sigma^2$, where f_X is the marginal density of X_i. At the limit as $\sigma^2 \to 0$, the conditional expectation of the log-likelihood, given the X_i's, amounts therefore to replacing U_i by $\psi_f^{-1}(X_i)$. We conclude that the limiting EM algorithm iterates therefore between the following steps.

The E step:

$$U_i \leftarrow \psi_f^{-1}(X_i), \qquad i = 1, \ldots, n, \tag{2.3}$$

The M step:

$$f \leftarrow \operatorname{argmax} \left[\sum_{i=1}^{n} \Big\{ \log q\{\psi_f(U_i)\} + 2 \log\{\psi_f'(U_i)\} \Big\} - nC_f \right].$$

Let $\boldsymbol{U} = (U_1, \ldots, U_n)$, $\boldsymbol{X} = (X_1, \ldots, X_n)$, and denote the E-step by $\boldsymbol{U} = \psi_f^{-1}(\boldsymbol{X})$. The M-step can be accomplished by solving the likelihood equation:

$$0 = \dot{\ell}_f^M(\xi; \boldsymbol{U}) = \sum_{i=1}^{n} \left[\frac{q'\{\psi_f(U_i)\}}{q\{\psi_f(U_i)\}} g(U_i; \xi) + \frac{2}{\psi_f'(U_i)} \frac{\partial}{\partial u} g(U_i, \xi) - \dot{C}_f(\xi) \right], \tag{2.4}$$

for all $\xi \in \operatorname{suppf}$, where

$$\begin{aligned} \dot{C}_f(\xi) &= \frac{\int [(q'\{\psi_f(v)\}/q\{\psi_f(v)\}) g(v; \xi) + (2/\psi_f'(v)) \frac{\partial}{\partial u} g(v, \xi)] q\{\psi_f(v)\}\{\psi_f'(v)\}^2 dv}{\int q\{\psi_f(v)\}\{\psi_f'(v)\}^2 dv} \\ &= \mathrm{E}_f \left[\frac{q'\{\psi_f(U)\}}{q\{\psi_f(U)\}} g(U; \xi) + \frac{2}{\psi_f'(U)} \frac{\partial}{\partial u} g(U, \xi) \right] \\ &= \mathrm{E}_f\{T_f(U; \xi)\}, \quad \text{say.} \end{aligned}$$

However, there is no need in the M-step to find the exact maximizer of the log-likelihood. All that is needed is that the likelihood be strictly increasing (if possible at all) at every M-step. Therefore, the exact M-step given above can be replaced by an approximate M-step, that is obtained by considering an approximate Newton-Raphson solution of (2.4), where the $\mathcal{O}_p(\sqrt{n})$ terms in the Hessian of the log-likelihood are discarded. That is the term

$$\sum_{i=1}^{n} \Big\{ \nabla_f T_f(U_i;\xi) - \mathrm{E}_f \nabla_f T_f(U;\xi) \Big\}.$$

We consider therefore the Newton-Raphson EM (NR-EM) algorithm:

$$f_{i+1} = \begin{cases} \tilde{f}_i \triangleq f_i + H_{f_i}^{-1} \ell_{f_i}^M \{\cdot; \psi_{f_i}^{-1}(\boldsymbol{X})\} & \ell_{\tilde{f}_i} > \ell_{f_i} \\ \text{the solution of (2.3)} & \text{otherwise,} \end{cases}$$

where $H_f : L_2(\mu) \to L_2(\mu)$ is the operator $H_f(\xi,\zeta) = \mathrm{Cov}_f\{T_f(U;\xi), T_f(U;\zeta)\}$.

3. EPK: Rates of Convergence

In the previous section we considered the MLE estimate of f. In this section we consider simple estimators of the type suggested by DHM. Using these estimators we will be able to discuss possible minimax rates of convergence. In essence, we start with a naive nonparametric estimator of the mixture, and in the second step we improve it or demix it for f.

One simple method for demixing the EPK is to start with (1.4) which can be written as

$$1 = c \int \frac{\partial}{\partial u} g(u;\xi) f(\xi) d\xi \frac{q}{p} \Big\{ \int g(u;\xi) f(\xi) d\xi \Big\} = c \frac{\partial}{\partial u} \frac{q}{p} \Big\{ \int g(u;\xi) f(\xi) d\xi \Big\}.$$

Hence $q/p\{\int g(u;\xi) f(\xi) d\xi\} = \alpha + \beta u$ for some α and β, or

$$\int g(u;\xi) f(\xi) d\xi = \Big(\frac{p}{q}\Big)^{-1} (\alpha + \beta u). \tag{3.1}$$

The utility function of an individual is defined up to affine transformation. To assure that it is well defined, we assume that that at the return of 1 the value of the utility is 0, and that of the derivative is 1. In terms of the inverse utility function this translates to $g(0,\xi) \equiv \frac{\partial}{\partial u} g(0,\xi) \equiv 1$. Hence

$$\begin{aligned} \alpha &= \frac{p(1)}{q(1)} \\ \beta &= \frac{p'(1)}{q(1)} - \frac{p(1)}{q(1)} \frac{q'(1)}{q(1)}. \end{aligned} \tag{3.2}$$

The parameter f is therefore the solution of

$$\int g(u;\xi)f(\xi)d\xi = \psi(u) \tag{3.3}$$

for some ψ given explicitly by (3.1) and (3.2). Since q is estimated as a parametric density (based on a much larger sample), and p can be estimated at a standard non-parametric rate based on a direct sample from p, ψ can as well be estimated at a regular density estimation rate.

The analysis of this section starts with (3.3). We assume that ψ and its relevant derivatives can be estimated at a polynomial rate $\|\hat{\psi}^{(i)} - \psi^i\|_\infty = \mathcal{O}_p(n^{-\alpha_i})$ for some $\alpha_i > 0$. The natural estimator suggested by DHM is given by the inverse function of a weighed density estimator. Under strict monotonicity and boundness, the inverse function inherits most properties from the density kernel estimator.

Note that model (3.3) looks like a linear model. For example, if f is approximated by a finite distribution with point mass at $\xi_1, \ldots, \xi_m$, and (3.3) is considered at the k points $u_1, \ldots, u_k$, then it can be written as

$$\hat{\psi}(u_i) = \sum_{j=1}^{m} \beta_j g(u_i;\xi_j) + \varepsilon_i, \qquad i = 1, \ldots, k. \tag{3.4}$$

(3.4) looks like a standard linear model and, indeed, we suggest estimating f by solving it. However, it is not. Most linear model assumptions are violated, e.g., $\varepsilon_1, \ldots, \varepsilon_k$ are not i.i.d. and they are not independent of the random $u_1, \ldots, u_k$.

The basic idea of this section is as follow. We assume that we have some naive nonparametric estimator of ψ. We then proceed to use the pseudo linear model (3.4) to to estimate the mixing distribution and to improve the estimate of ψ itself. We show that this method yields the minimax rates.

How fast can f be estimated? In the rest of the section we present simple examples following DHM. These examples show that in a very similar models very different types of behavior can be obtained. It can be that (i) There is no consistent estimator of f; (ii) f can be estimated at a regular nonparametric rate of $n^{-\alpha}$; (iii) f can be estimated but at a very slow rate. Thus one can suspect that any optimistic result of demixing depends too heavily on assumptions, and are *a priori* not robust (at least in the minimax sense). In particular, any result should be checked to stand against different changes in the model.

3.1. Switching between two utilities

Following DHM assume that for $x, \xi > 0$,

$$U(x;\xi) = \alpha_2(1-c)^{1-1/\alpha_2}\left\{[x-\xi]_+^{1/\alpha_1} \vee (x-c)^{1/\alpha_2}\right\} - \alpha_2(1-c), \tag{3.5}$$

where $\alpha_2 > \alpha_1 > 1$ are given, $c < 0$, and $[x]_+ = x\mathbf{1}(x > 0)$. See Figure 2. Then

$$g(u;\xi) = \min\left\{\beta^{\alpha_2}\{u + \alpha_2(1-c)\}^{\alpha_2} + c,\ \beta^{\alpha_1}\{u + \alpha_2(1-c)\}^{\alpha_1} + \xi\right\},$$

where $\beta = \alpha_2^{-1}(1-c)^{-1+1/\alpha_2}$. To simplify the notation and generalize the discussion, we consider a slightly more general case.

Theorem 3.1. *Suppose q is known and bounded away from 0 on a open interval, p has $s > 2$ bounded derivatives, and*

$$g(u;\xi) = \begin{cases} g_2(u) & -\infty < u \le h(\xi) \\ g_1(u) + \xi & \infty > u > h(\xi) \end{cases}, \qquad \xi > 0,$$

where g_1, g_2 are continuous with bounded derivatives, and h given by

$$h^{-1} = g_2 - g_1 \tag{3.6}$$

is a strictly increasing function. Then, f can be estimated with an $\mathcal{O}_p(n^{-(s-2)/(2s+1)})$ error.

Proof. Note that $g(u;\xi)$ is continuous in ξ. Equation (3.3) can be translated to

$$\psi(u) = \int^{h^{-1}(u)} \xi f(\xi)d\xi + g_2(u)F\{h^{-1}(u)\} + g_2(u)\Big\{1 - F\{h^{-1}(u)\}\Big\},$$

where F is the cdf corresponding to the pdf f. Changing variables and considering (3.6),

$$\psi\{h(s)\} = \int^{s} \xi f(\xi)d\xi - sF(s) + g_2\{h(s)\}.$$

Taking a derivative gives $F(s) = h'(s)\{g_2'\{h(s)\} - \psi'\{h(s)\}\}$. Hence estimating F at s is equivalent to the estimation of ψ' at $h(s)$. In other words, $f(\cdot)$ can be estimated at the same rate as the rate of the estimation of second derivative of ψ, which in turn is governed by the rate of estimation of the second derivative of p. Since, by assumption, p has s bounded derivatives, f can be estimated with an $\mathcal{O}_p(n^{-(s-2)/(2s+1)})$ error, cf. Silverman (1986).

3.2. Polynomial and exponential inverse utility function

Theorem 3.1 described a relatively optimistic example. However, modest changes in the inverse utility function may create situations in which f can hardly be estimated, or even not at all.

Here is a pessimistic example:

Theorem 3.2. *Suppose the CRRA (constant relative risk aversion) utility*

$$g(u;\zeta) = (\alpha\zeta^{\alpha-1})^{-1}\Big\{(u+\zeta)^{\alpha} - \zeta^{\alpha}\Big\} + 1, \quad u \in \mathbb{R},\ \zeta \in \mathbb{R}^{+}, \tag{3.7}$$

where α is a known integer. Then there is no consistent estimator of f.

Note that g in (3.7) is scaled such that both its value and its derivative at zero are equal to 1, that is, it represents one branch of (3.5). The proof of Theorem 3.2 is simple. Since α is an integer, $\psi(\cdot)$ is a function of f only through its first α moments. Hence, these moments can be estimated, but no other aspects of f can be estimated or identified.

Seemingly, more and more moments are revealed as $\alpha \to \infty$, and therefore, by the above argument, f is going to be identified at the limit. However, it is not clear that the high moments can be estimated effectively. We consider the limiting case explicitly. The limiting form of the inverse utility function, as $\alpha \to \infty$ and $\alpha/\zeta \to \xi$, is given by

$$g(u;\xi) \equiv \xi^{-1}(e^{u\xi} - 1) + 1. \tag{3.8}$$

The density f is now identified. For example, all its moments can be estimated, e.g., by $\int \xi^{i} f(\xi)d\xi = \psi^{(i+1)}(0)$. We are now going to analyze this model in some detail. We will argue that if $f(\cdot)$ is assumed to have two bounded derivatives, then its value at a point can indeed be estimated, but this can be done only at a very slow convergence rate, slower than any polynomial rate.

Theorem 3.3. *Assume that g is given by* (3.8) *and f is bounded and has two bounded derivatives. Suppose the minimax rate of estimation of ψ is n^{γ}, $\gamma \in (0,1/2)$. Then there is an estimator $\hat{f}$ such that $\hat{f}(s) - f(s) = \mathcal{O}_p(n^{-\alpha \log\log n/\log n})$ for some α, and for any $\alpha > 0$ there is no estimator $\tilde{f}(s)$ such that $\tilde{f}(s) - f(s) = \mathcal{O}_p(n^{-\alpha/\log\log n})$.*

The proof is given in the on-line supplement, see `http://www.stat.sinica.edu.tw/statistica`.

3.3. Smoothing the empirical estimate and an uncertainty principle

We start, as in the previous subsections, with a nonparametric $\hat{\psi}$. The purpose of this subsection is to show that this initial estimator can be improved considerably by a simple projection.

We argued in Subsection 3.2 that there is no reasonable estimator of f for g given in (3.8). If (3.8) is believed to be true, does this means that there is nothing to do? The surprising answer is no. Although f cannot be estimated per-se, many

of its functionals can be estimated quite easily and quite well. For example, as mentioned in Subsection 3.2, its moments. Similarly $\psi(u)$, another functional of f, can be estimated quite easily, considered as a simple linear functional.

Suppose that f is supported on some compact interval $[a, b]$. Then one can approximate $\psi(u) = \sum_{i=1}^{m} \beta_i u^i + R_m(u)$, where, for some $\tilde{u} \in (0, u)$;

$$0 \le R_m(u) = \frac{1}{(m+1)!}\psi^{m+1}(\tilde{u}) = \frac{1}{(m+1)!}\int_a^b \xi^m e^{\tilde{u}\xi} f(\xi)d\xi \le \frac{b^m e^{ub}}{(m+1)!}. \quad (3.9)$$

Generally speaking, the faster the coefficients β converge to 0, the easier it is to estimate ψ and the harder it is to estimate the mixing density g. As (3.9) shows, we need only a few terms to approximate ψ quite well. In fact we show that in this smooth case, where as on the one hand f can be hardly estimated, ψ can be estimated almost at the parametric rate. This is not an accident — these are two faces of one phenomena. The shape of the observable ψ hardly depends on the fine details of f, and essentially depends only on a few aspects of f. These aspects can be estimated well (and hence ψ can be estimated quite precisely). The other aspects can hardly be estimated and hence f cannot be estimated in a reasonable rate. This yields an uncertainty principle — the more you are certain about ψ the less certain you are about f.

Recall that a function g is called completely monotone if $(-1)^k g^{(k)} \ge 0$, and it is called a Bernstein function if its first derivative is completely monotone. It is well-known (Feller (1966)) that g is completely monotone if, and only if, $g(u) = \int_0^\infty e^{-u\xi} dF(\xi)$. In other words, ψ is a Bernstein function. Nonparametric maximum likelihood estimation for an exponential mixture (and hence completely monotone density) was discussed in Jewell (1982). Balabdaoui and Wellner (2007) discussed the estimation of a k-monotone density.

We assume that there is an estimate $\hat{\psi} = \hat{\psi}_n$ at our disposal. For any $u_1, \ldots, u_k > 0$, let $\Sigma(u_1, \ldots, u_k) \in \mathbb{R}^{k\times k}$, where $\Sigma_{ij}(u_1, \ldots, u_k) = \mathrm{Cov}\{\hat{\psi}(u_i), \hat{\psi}(u_j)\}$. Consider the following assumption:

Assumptions 1. For any n there is $k = k_n$ and $u_1, \ldots, u_k \in (c, d)$, $0 < c < d$, such that the spectral radius of $\Sigma(u_1, \ldots, u_k)$ is $\mathcal{O}(k/n)$, and $\max_i |\mathrm{E}\,\hat{\psi}(u_i) - \psi(u_i)|^2 = \mathcal{O}(\log n/n)$.

Assumption 1 is satisfied by many nonparametric density and regression estimators, when they strictly under-smooth. We care much more about bias than about variance of the original estimator $\hat{\psi}$. Thus, we have in mind a kernel estimator with bandwidth of order $n^{-1/4+\varepsilon}$. The spectral radius is based on the assumptions that the estimator at points that are a multiple of the bandwidth apart are (almost) independent, for example this is trivially the case with kernel estimators having a compact support. The relationships in the assumption

obtain when the bias of the estimator is $\mathcal{O}(\sigma^2)$, the variance is $\mathcal{O}(1/n\sigma)$, and $k = \mathcal{O}(\sigma^{-1})$.

Consider now the least squares regression of $Y = \{\hat{\psi}(u_1), \ldots, \hat{\psi}(u_k)\}^\top$ on the design matrix $Z \in \mathbb{R}^{k\times m}$, $Z_{ij} = u_i^j$. That is, $\hat{\beta} = (Z'Z)^{-1}Z'Y$, where $\hat{\beta} \in \mathbb{R}^m$. Finally let $\tilde{\psi}(u) = \sum_{j=1}^m \hat{\beta}_j u^j$, $u > 0$. We argue that the error achieved by $\tilde{\psi}$ is almost the parametric rate even though $\hat{\beta}$ can be estimated at a strictly lower rate.

Theorem 3.4. *Suppose* $g(u;\xi) \equiv \xi^{-1}(e^{u\xi} - 1)$ *and that* f *is supported on a compact interval. Assume* 1 *holds and* $m = m_n = \log n/\log\log n$. *Then* $k^{-1}\sum_{i=1}^k \{\tilde{\psi}(u_i) - \psi(u_i)\}^2 = \mathcal{O}_p\{(\log n)^2/n\}$.

Proof. Let β^0 be the true value $\beta_j^0 = \int \xi^{j-1} f(\xi)d\xi/j!$. Write $Y = Z\beta + \varepsilon$, where ε includes both the random error and the bias terms due to both the estimator and the truncation. The latter term is given in (3.9). By standard least squares results,

$$k^{-1}\mathrm{E}\sum_{i=1}^k \left\{\tilde{\psi}(u_i) - \psi(u_i)\right\}^2 = k^{-1}\mathrm{E}\left\{\varepsilon^\top Z(Z^\top Z)^{-1}Z^\top\varepsilon\right\}$$
$$= k^{-1}\,\mathrm{trace}\left\{Z(Z^\top Z)^{-1}Z^\top \mathrm{E}\,(\varepsilon\varepsilon^\top)\right\}.$$

Since $Z(Z^\top Z)^{-1}Z^\top$ is a projection matrix on a m-dimensional space, the RHS is bounded by the largest eigenvalue of $\mathrm{E}\,(\varepsilon\varepsilon^\top)$ times m/k. This has three components (variance and two biases) and hence

$$k^{-1}\mathrm{E}\sum_{i=1}^k \left\{\tilde{\psi}(u_i) - \psi(u_i)\right\}^2 = \mathcal{O}\left[\frac{m}{k}\left\{\frac{k}{n} + k\frac{\log n}{n} + k\left(\frac{b^m}{m!}\right)^2\right\}\right].$$

The factor k before the last two terms is due to the norm of the unit vector in $\mathbb{R}^k$, and, the last term is by (3.9). The theorem follows by taking $m = \log n/\log\log n$.

A more general result can be based on an assumption like the following.

Assumptions 2. For some c, d and each ε there are $h_{\varepsilon,1}, \ldots, h_{\varepsilon,M(\varepsilon)}$ such that

$$\sup_\xi \min_\gamma \max_{c<u<d} \left|g(u;\xi) - \sum_{j=1}^{M(\varepsilon)} \gamma_j h_j(u)\right| < \varepsilon.$$

Note that clearly the assumption ensures the existence of $\gamma(\cdot)$ such that $\max_{c<u<d} |g(u;\xi) - \sum_{j=1}^{M(\varepsilon)} \gamma_j(\xi)h_j(u)| < \varepsilon$, but then there are also $\beta_j = \int \gamma_j(\xi)f(\xi)d\xi$, $j = 1, \ldots, M(\varepsilon)$, such that $\max_{c<u<d} |\psi(u) - \sum_{j=1}^{M(\varepsilon)} \beta_j h_j(u)| < \varepsilon$.

The following theorem can be proved similarly to Theorem 3.4:

Theorem 3.5. *Suppose Assumptions* 1 *and* 2 *hold. Let* $\varepsilon_n = \operatorname{argmin}_\varepsilon\{M(\varepsilon)/n+\varepsilon\}$, *and let* $\tilde{\psi}$ *be the least squares estimate of the regression of* $\hat{\psi}$ *on* $h_{\varepsilon_n,1},\ldots,h_{\varepsilon_n,M(\varepsilon_n)}$. *Then* $k^{-1}\sum_{i=1}^{k}\{\tilde{\psi}(u_i)-\psi(u_i)\}^2 = \mathcal{O}_p(\varepsilon_n)$.

In practice, Theorems 3.4 and 3.5 may seem to be of limited use — a knowledge of the structure of the span of the individual utility functions is needed, and the regression is based on an identified efficient base, which may not be natural. For example, we used a polynomial base for the exponential utility function. The practical approach is a histogram or discrete approximation of f. Does such a procedure yield an effective estimator, an estimator which is both statistically speaking efficient, but at the same time easy to compute and can be be used in off-the-shelf manner?

This is indeed the case. Let $\xi_1,\ldots,\xi_{M(\varepsilon)}$ be reasonably spaced points in the support of f. With the notation introduced after Assumption 2, and by a similar argument, for a vector β on the simplex

$$\sup_u\left|\sum_{j=1}^{M(\varepsilon)}\beta_j g(u;\xi_j)-\sum_{j=1}^{M(\varepsilon)}\beta_j\sum_{l=1}^{M(\varepsilon)}\gamma_l(\xi_j)h_l(u)\right|\le\varepsilon.$$

Hence, one can use the base function $g(\cdot;\xi_1),\ldots,g(\cdot;\xi_{M(\varepsilon)})$ as well.

References

Ait-Sahalia, Y. and Lo, A. (2000). Nonparametric risk-management and implied risk aversion. *J. Econometrics* **94**.

Balabdaoui, F. and Wellner, J. A. (2007). Estimation of a k-monotone density: limit distribution theory and the spline connection. Manuscript.

Chabi-Yo, F., Garcia, R. M. and Renault, R. (2008). State dependence can explain the risk aversion puzzle. *Rev. Finan. Stud.* **21**, 973-1011.

Cochrane, J. H. (2005). *Asset Pricing (Revised)*. Princeton University Press, Princeton.

Detlefsen, K., Härdle, W. K. and Moro, R. A. (2007). Empirical pricing kernels and investor preferences. SFB649 Discussion paper 2007-017, `http://sfb649.wiwi.hu-berlin.de/fedc/discussionPapers_de.php`.

Feller, W. (1966). *An Introduction to Probability Theory and its Applications, Vol. II.* Wiley, New-York.

Friedman, M. and Savage, L. P. (1948). The utility analysis of choices involving risk. *J. Polit. Economy* **56**, 279-304.

Gallant, A. R. and Hong, H. (2007). A statistical inquiry into the plausibility of Epstein-Zin-Weil Utility. *J. Finan. Econom.* **5**, 523-559.

Gilbert, P. B. (2000). Large sample theory of maximum likelihood estimates in semiparametric biased sampling models. *Ann. Statist.* **28**, 151-194.

Gilbert, P. B., Lele, S. R. and Vardi, Y.(1999). Maximum likelihood estimation in semiparametric selection bias models with application to AIDS vaccine trials. *Biometrika* **86**, 27-43.

Gill, R. D., Vardi, Y. and Wellner, J. A. (1988). Large sample theory of empirical distributions in biased sampling models. *Ann. Statist.* **16**, 1069-1112.

Jewell, N. P. (1982). Mixtures of exponential distributions. *Ann. Statist.* **10**, 479-482.

Manski, C. F. and Lerman, S. R. (1977). The estimation of choice probabilities from choice based samples. *Econometrica* **45**, 1977-1988.

Mantel, N. (1973). Synthetic restropective studies and related topics. *Biometrics* **29**, 479-486.

Rosenberg, J. and Engle, R. (2002). Empirical pricing kernels. *J. Finan. Econom.* **64**, 341-372.

Silverman, B., (1986). *Density Estimation.* Chapman and Hall, London.

Vardi, Y. (1985). Empirical distributions in selection bias models. *Ann. Statist.* **13**, 178-203.

Walras, M.-E. L. (1874). *Éléments d'économie politique pure, ou théorie de la richesse sociale.*

Department of Statistics, The Hebrew University of Jerusalem 91905, Jerusalem, Israel.

E-mail: yaacov.ritov@gmail.com

CASE - Center for Applied Statistics and Economics, Institute for Statistics and Econometrics, Humboldt-Universität zu, 10178 Berlin, Germany.

E-mail: haerdle@wiwi.hu-berlin.de.

(Received February 2008; accepted February 2009)

The Bayesian Additive Classification Tree Applied to Credit Risk Modelling

Junni L. Zhang[1], Wolfgang K. Härdle[2]

[1]Department of Business Statistics and Econometrics, Guanghua School of Management, Peking University, Beijing 100871, P. R. China; email: zjn@gsm.pku.edu.cn.

[2]Center for Applied Statistics and Economics, Wirtschaftswissenschaftliche Fakultät, Humboldt-Universität zu Berlin, Spandauer Straβe 1, 10178, Berlin, Germany; email: haerdle@wiwi.hu-berlin.de.

Abstract: We propose a new nonlinear classification method based on a Bayesian "sum-of-trees" model, the Bayesian Additive Classification Tree (BACT), which extends the Bayesian Additive Regression Tree (BART) method into the classification context. Like BART, the BACT is a Bayesian nonparametric additive model specified by a prior and a likelihood in which the additive components are trees, and it is fitted by an iterative MCMC algorithm. Each of the trees learns a different part of the underlying function relating the dependent variable to the input variables, but the sum of the trees offers a flexible and robust model. Through several benchmark examples, we show that the BACT has excellent performance. We apply the BACT technique to classify whether firms would be insolvent. This practical example is very important for banks to construct their risk profile and operate successfully. We use the German Creditreform database and classify the solvency status of German firms based on financial statement information. We show that the BACT outperforms the logit model, CART and

the Support Vector Machine in identifying insolvent firms.

Key words and phrases: Classification and Regression Tree, Financial Ratio, Misclassification Rate, Accuracy Ratio

JEL-Codes: C14, C11, C45, C01

1 Introduction

Classification techniques have been popularly used in many fields. Standard classification tools include linear and quadratic discriminant analysis and the logistic model. The support vector machine (SVM) (Vapnik, 1995, 1997) recently arises as an important nonlinear classification tool. It maps the input space nonlinearly into a high dimensional feature space, and tries to find linear separating hyperplanes for the classes in the feature space, penalizing the distances of misclassified cases to the hyperplanes. The SVM has been widely and successfully applied to classification problems in many domains and often shown to have excellent performance compared to other classification methods.

Decision trees compose an important category of nonlinear classification methods. Ever since the introduction of the classification and regression tree (CART) by Breiman et al. (1984), it has attracted strong interest from researchers and practitioners. Figure 1 shows an example of a classification tree, where the root node (t_1) contains all training observations, and the training data are recursively partitioned by values of the input variables (x's) until reaching the leaf (terminal) nodes (t_3, t_4, t_6 and t_7) where the classification decision (for y) is made for all observations contained therein. For regression problems in which the dependent variable is continuous, a predicted value for the dependent variable would be assigned for all observations contained in each leaf node.

Traditional search methods for CART models use locally greedy algorithms to find the partitions. The Bayesian approaches for CART models (Chipman et al., 1998; Denison et al., 1998; Wu et al., 2007) specify a formal prior distribution for trees and other parameters and use Markov Chain Monte Carlo methods to sample them from the posterior distribution.

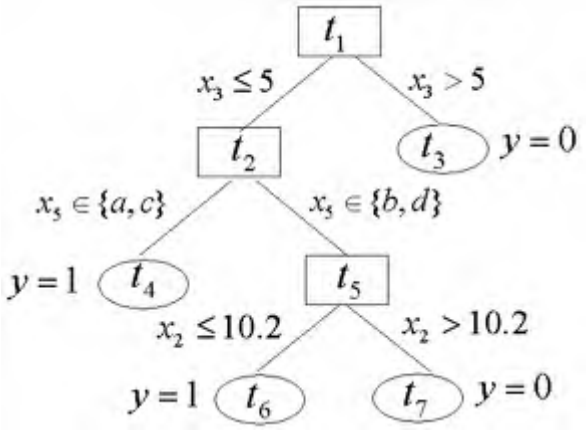


Figure 1: Example of a classification tree.

Chipman et al. (2006) proposed the Bayesian Additive Regression Tree (BART), in which the mean of a continuous dependent variable is approximated by a sum of trees rather than a single tree. This "sum-of-trees" model is defined by a prior and a likelihood, and fitted by iterative MCMC algorithm. Each individual tree explains a different portion of the underlying mean function, but the sum of these trees turns out to be a flexible and adaptive model. Chipman et al. (2006) showed that BART outperforms several competitive models, including LASSO (Efron et al., 2004), gradient boosting (Friedman, 2001), random forests (Breiman, 2001), and neural networks with one layer of hidden units. We will extend BART into the classification context, and therefore term the resulting classification technique as the Bayesian Additive Classification Tree (BACT).

To investigate the differences among the logit model, SVM, CART and BACT, we plot in Figure 2 the contours of these models trained to classify the solvency status of German firms using the German Creditreform database based on only two variables — the ratio of operating income to total assets ($x3$ in Figure 2) and the ratio of accounts payable to

total sales ($x24$ in Figure 2). Details of this application will be discussed in Section 4. The contours for the logit model are linear, thus making it inflexible for complex applications. The SVM finds flexible smooth curves in the input space (linear hyperplanes in the feature space) that can separate the classes. The CART is based on a single tree which recursively partitions the observations by the input variables, and hence the contours are piecewise linear. The BACT is based on the sum of many trees, so the contours are not constrained to be piecewise linear as in CART; although these contours are not as smooth as in SVM, they are quite flexible in explaining complex structure.

The rest of this paper is organized as follows. Section 2 will describe the BACT in detail. Section 3 will use several benchmark examples from the UCI Machine Learning Repository to compare the performance of the BACT with the logit model and the SVM. Section 4 will discuss our application to classification of solvency status of Germany firms using the German Creditreform database. Section 5 then concludes.

2 The Bayesian Additive Classification Tree (BACT)

2.1 The Model

Consider a binary classification problem in which an dependent variable $Y \in \{1, 0\}$ needs to be predicted based on a set of input variables $\boldsymbol{x} = (x_1, \cdots, x_p)^\top$. The majority of classification models assume that there is a latent continuous variable Y^* that determines

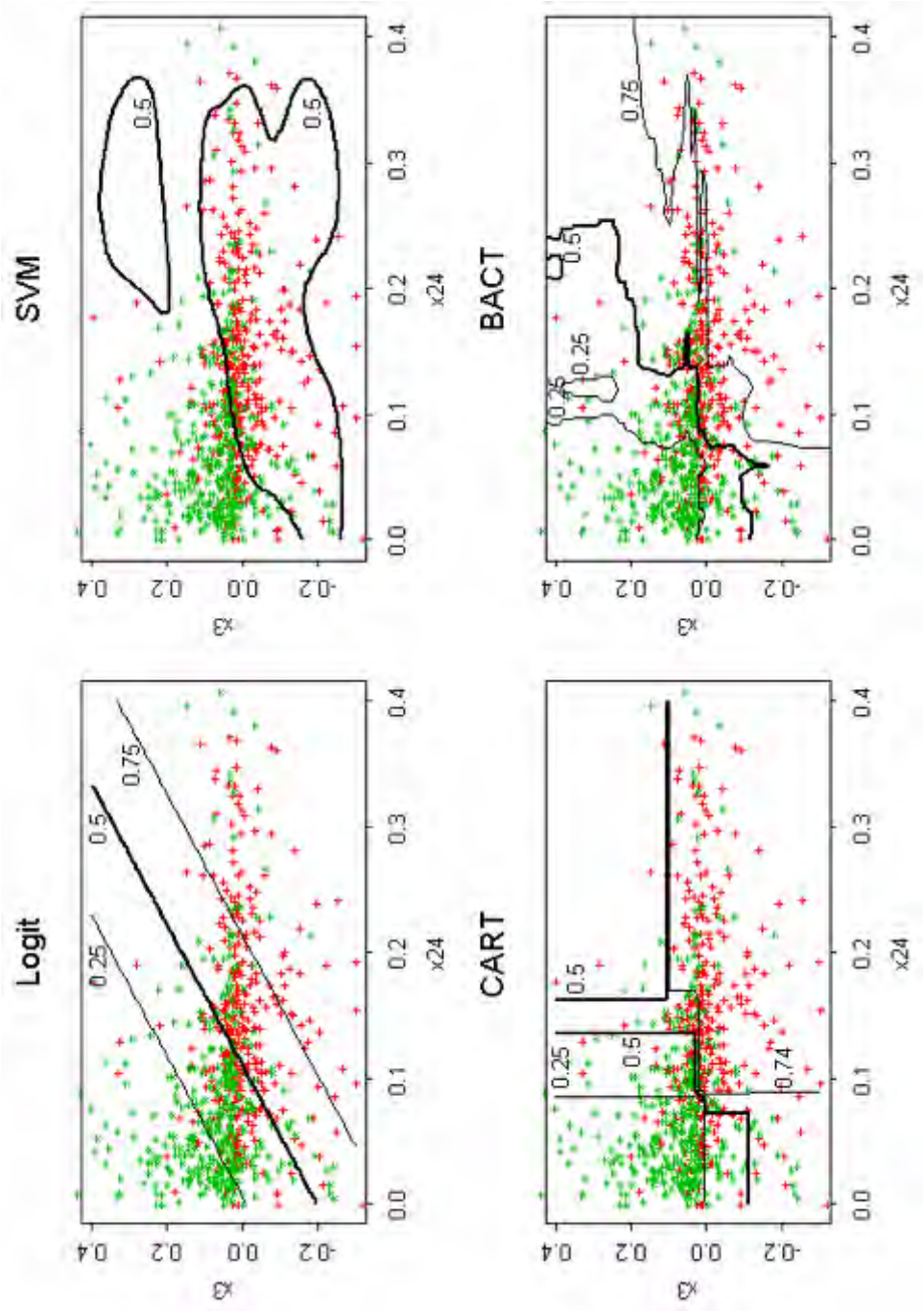


Figure 2: The contour plots for the logit model, SVM, CART, BACT. The pluses and stars represent insolvent firms and solvent firms respectively. The numbers by the contours indicate the probabilities of insolvency.

the value of Y as follows

$$\begin{cases} Y = 1 \ \text{ if } Y^* \geq 0 \\ Y = 0 \ \text{ if } Y^* < 0 \end{cases} \tag{1}$$

In the context of generalized linear models (GLM), the relationship of Y^* and $\boldsymbol{x}$ is

$$Y^* = \beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p + \varepsilon,$$

where the distribution of ε determines the link function, e.g. logit or probit. The generalized additive models (GAM, Hastie and Tibshirani (1990)) replace each linear term in the GLM by a more generalized functional form and relate Y^* to $\boldsymbol{x}$ by

$$Y^* = \beta_0 + f_1(x_1) + \cdots + f_p(x_p) + \varepsilon,$$

where each f_j is an unspecified smooth function.

Following the idea of the BART in Chipman et al. (2006), we assume that Y^* is related to $\boldsymbol{x}$ through an additive model, where each additive component is a tree based on all input variables (rather than a flexible function based on a single input variable as in GAM). In order to formally introduce the model, we first introduce some notation. Let m denote the number of trees to be used. For $j = 1, \cdots, m$, let T_j denote the j'th tree with a set of partition rules based on the input variables, and let L_j denote the number of leaf nodes in T_j; for $l = 1, \cdots, L_j$, let μ_{jl} denote the (continuous) predicted value associated with the l'th leaf node in T_j, and let $M_j = \{\mu_{j1}, \mu_{j2}, \cdots, \mu_{jL_j}\}$. For a given value of $\boldsymbol{x}$, let $g(\boldsymbol{x}, T_j, M_j)$ denote the predicted value associated with the leaf node that an observation with input variables being $\boldsymbol{x}$ would land in based on the partition rules for T_j. Thus Y^* is formally modelled as

$$Y^* = g(\boldsymbol{x}; T_1, M_1) + g(\boldsymbol{x}; T_2, M_2) + \cdots + g(\boldsymbol{x}; T_m, M_m) + \varepsilon, \tag{2}$$

and we further assume that $\varepsilon \sim N(0,1)$, using a probit-like link.

2.2 Prior Specification

In order to make inferences from the model given by (1) and (2) in a Bayesian way, we need to specify a joint prior distribution for the unknown tree structures and leaf nodes parameters. We assume a priori that the tree structures and the leaf node parameters have independent distributions, so the full prior distribution can be written as

$$p\{(T_1, M_1), (T_2, M_2), \cdots, (T_m, M_m)\} = \prod_{j=1}^{m} p(T_j) \prod_{j=1}^{m} \prod_{l=1}^{L_j} p(\mu_{jl}).$$

We further assume that every tree follows the same prior distribution, and every μ_{jl} follows the same prior distribution. So the task of prior specification is reduced to specifying the prior distribution for a single tree T and that for a single μ_{jl} parameter.

For a single tree T, we need to specify the prior distributions for its partition rules, including whether to further split a node or leave it as a leaf node, and if a further split is needed, which input variable and what values to be used for that split. We use the prior distribution for a single tree T as in Chipman et al. (2006). The prior probability of splitting any node n in tree T is

$$p_{split}(n, T) \propto \alpha(1 + d_n)^{-\beta},$$

where d_n is the depth of node n in tree T (the depth of node n is the length of the path from the root node to node n; e.g., in Figure 1, the node t_1 has depth 0, and the nodes t_2 and t_3 have depth 1). α and β here are positive hyperparameters, hence the deeper a node is, the smaller probability there is to further split it, or the larger probability that this node becomes a leaf node. It turns out that the performance of BACT is not very sensitive to the

Table 1: Prior distribution on number of terminal nodes based on different values of α and β.

	Setting 1	Setting 2	Setting 3
α	0.5	0.95	0.95
β	2	2	0.1
prior probability of trees with 1 terminal node	0.5	0.05	0.05
prior probability of trees with 2 terminal nodes	0.383	0.552	0.012
prior probability of trees with 3 terminal nodes	0.098	0.275	0.004
prior probability of trees with 4 terminal nodes	0.017	0.092	0.002
prior probability of trees with ≥ 5 terminal nodes	0.003	0.031	0.932

choice of *alpha* and *beta*. We tried three different settings listed in Table 1 where a priori the trees range from small size to large size, and the resulting performance was quite similar. So we just pick $\alpha = .95$ and $\beta = 2$ as in Chipman et al. (2006). If a node needs to be split, the prior for the associated splitting rules assigns equal probability to each available input variable and equal probability on each available rule given the variable.

The prior distribution of μ_{jl} is taken to be a conjugate normal distribution $\mu_{jl} \sim N(0, \sigma_\mu^2)$ (conjugate because ε in (2) follows a normal distribution). From (2), we can see that the expected value of Y^* is equal to the sum of m different μ_{jl} parameters (recall that $g(\boldsymbol{x}, T_j, M_j)$ is the μ_{jl} parameter associated with the leaf node that an observation with input variables being $\boldsymbol{x}$ would land in based on the partition rules for T_j); because of the a priori independence of μ_{jl}'s, the prior distribution for the expected value of Y^* is $N(0, m\sigma_\mu^2)$. Combining this with (1), it can be inferred that a priori each observation has probability 0.5 belonging to class 1 and probability 0.5 belonging to class 0.

To specify σ_μ^2, we use the following procedure. We first estimate the range of Y^* (to be explained soon), and then choose σ_μ^2 such that there is at least 95% prior probability that the

expected value of Y^* is in the estimated range. Let the training data be $\mathcal{D} = \{(\boldsymbol{x}_i, y_i)\}_{i=1}^N$, where N is the number of observations in the training data. We first randomly sample y_i^* for each observation i in the training data from truncated standard normal distributions such that the relationship in (1) holds between y_i^* and the observed y_i. Suppose that the sampled values are $\boldsymbol{y}^{*(0)} = \{y_i^{*(0)}\}_{i=1}^N$, and denote the minimum and maximum values of $y_i^{*(0)}$ as $\min(\boldsymbol{y}^{*(0)})$ and $\max(\boldsymbol{y}^{*(0)})$ respectively. Then $[\min(\boldsymbol{y}^{*(0)}), \max(\boldsymbol{y}^{*(0)})]$ is a very rough estimate of the range of Y^*. We choose an initial $\sigma_\mu^{2(0)}$ such that there is at least 95% prior probability that the expected value of Y^* is in this interval, i.e., $[-2\sqrt{m}\sigma_\mu^{2(0)}, 2\sqrt{m}\sigma_\mu^{2(0)}]$ covers $[\min(\boldsymbol{y}^{*(0)}), \max(\boldsymbol{y}^{*(0)})]$ and therefore $\sigma_\mu^{2(0)} = \max\big\{-\min(\boldsymbol{y}^{*(0)})/2\sqrt{m}, \max(\boldsymbol{y}^{*(0)})/2\sqrt{m}\big\}$. We then run the Markov Chain Monte Carlo (MCMC) algorithm to be described in Section 2.3 to generate posterior samples of y_i^*, and suppose that we obtain one posterior draw of $\boldsymbol{y}^{*(1)} = \{y_i^{*(1)}\}_{i=1}^N$ after dropping the first B_1 posterior draws used to reach convergence. We assume this set of y_i^* can be used to estimate reasonably the range of the true underlying Y^*, and choose the value of σ_μ^2 for further analysis such that there is at least 95% prior probability that the expected value of Y^* is in the interval $[\min(\boldsymbol{y}^{*(1)}), \max(\boldsymbol{y}^{*(1)})]$, i.e., $\sigma_\mu^2 = \max\big\{-\min(\boldsymbol{y}^{*(1)})/2\sqrt{m}, \max(\boldsymbol{y}^{*(1)})/2\sqrt{m}\big\}$.

2.3 Generation of Posterior Samples and Inference

We use the data augmentation method (Tanner and Wong, 1987) by treating $\boldsymbol{y}^* = \{y_i^*\}_{i=1}^N$ as missing data, and then use the Gibbs sampler to generate samples from the posterior distribution $p\{(T_1, M_1), (T_2, M_2), \cdots, (T_m, M_m), \boldsymbol{y}^* | \mathcal{D}\}$.

Let $T_{(j)}$ denote the $m-1$ trees other than T_j, and let $M_{(j)}$ denote the parameters

associated with the leaf nodes in $T_{(j)}$. The Gibbs sampler composes of drawing m successive draws of (T_j, M_j) for $j = 1, \cdots, m$ from $p\{(T_j, M_j)|T_{(j)}, M_{(j)}, \boldsymbol{y}^*, \mathcal{D}\}$ followed by draw of $\boldsymbol{y}^*$ from $p\{\boldsymbol{y}^*|(T_1, M_1), (T_2, M_2), \cdots, (T_m, M_m), \mathcal{D}\}$. The draws of (T_j, M_j) can be generated similar to Chipman et al. (2006). Let $\hat{y}_i^* = \sum_{j=1}^m g(\boldsymbol{x}_i; T_j, M_j)$ denote the fitted value for observation i from the m trees. Then y_i^* $(i = 1, \cdots, N)$ can be independently generated from truncated normal distributions:

$$\begin{cases} y_i^* \sim N(\hat{y}_i^*, 1) \text{ and } y_i^* \geq 0 & \text{if } y_i = 1 \\ y_i^* \sim N(\hat{y}_i^*, 1) \text{ and } y_i^* < 0 & \text{if } y_i = 0 \end{cases}$$

After σ_μ^2 has been chosen according to the procedure described in Section 2.2, we can drop the first B_2 posterior draws used to reach convergence, and use subsequent S posterior draws for inference. Denote these S posterior draws as $\{(T_1^{(s)}, M_1^{(s)}), \cdots, (T_m^{(s)}, M_m^{(s)})\}_{s=1}^S$. Given the s'th draw, the probability that an observation with input variables $\boldsymbol{x}$ belongs to class 1 is $\Phi\left\{\sum_{j=1}^m g(\boldsymbol{x}, T_j^{(s)}, M_j^{(s)})\right\}$, where Φ is the cumulative distribution function of standard normal distribution. Therefore, the posterior average probability that an observation with input variables $\boldsymbol{x}$ belongs to class 1 can be estimated as

$$\frac{1}{S}\sum_{s=1}^{S} \Phi\left\{\sum_{j=1}^{m} g(\boldsymbol{x}, T_j^{(s)}, M_j^{(s)})\right\}. \tag{3}$$

We can use (3) to classify observations in training data or other data: if the probability calculated from (3) is larger than 0.5, then the observation is classified into class 1; otherwise it is classified into class 0.

Table 2: For five benchmark data sets from the UCI Machine Learning Repository, the number of cases, the number of variables, and the average misclassification rates for the test data using the logit model, the SVM and the BACT.

Data Set	# Cases	# Variables	Logit	SVM	BACT
breast cancer	683	9	3.8%	2.8%	3.3%
ionosphere	351	34	12.8%	4.5%	7.2%
diabetes	768	8	21.8%	25.2%	24.8%
sonar	208	60	29.8%	19.4%	17.2%
German credit	1000	30	23.6%	27.3%	23.6%

3 Benchmark Examples

To compare the performance of the BACT with the logit model and SVM (in which radial basis function is used as the kernel, and the parameters are chosen by cross-validation), we use five data sets for binary classification from the UCI Machine Learning Repository (Asuncion and Newman, 2007): breast cancer, ionosphere, diabetes, sonar, and German credit. Columns 2-3 in Table 2 summarize the number of cases and the number of variables for these data sets. Throughout the rest of the paper, in the BACT method, we fix $m = 200$, $B_1 = 500$, $B_2 = 1000$ and $S = 1000$.

We partition each data set randomly into 80% of training data and 20% of test data. The training data is used to fit the models, and misclassification rate on the test data is calculated. Such procedure is repeated for 20 times, and columns 4-6 in Table 2 report the average misclassification rates on the test data using the logit model, the SVM and the BACT. We can see that the BACT has comparable performance with the SVM, and has no worse performance than the logit model except for the "diabetes" data set.

4 Classification of Solvency Status of German Firms

We use the German Creditreform database, which contains financial statement information on 20,000 solvent and 1,000 insolvent firms in Germany and spans the period from 1996 to 2002. Information on the insolvent firms were collected two years prior to insolvency. Chen et al. (2007); Härdle et al. (2008) applied SVM to classify the solvency status of German firms, with the former using the German Creditreform database. We will preprocess the data set in the same way as Chen et al. (2007) do, and compare the results of our BACT with those of the logit model, CART and SVM.

Following Chen et al. (2007), we clean the data of firms whose characteristics are very different from the others. We first eliminate firms within industries with small percentage in the industry composition and are left with 949 insolvent firms and 16583 solvent firms in four main industries — Construction, Manufacturing, Wholesale & Retail Trade and Real Estate. We then exclude those firms whose asset size is less than 10^5 EUR or greater than 10^8 EUR, because the credit quality of small firms often depends as much on the finances of a key individual as on the firm itself and largest firms rarely go bankrupt in Germany. We further exclude the solvent firms in 1996 due to lack of insolvent firms in that year. We also eliminate firms with zero value for some variables used as denominators in calculating financial ratios to be used in classification. Several apparent outliers are then deleted and we end up with a data set with 783 insolvent firms and 9,575 solvent firms (due to slightly different ways of deleting outliers, our remaining solvent firms differ a little from the 9,583 solvent firms in Chen et al. (2007)).

We adopt the same set of financial variables to be used for classification as in Chen et al.

(2007) and list them in Table 3. The five number summary of these financial variables are listed in Table 4 for insolvent firms and solvent firms separately. In order to avoid sensitivity to outliers in applying the SVM, Chen et al. (2007) truncated each financial variable to be between its 5% quantile and 95% quantile. The BACT, however, only uses the ordering of values of the input variables in the partition rules, so there is no need to do such truncation.

We use the data from 1997 to 1999 to train the model, and use the data from 2000 to 2002 to test the resulting model. The training set contains 387 insolvent firms and 3535 solvent firms, and the test set contains 396 insolvent firms and 6040 solvent firms. Because the density of insolvent firms is rather low, we need to oversample the insolvent firms in order for the models to pick up the patterns predictive of insolvency (e.g., Berry and Linoff (2000), chap. 5). This is done through the bootstrap technique (Efron and Tibshirani, 1993; Sobehart et al., 2001). For each bootstrap sample, a training subset is constructed as follows. We use all 387 insolvent firms in the training set and randomly sample 387 solvent firms from the training set. This subset of 774 firm with 50% being insolvent is then used to train the model. When training the CART model, the training subset is further randomly partitioned into two parts stratified by the solvency status of the firms. The first part comprises of 80% of the training subset and is used to grow the tree, and the second part comprises of the remaining 20% of the training subset and is used to prune the tree. Performance measures are then evaluated using all observations (396 insolvent firms and 6040 solvent firms) in the test set. The average performance measures over 30 bootstrap samples are then calculated. We can compare average performance measures across different models.

We consider two performance measures: Accuracy Ratio (AR) (Sobehart and Keenan,

Table 3: Definition of financial variables to be used for classification for the Creditreform data.

Var.	Definition
x1	Net Income/Total Assets
x2	Net Income/Total Sales
x3	Operating Income/Total Assets
x4	Operating Income/Total Sales
x5	Earnings before Interest and Tax/Total Assets
x6	Earnings Before Interest, Tax, Depreciation and Amortization/Total Assets
x7	Earnings before Interest and Tax/Total Sales
x8	Own Funds/Total Assets
x9	(Own Funds − Intangible Assets) /(Total Assets − Intangible Assets − Cash and Cash Equivalents − Lands and Buildings)
x10	Current Liabilities/Total Assets
x11	(Current Liabilities − Cash and Cash Equivalents)/Total Assets
x12	Total Liabilities/Total Assets
x13	Debt/Total Assets
x14	Earnings before Interest and Tax/Interest Expense
x15	Cash and Cash Equivalents/Total Assets
x16	Cash and Cash Equivalents/Current Liabilities
x17	(Cash and Cash Equivalents − Inventories)/Current Liabilities
x18	Current Assets/Current Liabilities
x19	(Current Assets − Current Liabilities)/Total Assets
x20	Current Liabilities/Total Liabilities
x21	Total Assets/Total Sales
x22	Inventories/Total Sales
x23	Accounts Receivable/Total Sales
x24	Accounts Payable/Total Sales
x25	log(Total Assets)
x26	Increase (Decrease) in Inventories/Inventories
x27	Increase (Decrease) in Liabilities/Total Liabilities
x28	Increase (Decrease) in Cash Flow/Cash and Cash Equivalents

Table 4: Five number summary (minimum, lower quartile, median, upper quartile, maximum) of the financial variables for insolvent firms and solvent firms.

	Insolvent Firms					Solvent Firms				
Var.	min	Q1	mdn.	Q3	max	min	Q1	mdn.	Q3	max
x1	-1.51	-0.02	0.00	0.02	1.13	-4.82	0.00	0.02	0.06	5.92
x2	-5.41	-0.02	0.00	0.01	6.10	-17.13	0.00	0.01	0.03	15.91
x3	-0.97	-0.04	0.00	0.03	1.14	-4.82	0.00	0.03	0.09	5.97
x4	-3.38	-0.02	0.00	0.02	10.15	-44.81	0.00	0.02	0.04	20.39
x5	-0.99	-0.01	0.02	0.05	1.15	-1.51	0.02	0.05	0.11	5.95
x6	-0.91	0.03	0.07	0.11	1.17	-1.46	0.06	0.11	0.18	5.95
x7	-3.55	-0.01	0.01	0.04	10.27	-39.63	0.01	0.02	0.05	14.53
x8	0.00	0.00	0.05	0.14	0.96	0.00	0.05	0.14	0.28	0.99
x9	-0.86	0.00	0.05	0.17	2.31	-2.68	0.05	0.16	0.37	49.18
x10	0.01	0.37	0.52	0.73	1.00	0.00	0.25	0.42	0.64	4.13
x11	-0.35	0.33	0.49	0.69	0.99	-0.86	0.17	0.36	0.58	4.12
x12	0.01	0.54	0.76	0.89	1.00	0.00	0.42	0.65	0.82	4.37
x13	0.00	0.09	0.21	0.37	0.91	0.00	0.02	0.15	0.33	0.98
x14	-17658.06	-0.56	1.05	1.92	433.40	-22796.04	0.86	2.16	6.55	516896.73
x15	0.00	0.00	0.02	0.06	0.44	0.00	0.01	0.03	0.11	0.90
x16	0.00	0.01	0.03	0.12	25.01	0.00	0.01	0.08	0.30	40.61
x17	0.01	0.43	0.68	0.97	57.44	0.00	0.59	0.94	1.58	238.37
x18	0.03	1.00	1.26	1.84	62.63	0.06	1.11	1.58	2.67	989.76
x19	-0.69	0.00	0.15	0.36	0.92	-3.45	0.06	0.25	0.47	0.98
x20	0.07	0.62	0.84	0.99	1.18	0.01	0.56	0.85	1.00	1.00
x21	0.07	0.40	0.61	0.94	97.26	0.02	0.32	0.48	0.74	828.76
x22	0.00	0.08	0.16	0.34	89.96	-0.14	0.05	0.11	0.21	451.09
x23	0.00	0.07	0.12	0.18	0.87	0.00	0.05	0.09	0.14	21.85
x24	0.00	0.09	0.14	0.19	43.96	0.00	0.04	0.07	0.11	61.29
x25	11.72	14.07	14.87	15.76	18.25	11.51	14.25	15.41	16.62	18.42
x26	-46.89	-0.09	0.00	0.26	2.83	-282.51	-0.01	0.00	0.06	145.12
x27	-12.75	-0.04	0.00	0.11	1.00	-28.91	-0.04	0.00	0.10	1.00
x28	-1283.20	-0.61	0.00	0.18	1.00	-2513.39	-0.27	0.00	0.26	1.75

2001; Engelmann et al., 2003) and misclassification rate. AR is calculated using the Cumulative Accuracy Profiles (CAP) (Sobehart and Keenan, 2001; Engelmann et al., 2003) curve. To obtain the CAP curve, the firms are first ordered by risk scores from riskiest to safest. For BACT and the Logit model, the risk score is simply the predicted probability of insolvency; for SVM, the risk score can be calculated as distance to the separating hyperplane. The higher the risk score is, the riskier the firm is. For a given fraction q of the total number of firms, the CAP curve is constructed by calculating the fraction $r(q)$ of the insolvent firms whose risk scores are equal to or larger than the minimum score at fraction q.

Figure 3 plots the CAP curve for the test set of the Creditreform data where the scoring model is the BACT model trained using one bootstrap training subset. In the ideal case, the insolvent firms will be assigned the highest risk scores, and therefore the CAP curve would be increasing linearly and then stay at one. For a random model without any discriminative power, the fraction q of all firms with the highest risk scores will contain fraction q of all insolvent firms, and therefore the corresponding CAP curve will be a straight line connecting the points $(0, 0)$ and $(1, 1)$. AR is defined as the ratio of the area between the CAP curve for a scoring model and that for the random model to the area between the CAP curve for the ideal case and that for the random model. The value of AR lies between zero and one, with zero indicating no discriminative power of the scoring model and one indicating perfect discriminative power. Mathematically, AR is defined as

$$AR \equiv \frac{\int_0^1 r_{model}(q)dq - \frac{1}{2}}{\int_0^1 r_{ideal}(q)dq - \frac{1}{2}}, \tag{4}$$

where $r_{model}(q)$ and $r_{ideal}(q)$ indicate $r(q)$ for the scoring model and the ideal case respectively, and the integrals can be approximated by $\frac{1}{N}\sum_{i=1}^{N} r(i/N)$ where N is the number of

observations in the test set.

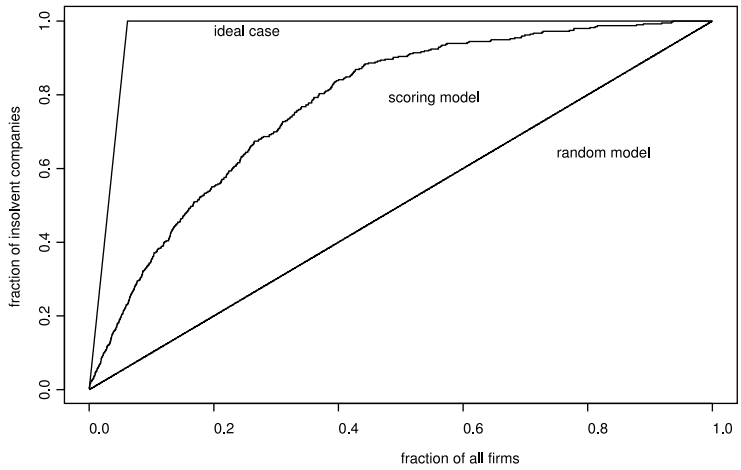


Figure 3: The CAP curve for the test set of the Creditreform data where the scoring model is the BACT model trained using one bootstrap training subset.

We also consider three types of misclassification rates: the overall misclassification rate, the type I misclassification rate and type II misclassification rate. Here type I misclassification refers to the case when the firm is in fact insolvent, but the model classifies the firm as solvent; whereas type II misclassification refers to the case when the firm is in fact solvent, but the model classifies the firm as insolvent. Financial institutions usually seek to keep either type of misclassification rate as low as possible (Sobehart et al., 2001).

Table 5 reports the average values of AR in (4) and the three types of misclassification rates for the Logit model, CART and BACT. Apparently, BACT outperforms the Logit model and CART in all aspects except for average Type I misclassification rate for which BACT is slightly worse than CART.

Table 5: The average values of AR and the three types of misclassification rates for the Logit model, CART and BACT.

Performance Measure	Logit	CART	BACT
AR	52.1%	58.7%	60.4%
Overall Misclassification Rate	30.2%	33.8%	26.6%
Type I Misclassification Rate	28.3%	27.2%	27.6%
Type II Misclassification Rate	30.3%	34.3%	26.5%

Rather than using all data from 2000 to 2002 as the test set, Chen et al. (2007) used a test subset for each bootstrap sample, which comprises of all insolvent firms and a random sample of the same number of solvent firms in the test set. They reported that the median AR value for 30 bootstrap samples was 60.5%, using $\frac{1}{10}\sum_{i=1}^{10} p(i/10)$ to approximate the integrals in calculating the AR value. The median overall misclassification rate was calculated as 28.2%. If we adopt the same procedure, BACT yields a median AR value of 66.5% and median overall classification rate as 27.2%. So BACT also outperforms SVM in identifying the insolvent firms.

5 Concluding Remarks

In this paper, we propose the Bayesian Additive Classification Tree as a general nonlinear classification method. We show that, based on the sum of many trees, the BACT can yield flexible class boundaries, and that it has excellent performance compared with the logit model, CART and SVM, as demonstrated through several benchmark examples and a real application to credit risk modelling.

Because the partitions in each tree depend only on the ordering of the values of the

input variables rather than the values themselves, the BACT is robust to extreme values in the input variables, and the results do not change with monotone transformation of any input variable. Hence little data processing is needed when using the BACT technique. Another thing to note is that although we only discuss binary classification in this paper, extension to multi-class classification is straightforward and left as future research.

Acknowledgement

This work was supported by Deutsche Forschungsgemeinschaft through the SFB 649 "Economic Risk". Junni L. Zhang's research was also sponsored by Chinese NSF grant 10401003 and USA NIH 1 R03 TW007197-01A2.

References

Asuncion, A. and Newman, D. (2007), "UCI Machine Learning Repository," Http://www.ics.uci.edu/∼mlearn/MLRepository.html, University of California, Irvine, School of Information and Computer Sciences.

Berry, M. and Linoff, G. (2000), *Mastering Data Mining*, John Wiley and Sons.

Breiman, L. (2001), "Random forests," *Machine Learning*, 5–32.

Breiman, L., Friedman, J. H., Olshen, R. A., and Stone, C. J. (1984), *Classification and Regression Trees*, CRC Press.

Chen, S., Härdle, W. K., and Moro, R. A. (2007), "Modeling Default Risk with Support Vector Machines," To appear in *Journal of Quantitative Finance*.

Chipman, H. A., George, E. I., and McCulloch, R. E. (1998), "Bayesian CART model search," *Journal of the American Statistical Association*, 935–948.

— (2006), "BART: Bayesian Addtive Regression Trees," Technical Report, Graduate School of Business, University of Chicago.

Denison, D., Mallick, B., and Smith, A. (1998), "A Bayesian CART Algorithm," *Biometrika*, 363–377.

Efron, B., Hastie, T., Johnstone, I., and Tibshirani, R. (2004), "Least angle regression," *Annals of Statitics*, 407–499.

Efron, B. and Tibshirani, R. J. (1993), *An introduction to the bootstrap*, Chapman and Hall.

Engelmann, B., Hayden, E., and Tasche, D. (2003), "Testing rating accuracy," *Risk*, 82–86.

Friedman, J. H. (2001), "Greedy function approximation: A gradient boosting machine," *The Annals of Statistics*, 1189–1232.

Härdle, W. K., Moro, R. A., and Schäfer, D. (2008), "Estimating Probabilities of Default With Support Vector Machines," *to appear in Journal of Banking and Finance*.

Hastie, T. J. and Tibshirani, R. J. (1990), *Generalized Additive Models*, Chapman and Hall.

Sobehart, J. and Keenan, S. (2001), "Measuring default risk accurately," *Risk*.

Sobehart, J., Keenan, S., and Stein, R. (2001), "Benchmarking Quantitative Default Risk Models: A Validation Methodology," *Algo Research Quarterly*.

Tanner, M. A. and Wong, W. H. (1987), "The calculation of posterior distributions by data augmentation (with discussion)," *Journal of American Statistical Association*, 528–550.

Vapnik, V. (1995), *The Nature of Statistical Learning Theory*, Springer, New York, NY.

— (1997), *Statistical Learning Theory*, Wiley, New York, NY.

Wu, Y., Tjelmeland, H., and West, M. (2007), "Bayesian CART: prior specification and posterior simulation," *Journal of Computational and Graphical Statistics*, in press.

Journal of Forecasting
J. Forecast. **29**, 406–433 (2010)
Published online 9 September 2009 in Wiley InterScience
(www.interscience.wiley.com) **DOI**: 10.1002/for.1134

Forecasting Volatility with Support Vector Machine-Based GARCH Model

SHIYI CHEN,[1]* WOLFGANG K. HÄRDLE[2] AND KIHO JEONG[3]

[1] *China Center for Economic Studies, School of Economics, Fudan University, Shanghai, China*
[2] *Center for Applied Statistics and Economics, Humboldt University, Berlin, Germany*
[3] *School of Economics and Trade, Kyungpook National University, Daegu, Republic of Korea*

ABSTRACT

Recently, support vector machine (SVM), a novel artificial neural network (ANN), has been successfully used for financial forecasting. This paper deals with the application of SVM in volatility forecasting under the GARCH framework, the performance of which is compared with simple moving average, standard GARCH, nonlinear EGARCH and traditional ANN-GARCH models by using two evaluation measures and robust Diebold–Mariano tests. The real data used in this study are daily GBP exchange rates and NYSE composite index. Empirical results from both simulation and real data reveal that, under a recursive forecasting scheme, SVM-GARCH models significantly outperform the competing models in most situations of one-period-ahead volatility forecasting, which confirms the theoretical advantage of SVM. The standard GARCH model also performs well in the case of normality and large sample size, while EGARCH model is good at forecasting volatility under the high skewed distribution. The sensitivity analysis to choose SVM parameters and cross-validation to determine the stopping point of the recurrent SVM procedure are also examined in this study.

KEY WORDS (recurrent) support vector machine; GARCH model; volatility forecasting; Diebold–Mariano test

INTRODUCTION

Volatility is important in financial markets since it is a key variable in portfolio optimization, securities valuation and risk management. Much attention of academics and practitioners has been focused on modeling and forecasting volatility in the last few decades (see Franses and McAleer, 2002, and Poon and Granger, 2003, for a comprehensive review). So far in the literature, the predominant model of the past is the GARCH model by Bollerslev (1986), who generalizes the seminal idea on

*Correspondence to: Shiyi Chen, China Center for Economic Studies, School of Economics, Fudan University, Guoquan Road 600, Shanghai, China 200433. E-mail: shiyichen@fudan.edu.cn

ARCH by Engle (1982), and its various extensions; see Li *et al.* (2002) for recent surveys of the models. The GARCH family models, together with the simplest historical price model prevalent in the pre-GARCH era[1] and stochastic volatility model studied a decade later than GARCH development,[2] comprise one of the two broad categories of methods widely used in volatility forecasting, the so-called time series volatility model; another is the market determined option implied volatility model.[3] This paper limits itself mainly to the analysis within the GARCH framework.

The popularity of the GARCH model is due to its ability to capture volatility persistence or clustering, supported by many studies (Akgiray, 1989; Bollerslev *et al.*, 1992; West and Cho, 1995; Andersen and Bollerslev, 1998; Marcucci, 2005). However, some empirical studies report that the GARCH model provides poor forecasting performance (Jorion, 1995, 1996; Brailsford and Faff, 1996; Figlewski, 1997; McMillan *et al.*, 2000; Choudhry and Wu, 2008). To improve the forecasting ability of the GARCH model, some alternative approaches have been advocated by innovating the model specification and estimation,[4] by using different evaluation metrics and definitions of realized volatility,[5] or by enriching the informational content of the model.[6]

As for GARCH model specification and estimation, for example, many financial returns are skewed distributed and nonlinearly dependent such that the linear GARCH model cannot cope with them and therefore forecast of symmetric GARCH model would be biased (Pagan and Schwert, 1990; Bollerslev *et al.*, 1992). To deal with this problem the regime-switching (RS) volatility model is proposed to detect nonlinear behavior in the variance by various tests for asymmetry or threshold

[1]This includes simple moving average method, exponential smoothing method, random walk model, ARMA model, exponentially weighted moving average (EWMA) method and its current extension of Riskmetrics™ model, etc.

[2]The stochastic volatility (SV) model has an additional innovative term in the volatility dynamics (Taylor, 1986). For a detailed discussion on the SV model and its relation to the GARCH class models, see the survey articles by Ghysels *et al.* (1996) and Chib *et al.* (2002), among others.

[3]The time series volatility model is based on historical price information only, while the option implied volatility (IV) model uses market traded option information alone or in addition to historical price sets to forecast volatility. Many studies examine the relative performance of the IV model to forecasting volatility (Day and Lewis, 1992; Lamoureux and Lastrapes, 1993; Pong *et al.*, 2004; Dotsis *et al.*, 2007; Becker *et al.*, 2009; Neely, 2009). This paper limits itself mainly to analysis within the GARCH framework.

[4]Except for the introduction below, other relatively sophisticated GARCH models and estimations include the multivariate GARCH model (Bauwens *et al.*, 2006; Rosenow, 2008), outlier-corrected GARCH model (Park, 2002; Zhang and King, 2005; Ané *et al.*, 2008), Markov chain Monte Carlo (MCMC) sampling techniques to estimate the GARCH model (Gerlach and Tuyl, 2006), other semiparametric or nonparametric specification and estimation such as genetic algorithm, wavelet smoother, kernel density etc. (Franke *et al.*, 2004; Lux and Schornstein, 2005; Renò, 2006; Chen *et al.*, 2008; Feng and McNeil, 2008; Corradi *et al.*, 2009) and combination forecasts from competing approaches (Hu and Tsoukalas, 1999; Dunis and Huang, 2002).

[5]Many studies find that the relative accuracy of various models is also highly sensitive to the measures used to evaluate them (Taylor, 1999; Brooks and Persand, 2003). Most comparisons are based on the average figure of mean absolute error (MAE) and mean square error (MSE) etc. Diebold and Mariano (1995) and West (1996) show how standard errors for MAE and MSE are derived taking into account serial correlation in the forecast errors for statistical inference. Lehar *et al.* (2002) applies value-at-risk (VaR)-oriented evaluation measures to compare the out-of-sample performance. In addition to the symmetric measures of MAE and MSE, Balaban (2004) also uses asymmetric evaluation criteria such as mean mixed error statistics to compare the forecasting performance, penalizing under/over-predictions of volatility more heavily. Recent research has also suggested that this relative failure of GARCH models arises not from a failure of the model but a failure to specify correctly the true volatility measure against which forecasting performance is measured. It is argued that the standard approach of using *ex post* daily squared returns as the measure of true volatility includes a large noisy component. An alternative measure for true volatility has therefore been suggested based on the cumulative squared returns from intra-day data, also referred to as realized, or integrated volatility (Andersen and Bollerslev, 1998; Andersen *et al.*, 2003; Meddahi, 2003; McMillan and Speight, 2004; Galbraith and Kisinbay, 2005; Ghysels *et al.*, 2006).

[6]In many instances, the researchers find the inclusion of implied volatility or trade volume as an exogenous variable in the framework of the GARCH model to be beneficial (Brooks, 1998; Fleming, 1998; Blair *et al.*, 2001; Koopman *et al.*, 2005; Gospodinov *et al.*, 2006; Becker *et al.*, 2007).

DOI: 10.1002/for

nonlinearity (Franses and Dijk, 2000). The first class of RS volatility model assumes that the regime can be determined by an observable variable, including the nonlinear exponential GARCH (EGARCH) model of Nelson (1991), threshold GJR-GARCH model of Glosten *et al.* (1992) and quadratic GARCH model of Engle *et al.* (1993) and Sentana (1995). The second class of RS model for volatility implements GARCH with a Hamilton (1989) type framework that assumes the regime is the realization of a hidden Markov chain, such as (double) Markov switching GARCH model of Gray (1996), Klaassen (2002) and Chen *et al.* (2008).

Both the linear and nonlinear GARCH model described above are parametric and normally estimated jointly by maximum likelihood estimation (MLE). That is, they make specific assumptions about the functional form of the data generation process and the distribution of error terms that is necessary for MLE. Such parametric models are easy to estimate and readily interpretable, but these advantages may come at a cost. Perhaps nonparametric models are better representations of the underlying data generation process. Instead of specifying a particular functional form and making *a priori* distributional assumption, the nonparametric model will search for the best fit over a large set of alternative functional forms. Thus, in the literature, many nonlinear nonparametric GARCH models are developed and still developing fast, among which the artificial neural network (ANN) is extensively used. This paper focuses on one of the neural network algorithms, the support vector machine (SVM), and investigates its forecasting ability of volatility as compared with the simplest moving average method, standard linear GARCH model, nonlinear EGARCH model and traditional recurrent ANN-based nonlinear GARCH model. The moving average method is chosen as the benchmark because some studies find that it provides more accurate forecasts than GARCH models (Dimson and Marsh, 1990; Tse and Tung, 1992; Figlewski, 1997). Among the number of nonlinear parametric GARCH models the EGARCH model is also the most commonly used (Cao and Tsay, 1992; Cumby *et al.*, 1993; Heynen and Kat, 1994; Chong *et al.*, 1999; Hu and Tsoukalas, 1999; Gokcan, 2000; Balaban, 2004).

In recent years, ANN has been successfully used for forecasting financial time series; for recent work, see Fernandez-Rodriguez *et al.* (2000), Qi and Wu (2003), and Pantelidaki and Bunn (2005). The studies in favor of ANN-based GARCH model as opposed to parametric GARCH model in forecasting conditional volatility include Donaldson and Kamstra (1997), Schittenkopf *et al.* (2000), Taylor (2000), Dunis and Huang (2002), Hamid and Iqbal (2004), Ferland and Lalancette (2006), Tseng *et al.* (2008). However, the traditional ANN algorithm also suffers from its own weaknesses such as the need for many controlling parameters, difficulty in obtaining a global solution and the danger of over-fitting (Tay and Cao, 2001). Thus, SVM that can obtain a unique global solution by solving a quadratic programming is developed by Vapnik and his coworkers (1995, 1997). Naturally, SVM also keeps the advantages of conventional ANN such as the flexibility in approximating any nonlinear function arbitrarily well, without *a priori* assumptions about the properties of the data and without the requirement of large sample size that MLE-based parametric GARCH models have. Unlike traditional ANN implementing the empirical risk minimization (ERM) principle, the most particular principle of SVM is to implement the structural risk minimization (SRM), which seeks to achieve a balance between the training error and generalization error, leading, theoretically, to better forecasting performance than traditional ANN (Gunn, 1998; Haykin, 1999). Recently, SVM has gained popularity in predicting financial variables owing to such attractive features (Cao and Tay, 2001; Härdle *et al.*, 2005, 2007; Chen *et al.*, 2009). Pérez-Cruz *et al.* (2003) also propose an SVM-based GARCH (1, 1) model and shows that it provides better volatility forecasts than the standard GARCH model. However, they use the feedforward SVM procedure, which has the same structure as the autoregressive (AR) process and has poor ability

DOI: 10.1002/for

to model a long-time memory. Inspired by the merit of recurrent ANN (Kuan and Liu, 1995; Dunis and Huang, 2002; Bekiros and Georgoutsos, 2008), in this paper we propose a recurrent SVM procedure which can model the ARMA process and apply it to forecast the conditional variance equation of the GARCH model in real data analysis.

The forecasting accuracy of the recurrent SVM-based GARCH model in one-period-ahead volatility forecasting is compared with the competing models in terms of two evaluation metrics of mean absolute error (MAE) and directional accuracy (DA). The statistical hypothesis of equal forecasting accuracy between pairwise models is also investigated by using the Diebold and Mariano (1995) test, calculated according to the Newey–West procedure (Newey and West, 1987). The Diebold and Mariano (DM) test is one of the most important contributions to the study of out-of-sample forecasting accuracy evaluation over the past two decades, and has been further generalized and extensively used in many studies since then (Corradi and Swanson, 2004; Awartani and Corradi, 2005; Preminger and Franck, 2007; Taylor, 2008; Groen *et al.*, 2009; Wong and Tu, 2009).

This paper is organized as follows. The next section briefly introduces the theory of SVM. The third section specifies the empirical model and forecasting scheme. The fourth section uses the Monte Carlo simulation to evaluate how the models perform under controlled conditions. The fifth section describes the GBP exchange rates and NYSE composite index data and discusses the volatility forecasting performance of all models for the real data. The paper concludes with the sixth section.

SUPPORT VECTOR MACHINE

The support vector machine (SVM) originates from Vapnik's statistical learning theory (Vapnik, 1995, 1997), which has the design of a feedforward network with an input layer, a single hidden layer of nonlinear units and an output layer, and formulates the regression problem as a quadratic programming (QP) problem. SVM estimates a function by nonlinearly mapping the input space into a high-dimensional hidden space and then running the linear regression in the output space. Thus, the linear regression in the output space corresponds to a nonlinear regression in the low-dimensional input space. The theory denotes that if the dimensions of feature space (or hidden space) are high enough, SVM may approximate any nonlinear mapping relations. As the name implies, the design of the SVM hinges upon the extraction of a subset of the training data that serves as support vectors, which represent a stable characteristic of the data.

Given a training dataset $(\mathbf{x_t}, y_t)$, where input vector $\mathbf{x_t} \in \mathbb{R}^p$ and output scalar $y_t \in \mathbb{R}^1$. Indeed, the desired response y, known as a 'teacher', represents the optimum action to be performed by the SVM. We aim at finding a sample regression function $f(\mathbf{x})$, or denoted by $\hat{y}$, as below to approximate the latent, unknown decision function $g(\mathbf{x})$:

$$f(\mathbf{x}) = \mathbf{w}^T \phi(\mathbf{x}) + b \tag{1}$$

where the superscript T is a transposing operator that should be differentiated from the sample size T of the time series used later in this paper. In equation (1), $\phi(\mathbf{x}) = [\phi_1(\mathbf{x}), \ldots, \phi_l(\mathbf{x})]^T$, $\mathbf{w} = [w_1, \ldots, w_l]^T$. The $\phi(\mathbf{x})$ is known as the nonlinear transfer function which represents the features of the input space and projects the inputs into the feature space. The dimension of the feature space is l, which is directly related to the capacity of the SVM to approximate a smooth input–output mapping; the higher the dimension of the feature space, the more accurate the approximation will be. Parameter

DOI: 10.1002/for

w denotes a set of linear weights connecting the feature space to the output space, and b is the threshold.

To get the function $f(\mathbf{x})$, the optimal $\mathbf{w}^*$ and b^* have to be estimated from the data. First, we define a linear ε-insensitive loss function, L_ε, originally proposed by Vapnik (1995):

$$L_\varepsilon(\mathbf{x}, y, f(\mathbf{x})) = \begin{cases} |y - f(\mathbf{x})| - \varepsilon & \text{for } |y - f(\mathbf{x})| \geqslant \varepsilon \\ 0 & \text{otherwise} \end{cases} \tag{2}$$

This function indicates the fact that it does not penalize errors below ε. The training points within the ε-tube have no loss and do not provide any information for decision. Therefore, these points do not appear in the decision function $f(\mathbf{x})$. Only those data points located on or outside the ε-tube will serve as the support vectors and are finally used to construct the $f(\mathbf{x})$. This property of sparseness algorithm results only from the ε-insensitive loss function and greatly simplifies the computation of SVM. The non-negative slack variables, ξ and ξ' (below or above the ε-tube, or denoted together by $\xi^{(\prime)}$; see Figure 1) are employed to describe this kind of ε-insensitive loss.

The derivation of SVM follows the principle of structural risk minimization (SRM) that is rooted in the Vapnik–Chervonenkis (VC) dimension theory (Haykin, 1999). Structural risk is the upper boundary of empirical loss, denoted by ε-insensitive loss function, plus the confidence interval (or called margin), which is constructed in equation (3). The primal constrained optimization problem of SVM is obtained below:

$$\min_{\mathbf{w}\in\mathbb{R}^l, \xi(')\in\mathbb{R}^{2T}, b\in\mathbb{R}} \mathbf{C}(\mathbf{w}, b, \xi_t, \xi_t') = \frac{1}{2}\|\mathbf{w}\|^2 + C\sum_{t=1}^{T}(\xi_t + \xi_t') \tag{3}$$

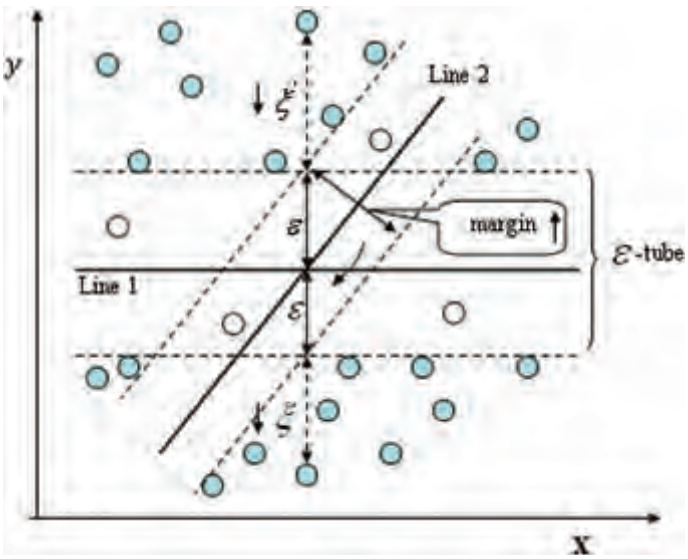


Figure 1. Principle of structural risk minimization (SRM) of SVM

DOI: 10.1002/for

such that

$$\mathbf{w}^T\phi(\mathbf{x_t}) + b - y_t \leqslant \varepsilon + \xi_t \tag{4}$$

$$y_t - \mathbf{w}^T\phi(\mathbf{x_t}) - b \leqslant \varepsilon + \xi_t' \tag{5}$$

$$\xi_t \geqslant 0, \xi_t' \geqslant 0, t = 1, 2, \ldots, T \tag{6}$$

The formulation of the cost function $\mathbf{C}(\cdot)$ in equation (3) is in perfect accord with the SRM principle, which is illustrated in Figure 1 (in which the dark circles are data points extracted as support vectors). In equation (3), the first term indicates the Euclidean norm of the weight vector $\mathbf{w}(\|\mathbf{w}\|^2 = \mathbf{w}^T\mathbf{w})$ and measures the function flatness; to minimize it is equivalent to maximizing the separation margin $(2/\|\mathbf{w}\|)$, that is, maximizing the generalization ability. The second term represents the empirical risk loss determined by the ε-insensitive loss function and is similar to the sum of residual squares in the objective function of ANN. Finally, SVM obtains the tradeoff between the two terms; as a result, it not only fits the historical data well but also forecasts the future data excellently. As shown in Figure 1, both regression lines 1 and 2 can classify the data points correctly and then minimize the empirical loss; however, the separation margin of the two lines are different, in which the regression line 1 has the larger margin. It is the special design of minimizing the structural risk that endows SVM with the excellent forecasting ability among all candidates. In addition, the convex quadratic programming and linear restrictions in the above primal problem ensure that SVM can always obtain the global unique optimal solution, which is different from the usual networks that easily get trapped in local minima. The penalty parameter $C > 0$ controls the penalizing extent on the sample points which lie outside ε-tube. Both ε and C, the free parameter of SVM, must be selected by the user.

The corresponding dual problem of the SVM can be derived from the primal problem by using the Karush–Kuhn–Tucker conditions as follows:

$$\min_{\alpha_t^{(\prime)} \in \mathbb{R}^{2T}} \frac{1}{2}\sum_{s=1}^{T}\sum_{t=1}^{T}(\alpha_s' - \alpha_s)(\alpha_t' - \alpha_t)K(x_s \cdot x_t) + \varepsilon\sum_{t=1}^{T}(\alpha_t' + \alpha_t) - \sum_{t=1}^{T} y_t(\alpha_t' - \alpha_t) \tag{7}$$

such that

$$\sum_{t=1}^{T}(\alpha_t - \alpha_t') = 0 \tag{8}$$

$$0 \leqslant \alpha_t, \alpha_t' \leqslant Cs, t = 1, 2, \ldots, T \tag{9}$$

where α_t and α_t' (or $\alpha_t^{(\prime)}$) are the Lagrange multipliers. The dual problem can be solved more easily than the primal problem (Scholkopf and Smola, 2001; Deng and Tian, 2004). Making use of any solution of α_t and α_t', the optimal solutions of the primal problem can be calculated in which $\mathbf{w}^*$ is unique and expressed as follows:

$$\mathbf{w}^* = \sum_{t=1}^{T}(\alpha_t' - \alpha_t)\phi(\mathbf{x_t}) \tag{10}$$

DOI: 10.1002/for

However, b^* is not unique and formulated in terms of different cases. If $i \in \{t | \alpha_t \in (0, C)\}$, then

$$b^* = y_t - \sum_{t=1}^{T} (\alpha_t' - \alpha_t) K(\mathbf{x_t} \cdot \mathbf{x_i}) + \varepsilon \tag{11}$$

If $j \in \{t | \alpha_t' \in (0, C)\}$, then

$$b^* = y_j - \sum_{t=1}^{T} (\alpha_t' - \alpha_t) K(\mathbf{x_t} \cdot \mathbf{x_j}) - \varepsilon \tag{12}$$

The cases of both $i, j \in \{t | \alpha_t^{(\prime)} = 0\}$ and $i, j \in \{t | \alpha_t^{(\prime)} = C\}$ rarely occur in reality.

Thus the regression decision function $f(\mathbf{x})$ will be computed by using $\mathbf{w}^*$ and b^* in the following forms:

$$\begin{aligned} f(\mathbf{x}) &= \mathbf{w}^{*\mathbf{T}} \phi(\mathbf{x}) + b^* \\ &= \sum_{t=1}^{T} (\alpha_t' - \alpha_t) \phi^T(\mathbf{x_t}) \phi(\mathbf{x}) + b^* \\ &= \sum_{t=1}^{T} (\alpha_t' - \alpha_t) K(\mathbf{x_t}, \mathbf{x}) + b^* \end{aligned} \tag{13}$$

where $K(\mathbf{x_t}, \mathbf{x}) = \phi^T(\mathbf{x_t})\phi(\mathbf{x})$ is the inner-product kernel function. In fact, the SVM theory considers only the form of $K(\mathbf{x_t}, \mathbf{x})$ in the feature space without specifying explicitly $\phi(\mathbf{x})$ and without computing all corresponding inner products. Therefore, the kernel function greatly reduces the computational complexity of high-dimensional hidden space and becomes the crucial part of SVM. The function which satisfies the Mercer theorem can be chosen as the SVM kernel. No analytical method is currently available to determine the most suitable kernel for a particular dataset. This paper experiments with three different kernels to investigate the effect of a kernel type in Monte Carlo simulation:

$$\text{Linear:} \quad K(\mathbf{x_t}, \mathbf{x}) = \mathbf{x}_t^T \mathbf{x} \tag{14}$$

$$\text{Polynomial:} \quad K(\mathbf{x_t}, \mathbf{x}) = \left(\mathbf{x}_t^T \mathbf{x} + 1\right)^d \tag{15}$$

$$\text{Gaussian:} \quad K(\mathbf{x_t}, \mathbf{x}) = \exp\left(\frac{-\|\mathbf{x} - \mathbf{x_t}\|^2}{2\sigma^2}\right) \tag{16}$$

where d and σ^2 are the parameters for the polynomial and Gaussian kernel. Before implementation of the SVM, the appropriate values of the coefficients ε, C, d and σ^2 must be determined in advance through cross-validation. The sensitivity analysis of the parameters and the kernel type will be illustrated by using the simulated data below ('Monte Carlo Simulation').

DOI: 10.1002/for

EMPIRICAL MODELING

In this study, the forecasts are obtained first by applying the Monte Carlo Simulation, following the suggestions in Andersen and Bollerslev (1998) and Clements and Smith (1999, 2001). The main motivation for conducting a simulation experiment is that, since the true volatility is known, the candidate volatility measures can be compared with certainty. We then fit each of the models to the daily returns on the GBP exchange rate and NYSE stock indexes and forecast their respective volatility. The empirical modeling and forecasting scheme described below are employed for both simulation and real data.

Model specification

In this paper the real data we analyze are the daily financial returns, y_t, converted from the corresponding price or index, I_t, using continuous compounding transformation as

$$y_t = 100 \times (\log I_{t+1} - \log I_t) \tag{17}$$

Empirical findings suggest that GARCH is a more parsimonious model than ARCH, and GARCH (1, 1) specification is sufficient to model the variance changing over long sample periods and has become the most popular structure when capturing financial volatility (Akgiray, 1989; Franses and Dijk, 1996; Brooks, 1998; Gokcan, 2000; Andersson, 2001; Brooks and Persand, 2003; Poon and Granger, 2003; Gerlach and Tuyl, 2006). As such, throughout the paper, the analysis is restricted to the case of the GARCH (1, 1) process for the second conditional variance function and the AR(1)[7] process for the conditional mean equation, for the sake of candidate comparison under the same conditions.

Thus the linear standard GARCH (1, 1) model is specified as follows:

$$y_t = c + \phi_1 y_{t-1} + u_t \quad u_t \sim N(0, h_t) \tag{18a}$$

$$h_t = \kappa + \delta_1 h_{t-1} + \alpha_1 u_{t-1}^2 \tag{18b}$$

where c, ϕ_1, κ, δ_1 and α_1 are constant parameters. Such restrictions on the parameters that κ, δ_1 and α_1 are non-negative and $\delta_1 + \alpha_1 < 1$ prevent negative variances (Bollerslev, 1986).

All odd moments of u_t in the standard GARCH model equal zero, and hence u_t and y_t are symmetric time series. The nonlinear EGARCH (1, 1) model that is able to capture the asymmetry is similar to the linear GARCH model but the h_t process is given by

$$\log(h_t) = \kappa + \delta_1 \log(h_{t-1}) + \alpha_1 \left(\frac{|u_{t-1}|}{\sqrt{h_{t-1}}} - \sqrt{2/\pi} \right) + \beta_1 \frac{u_{t-1}}{\sqrt{h_{t-1}}} \tag{19}$$

where κ, δ_1, α_1 and β_1 are the constant parameters. The EGARCH model is fundamentally different from the standard GARCH model in that the standardized innovation serves as the forcing variable for the conditional variance. Also, there are no restrictions on the parameters to ensure non-negativity

[7] Franses and Dijk (1996) also denote that the order of autoregression in the first conditional mean equation of the GARCH framework is usually 0 or small. Thus, the order 1 is specified for this study.

DOI: 10.1002/for

of the variances. The coefficient β_1 is introduced to capture the asymmetry. If $\beta_1 = 0$, a positive return shock has the same effect on h_t as the negative return shock of the same amount; if $\beta_1 < 0$, a positive return shock actually reduces h_t; if $\beta_1 > 0$, then a positive return shock increases h_t. Previous studies have viewed this coefficient as typically negative, indicating that negative return shocks normally generate more volatility than positive return shocks, so generating the so-called leverage effect.

The conditional variance of u_t is given by $h_t = E_{t-1}u_t^2 = \hat{u}_{t|t-1}^2$. Roughly speaking, in a GARCH process the conditional variances can be modeled by an ARMA type process (Franses and Dijk, 1996). For instance, the ARMA process of the conditional variance of u_t in a linear GARCH model can be expressed as below (Hamilton, 1997; Enders, 2004):

$$u_t^2 = \kappa + (\delta_1 + \alpha_1)u_{t-1}^2 + w_t - \delta_1 w_{t-1} \tag{20}$$

where $w_t \equiv u_t^2 - \hat{u}_{t|t-1}^2 = u_t^2 - h_t$, which is white noisy error. Inspired by this, the nonparametric recurrent ANN and SVM based nonlinear GARCH (1, 1) model is specified as the following form:

$$y_t = f(y_{t-1}) + u_t \tag{21a}$$

$$u_t^2 = g\left(u_{t-1}^2, w_{t-1}\right) + w_t \tag{21b}$$

where $f(\cdot)$ and $g(\cdot)$ are nonlinear nonparametric function forms for conditional mean and variance equations, respectively. Note that equation (21b) is adopted for the analysis of real data because the actual volatility h_t is unobservable, while in the case of simulation the conditional variance equation is just specified as $h_t = f(h_{t-1}, u_{t-1}^2)$ due to h_t being known. Because of the way GARCH (1, 1) class models are constructed, the volatility is known at time $t - 1$. Thus the one-step-ahead forecast of volatility is readily available.

The moving average method uses weighted moving averages of past squared innovations to forecast volatility (Niemira and Klein, 1994). For simulated data, the moving average forecast for the next-day volatility, using the five most recent observations, is expressed as

$$\hat{u}_{t+1}^2 = \frac{1}{5}\sum_{j=t-4}^{t} u_j^2 \tag{22}$$

For real data, the moving average forecast for the next-day volatility is expressed as (Engle *et al.*, 1993)

$$\hat{u}_{t+1}^2 = \frac{1}{5}\sum_{j=t-4}^{t} \left(y_j - \bar{y}_{5,t}\right)^2 \tag{23}$$

where

$$\bar{y}_{5,t} = \frac{1}{5}\sum_{j=t-4}^{t} y_j$$

The recurrent ANN used in this study is the feedback multilayer perceptrons (MLP) network with the addition of a global feedback connection from the output layer to its input space. We specify

DOI: 10.1002/for

this kind of recurrent back-propagation network with the following architecture: one nonlinear hidden layer with four neurons, each using a tan-sigmoid differentiable transfer function to generate the output, and one linear output layer with one neuron. As a training algorithm, the fast training Levenberg–Marquardt algorithm is chosen. The value of the learning rate parameter used in the training process is set to be 0.05. These specifications and choices are standard in the neural network literature.

Recurrent SVM procedure

As Haykin (1999) said, the standard SVM described above usually appears in the design of a simple network in which an input layer of source nodes projects onto an output layer of computation node, but not vice versa (see Figure 2(a)). This process is known as feedforward SVM and could be easily employed to estimate such AR process as the first conditional mean function (21a), $y_t = f(y_{t-1}) + u_t$, and the second conditional variance function in the situation of simulation, $h_t = f(h_{t-1}, u_{t-1}^2)$. However, because the unobservable error term w_t is introduced into the GARCH model which indeed exhibits the nonlinear ARMA process, how to estimate the conditional volatility model (21b) for real data?

To estimate the nonlinear ARMA model, a feedback process of SVM with unobservable moving average part as inputs, not addressed before our application[8], has to be described, which distinguishes itself from feedforward SVM in that it has at least one feedback loop (see Figure 2(b)). In this paper, we abuse terminology and refer to this process as 'recurrent SVM'. The feedback loops involve the use of particular branches composed of *one-delay operator*, z^{-1}, which result in nonlinear dynamical behavior and have a profound impact on the learning capability of SVM. Thus the recurrent SVM will capture more dynamic characteristics of y_t than does feedforward SVM.

To overcome the problem that the series of error term w_t is unavailable, we employ the model residuals as estimates of the errors in an iterative way, which is similar to the way that the linear ARMA model is iteratively estimated by MLE (Box *et al.*, 1994; Hamilton, 1997). Likewise, the

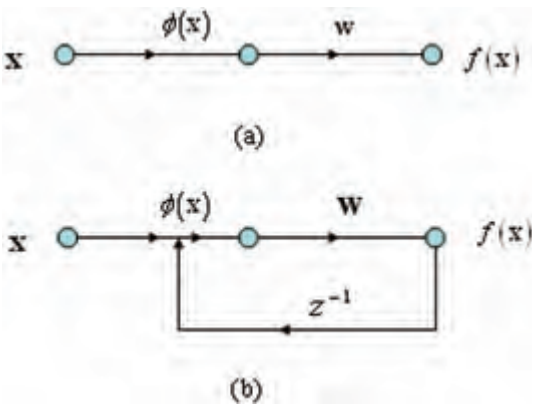


Figure 2. Signal-flow graphs of feedforward and recurrent SVM. (a) Signal-flow graph of a feedforward SVW. (b) Signal-flow graph of a single-loop recurrent SVW

[8] Suykens and Vandewalle (2000) proposed the algorithm of recurrent least squares SVM. The difference between the two recurrent SVM algorithms is their sparseness solutions.

DOI: 10.1002/for

error term is initially set to be its expectation: zero. The empirical procedure of the recurrent SVM executed during the training phase is described as follows. The letter i indicates the iterative epoch and t denotes the period:

- Step 1: Set $i = 1$ and star with all residuals at zero: $w_t^{(1)} = 0$.
- Step 2: Run an SVM procedure to get the decision function $f^{(i)}$ to the points $\{x_t, y_t\} = \{u_{t-1}^2, u_t^2\}$ with all inputs $x_t = \{u_{t-1}^2, w_{t-1}^{(i)}\}$.
- Step 3: Compute the new residuals $w_t^{(i+1)} = u_t^2 - f^{(i)}$.
- Step 4: Terminate the computational process when the stopping criterion is satisfied; otherwise, set $i = i + 1$ and go back to Step 2.

Note that the first iterative epoch is in fact a feedforward SVM process and results in an AR (1) model and that the following epochs provide results of the ARMA (1, 1) model, being estimated by the recurrent SVM.

In general, the procedure cannot be shown to converge, and there are no well-defined criteria for stopping its operation. Rather, some reasonable criteria can be found, although with its own practical drawback, which may be used to terminate the computational process.

To formulate such a criterion, it is logical to think in terms of the properties of the estimated residual series. After sufficiently long iterative steps, the autocorrelation displayed behind the residuals during the first AR epoch should disappear, and the information in the residual behavior has been completely adopted and the final residual series should be white noisy. Accordingly, we may suggest a sensible convergence criterion for the recurrent SVM procedure as follows:

> *The recurrent SVM procedure is considered to have converged when the corresponding residuals become white noisy, or has no autocorrelation.*

To quantify the measurement of white noise, we use the formal hypothesis test, the Ljung–Box–Pierce Q-test, to investigate a departure from randomness based on the ACF of the residuals. Under the null hypothesis of no autocorrelation in residuals, the Q-test statistic is asymptotically distributed as chi-square. In fact, we just check the actual p-values (exact level of significance) of the Q-test of lag 1. It is reasonable to think there is no higher-order autocorrelation if there is no one-order autocorrelation in residuals. Only if the p-values of the Q-test for five consecutive epochs are simultaneously higher than 0.1 is the iterative computational process stopped. To overcome the drawback of this convergence criterion, we use cross-validation to avoid the possible over-fitting problem; see 'Real data analysis' below for the iterative process in detail.

Forecasting scheme

To illustrate the forecasting scheme, the SVM-GARCH model is also exemplified. First, estimate the conditional mean equation (21a) by using the feedforward SVM in the full sample period $T(1, 2, \ldots, T)$ to obtain residuals, $u_1, u_2, \ldots, u_T$. Then, recursively run the SVM-GARCH (1, 1) model for squared residuals thus obtained to forecast the one-period-ahead volatility. The recursive forecasting scheme is employed with an updating sample window; the estimating and forecasting process is carried out recursively by updating the sample with one observation each time, rerunning the SVM approach and recalculating the model parameters and corresponding forecasts. Here, the SVM approach to estimate the conditional volatility is feedforward for simulation and recurrent, as described in the above subsection, for real data. The first training sample is $u_1^2, u_2^2, \ldots, u_{T_1}^2$ ($T_1 < T$). The observations of $T - T_1$ are retained as a forecasting or test sample.

DOI: 10.1002/for

Therefore, we can estimate and forecast the SVM-based conditional volatility equation for $n = T - T_1$ times. We set $n = 60$ for both simulation and real data in this study. Thus, 60 one-period-ahead forecast volatilities, $\hat{u}^2_{T-59}$, $\hat{u}^2_{T-58}, \ldots, \hat{u}^2_{T-1}$, $\hat{u}^2_T$, will be acquired for out-of-sample forecasting evaluation.

Evaluation measures and pairwise comparison of competing models

We evaluate the forecasting performance using two standard statistical criteria: mean absolute forecast error (MAE) and directional accuracy (DA), expressed as follows (Brooks, 1998; Moosa, 2000):

$$\mathrm{MAE} = \frac{1}{n}\sum_{t=T_1}^{T-1}\left|u^2_{t+1} - \hat{u}^2_{t+1}\right| \tag{24}$$

$$\mathrm{DA}(\%) = \frac{100}{n}\sum_{t=T_1}^{T-1} a_t \tag{25}$$

where

$$a_t = \begin{cases} 1 & \left(u^2_{t+1} - u^2_t\right)\left(\hat{u}^2_{t+1} - \hat{u}^2_t\right) \geqslant 0 \\ 0 & \text{otherwise} \end{cases}$$

MAE measures the average magnitude of forecasting error which disproportionately weights large forecast errors more gently relative to MSE; and DA measures the correctness of the turning point forecasts, which gives a rough indication of the average direction of the forecast volatility.

The fundamental problem with the evaluation of volatility forecasts of real data is that volatility is unobservable and so actual values with which to compare the forecasts do not exist. Therefore, researchers are necessarily required to make an auxiliary assumption about how the actual *ex post* volatility is calculated. In this paper, we use the square of the return minus its mean value as the surrogate of actual volatility against which MAE and DA can be calculated. This approach is similar to the standard one, squared returns, because the mean of returns is usually close to zero. The proxy of actual volatility in real data is expressed as

$$u^2_t = (y_t - \bar{y})^2 \tag{26}$$

where y_t is returns and $\bar{y}$ is mean of returns. This proxy has been used in many recent papers, such as Pagan and Schwert (1990), Day and Lewis (1992), Chan *et al.* (1995), West and Cho (1995), Chong *et al.* (1999), Brooks (2001) and Brooks and Persand (2003).

To test for equal forecasting accuracy of two competing models, we use the two-sided DM test statistic proposed by Diebold and Mariano (1995) for the difference of MAE loss function. The null and alternative hypotheses in this case are

$$H_0: \mathrm{MAE}_1 - \mathrm{MAE}_0 = 0 \textit{ versus } H_1: \mathrm{MAE}_1 - \mathrm{MAE}_0 \neq 0$$

DOI: 10.1002/for

where the subscript 0 denotes the benchmark model and 1 the competing model. The DM statistic in a robust form is then based on the following large sample statistic:

$$\mathrm{DM} = \frac{1}{\sqrt{n}} \frac{1}{\sqrt{\hat{S}^2}} \sum_{t=T_1}^{T-1} \left(\left| u_{t+1}^2 - \hat{u}_{1,t+1}^2 \right| - \left| u_{t+1}^2 - \hat{u}_{0,t+1}^2 \right| \right) \sim N(0,1) \tag{27}$$

where $\hat{S}^2$ denotes a heteroscedasticity and autocorrelation consistent (HAC) robust (co)variance matrix which is estimated according to the Newey–West procedure (Newey and West, 1987). We use Andrews' (1991) approximation rule to automatically select the number of lags for the HAC matrix. If n grows at a rate such that as $T \to \infty$, $n \to \infty$ and $n/T_1 \to 0$, then the DM statistic converges in distribution to a standard normal.

MONTE CARLO SIMULATION

Data-generating process

In this section we investigate the forecasting performance of all candidates using artificial simulated data under controlled conditions. To generate the data, we first need to parameterize the GARCH (1, 1) model in equation (18) with the following settings (c, ϕ_1, κ, δ_1, α_1) = (0, 0.5, 0.0005, 0.8, 0.1) for medium persistence and a disturbance term u_t distributed first as Gaussian and then as a Student's t with five degrees of freedom (kurtosis = 5). The second distribution tries to model the skewness and excess of kurtosis that usually appears in real financial series. Using the same specified models, two artificial samples of size 500 and 1000 are created under a two-distributions assumption, giving a total of four situations. To limit the computational burden, each situation is replicated only 50 times. Then the multiple simulated y_t and h_t are 500 × 50 and 1000 × 50 element matrices for different distribution.

Parameter selection

The use of cross-validation is appealing particularly when we have to design a somewhat complex approach with good generalization as the goal. For example, here we may use cross-validation to determine the values of free parameters of SVM with the best performance. One series of 50 simulated returns and volatility of 1000 size and Student's t distribution, one of the four situations, is exemplified as below. The first training data, that is, the former 940 observations, are used to determine the appropriate values taken by the free parameters. The training data are further randomly partitioned into two disjoint subsets: estimating sample and validating sample (700 and 240 observations, respectively).

As shown above, two free parameters (ε and C) and two kernel coefficients (d and σ^2) have to be selected by users before running the SVM procedure. The motivation for using cross-validation here is to validate the model on a dataset different from the one used for parameter estimation. In this way we may use the training set to assess the performance of various values of parameters, and thereby choose the best one. The sensitivity investigation of SVM (represented by the generalization error, MAE) with respect to four parameters is illustrated in Figures 3 and 4 for conditional mean and variance estimation, respectively.

Figure 3 describes the sensitivity analysis for the conditional mean equation. Parameter C varies from a very small value of 0.0001 to infinity, with ε being fixed at 0.0001 and σ^2 0.4. Clearly, when

DOI: 10.1002/for

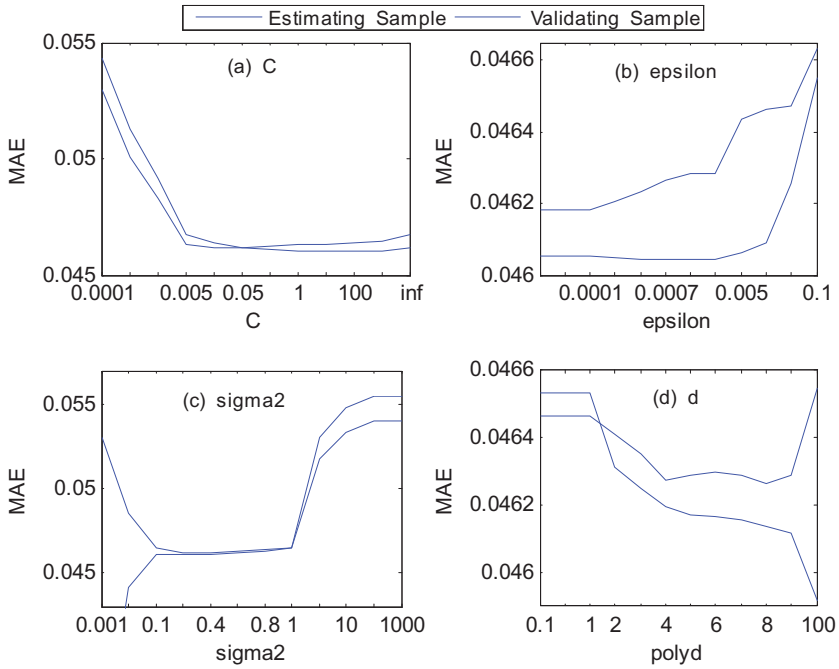


Figure 3. Sensitivity analysis of SVM in conditional mean estimation

$C = 0.05$, MAE of the validation sample obtains the lowest value, 0.046. Parameter ε takes values in the range [0.00001, 0.00005, 0.0001, 0.0003, 0.0005, 0.0007, 0.0009, 0.001, 0.005, 0.01, 0.05, 0.1], with $C = 0.05$ and $\sigma^2 = 0.4$. The values of ε to the left of the point = 0.0001 have no influence on the performance of SVM. Coefficient σ^2 varies from values of 0.001 to 1000, with C being 0.05 and 0.0001. Obviously, the value of $\sigma^2 = 0.4$ leads to the best validation performance. If we set $C = 0.05$ and 0.0001 and the polynomial kernel parameter d = [0.1, 0.5, 1, 2, 3, 4, 5, 6, 7, 8, 10, 100], the validating MAE attains the minima when $d = 8$; after that, over-fitting the training set occurs. Note that the polynomial kernel with $d = 1$ is similar to the linear kernel. Thus, the appropriate parameters of SVM for the conditional mean returns are: $C = 0.05$, $\varepsilon = 0.0001$, $\sigma^2 = 0.4$ and $d = 8$.

Figure 4 describes the parameter selection process for conditional variance series. Similar to the return series, the MAE of both estimating and validating sample decreases as the values of C increase and become stable when C takes a value greater than 10; in contrast to C, as the values of ε increase, both MAE of SVM are considerably more stable before the point of $\varepsilon = 0.0001$ and increase slowly, and sharply after $\varepsilon = 0.001$. The value of $\sigma^2 = 0.01$ results in the best validation performance; namely, its MAE reaches the minimum value, about 0.000065. The values of d taken between 100 and 1000 have not much effect on the performance of SVM but after that range the over-fitting phenomenon becomes serious. Likewise, when one parameter is analyzed, the others are set to be fixed. Therefore, the correct parameters chosen for the conditional variance series are $C = 10$, $\varepsilon = 0.00005$, $\sigma^2 = 0.01$ and $d = 250$, respectively.

DOI: 10.1002/for

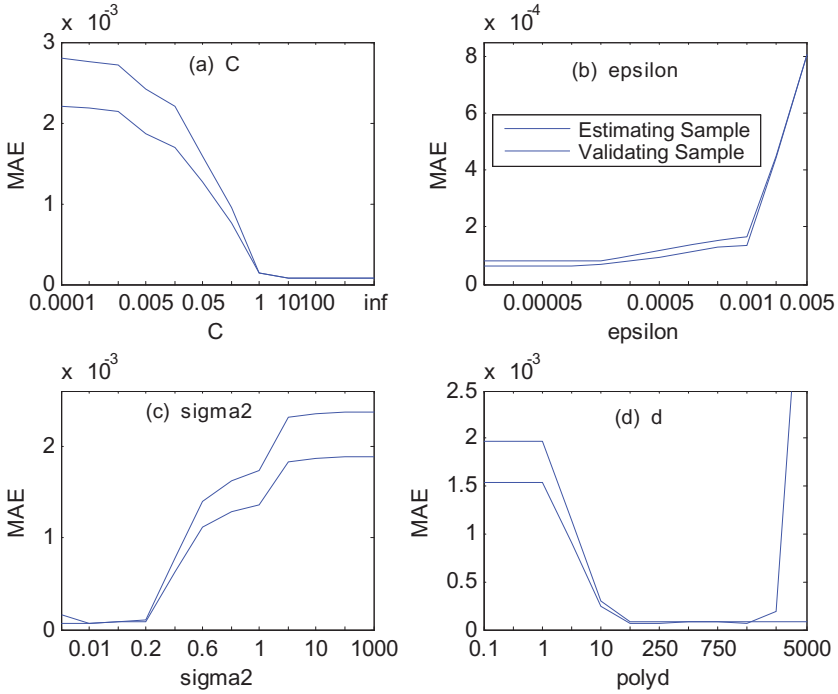


Figure 4. Sensitivity analysis of SVM in conditional variance estimation

Thus far we discuss the sensitivity investigation of parameters by using the simulated data with 1000 observations and *t* distribution. The parameter selection for the other three random samples is similar to this and not reported here to save space.

EFFECT OF KERNEL TYPE AND FORECASTING EVALUATION

There is still the possibility of over-fitting after training. Therefore, the generalization performance of the competing models is further measured and evaluated on the test set, which is different from the validation subset. For the simulated data, the forecasting sample is the last 60 observations. For each replication, the SVM-based GARCH (1, 1) model and the others are estimated, and the forecasting errors are calculated using the forecasting schemes described above. The results of out-of-sample one-period-ahead volatility forecasting measures for four situations are shown in Table I. The reported results are the mean values of 50 independent replications. Table II presents the *p*-values of Diebold-Mariano (DM) test for the MAE difference, which are defined as the significance levels at which the null hypothesis under investigation can be rejected. In calculating the DM statistic, the null hypothesis of equal forecasting ability is related to the four benchmark models: moving average, standard GARCH, EGARCH and traditional ANN models. We report the results of the DM test, say DM1, in the third and seventh columns for two simulated series, respectively, under the null hypothesis that the absolute forecast error produced by the moving average method is equal to those obtained using the other models. DM2, DM3 and DM4 are organized in the same manner and show the test results when the benchmark models are respectively the standard GARCH, EGARCH and recurrent ANN models. The DM tests in this study are investigated in a robust form, by simply

DOI: 10.1002/for

Table I. Diebold–Mariano test for the MAE difference on real data

Models	Sample Size = 500				Sample Size = 1000			
	Normality		Student's *t*		Normality		Student's *t*	
	MAE	DA	MAE	DA	MAE	DA	MAE	DA
Moving Average	0.0001276	44.07	0.0001747	59.32	0.0001198	54.24	0.0002130	40.68
Standard GARCH	0.0000972	76.27	0.0001765	55.93	0.0000488	79.66	0.0001083	59.32
EGARCH	0.0001312	67.80	0.0002075	64.41	0.0000730	57.63	0.0001864	74.58
ANN-GARCH	0.0001517	72.88	0.0002481	57.63	0.0000904	62.71	0.0001442	67.80
SVMl-GARCH	0.0000960	76.27	0.0001369	71.19	0.0000501	74.58	0.0000715	72.88
SVMp-GARCH	0.0000924	76.27	0.0001371	71.19	0.0000479	71.19	0.0000714	77.97
SVMg-GARCH	0.0000796	86.44	0.0001397	81.36	0.0000456	83.05	0.0000769	98.31

Note: SVMl, SVMp and SVMg represent the SVM with linear, polynomial and Gaussian kernel, respectively, for short.

scaling the numerator by a heteroscedasticity and autocorrelation consistent (HAC) (co)variance matrix calculated according to Newey-West procedures (Newey and West, 1987).

Table I firstly shows the effect of kernel functions on out-of-sample forecasting performance of SVM. The linear kernel behaves better in the sample with 500 sizes and *t* distribution based on DA measure. The polynomial kernel is the most suitable for forecasting the *t*-distributed 1000 sample size also based on DA. For all the other six cases, the Gaussian kernel looks promising, however, which is not a general conclusion but only true for the case we are studying. As a whole, three types of kernel-based SVM have a similar volatility forecasting performance and almost behave better than the benchmarks. Since no single kernel function dominates all volatility predictions, practitioners could try any kernel function. In the real data analysis later, for example, we only investigate the performance of the Gaussian kernel-based SVM-GARCH model.

Now, based on Table I, we revert to comparing the volatility forecasting ability among all competing models. In terms of the average ranking of MAE measures, the order of the forecasting ability of the different methods from highest to lowest is displayed in turn as follows: SVMp-GARCH, SVMg-GARCH, SVMl-GARCH[9], standard GARCH, EGARCH, moving average and ANN-GARCH model. Concretely, in the situation of normal distribution, the standard GARCH model behaves not badly, which is ranked fourth (only inferior to three SVM models) in the 500 sizes and even ranked third (only defeated by Gaussian and polynomial SVM models) in the series of 1000 sizes. Even though the data satisfy the normality assumption that is required for MLE in the standard GARCH model, the SVM-GARCH models still outperform it in forecasting the magnitude of the volatility error. Nonlinear EGARCH and ANN-GARCH models perform worse than the linear GARCH model. In the situation of *t* distribution, the forecasting performance of the linear GARCH model grows poorer and the difference of MAE values between SVM-GARCH and standard MLE-GARCH models becomes larger than that under normality. Possibly this results from the fact that the normality assumption required for MLE is violated but it is not necessary for the SVM method. Not as expected, the asymmetric EGARCH model is weak in reducing the forecasting error even in the case of skewed distribution.

Based on the DA measures in Table I, on average, the Gaussian SVM-GARCH model ranks highest (for all four situations) in forecasting volatility directions, followed by polynomial and linear

[9]That is, corresponding to SVM-based GARCH models with polynomial, Gaussian and linear kernel function, respectively.

DOI: 10.1002/for

SVM-GARCH models, linear GARCH model, EGARCH model, ANN-GARCH model and moving average, in turn. In the situation of the normal distribution, the standard GARCH model behaves even better than forecasting error magnitude—ranked second for both the series of 500 sizes (only inferior to Gaussian but equal to linear and polynomial SVM models) and 1000 sizes (worse than Gaussian but better than the other two SVM type models). In the case of normality and large sample sizes, particularly favorable for MLE, the standard GARCH model still cannot defeat the Gaussian-based SVM-GARCH model. It is not surprising for EGARCH to behave badly in this case. As for the situation of *t* distribution, the linear GARCH model is ranked last for the 500 sizes (55.93%) and second last for the 1000 sizes (59.32%); while the asymmetric EGARCH model is good at forecasts of volatility turning points—ranked fourth for short series (only behind the three SVM models) and even third for long series (inferior to Gaussian and polynomial but better than the linear SVM-GARHC model). This time the ANN-GARCH model defeats the linear GARCH model. As for the linear GARCH model and moving average method, in the situation of 500 sizes and *t* distribution the standard GARCH model performs worse than the moving average, the simplest time series method, in terms of both MAE and DA measures. The conclusions described above are obtained on average based on 50 replications.

Table II displays the *p*-values of the DM test when the moving average method, standard GARCH, EGARCH and ANN models are compared with each of the other models considered in the study. We denote these tests DM1, DM2, DM3 and DM4, respectively. For instance, DM1 presents the test results for the simple moving average, where a *p*-value no greater than 0.05 indicates that the moving average method yields a higher forecast error (in terms of absolute error) relative to the competing model at 5% significance level, a *p*-value no smaller than 0.95 means that the moving average produces a lower forecast error at the 5% level, while a *p*-value between 0.05 and 0.95 implies that the benchmark and competing model have equivalent forecasting accuracy from the viewpoint of statistics. The same interpretation applies to the *p*-values reported for DM2-DM4.

Table II. Diebold–Mariano test for the MAE difference on Monte Carlo simulation

Distribution	Models	Sample size = 500				Sample size = 1000			
		DM1	DM2	DM3	DM4	DM1	DM2	DM3	DM4
Normality	Moving average		0.976	0.401	0.070		1.000	0.999	0.875
	Standard GARCH	0.024		0.001	0.000	0.000		0.001	0.000
	EGARCH	0.600	0.999		0.005	0.001	0.999		0.033
	ANN-GARCH	0.930	1.000	0.995		0.125	1.000	0.967	
	SVMl-GARCH	0.018	0.460	0.002	0.000	0.000	0.574	0.002	0.000
	SVMp-GARCH	0.023	0.413	0.004	0.000	0.000	0.420	0.003	0.000
	SVMg-GARCH	0.002	0.097	0.000	0.000	0.000	0.354	0.000	0.000
Student's *t*	Moving average		0.480	0.036	0.000		1.000	0.822	0.984
	Standard GARCH	0.520		0.054	0.003	0.000		0.000	0.001
	EGARCH	0.964	0.946		0.021	0.178	1.000		0.966
	ANN-GARCH	1.000	0.997	0.979		0.016	0.999	0.034	
	SVMl-GARCH	0.043	0.037	0.002	0.000	0.000	0.019	0.000	0.000
	SVMp-GARCH	0.056	0.043	0.001	0.000	0.000	0.025	0.000	0.000
	SVMg-GARCH	0.070	0.050	0.000	0.000	0.000	0.033	0.000	0.000

Note: DM1, DM2, DM3 and DM4 are the robust Diebold and Mariano (1995) test by using the Newey–West procedures (Newey and West, 1987) when the benchmark models are the moving average, linear GARCH model, EGARCH model and traditional ANN-GARCH model, respectively. For each test we consider the MAE loss functions.

DOI: 10.1002/for

Under the normal distribution, DM1 tests indicate that there is equivalent forecasting ability between moving average and EGARCH for short series, and between moving average and ANN-GARCH for long series. Such models as standard GARCH and the three SVM-GARCH all have higher volatility forecasting accuracy than moving average for both series at least at the 5% significance level. Moving average outperforms the ANN-GARCH model at the 10% level for a series of 500 size and EGARCH outperforms moving average at the 0.1% significance level for long series. According to DM2, three SVM type models have statistically equivalent forecasting ability to standard GARCH model for both series, with only one exception that the Gaussian SVM-GARCH model behaves better than the standard GARCH model at 10% significance level for short series. For both series, the standard GARCH model outperforms EGARCH and ANN-GARCH models at extremely low significance level. The DM3 statistic reveals that, for two series, three SVM-GARCH models perform better than the EGARCH model and EGARCH better than the ANN-GARCH model all at extremely significant levels. Finally, the ANN-GARCH model is found statistically and consistently inferior to the three SVM models for any series based on DM4 tests.

In the case of Student's t distribution, the out-of-sample performance of the standard GARCH model deteriorates. Now, according to DM2, the three SVM-GARCH models forecast volatility significantly better than the standard GARCH model at the 5% level for both series. The standard GARCH model cannot statistically defeat the moving average, either, for short series. However, both EGARCH and ANN-GARCH models are still statistically inferior to the standard GARCH model. In fact, according to DM1, DM3 and DM4, the three SVM-GARCH models all consistently outperform such benchmarks as moving average, EGARCH and ANN-GARCH models in forecasting volatility for any series. In terms of DM1, furthermore, the null hypothesis of equal forecasting accuracy between moving average and EGARCH cannot be rejected for a series of 1000 size rather 500 size. Moving average is significantly better than the ANN-GARCH model for short series, but the case is reversed for long series. In a series of 500 sizes, the ANN-GARCH model is significantly outperformed by the EGARCH model, while for the series of 1000 size the ANN type model statistically defeats the EGARCH model.

In summary, it appears that the three SVM-GARCH models do a better job of forecasting volatility than the moving average, standard GARCH, EGARCH and ANN-GARCH models in terms of MAE measures, which is statistically supported by the DM1, DM3, DM4 tests and DM2 in the case of t distribution. The DM2 test reveals that under the normal distribution the three SVM-GARCH models and standard GARCH model have similar volatility forecasting ability. Based on DA measures, the standard GARCH model too has a better ability in forecasting volatility turning points under normality and large sample sizes, while the asymmetric EGARCH model behaves better under the skewed t distribution. But both linear GARCH and nonlinear EGARCH cannot defeat all SVM-type models, at least the Gaussian-based SVM-GARCH model, in forecasting volatility directions.

REAL DATA ANALYSIS

In this section, we investigate the volatility forecasting performance of all candidates by using real data for two kinds of financial variables: GBP/USD exchange rates and NYSE average index.

Data description

The first dataset consists of the daily nominal bilateral exchange rates of British pounds (GBP) against the US dollar for the period January 5, 2004 to December 31, 2007. The data are obtained

DOI: 10.1002/for

from a database provided by Policy Analysis Computing and Information Facility in Commerce (PACIFIC) at the University of British Columbia, which contains the closing rates for a total of 81 currencies and commodities. The second dataset consists of the daily closing price of the New York Stock Exchange (NYSE) composite stock index for the period January 8, 2004 to December 31, 2007. The data are downloaded directly from the Market Information section of the NYSE web page.

It has been widely accepted that a variety of financial variables including foreign exchange rates and stock prices are integrated of order one. To avoid the issue of possible nonstationarity, both sets of raw real data are transformed into daily returns via equation (17), giving a returns series of 1001 observations and then a residual series is obtained from a fitted conditional mean equation of the GARCH class models. For the squared residuals of 1000 observations, the recursive estimating samples for the conditional volatility function are updated from the former 940 observations through the former 999 and then 60 numbers of one-period-ahead volatility forecasts are obtained, corresponding to an evaluation sample spanned from the 941st through the 1000th data points, that is, out-of-sample period of October 3, 2007 to December 31, 2007 for GBP and October 5, 2007 to December 31, 2007 for NYSE data.

The daily series for the log-levels and the returns of the GBP and NYSE are depicted in Figure 5. This figure shows that the returns series are mean-stationary, and exhibit the typical volatility clustering phenomenon with periods of unusually large volatility followed by periods of relative tranquility. Table III reports the summary of the descriptive statistics for the GBP and NYSE returns. Both series are typically characterized by excessive kurtosis and asymmetry. The Bera and Jarque (1981) tests all strongly reject the normality hypothesis. For GBP series, the Ljung–Box Q(6) statistic

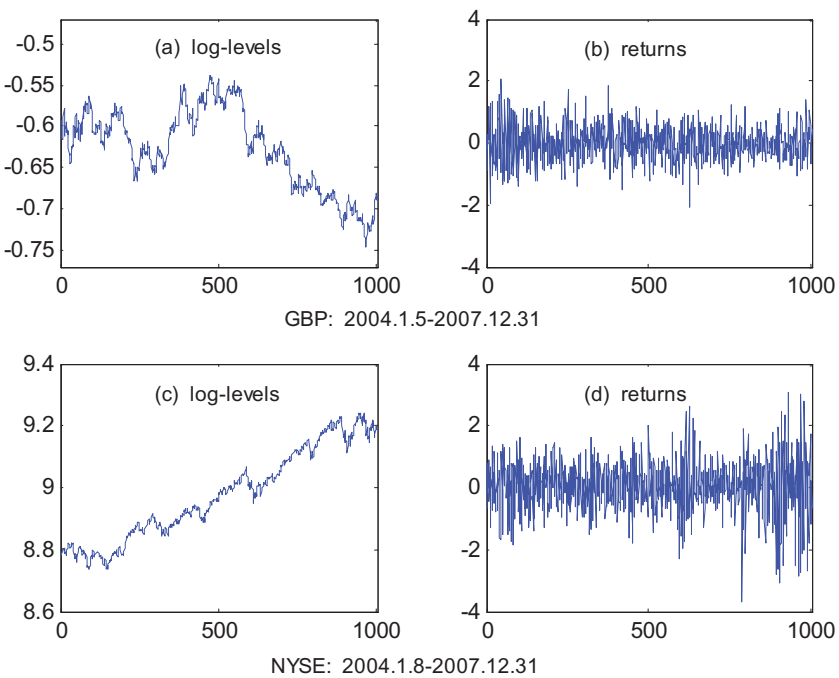


Figure 5. Log levels and returns of GBP exchange rates and NYSE stock index

DOI: 10.1002/for

Table III. Descriptive statistics for daily financial returns

Returns	GBP		NYSE	
	Statistics	*p*-value	Statistics	*p*-value
Mean	–0.0092		0.0393	
Variance	0.2827		0.6197	
Skewness	0.1206		–0.3489	
Kurtosis	3.7130		4.9343	
Normality	23.1860	0.00001	174.7200	0.00000
$Q(6)$	3.0313	0.80490	12.7100	0.04788
$Q(6)^*$	31.6390	0.00002	150.2400	0.00000
ARCH(6)	28.9280	0.00006	101.8400	0.00000

Notes: Normality is the Bera-Jarque (1981) normality test; Q(6) is the Ljung-Box Q test at 6 order for raw returns; Q(6)* is LB Q test for squared returns; ARCH(6) is Engle's (1982) LM test for ARCH effect.

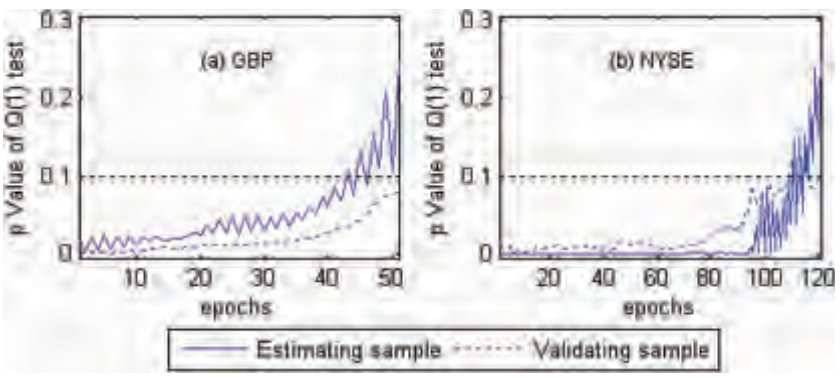


Figure 6. Iterative epochs of recurrent SVR procedure for real data

of raw returns indicates no significant correlation, but the Q(6) value of the squared returns reveals that there is significant autocorrelation in the squared returns. The Q(6) tests of both raw and squared returns of NYSE are all significant. Engle's (1982) LM tests for ARCH effect show significant evidence in support of GARCH effects (i.e., heteroscedasticity) for both series. Note that the number in parentheses indicates testing at 6 lag order. This examination of daily returns on the GBP and NYSE data reveals that returns can be characterized by heteroscedasticity and time-varying autocorrelation; therefore, we expect the GARCH class models to capture it adequately. Furthermore, as seen from Figure 5 and Table III, it seems that NYSE returns exhibit more variability, skewness, kurtosis and volatility clustering than GBP series such that nonlinear asymmetric EGARCH model should fit it more accurately.

Iterative epochs of recurrent SVM

Because the actual volatility h_t is unobservable for real data analysis, the second conditional variance equation (21b) of the GARCH (1, 1) model should be estimated by using the recurrent SVM procedure, as proposed above. Again, we use cross-validation to determine when the procedure is stopped.

With good forecasting performance as the goal, it is very difficult to figure out when it is best to stop training only in terms of fitting performance. It is possible for the procedure to end up

DOI: 10.1002/for

over-fitting the training data if the training session is not stopped at the right point. We can identify the onset of over-fitting and the stopping point through the use of cross-validation. Figure 6(a) and (b) describes the iterative epochs for volatility prediction of the first training sample of GBP and NYSE, respectively. For the GBP series, the iterative process of recurrent SVM procedure is stopped at the 51st epoch; while, for NYSE, the iterative process is longer and stopped after 121 iterative steps, possibly due to higher kurtosis and more variability and noise behind the NYSE series. Now, we could say, at about the 10% level of significance, the final residuals of equation (21b) obtained from the recurrent SVM procedure have no autocorrelation. In addition, the *p*-value curves of both estimating and validating samples exhibit a similar pattern (namely, increase with an increasing number of epochs) and point to almost the same stopping point. That is to say, there is no over-fitting phenomenon for the examples illustrated here; the recurrent SVM model does as well on the validating subset as it does on the estimating subset, on which its design is based.

The values taken by the free parameter of SVM and kernel coefficients are also selected according to the sensitivity investigation, similar to that done in Monte Carlo simulation. We do not report the parameter selection process here but present the formal results throughout the real data analysis. For both conditional mean and variance estimation of GBP and NYSE series, fortunately, similar parameter values of feedforward and recurrent SVM procedure could be found as follows: $C = 0.005$, $\varepsilon = 0.05$ and $\sigma^2 = 0.2$. Note that in the analysis of financial returns only the Gaussian kernel is employed for the sake of simplicity due to its best performance among linear, polynomial and Gaussian kernels, as described in Monte Carlo simulation.

Comparing the forecasting ability

The results of out-of-sample volatility forecasting accuracy for each model by using real data are presented in Table IV. Table V reports the *p*-values of the Diebold– Mariano (DM) test for the difference of MAE loss function in a robust HAC form from Newey–West procedures. In calculating the DM statistic, the null hypothesis of equal forecasting accuracy is related to the four benchmark

Table IV. Measure of volatility forecasting performance for real data

Models	Measures	Moving average	Standard GARCH	EGARCH	ANN-GARCH	SVM-GARCH
GBP	MAE	0.28895	0.24713	0.25719	0.24691	0.23257
	DA	37.29	38.98	49.15	38.98	45.76
NYSE	MAE	1.69610	1.51000	1.44880	1.62980	1.50410
	DA	32.20	42.37	55.93	32.20	57.63

Table V. Diebold–Mariano test for the MAE difference on real data

Models	GBP				NYSE			
	DM1	DM2	DM3	DM4	DM1	DM2	DM3	DM4
Moving average		0.990	0.970	0.981		0.935	0.970	0.813
Standard GARCH	0.010		0.017	0.583	0.065		0.902	0.061
EGARCH	0.030	0.983		0.980	0.030	0.098		0.044
ANN-GARCH	0.019	0.417	0.020		0.187	0.939	0.956	
SVM-GARCH	0.001	0.076	0.000	0.067	0.047	0.054	0.885	0.042

Note: DM1, DM2, DM3 and DM4 are the robust Diebold and Mariano (1995) test by using the Newey–West procedures (Newey and West, 1987) when the benchmark models are the moving average, linear GARCH model, EGARCH model and traditional ANN-GARCH model, respectively. For each test we consider the MAE loss functions.

DOI: 10.1002/for

models: moving average, standard GARCH, EGARCH and ANN models. We specify them as DM1, DM2, DM3 and DM4, respectively. A *p*-value no greater than 0.05 indicates that the benchmark model yields a higher forecast error (in terms of absolute error) relative to the competing model at the 5% significance level, a *p*-value no smaller than 0.95 means that benchmark model produces a lower forecast error at 5% level, while a *p*-value between 0.10 and 0.90 implies that the benchmark and competing models have the equal forecasting accuracy at 10% significance level.

According to MAE measures in Table IV, the SVM-GARCH model is the best one for the GBP series and second for the NYSE series in forecasting the magnitude of volatility error. DM tests in Table V almost statistically favor the SVM-GARCH model as the best model, too, at least at 10% significance level. Even though the MAE metric reveals that the EGARCH model outperforms the SVM-GARCH model for the NYSE series, it is not supported by the DM3 test, which means both models have equal forecasting ability. The better performance of the EGARCH model for NYSE is perhaps due to its ability to capture higher skewness and asymmetry occurring in the SYSE series than in GBP. The standard GARCH model performs modestly in terms of MAE measures, statistically inferior to EGARCH and superior to the ANN-GARCH model for NYSE and significantly better than EGARCH and similar to the ANN-GARCH model for GBP according to DM2 tests. The moving average method is always ranked last in forecasting the magnitude of volatility error, the evidence being significantly supported at least at the 10% level by the DM1 tests in Table V with just one exception, that for NYSE series moving average and ANN-GARCH model have equal forecasting ability. MAE measures and DM3 and DM4 tests denote that the EGARCH model also significantly outperforms the ANN-GARCH model for highly skewed NYSE series but the case is totally reverse for the GBP sample.

Based on DA measures in Table IV, on average, the moving average method is still ranked last, the ANN-GARCH model is ranked second last and the standard GARCH model is ranked at the middle position in forecasting volatility directions. For the GBP series, EGARCH performs best with DA value to be highest 49.15%, followed closely by the SVM-GARCH model; while, for the NYSE model, the best model to forecast volatility turning points is the SVM-GARCH model, with the asymmetric EGARCH model is ranked second, their DA values being 57.63% and 55.93%, respectively.

The empirical evidence of real data also confirms the conclusion obtained in Monte Carlo simulation and favors the theoretical advantage of the SVM-GARCH model. Due to high skewness in financial returns, the asymmetric EGARCH model normally behaves better than the standard GARCH model, particularly in the case of higher skewness or in forecasting volatility turning points. The moving average method always behaves worst and the ANN-GARCH model sometimes good in forecasting one-period-ahead financial volatilities among all candidates.

CONCLUSIONS

In many applications, SVM has shown excellent forecasting performance due to its particular structural design of SRM principle rather than ERM employed by conventional ANN and MLE methods. This inspires us to use it to improve the volatility forecasting ability of the parametric GARCH models. Empirical applications are made for forecasting the simulated data and the real data of daily GBP exchange rates and NYSE stock index.

To avoid the problem that the actual volatility for real data is unobservable, we propose a recurrent SVM procedure with a global feedback loop from the output layer to the input, as opposed to

DOI: 10.1002/for

the feedforward one for simulation, to estimate the conditional volatility equation, that is the ARMA process in nature, of the nonlinear GARCH model. The forecasting performance of the SVM-GARCH model is compared with the moving average, standard GARCH, asymmetric EGARCH and traditional ANN-GARCH models based on two quantitative evaluation measures and robust Diebold–Mariano tests following the Newey–West procedure.

The real data results, together with the simulation evidence, consistently and significantly support the use of the feedforward and recurrent SVM-based GARCH (1, 1) models in forecasting the one-period-ahead volatility error magnitude and direction. The standard GARCH model also performs well in the case of normality and large sample size, while the asymmetric EGARCH model is good at forecasting volatility under the high skewed distribution; but they rarely exceed SVM-GARCH models, at least the Gaussian-type SVM. The recurrent ANN-GARCH model and moving average method behave well only in a few cases. Overall, empirical analysis is in favor of the theoretical advantage of the SVM.

How to choose the appropriate values of free parameters and kernel coefficients and what effect of kernel type in the SVM procedure are investigated by using the sensitivity analysis in Monte Carlo simulation. The iterative process of the proposed recurrent SVM procedure in real data analysis is also examined in detail by the cross-validation method, which is shown to be implemented very easily and could be adopted as another standard SVM construction procedure in other applications.

ACKNOWLEDGEMENTS

The authors acknowledge the editor, Derek Bunn, and the referees for their constructive comments. Thanks also goes to the production editor, Ivry Tan, and my student, Qian Feng, who print and proofread the manuscript. This work is sponsored by Deutsche Forschungsgemeinschaft through the SFB 649 'Economic Risk'. Shiyi Chen is also supported by Kyungpook National University Graduate Scholarship for Excellent International Students, Shanghai Leading Academic Discipline Project (No. B101) and State Innovative Institute of Project 985 at Fudan University.

REFERENCES

Akgiray V. 1989. Conditional heteroskedasticity in time series models of stock returns: evidence and forecasts. *Journal of Business* **62**(1): 55–80.

Andersen T, Bollerslev T. 1998. Answering the skeptics: yes, standard volatility models do provide accurate forecasts. *International Economic Review* **39**: 885–905.

Andersen T, Bollerslev T, Diebold F, Labys P. 2003. Modeling and forecasting realized volatility. *Econometrica* **71**: 579–625.

Andersson J. 2001. On the normal inverse gaussian stochastic volatility model. *Journal of Business and Economic Statistics* **19**: 44–54.

Andrews D. 1991. Heteroskedasticity and autocorrelation consistent covariance matrix estimation. *Econometrica* **59**: 817–858.

Ané T, Ureche-Rangau L, Gambet J, Bouverot J. 2008. Robust outlier detection for Asia-Pacific stock index returns. *Journal of International Financial Markets, Institutions and Money* **18**(4): 326–343.

Awartani B, Corradi V. 2005. Predicting the volatility of the S&P-500 stock index via GARCH models: the role of asymmetries. *International Journal of Forecasting* **21**(1): 167–183.

Balaban E. 2004. Comparative forecasting performance of symmetric and asymmetric conditional volatility models of an exchange rate. *Economics Letters* **83**(1): 99–105.

DOI: 10.1002/for

Bauwens L, Sebastien L, Jeroen R. 2006. Multivariate GARCH models: a survey. *Journal of Applied Econometrics* **21**: 79–109.
Becker R, Clements A, White S. 2007. Does implied volatility provide any information beyond that captured in model-based volatility forecasts? *Journal of Banking and Finance* **31**(8): 2535–2549.
Becker R, Clements A, McClelland A. 2009. The jump component of S&P 500 volatility and the VIX index. *Journal of Banking and Finance* **33**(6): 1033–1038.
Bekiros S, Georgoutsos D. 2008. Direction-of-change forecasting using a volatility-based recurrent neural network. *Journal of Forecasting* **27**(5): 407–417.
Bera A, Jarque C. 1981. An efficient large-sample test for normality of observations and regression residuals. *Australian National University Working Papers in Econometrics*, 40. Canberra.
Blair B, Poon S,-H, Taylor S. 2001. Forecasting S&P 100 volatility: the incremental information content of implied volatilities and high-frequency index returns. *Journal of Econometrics* **105**: 5–26.
Bollerslev T. 1986. Generalized autoregressive conditional heteroskedasticity. *Journal of Econometrics* **31**: 307–327.
Bollerslev T, Chou R, Kroner K. 1992. Arch modeling in finance: a review of the theory and empirical evidence. *Journal of Econometrics* **52**: 5–59.
Box G, Jenkins G, Reinsel G. 1994. *Time Series Analysis: Forecasting and Control*. Prentice Hall: Englewood Cliffs, NJ.
Brailsford T, Faff R. 1996. An evaluation of volatility forecasting techniques. *Journal of Banking and Finance* **20**: 419–438.
Brooks C. 1998. Predicting stock index volatility: can market volume help? *Journal of Forecasting* **17**: 59–80.
Brooks C. 2001. A double-threshold GARCH model for the french franc/deutschmark exchange rate. *Journal of Forecasting* **20**: 135–143.
Brooks C, Persand G. 2003. Volatility forecasting for risk management. *Journal of Forecasting* **22**: 1–22.
Cao C, Tsay R. 1992. Nonlinear time-series analysis of stock volatilities. *Journal of Applied Econometrics* **1**: 165–185.
Cao L, Tay F. 2001. Financial forecasting using support vector machines. *Neural Computation and Application* **10**: 184–192.
Chan K, Christie W, Schultz P. 1995. Market structure and the intraday pattern of bid-ask spreads for NASDAQ securities. *Journal of Business* **68**(1): 35–40.
Chen G, Choi Y, Zhou Y. 2008. Detections of changes in return by a wavelet smoother with conditional heteroscedastic volatility. *Journal of Econometrics* **143**(2): 227–262.
Chen S, Härdle W, Moro R. 2009. Modeling default risk with support vector machines. *Quantitative Finance* (accepted for publication).
Chib S, Nardari F, Shephard N. 2002. Markov chain Monte Carlo methods for generalized stochastic volatility models. *Journal of Econometrics* **108**: 281–316.
Chong C, Ahmad M, Abdullah M. 1999. Performance of GARCH models in forecasting stock market volatility. *Journal of Forecasting* **18**: 333–343.
Choudhry T, Wu H. 2008. Forecasting ability of GARCH vs kalman filter method: evidence from daily UK time-varying beta. *Journal of Forecasting* **27**(8): 670–689.
Clements M, Smith J. 1999. A Monte Carlo study of the forecasting performance of empirical SETAR models. *Journal of Applied Econometrics* **14**: 123–141.
Clements M, Smith J. 2001. Evaluating forecasts from SETAR models of exchange rates. *Journal of International Money and Finance* **20**: 133–148.
Corradi V, Swanson N. 2004. Some recent developments in predictive accuracy testing with nested models and (generic) nonlinear alternatives. *International Journal of Forecasting* **20**(2): 185–199.
Corradi V, Distaso W, Swanson N. 2009. Predictive density estimators for daily volatility based on the use of realized measures. *Journal of Econometrics* **150**(2): 119–138.
Cumby R, Figlewski S, Hasbrouck J. 1993. Forecasting volatility and correlations with EGARCH models. *Journal of Derivatives* Winter: 51–63.
Day T, Lewis C. 1992. Stock market volatility and the information content of stock index options. *Journal of Econometrics* **52**: 267–287.
Deng N, Tian Y. 2004. *New Methods in Data Mining: Support Vector Machine*. Science Press: Beijing.
Diebold F, Mariano R. 1995. Comparing predictive accuracy. *Journal of Business and Economic Statistics* **13**: 253–265.

DOI: 10.1002/for

Dimson E, Marsh P. 1990. Volatility forecasting without data-snooping. *Journal of Banking and Finance* **44**: 399–421.
Donaldson R, Kamstra M. 1997. An artificial neural network-GARCH model for international stock return volatility. *Journal of Empirical Finance* **4**: 17–46.
Dotsis G, Psychoyios D, Skiadopoulos G. 2007. An empirical comparison of continuous-time models of implied volatility indices. *Journal of Banking and Finance* **31**: 3584–3603.
Dunis C, Huang X. 2002. Forecasting and trading currency volatility: an application of recurrent neural regression and model combination. *Journal of Forecasting* **21**: 317–354.
Enders W. 2004. *Applied Econometric Time Series* (2nd edn). Wiley: New York.
Engle R. 1982. Autoregressive conditional heteroskedasticity with estimates of the variance of UK inflation. *Econometrica* **50**: 957–1008.
Engle R, Hong C-H, Kane A, Noh J. 1993. *Advances in Futures and Options Research*, Vol. 6. JAI Press: Greenwich, CT; 393–415.
Feng Y, McNeil A. 2008. Modelling of scale change, periodicity and conditional heteroskedasticity in return volatility. *Economic Modelling* **25**(5): 850–867.
Ferland R, Lalancette S. 2006. Dynamics of realized volatilities and correlations: an empirical study. *Journal of Banking and Finance* **30**(7): 2109–2130.
Fernandez-Rodriguez F, Gonzalez-Martel C, Sosvilla-Rivero S. 2000. On the profitability of technical trading rules based on artificial neural networks: evidence from the Madrid stock market. *Economics Letters* **69**(1): 89–94.
Figlewski S. 1997. Forecasting volatility. *Financial Markets, Institutions and Instruments* **6**: 1–88.
Fleming J. 1998. The quality of market volatility forecasts implied by S&P 100 index option prices. *Journal of Empirical Finance* **5**: 317–345.
Franke J, Neumann M, Stockis J. 2004. Bootstrapping nonparametric estimators of the volatility function. *Journal of Econometrics* **118**: 189–218.
Franses P, Dijk DV. 1996. Forecasting stock market volatility using (non-linear) GARCH models. *Journal of Forecasting* **15**(3): 229–235.
Franses P, Dijk DV. 2000. *Nonlinear Time Series Models in Empirical Finance*. Cambridge University Press: Cambridge, UK.
Franses P, McAleer M. 2002. Financial volatility: an introduction. *Journal of Applied Econometrics* **17**: 419–424.
Galbraith J, Kisinbay T. 2005. Content horizons for conditional variance forecasts. *International Journal of Forecasting* **21**: 249–260.
Gerlach R, Tuyl F. 2006. MCMC methods for comparing stochastic volatility and GARCH models. *International Journal of Forecasting* **22**(1): 91–107.
Ghysels E, Harvey A, Rebault E. 1996. *Handbook of Statistics: Statistical Methods in Finance*, Vol. 14. Elsevier Science: Amsterdam; 119–191.
Ghysels E, Santa-Clara P, Valkanov R. 2006. Predicting volatility: how to get most out of returns data sampled at different frequencies. *Journal of Econometrics* **131**: 59–95.
Glosten L, Jagannathan R, Runkle D. 1992. On the relation between the expected value and the volatility of nominal excess return on stocks. *Journal of Finance* **46**: 1779–1801.
Gokcan S. 2000. Forecasting volatility of emerging stock markets: linear versus non-linear GARCH models. *Journal of Forecasting* **19**(6): 499–504.
Gospodinov N, Gavala A, Jiang D. 2006. Forecasting volatility. *Journal of Forecasting* **25**(6): 381–340.
Gray S. 1996. Modeling the conditional distribution of interest rates as a regime-switching process. *Journal of Financial Economics* **42**: 27–62.
Groen J, Kapetanios G, Price S. 2009. A real time evaluation of bank of England forecasts of inflation and growth. *International Journal of Forecasting* **25**(1): 74–80.
Gunn S. 1998. Support vector machines for classification and regression. *Isis-1-98*. Technical report, Image Speech and Intelligent Systems Group, University of Southampton, UK.
Hamid S, Iqbal Z. 2004. Using neural networks for forecasting volatility of S&P 500 index futures prices. *Journal of Business Research* **57**: 1116–1125.
Hamilton J. 1989. *A* new approach to the economic analysis of nonstationary time series and the business cycle. *Econometrica* **57**: 357–384.
Hamilton J. 1997. *Time Series Analysis*. Princeton University Press: Princeton, NJ.
Härdle, W, Moro R, Schäfer D. 2005. *Statistical Tools for Finance and Insurance*. Springer: Berlin.

DOI: 10.1002/for

Härdle W, Moro R, Schäfer D. 2007. *Handbook for Data Visualization*. Springer: Berlin.
Haykin S. 1999. *Neural Networks: A Comprehensive Foundations* (2nd edn). Prentice Hall: Englewood Chiffs, NJ.
Heynen R, Kat H. 1994. Volatility prediction: a comparison of stochastic volatility, GARCH(1, 1) and EGARCH(1, 1) models. *Journal of Derivatives* 50–65.
Hu M, Tsoukalas C, 1999. Combining conditional volatility forecasts using neural networks: an application to the EMS exchange rates. *Journal of International Financial Markets, Institution and Money* **9**: 407–422.
Jorion P. 1995. Predicting volatility in the foreign exchange market. *Journal of Finance* **50**: 507–528.
Jorion P. 1996. *The Microstructure of Foreign Exchange Markets*. Chicago University Press: Chicago, IL.
Klaassen F. 2002. Improving GARCH volatility forecasts with regime-switching GARCH. *Empirical Economics* **27**: 363–394.
Koopman S, Jungbacker B, Hol E. 2005. Forecasting daily variability of the S&P 100 stock index using historical, realised and implied volatility measurements. *Journal of Empirical Finance* **12**: 445–475.
Kuan C, Liu T. 1995. Forecasting exchange rates using feedforward and recurrent neural networks. *Journal of Applied Econometrics* **10**: 347–364.
Lamoureux C, Lastrapes W. 1993. Forecasting stock-return variance: understanding of stochastic implied volatilities. *Review of Financial Studies* **6**: 293–326.
Lehar A, Scheicher M, Schittenkopf C. 2002. GARCH vs. stochastic volatility: option pricing and risk management. *Journal of Banking and Finance* **26**: 323–345.
Li W, Ling S, McAleer M. 2002. Recent theoretical results for time series models with GARCH errors. *Journal of Economic Surveys* **16**: 245–269.
Lux T, Schornstein S. 2005. Genetic learning as an explanation of stylized facts of foreign exchange markets. *Journal of Mathematical Economics* **41**: 169–196.
Marcucci J. 2005. *Studies in Nonlinear Dynamics and Econometrics*, Vol. 9. Berkeley Electronic Press: Berkeley, CA; 1145.
McMillan D, Speight A. 2004. Daily volatility forecasts: reassessing the performance of GARCH models. *Journal of Forecasting* **23**(6): 449–460.
McMillan D, Speight A, Gwilym O. 2000. Forecasting UK stock market volatility: a comparative analysis of alternate methods. *Applied Financial Economics* **10**: 435–448.
Meddahi N. 2003. ARMA representations of integrated and realized variances. *Econometrics Journal* **6**: 334–355.
Moosa I. 2000. *Exchange Rate Forecasting: Techniques and Applications*. Macmillan Press: London.
Neely C. 2009. Forecasting foreign exchange volatility: why is implied volatility biased and inefficient? And does it matter? *Journal of International Financial Markets, Institutions and Money* **19**(1): 188–205.
Nelson D. 1991. Conditional heteroskedasticity in asset returns: a new approach. *Econometrica* **59**: 347–370.
Newey W, West K. 1987. A simple positive, semi-definite, heteroskedasticity and autocorrelation consistent covariance matrix. *Econometrica* **55**(3): 703–708.
Niemira M, Klein P. 1994. *Forecasting Financial and Economic Cycles*. Wiley: New York.
Pagan A, Schwert G. 1990. Alternative models for conditional stock volatility. *Journal of Econometrics* **45**: 267–290.
Pantelidaki S, Bunn D. 2005. Development of a multifunctional sales response model with the diagnostic aid of artificial neural networks. *Journal of Forecasting* **24**: 505–521.
Park B. 2002. An outlier robust GARCH model and forecasting volatility of exchange rate returns. *Journal of Forecasting* **21**(5): 381–393.
Pérez-Cruz F, Afonso-Rodriguez J, Giner J. 2003. Estimating GARCH models using SVM. *Quantitative Finance* **3**: 163–172.
Pong S, Shackleton M, Taylor S, Xu X. 2004. Forecasting currency volatility: a comparison of implied volatilities and AR(FI) MA models. *Journal of Banking and Finance* **28**: 2541–2563.
Poon S-H, Granger C. 2003. Forecasting volatility in financial markets: a review. *Journal of Economic Literature* **41**: 478–539.
Preminger A, Franck R. 2007. Forecasting exchange rates: a robust regression approach. *International Journal of Forecasting* **23**(1): 71–84.
Qi M, Wu Y. 2003. Nonlinear prediction of exchange rates with monetary fundamentals. *Journal of Empirical Finance* **10**: 623–640.

DOI: 10.1002/for

Renò R. 2006. Nonparametric estimation of stochastic volatility models. *Economics Letters* **90**(3): 390–395.
Rosenow B. 2008. Determining the optimal dimensionality of multivariate volatility models with tools from random matrix theory. *Journal of Economic Dynamics and Control* **32**(1): 279–302.
Schittenkopf C, Dorffner G, Dockner E. 2000. Forecasting time-dependent conditional densities: a semi-non-parametric neural network approach. *Journal of Forecasting* **19**: 355–374.
Scholkopf B, Smola A. 2001. *Learning with Kernels*. MIT Press: Cambridge, MA.
Sentana E. 1995. Quadratic ARCH models. *Review of Economic Studies* **62**: 639–661.
Suykens J, Vandewalle J. 2000. Recurrent least squares support vector machines. *IEEE Transactions on Circuits and Systems I* **47**(7): 1109–1114.
Tay F, Cao L. 2001. Application of support vector machines in financial time series forecasting. *Omega* **29**: 309–317.
Taylor J. 1999. Evaluating volatility and interval forecasts. *Journal of Forecasting* **18**: 111–128.
Taylor J. 2000. A quantile regression neural network approach to estimating the conditional density of multiperiod returns. *Journal of Forecasting* **19**: 299–311.
Taylor N. 2008. Can idiosyncratic volatility help forecast stock market volatility? *International Journal of Forecasting* **24**(3): 462–479.
Taylor S. 1986. *Modelling Financial Time Series*. Wiley: Chichester.
Tse Y, Tung S. 1992. Forecasting volatility in the singapore stock market. *Asia Pacific Journal of Management* **9**: 1–13.
Tseng C, Cheng S, Wang Y, Peng J. 2008. Artificial neural network model of the hybrid EGARCH volatility of the taiwan stock index option prices. *Physica A: Statistical Mechanics and its Applications* **387**(13): 3192–3200.
Vapnik V. 1995. *The Nature of Statistical Learning Theory*. Springer: New York.
Vapnik V. 1997. *Statistical Learning Theory*. Wiley: New York.
West K. 1996. Asymptotic inference about predictive ability. *Econometrica* **64**: 1067–1084.
West K, Cho D. 1995. The predictive ability of several models of exchange rate volatility. *Journal of Econometrics* **69**: 367–391.
Wong W, Tu A. 2009. Market imperfections and the information content of implied and realized volatility. *Pacific Basin Finance Journal* **17**(1): 58–79.
Zhang X, King M. 2005. Influence diagnostics in generalized autoregressive conditional heteroscedasticity processes. *Journal of Business and Economic Statistics* **23**: 118–129.

Authors' biographies:

Shiyi Chen started teaching at the School of Economics of Fudan University in China as an Assistant Professor after receiving his PhD degree in econometrics from the School of Economics and Trade at Kyungpook National University in the Republic of Korea in February 2006. From November 2008, Shiyi Chen became an Associate Professor of Econometrics. His research interests are time series forecasting, nonparametric econometrics, and energy and emission economics. One of his articles, Modeling Default Risk with Support Vector Machines, co-authored with Wolfgang K. Härdle and Rouslan A. Moro, was accepted by the *Journal of Quantitative Finance* in January 2009.

Wolfgang K. Härdle gained his Dr rer. nat. in mathematics at Universität Heidelberg in 1982 and his Habilitation at Universität Bonn in 1988. He is currently Chair Professor of Statistics at the Department of Economics and Business Administration, Humboldt-Universität zu Berlin. He is also director of CASE (Center for Applied Statistics and Economics) and of the Collaborative Research Center 'Economic Risk'. His research focuses on dimension reduction techniques, computational statistics and quantitative finance. He has published 34 books and more than 200 papers in leading statistical, econometrics and finance journals. He is one of the 'Highly Cited Scientists' according to the Institute of Scientific Information.

Kiho Jeong received a PhD degree in econometrics from the University of Wisconsin at Madison in 1991. After working for two years at the Korea Energy Economic Institute as an economist, he joined the School of Economics and Trade at Kyungpook National University in 1994 as an assistant professor, where he is now a full professor. His research interests are forecasting energy/financial markets, modelling climate change effects and nonparametric kernel methods.

DOI: 10.1002/for

Authors' addresses:
Shiyi Chen, School of Economics, Fudan University, 600 Guoquan Road, Shanghai 200433, China.

Wolfgang K. Härdle, Center for Applied Statistics and Economics, Humboldt University, Berlin, Germany.

Kiho Jeong, School of Economics and Trade, Kyungpook National University, Daegu, Republic of Korea.

J. Forecast. **29**, 406–433 (2010)
DOI: 10.1002/for

Journal of the American Statistical Association
Publication details, including instructions for authors and subscription information:
http://amstat.tandfonline.com/loi/uasa20

Localized Realized Volatility Modeling

Ying Chen, Wolfgang Karl Härdle and Uta Pigorsch
Ying Chen is Assistant Professor, Department of Statistics & Applied Probability, National University of Singapore, 6 Science Drive 2, Singapore 117546 . Wolfgang Karl Härdle is Ladislaus von Bortkiewicz Chair of Statistics, C.A.S.E.—Center for Applied Statistics and Economics, School of Business and Economics, Humboldt-Universität zu Berlin, Spandauerstr. 1, 10178 Berlin, Germany. Uta Pigorsch is Junior Professor, Department of Economics, Universität Mannheim, L7, 3-5, 68131 Mannheim, Germany. We are grateful to two editors, the associate editor, and three anonymous referees for their valuable comments. This research was supported by the Deutsche Forschungsgemeinschaft through the SFB 649 “Economic Risk” and the SFB 884 “Political Economy of Reforms,” and by the Berkeley-NUS Risk Management Institute at the National University of Singapore.

Available online: 01 Jan 2012

To cite this article: Ying Chen, Wolfgang Karl Härdle and Uta Pigorsch (2010): Localized Realized Volatility Modeling, Journal of the American Statistical Association, 105:492, 1376-1393

To link to this article: http://dx.doi.org/10.1198/jasa.2010.ap09039

Localized Realized Volatility Modeling

Ying CHEN, Wolfgang Karl HÄRDLE, and Uta PIGORSCH

With the recent availability of high-frequency financial data the long-range dependence of volatility regained researchers' interest and has led to the consideration of long-memory models for volatility. The long-range diagnosis of volatility, however, is usually stated for long sample periods, while for small sample sizes, such as one year, the volatility dynamics appears to be better described by short-memory processes. The ensemble of these seemingly contradictory phenomena point towards short-memory models of volatility with nonstationarities, such as structural breaks or regime switches, that spuriously generate a long memory pattern. In this paper we adopt this view on the dependence structure of volatility and propose a localized procedure for modeling realized volatility. That is at each point in time we determine a past interval over which volatility is approximated by a local linear process. A simulation study shows that long memory processes as well as short memory processes with structural breaks can be well approximated by this local approach. Furthermore, using S&P500 data we find that our local modeling approach outperforms long-memory type models and models with structural breaks in terms of predictability.

KEY WORDS: Adaptive procedure; Localized autoregressive modeling.

1. INTRODUCTION

Volatility is one of the key elements in modeling the stochastic dynamic behavior of financial assets. It is not only a measure of uncertainty about returns but also an important input parameter in derivative pricing, hedging, and portfolio selection. Accurate volatility modeling is therefore in the focus of financial econometrics and quantitative finance research. With the availability of high-frequency data, so-called realized volatility estimators (sums of squared high-frequency returns) have been proposed and have been shown to provide better volatility forecasts than the concurrent volatility estimators based on a coarser (e.g., daily) sampling frequency; see, for example, Andersen et al. (2001b).

Realized volatility together with other volatility measures exhibit significant autocorrelation which is the basis for the statistical predictability of volatility. In fact, the sample autocorrelation function has typically a hyperbolically-like decaying shape, also known as "long memory." A strand of literature focused on this kind of correlation phenomenon. The long memory "diagnosis," however, is usually stated for long sample periods such as three to 10 years. Over shorter sample periods, however, the autocorrelation function usually exhibits less persistence. This is also illustrated in Figure 1, which depicts the daily sample autocorrelation functions of daily logarithmic realized volatility of the S&P500 index futures for a long sample period (1985–2005) and for a short sample period (1995). The different degrees of persistence suggest that the diagnosis can also be generated by a simple model with structural change inside such a rather long interval; the possibility of such intermediate changes provides an alternative view on the described phenomenon. Like in the physical sciences, where one uses wave and particle theory to explain the emission of light, we have here a duality of theories for the emission of volatility. It is the objective of our study to investigate this dual view on volatility phenomenon.

In the literature of the long memory view of volatility, fractionally integrated [$I(d)$] processes have frequently been under consideration due to their hyperbolically decaying shock propagation for $0 < d < 1$. These processes have been proposed by, for example, Granger (1980), Granger and Joyeux (1980), and Hosking (1981). When applied to volatility they seem to provide a better description and predictability than short-memory models estimated over (the same) long sample periods. A typical example is the empirically better performance of the fractional integrated generalized autoregressive conditional heteroscedaticity (FIGARCH) model of Baillie, Bollerslev, and Mikkelsen (1996) as opposed to a standard GARCH model. For realized volatility, the autoregressive fractional integrated moving average (ARFIMA) process emerged as a standard model; see, for example, Andersen et al. (2003) and Pong et al. (2004). An alternative and quite popular model that does not belong to the class of fractionally integrated processes but approximates the long-range dependence by a sum of several multiperiod volatility components is the heterogenous autoregressive (HAR) model proposed by Corsi (2009).

The question on the true source of the long-memory diagnosis, however, still remains. Long memory in realized volatility may in fact be due to its construction, that is, by the aggregation over squared intraday returns, which are well known to exhibit also long-range dependence. Liebermann and Phillips (2008) therefore develop refined methods for conducting inference on long memory. Their empirical results, however, support the general finding on long memory in realized volatility.

Moreover, the presence of structural breaks may result in misleading inference on the long memory diagnosis, as has already been noted in Diebold (1986) and Lamoureux and Lastrapes (1990). In fact, the theoretical results provided in Diebold and Inoue (2001) and Granger and Hyung (2004) show that this phenomenon can also be spuriously generated by a short-memory model with structural breaks or regime shifts. More

Ying Chen is Assistant Professor, Department of Statistics & Applied Probability, National University of Singapore, 6 Science Drive 2, Singapore 117546 (E-mail: *stacheny@nus.edu.sg*). Wolfgang Karl Härdle is Ladislaus von Bortkiewicz Chair of Statistics, C.A.S.E.—Center for Applied Statistics and Economics, School of Business and Economics, Humboldt-Universität zu Berlin, Spandauerstr. 1, 10178 Berlin, Germany. Uta Pigorsch is Junior Professor, Department of Economics, Universität Mannheim, L7, 3-5, 68131 Mannheim, Germany. We are grateful to two editors, the associate editor, and three anonymous referees for their valuable comments. This research was supported by the Deutsche Forschungsgemeinschaft through the SFB 649 "Economic Risk" and the SFB 884 "Political Economy of Reforms," and by the Berkeley–NUS Risk Management Institute at the National University of Singapore.

DOI: 10.1198/jasa.2010.ap09039

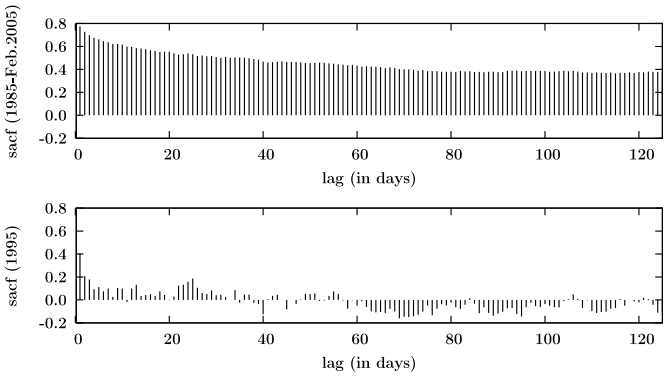


Figure 1. Sample ACF plots of daily logarithmic realized volatility of the S&P500 index futures for the sample from 1985–2005 (upper panel) and for the year 1995 (lower panel).

generally, Mikosch and Stărică (2004b) even argue independently of any particular model assumptions that nonstationarities in the data, such as changes in the unconditional mean or variance, can lead to the diagnosis of long-range dependencies. Such findings have led to the development of structural break detection methods and their application to financial volatility, where breaks are found in volatility processes using real data; see, for example, Chen and Gupta (1997), Mikosch and Stărică (2004a), Liu and Maheu (2008), and Čížek, Härdle, and Spokoiny (2009). Similarly, volatility models with time-varying coefficients have been proposed, which allow some or all of the model parameters to vary over time either in an abrupt fashion—for example, via Markov-Switching (see, e.g., Hamilton and Susmel 1994 and So, Lam, and Li 1998) and mixture multiplicative error specifications (see Lanne 2006)—or via a smooth function of time or other transition variables; see, for example, Baillie and Morana (2009b) and Scharth and Medeiros (2009), who show that nonlinearities, such as structural breaks and regimes induced by asymmetries like the leverage effect, may generate the observed long-range dependence. Such methods are also applied to long-memory models, addressing the possibility of the coexistence of long memory and structural breaks; see, for example, Baillie and Morana (2009a), Hillebrand and Medeiros (2008), and McAleer and Medeiros (2008b). The number of breaks in long memory realized volatility models is usually found to be one or two. Most of these studies, however, focus on sample periods covering at least 10 years. Given such a long time span of data, the presence of breaks even in long-memory models may be expected. Noteworthy, when it comes to forecasting, the more complicated models with breaks are often unable to significantly outperform the no-break long memory alternatives; see Hillebrand and Medeiros (2008), McAleer and Medeiros (2008b), and Martens, Dijk, and de Pooter (2009). Moreover, in some cases short memory models with breaks have been found to provide superior realized volatility forecasts than alternative long-memory models and regime switching ARFIMA models; see, for example, Lanne (2006) and Morana and Beltratti (2004).

In this paper we introduce the *localized realized volatility modeling* approach to describe realized volatility. In this approach the time-varying (local) structure of volatility is conveniently determined via adaptive statistical techniques, that allow us to find for each time point a past time interval, over which a local volatility model is a good approximator. Thus, in contrast to the previously cited literature our approach is local rather than global. The parameters of the local model as well as the length of the past time interval are determined at each point in time and may, therefore, differ from period to period. The method basically tries to adapt to local volatility. In doing so, it does not require any prior information or modeling assumptions on the number of break points, the potential (economic) sources of the break, its magnitude nor on its type (e.g., abrupt or smooth). This makes it very appealing. Moreover, it also allows to straightforwardly account for time-varying volatility of volatility, a feature that currently attracts researcher's interest, like Barndorff-Nielsen and Veraart (2009), and has been recognized to be important also for realized volatility; see Corsi et al. (2008) and Allen, McAleer, and Scharth (2010).

Although localized realized volatility modeling is a quite general concept that can be applied to various types of local parametric volatility models, we investigate it here based on autoregressive processes. In particular, we push here the alternative view on long memory to its limit by assuming a local linear short-memory model. Estimation and forecasting based on our approach is thus computationally straightforward.

The flexibility of our procedure is demonstrated within a simulation study, which shows that both, short-memory processes with breaks as well as long-memory processes, can be well described by the local approach. We additionally apply localized realized volatility modeling to S&P500 data and compare it to (approximate) long-memory techniques, such as the ARFIMA and HAR models, and to models with breaks. We find that our technique provides improved volatility forecasts.

The remainder of the paper is structured as follows. The next section reviews the concept of realized volatility, its construction, and the empirical properties of realized volatility of the S&P500 index futures. Section 3 presents in detail the localized realized volatility modeling approach along with a simulation study. Section 4 briefly reviews the alternative models considered in this paper, and Section 5 empirically compares the various models within a forecasting exercise. Section 6 concludes.

2. REALIZED VOLATILITY

Measuring the volatility of a financial asset based on high-frequency data has been one of the major focuses in the recent financial econometrics literature. The idea is to measure ex post the variation of asset prices over a lower frequency, commonly a day, by summing over products of high frequency, that is, intradaily returns. The approach is motivated by the theory of quadratic variation of semimartingales. For the ease of exposition, consider the case where the log price of a financial asset, p, follows a Brownian semimartingale—an assumption that is very popular in the asset pricing literature, that is,

$$p_t = \int_0^t \mu(s)\,ds + \int_0^t \sigma(s)\,dW(s) \qquad \forall t \geq 0, \tag{1}$$

where the instantaneous mean process $\{\mu(t)\}_{t\geq 0}$ is continuous and of finite variation, $\{\sigma(t)\}_{t\geq 0}$ with $\sigma(t) > 0$ $\forall t$ denotes the càdlàg instantaneous volatility, and $\{W(t)\}_{t\geq 0}$ is a standard Brownian Motion. Then the quadratic variation process of (1),

$$[p]_t = \operatorname{plim} \sum_{j=0}^{l-1} \left(p_{\tau_{j+1}} - p_{\tau_j}\right)^2, \tag{2}$$

where $\tau_0 = 0 \le \tau_1 \le \cdots \le \tau_l = t$ denotes a sequence of partitions with $\sup_j\{\tau_{j+1} - \tau_j\} \to 0$ for $l \to \infty$, is given by

$$[p]_t = \int_0^t \sigma^2(s)\, ds, \tag{3}$$

that is, as the integrated variance $\int_0^t \sigma^2(s)\, ds$ of the price process.

The theory of quadratic variation, thus, suggests that the sum over squared high-frequency returns may provide an ex post measure of the integrated variance and this is what is, oftentimes interchangeably, referred to as realized variance or realized volatility. Suppose we are interested in measuring volatility over a day t using $M+1$ intraday prices observed at time points $n_0, \ldots, n_M$. Furthermore, let p_{t,n_j} denote the logarithmic price observed at time point n_j of trading day t. The continuously compounded jth within-day return of day t is therefore given by

$$r_{t,j} = p_{t,n_j} - p_{t,n_{j-1}}, \qquad j = 1, \ldots, M. \tag{4}$$

Then *daily realized volatility* is defined as

$$\widetilde{RV}_t = \sum_{j=1}^{M} r_{t,j}^2. \tag{5}$$

Now, if $M \to \infty$, that is, the intraday sampling frequency goes to infinity, realized volatility converges to the quadratic variation of the price process; see, for example, Andersen and Bollerslev (1998) and Barndorff-Nielsen and Shephard (2002b). This implies that if the price follows a pure diffusion process as given in (1), realized volatility converges to the daily integrated variance, that is, $\widetilde{RV}_t \to IV_t$ for $M \to \infty$ with $IV_t = \int_{t-1}^t \sigma^2(s)\, ds$, which is oftentimes the main object of interest. Consistency and asymptotic distribution of realized volatility as an estimator of the integrated variance are derived in Barndorff-Nielsen and Shephard (2002a).

The theoretical results on realized volatility obviously build on the notion of an infinite sampling frequency. In practice, however, the sampling frequency is invariably limited by the actual quotation, or transaction frequency. Moreover, the observed high-frequency prices are further contaminated by market microstructure effects, such as the bid-and-ask bounce effect and price discreteness, which are due to the particular design and trading mechanism of financial markets; see, for example, Hasbrouck (2007). These effects introduce biases into realized volatility; see, for example, Andersen et al. (2001a) and Barndorff-Nielsen and Shephard (2002a). A common approach to reduce their impact is to simply construct realized volatility based on lower frequency returns (e.g., 10 to 30 minutes), at which market microstructure effects are negligible. However, such a procedure comes at the cost of a less precise volatility estimate, as it makes no use of all available data. Various alternative methods have therefore been proposed to solve this bias-variance trade-off. For a review, see McAleer and Medeiros (2008a) and Pigorsch, Pigorsch, and Popov (2010).

In this paper we compute a market microstructure noise robust version of realized volatility based on the approach of Barndorff-Nielsen et al. (2008). The reason for our choice is that their class of so-called *realized kernel estimators* of quadratic variation have very attractive properties. In particular, they are consistent and efficient and they are robust to a host of different market microstructure effects.

2.1 Noise-Corrected Realized Volatility

The idea of the realized kernel estimators is similar to that of autocorrelation and heteroscedasticity robust variance and covariance estimators, like the Newey–West estimator, that is, the correction is based on the sum of weighted autocovariances. Define the hth realized autocovariance for day t by $\gamma_{t,h} = \sum_{j=1}^{M} r_{t,j} r_{t,j-h}$. In the realized kernel estimators, realized volatility is then corrected by the weighted sum of those realized autocovariances. In particular, the flat-top realized kernel estimator, that we employ in this paper, provides a noise-corrected realized volatility RV_t by

$$RV_t = \widetilde{RV}_t + \sum_{h=1}^{H_t^*} k\left(\frac{h-1}{H_t^*}\right)(\gamma_{t,h} + \gamma_{t,-h}), \tag{6}$$

where the weights are given by the kernel function k being twice continuously differentiable on $[0, 1]$ and satisfying $k(0) = 1$, and $k(1) = k'(0) = k'(1) = 0$. The bandwith parameter H_t^* denotes the optimal number of lags to be considered for day t. It is optimal in the sense that it minimizes the asymptotic variance of the noise-corrected realized volatility. Barndorff-Nielsen et al. (2008) show that H_t^* depends on the chosen kernel weight function and on the noise-to-signal ratio $\xi_t = \omega_t^2/IV_t$, that relates the (daily) variance of the market microstructure noise, ω_t^2, to the (daily) integrated variance. In particular, $H_t^* = c^* \xi_t \sqrt{M}$, where c^* is a constant that depends, inter alia, on the specific kernel weight function. Its value is chosen such that it minimizes the asymptotic variance. The bandwith selection H_t^* and the computation of the noise-corrected realized volatility, thus, involve the precise specification of the kernel weight function and the estimation of the noise-to-signal ratio. We now turn to these issues.

For our empirical application we consider the modified Tukey–Hanning kernel with weight function $k(x) = \sin^2\{\frac{\pi}{2}(1-x)^a\}$, as it is most efficient among the finite lag kernels analyzed in Barndorff-Nielsen et al. (2008). Moreover, for increasing a the noise-corrected realized volatility approaches the (parametric) efficiency bound. As such, a large value of a might be preferable. However, an increasing number of a also leads to an increase in the number of autocovariances H_t^* considered in the noise correction (6), as c^* is increasing with a; see Barndorff-Nielsen et al. (2008). In practice, this imposes some limitations as the computation of the autocovariances $\gamma_{t,h}$ then involves an increasing number of returns outside the daily time interval. Note that in our application we make exclusive use of price observations within a day, such that fewer observations are available for the estimation of $\gamma_{t,h}$ as h increases. An increase in a therefore implies the use of less precisely estimated autocovariance terms. We therefore follow Barndorff-Nielsen et al. (2008) and choose $a = 2$ for our empirical application. For this kernel specification $c^* = 5.74$, see Barndorff-Nielsen et al. (2008). Noteworthy, the chosen realized kernel estimator is still close to efficient.

To finally determine H_t^* we estimate the noise-to-signal ratio ξ_t in the following way: we employ the estimator of the noise variance suggested by Bandi and Russell (2005) and compute the (scaled) conventional realized volatility estimator based on one minute returns, that is, $\hat{\omega}_t^2 = \widetilde{RV}^{1\,\mathrm{min}}/2M^{1\,\mathrm{min}}$, where the superscripts indicate the used sampling interval. An estimate

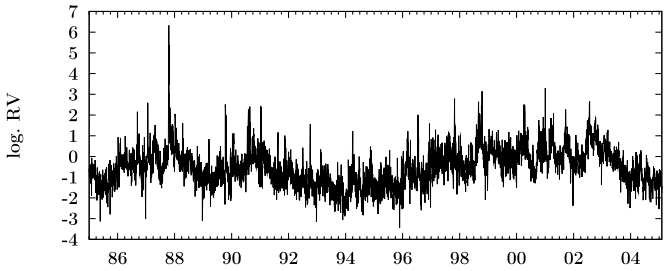


Figure 2. Time evolvement of logarithmic realized volatility of the S&P500 index futures.

of the variance of the "signal" (the integrated variance) is obtained by the realized volatility computed at a low, that is, 15 minutes, sampling interval, at which market microstructure effects should be negligible, thus $\widehat{IV}_t = \widetilde{RV}^{15\,\mathrm{min}}$. The optimal bandwidth is thus based on the estimate

$$\hat{H}_t^* = 5.74 \frac{\hat{\omega}_t^2}{\widehat{IV}_t} \sqrt{M^{1\,\mathrm{min}}}. \tag{7}$$

Rounding $\hat{H}_t^*$ to the nearest integer gives the final value of the bandwith. Given this bandwith, the noise-corrected realized volatility RV_t is then finally computed according to (6). Note that we estimate the realized autocovariances $\gamma_{t,h}$ and the market microstructure noise uncorrected realized volatility, $\widetilde{RV}_t$, based on one minute returns. Moreover, all intraday returns are constructed using the previous-tick method and by excluding overnight returns.

2.2 Data Description

Our empirical analysis focuses on the noise-corrected realized volatility of the S&P500 index futures ranging from January 2, 1985 to February 4, 2005; see Figure 2. From the various S&P500 Index futures with maturity dates in March, June, September, and December, we consider only the most liquid contracts. In addition, we have removed one day, February 18, 1990, from our dataset as there are only two transactions reported.

The descriptive statistics of the resulting realized volatility series are presented in Table 1. In summary, the empirical characteristics of the series are in line with the findings reported in the earlier literature on realized volatility. In particular, realized volatility is strongly skewed and fat-tailed, while its logarithmic version is much closer to Gaussianity. This is also confirmed by the kernel density estimate of logarithmic realized volatility, which is presented in Figure 3 along with the kernel density estimate of iid random variables simulated from the fitted normal distribution (with a sample size corresponding to the empirical one). Moreover, the sample autocorrelation function

Table 1. Descriptive statistics of realized volatility

Series	Mean	Std.Dev.	Skewness	Kurtosis	Ljung–Box(21)[(1)]
RV_t	1.0880	8.6961	55.5857	3412	1204
$\log(RV_t)$	−0.5314	0.8875	0.5343	4.9912	4.6861

[(1)]The critical value of the Ljung–Box test statistic of no autocorrelation up to approximately 1 month is 32.671.

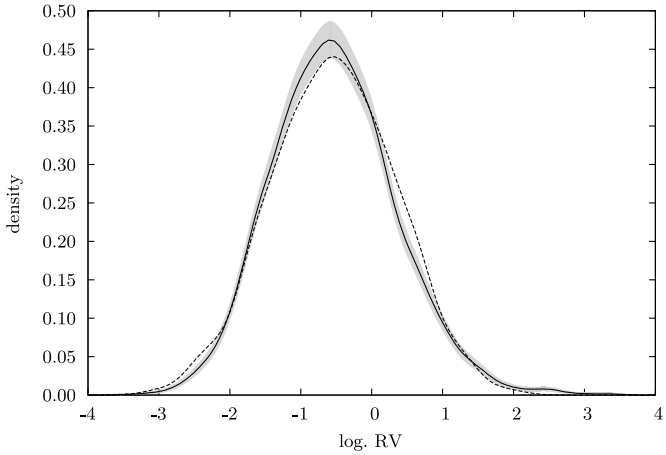


Figure 3. Kernel density estimate of logarithmic realized volatility of the S&P500 index futures (solid line). The shaded area corresponds to the pointwise 95% confidence intervals and the dashed line represents the kernel density estimate of iid random variables simulated from the fitted normal distribution.

of (logarithmic) realized volatility, Figure 1, exhibits the aforementioned hyperbolic decay. We evaluate this long-memory diagnosis in more detail in the empirical application. In the following, however, we first introduce our localized approach to realized volatility modeling.

3. THE LOCALIZED REALIZED VOLATILITY APPROACH

In this paper we adopt a local view on realized volatility modeling. The idea is simple. It is assumed that at each point in time there exists a past-time interval over which volatility can be well approximated by a local autoregressive (LAR) model. In contrast to fitting a global volatility model, we obtain at each point in time a potentially new set of parameters, which is estimated based on the so-called *interval of homogeneity*. For each point in time, the interval of homogeneity is selected in a sequential testing procedure, which starts from a small interval, where the local approximation holds and the AR parameters are approximately constant. The procedure then iteratively extends this interval and tests for time homogeneity until a structural break is found or data is exhausted. The local model is then fitted and can be used for volatility predictions.

The local (time-varying) autoregressive scheme is defined through a time-varying parameter set $\theta_t = (\theta_{0t}, \theta_{1t}, \ldots, \theta_{pt}, \sigma_t)^\top$:

$$\log RV_t = \theta_{0t} + \sum_{i=1}^{p} \theta_{it} \log RV_{t-i} + \varepsilon_t, \qquad \varepsilon_t \sim \mathrm{N}(0, \sigma_t^2), \tag{8}$$

where the Gaussian distributed innovations ε_t have mean zero and variance σ_t^2. Note that the specification also allows for time-varying volatility of volatility by letting σ_t rely on time.

Time-varying parameters at any point in time t are of course too flexible to really constitute a practical dynamic model. We therefore need to strike a balance between model flexibility and dimensionality. Traditional ways either estimate the time-varying parameters nonparametrically by assuming that the parameters are smooth functions of time (see, e.g., Cai,

Fan, and Li 2000) or assume that the time-varying parameters are piecewise constant functions provided that the number of changes are given (see, e.g., Bai and Perron 1998 and Mikosch and Stărică 2004a). Here we follow a different strategy by localizing (in time) a low-dimensional time series dynamics in the high-dimensional model (8). The basic idea is to approximate (8) at a fixed time point τ by a constant parameter vector $\theta_\tau = (\theta_{0\tau}, \theta_{1\tau}, \ldots, \theta_{p\tau}, \sigma_\tau)^\top$ over $I_\tau = [\tau - l_\tau, \tau)$ with $p + 2 \le l_\tau < \tau$. The interval I_τ is called the interval of homogeneity, whose length depends on time point τ. In the estimation of (8) at a particular time point τ, we only assume that an I_τ exists over which the local parametric model (approximately) holds for the process. This assumption nests the abovementioned "smooth transition" and "regime switching" assumptions as special cases: parameters can either smoothly vary over time or change abruptly. The question now is how to find I_τ or the value of l_τ over which the model parameters can be estimated.

The next section discusses the estimation and the test statistics employed to determine the interval of homogeneity. The sequential testing procedure is described in Section 3.2, while Section 3.3 discusses the choice of parameters involved in the procedure. The performance and sensitivity of the procedure are demonstrated in a set of simulations in Section 3.4.

3.1 Estimation and Test of Homogeneity

The estimation of the local parametric model is carried out via maximum likelihood. In particular, given an interval of homogeneity I_τ for time point τ, over which the process can be safely described by an AR model with constant parameters, the maximum likelihood (ML) estimator $\tilde{\theta}_\tau$ is defined as

$$\begin{aligned}\tilde{\theta}_\tau &= \operatorname*{argmax}_{\theta\in\Theta} L(\log RV; I_\tau, \theta)\\ &= \operatorname*{argmax}_{\theta\in\Theta}\left\{-\frac{l_\tau - p}{2}\log 2\pi - (l_\tau - p)\log\sigma\right.\\ &\quad\left. - \frac{1}{2\sigma^2}\sum_{t=\tau-l_\tau+p}^{\tau-1}\left(\log RV_t - \theta_0 - \sum_{i=1}^{p}\theta_i \log RV_{t-i}\right)^2\right\},\end{aligned}$$

where Θ denotes the parameter space and $L(\log RV; I_\tau, \theta)$ the local conditional log-likelihood function, for which we also use the short notation $L(I_\tau, \theta)$. We refer to the estimator $\tilde{\theta}_\tau$ as the *local ML estimator*.

The question now is how the interval of homogeneity I_τ can be determined. To this end likelihood ratio testing ideas are employed. Suppose that (log) *RV* is driven by an AR(p) process with a constant set of true parameters θ^*_τ at time point τ. The accuracy of estimation can be measured by the log-likelihood ratio (LR) (under homogeneity)

$$\mathrm{LR}(I_\tau, \tilde{\theta}_\tau, \theta^*_\tau) = L(I_\tau, \tilde{\theta}_\tau) - L(I_\tau, \theta^*_\tau). \tag{9}$$

Polzehl and Spokoiny (2006) derived a bound for LR and its power transformation $|\mathrm{LR}(I_\tau, \tilde{\theta}_\tau, \theta^*_\tau)|^r$ with $r > 0$ for an iid sequence of Gaussian innovations [in our case this refers to the innovations of the LAR(p) process]:

$$\mathsf{E}_{\theta^*_\tau}|\mathrm{LR}(I_\tau, \tilde{\theta}_\tau, \theta^*_\tau)|^r \le \xi_r. \tag{10}$$

This bound is nonasymptotic and can be applied to any interval I_τ. It allows to construct a confidence interval that can be used for testing homogeneity. The null hypothesis of time homogeneity means that the process follows the model (8) with a constant parameter, which implies that the ML estimator $\tilde{\theta}_\tau$ and the corresponding LR fulfill the risk bound (10). Therefore, the test of homogeneity can be performed, for example, by using the LR test statistic

$$|\mathrm{LR}(I_\tau, \tilde{\theta}_\tau, \theta^*_\tau)|^r.$$

In practice, the *hypothetical* AR(p) *parameters* θ^*_τ and also the risk bound ξ_r are unknown but can be computed empirically. Details on the feasible test procedure are given in the next section. In the estimation we are searching for an interval of *homogeneity* over which the process is well approximated by a parametric model. In other words, we mimic the unknown data-generating process by a local parametric model and simultaneously require that the modeling bias under this local parametric assumption is small. There exists a well-established theory addressing this local parametric assumption under a small modeling bias condition; see, e.g., Chen and Spokoiny (2010). Belomestny and Spokoiny (2007) shows that an optimal choice of an interval of local homogeneity can be obtained via an adaptive procedure. In the following, we concentrate on the construction details and its application to the dual view on the dependence structure of volatility. However, details of the results can be found in the cited literature and a comprehensive simulation study in Section 3.4 illustrates the performance of the adaptively selected estimators.

3.2 Adaptive Identification of the Interval of Homogeneity

This section presents a feasible adaptive selection algorithm of the interval of homogeneity for a particular point in time. Nevertheless, the procedure is general and is applied at every time point. The aim of the algorithm is to select the longest interval of homogeneity over which the parametric model is a good approximator for the process. The number of possible interval candidates is large, for example, the first interval may include just a few past observations and the intervals considered thereafter may be increased by just one observation in each step up to including all past observations. As this results in a large number of candidate intervals, it is practical to consider only a finite set of intervals $\mathbf{I}_\tau = \{I^1_\tau, \ldots, I^K_\tau\}$ with K candidates as suggested in Chen and Spokoiny (2010). For computational tractability, the intervals are increasingly ordered according to their length, that is, $I^1_\tau \subset \cdots \subset I^K_\tau$. To each interval there corresponds a local ML estimator, denoted by $\tilde{\theta}^k_\tau$ with $k = 1, \ldots, K$. In statistical learning theory those are called weak learners. Note that we are using the parametric assumption where the LAR model is only a good approximator of the process. Referring to the nonparametric smoothing literature, an increase in the length of intervals in (8) leads to an increase in modeling bias while the variance of the estimators is decreasing; see, for example, Härdle et al. (2004). In accordance with the chosen $\mathbf{I}_\tau$, the K weak learners therefore exhibit an increasing modeling bias and decreasing variance. Under the assumption that the interval of local homogeneity exists, the first interval I^1_τ is required to be short such that the modeling bias is small. Our interest here is to select an optimal estimator that has the smallest variance without violating the small modeling bias condition.

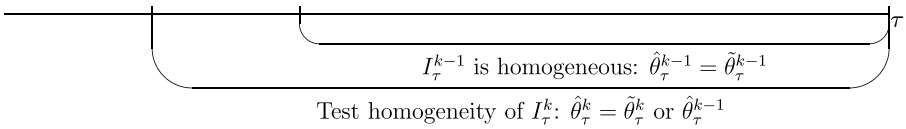


Figure 4. Sequential test of homogeneity: the longer interval I_τ^k is tested after the hypothesis of homogeneity over the shorter interval I_τ^{k-1} has been accepted.

The selection algorithm is based on a sequential testing procedure. The procedure starts from the shortest interval I_τ^1, over which local homogeneity holds by assumption. The weak learner $\tilde{\theta}_\tau^1$ is automatically accepted as an eligible local homogeneous estimator: $\hat{\theta}_\tau^1 = \tilde{\theta}_\tau^1$. Sequentially, at each step k with $2 \le k \le K$, we test the hypothesis of local homogeneity given that at the former step $k-1$ the null hypothesis has not been rejected; see Figure 4. The selected interval $\hat{I}_\tau$ corresponds to the largest accepted interval I_τ^k such that

$$|\,\mathrm{LR}(I_\tau^k, \tilde{\theta}_\tau^k, \hat{\theta}_\tau^{k-1})|^r \le \zeta_k, \tag{11}$$

where ζ_k is the critical value at step k and is described in more detail below. Note that this test (11) measures the difference of an estimator $\tilde{\theta}_\tau^k$ over a "possible" interval of local homogeneity I_τ^k to the most recently available optimal estimator $\hat{\theta}_\tau^{k-1}$. It differs from the LR test statistic implicitly linked to (10). Here the unknown hypothetical parameter θ_τ^* is replaced by the tentatively optimal estimator $\hat{\theta}_\tau^{k-1}$ since the latter is the possibly best estimator at the current step k. If there is no significant difference between the two estimators, it means that there is no significant change in the dynamics and the small modeling bias condition is not violated. We thus accept the null hypothesis of homogeneity and adopt the new estimator $\hat{\theta}_\tau^k = \tilde{\theta}_\tau^k$ as it has a smaller variance. On the other hand, if the test statistic is significant, it indicates that at least one structural change of the process exists and the LAR model is no longer a good approximator of the process. The sequential testing procedure terminates. This procedure then leads to the optimal estimator $\hat{\theta}_\tau$ that corresponds to the selected interval $\hat{I}_\tau$.

The formal definition of the procedure for a particular point in time τ is as follows:

1. Initialization: $\hat{\theta}_\tau^1 = \tilde{\theta}_\tau^1$.
2. $k = 2$: while $|\,\mathrm{LR}(I_\tau, \tilde{\theta}_\tau^k, \hat{\theta}_\tau^{k-1})|^r \le \zeta_k$ and $k \le K$,
$$k = k + 1,$$
$$\hat{\theta}_\tau^k = \tilde{\theta}_\tau^k.$$
3. Final estimate: $\hat{\theta}_\tau = \hat{\theta}_\tau^k$.

3.3 Choice of Parameters and Implementation Details

Clearly, the proposed procedure depends on a set of parameters, such as the lag order p in the LAR setup, the set of intervals, the power parameter r, and the critical values $\{\zeta_k\}_{k=1}^K$. In the following we address the choice of these parameters, and also discuss the computation of the critical values via Monte Carlo simulations.

3.3.1 Set of Intervals. We consider a finite set with $K = 13$ intervals in our study. This set is composed of the following interval lengths:

$$\{1\mathrm{w}, 1\mathrm{m}, 3\mathrm{m}, 6\mathrm{m}, 1\mathrm{y}, 1.5\mathrm{y}, 2\mathrm{y}, 2.5\mathrm{y}, 3\mathrm{y}, 3.5\mathrm{y}, 4\mathrm{y}, 4.5\mathrm{y}, 5\mathrm{y}\},$$

where w denotes a week (5 days), m refers to one month (21 days), and y to one year (252 days). In other words, $I_\tau^1 = [\tau - 1\mathrm{w}, \tau)$, $I_\tau^2 = [\tau - 1\mathrm{m}, \tau), \ldots, I_\tau^{13} = [\tau - 5\mathrm{y}, \tau)$. This choice is motivated by the practical reason that investors are often concerned about special investment horizons. As the set $\mathbf{I}_\tau = \{I_\tau^k\}_{k=1}^{13}$ is used for each time point τ, we drop the subscript in the following for notational convenience. Other sets of intervals may be considered (see also Section 3.4). However, it is important to assure homogeneity over the shortest interval.

3.3.2 Selection of the Lag Order. While the lag selection in the (global) AR models is straightforward, it is more complicated in the LAR approach as the identification of the local intervals of homogeneity depends on the assumed lag order. The selection of the lag order p can be based, for example, on the minimum average value of the information criteria obtained from the log-likelihood values of the selected optimal estimators or on the minimum root mean square forecast errors. Depending on the number of lags, such a procedure may of course be computationally demanding (but still feasible).

Alternatively, we can exploit the flexibility of the LAR procedure, where the local parametric model, that is, the LAR(p) model, is only required to provide a good approximation of the true latent DGP over the interval of local homogeneity. The small modeling bias guarantees that the confidence set, built on the basis of the upper risk bound given in Equation (10), continues to hold with a slightly smaller coverage probability; see also Čížek, Härdle, and Spokoiny (2009). In other words, even if the assumed lag order of the LAR model is not the true one, but close to it, the procedure is appropriate. Section 3.4.3 addresses the issue of a wrong lag selection within a simulation study, which supports our expectation. To investigate the dual view on long memory, we therefore adopt in the empirical application the most extreme case of a short-memory model, that is, an AR(1) specification.

3.3.3 Parameter r. Belomestny and Spokoiny (2007) suggest to choose $r = 1/2$ in order to provide a stable performance and to minimize the computation error in the Monte Carlo simulation. We follow their recommendation in our empirical application. The sensitivity of the LAR procedure to different values of r is also assessed within a simulation study.

3.3.4 Critical Values. In the testing procedure, critical values measure the significance of ML estimators under the hypothesis of local homogeneity. The critical values are selected

using the general approach of testing theory: to provide a prescribed performance of the procedure under the null hypothesis. In particular, we generate global homogeneous processes, that is, AR(p) models with constant parameters in (8), ensuring homogeneity for every past interval. The critical values are then selected so that the ML estimators under homogeneity fulfill the risk bound (10) over each interval.

As an illustration, we calculate critical values for LAR(1) based on 100,000 generated AR(1) processes with $\theta_t = \theta^* = (\theta_0^*, \theta_1^*, \sigma^*)^\top$ for all t:

$$y_t = \theta_0^* + \theta_1^* y_{t-1} + \varepsilon_t, \qquad \varepsilon_t \sim \mathrm{N}(0, \sigma^{*2}).$$

The starting value is set to $y_0 = \theta_0^*/(1-\theta_1^*)$. The sample size of each process is 1261 in correspondence to the largest interval of $I^K = I^{13} = 5\mathrm{y}$. Under homogeneity, the ML estimator with respect to the largest interval is the optimal estimator (with the smallest variance among others), that is, $\hat\theta_t = \hat\theta_t^K = \tilde\theta_t^K$. Given a reasonable set of critical values, the risk bound (10) holds over the longest interval of homogeneity

$$\mathsf{E}_{\theta^*}\left|\mathrm{LR}\left(I^K, \tilde\theta_t^K, \hat\theta^K_{t(\zeta_1,\ldots,\zeta_K)}\right)\right|^r \le \xi_r. \tag{12}$$

We mimic here the environment of the sequential testing by replacing the unknown hypothetical AR(p) parameter θ^* with the most recently available optimal estimator $\hat\theta_t^k$. In addition, we use the notation $\hat\theta^k_{t(\zeta_1,\ldots,\zeta_k)}$ to emphasize that the adaptively selected estimator depends on the critical values $\{\zeta_1, \ldots, \zeta_k\}$. The bound $\xi_r = \mathsf{E}_{\theta^*}|\,\mathrm{LR}(I^K, \tilde\theta_t^K, \theta^*)|^r$ is empirically calculated. We also notice that the sequential testing procedure accumulates uncertainty in estimation due to the increase in the degrees of freedom. To take this into account, a condition similar to (12) is imposed at each step:

$$\mathsf{E}_{\theta^*}\left|\mathrm{LR}\left(I^k, \tilde\theta_t^k, \hat\theta^k_{t(\zeta_1,\ldots,\zeta_k)}\right)\right|^r \le \frac{k-1}{K-1}\xi_r, \qquad k = 1, \ldots, K. \tag{13}$$

The sequential testing procedure is adopted to compute the critical values. At step $k = 1$, we set $\zeta_1 = \infty$ in agreement with the local homogeneity in the shortest interval I^1 leading to $\hat\theta_t^1 = \tilde\theta_t^1$. In the computation of ζ_2 we set all the remaining $\zeta_k = \infty$ for $k \ge 3$ to specify the contribution of ζ_2 and choose the minimal value of ζ_2 that delivers the estimator satisfying the following risk function:

$$\mathsf{E}_{\theta^*}\left|\mathrm{LR}\left(I^k, \tilde\theta_t^k, \hat\theta^k_{t(\zeta_1,\zeta_2)}\right)\right|^r \le \frac{1}{K-1}\xi_r, \qquad k = 2, \ldots, K.$$

Consequently with $\zeta_1, \zeta_2, \ldots, \zeta_{k-1}$ fixed, we select the minimal value of ζ_k for $k = 3, \ldots, K$ which fulfills

$$\mathsf{E}_{\theta^*}\left|\mathrm{LR}\left(I^q, \tilde\theta_t^q, \hat\theta^q_{t(\zeta_1,\zeta_2,\ldots,\zeta_k)}\right)\right|^r \le \frac{k-1}{K-1}\xi_r, \qquad q = k, \ldots, K.$$

3.3.5 Hypothetical Parameters. Clearly, critical values also depend on the hypothetical parameters θ^* used for generating the homogeneous processes. In our study, we consider two ways for selecting θ^*: a global selection where θ^* is estimated over the full sample period or an adaptive selection where θ^* is reestimated at each time point using a rolling window with a fixed length. For the adaptive selection, a large rolling window size means that we put more attention to a time homogeneous situation. Such a choice leads to a rather conservative procedure with possibly low accuracy of estimation. On the contrary, a rolling window including fewer observations is more sensitive to structural shifts. Alternatively, the size of rolling window can be selected in a data driven way by minimizing some objective function, for example, by minimizing the forecast error, which is however computationally more intensive. In our empirical analysis we consider the predictive performance of the LAR procedure using both the global selection scheme as well as the adaptive selection based on rolling windows of 1 month, 6 months, 1 year, and 2.5 years. As expected, using the time dependent critical values (slightly) increases the accuracy of prediction.

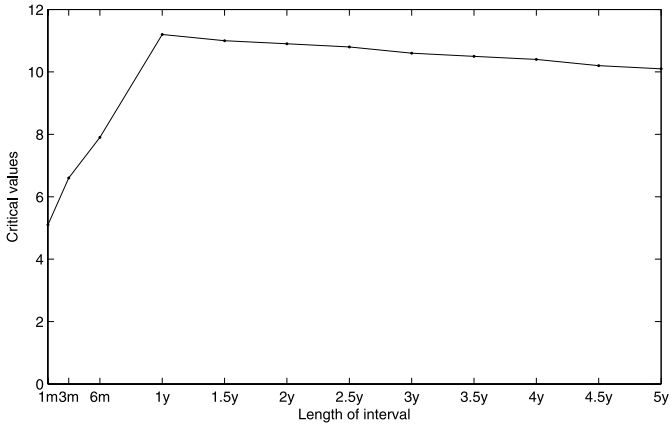


Figure 5. The set of critical values for LAR(1) model. They are based on $r = 1/2$ and on $\theta^* = (-0.1156, 0.7827, 0.5525)^\top$, which are calculated for the log realized volatility of the S&P500 index futures under the hypothesis of constant parameters in (8). The set of interval lengths is given on the x-axis.

Figure 5 depicts the global critical values calculated for a LAR(1) model with $r = 1/2$, the interval candidates given in Section 3.3.1 and the hypothetical AR(p) parameter $\theta^* = (-0.1156, 0.7827, 0.5525)^\top$, the estimates of an AR(1) model fitted to our real dataset—the logarithm of realized volatility of the S&P500 index data.

3.4 Simulation Experiments and Sensitivity Analysis

This section investigates the performance of the localized RV approach in a number of simulation studies focusing on the LAR(1) model. In particular, we assess its performance under different types of structural breaks, we analyze the impact of the parameters involved in the adaptive technique, and we assess the issue of model misspecification, such as a wrong lag selection.

3.4.1 Parameter Changes. In the following we consider the performance of the LAR(1) approach under various scenarios. Specifically, we simulate from an AR(1) with suddenly and gradually changing parameters in order to investigate the appropriateness of the LAR approach under different types of changes. The actual values of the parameters are again based on the estimates of an AR(1) model fitted to the full S&P500 realized volatility data. In each scenario, only one parameter varies over time while the other two remain constant. The processes of the changing parameters are displayed in Figures 6 to 8. The character S denotes a scenario with sudden changes of parameters, where big changes occur at time points $t = 1501$ and

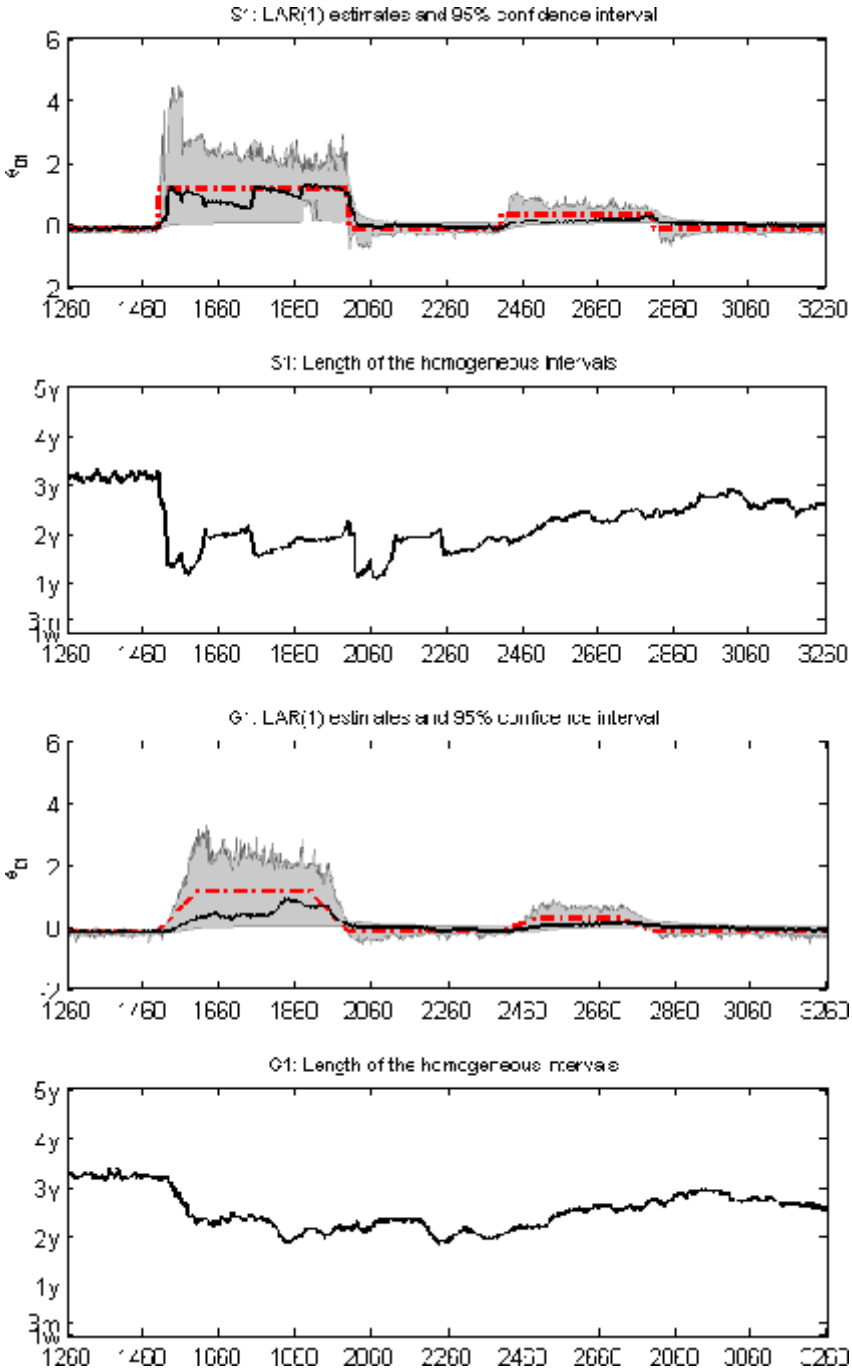


Figure 6. Simulation results for scenarios S1 and G1 (changing parameter: θ_{0t}). The red dashed line represents the process of the true time-varying parameter (S1: $\theta^*_{0t} = 1.1557$ for $t \in [1501, 2000]$, 0.3467 for $t \in [2401, 2800]$, -0.1156 otherwise) and the bold solid line gives the average value of the estimated parameter over 500 simulations. The shaded area corresponds to the pointwise 95% confidence intervals. The average values of the selected homogeneous intervals for each time point are presented in the bottom panel of each scenario. The online version of this figure is in color.

$t = 2000$ and small ones at $t = 2401$ and $t = 2800$, respectively. The G scenarios, on the contrary, denote gradual changes where the parameter gradually reaches to a new level within 100 steps after the change point. For example, in scenario G2 the autoregressive parameter θ_{1t} gradually changes from 0.7827 to -0.7827 over the period from 1501 to 1600, stays at the new level until it drops gradually back to 0.7827 over the period from 2001 to 2100. Similarly the small gradual changes occur over the periods $[2401, 2500]$ and $[2801, 2900]$. For each scenario, we generate 500 LAR(1) processes with 3261 observations. The first 1261 observations, corresponding to the largest interval $I^{13} = 5$ years, are used as training set.

The average value of the estimated parameters (solid line) and the pointwise 95% confidence intervals (shaded areas) are displayed in Figures 6 to 8 along with the true values of θ^* (dashed line). For each point in time the average value of the selected homogeneous intervals is also presented. Obviously, the selected homogeneous intervals are long when the parameters

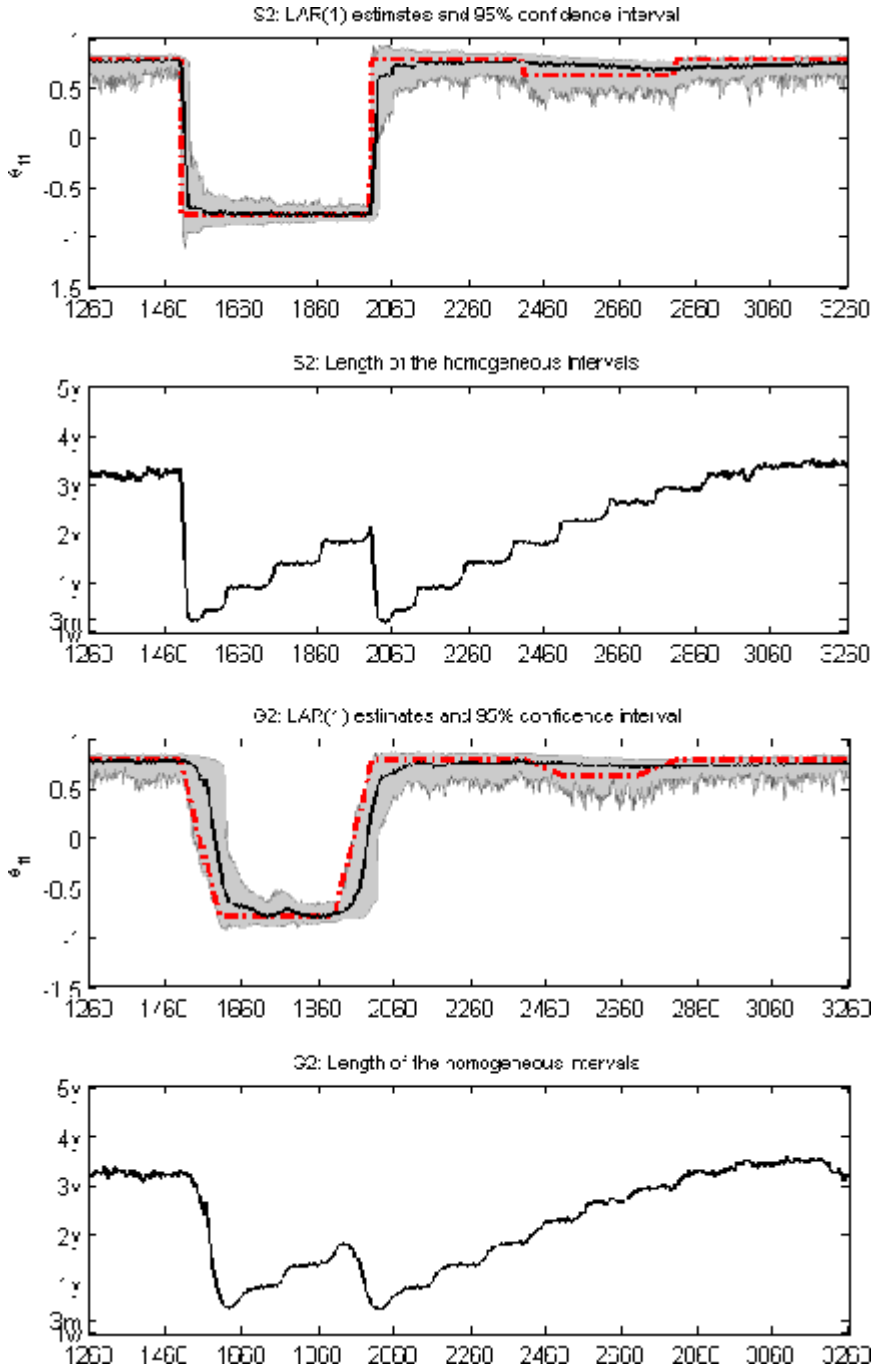


Figure 7. Simulation results for scenarios S2 and G2 (changing parameter: θ_{1t}). The red dashed line represents the process of the true time-varying parameter (S2: $\theta^*_{1t} = -0.7827$ for $t \in [1501, 2000]$, 0.6261 for $t \in [2401, 2800]$, 0.7827 otherwise) and the bold solid line gives the average values of the estimated parameter over 500 simulations. The shaded area corresponds to the pointwise 95% confidence intervals. The average value of the selected homogeneous intervals for each time point are presented in the bottom panel of each scenario. The online version of this figure is in color.

are constant over a long past time interval, but decline sharply when shifts occur. It indicates that the LAR procedure selects reasonable intervals of homogeneity.

In order to assess the performance of the local procedure in more detail, we additionally compute the detection speed, that is, the number of periods required for reaching 50% and 75% of the new level of the parameter. In the G scenarios the counting starts once the parameter has reached its new level, that is, after the gradual changes have finished. In the S scenarios the counting starts immediately from the change point. Table 2 reports the results. In general, the adaptive procedure works well. It shows that the procedure reacts quickly to a big shift, but slowly to a small shift. For example in the scenario S2, where the AR coefficient θ_{1t} jumps at $t = 1501$, the technique only needs 12 periods to catch up 50% of the big shift, while for the small shift at $t = 2401$ it takes 213 periods. This finding, however, is quite reasonable. After a small change of the parameters of the DGP the simulated observations may still

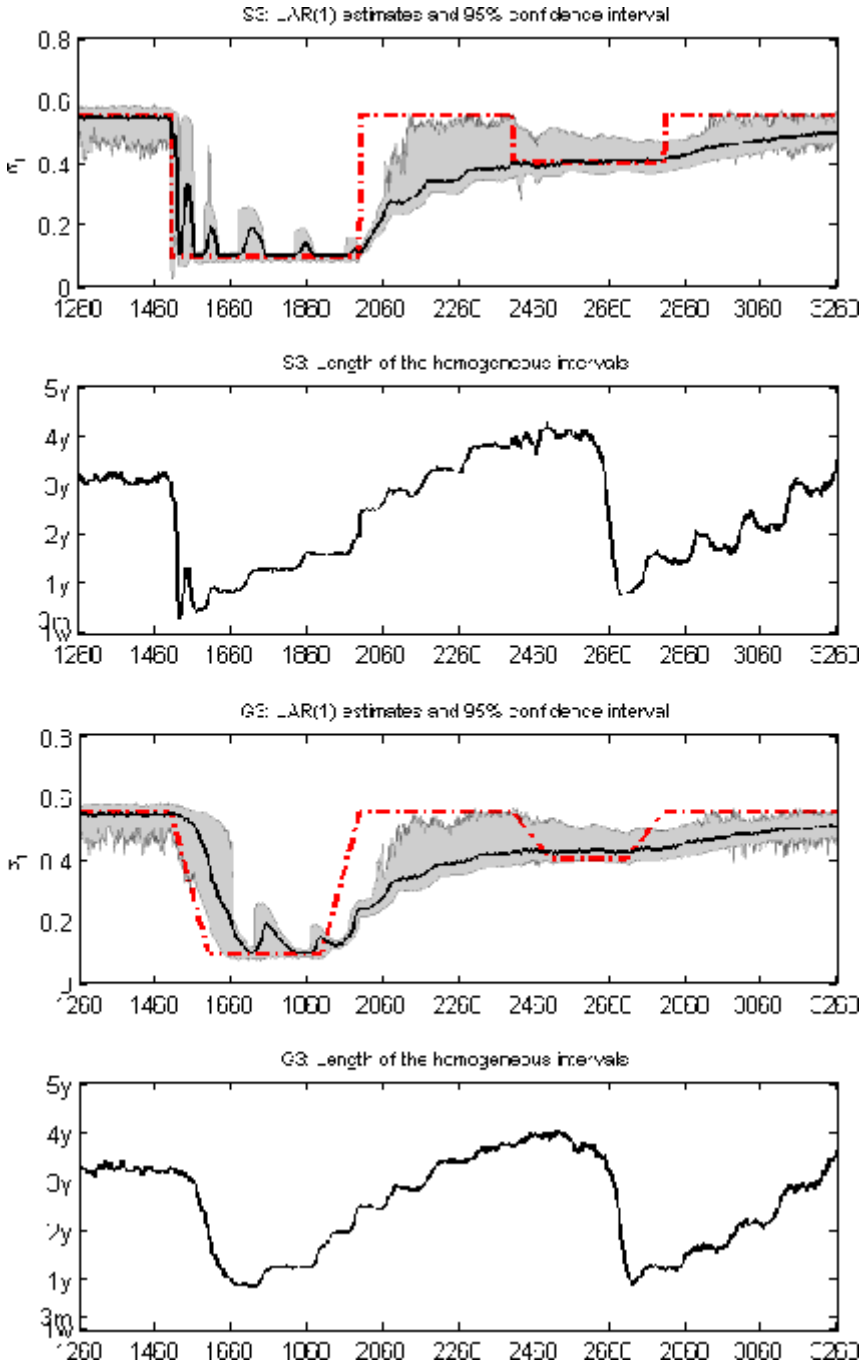


Figure 8. Simulation results for scenarios S3 and G3 (changing parameter: σ_t). The red dashed line represents the process of the true time-varying parameter (S3: $\sigma_t^* = 0.1000$ for $t \in [1501, 2000]$, 0.4000 for $t \in [2401, 2800]$, 0.5525 otherwise) and the bold solid line gives the average values of the estimated parameter over 500 simulations. The shaded area corresponds to the pointwise 95% confidence intervals. The average value of the selected homogeneous intervals for each time point are presented in the bottom panel of each scenario. The online version of this figure is in color.

be very close to those of the previous DGP and it is therefore hard for the procedure to differentiate between the two processes. In this case, more observations from the new DGP are needed for the identification of the parameter change. Nevertheless, the technique is able to detect the changes as more and more small shifts accumulate over time. Similar patterns are observed in the G scenarios which correspond to many small subsequent shifts. The results for the scenarios of σ_t further show that positive shifts, corresponding to an increase in the signal-to-noise ratio, can be only slowly detected; see also Figure 8.

3.4.2 Impact of Parameters. In the following we investigate the effect of the choice of parameters on the performance of the LAR procedure. Here we compute the detection speed of the LAR approach based on different sets of intervals $\{I_k\}_{k=1}^K$, different values of the power transformation parameter r and of the hypothetical AR(p) parameters θ^* used in the computation of the critical values. Moreover, we compute the root

Table 2. Detection speeds for the different scenarios

	S1		S2		S3			G1		G2		G3	
t	50%	75%	50%	75%	50%	75%	t	50%	75%	50%	75%	50%	75%
1501	21	23	12	18	18	20	1601	207	232	1	4	12	56
2001	9	19	13	19	169	>400	2101	1	1	1	6	88	>400
2401	66	374	213	>400	1	1	2501	169	>400	183	>400	1	1
2801	20	243	56	>400	293	>400	2901	1	166	1	346	176	>400

NOTE: Reported are the number of steps required for reaching 50% and 75% of the parameter change.

mean square forecast errors (RMSFEs) of LAR forecasts based on different choices of these parameters. The results are compared to our "default" case, where the parameters are set to the suggested values in Section 3.3, that is, the interval set is given by $\mathbf{I}_\tau = \{I_\tau^k\}_{k=1}^{13} = \{1\text{w}, 1\text{m}, 3\text{m}, 6\text{m}, 1\text{y}, 1.5\text{y}, 2\text{y}, 2.5\text{y}, 3\text{y}, 3.5\text{y}, 4\text{y}, 4.5\text{y}, 5\text{y}\}$, $r = 1/2$ and the vector of hypothetical AR(1) parameters $\theta^* = (-0.1156, 0.7827, 0.5525)^\top$. For the ease of exposition we only report the results for the scenarios with changes in the autoregressive parameter, that is, S2 and G2, as those are also particularly interesting in the model misspecification analysis discussed later. In particular, we consider two alternative sets of intervals. In order to assess the impact of the maximum length of the intervals, we truncate the default set of intervals at $K = 9$ (corresponding to 3 years), while the second scenario aims at investigating the sensitivity of the procedure towards a finer grid of intervals by including more intermediate subintervals, that is, introducing a three-months grid such that $K = 22$. We further evaluate the impact of a smaller value and a larger value of the power transformation parameter setting $r = 1/3$ and $r = 1$. As the critical values rely on the choice of the hypothetical parameters, we check the predictive performance using $80\%\theta^*$ and $120\%\theta^*$ to generate the homogeneous processes in the computation of critical values, which can be interpreted as an underestimation and overestimation of the actual parameter values, respectively.

Table 3 presents the results. In order to facilitate the comparison, we report here the relative average RSMFE of one-step ahead predictions, that is,

$$\text{R-RMSFE} = \sum_{j=1}^{500} \text{R-RMSFE}_j^{\text{nondefault}} \Big/ \sum_{j=1}^{500} \text{R-RMSFE}_j^{\text{default}},$$

where the average value of the RMSFEs with default choice is 0.5411 for S2 and 0.5374 for G2. We also define the relative detection speed as the difference of the average detection speed of the LAR procedure based on nondefault parameters to the average detection speed using the default choice. Thus, a positive/negative value indicates a slower/faster reaction of the technique with nondefault choices. The results illustrate well, that the LAR procedure is quite robust to the choice of the parameters. The "worst" cases appear when CVs are calculated based on imprecise hypothetical AR(p) parameters: a 2.74% improved predictability for $0.8\theta^*$ and a 3.95% worse performance for $1.2\theta^*$. It suggests that using alternative choices of the parameters delivers only small deviations from the default choices. Moreover, there are no crucial changes in the detection speed in the presence of large parameter changes, although

Table 3. Sensitivity analysis: impact of parameters

	Choice of parameters											
	$K = 9$		$K = 22$		$r = 1/3$		$r = 1$		$0.8\theta^*$		$1.2\theta^*$	
					Scenario S2							
R-RMSFE:	0.9956		1.0128		1.0102		0.9974		0.9726		1.0395	
R-DS:	50%	75%	50%	75%	50%	75%	50%	75%	50%	75%	50%	75%
$t = 1501$	−1	−1	2	2	0	0	2	2	−1	−4	6	5
$t = 2001$	0	0	3	0	1	0	2	0	−3	0	5	1
$t = 2401$	0	–	5	–	5	–	0	–	0	–	47	–
$t = 2801$	>344	–	>344	–	>344	–	>344	–	>344	–	>344	–
					Scenario G2							
R-RMSFE:	0.9946		1.0132		1.0143		0.9922		0.9728		1.0506	
R-DS:	50%	75%	50%	75%	50%	75%	50%	75%	50%	75%	50%	75%
$t = 1601$	0	0	0	0	0	0	0	0	0	0	0	4
$t = 2001$	0	0	0	0	0	0	0	0	0	0	0	6
$t = 2501$	0	–	0	–	0	–	0	–	0	–	184	–
$t = 2801$	>399	>54	>399	>54	>399	>54	>399	>54	>399	>54	>399	>54

NOTE: Reported are the relative one-step-ahead RMSFEs and the relative detections speeds (R-DS) in the scenarios S2 and G2 for different choices of the parameters. The default choice (i.e., the benchmark) is given by $K = 13$, $r = 1/2$, and $\theta^* = (-0.1156, 0.7827, 0.5525)^\top$. In the alternative choices only one parameter is changed fixing the other ones to the default values. $K = 9$ corresponds to a scenario with reduced maximum interval length while $K = 22$ is characterized by a finer grid of intervals. More details are given in the text. Scenarios S2 and G2 are displayed in Figure 7. "–" indicates that the detection speeds in both scenarios are greater than 400.

for small parameter changes the detection speed slows down. In general, the sensitivity analysis supports our default choice of parameters and the results suggest that for an adaptive, data-driven computation of the critical values the selection of the parameters may become even less important with respect to predictability.

3.4.3 Model Misspecification. In this section we investigate the robustness of the LAR procedure towards model misspecification, that is, if the true DGP has a different lag structure than the assumed one or, even worse, if the true DGP follows a different dynamic structure. The analysis is twofold: we first focus on short-memory models, which allow us to evaluate the impact of the lag order, and then consider the performance of the LAR procedure if the true DGP is a long-memory process.

For the short-memory scenarios we consider the local constant model, that is, $y_t = \theta_{0t} + \varepsilon_t, \varepsilon_t \sim \mathrm{N}(0, \sigma_t^2)$ and the LAR model with lag order $p = 2, 5$, and 10, $y_t = \theta_{0t} + \sum_{j=1}^{p} \theta_{jt} y_{t-j} + \varepsilon_t, \varepsilon_t \sim \mathrm{N}(0, \sigma_t^2)$, in order to account for the situations where the true DGPs either have less or more lags than is assumed in the LAR(1). The actual parameter values are again set to the ML estimates obtained by fitting the globally constant version of these models to the full S&P500 data sample. The design of the time variation in the parameters is similar to the scenarios in Figures 6 to 8, where the big and small changes are determined by a new level of the parameters (e.g., $-1\theta_p$ and $0.8\theta_p$ respectively). As the focus is on the impact of a misspecification in the lag order, we consider here only cases with changes in θ_{0t} in the local constant model and changes in θ_{pt}, the pth autoregressive part of the LAR(p) model. In the long-memory scenario we simulate from an ARFIMA$(2, 0.47, 0)$ with constant parameters. The specification of the ARFIMA model is also guided by the empirical results obtained for the full S&P500 sample; see Section 4. For each DGP, 500 series are simulated, each with a length of 3261 observations.

The sensitivity of the LAR procedure towards model misspecification is assessed in terms of predictability. In particular, we compute for each simulated series 2000 one-step-ahead forecasts based on: (i) the "wrong" but flexible LAR(1) approach, where the intervals of local homogeneity are selected by using the adaptive technique; (ii) the true data-generating model using optimally time-varying window size. More precisely, for a particular time point the optimal window is either identified using the LAR(p) procedure (for LAR scenarios) or assumed to be the interval used for generating the process (otherwise). In addition, the shortest/longest length of the intervals is set to include 15/1250 observations, which is in line with the assumption of homogeneity in the LAR procedure and assures the feasibility of estimation. The average value of the RMSFEs for different scenarios are reported in Table 4. In most cases, the forecasts based on the LAR(1) specification yield only slightly bigger RSMFEs than the true DGPs. It supports that the LAR procedure with the lag order $p = 1$ can provide a quite accurate approximation. In other words, the LAR procedure is quite robust to the selection of the lag order p. Moreover, the LAR(1) performs also well if the true source of the long-range dependence is a long-memory process, confirming that long memory can well be approximated by a short-memory model with breaks. In summary, the simulation shows that the local adaptive procedure with lag order $p = 1$ is a reasonable approximation, even if the underlying process deviates from the LAR(1) setup.

Table 4. Sensitivity analysis: model misspecification

DGP:	Local const. θ_{0t}	LAR(2) θ_{2t}	LAR(5) θ_{5t}	LAR(10) θ_{10t}	ARFIMA
$\widehat{\mathrm{DGP}}$	1.0225	0.6247	0.5664	0.5568	0.5105
LAR(1)	0.9339	0.6293	0.5848	0.5724	0.5619

NOTE: Reported are the average RMSFEs based on the LAR(1) procedure and the estimated data-generating processes, $\widehat{\mathrm{DGP}}$.

4. ALTERNATIVE MODELS

As we aim at a comparison of the LAR procedure to the long memory view of volatility, we primarily consider alternative models that emanate from this view. Nevertheless, we also compare our procedure to the smooth transition regression tree (STR-Tree) model, that is, a model with breaks.

The ARFIMA model is one of the standard models used in the realized volatility literature; see, for example, Andersen et al. (2003). Under the ARFIMA(p, d, q) model, the dynamics of logarithmic realized volatility is given by

$$\phi(L)(1-L)^d(\log RV_t - \mu) = \psi(L)u_t, \tag{14}$$

with $\phi(L) = 1 - \phi_1 L - \cdots - \phi_p L^p$, $\psi(L) = 1 + \psi_1 L + \cdots + \psi_q L^q$, L denoting the lag operator, and $d \in (0, 0.5)$ is the fractional difference parameter. Given the empirical distributional properties of logarithmic realized volatility, u_t is usually assumed to be a Gaussian white noise process, which facilitates the exact maximum-likelihood estimation of the model.

The HAR model aims at reproducing the observed volatility phenomenon. However, in contrast to the ARFIMA model, the HAR model is formally not a long-memory model. Instead, the correlation structure is approximated by the sum of a few multiperiod volatility components. The use of such components is motivated by the existence of heterogenous agents having different investment horizons; see Corsi (2009) and Müller et al. (1997). In particular, the HAR model put forward by Corsi (2009) builds on a daily, weekly, and monthly component, which are defined by

$$RV_{t+1-k:t} = \frac{1}{k}\sum_{j=1}^{k} RV_{t-j}$$

with $k = 1, 5, 21$, respectively. The HAR model is then given by

$$\log RV_t = \alpha_0 + \alpha_d \log RV_{t-1} + \alpha_w \log RV_{t-5:t-1} + \alpha_m \log RV_{t-21:t-1} + u_t \tag{15}$$

with u_t typically being also Gaussian white noise. Maximum-likelihood estimation is straightforward. Interestingly, the HAR and ARFIMA models have been found to obtain a similar forecasting performance with both models outperforming the traditional volatility models based on daily returns; see, for example, Andersen, Bollerslev, and Diebold (2007) and Koopman, Jungbacker, and Hol (2005).

It is sometimes argued that volatility exhibits both long memory and structural breaks. We therefore compare our procedure

also to the adaptive ARFIMA model, that has recently been developed in Baillie and Morana (2009a) for modeling inflation dynamics. The model is based on a time-dependent intercept that is given by a Flexible Fourier Form representation, and an innovation term that follows a stationary long-memory process. The flexible functional form of the intercept allows for smooth as well as sharp nonlinearities without the need to identify break points and the magnitude of the breaks. Baillie and Morana (2009b) have shown that a FIGARCH model with such a time dependent intercept provides superior volatility forecasts in comparison to alternative GARCH and adaptive GARCH specifications. We therefore adopt the adaptive ARFIMA model for modeling logarithmic realized volatility, which is given by

$$\log RV_t = \mu + \sum_{j=1}^{k} (\sin(2\pi jt/T) + \delta_j \cos(2\pi jt/T)) + u_t, \quad (16)$$

where

$$\phi(L)(1-L)^d u_t = \psi(L)\epsilon_t.$$

It is characterized as A-ARFIMA(p, d, q, k).

The STR-Tree model proposed in da Rosa, Veiga, and Medeiros (2008) provides an interesting alternative to the LAR procedure. It builds on the methodology of classification and regression tress, where it is assumed that the dependent variable is given by the sum of regression models, each of which is determined by recursive partitions of the covariate space. The structure of a regression tree model is usually represented in the format of a binary choice decision tree with a set of parent and terminal nodes, denoted here by $\mathbb{J}$ and $\mathbb{K}$, respectively. The splits at the parent nodes are sharp. The STR-Tree model instead smoothes the splits by replacing the indicator function by a logistic function

$$G(x; \gamma, c) = \frac{1}{1 + e^{-\gamma(x-c)}}.$$

Scharth and Medeiros (2009) advocate the use of the STR-Tree approach for modeling logarithmic realized volatility, that is,

$$\log RV_t = \boldsymbol{\alpha}^\top \mathbf{w}_t + \sum_{k \in \mathbb{K}} \boldsymbol{\theta}_k^\top \mathbf{z}_t B_{\mathbb{J}k}(\mathbf{x}_t, \boldsymbol{\beta}_k) + u_t, \quad (17)$$

where $\mathbf{x}_t = (x_{1,t}, \ldots, x_{m,t})^\top$ denotes the vector of explanatory variables, or in terms of the smooth transition literature the so-called transition variables, $\mathbf{z}_t = (\log RV_{t-1}, \ldots, \log RV_{t-p})^\top$, and $\mathbf{w}_t$ is a vector of linear regressors that are unaffected by the tree, $\mathbf{w}_t \nsubseteq \mathbf{x}_t$. Moreover,

$$B_{\mathbb{J}k}(\mathbf{x}_t, \boldsymbol{\beta}_k) = \prod_{j \in \mathbb{J}} G\big(x_{s_j,t}; \gamma_j, c_j\big)^{n_{k,j}(1+n_{k,j})/2} \times \big[1 - G\big(x_{s_j,t}; \gamma_j, c_j\big)\big]^{(1-n_{k,j})(1+n_{k,j})}, \quad (18)$$

where $s_j \in \{1, \ldots, m\}$ gives the transition variable being relevant at node j and

$$n_{k,j} = \begin{cases} -1 & \text{if parent node } j \text{ is not included} \\ & \text{in the path to terminal node } k \\ 0 & \text{if the right-child node of parent node } j \\ & \text{is included in the path to terminal node } k \\ 1 & \text{if the left-child node of parent node } j \\ & \text{is included in the path to terminal node } k. \end{cases}$$

The spirit of the STR-Tree model is similar to the LAR procedure in the sense that realized volatility is approximated by local AR(p) models. However, in the STR-Tree model the regimes are due to partitions of the transition variables, such as lagged returns (capturing the well-known leverage effect), which are determined globally, that is, over the full sample period. The LAR instead is more flexible, as the interval of homogeneity is determined locally. Moreover, it does not require the specification of a set of variables that may lead to parameter changes. In fact, any event or changes in variables that affect the parameters of the AR(p) model such that local homogeneity is rejected are automatically encountered in the procedure.

5. EMPIRICAL ANALYSIS

We now turn to the empirical investigation of the dual views on the dynamics of volatility. We focus our analysis on realized volatility of the S&P500 index futures from January 2, 1985 to February 4, 2005 (see Section 2). Like in the simulation exercise we use the first 5 years of our sample as a training set. For the local autoregressive procedure this means that January 2, 1990 is the first time point for which we estimate the LAR model and that we allow the longest interval of homogeneity ($K = 13$) to be 5 years with the remaining set of subintervals given as in the Section 3.3, that is, 1 week ($k = 1$), 1 month ($k = 2$), …, 4.5 years ($k = 12$).

The estimation of the LAR model is conducted for different sets of critical values, in order to assess also the empirical sensitivity of the approach with respect to the choice of the critical values. We therefore consider critical values obtained from a Monte Carlo simulation based on the parameter values of the AR model being estimated over the full sample period. We refer to this as the global LAR model. The other sets of critical values are obtained adaptively using a 1 month, 6 months, 1 year, and 2.5 years sample period. Figure 9 shows the distribution of the lengths of the selected homogenous intervals of the LAR(1) model over the evaluation period (January 2, 1990 to February 4, 2005) based on the global and the adaptive

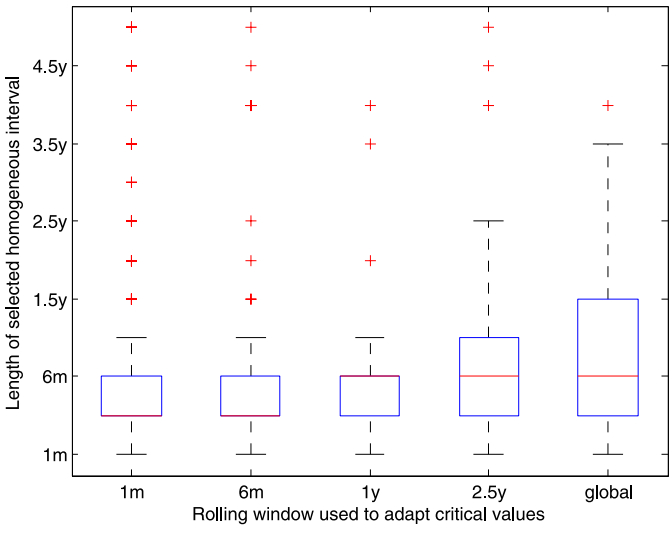


Figure 9. Boxplot of the homogenous intervals selected by the LAR(1) procedure with 1 month, 6 months, 1 year, 2.5 years adaptive critical values and the global LAR(1) procedure.

critical values. Obviously, the global LAR(1) model exhibits a slightly higher variation in the length of the selected intervals. Interestingly, with the exception of the adaptive 1 month and 6 months LAR(1) models for which the median interval length is at $k = 3$, we find that the median is $k = 4$, which corresponds to 6 months of homogeneity. Furthermore, note that the average interval length is for nearly all LAR(1) models about 6 months, which indicates only a weak sensitivity of the interval selection procedure to the sample size used in the computation of the critical values.

In our analysis we assess the forecasting performance for several periods into the future. Such multiperiod predictions may seem to be at odds with the idea of the LAR procedure, which builds on local homogeneity. Local homogeneity has the advantage that forecasts are based only on the most recent information being relevant at the particular forecast origins. But for iterative long-term predictions it also implies that the procedure may perform poor as for increasing forecast horizons it becomes more likely that the assumption of local homogeneity is violated. Nevertheless, the advantage of local homogeneity can also be transferred to the case of multiperiod predictions by incorporating the forecast horizon into the adaptive selection via a restricted LAR(h) specification:

$$\log RV_{t+h} = \theta_{0t} + \theta_{ht} \log RV_t + \varepsilon_t, \qquad \varepsilon_t \sim \mathrm{N}(0, \sigma_t^2), \quad (19)$$

which leads to a direct forecasting approach. We adopt this specification in the empirical analysis.

Table 5 presents the RMSFEs of the LAR model for the 1-day, 5-days, and 10-days ahead forecasts using the different sets of critical values. The empirical results reveal that an adaptive approach and a reduction of the sample period underlying the computation of the critical values introduces more flexibility into the procedure, which seems to result in an increase in forecast accuracy.

We investigate the dual views by comparing the forecasting performance of the LAR procedure to the alternative models. To this end, we recursively compute (logarithmic) realized volatility forecasts from all model types over the evaluation period. Moreover, as we have observed different degrees of persistence in log realized volatility for different lengths of the sample period (see Figure 1), we consider for each of the alternative models forecasts conditional on different information sets, that is, different sample sizes.

Table 5. Root mean square forecast errors of the LAR model based on different sets of critical values

Sample size used in the critical values	$h = 1$	$h = 5$	$h = 10$
1m	0.4823	**0.4619**	**0.4615**
6m	**0.4791**	0.4791	0.4873
1y	0.4842	0.4881	0.4945
2.5y	0.4898	0.5027	0.5056
Global	0.4986	0.5660	0.5884

NOTE: The table reports the root mean square forecast errors (RMSFE) of the h-day ahead logarithmic realized volatility forecasts of the S&P500 index futures based on the LAR(h) models. The first column refers to the information set that is used in the computation of the critical values. For example, the number reported in the first upper-left cell gives the RMSFE of forecasts based on the LAR(1) approach with critical values being computed adaptively over the previous month. Global indicates that the critical values have been computed based on the full sample. Bold numbers indicate the minimum RMSFE for each forecast horizon.

More precisely, forecasts of the ARFIMA, adaptive ARFIMA, and HAR models are based on a rolling window scheme, with rolling window sizes ranging from 3 months to 5 years, which is broadly consistent with our choice of subintervals in the LAR procedure. The conditioning on the different sample sizes is also an attempt to account for the possibility that both long memory and structural breaks are driving volatility. For the STR-Tree model we follow Scharth and Medeiros (2009), and form forecasts based on the recursive scheme. We additionally compute forecasts from constant AR models conditional on the set of rolling windows used also in the HAR and ARFIMA models, as this allows for a direct evaluation of the relevance of the local selection of the interval length employed in the LAR procedure. Such an evaluation requires that forecasts from AR models are also based on the direct forecasting approach.

The forecasts of the other models are computed iteratively, such that their specifications remain the same for all forecast horizons. In particular, the ARFIMA forecasts are based on an ARFIMA$(2, d, 0)$ specification, which was selected according to the Akaike as well as the Bayesian information criteria using the full sample period. For the adaptive ARFIMA model we obtain an A-ARFIMA$(1, d, 1, 2)$ specification with $\gamma_2 = 0$. Estimation and forecasting is carried out using the Ox ARFIMA 1.04 package; see Doornik and Ooms (2004, 2006). For the STR-Tree model we consider the daily lagged return as the transition variable in order to account for the most popular leverage specification. Moreover, for consistency with the short-memory models considered in this paper, we set $p = 1$, and let only the AR(1) coefficient be affected by the tree as indicated by statistical tests on the relevance of explanatory variables in the tree based on the full sample period. Over this period the model is characterized by two splits. In computing the forecasts we respecify the tree structure and reestimate the model every period. We are grateful to Marcel Scharth for providing us with his code. Multistep forecasts are based on conditional simulations as explained in the appendix of Scharth and Medeiros (2009).

For the ease of exposition we do not report all forecasting results but instead focus only on those models that yielded the minimal RMSFE within each model class. Table 6 thus reports the RMSFE of the "best" models along with the corresponding conditioning information set for which the forecasts have

Table 6. Root mean square forecast errors and information sets of the best models

	$h = 1$		$h = 5$		$h = 10$	
Model	RMSFE	Info set	RMSFE	Info set	RMSFE	Info set
LAR	0.4791	6m	0.4619	1m	0.4615	1m
AR	0.5047	3m	0.5712	3m	0.5873	3m
STR-Tree	0.5547	Rec.	0.7746	Rec.	0.8738	rec.
ARFIMA	0.4991	3y	0.5827	3y	0.6207	3y
A-ARFIMA	0.5020	4.5y	0.5904	4.5y	0.6312	4y
HAR	0.5014	3y	0.5848	2.5y	0.6232	2.5y

NOTE: The table reports the root mean square forecast errors (RMSFE) of the h-day ahead logarithmic realized volatility forecasts of the S&P500 index futures based on the various models. Reported are the results for the models yielding minimal RMSFE within each model class. "Info set" refers to the corresponding sample size used in the computation of the critical values (for the LAR procedure) or to the size of the rolling window used in model estimation and prediction (for the AR, ARFIMA, and HAR models). "Rec." refers to forecasts based on the STR-Tree model, for which the recursive forecasting scheme is employed.

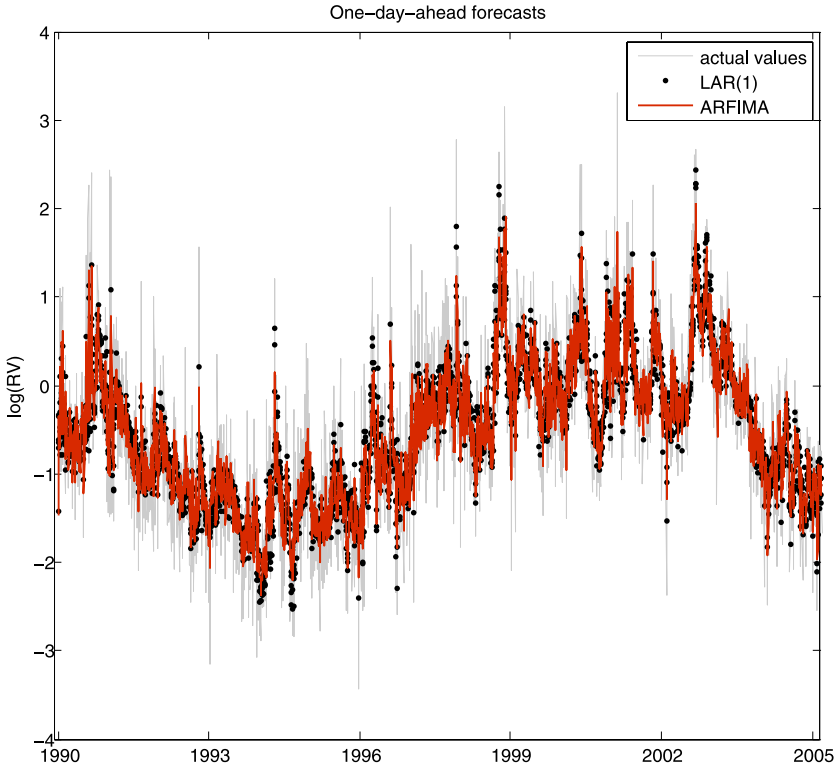


Figure 10. Time-evolvement of the actual log realized volatility (grey line in the background) and the one-step ahead forecasts of (i) the LAR(1) model with critical values being computed over 6 months and (ii) the ARFIMA model based on a 3-year rolling window. These model specifications yield the minimum RMSFE within each model class (see Table 6).

been found to minimize the RMSFE. That is the information set reports either the rolling window size or the sample size used in the computation of the critical values. An illustration of the time-evolvement of the forecasts is presented in Figure 10 which depicts the one-day ahead forecasts of the LAR(1) and ARFIMA models having minimal RMSFEs.

Interestingly, according to the RMSFEs our LAR procedure provides the most accurate forecasts at all forecast horizons. Note that this already holds for the forecasts based on the LAR model with globally computed critical values, which can be readily inferred by comparing the results reported in Tables 5 and 6.

The direct comparison of the LAR forecasts with those based on the constant AR models also reveals, that the selection of the locally homogenous intervals is indeed important. The adaptive procedure, which determines at each time point the adequate length of the time interval over which the AR model is appropriate, is superior. Note that for increasing window sizes, that is, larger information sets, the predictability of the constant AR model worsens (results are not reported gere, but are available from the authors upon request). This might be expected as for larger sample sizes, for example, more than 2 years, the autocorrelation function of realized volatility exhibits more persistence and, thus, an AR model tends to be misspecified. The STR-Tree model, instead, is better suited to generate long-range dependence as it picks local AR(1) specifications that depend on the state of the lagged daily return. It is therefore surprising that the model performs worse than those without leverage effect. However, this may be due to our model specification that makes only use of past daily returns. For a different dataset, Scharth and Medeiros (2009), for example, find a superior performance of the STR-Tree model where the splits are determined by returns accumulated over the past 90, 39, 5, and 2 days, indicating that long-term returns are important when modeling and forecasting realized volatility. A more thorough treatment of the leverage effect is the subject of future research.

In accordance to the empirical results reported in the realized volatility literature so far, the HAR and ARFIMA models exhibit similar forecast accuracy with a slight tendency of the ARFIMA model to outperform the HAR model. Interestingly, the results indicate that the inclusion of structural changes in the form of the adaptive ARFIMA model does not lead to improvements in the predictability of the S&P500 realized volatility. Moreover, all long-memory models are outperformed by the LAR method. This becomes even more pronounced for larger forecast horizons. In order to get a feeling of whether this is due to a comparison of direct with iterated forecasts we have additionally computed direct forecasts for the HAR model. We find that the iterated method provides better forecasts than the direct one (e.g., the RSMFE of the direct HAR forecasts based on a 2.5 year rolling window size is 0.5857 for $h = 5$ and 0.6240 for $h = 10$), which is consistent with the recent empirical findings reported in Ghysels, Rubia, and Valkanov (2009) and Marcellino, Stock, and Watson (2005).

We further evaluate the predictive performance of the different realized volatility models on the grounds of the so-called Mincer–Zarnowitz regressions, that is, by regressing the observed log realized volatility on the corresponding forecasts of model i:

$$\log RV_t = \alpha + \beta \log \widehat{RV}_{t,i} + \nu_t. \tag{20}$$

Table 7. Results of the Mincer–Zarnowitz regressions and Diebold–Mariano tests for the volatility models with minimal RMSFEs

Model	p-value	R^2	*DM* (best LAR) t-stat.
	$h = 1$		
LAR, 6m	0.7242	0.7180	
3m AR(1)	0.6203	0.6872	−9.5125
STR-Tree	0.0014	0.6242	−14.8673
3y ARFIMA	0.6375	0.6942	−6.2782
4.5y A-ARFIMA	0.0388	0.6909	−6.8551
3y HAR	0.8842	0.6910	−6.9275
	$h = 5$		
LAR, 1m	0.1265	0.7377	
3m AR(5)	0.2326	0.6004	−14.4786
STR-Tree	0.0000	0.4860	−15.3163
3y ARFIMA	0.5656	0.5835	−14.2157
4.5y A-ARFIMA	0.0233	0.5745	−14.1834
2.5y HAR	0.7427	0.5803	−14.7150
	$h = 10$		
LAR, 1m	0.0705	0.7392	
3m AR(10)	0.0897	0.5789	−13.8568
STR-Tree	0.0000	0.1911	−14.7564
3y ARFIMA	0.7593	0.5273	−12.8450
4y A-ARFIMA	0.2201	0.5123	−12.4366
2.5y HAR	0.6611	0.5236	−13.0511

NOTE: Reported are results of the Mincer–Zarnowitz regressions and of the modified Diebold–Mariano tests for the models yielding the minimum RMSFE within each model class (see Tables 5 to 6). The results are reported for different forecast horizons h (in days). The second column reports the p-value of a F-test for $H_0 : \alpha = 0$ and $\beta = 1$, and the third column reports the coefficient of determination (R^2) of the Mincer–Zarnowitz regression given in Equation (20). The last column gives the modified t-statistics of the Diebold–Mariano test on equal forecast performance, that is, $H_0 : \mu = 0$ in the regression $e^2_{t,\text{LAR}} - e^2_{t,i} = \mu + \nu_t$ with $e_{t,i}$ denoting the forecast error of model i. Results are based on heteroscedasticity and autocorrelation robust Newey–West (co)variances.

This allows to test for the unbiasedness of the different forecasts. Table 7 reports the coefficients of determination (R^2s) of this regression along with the p-value of the F-test on unbiased forecasts, i.e., $H_0 : \alpha = 0$ and $\beta = 1$. Note that for the ease of exposition we again solely present here the comparison of the models performing best in terms of the RMSFE.

The results indicate that, with the exception of the forecasts of the STR-Tree model, none of the forecasts is significantly biased at the 5% significance level. The coefficients of determination reported in Table 7 indicate a superior forecasting performance of the adaptive LAR models. We investigate this result further and test for the significance of the observed differences in the forecast accuracies. In particular, we conduct a pairwise test on the equality of the mean square forecast errors (MSFE) of the LAR procedure and the other models; see Diebold and Mariano (1995). To this end, we regress the difference between the squared forecast errors of the LAR model and those of the competing model i, that is, $e^2_{t,\text{LAR}} - e^2_{t,i}$, on a constant μ. The null hypothesis of equal MSFEs is equivalent to $H_0 : \mu = 0$. Table 7 reports the modified Diebold–Mariano test statistics proposed in Harvey, Leybourne, and Newbold (1997). Obviously, the null hypothesis is always strongly rejected in favor of a significant better forecasting performance of the adaptive LAR model, as indicated by the significant negative sign of the t-statistic. Overall, the LAR approach seems to be superior.

However, it should be noted that this conclusion is based on a pairwise comparison of the best models and there may be LAR models for which this is not the case. A simultaneous comparison of the predictive ability of all competing models would be desirable at this stage. However, the corresponding existing tests, like the test for superior predictive ability (SPA) of Hansen (2005) and the model confidence set approach of Hansen, Lunde, and Nason (2010) are not applicable here, as the forecasts are based on time varying window sizes (given by the locally selected interval of homogeneity and the recursive forecasting scheme employed in the STR-Tree model), which violates the assumption of strict stationarity of the loss differential. The Diebold–Mariano test, in contrast, can still be applied; see Giacomini and White (2006). To obtain a broader picture on the performance of the LAR procedure, we therefore extend the pairwise comparisons. In particular, we additionally conduct a pairwise comparison of forecasts of the alternative models conditional on a moderately small sample (1 year) and on a large sample (5 years) with forecasts from the LAR models based on 1 year adaptively and on globally computed critical values. Note that for the ease of exposition we do not report the corresponding results here, however, they are available from the authors upon request. Overall, the results are similar to the ones reported in Table 7. Only for the global LAR model, we fail to reject the null in the comparison with the one-step-ahead forecasts of the long-memory models. But also in those cases the t-statistics are negative.

6. CONCLUSION

This paper investigates a dual view on the long-range dependence of realized volatility. While the current realized volatility literature primarily advocates the use of long-memory models to explain this phenomenon, we argue that volatility can alternatively be described by short-memory models with structural breaks. To this end we propose localized realized volatility modeling where we consider the case of a dynamic short-memory model. In particular, at each point in time we determine an interval of homogeneity over which the volatility is approximated by an AR process. Our approach is based on local adaptive techniques developed in Belomestny and Spokoiny (2007), which make it flexible and allow for time-varying coefficients. It does neither require the specification of the type, magnitude or reasons of breaks. This contrasts to smooth transition or regime switching models.

Our procedure relies on parameters, that have to be predetermined. A simulation study, however, shows that the procedure is quite robust to the choice of parameters and to model misspecification. Interestingly, the method performs also well, even if the true source of the long-range dependence is a long-memory process. Moreover, we show, that an adaptive view on intervals of local homogeneity (and a decrease in the respective underlying sample size) is increasing the procedure's flexibility, yielding higher accuracy in estimation and a better forecasting performance. Furthermore, the choice of the underlying parameters can also be based upon criteria reflecting the user's objective, such as in sample fit or forecasting criteria. Although we have refrained from doing so in our empirical application, we find that our adaptive localized realized volatility procedure

provides accurate volatility forecasts and significantly outperforms the standard long-memory realized volatility models and two alternative models with breaks. It seems that our view on volatility is practical and realistic.

Extensions of the local parametric model to explicitly account for other important data characteristics, such as the leverage effect, are left for future research.

[Received January 2009. Revised June 2010.]

REFERENCES

Allen, D. E., McAleer, M., and Scharth, M. (2010), "Realized Volatility Risk," Discussion Paper CIRJE F-693, CIRJE, University of Tokyo, Faculty of Economics. [1377]

Andersen, T. G., and Bollerslev, T. (1998), "Answering the Skeptics: Yes, Standard Volatility Models Do Provide Accurate Forecasts," *International Economic Review*, 39, 885–905. [1378]

Andersen, T. G., Bollerslev, T., and Diebold, F. X. (2007), "Roughing It Up: Including Jump Components in the Measurement, Modeling and Forecasting of Return Volatility," *Review of Economics and Statistics*, 89, 701–720. [1387]

Andersen, T., Bollerslev, T., Diebold, F., and Ebens, H. (2001a), "The Distribution of Realized Stock Return Volatility," *Journal of Financial Economics*, 61, 43–76. [1378]

Andersen, T. G., Bollerslev, T., Diebold, F. X., and Labys, P. (2001b), "The Distribution of Realized Exchange Rate Volatility," *Journal of the American Statistical Association*, 96, 42–55. [1376]

——— (2003), "Modeling and Forecasting Realized Volatility," *Econometrica*, 71, 579–625. [1376,1387]

Bai, J., and Perron, P. (1998), "Estimating and Testing Linear Models With Multiple Structural Changes," *Econometrica*, 66, 47–78. [1380]

Baillie, R. T., and Morana, C. (2009a), "Investigating Inflation Dynamics and Structural Change With an Adaptive ARFIMA Approach," Working Paper 6/09, International Centre for Economic Research. [1377,1388]

——— (2009b), "Modelling Long Memory and Structural Breaks in Conditional Variances: An Adaptive FIGARCH Approach," *Journal of Economics Dynamics and Control*, 33, 1577–1592. [1377,1388]

Baillie, R. T., Bollerslev, T., and Mikkelsen, H. O. (1996), "Fractionally Integrated Generalized Autoregressive Conditional Heteroskedasticity," *Journal of Econometrics*, 74, 3–30. [1376]

Bandi, F. M., and Russell, J. R. (2005), "Microstructure Noise, Realized Volatility, and Optimal Sampling," *Review of Economic Studies*, 75, 339–369. [1378]

Barndorff-Nielsen, O. E., and Shephard, N. (2002a), "Econometric Analysis of Realized Volatility and Its Use in Estimating Stochastic Volatility Models," *Journal of the Royal Statistical Society, Ser. B*, 64, 253–280. [1378]

——— (2002b), "Estimating Quadratic Variation Using Realized Variance," *Journal of Applied Econometrics*, 17, 457–477. [1378]

Barndorff-Nielsen, O. E., and Veraart, A. E. D. (2009), "Stochastic Volatility of Volatility in Continuous Time," Research Paper 2009-25, CREATES. [1377]

Barndorff-Nielsen, O. E., Hansen, P. R., Lunde, A., and Shephard, N. (2008), "Designing Realised Kernels to Measure the Ex-Post Variation of Equity Prices in the Presence of Noise," *Econometrica*, 76, 1481–1536. [1378]

Belomestny, D., and Spokoiny, V. (2007), "Spatial Aggregation of Local Likelihood Estimates With Applications to Classification," *The Annals of Statistics*, 35, 2287–2311. [1380,1381,1391]

Cai, Z., Fan, J., and Li, R. (2000), "Efficient Estimation and Inferences for Varying Coefficient Models," *Journal of the American Statistical Association*, 95, 888–902. [1379]

Chen, J., and Gupta, A. (1997), "Testing and Locating Variance Changepoints With Application to Stock Prices," *Journal of the American Statistical Association*, 92, 739–747. [1377]

Chen, Y., and Spokoiny, V. (2010), "Modeling and Estimation for Nonstationary Time Series With Applications to Robust Risk Management," manuscript, C.A.S.E.—Center for Applied Statistics and Economics. [1380]

Čížek, P., Härdle, W., and Spokoiny, V. (2009), "Statistical Inference for Time-Inhomogeneous Volatility Models," *Econometrics Journal*, 12, 248–271. [1377,1381]

Corsi, F. (2009), "A Simple Approximate Long-Memory Model of Realized Volatility," *Journal of Financial Econometrics*, 7, 174–196. [1376,1387]

Corsi, F., Mittnik, S. M., Pigorsch, C., and Pigorsch, U. (2008), "The Volatility of Realized Volatility," *Econometric Reviews*, 27, 46–78. [1377]

da Rosa, J., Veiga, A., and Medeiros, M. C. (2008), "Tree-Structured Smooth Transition Regression Models," *Computational Statistics and Data Analysis*, 52, 2469–2488. [1388]

Diebold, F. X. (1986), Comment on "Modeling the Persistence of Conditional Variance," by R. F. Engle and T. Bollerslew, *Econometric Reviews*, 5, 51–56. [1376]

Diebold, F. X., and Inoue, A. (2001), "Long Memory and Regime Switching," *Journal of Econometrics*, 105, 131–159. [1376]

Diebold, F., and Mariano, R. (1995), "Comparing Predictive Accuracy," *Journal of Business & Economic Statistics*, 13, 253–263. [1391]

Doornik, J. A., and Ooms, M. (2004), "Inference and Forecasting for ARFIMA Models, With an Application to US and UK Inflation," *Studies in Nonlinear Dynamics and Econometrics*, 8, Article 14. [1389]

——— (2006), "A Package for Estimating, Forecasting and Simulating ARFIMA Models: ARFIMA," Ox package 1.04, available at *http://www.doornik.com/download.html*. [1389]

Ghysels, E., Rubia, A., and Valkanov, R. (2009), "Multi-Period Forecasts of Volatility: Direct, Iterated, and Mixed-Data Approaches," working paper, University of North Carolina, Chapel Hill. [1390]

Giacomini, R., and White, H. (2006), "Tests of Conditional Predictive Ability," *Econometrica*, 74, 1545–1578. [1391]

Granger, C. W. J. (1980), "Long Memory Relationships and the Aggregation of Dynamic Models," *Journal of Econometrics*, 14, 227–238. [1376]

Granger, C. W. J., and Hyung, N. (2004), "Occasional Structural Breaks and Long Memory With an Application to the S&P500 Absolute Stock Returns," *Journal of Empirical Finance*, 11, 399–421. [1376]

Granger, C. W., and Joyeux, R. (1980), "An Introduction to Long Memory Time Series Models and Fractional Differencing," *Journal of Time Series Analysis*, 1, 5–39. [1376]

Hamilton, J. D., and Susmel, R. (1994), "Autoregressive Conditional Heteroskedasticity and Changes in Regime," *Journal of Econometrics*, 64, 307–333. [1377]

Hansen, P. R. (2005), "A Test for Superior Predictive Ability," *Journal of Business & Economic Statistics*, 23, 365–380. [1391]

Hansen, P. R., Lunde, A., and Nason, J. M. (2010), "Model Confidence Sets for Forecasting Models," working paper, Federal Reserve Bank of Atlanta, available at *http://ssrn.com/abstract=522382*. [1391]

Härdle, W., Müller, M., Sperlich, S., and Werwatz, A. (2004), *Nonparametric and Semiparametric Models*, Berlin/Heidelberg/New York: Springer-Verlag. [1380]

Harvey, D., Leybourne, S., and Newbold, P. (1997), "Testing the Equality of Prediction Mean Squared Errors," *International Journal of Forecasting*, 13, 281–291. [1391]

Hasbrouck, J. (2007), *Empirical Market Microstructure*, New York: Oxford University Press. [1378]

Hillebrand, E., and Medeiros, M. C. (2008), "Asymmetries, Breaks, and Long-Range Dependence: An Estimation Framework for Time Series of Daily Realized Volatility," discussion paper, Pontifical Catholic University of Rio de Janeiro. [1377]

Hosking, J. R. M. (1981), "Fractional Differencing," *Biometrika*, 68, 165–176. [1376]

Koopman, S. J., Jungbacker, B., and Hol, E. (2005), "Forecasting Daily Variability of the S&P100 Stock Index Using Historical, Realised and Implied Volatility Measurements," *Journal of Empirical Finance*, 12, 445–475. [1387]

Lamoureux, C. G., and Lastrapes, W. D. (1990), "Persistence in Variance, Structural Change and the GARCH Model," *Journal of Business & Economic Statistics*, 8, 225–234. [1376]

Lanne, M. (2006), "A Mixture Multiplicative Error Model for Realized Volatility," *Journal of Financial Econometrics*, 4, 594–616. [1377]

Liebermann, O., and Phillips, P. C. B. (2008), "Refined Inference on Long Memory in Realized Volatility," *Econometric Reviews*, 27, 254–267. [1376]

Liu, C., and Maheu, J. M. (2008), "Are There Structural Breaks in Realized Volatility?" *Journal of Financial Econometrics*, 6, 326–360. [1377]

Marcellino, M., Stock, J. H., and Watson, M. (2005), "A Comparison of Direct and Iterated Multistep AR Methods for Forecasting Macroeconomic Time Series," *Journal of Econometrics*, 135, 499–526. [1390]

Martens, M., Dijk, D. v., and de Pooter, M. (2009), "Modeling and Forecasting S&P 500 Volatility: Long Memory, Level Shifts, Leverage Effects, Day-of-the-Week Seasonality, and Macroeconomic Announcements," *International Journal of Forecasting*, 25, 282–303. [1377]

McAleer, M., and Medeiros, M. (2008a), "Realized Volatility: A Review," *Econometric Reviews*, 27 (1), 10–45. [1378]

——— (2008b), "A Multiple Smooth Transition Heterogeneous Autoregressive Model for Long Memory and Asymmetries," *Journal of Econometrics*, 147, 104–119. [1377]

Mikosch, T., and Stărică, C. (2004a), "Changes of Structure in Financial Time Series and the GARCH Model," *REVSTAT Statistical Journal*, 2, 41–73. [1377,1380]

——— (2004b), "Non-Stationarities in Financial Time Series, the Long Range Dependence and the IGARCH Effects," *Review of Economics and Statistics*, 86, 378–390. [1377]

Morana, C., and Beltratti, A. (2004), "Structural Change and Long-Range Dependence in Volatility of Exchange Rates: Either, Neither or Both?" *Journal of Empirical Finance*, 11, 629–658. [1377]

Müller, U. A., Dacorogna, M. M., Dav, R. D., Olsen, R. B., Pictet, O. V., and von Weizsäcker, J. E. (1997), "Volatilities of Different Time Resolutions—Analyzing the Dynamics of Market Components," *Journal of Empirical Finance*, 4, 213–239. [1387]

Pigorsch, C., Pigorsch, U., and Popov, I. (2010), "Volatility Estimation Based on High-Frequency Data," in *Handbook of Computational Finance*, eds. J. C. Duan, J. E. Gentle, and W. K. Härdle, Heidelberg: Springer. [1378]

Polzehl, J., and Spokoiny, V. (2006), "Propagation-Separation Approach for Local Likelihood Estimation," *Probability Theory and Related Fields*, 135, 335–362. [1380]

Pong, S., Shackleton, M. B., Taylor, S. J., and Xu, X. (2004), "Forecasting Currency Volatility: A Comparison of Implied Volatilities and AR(FI)MA Models," *Journal of Banking & Finance*, 28, 2541–2563. [1376]

Scharth, M., and Medeiros, M. C. (2009), "Asymmetric Effects and Long Memory in the Volatility of Dow Jones Stocks," *International Journal of Forecasting*, 25, 304–327. [1377,1388-1390]

So, M. K. P., Lam, K., and Li, W. K. (1998), "A Stochastic Volatility Model With Markov Switching," *Journal of Business & Economic Statistics*, 16, 244–253. [1377]

Quantitative Finance
Publication details, including instructions for authors and subscription information:
http://www.tandfonline.com/loi/rquf20

Modeling default risk with support vector machines

Shiyi Chen [a] , W. K. Härdle [b] & R. A. Moro [c d]

[a] China Center for Economic Studies (CCES), Fudan University, 220 Handan Road, 200433 Shanghai, PR China

[b] Center for Applied Statistics and Economics (CASE), Humboldt-Universität zu Berlin, Spandauer Str. 1, 10178 Berlin, Germany

[c] Department of Economics and Finance, Brunel University, Uxbridge, Middlesex UB8 3PH, UK

[d] DIW econ, Mohrenstr. 58, 10117 Berlin, Germany

Version of record first published: 20 Apr 2010

To cite this article: Shiyi Chen, W. K. Härdle & R. A. Moro (2011): Modeling default risk with support vector machines, Quantitative Finance, 11:1, 135-154

To link to this article: http://dx.doi.org/10.1080/14697680903410015

Quantitative Finance, Vol. 11, No. 1, January 2011, 135–154

Modeling default risk with support vector machines

SHIYI CHEN*†, W. K. HÄRDLE‡ and R. A. MORO§¶

†China Center for Economic Studies (CCES), Fudan University, 220 Handan Road, 200433 Shanghai, PR China
‡Center for Applied Statistics and Economics (CASE), Humboldt-Universität zu Berlin, Spandauer Str. 1, 10178 Berlin, Germany
§Department of Economics and Finance, Brunel University, Uxbridge, Middlesex UB8 3PH, UK
¶DIW econ, Mohrenstr. 58, 10117 Berlin, Germany

(*Received 10 January 2007; in final form 12 January 2009*)

Predicting default risk is important for firms and banks to operate successfully. There are many reasons to use nonlinear techniques for predicting bankruptcy from financial ratios. Here we propose the so-called Support Vector Machine (SVM) to predict the default risk of German firms. Our analysis is based on the Creditreform database. In all tests performed in this paper the nonlinear model classified by SVM exceeds the benchmark logit model, based on the same predictors, in terms of the performance metric, AR. The empirical evidence is in favor of the SVM for classification, especially in the linear non-separable case. The sensitivity investigation and a corresponding visualization tool reveal that the classifying ability of SVM appears to be superior over a wide range of SVM parameters. In terms of the empirical results obtained by SVM, the eight most important predictors related to bankruptcy for these German firms belong to the ratios of activity, profitability, liquidity, leverage and the percentage of incremental inventories. Some of the financial ratios selected by the SVM model are new because they have a strong nonlinear dependence on the default risk but a weak linear dependence that therefore cannot be captured by the usual linear models such as the DA and logit models.

Keywords: Statistical learning theory; Applications to default risk; Capital asset pricing; Economics of risk

1. Introduction

Predicting default probabilities and deducing the corresponding risk classification is becoming more and more important in order for firms to operate successfully and for banks to clearly grasp their clients' specific risk class. In particular, the implementation of the Basel II capital accord will further exert pressure on firms and banks. As both the risk premium and the credit costs are determined by the default risk, the firms' ratings will have a deeper economic impact on banks as well as on the firms themselves than ever before. Thus, from a risk management perspective, the choice of a correct rating model that can capture consistent predictive information concerning the probabilities of default over some successive time periods is of crucial importance.

There are strands of the literature that deal with the statistical and stochastic analysis of default risk (Burnham and Anderson 1998, Caouette *et al*. 1998, Shumway 1998, Sobehart *et al*. 2000, Saunders and Allen 2002, Gaeta 2003, Chakrabarti and Varadachari 2004, Giesecke 2004, Zagst and Hocht 2006). One models default events using accounting data, whereas other models recommend using market information. Market-based models can be further classified into structural models and reduced form models. There is also a hybrid approach that uses accounting data as well as market information to predict the probability of default. The market-based approach relies on the time series of company market data. Unfortunately, time series long enough to reliably estimate the risk is not available for most companies. Moreover, the majority of German firms are not listed and, therefore, their market price is unknown. This justifies the choice of a model for which only cross-sectional or pooled accounting data would be required. For this study, accounting data for bankrupt and operating German companies was provided by Creditreform.

Among the accounting-based models, the first attempts to identify the difference between the financial ratios of

*Corresponding author. Email: shiyichen@fudan.edu.cn

Quantitative Finance
ISSN 1469–7688 print/ISSN 1469–7696 online © 2011 Taylor & Francis
http://www.informaworld.com
DOI: 10.1080/14697680903410015

solvent and insolvent firms were the studies of Ramser and Foster (1931), Fitzpatrick (1932), Winakor and Smith (1935) and Merwin (1942). These studies settled the fundamentals for bankruptcy prediction research. It was not until the 1960s that the traditional research was changed. Beaver (1966) pioneeringly presented the univariate approach to discriminant analysis (DA) for bankruptcy prediction. Altman (1968) expanded this analysis to multivariate analysis. Up to the 1980s, DA was the dominant method in bankruptcy prediction. However, there are obvious modeling restrictions of this approach, some of which are the assumptions of normality, homoscedasticity of the disturbances, fulfillment of conditional expectation of the dependent variable between 0 and 1, and no adjustment for multicollinearity. During the 1980s the DA method was replaced by logistic analysis, which fits the logistic regression model for binary or ordinal response data by the method of maximum likelihood estimation (MLE). In fact, the logit model uses the logistic cumulative distribution function in modeling the default probability. Among the first users of logit analysis in the context of bankruptcy were Ohlson (1980), Collins and Green (1982), Lo (1986) and Platt *et al*. (1994). The advantage of the logit model is that it does not assume multivariate normality and equal variance disturbance, and its probability lies between 0 and 1 (Härdle and Simar 2003). However, the logit model is also sensitive to the collinearity among the variables. In addition, the key assumption behind the logit model is that the logarithm of odds is linear in the underlying random variable; therefore, common to DA and logit modeling is a linear classifying hyperplane that separates insolvent and solvent firms. This works well if the data are linearly separable. A linear separating hyperplane is, however, not suitable if there is doubt that the separation mechanism is of a nonlinear kind. There are good reasons to take the linear non-separability case seriously (Falkenstein *et al*. 2000).

Many nonlinear numerical methodologies have been developed to solve the linear non-separability problem: Maximum Expected Utility (MEU), Artificial Neural Networks (ANN) and Support Vector Machines (SVM). The MEU model was proposed at Standard & Poor's Risk Solutions Group, which allows models to incorporate the nonlinearity, non-monotonicity, and interactions present in the data, reducing the risk of overfitting. Friedman and Sandow (2003a, b) and Friedman and Huang (2003) demonstrated how the MEU method outperforms the Logit model. ANN was introduced to analyse bankrupt firms in the 1990s (see Hertz *et al*. (1991), Refenes (1995) and Härdle *et al*. (2004) for more details). This method also discards the assumption of linearity and mutual independence of explanatory variables for the default prediction function (Serrano *et al*. 1993, Back *et al*. 1994, 1996, Wilson and Sharda 1994). ANN models built using *K*-fold cross-validation techniques can be very robust and reduce over-fitting. Although the nonlinear ANN can classify a dataset much better than the linear models, it has often been criticized to be vulnerable to the multiple minima problem. Common to the OLS and MLE for linear models, ANN also makes use of the principle of minimizing empirical risk, which usually leads to a poor level of classification for out-of-sample data (Haykin 1999).

Based on statistical learning theory, an alternative nonlinear separation method, the Support Vector Machine (SVM), was recently introduced in default risk analysis. The SVM yields a single minimum without undesirable local fits as often produced by ANN. This property results from the minimized target function that is convex quadratic and linearly restricted. In addition, the SVM is also able to handle the interactions between the ratios and does not need any parameter restrictions and prior assumptions such as that concerning the distribution for latent errors. Furthermore, the biggest advantage of SVM among all the alternatives is its ability to minimize the risk associated with model misspecification, which endows SVM with an excellent separating ability. The current literature in statistical learning theory has produced strong evidence that SVM systematically outperforms standard pattern recognition/classification, function regression and data analysis techniques (Vapnik 1995, Haykin 1999). The application of SVM to company default analysis is less reported in the management science and finance literature. Härdle *et al*. (2005, 2007) report that, compared with the traditional DA and logit models in predicting the probabilities of default and rating firms, the SVM has a superior performance. Gestel *et al*. (2005) combined SVM and the logistic regression model to capture the multivariate nonlinear relations. This combination technique balances the interpretability and predictability required to rating banks.

In this study, we investigate the applicability of this new technique to predicting the risk scores and the probabilities of defaults (PDs) of German firms from the Creditreform database spanning from 1996 through 2002. The aim is to investigate (1) which of the accounting ratios are meaningful and have predictive character for bankruptcy, and (2) does a well-specified SVM-based nonlinear model consistently outperform the benchmark logit model in predicting PDs as predicted by theory?

The rest of the paper is organized as follows. In the next section we give a theoretical introduction to the Support Vector Machine (SVM) for classification. Section 3 describes the Creditreform database and the variables and ratios used in this study. In section 4, we present the validation procedures, re-sampling technique, performance measures and the ratios selection methods. Section 5 analyses the empirical results, including the predictors related to bankruptcy, the sensitivity analysis of SVM parameters, and a comparison of the predictive performance between SVM and the logit model. Section 7 offers conclusions.

2. The Support Vector Machine

The term Support Vector Machine (SVM) originates from Vapnik's statistical learning theory (Vapnik 1995, 1997), which formulates the classification problem as a quadratic

programming (QP) problem. The principles on which the SVM is based, especially the regularization principle for solving ill-posed problems, are also described by Tikhonov (1963), Tikhonov and Arsenin (1977) and Vapnik (1979). The SVM transforms by nonlinear mapping the input space (of covariates) into a high-dimensional feature space and then solves a linear separable classification problem in this feature space. Thus, linear separable classification in the feature space corresponds to linearly non-separable classification in the lower-dimensional input space. As the name implies, the design of the SVM hinges on the extraction of a subset of the training data that serves as support vectors and that represents a stable characteristic of the data.

Given a training data set $\{x_i, y_i\}_{i=1}^n \{(\mathbf{x}_i, y_i)\}_{i=1}^n$, with input vector $\mathbf{x}_i \in R^d$ (company financial ratios in this study) $x_i \in R^d$ and output scalar $y_i \in \{+1, -1\}$ $y_i = \{+1, -1\} \in R^1$ (-1 = 'successful', $+1$ = 'bankrupt'), we aim to find a classifying (score) function $f(\mathbf{x})$ to approximate the latent, unknown decision function $g(\mathbf{x})$. In the logistic and the DA case, this is simply a linear function. In the SVM case, the classifying function is

$$f(\mathbf{x}) = \sum_{l=1}^{l} w_l \phi_l(\mathbf{x}) + b = \mathbf{w}^T \phi(\mathbf{x}) + b, \qquad (1)$$

where $\phi(\mathbf{x}) = [\phi_1(\mathbf{x}), \ldots, \phi_l(\mathbf{x})]^T$ and $\mathbf{w} = [w_1, \ldots, w_l]^T$.

The nonlinear functions $\phi(\mathbf{x})$ are the transformation functions from the input space to the feature space that represent the features of the input space. A simple example of features for a quadratic function in a two-dimensional space is $\phi_1 = x_1^2$, $\phi_2 = \sqrt{2}x_1x_2$ and $\phi_3 = x_2^2$. The dimension of the feature space is l, which is directly related to the capacity of the SVM to approximate a smooth input–output mapping; the higher the dimension of the feature space, the more accurate, at the cost of variability, the approximation will be. Parameter $\mathbf{w}$ denotes a set of linear weights connecting the feature space to the output space, and b is the bias or threshold. The optimal solution $\mathbf{w}^*$ and b^* can be used to construct the optimal hyperplane $\mathbf{w}^{*T}\phi(\mathbf{x}) + b^* = 0$ and the classification function $f(\mathbf{x}) = \mathbf{w}^{*T}\phi(\mathbf{x}) + b^*$. We can predict solvent and insolvent companies using the estimated function $f(\mathbf{x})$.

2.1. Advantage of SVM for classification in theory

The main superiority of nonlinear non-parametric SVM over the benchmarking methods in predicting company credit risk results from its special theoretical device in two ways: (1) it takes linearly non-separable situations into account, whereas the DA and logit models only work well if the data are linear separable; and (2) it adopts the principle of structural risk minimization rather than empirical risk minimization employed by the OLS, MLE, ANN (and other) models. We illustrate the principle in figure 1 using the simplest classifying function $f(\mathbf{x}) = -x_1 - 2x_2 + 2$, where $\mathbf{x} = (x_1, x_2)^T$, $\mathbf{w} = (-1, -2)$ and $b = 2$.

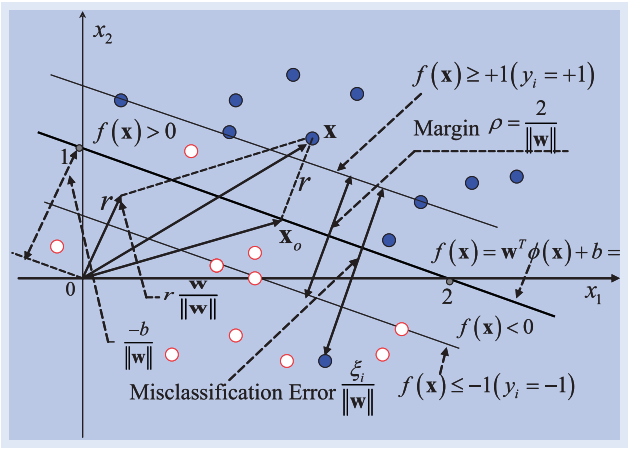


Figure 1. Separation margin, misclassification error and structural risk minimization for the SVM in two-dimensional input space.

The statistical problem is how to construct a classifying hyperplane (hypersurface) and obtain the classifying function $f(\mathbf{x})$. If the data set is linearly separable, the perfect classification hyperplane does exist. The function $f(\mathbf{x})$ gives an algebraic measure of the distance from $\mathbf{x}$ to the optimal hyperplane. Perhaps the easiest way to see this is to express $\mathbf{x}$ as $\mathbf{x} = \mathbf{x}_0 + r(\mathbf{w}/||\mathbf{w}||)$, where $\mathbf{x_0}$ is the normal projection of $\mathbf{x}$ onto the optimal hyperplane, r is the desired algebra distance from any point $\mathbf{x}$ to the optimal hyperplane (positive if $\mathbf{x}$ is on the positive side of the optimal hyperplane and negative otherwise), and $||\mathbf{w}||$ is the Euclidean norm of the weight vector $\mathbf{w}$. Since, by definition, $f(\mathbf{x}_0) = 0$, it follows that

$$f(\mathbf{x}) = \mathbf{w}^T\mathbf{x}_0 + \mathbf{w}^T r\frac{\mathbf{w}}{||\mathbf{w}||} + b = f(\mathbf{x}_0) + r||\mathbf{w}|| = r||\mathbf{w}||$$

or

$$r = \frac{f(\mathbf{x})}{||\mathbf{w}||}.$$

Because of the values of y_i being ± 1, the parameters $(\mathbf{w}, b)$ for the optimal hyperplane must satisfy the constraints $f(\mathbf{x}) \geq 1$ for $y_i = +1$ (insolvent) or $f(\mathbf{x}) \leq -1$ for $y_i = -1$ (solvent), that is $y_i \cdot f(\mathbf{x}) \geq 1$. The particular data points for which the constraint is satisfied with the equality sign are called *support vectors*, hence the name 'Support Vector Machine'. In conceptual terms, the support vectors are those data points that lie closest to the decision surface and are therefore the most difficult to classify. As such, they have a direct bearing on the optimum location of the classification hyperplane and play a prominent role in the operation of SVM. Now consider the support vectors; they are located on the upper and lower separation band for which $f(\mathbf{x}) = \pm 1$. Therefore, the algebraic distance from the support vectors to the optimal hyperplane is

$$r = \frac{f(\mathbf{x})}{||\mathbf{w}||} = \frac{\pm 1}{||\mathbf{w}||}.$$

Let ρ denote the optimum value of the margin of separation between solvent and insolvent companies.

Then it follows that $\rho = 2r = 2/||\mathbf{w}||$, which states that maximizing the margin of separation between classes is equivalent to minimizing the Euclidean norm of $\mathbf{w}$, $\|\mathbf{w}\|$. Thus, the classifying function in the linear separable case can be derived from maximizing the separation margin directly. Likewise, the distance from the origin to the optimal hyperplane is given by $-b/||\mathbf{w}||$, as shown in figure 1.

If the training set is linearly non-separable, the hyperplane that can correctly classify the training set no longer exists and, naturally, we need to find a hypersurface instead. For the hypersurface, however, we know less about the concept of the geometrical margin that is particular for the hyperplane; therefore, it is more difficult to find a hypersurface than a hyperplane. The transformation from the input space into higher-dimensional feature space, i.e. $\mathbf{x} \mapsto \phi(\mathbf{x})$, is then introduced in the SVM. It is possible that the new training set in the feature space $\{\phi(\mathbf{x}_i), y_i\}_{i=1}^n$ becomes linearly separable. Accordingly, the problem of finding a hypersurface in the input space is transformed into finding a hyperplane in the feature space and letting its margin or the 'safe' distance between classes, where in the perfectly separable case no observation can lie, be maximized.

It is not possible to construct a separating hyperplane without encountering classification errors. The margin of separation between classes is said to be soft if a data point violates the condition $y_i \cdot f(\mathbf{x}) \geq 1$. This violation can arise in one of two ways: (1) the data point falls inside the region of separation but on the right side of the decision surface; and (2) the data points falls on the wrong side of the decision surface. Note that we have correct classification in case (1), but misclassification in case (2). Therefore, a new set of non-negative slack variables $\{\xi_i\}_{i=1}^n$ are introduced and the condition is softened to $y_i \cdot f(\mathbf{x}) \geq 1 - \xi_i$. Note $0 < \xi_i \leq 1$ for case (1), $\xi_i \geq 1$ for case (2), and $\xi_i = 0$ for the linearly separable case. The support vectors are those particular data points that satisfy the soft condition precisely even if $\xi_i > 0$. The support vectors are thus defined in exactly the same way for both linearly separable and non-separable cases. In fact, using the soft constraints and the condition $\xi_i \geq 0$, the slack variables ξ_i can be represented as a hinge loss function which is the tightest convex upper bound of the misclassification loss and special and preferred to the loss function of the logit model because it allows a sparse solution, in the sense that some observations of the training set, if they are classified correctly, may not be necessary to construct the separating boundary. Sparseness of the solution also greatly simplifies the computation of SVM because then usually only few observations, so-called support vectors, are required to restore the solution, while for the logit regression, all observations are necessary.

The algebraic distance from the misclassification point to the optimal hyperplane is $r = [(1 - \xi_i)/\|\mathbf{w}\|]$, which can be derived making use of the same algebraic manipulation as in the linear separable case. Thus, the distance between the misclassification point and the upper band, the case in figure 1, is $\xi_i/\|\mathbf{w}\|$ and the tolerance to misclassification errors on the training set can be measured by $\sum_{i=1}^n \xi_i/||\mathbf{w}||$. Our goal is to find a separating hyperplane for which the misclassification error, averaged on the training set, is minimized, which is similar to minimize the sum of residual squares, the empirical risk in OLS and MLE estimation.

Thus, two targets exist for SVM in the linear non-separable case: still maximize the separation margin $2/||\mathbf{w}||$ and simultaneously minimize the misclassification distance $\sum_{i=1}^n \xi_i/||\mathbf{w}||$. The most intuitive form of the objective function to be minimized is

$$\min_{\mathbf{w}, b, \xi} \frac{1}{2}||\mathbf{w}|| + C \sum_{i=1}^{n} \frac{\xi_i}{||\mathbf{w}||}. \tag{2}$$

As shown above, the second term is the margin-based loss function, which is the sum of errors measured as the distance from a misclassified observation to the hyperplane boundary, its class weighted with the parameter C. Equation (2) exhibits the so-called structural risk minimizing principle held by the SVM method. The benchmark models such as the DA and logit estimated by OLS and MLE, and simple ANN-based nonlinear models with no constraints usually employ the principle of minimizing error functions calculated on the training sample. Therefore, SVM not only minimizes the traditional empirical risk, but also maximizes the separating margin, and finally obtains a trade-off between two targets. It is this kind of special design of minimizing the structural risk that endows SVM with stronger classifying ability than the benchmark methods.

2.2. *SVM algorithm*

To minimize the cost function (2), an equivalent quadratic cost function, $(1/2)||\mathbf{w}||^2 + C\sum_{i=1}^n \xi_i$, can be obtained from equation (2) multiplied by $||\mathbf{w}||$ ($||\mathbf{w}|| > 0$). Thus, the primary problem of the SVM for the non-separable case is expressed as

$$\min_{\mathbf{w}, b, \xi} \frac{1}{2}||\mathbf{w}||^2 + C \sum_{i=1}^{n} \xi_i, \tag{3}$$

s.t.

$$y_i \times \{\mathbf{w}^T \phi(\mathbf{x}_i) + b\} + \xi_i \geq 1, \tag{4}$$

$$\xi_i \geq 0, \quad i = 1, 2, \ldots, n. \tag{5}$$

As before, minimizing the first term of equation (3) is equivalent to maximizing the separation margin. The scaling factor 1/2 is included here for convenience of presentation. As for the second term, it is an upper bound on the number of misclassification errors. The formulation of the cost function in equation (3) is also therefore in perfect accord with the principle of structural risk minimization. The penalty parameter $C > 0$ is introduced to integrate the weights of two targets. It controls the trade-off between the complexity of the machine and the number of non-separable points; that is, the penalty parameter C controls the extent of penalization

(or the tolerance) to misclassification errors on the training set. Partially the optimization function is derived from the problem of separating the population of defaulters from non-defaulters. However, it contains a second part responsible for margin maximization that is introduced artificially. Although it introduces a bias to the original optimization problem, it reduces the complexity of the SVM and increases its accuracy on out-of-sample data. The value of parameter C has to be selected by the user (Haykin 1999). The optimization problem for non-separable patterns stated above includes the optimization problem for linearly separable patterns as a special case. Specifically, setting $\xi_i = 0$ for all i in both equations (3) and (4) reduces them to the corresponding forms for the linearly separable case.

The corresponding dual problem of SVM for non-separable patterns can be derived using the Karush–Kuhn–Tucker conditions (Fletcher 1987, Bertsekas 1995) as follows:

$$\min_{\alpha} \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n} y_i y_j \alpha_i \alpha_j K(\mathbf{x}_i, \mathbf{x}_j) - \sum_{i=1}^{n} \alpha_j, \tag{6}$$

s.t

$$\sum_{i=1}^{n} y_i \alpha_i = 0, \tag{7}$$

$$0 \le \alpha_i \le C, \quad i = 1, 2, \ldots, n, \tag{8}$$

where α_i and α_j are Lagrange multipliers. Note that neither the slack variables ξ_i nor their Lagrange multipliers appear in the dual problem. Thus, the objective function (6) to be minimized is the same in both the linear separable and non-separable cases. Deng and Tian (2004) demonstrate that the dual problem is easier to solve than the primal problem. We can then use the optimal solution α_i^* to obtain the solution of the primal problem:

$$\mathbf{w}^* = \sum_{i=1}^{n} y_i \alpha_i^* \phi(\mathbf{x}_i), \tag{9}$$

$$b^* = y_j - \sum_{i=1}^{n} y_i \alpha_i^* K(\mathbf{x}_i, \mathbf{x}_j), \quad \forall j \in \{j \,|\, 0 < \alpha_j^* < C\}. \tag{10}$$

By substitution, the nonlinear classifying (score) function can be obtained:

$$\begin{aligned} f(\mathbf{x}_j) &= \mathbf{w}^{*T}\phi(\mathbf{x}_j) + b^* = \sum_{i=1}^{n} y_i \alpha_i^* \phi^T(\mathbf{x}_i)\phi(\mathbf{x}_j) + b^* \\ &= \sum_{i=1}^{n} y_i \alpha_i^* K(\mathbf{x}_i, \mathbf{x}_j) + b^*, \end{aligned} \tag{11}$$

where $K(\mathbf{x}_i, \mathbf{x}_j) = \phi^T(\mathbf{x}_i)\phi(\mathbf{x}_j)$ is the inner product kernel function in which $\mathbf{x}_i$ belongs to the training set and $\mathbf{x}_j$ is the new company financial ratio, either in the training set or validating and forecasting set. For the classification problem, the decision function (11) is constructed to help us deduce in what kind of category, say $+1$ or -1, the new output $f(\mathbf{x}_j)$ corresponding to $\mathbf{x}_j$ is located. To the end, the intuitive way is to compare $\mathbf{x}_j$ with $\mathbf{x}_i$ pairwise; if $\mathbf{x}_j$ is closer to $\mathbf{x}_i$ on the positive side, then the new output $f(\mathbf{x}_j)$ nears $+1$, if $\mathbf{x}_j$ is closer to $\mathbf{x}_i$ on the negative side $f(\mathbf{x}_j)$ falls into the category -1. This is reasonable because a similar input should lead to the same output. Therefore, the decision function only depends on the proximity between two observations and the classification is in fact a proximity problem. In SVM, the inner product kernel function $K(\mathbf{x}_i, \mathbf{x}_j)$ is the key tool to measure this kind of proximity. In addition, the SVM theory considers the form of $K(\mathbf{x}_i, \mathbf{x}_j)$ in the Hilbert space without specifying $\phi(\cdot)$ explicitly and without computing all corresponding inner products, which provides the flexibility of the high-dimensional Hilbert space for low computational costs and greatly reduces the computational complexity. Thus, the kernel becomes the crucial part of SVM.

It is necessary to find an appropriate kernel in order to solve the optimization problem of SVM. The requirement on the kernel function is to satisfy Mercer's theorem (Mercer 1908, Courant and Hilbert 1970), such that the Kernel matrix, $\{K(\mathbf{x}_i, \mathbf{x}_j)\}_{i,j=1}^{n}$, is symmetric and semi-positive definite. Mercer's theorem tells us whether or not a candidate kernel is actually an inner-product kernel in some space and therefore admissible for use in a support vector machine. Within this requirement there is some freedom in how it is chosen. The usual chosen kernels are linear, polynomial and Gaussian kernel functions. A different kernel requires estimating the extent of proximity based on a different metric criterion. In this study, we choose an anisotropic Gaussian kernel for the SVM:

$$K(\mathbf{x}_i, \mathbf{x}_j) = \exp(-(\mathbf{x}_i - \mathbf{x}_j)^T r^{-2} \Sigma^{-1} (\mathbf{x}_i - \mathbf{x}_j)/2), \tag{12}$$

where Σ is the variance–covariance matrix of the data and r is the Gaussian, also known as the radial basis kernel coefficient which implicitly controls the complexity of the feature space and the solution—the larger r, the less the complexity. Therefore, based on expression (11), for any new company $\mathbf{x}_j$, those companies from the training sample $\mathbf{x}_i$ will have a greater impact on $f(\mathbf{x}_j)$ if $\mathbf{x}_j$ are closer to $\mathbf{x}_i$. The anisotropic Gaussian kernel offers a way of measuring the proximity between two companies; it is higher when the companies are close and smaller when they are far from each other.

3. Data and financial ratios

3.1. Data description

The data used in this study is the Creditreform database. It contains a random sample of 20,000 solvent and 1000 insolvent firms in Germany and spans the period from 1996 to 2002, although the data are concentrated in 2001 and 2002 with approximately 50% of the observations coming from this period. Most firms appear in the database several times in different years. Each firm is described by a set of financial statement variables such as those in balance sheets and

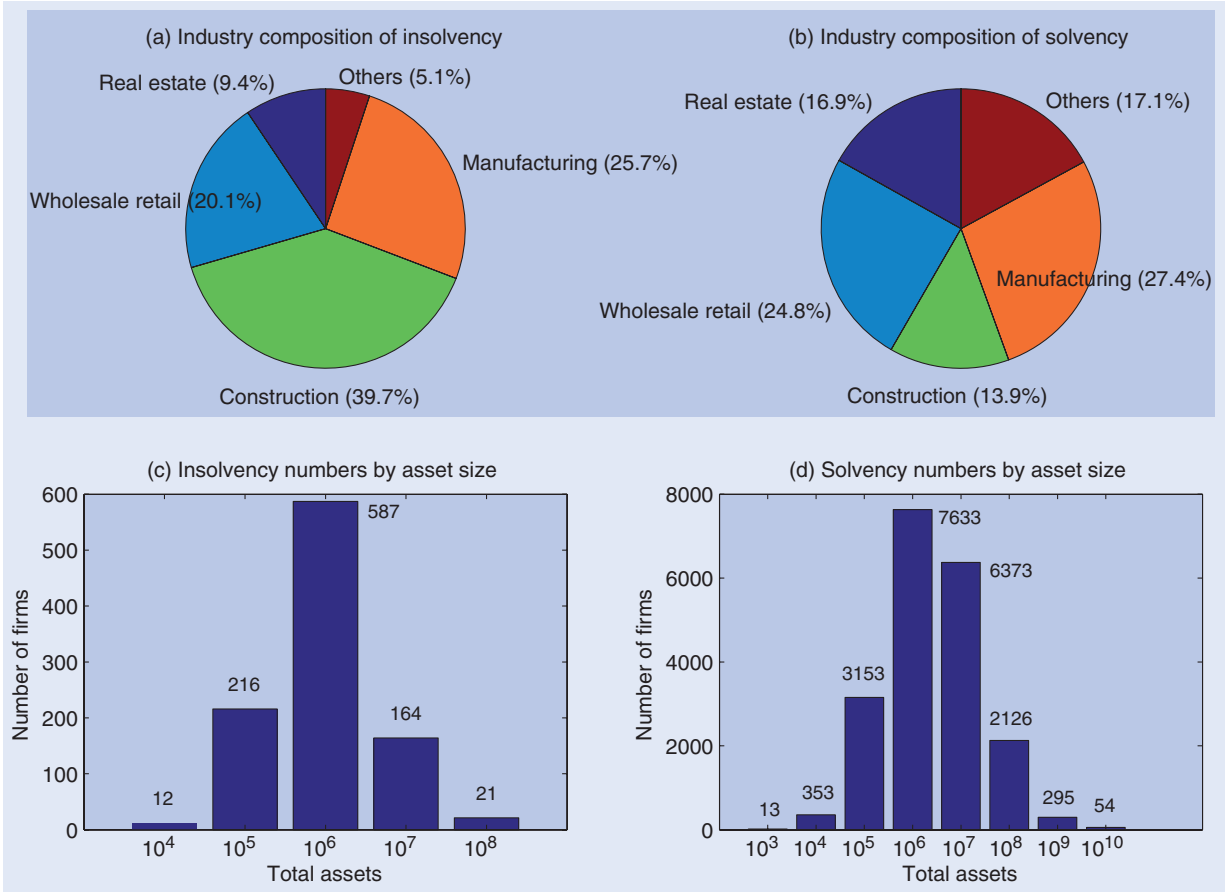


Figure 2. Industry composition and size distribution of the companies in the Creditreform database.

income statements. The data for the insolvent firms were collected two years prior to insolvency.

Figure 2 shows the industry composition and size distribution of the database. The industries to which each firm belongs can be systematically classified according to an internationally recognized system—Classification of Economic Activities, Edition 1993 (WZ 93)—published by the German Federal Statistical Office. WZ 93 uses a hierarchy of five different levels. The higher the level, the more precise the description of the main activity. In terms of the classification industry codes of WZ 93, as shown in figures 2(a) and (b), the 1000 insolvent firms consist of about 39.7% construction, 25.7% manufacturing, 20.1% the wholesale and retail trade, 9.4% real estate and 5.1% others. The others among the 1000 insolvent firms include agriculture, mining, electricity, gas and water supply, hotels and restaurants, transport and communication, financial intermediation and social service activities. The industries of the 20,000 solvent firms are manufacturing (27.4%), wholesale and retail trade (24.8%), real estate (16.9%), construction (13.9%) and others (17.1%). Different from the 'others' of insolvent firms, the others in solvency contain additional industries such as publishing, administration and defense, education and health.

The distribution of total assets can be regarded as being representative of the distribution of the firm size. In figures 2(c) and (d), the 1000 insolvent sample comprises 12 firms located in the size category 10^4 EUR, 216 in 10^5 EUR, 587 in 10^6 EUR, 164 in 10^7 EUR and 21 in 10^8 EUR. (Here, 10^4 EUR represents one category of asset size in which the firms have total assets of between 10,000 and 99,999 EUR. The definition of the other size categories is similar to that for 10^4 EUR.) The number of firms corresponding to each asset size category of the 20,000 solvent firms is 13 (10^3 EUR and below), 353 (10^4 EUR), 3153 (10^5 EUR), 7633 (10^6 EUR), 6373 (10^7 EUR), 2126 (10^8 EUR), 295 (10^9 EUR) and 54 (10^{10} EUR and above).

In an attempt to obtain a more homogeneous company sample, we cleaned the database of companies whose characteristics are very different from the others. That is to say, we do not attempt to cover all firms in the database for our study because of the very different nature of some firms. Thus, in focusing on predicting the PDs of German firms we eliminated the following types of firms from the whole sample.

- *Firms with a small percentage composition of industry*—that is, we eliminate the firms that belong to the 'other' industries in the insolvent and solvent databases, for example financial intermediation and public institutions. Thus only four main types of industry (Construction, Manufacturing, Wholesale & Retail Trade and Real Estate) remain in the study.
- *Smallest and largest firms*—that is, we exclude those firms that, because of their asset size,

Table 1. Variables used in the study.

Abbreviation	Variable	Abbreviation	Variable
CASH	Cash and cash equivalents	DEBT	Debt
INV	Inventories	AP	Accounts payable
CA	Current assets	SALE	Sales
ITGA	Intangible assets	AD	Amortization and depreciation
TA	Total assets	INTE	Interest expense
QA	Quick assets (=CA-INV)	EBIT	Earnings before interest and tax
AR	Accounts receivable	OI	Operating income
LB	Lands and buildings	NI	Net income
OF	Own funds	IDINV	Increase (decrease) inventories
CL	Current liabilities	IDL	Increase (decrease) liabilities
TL	Total liabilities	IDCASH	Increase (decrease) cash
WC	Working capital (=CA-CL)		

are not located in the categories 10^5, 10^6 and 10^7 EUR. As Khandani *et al*. (2001) noted, the credit quality of the smallest firms is often as dependent on the finances of a key individual as on the firm itself; the number of largest firms that go bankrupt is usually very small in Germany.

We further clean the database to ensure that the value of some variables, such as the denominator when calculating the ratios, should not be zero. We also exclude the firms solvent in 1996 because of missing insolvency values for this year.

Thus, 783 insolvent firms and 9583 solvent firms were chosen and analysed. The bankrupt firms are paired with non-bankrupt firms with a similar industry and total asset size. Correspondingly, the predicted default probabilities and rating results in this study are only suitable for German firms from four main industry sectors (Construction, Manufacturing, Wholesale & Retail Trade and Real Estate) and with medium asset size (lying within the categories 10^5, 10^6, and 10^7 EUR).

3.2. Ratio definitions

The Creditreform database provides many financial statement variables for each firm. In accordance with the existing literature, 28 ratios were selected for the bankruptcy analysis. In summary, there are 28 financial ratios (including one size variable) and a binary response, which records whether the firm went bankrupt within two years of the financial statements or not. There is also information on the industry distribution and on the year of the accounts. There are no missing values. These ratios can be grouped into the following six broad categories (factors): profitability, leverage, liquidity, activity, firm size and the percentage change for some variables. The variables applied to calculate these ratios are shown in table 1. Table 2 describes these ratios and how they were calculated. For simplicity, we provide short names for some ratios that capture the essence of what they measure. Table 3 summarizes the descriptive statistics of the 28 ratios for both the insolvency and solvency sample.

In previous studies, profitability ratios have appeared to be strong predictors related to bankruptcy. In addition, among all the potential risk factors, there are more profitability ratios than any other factor. The profitability ratios employed in our study are return on assets (ROA, NI/TA), net profit margin (NI/SALE), OI/TA, operating profit margin (OI/SALE), EBIT/TA, EBITDA and EBIT/SALE, denoted respectively as x1, x2, x3, x4, x5, x6 and x7.

The ROA figure gives investors an idea of how effectively the firm is deploying its assets to generate income. The higher the ROA number, the better, because the firm is earning more money on less investment. Net profit margin measures how much of every dollar of sales a firm actually keeps in earnings. A higher profit margin indicates a more profitable firm that has better control over its costs compared with its competitors. Some investors add extraordinary items back into net income when performing this calculation because they would like to use operating returns on assets, which represent a firm's true operating performance. Operating income is also required to calculate operating profit margin, which describes a firm's operating efficiency and pricing strategy. EBIT is all profits before taking into account interest payments and income taxes. An important factor contributing to the widespread use of EBIT is the way in which it nullifies the effects of different capital structures and tax rates used by different firms. By excluding both taxes and interest expenses the figure homes in on the firm's ability to profit and thus makes for easier cross-firm comparisons. EBIT is the precursor to EBITDA, which takes the process further by removing two non-cash items from the equation (depreciation and amortization). Thus, defaulting firms usually have lower profitability values; however, firms with extremely large and volatile profitability may also be likely to translate into higher default probabilities. We will try to capture this kind of complex nonlinear dependence in our database.

Leverage is also a key measure of firm risk. In this study, seven leverage ratios are analysed. They are simple and adjusted own funds ratio, CL/TA, net indebtedness, TL/TA, debt ratio (DEBT/TA) and interest coverage ratio (EBIT/INTE), represented by x8 through x14.

The own funds ratio measures the ratio of a firm's internal capital to its assets. The simple version is widely used in credit models, which is basically the mirror image

Table 2. Definitions of accounting ratios.

Ratio No.	Definition	Ratio	Category
x1	NI/TA	Return on assets (ROA)	Profitability
x2	NI/SALE	Net profit margin	Profitability
x3	OI/TA		Profitability
x4	OI/SALE	Operating profit margin	Profitability
x5	EBIT/TA		Profitability
x6	(EBIT + AD)/TA	EBITDA	Profitability
x7	EBIT/SALE		Profitability
x8	OF/TA	Own funds ratio (simple)	Leverage
x9	(OF-ITGA)/(TA-ITGA-CASH-LB)	Own funds ratio (adjusted)	Leverage
x10	CL/TA		Leverage
x11	(CL-CASH)/TA	Net indebtedness	Leverage
x12	TL/TA		Leverage
x13	DEBT/TA	Debt ratio	Leverage
x14	EBIT/INTE	Interest coverage ratio	Leverage
x15	CASH/TA		Liquidity
x16	CASH/CL	Cash ratio	Liquidity
x17	QA/CL	Quick ratio	Liquidity
x18	CA/CL	Current ratio	Liquidity
x19	WC/TA		Liquidity
x20	CL/TL		Liquidity
x21	TA/SALE	Asset turnover	Activity
x22	INV/SALE	Inventory turnover	Activity
x23	AR/SALE	Account receivable turnover	Activity
x24	AP/SALE	Account payable turnover	Activity
x25	Log(TA)		Size
x26	IDINV/INV	Percentage of incremental inventories	Percentage
x27	IDL/TL	Percentage of incremental Liabilities	Percentage
x28	IDCASH/CASH	Percentage of incremental cash flow	Percentage

Table 3. Descriptive statistics of the 28 accounting ratios. IQR is the interquartile range.

	Insolvent				Solvent			
Ratio	q0.05	Med.	q0.95	IQR	q0.05	Med.	q0.95	IQR
NI/TA	−0.19	0.00	0.09	0.04	−0.09	0.02	0.19	0.06
NI/SALE	−0.15	0.00	0.06	0.03	−0.07	0.01	0.10	0.03
OI/TA	−0.22	0.00	0.10	0.06	−0.11	0.03	0.27	0.09
OI/SALE	−0.16	0.00	0.07	0.04	−0.08	0.02	0.13	0.04
EBIT/TA	−0.19	0.02	0.13	0.07	−0.09	0.05	0.27	0.09
EBITDA	−0.13	0.07	0.21	0.08	−0.04	0.11	0.35	0.12
EBIT/SALE	−0.14	0.01	0.10	0.04	−0.07	0.02	0.14	0.05
OF/TA	0.00	0.05	0.40	0.13	0.00	0.14	0.60	0.23
(OF-ITGA) / (TA-ITGA-CASH-LB)	−0.01	0.05	0.56	0.17	0.00	0.16	0.95	0.32
CL/TA	0.18	0.52	0.91	0.36	0.09	0.42	0.88	0.39
(CL-CASH)/TA	0.12	0.49	0.89	0.36	−0.05	0.36	0.83	0.41
TL/TA	0.29	0.76	0.98	0.35	0.16	0.65	0.96	0.40
DEBT/TA	0.00	0.21	0.61	0.29	0.00	0.15	0.59	0.31
EBIT/INTE	−7.90	1.05	7.20	2.47	−6.78	2.16	73.95	5.69
CASH/TA	0.00	0.02	0.16	0.05	0.00	0.03	0.32	0.10
CASH/CL	0.00	0.03	0.43	0.11	0.00	0.08	1.40	0.29
QA/CL	0.18	0.68	1.90	0.54	0.25	0.94	4.55	1.00
CA/CL	0.56	1.26	3.73	0.84	0.64	1.58	7.15	1.56
WC/TA	−0.32	0.15	0.63	0.36	−0.22	0.25	0.73	0.41
CL/TL	0.34	0.84	1.00	0.37	0.22	0.85	1.00	0.44
SALE/TA	0.43	1.63	4.15	1.41	0.50	2.08	6.19	1.76
INV/SALE	0.02	0.16	0.89	0.26	0.01	0.11	0.56	0.16
AR/SALE	0.02	0.12	0.33	0.11	0.00	0.09	0.25	0.09
AP/SALE	0.03	0.14	0.36	0.10	0.01	0.07	0.24	0.08
Log(TA)	13.01	14.87	17.16	1.69	12.82	15.41	17.95	2.37
IDINV/INV	−1.20	0.00	0.75	0.34	−0.81	0.00	0.56	0.07
IDL/TL	−0.44	0.00	0.48	0.15	−0.53	0.00	0.94	0.14
IDCASH/CASH	−12.71	0.00	0.94	0.79	−7.13	0.00	0.91	0.52

of TL/TA, as expected: they are mathematical complements. We have made some adjustments to the simple own funds ratio to counter creative accounting practices, and to try to generate a better measure of firm credit strength. The adjustments are also used by Khandani *et al.* (2001). Net indebtedness measures the level of short-term liabilities not covered by the firm's most liquid assets as a proportion of its total assets. Thus, in addition to measuring the short-term leverage of a firm, it also provides a measure of the liquidity of a firm. While the debt ratio performs about as well as TL/TA for public firms, it does considerably worse for private firms, which makes TL/TA preferred. The difference between debt and liabilities is that liabilities is a more inclusive term that includes debt, deferred taxes, minority interest, accounts payable, and other liabilities. The interest coverage ratio is highly predictive. Falkenstein *et al.* (2000) argue that the interest coverage ratio turns out to be one of the most valuable explanatory variables in the public firm dataset in a multivariate context, although in the private firm database its relative power decreases significantly.

Six liquidity ratios, CASH/TA, cash ratio, quick ratio, current ratio, WC/TA and CL/TA (x15 through x20), are analysed in this paper. Liquidity is a common variable in most credit decisions and represents the ability to convert an asset into cash quickly. In the private dataset, CASH/TA is the most important single variable relative to default. Quick ratio is an indicator of a firm's short-term liquidity and measures a firm's ability to meet its short-term obligations with its most liquid assets. The larger the quick ratio, the better the position of the firm. The quick ratio is more conservative than the current ratio because it excludes inventory from current assets. Current ratio is mainly used to give an idea of the firm's ability to pay back its short-term liabilities (debt and payables) with its short-term assets (cash, inventory, receivables). If a firm is in default, its current ratio must be low. Yet, just as the cash in your wallet does not necessarily imply wealth, a high current ratio does not necessarily imply health. Working capital measures both a firm's efficiency and its short-term financial health. Altman (1968) reported that the WC/TA ratio is a measure of the net liquid assets of the firm relative to the total capitalization and proved to be more valuable than the current ratio and the quick ratio. Falkenstein *et al.* (2000) showed that, firstly, the CL/TL ratio appears of little use in forecasting, second that the quick ratio appears slightly more powerful than the WC/TA ratio, and third, the quick ratio and current ratio carry roughly similar information.

Activity ratios also capture important bankruptcy information and are frequently used when performing fundamental analysis for different firms. We analyse four different activity ratios: the asset turnover (TA/SALE, x21), the inventory turnover (INV/SALE, x22), the account receivable and payable turnover (AR/SALE, x23; AP/SALE, x24).

The asset turnover ratio is a standard financial ratio illustrating the sales-generating ability of the firm's assets. Usually, the asset turnover is non-monotonic and very flat. Note that some studies report that the asset turnover degrades model predictability, for example the *Z*-score that reduces the asset turnover performs better than the one that keeps it. The reciprocal of the inventory turnover shows how many times a firm's inventory is sold and replaced over a period. A high turnover implies poor sales and, therefore, excess inventory. High inventory levels are unhealthy because they represent an investment with a rate of return of zero. Accounts payable and receivable turnover ratios are more powerful predictors, the reciprocals of which also display how many times the firm's accounts are converted into sales over a period. The former is a short-term liquidity measure used to quantify the rate at which a firm pays off its suppliers. The latter is a measure used to quantify a firm's effectiveness in extending credit as well as collecting debts. By maintaining accounts receivable, firms are indirectly extending interest-free loans to their clients. The above description of the activity ratios is usually true in the manufacturing industry but is not the case for other industries. For instance, service firms may have no inventory to turn over.

Sales or total assets are almost indistinguishable as indicators of size risk, which makes the choice between the two measures arbitrary. In this study, we use the natural logarithm of total assets (log(TA), x25) to represent the firm size to investigate the default risk of small, medium (SMEs) and large firms. For example, access to capital for these firms is very different and may affect the prediction ability of some financial ratios and, consequently, the performance of the SVM model. Due to the available variables provided by the Creditreform database, we also compute three ratios of the percentage of incremental inventories, liabilities and cash flow (x26, x27, x28), respectively. For example, the increased (decreased) cash flow is the additional operating cash flow that an organization receives from taking on a new project. A positive incremental cash flow means that the firm's cash flow will increase with the acceptance of the project, the ratio of which is a good indication that an organization should spend some time and money investing in the project.

Previous empirical research has found that a firm is more likely to go bankrupt if it is unprofitable, highly leveraged, and suffers cashflow difficulties (Myers 1977, Aghion and Bolton 1992, Lennox 1999). Moreover, large firms are less likely to encounter credit constraints because of reputation effects. This is clearly demonstrated by the statistical description of financial ratios in table 3, which shows that insolvent firms are typically small, have poor profitability and liquidity, and are highly leveraged, compared with solvent firms, with only a few exceptions such as EBIT/SALE, OF/TA and EBIT/INTE. In addition, the firms that go on to default have higher values for the activity ratio. Except for the last three, all ratios for insolvent firms vary less than for solvent firms because of the smaller number of observations.

The statistics described in table 3 reveal that several of the ratios are highly skewed and there are many outliers; this may affect whether they can be of much help in

identifying insolvent and solvent firms. It is also possible that many of these outliers are errors of some kind. Therefore, the ratios used in the following analysis are processed as follows: if $x_i < q_{0.05}(x_i)$, then $x_i = q_{0.05}(x_i)$, and if $x_i > q_{0.95}(x_i)$, then $x_i = q_{0.95}(x_i)$, $i = 1, 2, \ldots, 28$. $q_\alpha(x_i)$ is an α quantile of x_i. Thus, the discriminating results obtained from both the SVM and the logit model are robust and not sensitive to outliers.

4. Prediction framework

4.1. The validation procedure

To compare the SVM and the logit models in a setting most close to the real situation in which these models are used in practice, the holdout method is chosen in this study for cross validation, namely training of the model on all available data up to the present period and the forecasting of default events for the next period. In this study, the training data are chosen from 1997 through 1999, and the validating set are selected from 2000 through 2002. Then the model is first estimated using the training data; once the model form and parameters are established, the model is used to identify insolvencies among all the firms available during the holdout period (2000–2002). Note that the predicted outputs for 2000 through 2002 are out of time for firms existing in the previous three years, and out of sample for all the firms whose data become available only after 2000. Such out-of-sample and out-of-time tests are the most appropriate way to compare model performance. The validation result set is the collection of all the out-of-sample and out-of-time model predictions that can then be used to analyse the performance of the model in more detail. For an introduction to the validation framework, see Sobehart *et al.* (2001).

Following the holdout validation procedure, we construct a training set containing 387 insolvent and 3534 solvent companies and a validation set containing 396 default events and 6049 non-defaulters. Note that the training and validation sets are themselves a subsample of the population and, therefore, may yield spurious model performance differences based only on data anomalies. A common approach to overcome this problem is to use the re-sampling techniques to leverage the available data and reduce the dependency on the particular sample at hand (Efron and Tibshirani 1993, Herrity *et al.* 1999, Horowitz 2001). Re-sampling approaches provide two related benefits (Sobehart *et al.* 2001). First, they give an estimate of the variability around the actual reported model performance. This variability can be used to determine whether differences in model performance are statistically significant, using familiar statistical tests. Second, because of the low numbers of defaults, re-sampling approaches decrease the likelihood that individual defaults (or non-defaults) will overly influence the chances of a particular model being ranked higher or lower than another model. Similar to previous bankruptcy studies, this paper also adopts a matched pairs approach for drawing subsamples for both the training and validation set. The advantage of the matching procedure is that it helps to cut the cost of data collection, as the proportion of insolvent firms in the population is very small. The problem that the use of relatively small samples could lead to over-fitting can be avoided by the re-sample techniques.

The re-sampling technique employed in this analysis is the bootstrap, which proceeds as follows. We use all insolvent firms, 387 in the training set and 396 in the validation set, and randomly select a subsample with the same number of solvencies from the 3534 solvencies in the training set and the 6049 solvencies in the validation set, respectively.

For the selected validation subset the performance measure is calculated and recorded. Then we perform a Monte Carlo experiment: another subsample is drawn, and the process is repeated. This continues for many repetitions until a distribution for each performance measure is established. In this paper the process will be repeated 30 times.

4.2. Performance measures

We now introduce two metrics for measuring and comparing the performance of credit risk models: the Accuracy Ratio (AR) and the misclassification error. These two measures aim to determine the power of discrimination that a model exhibits in warning of default risk. These techniques are quite general and can be used to compare different types of models even when the model outputs differ and are difficult to compare directly.

AR is a valuable and simple tool to determine the discriminative power of risk models. AR can be derived from the Cumulative Accuracy Profile (CAP) curve, which is particularly useful in that it simultaneously measures Type I and Type II errors (Herrity *et al.* 1999, Engelmann *et al.* 2003, Basle Committee on Banking Supervision 2005). In statistical terms, the CAP curve represents the cumulative probability distribution of default events for different percentiles of the risk score scale. To obtain CAP curves, firms are first ordered by their risk scores. For a given fraction $x\%$ of the total number of firms, a CAP curve is constructed by calculating the percentage $y(x)$ of the defaulters whose risk score is equal to or smaller than that for fraction x. In other words, for a given x, $y(x)$ measures the fraction of defaulters (of the total defaulters) whose risk scores are equal to or smaller than those of fraction x (of the total firms). One would expect a concentration of non-defaulters at the highest scores and defaulters at the lowest scores.

Figure 3 shows a CAP plot. The random CAP represents the case of zero information (which is equivalent to a random assignment of scores). The ideal CAP represents the case in which the model is able to discriminate perfectly, and all defaults are caught at the lowest model output. The actual CAP shows the performance of the model being evaluated. It depicts the percentage of defaults captured by the model.

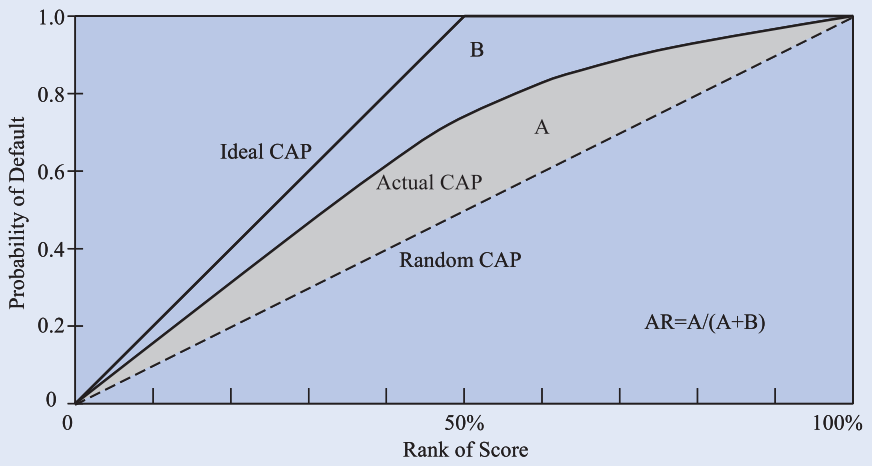


Figure 3. Cumulative accuracy profile (CAP) curve.

Therefore, AR is defined as the ratio of the area between a model's CAP curve and the random CAP curve to the area between the perfect CAP curve and the random CAP curve (see figure 3). The AR value is a fraction between zero and one. Risk measures with AR that approach zero have little advantage over a random assignment of risk scores, whereas those close to one display good predictive power. Mathematically, the AR value is defined as

$$AR = \frac{\int_0^1 y(x)\,\mathrm{d}x - (1/2)}{\int_0^1 y_{\text{ideal}}(x)\,\mathrm{d}x - (1/2)}. \tag{13}$$

If the number of bankruptcies equals the number of operating companies in the sample, then the AR becomes

$$AR \approx 2\int_0^1 y(x)\,\mathrm{d}x - 1. \tag{14}$$

In addition, when evaluating the explanatory power of the bankruptcy models, it is helpful to define two types of prediction error: a type I error, which indicates low default risk when in fact the risk is high, and a type II error, which conversely indicates a high default risk when in fact the risk is low. Usually, minimizing one type of error comes at the expense of increasing the other type of error. Clearly, the type I and type II error rates depend on the number of firms predicted to fail. The higher (lower) the number of firms predicted to go bankrupt, the smaller (larger) is the type I error rate and the larger (smaller) is the type II error rate. The number of predicted bankruptcies depends on the cut-off probability, which is equal to 0.5 in our study. From a supervisory viewpoint, type I errors are more problematic as they produce higher costs. Usually, the cost of a default is higher than the loss of prospective profits. Altman *et al.* (1977) estimated the relative costs of type I and type II errors for commercial bank loans as being 7:1. Sobehart *et al.* (2001) also described the cost scenarios schematically.

For more details on the performance measures, we refer to DeLong *et al.* (1988), Swets (1998), Keenan and Sobehart (1999), Swets *et al.* (2000), Sobehart *et al.* (2001) and Sobehart and Keenan (2004).

4.3. Predictor selection

In this study, the benchmark linear parametric probability model is the conditional logit model estimated by MLE, which is described as follows:

$$\Pr(y_i = 1|\mathbf{x}_{i1},\ldots,\mathbf{x}_{id}) = \frac{1}{1+\mathrm{e}^{-(\beta_0+\beta_1\mathbf{x}_{i1}+\cdots+\beta_d\mathbf{x}_{id}+\varepsilon_i)}}. \tag{15}$$

Based on equation (1) or (11), the target nonlinear non-parametric probability model estimated by the SVM can also be expressed in the following form:

$$\Pr(y_i = 1|\mathbf{x}_{i1},\ldots,\mathbf{x}_{id}) = f(\mathbf{x}_{i1},\ldots,\mathbf{x}_{id}) + \varepsilon_i, \tag{16}$$

where $y_i = 1$ indicates the bankrupt company, and $y_i = 0$ for the logit case and $y_i = -1$ for SVM represent the successful firm; the input vectors $\mathbf{x}_i$ are the relevant company financial ratios explaining the probability of bankrupcy. Before we begin to estimate the models, the process of predictor selection is illustrated.

For a parametric model we can estimate the distribution of the coefficients of the predictors and their confidence intervals. However, we cannot do so for non-parametric models. Instead, we can use the bootstrap technique, as described in the subsection on the validation procedure, to empirically estimate the distribution of the AR on many subsamples. In this study we randomly select 30 subsamples and compute the corresponding ARs 30 times. The median AR provides a robust measure to compare different ratios as predictors.

There are so many possible financial ratios that can be used as explanatory variables in credit scoring models that selection criteria are needed to obtain a parsimonious model. There are two main methods for selecting the appropriate ratios (Falkenstein *et al.* 2000). The first is forward stepwise selection. Start with the predictor that has the highest performance accuracy and then sequentially add the next predictor that also has the highest accuracy in the group and higher than the former until additional predictors have no additional improvement.

Table 4. Median of the AR measure for a univariate SVM model. Accounts payable turnover (AP/SALE, x24) produces the highest AR median.

No.	Ratio	AR median	No.	Ratio	AR median
x1	NI/TA	28.428	x15	CASH/TA	22.140
x2	NI/SALE	22.985	x16	CASH/CL	25.821
x3	OI/TA	36.358	x17	QA/CL	28.746
x4	OI/SALE	31.413	x18	CA/CL	16.983
x5	EBIT/TA	29.941	x19	WC/TA	14.264
x6	EBITDA	29.155	x20	CL/TL	−7.608
x7	EBIT/SALE	19.447	x21	SALE/TA	17.414
x8	OF/TA	32.941	x22	INV/SALE	24.764
x9	(OF-ITGA) / (TA-ITGA-CASH-LB)	31.938	x23	AR/SALE	17.468
x10	CL/TA	18.020	**x24**	**AP/SALE**	**49.174**
x11	(CL-CASH)/TA	23.319	x25	Log(TA)	23.816
x12	TL/TA	22.477	x26	IDINV/INV	15.493
x13	DEBT/TA	16.528	x27	IDL/TL	−9.528
x14	EBIT/INTE	28.270	x28	IDCASH/CASH	−6.562

The second is backward elimination in which one starts with all predictors, then reduces all of the poor variables. In this study, forward selection is preferred for the SVM method due to its relatively lower computational cost. The logit model, with forward selection, together with the investigation of the statistical significance and correct sign of the individual parameters of the predictors, is likely to choose different explanatory variables than the SVM. To compute and compare each method more conveniently, we will only report the results of the logit model with the same predictors as the SVM-based model. The discriminating power of each ratio is assessed using the median of the AR performance measures.

5. Empirical results

This section discusses the empirical results for each stage of the analysis of the German bankruptcy data using an SVM model. The prediction horizon in each case is two years, i.e. the data were recorded two years prior to bankruptcy for the companies that would become bankrupt. The balance sheet and income statement data for 20,000 solvent and 1000 insolvent firms in Germany were selected randomly by Creditreform. These data are represented as the financial ratios listed in table 4. They cover the period from 1996 to 2002. Each company may appear several times in different years.

5.1. Selection of the first predictor and the sensitivity of the SVM parameters

The first stage of analysing default risk is the selection of the first best predictor related to bankruptcy among the 28 ratios using the median of the AR metric in which the SVM model has one input. It is often argued that the SVM lacks interpretability of the results as is the case for the logit model. Most importantly, since there are no distributional assumptions underlying the SVM modeling, it is impossible to test the significance of variables within the SVM framework. Therefore, we will identify the most significant variable in an additional procedure before analysing the SVM model.

Based on table 4 we can see that Accounts Payable Turnover (AP/SALE, x24) provides the highest median AR of 49.17%. We can also see that CL/TL (x20), IDL/TL (x27) and IDCASH/CASH (x28) have a very low accuracy: their median AR values are below zero. For the next step we will select Accounts Payable Turnover (x24) as the first best single predictor related to German default firms, which is somewhat different from previous studies in which it was usually argued that the most significant predictors were profitability or leverage ratios. In fact, the SVM-based nonlinear model is able to search the nonlinear dependence of the data automatically as opposed to the logit model and it is Accounts Payable Turnover selected by SVM as the first predictor that greatly improves the classifying performance of SVM by more than 10%. Using most of the other ratios as the first predictor, the SVM-based model does not exceed the logit model by much in modeling the default risk.

The accounts payable turnover ratio is calculated by taking the average accounts payable and dividing it by the total sales during the same period. Its reciprocal shows investors how many times per period the firm pays its average payable amount. If the turnover ratio increases from one period to another, this is a sign that it takes the firm longer to pay off its suppliers than before. The opposite is true when the turnover ratio is falling, which means that the firm is paying off suppliers at a faster rate. Therefore, the firms with higher accounts payable turnover values will have less ability to convert their accounts into sales, have lower revenues, and go bankrupt more readily.

The SVM model has two control parameters, the influence of which was investigated in this study: the penalty parameter C and the Gaussian kernel coefficient r. C controls the tolerance to misclassification errors on the training set, while r represents the complexity of classifying functions. The possibility of fine-tuning SVM using these parameters, besides the flexibility of its classification function, further contributed to the higher performance of the SVM compared with the logit model,

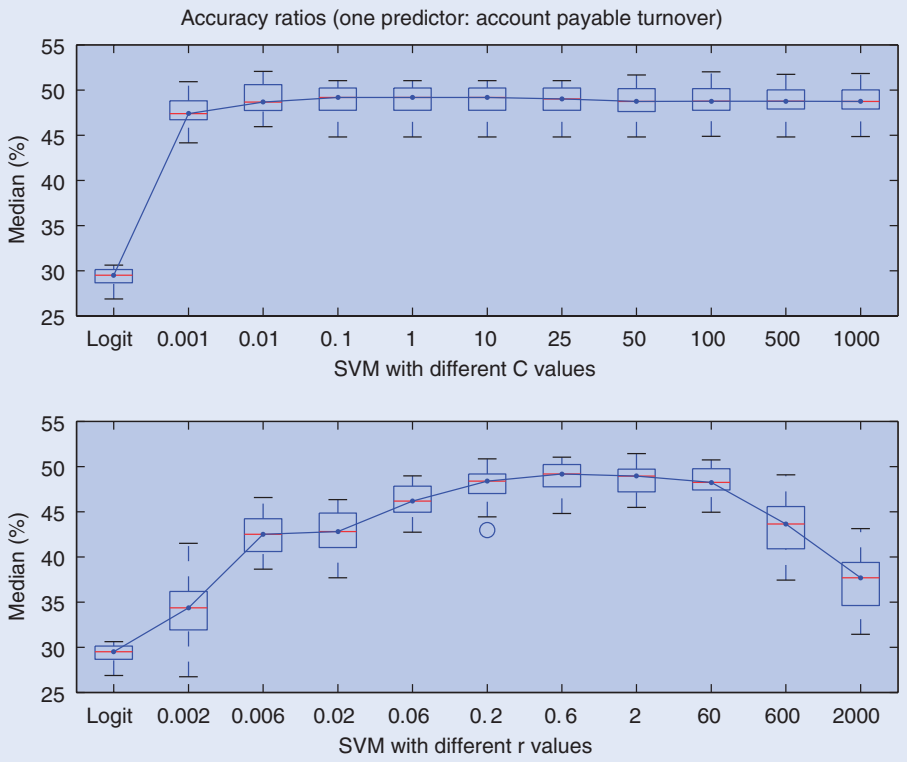


Figure 4. Sensitivity of the SVM to different parameters.

Table 5. Misclassification error (30 randomly selected samples; one predictor AP/SALE, x24).

	Parameter		Type I error		Type II error		Total error	
Model	*C*	*r*	Mean	Std	Mean	Std	Mean	Std
SVM	0.001	0.6	40.57	0.1167	23.43	0.9812	32.01	0.5723
	0.1	0.6	38.42	0.5125	24.45	1.1938	31.44	0.7014
	10	0.6	34.43	1.2126	27.86	1.637	31.15	0.9433
	100	0.6	25.22	0.6176	34.66	1.3541	29.94	0.8086
	1000	0.6	25.76	0.7705	34.26	1.3805	30.01	0.8712
	10	0.002	37.2	2.4512	32.79	2.5753	34.99	1.7611
	10	0.06	31.86	3.1527	29.25	2.2887	30.56	1.1405
	10	0.6	34.43	1.2126	27.86	1.637	31.15	0.9433
	10	60	37.27	0.5112	25.87	1.2134	31.57	0.7798
	10	2000	41.09	0.0791	24.85	0.3265	32.97	0.1123
Logit			38.15	0.5625	32.77	1.1888	35.46	0.7151

which has no similar adjustment parameters. Moreover, a grater SVM performance is a consequence of the SVM loss function, which is a tighter upper bound on the {0,1} step loss function. For univariate models, as figure 4 illustrates, the gain in performance of the SVM over the logit model is substantial and greater than for multivariate models since the former intrinsically has a larger number of degrees of freedom than the latter, which is limited by the number of variables.

The results in table 4 were obtained from the SVM with parameters $C=10$ and $r=0.6$, which were chosen according to the following sensitivity investigation of the SVM parameters (see box plot in figure 4 and table 5). That is to say, the values of parameters C and r could be determined experimentally via the standard use of a re-sampling training data set. Obviously, the SVM differs in different values of the penalty parameter C and the Gaussian kernel coefficient r. The ratio AP/SALE (x24) is exemplified here and the result for the benchmark logit model is also reported.

Here the median ARs are also estimated on 30 bootstrapped subsamples. On the whole, the discriminating ability of the SVM seems to be more sensitive to the value of r rather than to that of C. In figure 4(top), with fixed $r=0.6$, the median of the AR starts from 47.4% for $C=0.001$ and reaches the highest value 49.2% for $C=10$ and slightly decreases to 48.7% when $C=1000$. The varying range of AR is very small. Figure 4(bottom) illustrates the AR of the SVM versus r with fixed $C=10$. Within the interval, r is found to have a strong impact on the AR value, which starts at 34.4% when $r=0.002$ and drastically increases to the highest value 49.2%

when $r = 0.6$ and then decreases to 37.7% when $r = 2000$. In both parts of the figure the discriminating performance of the logit model is inferior to that of the SVM-based model with different parameter values.

As we have seen, $C = 10$ and $r = 0.6$ seem to be the best choice of parameter combination for the study in this paper. Thus, if we do not mention it particularly, the results of the SVM in the remaining part of this paper are all obtained using these parameter values. Note that this is not the case for the other data sample. The appropriate values of the C and r parameters will vary from sample to sample, therefore the sensitivity investigation of the SVM parameters should be carried out before classifying different data samples.

Table 5 shows the percentage of misclassified out-of-sample observations for the logit model and the SVM-based model with different parameters using a single predictor, the Account Payable Turnover. These errors are also obtained by bootstrap, and are all significant according to the standard deviations listed in table 5. Smaller values indicate better model accuracy. As shown in the table, the logit model has higher type I, type II and total error rates than the SVM-based model with only a few exceptions, suggesting that a well-specified SVM-based nonlinear model is superior to a logit model. For the SVM, with an increase of C from 0.001 to 1000, type II errors also increase, but type I errors decrease, and the total errors first decrease and then increase slightly. With increasing r values, type I and total errors also follow a U-shaped trend and type II errors have a monotonic negative relation with the r value. Therefore, $C = 10$ and $r = 0.6$ also appear to be the appropriate trade-off choice for our study in the following part of this paper. They produce only 34.43% type I errors, 27.86% type II errors and 31.15% total errors, whereas logit analysis produces 38.15% type I errors, 32.77% type II errors and 35.46% total errors.

As is evident from figure 5, which shows a univariate dependence of PD on AP/SALE, this dependence is not monotonously increasing or following any distinctive pattern, e.g. a logistic function. The SVM, being a more flexible non-parametric approach, is better suited for describing a broader class of dependence, such as this one, than the logit model. Another advantage of the SVM is its smaller bias in the estimation of the boundary between the solvent and insolvent companies in a situation when the number of the former is much larger than the number of the latter, as is almost always the case. The score of the logit model, which is interpreted as a PD, can be significantly biased for score values much lower or higher than 0.5. Subsequently, the threshold score for the boundary between solvent and insolvent companies is also biased. This is one reason for the substantial improvement in accuracy of the SVM compared with the logit model, as illustrated in figure 4. Because of this feature the SVM gains an additional improvement over the logit model if instead of subsamples with a 50/50 ratio of insolvent versus solvent companies we use subsamples where solvent companies prevail.

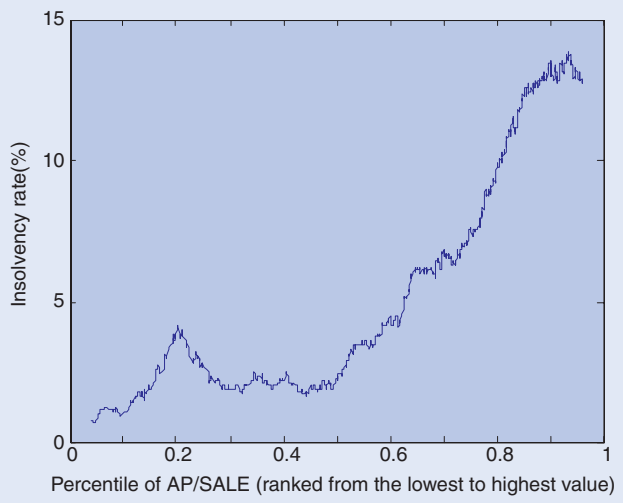


Figure 5. Insolvency rate evaluated for the financial ratio AP/SALE (x24) from the German Creditreform database. The k-nearest-neighbors procedure was used with the size of the window around 1/12 of 18,800 observations (the observations with zero values of sales used as the denominator to calculate the ratios were deleted from all 21,000 observations).

5.2. *Comparison of models with two predictors and PD visualization*

Table 6 shows the identifying performance of bivariate SVM-based models using the best predictor from the univariate model (AP/SALE) and one other. The values of the median of the AR direct us to the profitability ratio OI/TA (x3), the value of which increases to the highest of 56.46%, which indicates that OI/TA (x3) is the best choice for the second predictor.

Therefore, different from the usual result that NI/TA dominates other profitability ratios related to default risk, our study reveals that OI/TA performs better than the others in identifying bankrupt German firms. As the operating income does not include items such as investments in other firms, taxes, interest expenses and depreciation, the ratio represents a firm's true operating performance.

For two dimensions (i.e. two predictors), graphs are obviously an extremely useful tool for studying the data and assessing the quality of different default risk models. In addition, because of its nonlinearity it is more necessary for the SVM-based model to use visual tools than for the logit model to represent classification results. We demonstrate an application of visualization techniques for default analysis and parameter sensitivity investigation based on the SVM in figure 6. In the case of the logit model, the scores can be directly explained as the default probabilities, whereas for the SVM-based model the probabilities of default need to be calculated using the risk scores predicted by the estimated classifying function. Making use of the monotonic logistic cumulative distribution function, the default probabilities of German companies by SVM are calculated from the scores and then plotted as the background contour in figure 6 (corresponding to the right-hand bar in each sub-figure). The two predictors are the ratios AP/SALE (x24) and OI/TA (x3). These graphs are a subset of those used in

Table 6. Median of AR measure for a bivariate SVM model. AP/SALE (x24) and OI/TA (x3) produce the highest AR median.

No.	Ratio	AR median	No.	Ratio	AR median
x1	NI/TA	54.362	x15	CASH/TA	53.011
x2	NI/SALE	53.809	x16	CASH/CL	52.233
x3	**OI/TA**	**56.460**	x17	QA/CL	50.553
x4	OI/SALE	55.652	x18	CA/CL	44.678
x5	EBIT/TA	54.409	x19	WC/TA	48.676
x6	EBITDA	53.847	x20	CL/TL	49.725
x7	EBIT/SALE	52.948	x21	SALE/TA	49.624
x8	OF/TA	51.907	x22	INV/SALE	51.305
x9	(OF-ITGA) / (TA-ITGA-CASH-LB)	51.316	x23	AR/SALE	49.604
x10	CL/TA	48.197	**x24**	**AP/SALE**	
x11	(CL-CASH)/TA	49.680	x25	Log(TA)	51.545
x12	TL/TA	51.080	x26	IDINV/INV	49.904
x13	DEBT/TA	52.231	x27	IDL/TL	49.013
x14	EBIT/INTE	46.517	x28	IDCASH/CASH	46.617

the study. White and black points represent the 396 insolvent and 396 solvent firms from one random subsample of the validation set. The outliers were capped at the 5% and 95% quantiles as described in section 3.2 and kept in the subsample. In most panels of figure 6 they appear at the border. The classifying decision function (optimal hyperplane) is represented by the line denoted 0.5, along which the default probability is 0.5 and the risk scores are zero for SVM. The lines denoted 0.3 and 0.7 (or, more accurately, 0.27 and 0.73) are the lower and upper boundaries of the separation margin corresponding to scores of -1 and $+1$ in SVM. As shown in figure 6, clearly most successful firms lying in the blue area have positive profitability (OI/TA) and relatively lower account payable turnover (AP/SALE), while a majority of bankrupt firms is located in the opposite area. As known, low profitability usually indicates a high default risk, but extremely high profitability may also indicate a high cash flow volatility that is likely to translate into a higher default probability. Although the SVM-based model is sufficiently flexible to reveal a nonlinear dependence between profitability and PD, different from the logit model, for the Creditreform data in this study, the dependence could be too weak to be captured by SVM. Also, the sensitivity investigation results of the free parameters, C and r, of SVM could easily be determined from the figure.

Figure 6(a) shows the classification results for the logit model. Because the disadvantage of the logit model is the linearity of its solution, we see a straight classification line that is the linear combination of two predictors. Figure 6(b) shows the discriminating results obtained with the SVM-based model using a classifying function of moderate complexity ($r = 0.6$) and $C = 10$. This nonlinear classifying line (score 0 and PD 0.5) seems to identify the two types of firms very well with the areas in which solvent and insolvent firms are localized.

Fix $r = 0.6$. If the penalty is too low (C decreases to 0.01 and 0.1 as in figures 6(c) and (d)), the discriminating curve becomes flatter than that in figure 6(b). The calculated default probabilities are too small to display the two boundaries. That is, most of the firms fall inside the separation region but the insolvent and solvent firms are still clustered in their own areas. If the penalty increases, for example $C = 500$ as in figure 6(e), the identifying ability of SVM cannot be increased further than shown in figure 6(b).

Fix $C = 10$. If the complexity of the classifying functions increases (the r value decreases to 0.06 as illustrated in figure 6(f)), the SVM will try to capture each observation, although the majority of the insolvent firms still lie inside the band (0.5, 0.7) and above, with the solvent firms inside (0.5, 03) and below. The complexity in this case is too high for the given sample. If the r value increases to 60 (figure 6(g)), the classifying curve becomes flatter than that with $r = 0.6$; if r increases further to 2000 (figure 6(h)), the discriminating curve can be approximated as a linear combination of two predictors and is similar to the benchmark logit model, although the coefficients of the predictors may be different. The calculated default probabilities are also very small. The complexity here is too low to obtain a more detailed picture.

Although two cases of high complexity clearly demonstrate overfitting, (f) when $C = 10$ and $r = 0.06$, and (e) when $C = 500$ and $r = 0.6$, in all other cases the separating line is moderately nonlinear and for the case of a virtually linear SVM (h) with $C = 10$ and $r = 2000$ the separating line resembles that for the logit regression (a), with a different slope. Perfect separation for out-of-sample observations is not possible in any case. Nevertheless, comparing panel (a) for the logit with panel (f) for the SVM that achieved the maximum separation power, we observe that the most important difference between the two is in the area where the density of observations is the highest and even a small change in shape can lead to a substantial change in the classification ability.

The sensitivity analysis information obtained from this graphical analysis is similar to Härdle *et al.* (2005) and also confirms the choice combination of parameters as described in the sensitivity investigation of section 5.1. A set of alternative random subsamples as extracted from the validation set also display similar findings using the same visualization technique.

While the analysis here has been restricted to only two classes, namely bankruptcy and solvency, it can easily be

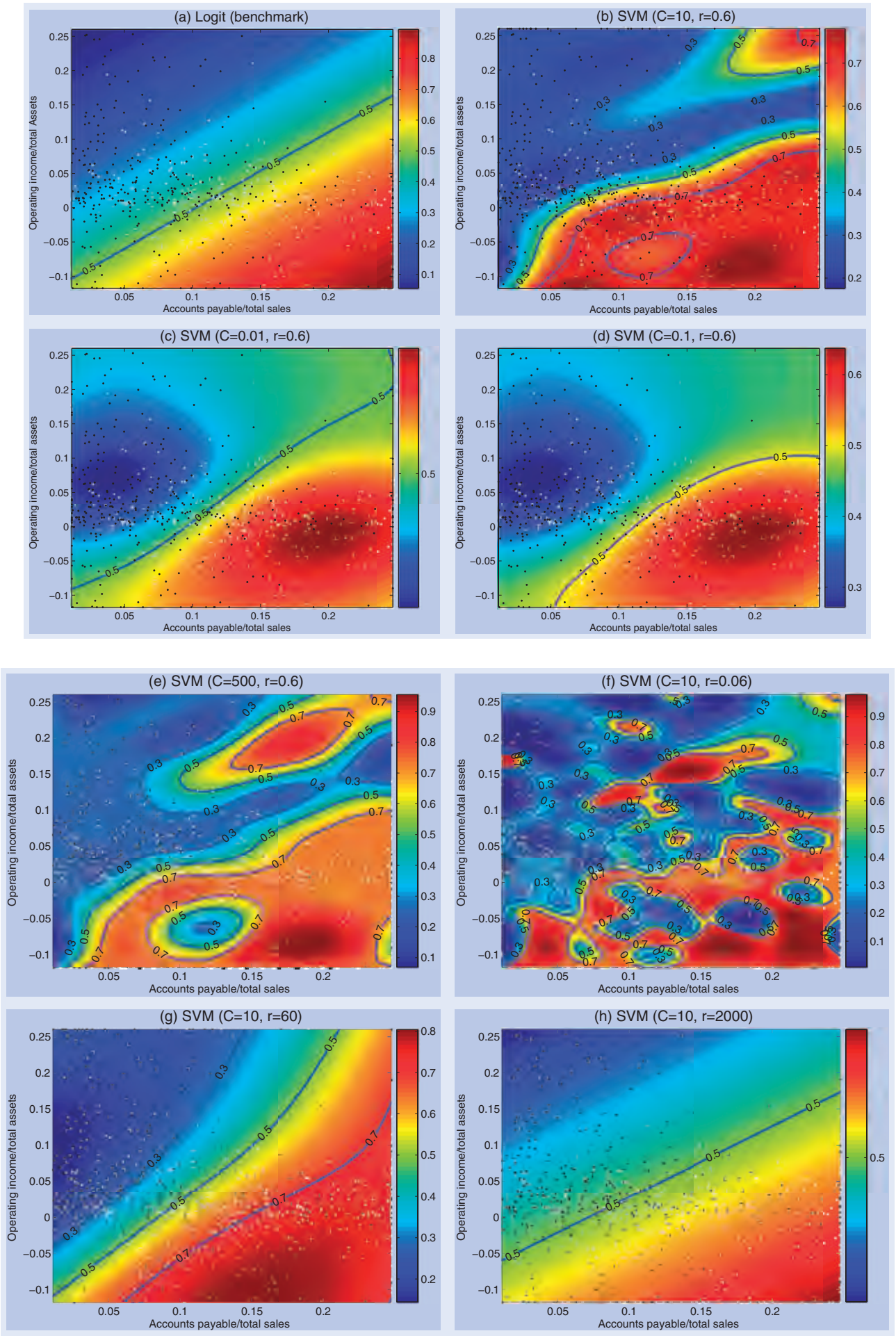


Figure 6. Default probabilities predicted for one random subsample and sensitivity analysis for the SVM.

Table 7. Median of AR measure for the best SVM model with eight important financial ratios calculated on 50/50 subsamples.

		AR median					AR median		
No.	Ratio	Logit	SVM	Predictors	No.	Ratio	Logit	SVM	Predictors
x1	NI/TA	35.12	59.93		**x15**	**CASH/TA**			**3**
x2	**NI/SALE**	**35.15**	**60.51**	**8**	x16	CASH/CL	34.87	59.42	
x3	**OI/TA**			**2**	x17	QA/CL	34.66	55.62	
x4	OI/SALE	35.06	60.44		x18	CA/CL	34.41	54.93	
x5	**EBIT/TA**			**7**	x19	WC/TA	34.72	59.48	
x6	EBITDA	34.93	59.85		x20	CL/TL	33.91	57.45	
x7	EBIT/SALE	35.14	60.4		x21	SALE/TA	35.05	56.61	
x8	OF/TA	35.04	59.64		**x22**	**INV/SALE**			**6**
x9	(OF-ITGA)/(TA-ITGA-CASH-LB)	34.94	59.42		x23	AR/SALE	35.15	59.81	
x10	CL/TA	33.94	58.19		**x24**	**AP/SALE**			**1**
x11	(CL-CASH)/TA	34.01	57.76		x25	Log(TA)	36.14	55.77	
x12	**TL/TA**			**4**	**x26**	**IDINV/INV**			**5**
x13	DEBT/TA	34.97	59.07		x27	IDL/TL	35.22	58.88	
x14	EBIT/INTE	35.03	54.37		x28	IDCASH/CASH	35.06	55.08	

generalized to multiple classes. In a multiple class case, financial analysts usually pre-specify rating classes (i.e. AAA, A, BB, C, etc.). A certain range of scores and default probabilities is associated with each rating class. The ranges are computed on the basis of historical data. According to the similarity of the scores, a new firm is assigned to one particular class. Therefore, we can draw more than one classifying function in the figure above to separate different rating classes.

5.3. *Powerful predictors related to insolvent German firms*

The selection procedure will be repeated for each new ratio added. The values of the AR increase until the model includes eight ratios, then they slowly decline. The medians of the AR for the models with eight ratios are shown in table 7. Most of the models tested here had AR values in the range 43.50–60.51% for out-of-sample and out-of-time tests. The results reported here are the product of the bootstrap approach described in the previous section. Obviously, the SVM-based model including ratios AP/SALE (x24), OI/TA (x3), CASH/TA (x15), TL/TA (x12), IDINV/INV (x26), INV/SALE (x22), EBIT/TA (x5) and NI/SALE (x2) attains the highest median AR, 60.51%. For comparison, we also report the median AR for the benchmark logit model with the same ratios. We can see that, for models containing the former seven ratios and one of the remaining, the medians of the AR are always higher for the SVM. This clearly reveals that the SVM-based model is always consistently superior to the benchmark logit model in identifying bankrupt firms and confirms the theoretical advantage of SVM for classification in the linear non-separable case. With respect to the percentage of correctly classified out-of-sample observations, a similar result is achieved (71.85% for the SVM-based model vs. 67.24% for the logit model).

It is noteworthy that, because the insolvency data was collected two years prior to insolvency, the predicted risk scores and calculated performance metrics in this study measure the model's ability to identify the firms that are going to default within the next two years. For example, the predicted default probability for 2002 denotes the probability that a firm defaults in 2003 or 2004.

We could not significantly improve upon our results by adding more ratios, and no model with fewer ratios performed as well. The eight selected predictors related to bankrupt German firms are AP/SALE (account payable turnover, x24), OI/TA (x3), CASH/TA (x15), TL/TA (x12), IDINV/INV (percentage of changing inventories, x26), INV/SALE (inventory turnover, x22), EBIT/TA (x5) and NI/SALE (net profit margin, x2). The size of the company was controlled in the analysis by the logarithm of the total assets (log(TA), x25). This can serve as a proxy for the cost of capital. In contrast to other studies, firm size has been shown to have no important effects on the probability of bankruptcy, which could be the result of pre-selecting only medium-sized companies.

Among the powerful predictors in identifying bankrupt German firms, there are two activity ratios (Account Payable Turnover and Inventory Turnover), three profitability ratios (OI/TA, EBIT/TA and Net Profit Margin), one liquidity ratio (CASH/TA), one leverage ratio (TL/TA) and one percentage of change ratio (Percentage of Incremental Inventories). It seems that activity ratios play the most important role in predicting the default probabilities of German firms. The activity ratio measures a firm's ability to convert different positions of their balance sheets into cash or sales. German firms will typically try to turn their accounts payable and inventories into sales as fast as possible because these will actually lead to higher revenues. Instead of ROA, EBIT/TA has a more powerful impact on insolvent German firms. In essence, it measures the operating performance and true productivity of firm assets on whose earning power the existence of the firm is based. Of course, the earnings of a firm only cannot tell the entire story. High earnings are good, but an increase in earnings does not mean that the net profit margin of a firm is improving.

For instance, if a firm has costs that have increased at a greater rate than sales, it leads to a lower profit margin. This is an indication that costs need to be under better control. Therefore, net profit margin is also very useful when analysing German bankruptcy data. In our study the liquidity ratio CASH/TA is only inferior to activity and profitability ratios when explaining German bankruptcies. Its strong explanatory power may result because the sample used in this study is mainly composed of private firms and this might not be true for public firms used in previous studies. The leverage ratio TL/TA also has a powerful influence on the identification of German bankruptcies. This metric is used to measure a firm's financial risk by determining how much of its assets have been financed by debt. This is a very broad ratio as it includes short- and long-term liabilities (debt) as well as all types of both tangible and intangible assets. The higher a firm's degree of leverage, the more the firm is considered risky. A firm with high leverage is more vulnerable to downturns in the business cycle because the firm must continue to service its debt regardless of how bad sales are. The incremental inventories provided by the Creditreform database also contain useful information for studying insolvent German firms.

To summarize our results, a German firm is most likely to go bankrupt when it has high turnover, low profits, low cash flows, is highly leveraged and has a high percentage of changing inventories. Although these results are similar to those of previous studies, the discovery of significant effects of the activity ratio and incremental inventories for predicting defaults in Germany is new.

6. Conclusions

We use a discrimination technique, the Support Vector Machine for classification, to analyse the German bankrupt company database spanning from 1996 through 2002. The identifying ability of an SVM-based nonlinear and non-parametric model is compared with that of the benchmark logit model with regard to two performance metrics (AR and misclassification error) on the basis of bootstrapped subsamples. The evidence from empirical results consistently shows that a credit risk model based on SVM significantly outperforms the benchmark linear parametric model in modeling the default risk of German firms out of sample and out of time. The sensitivity of the SVM to the penalty parameter C and Gaussian kernel coefficient r is examined according to the median of the AR using box plots (see figure 4), classification errors (see table 5) and two-dimensional visualization tools (figure 6). It is found that the discriminating ability of the SVM seems to be more sensitive to the values of r than C. Thus, appropriate trade-off values of parameters C and r should be chosen for bankruptcy analysis; for example, $C = 10$ and $r = 0.6$ in this study for the formal empirical analysis.

In addition to the unique minimum, no prior assumptions and it not being necessary to adjust the collinearity between the ratios, in particular the principle of structural risk minimization, endows the SVM approach with the most excellent classifying ability among all alternatives. Also, the SVM-based model is good at searching the linear non-separable hypersurface, which the logit model cannot do. As shown in table 4, the ratio Account Payable Turnover was selected by SVM among 28 candidates as the first best predictor to model the risk, which drastically upgrades the classifying accuracy, AR, of SVM by more than 10% as opposed to most of the other ratios selected. Otherwise, the performance gap between the SVM-based and logit model would not be so great, as shown in table 7. If the data are nonlinear, e.g. the Creditreform database, no linear model is able to separate the populations optimally, regardless of the DA, and the logit and probit models. The SVM method (as well as other pattern-recognition techniques) provides a more consistent way of finding the nonlinearities in the data, as opposed to performing an *ad-hoc* search of all possible combinations of the logit model. The holdout validation method, the most appropriate for modeling the real risk in practice, and the bootstrap re-sampling technique, guarantee the robustness and stability of the SVM approach. Due to the application of a kernel function and the sparseness of the algorithm, the achievement of such an improvement by SVM is not at a cost of much computational time, just a few seconds. Therefore, the empirical evidence confirms the theoretical advantage of SVM for classification and justifies it as applicable in practice. Of course, the non-parametric nature behind the SVM will come at the expense of understanding and insight; that is, the impact (the magnitude and direction and its significance) of the predictors on the default probabilities cannot be interpreted explicitly, in contrast to the parametric logit model. What the SVM is good at is capturing the nonlinearities better and forecasting the default probabilities more accurately than the benchmark.

As described in section 5.3, there are eight accounting ratios that are powerful predictors related to the bankruptcy of German companies. It turns out that activity ratios such as Account Payable and Inventory Turnover play the most important role in predicting the default probabilities. The percentage of incremental inventories provided by the Creditreform database also contains useful information for German bankruptcy analysis. These findings are new and somewhat different from the other default risk studies. The ability to automatically find the nonlinear dependence of the SVM model and the application of a widely accepted forward stepwise selection procedure in our case provides adequate selection that cannot be done by the usual linear classifying techniques such as the DA, logit model. That is to say, for German companies, Account Payable and Inventory Turnover, the percentage of incremental inventories selected have a strong nonlinear dependence on PDs, but a weak linear dependence that may lead to their unpopularity. Consistent with previous research, the profitability ratios, e.g. OI/TA, EBIT/TA and NI/SALE (net profit margin), are also powerful predictors related to German insolvency. Other results are similar to published research, e.g. that liquidity and leverage ratios also have

important effects on the probability of default for German companies. But, in contrast to the others, firm size (log(TA), x25) was not chosen by the forward selection procedure as a predictor, which could be the result of pre-selecting only medium-sized companies.

Acknowledgements

The authors thank the Editors, Philip Angell, David Burgoyne, Beth Cawte, Collette Teasdale, and two referees for their constructive suggestions that significantly improved the paper. This work was supported by Deutsche Forschungsgemeinschaft through SFB 649 'Economic Risk'. Shiyi Chen was also sponsored by the Shanghai Pujiang Program, the Shanghai Leading Academic Discipline Project (No. B101) and the State Innovative Institute of Project 985 at Fudan University. W.K. Härdle was also partially supported by the National Center for Theoretical Sciences (South), Taiwan. R.A. Moro was supported by the German Academic Exchange Service (DAAD).

References

Aghion, P. and Bolton, P., An 'incomplete contracts' approach to financial contracting. *Rev. Econ. Stud.*, 1992, **59**(3), 473–494.

Altman, E.I., Financial ratios, discriminant analysis and the prediction of corporate bankruptcy. *J. Finance*, 1968, **23**(4), 589–609.

Altman, E.I., Haldeman, R. and Narayanan, P., Zeta analysis: a new model to identify bankruptcy risk of corporations. *J. Bank. Finance*, 1977, **1**(1), 29–54.

Back, B., Laitinen, T. and Sere, K., Neural networks and bankruptcy prediction, in *17th Annual Congress of the European Accounting Association*, Venice, Italy, 1994. Abstract in *Collected Abstracts of the 17th Annual Congress of the European Accounting Association* 116.

Back, B., Laitinen, T., Sere, K. and Wezel, M., Choosing bankruptcy predictors using discriminant analysis, logit analysis, and genetic algorithms. Technical Report 40, TUCS Research Group, 1996.

Basle Committee on Banking Supervision, Studies on the validation of internal rating systems. AIG/RTF BIS Working Paper No. 14, 2005.

Beaver, W., Financial ratios as predictors of failures. Empirical research in accounting: Selected studies. *J. Account. Res.*, 1966, **5**(suppl.), 71–111.

Bertsekas, D.P., *Nonlinear Programming*, 1995 (Athenas Science: Belmont, MA).

Burnham, K.P. and Anderson, D.R., *Model Selection and Inference*, 1998 (Springer: New York).

Caouette, J.B., Altman, E.I. and Narayanan, P., *Managing Credit Risk: The Next Great Financial Challenge*, 1998 (Wiley: New York).

Chakrabarti, B. and Varadachari, R., Quantitative methods for default probability estimation – a first step towards Basel II. i-flex solutions, 2004.

Collins, R. and Green, R., Statistical methods for bankruptcy prediction. *J. Econ. Business*, 1982, **34**(4), 349–354.

Courant, R. and Hilbert, D., *Methods of Mathematical Physics*, Vol. I and II, 1970 (Wiley Interscience: New York).

DeLong, E.R., DeLong, D.M. and Clarke-Pearson, D.L., Comparing the areas under two or more correlated receiver operating characteristic curves: a non-parametric approach. *Biometrics*, 1988, **44**(3), 837–845.

Deng, N.Y. and Tian, Y.J., *New Methods in Data Mining: Support Vector Machine*, 2004 (Science Press: Beijing).

Efron, B. and Tibshirani, R.J., *An Introduction to the Bootstrap*, 1993 (Chapman & Hall: New York).

Engelmann, B., Hayden, E. and Tasche, D., Testing rating accuracy. *Risk*, 2003, **January**, 82–86.

Falkenstein, E., Boral, A. and Carty, L., Riskcalc for private companies: Moody's default model, Report Number: 56402, Moody's Investors Service, Inc., New York, 2000.

Fitzpatrick, P., *A Comparison of the Ratios of Successful Industrial Enterprises With Those of Failed Companies,* 1932 (The Accountants Publishing Company: Washington, DC).

Fletcher, R., *Practical Methods of Optimization*, 2nd ed., 1987 (Wiley: New York).

Friedman, C. and Sandow, S., Model performance measures for expected utility maximizing investors. *Int. J. Theor. Appl. Finance*, 2003a, **6**(4), 355–401.

Friedman, C. and Sandow, S., Learning probabilistic models: an expected utility maximization approach. *J. Mach. Learn. Res.*, 2003b, **4**, 257–291.

Friedman, C. and Huang, J., Default probability modeling: a maximum expected utility approach. Standard & Poor's Risk Solutions Group, New York, 2003.

Gaeta, G., editor, *The Certainty of Credit Risk: Its Measurement and Management*, 2003 (Wiley Finance (Asia): Singapore).

Gestel, T.V., Baesens, B., Dijcke, P.V., Suykens, J., Garcia, J. and Alderweireld, T., Linear and nonlinear credit scoring by combining logistic regression and support vector machines. *J. Credit Risk*, 2005, **1**(4), 31–60.

Giesecke, K., Credit risk modeling and valuation: An introduction. In *Credit Risk: Modeling and Management*, 2nd ed., edited by D. Shimko, pp. 487–526, 2004 (Risk Books: London).

Hanley, A. and McNeil, B., The meaning and use of the area under a receiver operating characteristics (ROC) curve. *Diagn. Radiol.*, 1982, **143**(1), 29–36.

Härdle, W., Moro, R.A. and Schäfer, D., Predicting bankruptcy with support vector machines. In *Statistical Tools for Finance and Insurance*, edited by P. Cizek, W. Härdle, and R. Weron, pp. 225–248, 2005 (Springer: Berlin).

Härdle, W., Moro, R.A. and Schäfer, D., Graphical data representation in bankruptcy analysis. In *Handbook for Data Visualization*, edited by Ch.-H. Chen, W. Härdle, and A. Unwin, pp. 853–872, 2007 (Springer: Berlin).

Härdle, W., Müller, M., Sperlich, S. and Werwatz, A., *Nonparametric and Semiparametric Models*, 2004 (Springer: Heidelberg).

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

Haykin, S., *Neural Networks: A Comprehensive Foundation*, 1999 (Prentice-Hall: Engelwood Cliffs, NJ).

Herrity, J.V., Keenan, S.C., Sobehart, J.R., Carty, L.V. and Falkenstein, E.G., Measuring private firm default risk. Moody's Investors Service Special Comment, 1999.

Hertz, J., Krogh, A. and Palmer, R.G., *The Theory of Neural Network Computation*, 1991 (Addison Welsey: Redwood, CA).

Horowitz, J.L., *The Bootstrap*, Vol. 5, 2001 (Elsevier: Amsterdam).

Keenan, S.C. and Sobehart, J.R., Performance measures for credit risk models. Research report #1-10-10-99, Moody's Risk Management Services, 1999.

Khandani, B., Lozano, M. and Carty, L., Moody's riskcalc for private companies: The German model. Rating Methodology, Moody's Investors Service, 2001.

Lennox, C., Identifying failing companies: A re-evaluation of the logit, probit and DA approaches. *J. Econ. Business*, 1999, **51**, 347–364.

Lo, A.W., Logit versus discriminant analysis: A specification test and application to corporate bankruptcies. *J. Econometr.*, 1986, **31**(2), 151–178.

Mercer, J., Functions of positive and negative type, and their connection with the theory of integral equations. *Trans. London Philos. Soc. A*, 1908, **209**, 415–446.
Merwin, C., Financing small corporations in five manufacturing industries, 1926–36. National Bureau of Economic Research, 1942.
Myers, S., Determinants of corporate borrowing. *J. Financial Econ.*, 1977, **5**(2), 147–175.
Ohlson, J., Financial ratios and the probabilistic prediction of bankruptcy. *J. Account. Res.*, 1980, **18**, 109–131.
Platt, H., Platt, M. and Pedersen, J., Bankruptcy discrimination with real variables. *J. Business, Finance Account.*, 1994, **21**(4), 491–510.
Ramser, J. and Foster, L., A demonstration of ratio analysis. Bulletin No. 40, Bureau of Business Research, University of Illinois, 1931.
Refenes, A.P., *Neural Networks in the Capital Markets*, 1995 (Wiley: Chichester).
Saunders, A. and Allen, L., *Credit Risk Measurement*, 2nd ed., 2002 (Wiley: New York).
Serrano, C., Martin, B. and Gallizo, J.L., Artificial neural networks in financial statement analysis: Ratios versus accounting data. Technical report, paper presented at the 16th Annual Congress of the European Accounting Association, Turku, Finland, April 28–30, 1993.
Shumway, T., Forecasting bankruptcy more accurately: a simple hazard model. Working Paper, University of Michigan Business School, 1998.
Sobehart, J.R., Stein, R.M., Mikityanskaya, V. and Li, L., Moody's public firm risk model: a hybrid approach to modeling default risk. Moody's Investors Service Rating Methodology, 2000.
Sobehart, J., Keenan, S. and Stein, R., Benchmarking quantitative default risk models: A validation methodology. *Algo Res. Q.*, 2001, **4**(1/2), 57–72.
Sobehart, J.R. and Keenan, S.C., Performance evaluation for credit spread and default risk models. In *Credit Risk: Models and Management*, 2nd ed., edited by D. Shimko, pp. 275–305, 2004 (Risk Books: London).
Swets, J.A., Measuring the accuracy of diagnostic systems. *Science*, 1998, **240**(4857), 1285–1293.
Swets, J.A., Dawes, R.M. and Monahan, J., Better decisions through science. *Sci. Am.*, 2000, **October**, 82–87.
Tikhonov, A.N., On solving ill-posed problem and method regularization. *Dokl. Akad. Nauk USSR*, 1963, **153**, 501–504.
Tikhonov, A.N. and Arsenin, V.Y., *Solution of Ill-posed Problems*, 1977 (W.H. Winston: Washington, DC).
Vapnik, V., *Estimation of Dependencies Based on Empirical Data*, 1979 (Nauka: Moscow).
Vapnik, V., *The Nature of Statistical Learning Theory*, 1995 (Springer: New York).
Vapnik, V., *Statistical Learning Theory*, 1997 (Wiley: New York).
Wilson, R.L. and Sharda, R., Bankruptcy prediction using neural networks. *Decis. Supp. Syst.*, 1994, **11**, 545–557.
Winakor, A. and Smith, R., Changes in the financial structure of unsuccessful industrial corporations. Bulletin No. 51, Bureau of Business Research, University of Illinois, 1935.
Zagst, R. and Hocht, S., Comparing default probability models. Working Paper, Munich University of Technology, 2006.

AStA Adv Stat Anal
DOI 10.1007/s10182-011-0182-1

ORIGINAL PAPER

Simultaneous confidence bands for expectile functions

Mengmeng Guo · Wolfgang Karl Härdle

Received: 25 February 2011 / Accepted: 9 November 2011

Abstract Expectile regression, as a general M smoother, is used to capture the tail behaviour of a distribution. Let $(X_1, Y_1), \ldots, (X_n, Y_n)$ be i.i.d. rvs. Denote by $v(x)$ the unknown τ-expectile regression curve of Y conditional on X, and by $v_n(x)$ its kernel smoothing estimator. In this paper, we prove the strong uniform consistency rate of $v_n(x)$ under general conditions. Moreover, using strong approximations of the empirical process and extreme value theory, we consider the asymptotic maximal deviation $\sup_{0\le x\le 1} |v_n(x) - v(x)|$. According to the asymptotic theory, we construct simultaneous confidence bands around the estimated expectile function. Furthermore, we apply this confidence band to temperature analysis. Taking Berlin and Taipei as an example, we investigate the temperature risk drivers to these two cities.

Keywords Expectile regression · Consistency rate · Simultaneous confidence bands · Asymmetric least squares · Kernel smoothing

1 Introduction

In regression function estimation, most investigations are concerned with the conditional mean. Geometrically, the observations $\{(X_i, Y_i),\ i = 1, \ldots, n\}$ form a cloud of points in a Euclidean space. The mean regression function focuses on the center of the point-cloud, given the covariate X, see Efron (1991). However, more insights

M. Guo (✉)
Institute for Statistics and Econometrics, Humboldt-Universität zu Berlin, Unter den Linden 6, 10099, Berlin, Germany
e-mail: guomengm@cms.hu-berlin.de

W.K. Härdle
C.A.S.E.—Center for Applied Statistics and Economics, Humboldt-Universität zu Berlin, Unter den Linden 6, 10099, Berlin, Germany
e-mail: haerdle@wiwi.hu-berlin.de

Published online: 03 December 2011

about the relation between Y and X can be gained by considering the tails of the conditional distribution.

Asymmetric least squares estimation provides a convenient and relatively efficient method of summarizing the conditional distribution of a dependent variable given the regressors. It turns out that similar to conditional percentiles, the conditional expectiles also characterize the distribution. Breckling and Chambers (1988) proposed M-quantiles, which extend this idea by a "quantile-like" generalization of regression based on asymmetric loss functions. Expectile regression, and more general M-quantile regression, can be used to characterize the relationship between a response variable and explanatory variables when the behaviour of "non-average" individuals is of interest. Jones (1994) described that expectiles and M-quantiles are related to means and quantiles are related to the median, and moreover expectiles are indeed quantiles of a transformed distribution. However, Koenker (2005) pointed out that expectiles have a more global dependence on the form of the distribution.

The expectile curves can be key aspects of inference in various economic problems and are of great interest in practice. Kuan et al. (2009) considered the conditional autoregressive expectile (CARE) model to calculate the VaR. Expectiles are also applied to calculate the expected shortfall in Taylor (2008). Moreover, Schnabel and Eilers (2009a) analyzed the relationship between gross domestic product per capita (GDP) and average life expectancy using expectile curves. Several well-developed methods already existed to estimate expectile curves. Schnabel and Eilers (2009b) combined asymmetric least square and P-splines to calculate a smooth expectile curve. In this paper, we apply the kernel smoothing techniques for the expectile curve, and construct the simultaneous confidence bands for the expectile curve, which describes a picture about the global variability of the estimator.

Let $(X_1, Y_1), \ldots, (X_n, Y_n)$ be i.i.d. rvs. We denote the joint probability density function (pdf) of the rvs is $f(x, y)$, $F(x, y)$ is the joint cumulative distribution function (cdf), conditional pdf is $f(y|x)$, $f(x|y)$ and conditional cdf $F(y|x)$, $F(x|y)$. Further, $x \in J$ with J a possibly infinite interval in $\mathbb{R}^d$ and $y \in \mathbb{R}$. In general, X may be a multivariate covariate.

From an optimization point of view, both quantile and expectile can be expressed as minimum contrast parameter estimators. Define $\rho_\tau(u) = |\mathbf{I}(u \le 0) - \tau||u|$ for $0 < \tau < 1$, then the τth quantile is expressed as $\arg\min_\theta \mathsf{E}\,\rho_\tau(y - \theta)$, where

$$\mathsf{E}\,\rho_\tau(y-\theta) = (1-\tau)\int_{-\infty}^{\theta} |y-\theta|\,dF(y|x) + \tau\int_{\theta}^{\infty} |y-\theta|\,dF(y|x)$$

where θ is the estimator of the τ expectile, and define $\theta \in I$, where the compact set $I \subset \mathbb{R}$. With the interpretation of the contrast function $\rho_\tau(u)$ as the negative log likelihood of asymmetric Laplace distribution, we can see the τth quantile as a quasi maximum estimator in the location model. Changing the loss (contrast) function to

$$\rho_\tau(u) = \big|\mathbf{I}(u \le 0) - \tau\big|u^2, \quad \tau \in (0,1) \tag{1}$$

leads to expectile. Note that for $\tau = \frac{1}{2}$, we obtain the mean respective to the sample average. Putting this into a regression framework, we define the conditional expectile

function (to level τ) as

$$v(x) = \arg\min_{\theta} \mathsf{E}\big\{\rho_\tau(y-\theta)|X=x\big\} \tag{2}$$

Inserting (1) into (2), we obtain the expected loss function:

$$\mathsf{E}\big\{\rho_\tau(y-\theta)|X=x\big\} = (1-\tau)\int_{-\infty}^{\theta}(y-\theta)^2\,dF(y|x) + \tau\int_{\theta}^{\infty}(y-\theta)^2\,dF(y|x) \tag{3}$$

From now on, we silently assume τ is fixed therefore we suppress the explicit notion. Recall that the conditional quantile $l(x)$ at level τ can be considered as

$$l(x) = \inf\big\{y \in \mathbb{R}|F(y|x) \geq \tau\big\}$$

Therefore, the proposed estimate $l_n(x)$ can be expressed:

$$l_n(x) = \inf\big\{y \in \mathbb{R}|\widehat{F}(y|x) \geq \tau\big\}$$

where $\widehat{F}(y|x)$ is the kernel estimator of $F(y|x)$:

$$\widehat{F}(y|x) = \frac{\sum_{i=1}^n K_h(x-X_i)\mathbf{I}(Y_i \leq y)}{\sum_{i=1}^n K_h(x-X_i)}$$

In the same spirit, define $G_{Y|x}(\theta)$ as

$$G_{Y|x}(\theta) = \frac{\int_{-\infty}^{\theta}|y-\theta|\,dF(y|x)}{\int_{-\infty}^{\infty}|y-\theta|\,dF(y|x)}$$

Replacing θ by $v(x)$, we get

$$G_{Y|x}(v) = \frac{\int_{-\infty}^{v(x)}|y-v(x)|\,dF(y|x)}{\int_{-\infty}^{\infty}|y-v(x)|\,dF(y|x)} = \tau$$

so $v(x)$ can be equivalently seen as solving: $G_{Y|x}(\theta) - \tau = 0$ (w.r.t. θ). Therefore,

$$v(x) = G_{Y|x}^{-1}(\tau)$$

with the τth expectile curve kernel smoothing estimator:

$$v_n(x) = \hat{G}_{Y|x}^{-1}(\tau)$$

where the nonparametric estimate of $G_{Y|x}(v)$ is

$$\hat{G}_{Y|x}(\theta) = \frac{\sum_{i=1}^n K_h(x-X_i)\,\mathbf{I}(Y_i < y)|y-\theta|}{\sum_{i=1}^n K_h(x-X_i)|y-\theta|}$$

Quantiles and expectiles both characterize a distribution function although they are different in nature. As an illustration, Fig. 1 plots curves of quantiles and expectiles

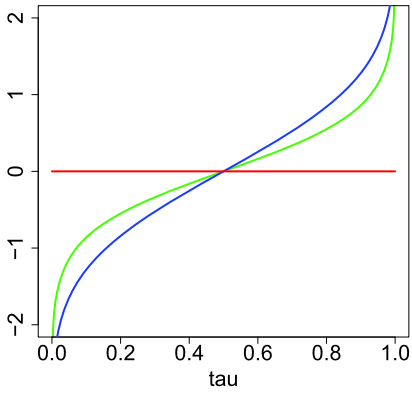


Fig. 1 (Color online) Quantile curve (*blue*) and expectile curve (*green*) for standard normal distribution

of the standard normal $N(0,1)$. Obviously, there is a one-to-one mapping between quantile and expectile, see Yao and Tong (1996). For fixed x, define $w(\tau)$ such that $v_{w(\tau)}(x) = l(x)$, then $w(\tau)$ is related to the τth quantile curve $l(x)$ via

$$w(\tau) = \frac{\tau l(x) - \int_{-\infty}^{l(x)} y\, dF(y|x)}{2\,\mathsf{E}(Y|x) - 2\int_{-\infty}^{l(x)} y\, dF(y|x) - (1-2\tau)l(x)} \tag{4}$$

$l(x)$ is an increasing function of τ, therefore, $w(\tau)$ is also a monotonically increasing function. Expectiles correspond to quantiles with this transformation w. However, it is not straightforward to apply (4), since it depends on the conditional distribution of the regressors. For very simple distributions, it is not hard to calculate the transformation $w(\tau)$, for example, $Y \sim U(-1,1)$, then $w(\tau) = \tau^2/(2\tau^2 - 2\tau + 1)$. However, if the distribution is more complicated, even worse, the conditional distribution is unknown, it is hard to apply this transformation, see Jones (1994). Therefore, it is not feasible to calculate expectiles from the corresponding quantiles.

In the current paper, we apply the methodology to weather studies. Weather risk is an uncertainty caused by weather volatility. Energy companies take positions in weather risk if it is a source of financial uncertainty. However, weather is also a local phenomenon, since the location, the atmosphere, human activities and some other factors influence the temperature. We investigate whether such local factors exist. Taking two cities, Berlin and Taipei, as an example, we check whether the performance of high expectiles and low expectiles of temperature varies over time. To this end, we calculate the expectiles of trend and seasonality corrected temperature.

The structure of this paper is as follows. In Sect. 2, the stochastic fluctuation of the process $\{v_n(x) - v(x)\}$ is studied and the simultaneous confidence bands are presented through the equivalence of several stochastic processes. We calculate the asymptotic distribution of $v_n(x)$, and the strong uniform consistency rate of $\{v_n(x) - v(x)\}$ is discussed in this section. In Sect. 3, a Monte Carlo study is to investigate the behaviour of $v_n(x)$ when the data are generated with the error terms standard normally distributed. Section 4 considers an application in the temperature of Berlin and Taipei. All proofs are attached in Appendix.

2 Results

In light of the concepts of M-estimation as in Huber (1981), if we define $\psi(u)$ as

$$\begin{aligned}\psi(u) &= \frac{\partial \rho(u)}{\partial u} \\ &= \big|\mathbf{I}(u \le 0) - \tau\big| u \\ &= \big\{\tau - \mathbf{I}(u \le 0)\big\}|u|\end{aligned}$$

$v_n(x)$ and $v(x)$ can be treated as a zero (w.r.t. θ) of the function:

$$H_n(\theta, x) \stackrel{\text{def}}{=} n^{-1} \sum_{i=1}^{n} K_h(x - X_i)\psi(Y_i - \theta) \tag{5}$$

$$H(\theta, x) \stackrel{\text{def}}{=} \int_{\mathbb{R}} f(x, y)\psi(y - \theta)\, dy \tag{6}$$

respectively.

Härdle (1989) has constructed the uniform confidence bands for general M-smoothers. Härdle and Song (2009) studied the uniform confidence bands for quantile curves. In our paper, we investigate expectile curves, one kind of M-smoother. The loss function for quantile regression is not differentiable, however it is differentiable for expectile when it is in the asymmetric quadratic form. Therefore, by employing similar methods as those developed in Härdle (1989), it is shown in this paper that

$$\begin{aligned}&\mathrm{P}\Big[(2\delta \log n)^{1/2}\Big\{\sup_{x \in J} r(x)\big|v_n(x) - v(x)\big|/\lambda(K)^{1/2} - d_n\Big\} < z\Big] \\ &\quad \longrightarrow \quad \exp\big\{-2\exp(-z)\big\}, \quad \text{as } n \to \infty\end{aligned} \tag{7}$$

with some adjustment of $v_n(x)$, we can see that the supreme of $v_n(x) - v(x)$ follows the asymptotic Gumbel distribution, where $r(x)$, δ, $\lambda(K)$, d_n are suitable scaling parameters. The asymptotic result (7) therefore allows the construction of simultaneous confidence bands for $v(x)$ based on specifications of the stochastic fluctuation of $v_n(x)$. The strong approximation with Brownian bridge techniques is applied in this paper to prove the asymptotic distribution of $v_n(x)$.

To construct the confidence bands, we make the following necessary assumptions about the distribution of (X, Y) and the score function $\psi(u)$ in addition to the existence of an initial estimator whose error is a.s. uniformly bounded.

(A1) The kernel $K(\cdot)$ is positive, symmetric, has compact support $[-A, A]$ and is Lipschitz continuously differentiable with bounded derivatives.

(A2) $(nh)^{-1/2}(\log n)^{3/2} \to 0$, $(n \log n)^{1/2}h^{5/2} \to 0$, $(nh^3)^{-1}(\log n)^2 \le M$, M is a constant.

(A3) $h^{-3}(\log n) \int_{|y|>a_n} f_Y(y)\, dy = \mathcal{O}(1)$, $f_Y(y)$ the marginal density of Y, $\{a_n\}_{n=1}^{\infty}$ a sequence of constants tending to infinity as $n \to \infty$.

(A4) $\inf_{x \in J} |p(x)| \ge p_0 > 0$, where $p(x) = \partial\, \mathsf{E}\{\psi(Y - \theta)|x\}/\partial\theta|_{\theta=v(x)} \cdot f_X(x)$, where $f_X(x)$ is the marginal density of X.

(A5) The expectile function $v(x)$ is Lipschitz twice continuously differentiable, for all $x \in J$.

(A6) $0 < m_1 \leq f_X(x) \leq M_1 < \infty$, $x \in J$, and the conditional density $f(\cdot|y)$, $y \in \mathbb{R}$, is uniform locally Lipschitz continuous of order $\tilde{\alpha}$ (ulL-$\tilde{\alpha}$) on J, uniformly in $y \in \mathbb{R}$, with $0 < \tilde{\alpha} \leq 1$, and $\psi(x)$ is piecewise twice continuously differentiable.

Define also

$$\sigma^2(x) = \mathsf{E}\big[\psi^2\big\{Y - v(x)\big\}|x\big]$$

$$H_n(x) = (nh)^{-1} \sum_{i=1}^{n} K\big\{(x - X_i)/h\big\}\psi\big\{Y_i - v(x)\big\}$$

$$D_n(x) = (nh)^{-1} \frac{\partial \sum_{i=1}^{n} K\{(x - X_i)/h\}\psi\{Y_i - \theta\}}{\partial\theta}\bigg|_{\theta=v(x)}$$

and assume that $\sigma^2(x)$ and $f_X(x)$ are differentiable.

Assumption (A1) on the compact support of the kernel could possibly be relaxed by introducing a cutoff technique as in Csörgö and Hall (1982) for density estimators. Assumption (A2) has purely technical reasons: to keep the bias at a lower rate than the variance and to ensure the vanishing of some non-linear remainder terms. Assumption (A3) appears in a somewhat modified form also in Johnston (1982). Assumption (A4) guarantees that the first derivative of the loss function, i.e. $\psi(u)$ is differentiable. Assumptions (A5) and (A6) are common assumptions in robust estimation as in Huber (1981), Härdle et al. (1988) that are satisfied by exponential, and generalized hyperbolic distributions.

Zhang (1994) has proved the asymptotic normality of the nonparametric expectile. Under the Assumptions (A1) to (A4), we have

$$\sqrt{nh}\big\{v_n(x) - v(x)\big\} \xrightarrow{\mathcal{L}} \mathrm{N}\big\{0, V(x)\big\} \tag{8}$$

with

$$V(x) = \lambda(K) f_X(x)\sigma^2(x)/p(x)^2$$

where we can denote

$$\begin{aligned}
\lambda(K) &= \int_{-A}^{A} K^2(u)\,du \\
\sigma^2(x) &= \mathsf{E}\big[\psi^2\big\{Y - v(x)\big\}|x\big] \\
&= \int \psi^2\big\{y - v(x)\big\}\,dF(y|x) \\
&= \tau^2 \int_{v(x)}^{\infty} \big\{y - v(x)\big\}^2\,dF(y|x) + (1-\tau)^2 \int_{-\infty}^{v(x)} \big\{y - v(x)\big\}^2\,dF(y|x)
\end{aligned} \tag{9}$$

$$p(x) = \mathsf{E}\big[\psi'\big\{Y - v(x)\big\}|x\big] \cdot f_X(x)$$
$$= \left\{\tau \int_{v(x)}^{\infty} dF(y|x) + (1-\tau)\int_{-\infty}^{v(x)} dF(y|x)\right\} \cdot f_X(x) \tag{10}$$

For the uniform strong consistency rate of $v_n(x) - v(x)$, we apply the result of Härdle et al. (1988) by taking $\beta(y) = \psi(y-\theta)$, $y \in \mathbb{R}$, for $\theta \in I$, $q_1 = q_2 = -1$, $\gamma_1(y) = \max\{0, -\psi(y-\theta)\}$, $\gamma_2(y) = \min\{0, -\psi(y-\theta)\}$ and $\lambda = \infty$ to satisfy the representations for the parameters there. We have the following lemma under some specified assumptions:

Lemma 1 *Let $H_n(\theta, x)$ and $H(\theta, x)$ be given by (5) and (6). Under Assumption* (A6) *and $(nh/\log n)^{1/2} \to \infty$ through Assumption* (A2), *for some constant A^* not depending on n, we have a.s. as $n \to \infty$*

$$\sup_{\theta\in I}\sup_{x\in J}\big|H_n(\theta, x) - H(\theta, x)\big| \le A^* \max\big\{(nh/\log n)^{-1/2}, h^{\tilde{\alpha}}\big\} \tag{11}$$

For our result on $v_n(\cdot)$, we shall also require

$$\inf_{x\in J}\left|\int \psi\big\{y - v(x) + \varepsilon\big\}\, dF(y|x)\right| \ge \tilde{q}|\varepsilon|, \quad \text{for } |\varepsilon| \le \delta_1 \tag{12}$$

where δ_1 and $\tilde{q}$ are some positive constants, see also Härdle and Luckhaus (1984). This assumption is satisfied if there exists a constant $\tilde{q}$ such that $f\{v(x)|x\} > \tilde{q}/p$, $x \in J$.

Theorem 1 *Under the conditions of Lemma 1 and also assuming (12) holds, we have a.s. as $n \to \infty$*

$$\sup_{x\in J}\big|v_n(x) - v(x)\big| \le B^* \max\big\{(nh/\log n)^{-1/2}, h^{\tilde{\alpha}}\big\} \tag{13}$$

with $B^* = A^*/m_1\tilde{q}$ not depending on n and m_1 a lower bound of $f_X(x)$. If additionally $\tilde{\alpha} \ge \{\log(\sqrt{\log n}) - \log(\sqrt{nh})\}/\log h$, it can be further simplified to

$$\sup_{x\in J}\big|v_n(x) - v(x)\big| \le B^*\big\{(nh/\log n)^{-1/2}\big\}$$

Theorem 2 *Let $h = n^{-\delta}$, $\frac{1}{5} < \delta < \frac{1}{3}$ with $\lambda(K)$ as defined before, and*

$$d_n = (2\delta\log n)^{1/2} + (2\delta\log n)^{-1/2}\left[\log\big\{c_1(K)/\pi^{1/2}\big\} + \frac{1}{2}(\log\delta + \log\log n)\right]$$
$$\text{if } c_1(K) = \big\{K^2(A) + K^2(-A)\big\}/\big\{2\lambda(K)\big\} > 0$$
$$d_n = (2\delta\log n)^{1/2} + (2\delta\log n)^{-1/2}\log\big\{c_2(K)/2\pi\big\}$$
$$\text{otherwise with } c_2(K) = \int_{-A}^{A}\big\{K'(u)\big\}^2\, du/\big\{2\lambda(K)\big\}$$

Then (7) *holds with*

$$r(x) = (nh)^{-\frac{1}{2}} p(x) \left\{ \frac{f_X(x)}{\sigma^2(x)} \right\}^{\frac{1}{2}}$$

This theorem can be used to construct uniform confidence intervals for the regression function as stated in the following corollary.

Corollary 1 *Under the assumptions of the theorem above, an approximate* $(1-\alpha) \times 100\%$ *confidence band over* $[0, 1]$ *is*

$$v_n(x) \pm (nh)^{-1/2} \left\{ \hat{\sigma}^2(x) \lambda(K) / \hat{f}_X(x) \right\}^{1/2} \hat{p}^{-1}(x) \left\{ d_n + c(\alpha)(2\delta \log n)^{-1/2} \right\}$$

where $c(\alpha) = \log 2 - \log |\log(1-\alpha)|$ *and* $\hat{f}_X(x)$, $\hat{\sigma}^2(x)$ *and* $\hat{p}(x)$ *are consistent estimates for* $f_X(x)$, $\sigma^2(x)$ *and* $p(x)$.

With $\sqrt{V(x)}$ introduced, we can further write Corollary 1 as

$$v_n(x) \pm (nh)^{-1/2} \left\{ d_n + c(\alpha)(2\delta \log n)^{-1/2} \right\} \sqrt{\hat{V}(x)}$$

where $\hat{V}(x)$ is the nonparametric estimator of $V(x)$. Bandwidth selection is quite crucial in kernel smoothing. In this paper, we use the optimal bandwidth discussed in Zhang (1994), which has the following form

$$h_n^{\text{opt}} = \left(\frac{\sigma^2(x) \lambda(K)}{n[\Lambda\{v(x)|x\}]^2 [\int \{y - v(x)\}^2 K^2\{y - v(x)\} \, dF(y|x)]^2} \right)^{1/5} \tag{14}$$

where

$$\Lambda(\theta|x) = \frac{\partial^2 \psi(\theta|x-u)}{\partial u^2}|_{u=0}$$

The proof is essentially based on a linearization argument after a Taylor series expansion. The leading linear term will then be approximated in a similar way as in Johnston (1982), Bickel and Rosenblatt (1973). The main idea behind the proof is a strong approximation of the empirical process of $\{(X_i, Y_i)_{i=1}^n\}$ by a sequence of Brownian bridges as proved by Tusnady (1977).

As $v_n(x)$ is the zero (w.r.t. θ) of $H_n(\theta, x)$, it follows by applying second-order Taylor expansions to $H_n(\theta, x)$ around $v(x)$ that

$$v_n(x) - v(x) = \left\{ H_n(x) - \mathsf{E}\, H_n(x) \right\} \big/ p(x) + R_n(x) \tag{15}$$

where $\{H_n(x) - \mathsf{E}\, H_n(x)\}/p(x)$ is the leading linear term and the remainder term is written as

$$\begin{aligned} R_n(x) = {} & H_n(x) \left\{ p(x) - D_n(x) \right\} \big/ \left\{ D_n(x) \cdot p(x) \right\} + \mathsf{E}\, H_n(x)/p(x) \\ & + \frac{1}{2} \left\{ v_n(x) - v(x) \right\}^2 \cdot \left\{ D_n(x) \right\}^{-1} \end{aligned} \tag{16}$$

$$\cdot (nh)^{-1} \sum_{i=1}^{n} K\{(x - X_i)/h\} \psi''\{Y_i - v(x) + r_n(x)\}, \tag{17}$$

$$|r_n(x)| < |v_n(x) - v(x)|.$$

We show in Appendix that (Lemma 4) that $\|R_n\| = \sup_{x \in J} |R_n(x)| = \mathcal{O}_p\{(nh \log n)^{-1/2}\}$.

Furthermore, the rescaled linear part

$$Y_n(x) = (nh)^{1/2} \{\sigma^2(x) f_X(x)\}^{-1/2} \{H_n(x) - \mathsf{E}\, H_n(x)\}$$

is approximated by a sequence of Gaussian processes, leading finally to the Gaussian process

$$Y_{5,n}(x) = h^{-1/2} \int K\{(x - t)/h\}\, dW(x) \tag{18}$$

Drawing upon the result of Bickel and Rosenblatt (1973), we finally obtain asymptotically the Gumbel distribution.

We also need the Rosenblatt (1952) transformation,

$$T(x, y) = \{F_{X|y}(x|y), F_Y(y)\}$$

which transforms (X_i, Y_i) into $T(X_i, Y_i) = (X_i', Y_i')$ mutually independent uniform rv's. In the event that x is a d-dimension covariate, the transformation becomes

$$T(x_1, x_2, \ldots, x_d, y) = \{F_{X_1|y}(x_1|y), F_{X_2|y}(x_2|x_1, y), \ldots, F_{X_k|x_{d-1},\ldots,x_1,y}(x_k|x_{d-1}, \ldots, x_1, y), F_Y(y)\} \tag{19}$$

With the aid of this transformation, Theorem 1 of Tusnady (1977) may be applied to obtain the following lemma.

Lemma 2 *On a suitable probability space a sequence of Brownian bridges B_n exists that*

$$\sup_{x \in J, y \in \mathbb{R}} |Z_n(x, y) - B_n\{T(x, y)\}| = \mathcal{O}\{n^{-1/2} (\log n)^2\} \quad a.s.$$

where $Z_n(x, y) = n^{1/2}\{F_n(x, y) - F(x, y)\}$ denotes the empirical process of $\{(X_i, Y_i)\}_{i=1}^n$.

For $d > 2$, it is still an open problem which deserves further research.

Before we define the different approximating processes, let us first rewrite (18) as a stochastic integral w.r.t. the empirical process $Z_n(x, y)$,

$$Y_n(x) = \{hg'(x)\}^{-1/2} \iint K\{(x - t)/h\} \psi\{y - v(x)\}\, dZ_n(t, y)$$

$$g'(x) = \sigma^2(x) f_X(x)$$

The approximating processes are now

$$Y_{0,n}(x) = \{hg(x)\}^{-1/2} \iint_{\Gamma_n} K\{(x-t)/h\}\psi\{y - v(x)\}\, dZ_n(t, y)$$

where $\Gamma_n = \{|y| \le a_n\}$,

$$g(t) = \mathsf{E}\big[\psi^2\{y - v(x)\} \cdot \mathbf{I}(|y| \le a_n)\big|X = x\big] \cdot f_X(x) \tag{20}$$

$$Y_{1,n}(x) = \{hg(x)\}^{-1/2} \iint_{\Gamma_n} K\{(x-t)/h\}\psi\{y - v(x)\}\, dB_n\{T(t, y)\}$$

$\{B_n\}$ being the sequence of Brownian bridges from Lemma 2 (21)

$$Y_{2,n}(x) = \{hg(x)\}^{-1/2} \iint_{\Gamma_n} K\{(x-t)/h\}\psi\{y - v(x)\}\, dW_n\{T(t, y)\}$$

$\{W_n\}$ being the sequence of Wiener processes satisfying

$$B_n(t', y') = W_n(t', y') - t'y'W_n(1, 1) \tag{22}$$

$$Y_{3,n}(x) = \{hg(x)\}^{-1/2} \iint_{\Gamma_n} K\{(x-t)/h\}\psi\{y - v(t)\}\, dW_n\{T(t, y)\} \tag{23}$$

$$Y_{4,n}(x) = \{hg(x)\}^{-1/2} \int g(t)^{1/2} K\{(x-t)/h\}\, dW(t) \tag{24}$$

$$Y_{5,n}(x) = h^{-1/2} \int K\{(x-t)/h\}\, dW(t)$$

$\{W(\cdot)\}$ being the Wiener process (25)

Lemmas 5 to 10 ensure that all these processes have the same limit distributions. The result then follows from

Lemma 3 (Theorem 3.1 in Bickel and Rosenblatt 1973) *Let d_n, $\lambda(K)$, δ as in Theorem 2. Let*

$$Y_{5,n}(x) = h^{-1/2} \int K\{(x-t)/h\}\, dW(t)$$

Then, as $n \to \infty$, the supremum of $Y_{5,n}(x)$ has a Gumbel distribution.

$$\mathrm{P}\Big\{(2\delta \log n)^{1/2}\Big[\sup_{x\in J}|Y_{5,n}(x)|/\{\lambda(K)\}^{1/2} - d_n\Big] < z\Big\} \to \exp\{-2\exp(-z)\}$$

Same as quantile, the supremum of a nonparametric expectile converge to its limit at a rate $(\log n)^{-1}$. We do not check the bootstrap confidence bands in this paper, which can be future work. Instead, we point out several well documented literature about this issue. For example, Claeskens and Keilegom (2003) discussed the bootstrap confidence bands for regression curves and their derivatives. Partial linear quantile regression and bootstrap confidence bands are well studied in Härdle et al. (2010). They proved that the convergence rate by bootstrap approximation to the dis-

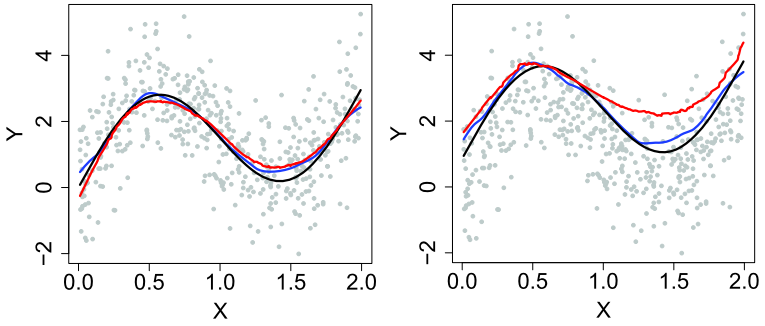


Fig. 2 (Color online) $\tau = 0.5$ (*left*) and $\tau = 0.9$ (*right*) estimated quantile and expectile plot. Quantile curve, theoretical expectile curve, estimated expectile curve

tribution of the supremum of a quantile estimate has been improved from $(\log n)^{-1}$ to $n^{-2/5}$.

3 A Monte Carlo study

In the design of the simulation, we generate bivariate random variables $\{(X_i, Y_i)\}_{i=1}^{n}$ with sample size $n = 50, n = 100, n = 200, n = 500$. The covariate X is uniformly distributed on $[0, 2]$

$$Y = 1.5X + 2\sin(\pi X) + \varepsilon \tag{26}$$

where $\varepsilon \sim \mathrm{N}(0, 1)$.

Obviously, the theoretical expectiles (fixed τ) are determined by

$$v(x) = 1.5x + 2\sin(\pi x) + v_N(\tau) \tag{27}$$

where $v_N(\tau)$ is the τth expectile of the standard Normal distribution.

Figure 2 (in the left part) describes the simulated data (the grey points), together with the 0.5 estimated quantile and estimated expectile and theoretical expectile curves, which represents, respectively, the conditional median and conditional mean. The conditional mean and conditional median coincide with each other, since the error term is symmetrically distributed, which is obvious in Fig. 2. In the right part of the figure, we consider the conditional 0.9 quantile and expectile curves. Via a transformation (4), there is a gap between the quantile curve and the expectile curve. By calculating $w(\tau)$ for the standard normal distribution, the 0.9 quantile can be expressed by the around 0.96 expectile. The estimated expectile curve is close to the theoretical one.

Figure 3 shows the 95% uniform confidence bands for expectile curve, which are represented by the two red dashed lines. We calculate both 0.1 (left) and 0.9 (right) expectile curves. The black lines stand for the corresponding 0.1 and 0.9 theoretical expectile curves, and the blue lines are the estimated expectile curves. Obviously, the theoretical expectile curves locate in the confidence bands.

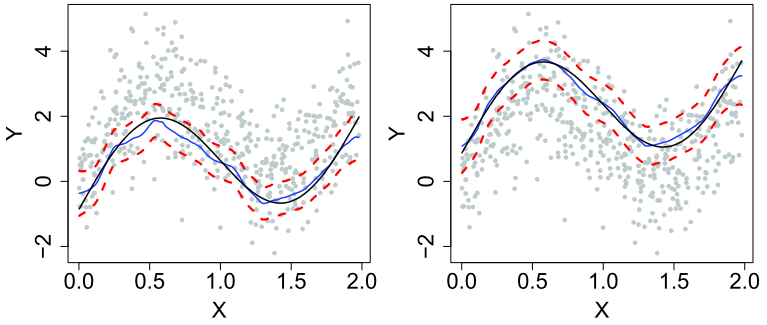


Fig. 3 Uniform confidence bands for expectile curve for $\tau = 0.1$ (*left*) and $\tau = 0.9$ (right). Theoretical expectile curve, estimated expectile curve and 95% uniform confidence bands

Table 1 Simulated coverage probabilities of 95% confidence bands for 0.9 expectile with 500 runs of simulation. *cp* stands for the coverage probability, and *h* is the width of the band

n	*cp*	*h*
50	0.526	1.279
100	0.684	1.093
200	0.742	0.897
500	0.920	0.747

Table 2 Simulated coverage probabilities of 95% confidence bands for 0.1 expectile with 500 runs of simulation. *cp* stands for the coverage probability, and *h* is the width of the band

n	*cp*	*h*
50	0.386	0.859
100	0.548	0.768
200	0.741	0.691
500	0.866	0.599

To check the performance of the calculated confidence bands, we compare the simulated coverage probability with the nominal values for coverage probability 95% for different sample sizes. We apply this method to both 0.9 and 0.1 expectile. Table 1 and Table 2 present the corresponding results. We run the simulation 500 times for each scenario. Obviously, the coverage probabilities improve with the increased the sample size, and the width of the bands h becomes smaller for both 0.9 and 0.1 expectile. It is noteworthy that when the number of observation is large enough, for example $n = 500$, the coverage probability is very close to the nominal probability, especially for the 0.9 expectile.

4 Application

In this part, we apply the expectile into the temperature study. We consider the daily temperature both of Berlin and Taipei, ranging from 19480101 to 20071231, together 21900 observations for each city. The statistical properties of the temperature are

Table 3 Statistical summary of the temperature in Berlin and Taipei

	Mean	SD	Skewness	Kurtosis	Max	Min
Berlin	9.66	7.89	−0.315	2.38	30.4	−18.5
Taipei	22.61	5.43	−0.349	2.13	33.0	6.5

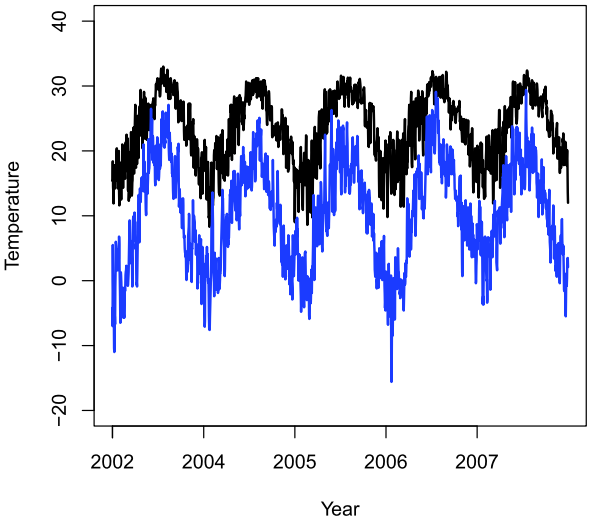


Fig. 4 (Color online) The time series plot of the temperature in Berlin and Taipei from 2002–2007. *The black line* stands for the temperature in Taipei, and *the blue line* is in Berlin

summarized in Table 3. The Berlin temperature data were obtained from Deutscher Wetterdienst, and the Taipei temperature data were obtained from the center for adaptive data analysis in National Central University.

Before proceeding to detailed modeling and forecasting results, it is useful to get an overall view of the daily average temperature data. Figure 4 displays the average temperature series of the sample from 2002 to 2007. The black line stands for the temperature in Taipei, and the blue line describes for the temperature in Berlin. The time series plots reveal strong and unsurprising seasonality in average temperature: in each city, the daily average temperature moves repeatedly and regularly through periods of high temperature (summer) and low temperature (winter). It is well documented that seasonal volatility in the regression residuals appears highest during the winter months where the temperature shows high volatility. Importantly, however, the seasonal fluctuations differ noticeably across cities both in terms of amplitude and detail of pattern.

Based on the observed pattern, we apply a stochastic model with seasonality and inter temporal autocorrelation, as in Benth et al. (2007). To understand the model clearly, let us introduce the time series decomposition of the temperature, with $t =$

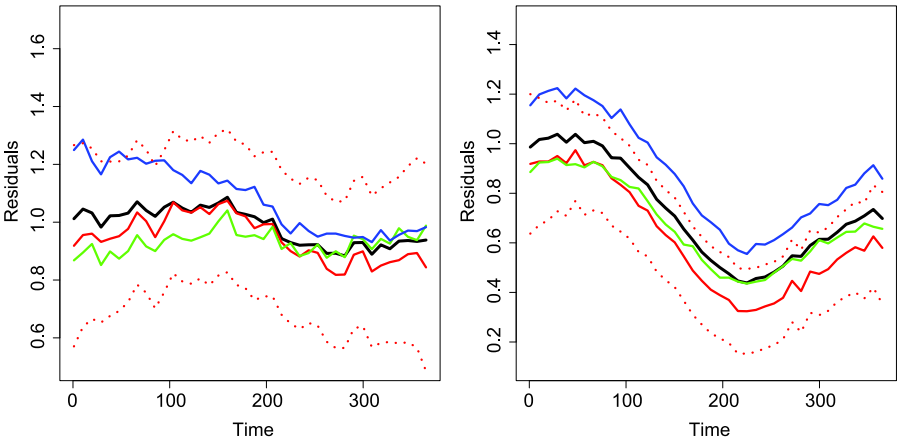


Fig. 5 0.9 expectile curves for Berlin (*left*) and Taipei (*right*) daily temperature residuals from 1948–2007 with the 95% uniform confidence bands for the first 20 years expectile

$1, \ldots, 365$ days, and $j = 0, \ldots, J$ years:

$$\begin{aligned} X_{365j+t} &= T_{t,j} - \Lambda_t \\ X_{365j+t} &= \sum_{l=1}^{L} \beta_{lj} X_{365j+t-l} + \varepsilon_{t,j} \\ \Lambda_t &= a + bt + \sum_{m=1}^{M} c_l \cos\left\{\frac{2\pi(t-d_m)}{l \cdot 365}\right\} \end{aligned} \tag{28}$$

where $T_{t,j}$ is the temperature at day t in year j, and Λ_t denotes the seasonality effect. Motivation of this modeling approach can be found in Diebold and Inoue (2001). Further studies as Campbell and Diebold (2005) has provided evidence that the parameters β_{lj} are likely to be j independent and hence estimated consistently from a global autoregressive process $AR(L_j)$ model with $L_j = L$. The analysis of the partial autocorrelations and Akaike's Information Criterion (AIC) suggests that a simple $AR(3)$ model fits well the temperature evolution both in Berlin and Taipei.

In this paper, the risk factor of temperature, which is the residual $\hat{\varepsilon}_{t,j}$ from (28), is studied in the expectile regression. We intend to construct the confidence bands for the 0.01 and 0.9 expectile curves for the volatility of temperature. It is interesting to check whether the extreme values perform differently in different cities.

The left part of the figures describes the expectile curves for Berlin, and the right part is for Taipei. In each figure, the thick black line depicts the average expectile curve with the data from 1948 to 2007. The red line is the expectile for the residuals from (28) with the data of the first 20 years temperature, i.e. in the period from 1948 to 1967. The 0.9 expectile for the second 20 years (1968–1987) residuals is described by the green line, and the blue line stands for the expectile curve in the latest 20 years (1988–2007). The dotted lines are the 95% confidence bands corresponding to the expectile curve with the same color. Figures 5, 6 and 7 describe the 0.9 expectile curves

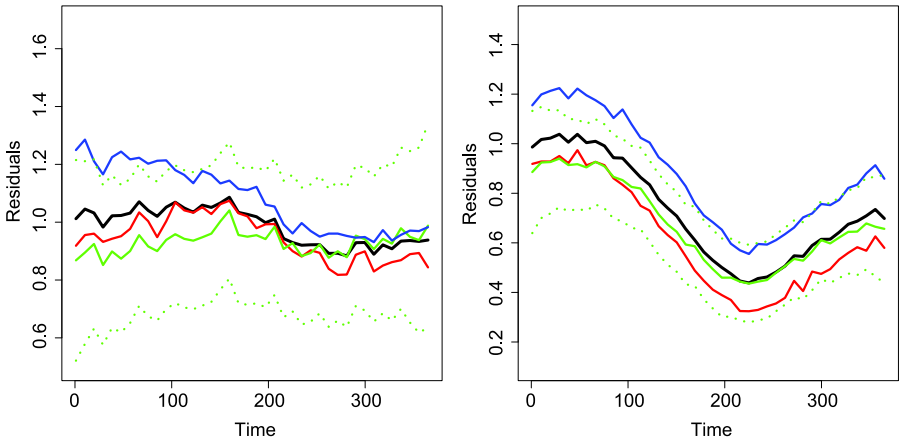


Fig. 6 0.9 expectile curves for Berlin (*left*) and Taipei (*right*) daily temperature residuals from 1948–2007 with the 95% uniform confidence bands for the second 20 years expectile

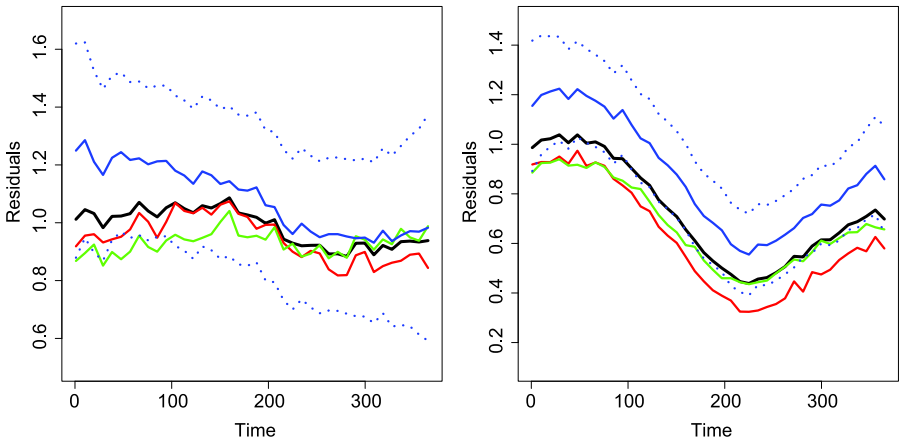


Fig. 7 0.9 expectile curves for Berlin (*left*) and Taipei (*right*) daily temperature residuals from 1948–2007 with the 95% uniform confidence bands for the latest 20 years expectile

for Berlin and Taipei, as well as their corresponding confidence bands. Obviously, the variance is higher in winter–earlier summer both in Berlin and Taipei.

Note that the behaviour of expectile curves in Berlin and Taipei is quite different. Firstly, the variation of the expectiles in Berlin is smaller than that of Taipei. All the expectile curves cross with each other in the last 100 observations of the year for Berlin, and the variance in this period is smaller. Moreover, all of these curves nearly locate in the corresponding three confidence bands. However, the performance of the expectile in Taipei is quite different from that of Berlin. The expectile curves for Taipei have similar trends for each 20 years. They have highest volatilities in January, and lowest volatility in July. More interestingly, the expectile curve for the latest 20 years does not locate in the confidence bands constructed using the data from the

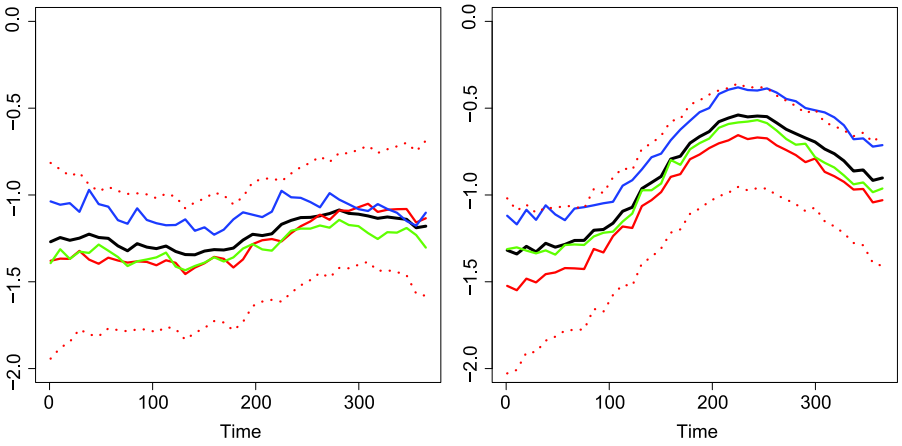


Fig. 8 0.01 expectile curves for Berlin (*left*) and Taipei (*right*) daily temperature residuals from 1948–2007 with the 95% uniform confidence bands for the first 20 years expectile

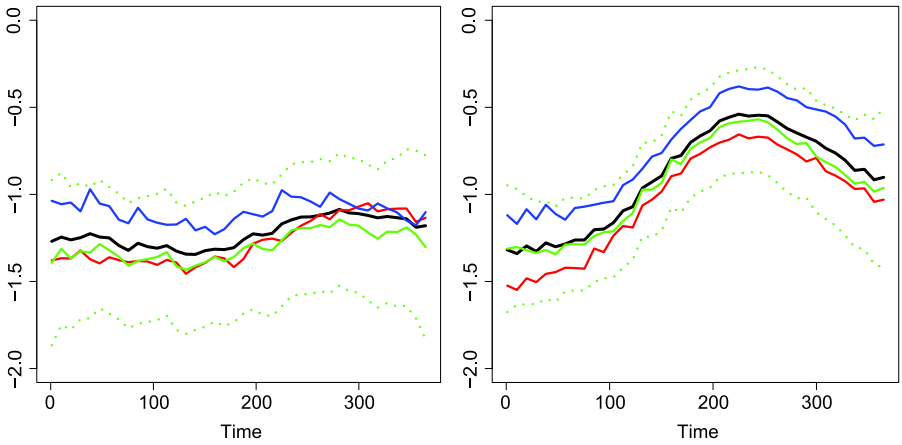


Fig. 9 0.01 expectile curves for Berlin (*left*) and Taipei (*right*) daily temperature residuals from 1948–2007 with the 95% uniform confidence bands for the second 20 years expectile

first 20 years and second 20 years, see Figs. 5 and 7. Similarly, the expectile curve for the first 20 years does not locate in the confidence bands constructed using the information from the latest 20 years.

Further, let us study low expectile for the residuals of the temperature in Berlin and Taipei. It is hard to calculate very small percentage of quantile curves, due to the sparsity of the data, expectiles though can overcome this drawback. One can calculate very low or very high expectiles, such as 0.01 and 0.99 expectile curves, even when there are not so many observations. Display of the 0.01 expectiles for the residuals and their corresponding confidence bands is given in Figs. 8, 9 and 10. One can detect that the shapes of the 0.01 expectile for Berlin and Taipei are different. It does not fluctuate a lot during the whole year in Berlin, while the variation in Taipei

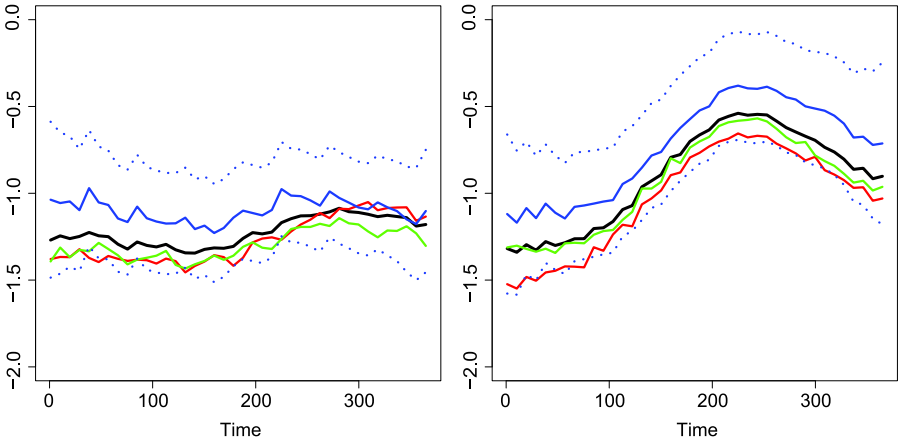


Fig. 10 0.01 expectile curves for Berlin (*left*) and Taipei (*right*) daily temperature residuals from 1948–2007 with the 95% uniform confidence bands for the latest 20 years expectile

is much bigger. However, all the curves both for Berlin and Taipei locate in their corresponding confidence bands.

As depicted in the figures, the performance of the residuals are quite different from Berlin and Taipei, especially for high expectiles. The variation of the temperature in Taipei is more volatile. One interpretation is that in the last 60 years, Taiwan has been experiencing a fast developing period. Industrial expansion, burning of fossil fuel and deforestation and other sectors, could be an important factor for the bigger volatility in the temperature of Taipei. However, Germany is well-developed in this period, especially in Berlin, where there are no intensive industries. Therefore, one may say the residuals reveals the influence of the human activities, which induce the different performance of the residuals of temperature.

Acknowledgements The financial support from the Deutsche Forschungsgemeinschaft via SFB 649 "Ökonomisches Risiko", Humboldt-Universität zu Berlin is gratefully acknowledged. China Scholarship Council (CSC) is gratefully acknowledged.

Appendix

Proof of Theorem 1 By the definition of $v_n(x)$ as a zero of (5), we have, for $\varepsilon > 0$,

$$\text{if } v_n(x) > v(x) + \varepsilon, \quad \text{and} \quad \text{then } H_n\big\{v(x) + \varepsilon, x\big\} > 0 \tag{29}$$

Now

$$H_n\big\{v(x) + \varepsilon, x\big\} \le H\big\{v(x) + \varepsilon, x\big\} + \sup_{\theta \in I} \big|H_n(\theta, x) - H(\theta, x)\big| \tag{30}$$

Also, by the identity $H\{v(x), x\} = 0$, the function $H\{v(x) + \varepsilon, x\}$ is not positive and has a magnitude $\ge m_1 \tilde{q} \varepsilon$ by assumption (A6) and (12), for $0 < \varepsilon < \delta_1$. That is, for

$0 < \varepsilon < \delta_1$,

$$H\{v(x) + \varepsilon, x\} \leq -m_1 \tilde{q} \varepsilon \tag{31}$$

Combining (29), (30) and (31), we have, for $0 < \varepsilon < \delta_1$:

$$\text{if } v_n(x) > v(x) + \varepsilon, \quad \text{and} \quad \text{then } \sup_{\theta \in I} \sup_{x \in J} \left| H_n(\theta, x) - H(\theta, x) \right| > m_1 \tilde{q} \varepsilon$$

With a similar inequality proved for the case $v_n(x) < v(x) + \varepsilon$, we obtain, for $0 < \varepsilon < \delta_1$:

$$\text{if } \sup_{x \in J} \left| v_n(x) - v(x) \right| > \varepsilon, \quad \text{and} \quad \text{then } \sup_{\theta \in I} \sup_{x \in J} \left| H_n(\theta, x) - H(\theta, x) \right| > m_1 \tilde{q} \varepsilon \tag{32}$$

It readily follows that (32) and (11) imply (13). □

Below we first show that $\|R_n\|_\infty = \sup_{x \in J} |R_n(x)|$ vanishes asymptotically faster than the rate $(nh \log n)^{-1/2}$; for simplicity we will just use $\|\cdot\|$ to indicate the sup-norm.

Lemma 4 *For the remainder term $R_n(t)$ defined in (16) we have*

$$\|R_n\| = \mathcal{O}_p\left\{(nh \log n)^{-1/2}\right\} \tag{33}$$

Proof First we have by the positivity of the kernel K,

$$\|R_n\| \leq \left[\inf_{0 \leq x \leq 1} \left\{ \left| D_n(x) \right| \cdot p(x) \right\} \right]^{-1} \left\{ \|H_n\| \cdot \|p - D_n\| + \|D_n\| \cdot \|\mathsf{E} H_n\| \right\}$$
$$+ C_1 \cdot \|v_n - l\|^2 \cdot \left\{ \inf_{0 \leq t \leq 1} \left| D_n(x) \right| \right\}^{-1} \cdot \|f_n\|$$

where $f_n(x) = (nh)^{-1} \sum_{i=1}^n K\{(x - X_i)/h\}$.

The desired result (4) will then follow if we prove

$$\|H_n\| = \mathcal{O}_p\left\{(nh)^{-1/2} (\log n)^{1/2}\right\} \tag{34}$$

$$\|p - D_n\| = \mathcal{O}_p\left\{(nh)^{-1/4} (\log n)^{-1/2}\right\} \tag{35}$$

$$\|\mathsf{E} H_n\| = \mathcal{O}\left(h^2\right) \tag{36}$$

$$\|v_n - v\|^2 = \mathcal{O}_p\left\{(nh)^{-1/2} (\log n)^{-1/2}\right\} \tag{37}$$

Since (36) follows from the well-known bias calculation

$$\mathsf{E} H_n(x) = h^{-1} \int K\left\{(x - u)/h\right\} \mathsf{E}\left[\psi\left\{y - v(x)\right\} | X = u\right] f_X(u)\, du = \mathcal{O}\left(h^2\right)$$

where $\mathcal{O}(h^2)$ is independent of x in Parzen (1962), we have from assumption (A2) that $\|\mathsf{E} H_n\| = \mathcal{O}_p\{(nh)^{-1/2} (\log n)^{-1/2}\}$.

According to Lemma A.3 in Franke and Mwita (2003),

$$\sup_{x\in J}\left|H_n(x)-\mathsf{E}\,H_n(x)\right|=\mathcal{O}\left\{(nh)^{-1/2}(\log n)^{1/2}\right\}$$

and the following inequality:

$$\begin{aligned}\|H_n\| &\le \|H_n-\mathsf{E}\,H_n\|+\|\mathsf{E}\,H_n\|\\ &=\mathcal{O}\left\{(nh)^{-1/2}(\log n)^{1/2}\right\}+\mathcal{O}_p\left\{(nh)^{-1/2}(\log n)^{-1/2}\right\}\\ &=\mathcal{O}\left\{(nh)^{-1/2}(\log n)^{1/2}\right\}\end{aligned}$$

Statement (34) thus is obtained.

Statement (35) follows in the same way as (34) using assumption (A2) and the Lipschitz continuity properties of K, ψ', l.

According to the uniform consistency of $v_n(x)-v(x)$ shown before, we have

$$\|v_n-v\|=\mathcal{O}_p\left\{(nh)^{-1/2}(\log n)^{1/2}\right\}$$

which implies (37).

Now the assertion of the lemma follows, since by tightness of $D_n(x)$, $\inf_{0\le t\le 1}|D_n(x)|\ge q_0$ a.s. and thus

$$\|R_n\|=\mathcal{O}_p\left\{(nh\log n)^{-1/2}\right\}\left(1+\|f_n\|\right)$$

Finally, by Theorem 3.1 of Bickel and Rosenblatt (1973), $\|f_n\|=\mathcal{O}_p(1)$; thus the desired result $\|R_n\|=\mathcal{O}_p\{(nh\log n)^{-1/2}\}$ follows. □

We now begin with the subsequent approximations of the processes $Y_{0,n}$ to $Y_{5,n}$.

Lemma 5

$$\|Y_{0,n}-Y_{1,n}\|=\mathcal{O}\left\{(nh)^{-1/2}(\log n)^2\right\}\quad a.s.$$

Proof Let x be fixed and put $L(y)=\psi\{y-v(x)\}$ still depending on x. Using integration by parts, we obtain

$$\begin{aligned}&\iint_{\Gamma_n} L(y)K\left\{(x-t)/h\right\}dZ_n(t,y)\\ &=\int_{u=-A}^{A}\int_{y=-a_n}^{a_n} L(y)K(u)\,dZ_n(x-h\cdot u,y)\\ &=-\int_{-A}^{A}\int_{-a_n}^{a_n} Z_n(x-h\cdot u,y)d\left\{L(y)K(u)\right\}\\ &\quad+L(a_n)(a_n)\int_{-A}^{A} Z_n(x-h\cdot u,a_n)\,dK(u)\\ &\quad-L(-a_n)(-a_n)\int_{-A}^{A} Z_n(x-h\cdot u,-a_n)\,dK(u)\end{aligned}$$

$$
\begin{aligned}
&+ K(A)\left\{\int_{-a_n}^{a_n} Z_n(x - h\cdot A, y)\, dL(y)\right.\\
&\left.+ L(a_n)(a_n)Z_{n_a}(x - h\cdot A, a_n) - L(-a_n)(-a_n)Z_n(x - h\cdot A, -a_n)\right\}\\
&- K(-A)\left\{\int_{-a_n}^{a_n} Z_n(x + h\cdot A, y)\, dL(y) + L(a_n)(a_n)Z_n(x + h\cdot A, a_n)\right.\\
&\left.- L(-a_n)(-a_n)Z_n(x + h\cdot A, -a_n)\right\}
\end{aligned}
$$

If we apply the same operation to $Y_{1,n}$ with $B_n\{T(x, y)\}$ instead of $Z_n(x, y)$ and use Lemma 2, we finally obtain

$$\sup_{0\le x\le 1} h^{1/2} g(x)^{1/2} \left|Y_{0,n}(x) - Y_{1,n}(x)\right| = \mathcal{O}\left\{n^{-1/2}(\log n)^2\right\} \quad \text{a.s.}$$

□

Lemma 6 $\|Y_{1,n} - Y_{2,n}\| = \mathcal{O}_p(h^{1/2})$.

Proof Note that the Jacobian of $T(x, y)$ is $f(x, y)$. Hence

$$
\begin{aligned}
&Y_{1,n}(x) - Y_{2,n}(x)\\
&\quad = \left|\left\{g(x)h\right\}^{-1/2} \iint_{\Gamma_n} \psi\left\{y - v(x)\right\} K\left\{(x - t)/h\right\} f(t, y)\, dt\, dy\right| \cdot \left|W_n(1, 1)\right|
\end{aligned}
$$

It follows that

$$
\begin{aligned}
h^{-1/2}\|Y_{1,n} - Y_{2,n}\| \le{}& \left|W_n(1, 1)\right| \cdot \left\|g^{-1/2}\right\|\\
&\cdot \sup_{0\le t\le 1} h^{-1} \iint_{\Gamma_n} \left|\psi\left\{y - v(x)\right\} K\left\{(x - t)/h\right\}\right| f(t, y)\, dt\, dy
\end{aligned}
$$

Since $\|g^{-1/2}\|$ is bounded by assumption, we have

$$h^{-1/2}\|Y_{1,n} - Y_{2,n}\| \le \left|W_n(1, 1)\right| \cdot C_4 \cdot h^{-1} \int K\left\{(x - t)/h\right\} dx = \mathcal{O}_p(1)$$

□

Lemma 7 $\|Y_{2,n} - Y_{3,n}\| = \mathcal{O}_p(h^{1/2})$.

Proof The difference $|Y_{2,n}(x) - Y_{3,n}(x)|$ may be written as

$$\left|\left\{g(x)h\right\}^{-1/2} \iint_{\Gamma_n} \left[\psi\left\{y - v(x)\right\} - \psi\left\{y - v(t)\right\}\right] K\left\{(x - t)/h\right\} dW_n\left\{T(t, y)\right\}\right|$$

If we use the fact that l is uniformly continuous, this is smaller than

$$h^{-1/2}\left|g(x)\right|^{-1/2} \cdot \mathcal{O}_p(h)$$

and the lemma thus follows. □

Lemma 8 $\|Y_{4,n} - Y_{5,n}\| = \mathcal{O}_p(h^{1/2})$.

Proof

$$\begin{aligned}
\big|Y_{4,n}(x) - Y_{5,n}(x)\big| &= h^{-1/2}\left|\int\left[\left\{\frac{g(t)}{g(x)}\right\}^{1/2} - 1\right]K\big\{(x-t)/h\big\}dW(x)\right| \\
&\le h^{-1/2}\left|\int_{-A}^{A} W(x-hu)\frac{\partial}{\partial u}\left[\left\{\frac{g(x-hu)}{g(x)}\right\}^{1/2} - 1\right]K(u)\,du\right| \\
&\quad + h^{-1/2}\left|K(A)W(t-hA)\left[\left\{\frac{g(x-Ah)}{g(x)}\right\}^{1/2} - 1\right]\right| \\
&\quad + h^{-1/2}\left|K(-A)W(x+hA)\left[\left\{\frac{g(x+Ah)}{g(x)}\right\}^{1/2} - 1\right]\right| \\
&\quad S_{1,n}(x) + S_{2,n}(x) + S_{3,n}(x), \quad \text{say}
\end{aligned}$$

The second term can be estimated by

$$h^{-1/2}\|S_{2,n}\| \le K(A)\cdot \sup_{0\le x\le 1}\big|W(x-Ah)\big|\cdot \sup_{0\le x\le 1} h^{-1}\left|\left[\left\{\frac{g(x-Ah)}{g(x)}\right\}^{1/2} - 1\right]\right|$$

by the mean value theorem it follows that

$$h^{-1/2}\|S_{2,n}\| = \mathcal{O}_p(1)$$

The first term $S_{1,n}$ is estimated as

$$\begin{aligned}
h^{-1/2}S_{1,n}(x) &= \left|h^{-1}\int_{-A}^{A} W(x-uh)K'(u)\left[\left\{\frac{g(x-uh)}{g(x)}\right\}^{1/2} - 1\right]du\right. \\
&\quad \left.\times\frac{1}{2}\int_{-A}^{A} W(x-uh)K(u)\left\{\frac{g(x-uh)}{g(x)}\right\}^{1/2}\left\{\frac{g'(x-uh)}{g(x)}\right\}du\right| \\
&= \big|T_{1,n}(x) - T_{2,n}(x)\big|, \quad \text{say}
\end{aligned}$$

$\|T_{2,n}\| \le C_5 \cdot \int_{-A}^{A}|W(t-hu)|\,du = \mathcal{O}_p(1)$ by assumption on $g(x) = \sigma^2(x)\cdot f_X(x)$. To estimate $T_{1,n}$ we again use the mean value theorem to conclude that

$$\sup_{0\le x\le 1} h^{-1}\left|\left\{\frac{g(x-uh)}{g(x)}\right\}^{1/2} - 1\right| < C_6\cdot|u|$$

hence

$$\|T_{1,n}\| \le C_6\cdot \sup_{0\le x\le 1}\int_{-A}^{A}\big|W(x-hu)\big|K'(u)u/\,du = \mathcal{O}_p(1)$$

Since $S_{3,n}(x)$ is estimated as $S_{2,n}(x)$, we finally obtain the desired result. □

The next lemma shows that the truncation introduced through $\{a_n\}$ does not affect the limiting distribution.

Lemma 9 $\|Y_n - Y_{0,n}\| = \mathcal{O}_p\{(\log n)^{-1/2}\}$.

Proof We shall only show that $g'(x)^{-1/2}h^{-1/2}\iint_{\mathbb{R}-\Gamma_n}\psi\{y - v(x)\} \times K\{(x-t)/h\}\,dZ_n(t,y)$ fulfills the lemma. The replacement of $g'(x)$ by $g(x)$ may be proved as in Lemma A.4 of Johnston (1982). The quantity above is less than $h^{-1/2}\|g^{-1/2}\| \cdot \|\iint_{\{|y|>a_n\}}\psi\{y - v(x)\}K\{(x-t)/h\}\,dZ(t,y)\|$. It remains to be shown that the last factor tends to zero at a rate $\mathcal{O}_p\{(\log n)^{-1/2}\}$. We show first that

$$V_n(x) = (\log n)^{1/2}h^{-1/2}\iint_{\{|y|>a_n\}}\psi\big\{y - v(x)\big\}K\big\{(x-t)/h\big\}\,dZ_n(t,y)$$

$$\xrightarrow{p} 0 \quad \text{for all } x$$

and then we show tightness of $V_n(x)$, the result then follows:

$$V_n(x) = (\log n)^{1/2}(nh)^{-1/2}\sum_{i=1}^{n}\big[\psi\big\{Y_i - v(x)\big\}\mathbf{I}\big(|Y_i| > a_n\big)K\big\{(x-X_i)/h\big\}$$

$$- \mathsf{E}\,\psi\big\{Y_i - v(x)\big\}\mathbf{I}\big(|Y_i| > a_n\big)K\big\{(x-X_i)/h\big\}\big]$$

$$= \sum_{i=1}^{n} X_{n,x}(x)$$

where $\{X_{n,x}(x)\}_{i=1}^{n}$ are i.i.d. for each n with $\mathsf{E}\,X_{n,x}(x) = 0$ for all $x \in [0,1]$. We then have

$$\mathsf{E}\,X_{n,x}^2(x) \le (\log n)(nh)^{-1}\,\mathsf{E}\,\psi^2\big\{Y_i - v(x)\big\}\mathbf{I}\big(|Y_i| > a_n\big)K^2\big\{(x-X_i)/h\big\}$$

$$\le \sup_{-A\le u\le A} K^2(u)\cdot(\log n)(nh)^{-1}\,\mathsf{E}\,\psi^2\big\{Y_i - v(x)\big\}\mathbf{I}\big(|Y_i| > a_n\big)$$

hence

$$\mathsf{Var}\big\{V_n(x)\big\} = \mathsf{E}\left\{\sum_{i=1}^{n} X_{n,x}(x)\right\}^2 = n\cdot\mathsf{E}\,X_{n,x}^2(x)$$

$$\le \sup_{-A\le u\le A} K^2(u)h^{-1}(\log n)\int_{\{|y|>a_n\}} f_y(y)\,dy\cdot M_\psi$$

where M_ψ denotes an upper bound for ψ^2. This term tends to zero by assumption (A3). Thus by Markov's inequality we conclude that

$$V_n(x) \xrightarrow{p} 0 \quad \text{for all } x \in [0,1]$$

To prove tightness of $\{V_n(x)\}$ we refer again to the following moment condition as stated in Lemma 4:

$$\mathsf{E}\big\{\big|V_n(x) - V_n(x_1)\big| \cdot \big|V_n(x_2) - V_n(x)\big|\big\} \le C' \cdot (x_2 - x_1)^2$$

C' denoting a constant, $x \in [x_1, x_2]$

We again estimate the left-hand side by Schwarz's inequality and estimate each factor separately,

$$\mathsf{E}\big\{V_n(x) - V_n(x_1)\big\}^2 = (\log n)(nh)^{-1}\,\mathsf{E}\Bigg[\sum_{i=1}^{n} \Psi_n(x, x_1, X_i, Y_i) \cdot \mathbf{I}\big(|Y_i| > a_n\big) - \mathsf{E}\big\{\Psi_n(x, x_1, X_i, Y_i) \cdot \mathbf{I}\big(|Y_i| > a_n\big)\big\}\Bigg]^2$$

where $\Psi_n(x, x_1, X_i, Y_i) = \psi\{Y_i - v(x)\}K\{(x - X_i)/h\} - \psi\{Y_i - v(x_1)\}K\{(x_1 - X_1)/h\}$. Since ψ, K are Lipschitz continuous except at one point and the expectation is taken afterwards, it follows that

$$\begin{aligned}&\big[\mathsf{E}\big\{V_n(x) - V_n(x_1)\big\}^2\big]^{1/2}\\ &\quad\le C_7 \cdot (\log n)^{1/2} h^{-3/2} |x - x_1| \cdot \left\{\int_{\{|y| > a_n\}} f_y(y)\,dy\right\}^{1/2}\end{aligned}$$

If we apply the same estimation to $V_n(x_2) - V_n(x_1)$ we finally have

$$\begin{aligned}&\mathsf{E}\big\{\big|V_n(x) - V_n(x_1)\big| \cdot \big|V_n(x_2) - V_n(x)\big|\big\}\\ &\quad\le C_7^2 (\log n) h^{-3} |x - x_1||x_2 - x| \times \int_{\{|y| > a_n\}} f_y(y)\,dy\\ &\quad\le C' \cdot |x_2 - x_1|^2 \quad \text{since } x \in [x_1, x_2] \text{ by (A3)}\end{aligned}$$

□

Lemma 10 *Let $\lambda(K) = \int K^2(u)\,du$ and let $\{d_n\}$ be as in the theorem. Then*

$$(2\delta \log n)^{1/2}\big[\|Y_{3,n}\|/\big\{\lambda(K)\big\}^{1/2} - d_n\big]$$

has the same asymptotic distribution as

$$(2\delta \log n)^{1/2}\big[\|Y_{4,n}\|/\big\{\lambda(K)\big\}^{1/2} - d_n\big]$$

Proof $Y_{3,n}(x)$ is a Gaussian process with

$$\mathsf{E}\, Y_{3,n}(x) = 0$$

and covariance function

$$
\begin{aligned}
r_3(x_1, x_2) &= \mathsf{E}\, Y_{3,n}(x_1) Y_{3,n}(x_2) \\
&= \big\{g(x_1)g(x_2)\big\}^{-1/2} h^{-1} \iint_{\Gamma_n} \psi^2\big\{y - v(x)\big\} K\big\{(x_1 - x)/h\big\} \\
&\quad \times K\big\{(x_2 - x)/h\big\} f(t, y)\, dt\, dy \\
&= \big\{g(x_1)g(x_2)\big\}^{-1/2} h^{-1} \iint_{\Gamma_n} \psi^2\big\{y - v(x)\big\} f(y|x) dy K\big\{(x_1 - x)/h\big\} \\
&\quad \times K\big\{(x_2 - x)/h\big\} f_X(x)\, dx \\
&= \big\{g(x_1)g(x_2)\big\}^{-1/2} h^{-1} \int g(x) K\big\{(x_1 - x)/h\big\} K\big\{(x_2 - x)/h\big\}\, dx \\
&= r_4(x_1, x_2)
\end{aligned}
$$

where $r_4(x_1, x_2)$ is the covariance function of the Gaussian process $Y_{4,n}(x)$, which proves the lemma. □

References

Benth, F., Benth, J., Koekebakker, S.: Putting a price on temperature. Scand. J. Stat. **34**(4), 746–767 (2007)

Bickel, P., Rosenblatt, M.: On some global measures of the deviation of density function estimates. Ann. Stat. **1**, 1071–1095 (1973)

Breckling, J., Chambers, R.: M-quantiles. Biometrika **74**(4), 761–772 (1988)

Campbell, S., Diebold, F.: Weather forecasting for weather derivatives. J. Am. Stat. Assoc. **100**, 6–16 (2005)

Claeskens, G., Keilegom, I.V.: Bootstrap confidence bands for regression curves and their derivatives. Ann. Stat. **31**(6), 1852–1884 (2003)

Csörgö, S., Hall, P.: Upper and lower classes for triangular arrays. Z. Wahrscheinlichkeitstheor. Verw. Geb. **61**, 207–222 (1982)

Diebold, F., Inoue, A.: Long memory and regime switching. J. Econom. **105**, 131–159 (2001)

Efron, B.: Regression percentiles using asymmetric squared loss. Stat. Sin. **1**, 93–125 (1991)

Franke, J., Mwita, P.: Nonparametric estimates for conditional quantiles of time series. Report in Wirtschaftsmathematik, 87, University of Kaiserslautern (2003)

Härdle, W.: Asymptotic maximal deviation of M-smoothers. J. Multivar. Anal. **29**, 163–179 (1989)

Härdle, W., Luckhaus, S.: Uniform consistency of a class of regression function estimators. Ann. Stat. **12**, 612–623 (1984)

Härdle, W., Song, S.: Confidence bands in quantile regression. Econom. Theory **3**, 1–21 (2009)

Härdle, W., Janssen, P., Serfling, R.: Strong uniform consistency rates for estimators of conditional functionals. Ann. Stat. **16**, 1428–1429 (1988)

Härdle, W., Ritov, Y., Song, S.: Partial linear regression and bootstrap confidence bands. SFB 649 Discussion Paper 2010-002. J. Multivar. Anal. (2010, submitted)

Huber, P.: Robust Statistics. Wiley, New York (1981)

Johnston, G.: Probabilities of maximal deviations of nonparametric regression function estimates. J. Multivar. Anal. **12**, 402–414 (1982)

Jones, M.: Expectiles and M-quantiles are quantiles. Stat. Probab. Lett. **20**, 149–153 (1994)

Koenker, R.: Quantile Regression. Cambridge University Press, Cambridge (2005)

Kuan, C.M., Yeh, Y.H., Hsu, Y.C.: Assessing value at risk with care, the conditional autoregressive expectile models. J. Econom. **150**, 261–270 (2009)

Parzen, M.: On estimation of a probability density function and mode. Ann. Math. Stat. **32**, 1065–1076 (1962)

Rosenblatt, M.: Remarks on a multivariate transformation. Ann. Math. Stat. **23**, 470–472 (1952)

Schnabel, S., Eilers, P.: An analysis of life expectancy and economic production using expectile frontier zones. Demogr. Res. **21**, 109–134 (2009a)

Schnabel, S., Eilers, P.: Optimal expectile smoothing. Comput. Stat. Data Anal. **53**, 4168–4177 (2009b)
Taylor, J.: Estimating value at risk and expected shortfall using expectiles. J. Financ. Econom. **6**, 231–252 (2008)
Tusnady, G.: A remark on the approximation of the sample distribution function in the multidimensional case. Period. Math. Hung. **8**, 53–55 (1977)
Yao, Q., Tong, H.: Asymmetric least squares regression estimation: a nonparametric approach. J. Nonparametr. Stat. **6**(2–3), 273–292 (1996)
Zhang, B.: Nonparametric regression expectiles. J. Nonparametr. Stat. **3**, 255–275 (1994)

The Annals of Statistics
2011, Vol. 39, No. 3, 1658–1688
DOI: 10.1214/10-AOS871

THE EFM APPROACH FOR SINGLE-INDEX MODELS

BY XIA CUI[1], WOLFGANG KARL HÄRDLE[2] AND LIXING ZHU[3]

Sun Yat-sen University, Humboldt-Universität zu Berlin and National Central University, and Hong Kong Baptist University and Yunnan University of Finance and Economics

Single-index models are natural extensions of linear models and circumvent the so-called curse of dimensionality. They are becoming increasingly popular in many scientific fields including biostatistics, medicine, economics and financial econometrics. Estimating and testing the model index coefficients $\boldsymbol{\beta}$ is one of the most important objectives in the statistical analysis. However, the commonly used assumption on the index coefficients, $\|\boldsymbol{\beta}\| = 1$, represents a nonregular problem: the true index is on the boundary of the unit ball. In this paper we introduce the EFM approach, a method of estimating functions, to study the single-index model. The procedure is to first relax the equality constraint to one with $(d-1)$ components of $\boldsymbol{\beta}$ lying in an open unit ball, and then to construct the associated $(d-1)$ estimating functions by projecting the score function to the linear space spanned by the residuals with the unknown link being estimated by kernel estimating functions. The root-n consistency and asymptotic normality for the estimator obtained from solving the resulting estimating equations are achieved, and a Wilks type theorem for testing the index is demonstrated. A noticeable result we obtain is that our estimator for $\boldsymbol{\beta}$ has smaller or equal limiting variance than the estimator of Carroll et al. [*J. Amer. Statist. Assoc.* **92** (1997) 447–489]. A fixed-point iterative scheme for computing this estimator is proposed. This algorithm only involves one-dimensional nonparametric smoothers, thereby avoiding the data sparsity problem caused by high model dimensionality. Numerical studies based on simulation and on applications suggest that this new estimating system is quite powerful and easy to implement.

1. Introduction. Single-index models combine flexibility of modeling with interpretability of (linear) coefficients. They circumvent the curse of dimensionality and are becoming increasingly popular in many scientific fields. The reduction of dimension is achieved by assuming the link function to be a univariate function applied to the projection of explanatory covariate vector on to some direction.

Received April 2010; revised December 2010.

[1]Supported by NNSF project (11026194) of China, RFDP (20100171120042) of China and "the Fundamental Research Funds for the Central Universities" (11lgpy26) of China.

[2]Supported by Deutsche Forschungsgemeinschaft SFB 649 "Ökonomisches Risiko."

[3]Supported by a Grant (HKBU2030/07P) from Research Grants Council of Hong Kong, Hong Kong, China.

MSC2010 subject classifications. 62G08, 62G08, 62G20.

Key words and phrases. Single-index models, index coefficients, estimating equations, asymptotic properties, iteration.

In this paper we consider an extension of single-index models where, instead of a distributional assumption, assumptions of only the mean function and variance function of the response are made. Let $(Y_i, \mathbf{X}_i)$, $i = 1, \ldots, n$, denote the observed values with Y_i being the response variable and $\mathbf{X}_i$ as the vector of d explanatory variables. The relationship of the mean and variance of Y_i is specified as follows:

$$E(Y_i|\mathbf{X}_i) = \mu\{g(\boldsymbol{\beta}^\top \mathbf{X}_i)\}, \qquad \mathrm{Var}(Y_i|\mathbf{X}_i) = \sigma^2 V\{g(\boldsymbol{\beta}^\top \mathbf{X}_i)\}, \tag{1.1}$$

where μ is a known monotonic function, V is a known covariance function, g is an unknown univariate link function and $\boldsymbol{\beta}$ is an unknown index vector which belongs to the parameter space $\Theta = \{\boldsymbol{\beta} = (\beta_1, \ldots, \beta_d)^\top : \|\boldsymbol{\beta}\| = 1, \beta_1 > 0, \boldsymbol{\beta} \in \mathbb{R}^d\}$. Here we assume the parameter space is Θ rather than the entire $\mathbb{R}^d$ in order to ensure that $\boldsymbol{\beta}$ in the representation (1.1) can be uniquely defined. This is a commonly used assumption on the index parameter [see Carroll et al. (1997), Zhu and Xue (2006), Lin and Kulasekera (2007)]. Another reparameterization is to let $\beta_1 = 1$ for the sign identifiability and to transform $\boldsymbol{\beta}$ to $(1, \beta_2, \ldots, \beta_d)/(1 + \sum_{r=2}^d \beta_r^2)^{1/2}$ for the scale identifiability. Clearly $(1, \beta_2, \ldots, \beta_d)/(1+\sum_{r=2}^d \beta_r^2)^{1/2}$ can also span the parameter space Θ by simply checking that $\|(1, \beta_2, \ldots, \beta_d)/(1 + \sum_{r=2}^d \beta_r^2)^{1/2}\| = 1$ and the first component $1/(1 + \sum_{r=2}^d \beta_r^2)^{1/2} > 0$. However, the fixed-point algorithm recommended in this paper for normalized vectors may not be suitable for such a reparameterization. Model (1.1) is flexible enough to cover a variety of situations. If μ is the identity function and V is equal to constant 1, (1.1) reduces to a single-index model Härdle, Hall and Ichimura (1993). Model (1.1) is an extension of the generalized linear model McCullagh and Nelder (1989) and the single-index model. When the conditional distribution of Y is logistic, then $\mu\{g(\boldsymbol{\beta}^\top \mathbf{X})\} = \exp\{g(\boldsymbol{\beta}^\top \mathbf{X})\}/[1 + \exp\{g(\boldsymbol{\beta}^\top \mathbf{X})\}]$ and $V\{g(\boldsymbol{\beta}^\top \mathbf{X})\} = \exp\{g(\boldsymbol{\beta}^\top \mathbf{X})\}/[1 + \exp\{g(\boldsymbol{\beta}^\top \mathbf{X})\}]^2$.

For single-index models: $\mu\{g(\boldsymbol{\beta}^\top \mathbf{X})\} = g(\boldsymbol{\beta}^\top \mathbf{X})$ and $V\{g(\boldsymbol{\beta}^\top \mathbf{X})\} = 1$, various strategies for estimating $\boldsymbol{\beta}$ have been proposed in the last decades. Two most popular methods are the average derivative method (ADE) introduced in Powell, Stock and Stoker (1989) and Härdle and Stoker (1989), and the simultaneous minimization method of Härdle, Hall and Ichimura (1993). Next we will review these two methods in short. The ADE method is based on that $\partial E(Y|\mathbf{X} = \mathbf{x})/\partial \mathbf{x} = g'(\boldsymbol{\beta}^\top \mathbf{x})\boldsymbol{\beta}$ which implies that the gradient of the regression function is proportional to the index parameter $\boldsymbol{\beta}$. Then a natural estimator for $\boldsymbol{\beta}$ is $\hat{\boldsymbol{\beta}} = n^{-1} \sum_{i=1}^n \widehat{\nabla G}(\mathbf{X}_i)/\|n^{-1} \sum_{i=1}^n \widehat{\nabla G}(\mathbf{X}_i)\|$ with $\nabla G(\mathbf{x})$ denoting $\partial E(Y|\mathbf{X} = \mathbf{x})/\partial \mathbf{x}$ and $\|\cdot\|$ being the Euclidean norm. An advantage of the ADE approach is that it allows estimating $\boldsymbol{\beta}$ directly. However, the high-dimensional kernel smoothing used for computing $\widehat{\nabla G}(\mathbf{x})$ suffers from the "curse of dimensionality" if the model dimension d is large. Hristache, Juditski and Spokoiny (2001) improved the ADE approach by lowering the dimension of the kernel gradually. The method of Härdle, Hall and Ichimura (1993) is carried out by minimizing a least squares criterion based on nonparametric estimation of the link g with respect

to $\boldsymbol{\beta}$ and bandwidth h. However, the minimization is difficult to implement since it depends on an optimization problem in a high-dimensional space. Xia et al. (2002) proposed to minimize average conditional variance (MAVE). Because the kernel used for computing $\boldsymbol{\beta}$ is a function of $\|\mathbf{X}_i - \mathbf{X}_j\|$, MAVE meets the problem of data sparseness. All the above estimators are consistent under some regular conditions. Asymptotic efficiency comparisons of the above methods have been discussed in Xia (2006) resulting in the MAVE estimator of $\boldsymbol{\beta}$ having the same limiting variance as the estimators of Härdle, Hall and Ichimura (1993), and claiming alternative versions of the ADE method having larger variance. In addition, Yu and Ruppert (2002) fitted the partially linear single-index models using a penalized spline method. Huh and Park (2002) used the local polynomial method to fit the unknown function in single-index models. Other dimension reduction methods that were recently developed in the literature are sliced inverse regression, partial least squares and canonical correlation method. These methods handle high-dimensional predictors; see Zhu and Zhu (2009a, 2009b) and Zhou and He (2008).

The main challenges of estimation in the semiparametric model (1.1) are that the support of the infinite-dimensional nuisance parameter $g(\cdot)$ depends on the finite-dimensional parameter $\boldsymbol{\beta}$, and the parameter $\boldsymbol{\beta}$ is on the boundary of a unit ball. For estimating $\boldsymbol{\beta}$ the former challenge forces us to deal with the infinite-dimensional nuisance parameter g. The latter one represents a nonregular problem. The classic assumptions about asymptotic properties of the estimates for $\boldsymbol{\beta}$ are not valid. In addition, as a model proposed for dimension reduction, the dimension d may be very high and one often meets the problem of computation. To attack the above problems, in this paper we will develop an estimating function method (EFM) and then introduce a computational algorithm to solve the equations based on a fixed-point iterative scheme. We first choose an identifiable parameterization which transforms the boundary of a unit ball in $\mathbb{R}^d$ to the interior of a unit ball in $\mathbb{R}^{d-1}$. By eliminating β_1, the parameter space Θ can be rearranged to a form $\{((1-\sum_{r=2}^{d}\beta_r^2)^{1/2}, \beta_2, \ldots, \beta_d)^\top : \sum_{r=2}^{d}\beta_r^2 < 1\}$. Then the derivatives of a function with respect to $(\beta_2, \ldots, \beta_d)^\top$ are readily obtained by the chain rule and the classical assumptions on the asymptotic normality hold after transformation. The estimating functions (equations) for $\boldsymbol{\beta}$ can be constructed by replacing $g(\boldsymbol{\beta}^\top\mathbf{X})$ with $\hat{g}(\boldsymbol{\beta}^\top\mathbf{X})$. The estimate $\hat{g}$ for the nuisance parameter g is obtained using kernel estimating functions and the smoothing parameter h is selected using K-fold cross-validation. For the problem of testing the index, we establish a quasi-likelihood ratio based on the proposed estimating functions and show that the test statistics asymptotically follow a χ^2-distribution whose degree of freedom does not depend on nuisance parameters, under the null hypothesis. Then a Wilks type theorem for testing the index is demonstrated.

The proposed EFM technique is essentially a unified method of handling different types of data situations including categorical response variable and discrete explanatory covariate vector. The main results of this research are as follows:

(a) *Efficiency*. A surprising result we obtain is that our EFM estimator for $\boldsymbol{\beta}$ has smaller or equal limiting variance than the estimator of Carroll et al. (1997).
(b) *Computation*. The estimating function system only involves one-dimensional nonparametric smoothers, thereby avoiding the data sparsity problem caused by high model dimensionality. Unlike the quasi-likelihood inference [Carroll et al. (1997)] where the maximization is difficult to implement when d is large, the reparameterization and the explicit formulation of the estimating functions facilitate an efficient computation algorithm. Here we use a fixed-point iterative scheme to compute the resultant estimator. The simulation results show that the algorithm adapts to higher model dimension and richer data situations than the MAVE method of Xia et al. (2002).

It is noteworthy that the EFM approach proposed in this paper cannot be obtained from the SLS method proposed in Ichimura (1993) and investigated in Härdle, Hall and Ichimura (1993). SLS minimizes the weighted least squares criterion $\sum_{j=1}^{n}[Y_j - \mu\{\hat{g}(\boldsymbol{\beta}^\top \mathbf{X}_j)\}]^2 V^{-1}\{\hat{g}(\boldsymbol{\beta}^\top \mathbf{X}_j)\}$, which leads to a biased estimating equation when we use its derivative if $V(\cdot)$ does not contain the parameter of interest. It will not in general provide a consistent estimator [see Heyde (1997), page 4]. Chang, Xue and Zhu (2010) and Wang et al. (2010) discussed the efficient estimation of single-index model for the case of additive noise. However, their methods are based on the estimating equations induced from the least squares rather than the quasi-likelihood. Thus, their estimation does not have optimal property. Also their comparison is with the one from Härdle, Hall and Ichimura (1993) and its later development. It cannot be applied to the setting under study. In this paper, we investigate the efficiency and computation of the estimates for the single-index models, and systematically develop and prove the asymptotic properties of EFM.

The paper is organized as follows. In Section 2, we state the single-index model, discuss estimation of g using kernel estimating functions and of $\boldsymbol{\beta}$ using profile estimating functions, and investigate the problem of testing the index using quasi-likelihood ratio. In Section 3 we provide a computation algorithm for solving the estimating functions and illustrate the method with simulation and practical studies. The proofs are deferred to the Appendix.

2. Estimating function method (EFM) and its large sample properties. In this section, which is concerned with inference based on the estimating function method, the model of interest is determined through specification of mean and variance functions, up to an unknown vector $\boldsymbol{\beta}$ and an unknown function g. Except for Gaussian data, model (1.1) need not be a full semiparametric likelihood specification. Note that the parameter space $\Theta = \{\boldsymbol{\beta} = (\beta_1, \ldots, \beta_d)^\top : \|\boldsymbol{\beta}\| = 1, \beta_1 > 0, \boldsymbol{\beta} \in \mathbb{R}^d\}$ means that $\boldsymbol{\beta}$ is on the boundary of a unit ball and it represents therefore a nonregular problem. So we first choose an identifiable parameterization which transforms the boundary of a unit ball in $\mathbb{R}^d$ to the interior of a unit ball in $\mathbb{R}^{d-1}$. By eliminating β_1, the parameter space Θ can be rearranged to a form

$\{((1-\sum_{r=2}^{d}\beta_r^2)^{1/2},\beta_2,\ldots,\beta_d)^\top:\sum_{r=2}^{d}\beta_r^2<1\}$. Then the derivatives of a function with respect to $\boldsymbol{\beta}^{(1)}=(\beta_2,\ldots,\beta_d)^\top$ are readily obtained by chain rule and the classic assumptions on the asymptotic normality hold after transformation. This reparameterization is the key to analyzing the asymptotic properties of the estimates for $\boldsymbol{\beta}$ and to facilitating an efficient computation algorithm. We will investigate the estimation for g and $\boldsymbol{\beta}$ and propose a quasi-likelihood method to test the statistical significance of certain variables in the parametric component.

2.1. *The kernel estimating functions for the nonparametric part* g. If $\boldsymbol{\beta}$ is known, then we estimate $g(\cdot)$ and $g'(\cdot)$ using the local linear estimating functions. Let h denote the bandwidth parameter, and let $K(\cdot)$ denote the symmetric kernel density function satisfying $K_h(\cdot)=h^{-1}K(\cdot/h)$. The estimation method involves local linear approximation. Denote by α_0 and α_1 the values of g and g' evaluating at t, respectively. The local linear approximation for $g(\boldsymbol{\beta}^\top\mathbf{x})$ in a neighborhood of t is $\tilde{g}(\boldsymbol{\beta}^\top\mathbf{x})=\alpha_0+\alpha_1(\boldsymbol{\beta}^\top\mathbf{x}-t)$. The estimators $\hat{g}(t)$ and $\hat{g}'(t)$ are obtained by solving the kernel estimating functions with respect to α_0,α_1:

$$
\begin{cases}
\displaystyle\sum_{j=1}^{n} K_h(\boldsymbol{\beta}^\top\mathbf{X}_j-t)\mu'\{\tilde{g}(\boldsymbol{\beta}^\top\mathbf{X}_j)\}V^{-1}\{\tilde{g}(\boldsymbol{\beta}^\top\mathbf{X}_j)\} \\
\qquad\qquad \times[Y_j-\mu\{\tilde{g}(\boldsymbol{\beta}^\top\mathbf{X}_j)\}]=0, \\
\displaystyle\sum_{j=1}^{n}(\boldsymbol{\beta}^\top\mathbf{X}_j-t)K_h(\boldsymbol{\beta}^\top\mathbf{X}_j-t)\mu'\{\tilde{g}(\boldsymbol{\beta}^\top\mathbf{X}_j)\}V^{-1}\{\tilde{g}(\boldsymbol{\beta}^\top\mathbf{X}_j)\} \\
\qquad\qquad \times[Y_j-\mu\{\tilde{g}(\boldsymbol{\beta}^\top\mathbf{X}_j)\}]=0.
\end{cases}
\tag{2.1}
$$

Having estimated α_0,α_1 at t as $\hat{\alpha}_0,\hat{\alpha}_1$, the local linear estimators of $g(t)$ and $g'(t)$ are $\hat{g}(t)=\hat{\alpha}_0$ and $\hat{g}'(t)=\hat{\alpha}_1$, respectively.

The key to obtain the asymptotic normality of the estimates for $\boldsymbol{\beta}$ lies in the asymptotic properties of the estimated nonparametric part. The following theorem will provide some useful results. The following notation will be used. Let $\mathcal{X}=\{\mathbf{X}_1,\ldots,\mathbf{X}_n\}$, $\rho_l(z)=\{\mu'(z)\}^lV^{-1}(z)$ and $\mathbf{J}=\frac{\partial\boldsymbol{\beta}}{\partial\boldsymbol{\beta}^{(1)}}$ the Jacobian matrix of size $d\times(d-1)$ with

$$
\mathbf{J}=\begin{pmatrix}-\boldsymbol{\beta}^{(1)\top}/\sqrt{1-\|\boldsymbol{\beta}^{(1)}\|^2}\\ \mathbf{I}_{d-1}\end{pmatrix},\qquad \boldsymbol{\beta}^{(1)}=(\beta_2,\ldots,\beta_d)^\top.
$$

The moments of K and K^2 are denoted, respectively, by, $j=0,1,\ldots,$

$$
\gamma_j=\int t^jK(t)\,dt\quad\text{and}\quad \nu_j=\int t^jK^2(t)\,dt.
$$

PROPOSITION 1. *Under regularity conditions* (a), (b), (d) *and* (e) *given in the Appendix*, *we have*:

(i) *With $h \to 0$, $n \to \infty$ such that $h \to 0$ and $nh \to \infty$, $\forall \boldsymbol{\beta} \in \Theta$, the asymptotic conditional bias and variance of $\hat{g}$ are given by*

$$
\begin{aligned}
&E\{\{\hat{g}(\boldsymbol{\beta}^\top \mathbf{x}) - g(\boldsymbol{\beta}^\top \mathbf{x})\}^2 | \mathcal{X}\} \\
&\quad = \{\tfrac{1}{2}\gamma_2 h^2 g''(\boldsymbol{\beta}^\top \mathbf{x})\}^2 \\
&\qquad + \nu_0 \sigma^2 / [nh f_{\boldsymbol{\beta}^\top \mathbf{x}}(\boldsymbol{\beta}^\top \mathbf{x}) \rho_2\{g(\boldsymbol{\beta}^\top \mathbf{x})\}] \\
&\qquad + \mathcal{O}_P(h^4 + n^{-1}h^{-1}).
\end{aligned} \tag{2.2}
$$

(ii) *With $h \to 0$, $n \to \infty$ such that $h \to 0$ and $nh^3 \to \infty$, for the estimates of the derivative g', it holds that*

$$
\begin{aligned}
&E\{\{\hat{g}'(\boldsymbol{\beta}^\top \mathbf{x}) - g'(\boldsymbol{\beta}^\top \mathbf{x})\}^2 | \mathcal{X}\} \\
&\quad = \{\tfrac{1}{6}\gamma_4 \gamma_2^{-1} h^2 g'''(\boldsymbol{\beta}^\top \mathbf{x}) \\
&\qquad + \tfrac{1}{2}(\gamma_4 \gamma_2^{-1} - \gamma_2) h^2 g''(\boldsymbol{\beta}^\top \mathbf{x}) \\
&\qquad\quad \times [\rho_2'\{g(\boldsymbol{\beta}^\top \mathbf{x})\}/\rho_2\{g(\boldsymbol{\beta}^\top \mathbf{x})\} + f'_{\boldsymbol{\beta}^\top \mathbf{x}}(\boldsymbol{\beta}^\top \mathbf{x})/f_{\boldsymbol{\beta}^\top \mathbf{x}}(\boldsymbol{\beta}^\top \mathbf{x})]\}^2 \\
&\qquad + \nu_2 \gamma_2^{-2} \sigma^2 / [nh^3 f_{\boldsymbol{\beta}^\top \mathbf{x}}(\boldsymbol{\beta}^\top \mathbf{x}) \rho_2\{g(\boldsymbol{\beta}^\top \mathbf{x})\}] \\
&\qquad + \mathcal{O}_P(h^4 + n^{-1}h^{-3}).
\end{aligned} \tag{2.3}
$$

(iii) *With $h \to 0$, $n \to \infty$ such that $h \to 0$ and $nh^3 \to \infty$, we have that*

$$
E\left\{\left\| \frac{\partial \hat{g}(\boldsymbol{\beta}^\top \mathbf{x})}{\partial \boldsymbol{\beta}^{(1)}} - g'(\boldsymbol{\beta}^\top \mathbf{x}) \mathbf{J}^\top \{\mathbf{x} - E(\mathbf{x}|\boldsymbol{\beta}^\top \mathbf{x})\} \right\|^2 \Big| \mathcal{X} \right\} = \mathcal{O}_P(h^4 + n^{-1}h^{-3}). \tag{2.4}
$$

The proof of this proposition appears in the Appendix. Results (i) and (ii) in Proposition 1 are routine and similar to Carroll, Ruppert and Welsh (1998). In the situation where $\sigma^2 V = \sigma^2$ and the function μ is identity, results (i) and (ii) coincide with those given by Fan and Gijbels (1996). From result (iii), it is seen that $\partial \hat{g}(\boldsymbol{\beta}^\top \mathbf{x})/\partial \boldsymbol{\beta}^{(1)}$ converges in probability to $g'(\boldsymbol{\beta}^\top \mathbf{x}) \mathbf{J}^\top \{\mathbf{x} - E(\mathbf{x}|\boldsymbol{\beta}^\top \mathbf{x})\}$, rather than $g'(\boldsymbol{\beta}^\top \mathbf{x}) \mathbf{J}^\top \mathbf{x}$ as if g were known. That is, $\lim_{n\to\infty}\{\partial \hat{g}(\boldsymbol{\beta}^\top \mathbf{x})/\partial \boldsymbol{\beta}^{(1)}\} \neq \partial\{\lim_{n\to\infty} \hat{g}(\boldsymbol{\beta}^\top \mathbf{x})\}/\partial \boldsymbol{\beta}^{(1)}$, which means that the convergence in probability and the derivation of the sequence $\hat{g}_n(\boldsymbol{\beta}^\top \mathbf{x})$ (as a function of n) cannot commute. This is primarily caused by the fact that the support of the infinite-dimensional nuisance parameter $g(\cdot)$ depends on the finite-dimensional projection parameter $\boldsymbol{\beta}$. In contrast, a semiparametric model where the support of the nuisance parameter is independent of the finite-dimensional parameter is a partially linear regression model having form $Y = \mathbf{X}^\top \boldsymbol{\theta} + \eta(T) + \varepsilon$. It is easy to check that the limit of $\partial \hat{\eta}(T)/\partial \boldsymbol{\theta}$ is equal to $E(\mathbf{X}|T)$, which is the derivative of $\lim_{n\to\infty} \hat{\eta}(T) = E(Y|T) - E(\mathbf{X}^\top|T)\boldsymbol{\theta}$ with respect to $\boldsymbol{\theta}$. Result (iii) ensures that the proposed estimator does not require undersmoothing of $g(\cdot)$ to obtain a root-n consistent estimator for $\boldsymbol{\beta}$ and it is also of its own interest in inference theory for semiparametric models.

2.2. *The asymptotic distribution for the estimates of the parametric part* $\boldsymbol{\beta}$. We will now proceed to the estimation of $\boldsymbol{\beta} \in \Theta$. We need to estimate the $(d-1)$-dimensional vector $\boldsymbol{\beta}^{(1)}$, the estimator of which will be defined via

$$\sum_{i=1}^{n} [\partial \mu\{\hat{g}(\boldsymbol{\beta}^\top \mathbf{X}_i)\}/\partial \boldsymbol{\beta}^{(1)}] V^{-1}\{\hat{g}(\boldsymbol{\beta}^\top \mathbf{X}_i)\}[Y_i - \mu\{\hat{g}(\boldsymbol{\beta}^\top \mathbf{X}_i)\}] = 0. \tag{2.5}$$

This is the direct analogue of the "ideal" estimating equation for known g, in that it is calculated by replacing $g(t)$ with $\hat{g}(t)$. An asymptotically equivalent and easily computed version of this equation is

$$\begin{aligned} \hat{\mathbf{G}}(\boldsymbol{\beta}) &\stackrel{\text{def}}{=} \sum_{i=1}^{n} \mathbf{J}^\top \hat{g}'(\boldsymbol{\beta}^\top \mathbf{X}_i)\{\mathbf{X}_i - \hat{\mathbf{h}}(\boldsymbol{\beta}^\top \mathbf{X}_i)\}\rho_1\{\hat{g}(\boldsymbol{\beta}^\top \mathbf{X}_i)\}[Y_i - \mu\{\hat{g}(\boldsymbol{\beta}^\top \mathbf{X}_i)\}] \\ &= 0 \end{aligned} \tag{2.6}$$

with $\mathbf{J} = \frac{\partial \boldsymbol{\beta}}{\partial \boldsymbol{\beta}^{(1)}}$ the Jacobian mentioned above, $\hat{g}$ and $\hat{g}'$ are defined by (2.1), and $\hat{\mathbf{h}}(t)$ the local linear estimate for $\mathbf{h}(t) = E(\mathbf{X}|\boldsymbol{\beta}^\top \mathbf{X} = t) = (h_1(t), \ldots, h_d(t))^\top$,

$$\hat{\mathbf{h}}(t) = \sum_{i=1}^{n} b_i(t)\mathbf{X}_i \Big/ \sum_{i=1}^{n} b_i(t),$$

where $b_i(t) = K_h(\boldsymbol{\beta}^\top \mathbf{X}_i - t)\{S_{n,2}(t) - (\boldsymbol{\beta}^\top \mathbf{X}_i - t)S_{n,1}(t)\}$, $S_{n,k} = \sum_{i=1}^{n} K_h(\boldsymbol{\beta}^\top \times \mathbf{X}_i - t)(\boldsymbol{\beta}^\top \mathbf{X}_i - t)^k$, $k = 1, 2$. We use (2.6) to estimate $\boldsymbol{\beta}^{(1)}$ in the single-index model, and then use the fact that $\beta_1 = \sqrt{1 - \|\boldsymbol{\beta}^{(1)}\|^2}$ to obtain $\hat{\beta}_1$. The use of (2.6) constitutes in our view a new approach to estimating single-index models; since (2.6) involves smooth pilot estimation of g, g' and $\mathbf{h}$ we call it the Estimation Function Method (EFM) for $\boldsymbol{\beta}$.

REMARK 1. The estimating equations $\hat{\mathbf{G}}(\boldsymbol{\beta})$ can be represented as the gradient vector of the following objective function:

$$\hat{Q}(\boldsymbol{\beta}) = \sum_{i=1}^{n} Q[\mu\{\hat{g}(\boldsymbol{\beta}^\top \mathbf{X}_i)\}, Y_i]$$

with $Q[\mu, y] = \int_{\mu}^{y} \frac{s-y}{V\{\mu^{-1}(s)\}}\, ds$ and $\mu^{-1}(\cdot)$ the inverse function of $\mu(\cdot)$. The existence of such a potential function makes $\hat{\mathbf{G}}(\boldsymbol{\beta})$ to inherit properties of the ideal likelihood score function. Note that $\{\|\boldsymbol{\beta}^{(1)}\| < 1\}$ is an open, connected subset of $\mathbb{R}^{d-1}$. By the regularity conditions assumed on $\mu(\cdot), g(\cdot), V(\cdot)$ (for details see the Appendix), we know that the quasi-likelihood function $\hat{Q}(\boldsymbol{\beta})$ is twice continuously differentiable on $\{\|\boldsymbol{\beta}^{(1)}\| < 1\}$ such that the global maximum of $\hat{Q}(\boldsymbol{\beta})$ can be achieved at some point. One may ask whether the so-

lution is unique and also consistent. Some elementary calculations lead to the Hessian matrix $\partial^2\hat{Q}(\boldsymbol{\beta})/\partial\boldsymbol{\beta}^{(1)}\partial\boldsymbol{\beta}^{(1)\top}$, because the partial derivative $\frac{\partial\mu\{\hat{g}(\boldsymbol{\beta}^\top\mathbf{X}_i)\}}{\partial\boldsymbol{\beta}^{(1)}} = \mu'\{\hat{g}(\boldsymbol{\beta}^\top\mathbf{X}_i)\}\hat{g}'(\boldsymbol{\beta}^\top\mathbf{X}_i)\{\mathbf{X}_i - \hat{\mathbf{h}}(\boldsymbol{\beta}^\top\mathbf{X}_i)\}$, then

$$
\begin{aligned}
&\frac{1}{n}\frac{\partial^2\hat{Q}(\boldsymbol{\beta})}{\partial\boldsymbol{\beta}^{(1)}\partial\boldsymbol{\beta}^{(1)\top}} \\
&\quad= \frac{1}{n}\frac{\partial\hat{\mathbf{G}}(\boldsymbol{\beta})}{\partial\boldsymbol{\beta}^{(1)}} \\
&\quad= \frac{1}{n}\sum_{i=1}^{n}\frac{\partial[\mathbf{J}^\top\hat{g}'(\boldsymbol{\beta}^\top\mathbf{X}_i)\{\mathbf{X}_i - \hat{\mathbf{h}}(\boldsymbol{\beta}^\top\mathbf{X}_i)\}\rho_1\{\hat{g}(\boldsymbol{\beta}^\top\mathbf{X}_i)\}]}{\partial\boldsymbol{\beta}^{(1)}}[Y_i - \mu\{\hat{g}(\boldsymbol{\beta}^\top\mathbf{X}_i)\}] \\
&\qquad- \frac{1}{n}\sum_{i=1}^{n}\mathbf{J}^\top\hat{g}'(\boldsymbol{\beta}^\top\mathbf{X}_i)\{\mathbf{X}_i - \hat{\mathbf{h}}(\boldsymbol{\beta}^\top\mathbf{X}_i)\}\rho_1\{\hat{g}(\boldsymbol{\beta}^\top\mathbf{X}_i)\}\frac{\partial\mu\{\hat{g}(\boldsymbol{\beta}^\top\mathbf{X}_i)\}}{\partial\boldsymbol{\beta}^{(1)}} \\
&\quad= \frac{1}{n}\sum_{i=1}^{n}\Bigg[-\frac{\partial\{\boldsymbol{\beta}^{(1)}/\sqrt{1-\|\boldsymbol{\beta}^{(1)}\|^2}\}}{\partial\boldsymbol{\beta}^{(1)}}\hat{g}'(\boldsymbol{\beta}^\top\mathbf{X}_i)\{\mathbf{X}_{1i} - \hat{h}_1(\boldsymbol{\beta}^\top\mathbf{X}_i)\}\rho_1\{\hat{g}(\boldsymbol{\beta}^\top\mathbf{X}_i)\} \\
&\qquad\qquad+\mathbf{J}^\top\{\mathbf{X}_i - \hat{\mathbf{h}}(\boldsymbol{\beta}^\top\mathbf{X}_i)\}\frac{\partial\hat{g}'(\boldsymbol{\beta}^\top\mathbf{X}_i)}{\partial\boldsymbol{\beta}^{(1)\top}}\rho_1\{\hat{g}(\boldsymbol{\beta}^\top\mathbf{X}_i)\} \\
&\qquad\qquad+\mathbf{J}^\top\hat{g}'(\boldsymbol{\beta}^\top\mathbf{X}_i)\{\mathbf{X}_i - \hat{\mathbf{h}}(\boldsymbol{\beta}^\top\mathbf{X}_i)\}\frac{\partial\rho_1\{\hat{g}(\boldsymbol{\beta}^\top\mathbf{X}_i)\}}{\partial\boldsymbol{\beta}^{(1)\top}} \\
&\qquad\qquad\qquad-\mathbf{J}^\top\hat{g}'(\boldsymbol{\beta}^\top\mathbf{X}_i)\frac{\partial\hat{\mathbf{h}}(\boldsymbol{\beta}^\top\mathbf{X}_i)}{\partial\boldsymbol{\beta}^{(1)}}\rho_1\{\hat{g}(\boldsymbol{\beta}^\top\mathbf{X}_i)\}\Bigg] \\
&\qquad\times[Y_i - \mu\{\hat{g}(\boldsymbol{\beta}^\top\mathbf{X}_i)\}] \\
&\qquad- \frac{1}{n}\sum_{i=1}^{n}\mathbf{J}^\top\hat{g}'^2(\boldsymbol{\beta}^\top\mathbf{X}_i)\{\mathbf{X}_i - \hat{\mathbf{h}}(\boldsymbol{\beta}^\top\mathbf{X}_i)\}\{\mathbf{X}_i - \hat{\mathbf{h}}(\boldsymbol{\beta}^\top\mathbf{X}_i)\}^\top\rho_2\{\hat{g}(\boldsymbol{\beta}^\top\mathbf{X}_i)\}\mathbf{J}.
\end{aligned}
$$

By the regularity conditions in the Appendix, the multipliers of the residuals $[Y_i - \mu\{\hat{g}(\boldsymbol{\beta}^\top\mathbf{X}_i)\}]$ in the first sum of (2.7) are bounded. Mimicking the proof of Proposition 1, the first sum can be shown to converge to 0 in probability as n goes to infinity. The second sum converges to a negative semidefinite matrix. If the Hessian matrix $\frac{1}{n}\frac{\partial^2\hat{Q}(\boldsymbol{\beta})}{\partial\boldsymbol{\beta}^{(1)}\partial\boldsymbol{\beta}^{(1)\top}}$ is negative definite for all values of $\boldsymbol{\beta}^{(1)}$, $\hat{\mathbf{G}}(\boldsymbol{\beta})$ has a unique root. At sample level, however, estimating functions may have more than one root. For the EFM method, the quasi-likelihood $\hat{Q}(\boldsymbol{\beta})$ exists, which can be used to distinguish local maxima from minima. Thus, we suppose (2.6) has a unique solution in the following context.

REMARK 2. It can be seen from the proof in the Appendix that the population version of $\hat{\mathbf{G}}(\boldsymbol{\beta})$ is

$$\mathbf{G}(\boldsymbol{\beta}) = \sum_{i=1}^{n} \mathbf{J}^\top g'(\boldsymbol{\beta}^\top \mathbf{X}_i)\{\mathbf{X}_i - \mathbf{h}(\boldsymbol{\beta}^\top \mathbf{X}_i)\}\rho_1\{g(\boldsymbol{\beta}^\top \mathbf{X}_i)\}[Y_i - \mu\{g(\boldsymbol{\beta}^\top \mathbf{X}_i)\}], \tag{2.7}$$

which is obtained by replacing $\hat{g}, \hat{g}', \hat{\mathbf{h}}$ with $g, g', \mathbf{h}$ in (2.6). One important property of (2.7) is that the second Bartlett identity holds, for any $\boldsymbol{\beta}$:

$$E\{\mathbf{G}(\boldsymbol{\beta})\mathbf{G}^\top(\boldsymbol{\beta})\} = -E\left\{\frac{\partial \mathbf{G}(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}^{(1)}}\right\}.$$

This property makes the semiparametric efficiency of the EFM (2.6) possible.

Let $\boldsymbol{\beta}^0 = (\beta_1^0, \boldsymbol{\beta}^{(1)0^\top})^\top$ denote the true parameter and $\mathbf{B}^+$ denote the Moore–Penrose inverse of any given matrix $\mathbf{B}$. We have the following asymptotic result for the estimator $\hat{\boldsymbol{\beta}}^{(1)}$.

THEOREM 2.1. *Assume the estimating function* (2.6) *has a unique solution and denote it by* $\hat{\boldsymbol{\beta}}^{(1)}$. *If the regularity conditions* (a)–(e) *in the Appendix are satisfied, the following results hold*:

(i) *With* $h \to 0$, $n \to \infty$ *such that* $(nh)^{-1}\log(1/h) \to 0$, $\hat{\boldsymbol{\beta}}^{(1)}$ *converges in probability to the true parameter* $\boldsymbol{\beta}^{(1)0}$.

(ii) *If* $nh^6 \to 0$ *and* $nh^4 \to \infty$,

$$\sqrt{n}\big(\hat{\boldsymbol{\beta}}^{(1)} - \boldsymbol{\beta}^{(1)0}\big) \xrightarrow{\mathcal{L}} N_{d-1}(\mathbf{0}, \boldsymbol{\Sigma}_{\boldsymbol{\beta}^{(1)0}}), \tag{2.8}$$

where $\boldsymbol{\Sigma}_{\boldsymbol{\beta}^{(1)0}} = \{\mathbf{J}^\top \boldsymbol{\Omega}\mathbf{J}\}^+|_{\boldsymbol{\beta}^{(1)}=\boldsymbol{\beta}^{(1)0}}$, $\mathbf{J} = \frac{\partial \boldsymbol{\beta}}{\partial \boldsymbol{\beta}^{(1)}}$ *and*

$$\boldsymbol{\Omega} = E[\{\mathbf{X}\mathbf{X}^\top - E(\mathbf{X}|\boldsymbol{\beta}^\top \mathbf{X})E(\mathbf{X}^\top|\boldsymbol{\beta}^\top \mathbf{X})\}\rho_2\{g(\boldsymbol{\beta}^\top \mathbf{X})\}\{g'(\boldsymbol{\beta}^\top \mathbf{X})\}^2/\sigma^2].$$

REMARK 3. Note that $\boldsymbol{\beta}^\top \boldsymbol{\Omega}\boldsymbol{\beta} = 0$, so the nonnegative matrix $\boldsymbol{\Omega}$ degenerates in the direction of $\boldsymbol{\beta}$. If the mean function μ is the identity function and the variance function is equal to a scale constant, that is, $\mu\{g(\boldsymbol{\beta}^\top \mathbf{X})\} = g(\boldsymbol{\beta}^\top \mathbf{X})$, $\sigma^2 V\{g(\boldsymbol{\beta}^\top \mathbf{X})\} = \sigma^2$, the matrix $\boldsymbol{\Omega}$ in Theorem 2.1 reduces to be

$$\boldsymbol{\Omega} = E[\{\mathbf{X}\mathbf{X}^\top - E(\mathbf{X}|\boldsymbol{\beta}^\top \mathbf{X})E(\mathbf{X}^\top|\boldsymbol{\beta}^\top \mathbf{X})\}\{g'(\boldsymbol{\beta}^\top \mathbf{X})\}^2/\sigma^2].$$

Technically speaking, Theorem 2.1 shows that an undersmoothing approach is unnecessary and that root-n consistency can be achieved. The asymptotic covariance $\boldsymbol{\Sigma}_{\boldsymbol{\beta}^{(1)0}}$ in general can be estimated by replacing terms in its expression by estimates of those terms. The asymptotic normality of $\hat{\boldsymbol{\beta}} = (\hat{\beta}_1, \hat{\boldsymbol{\beta}}^{(1)^\top})^\top$ will follow from Theorem 2.1 with a simple application of the multivariate delta-method,

since $\hat{\beta}_1 = \sqrt{1 - \|\hat{\boldsymbol{\beta}}^{(1)}\|^2}$. According to the results of Carroll et al. (1997), the asymptotic variance of their estimator is $\boldsymbol{\Omega}^+$. Define the block partition of matrix $\boldsymbol{\Omega}$ as follows:

$$\boldsymbol{\Omega} = \begin{pmatrix} \boldsymbol{\Omega}_{11} & \boldsymbol{\Omega}_{12} \\ \boldsymbol{\Omega}_{21} & \boldsymbol{\Omega}_{22} \end{pmatrix}, \tag{2.9}$$

where $\boldsymbol{\Omega}_{11}$ is a positive constant, $\boldsymbol{\Omega}_{12}$ is a $(d-1)$-dimensional row vector, $\boldsymbol{\Omega}_{21}$ is a $(d-1)$-dimensional column vector and $\boldsymbol{\Omega}_{22}$ is a $(d-1) \times (d-1)$ nonnegative definite matrix.

COROLLARY 1. *Under the conditions of Theorem 2.1, we have*

$$\sqrt{n}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^0) \xrightarrow{\mathcal{L}} N_p(\mathbf{0}, \boldsymbol{\Sigma}_{\boldsymbol{\beta}^0}) \tag{2.10}$$

with $\boldsymbol{\Sigma}_{\boldsymbol{\beta}^0} = \mathbf{J}\{\mathbf{J}^\top \boldsymbol{\Omega} \mathbf{J}\}^+ \mathbf{J}^\top|_{\boldsymbol{\beta}=\boldsymbol{\beta}^0}$. *Further,*

$$\boldsymbol{\Sigma}_{\boldsymbol{\beta}^0} \le \boldsymbol{\Omega}^+|_{\boldsymbol{\beta}=\boldsymbol{\beta}^0}$$

and a strict less-than sign holds when $\det(\boldsymbol{\Omega}_{22}) = 0$. *That is, in this case EFM is more efficient than that of Carroll et al.* (1997).

The possible smaller limiting variance derived from the EFM approach partly benefits from the reparameterization so that the quasi-likelihood can be adopted. As we know, the quasi-likelihood is often of optimal property. In contrast, most existing methods treat the estimation of $\boldsymbol{\beta}$ as if it were done in the framework of linear dimension reduction. The target of linear dimension reduction is to find the directions that can linearly transform the original variables vector into a vector of one less dimension. For example, ADE and SIR are two relevant methods. However, when the link function $\mu(\cdot)$ is identity, the limiting variance derived here may not be smaller or equal to the ones of Wang et al. (2010) and Chang, Xue and Zhu (2010) when the quasi-likelihood of (2.5) is applied.

2.3. *Profile quasi-likelihood ratio test.* In applications, it is important to test the statistical significance of added predictors in a regression model. Here we establish a quasi-likelihood ratio statistic to test the significance of certain variables in the linear index. The null hypothesis that the model is correct is tested against a full model alternative. Fan and Jiang (2007) gave a recent review about generalized likelihood ratio tests. Bootstrap tests for nonparametric regression, generalized partially linear models and single-index models have been systematically investigated [see Härdle and Mammen (1993), Härdle, Mammen and Müller (1998),

Härdle, Mammen and Proenca (2001)]. Consider the testing problem:

$$
\begin{aligned}
&H_0 : g(\cdot) = g\left(\sum_{k=1}^{r} \beta_k X_k\right) \\
&\longleftrightarrow \quad H_1 : g(\cdot) = g\left(\sum_{k=1}^{r} \beta_k X_k + \sum_{k=r+1}^{d} \beta_k X_k\right).
\end{aligned} \tag{2.11}
$$

We mainly focus on testing $\beta_k = 0, k = r+1, \ldots, d$, though the following test procedure can be easily extended to a general linear testing $\mathbf{B}\tilde{\boldsymbol{\beta}} = 0$ where $\mathbf{B}$ is a known matrix with full row rank and $\tilde{\boldsymbol{\beta}} = (\beta_{r+1}, \ldots, \beta_d)^\top$. The profile quasi-likelihood ratio test is defined by

$$
T_n = 2\Big\{\sup_{\boldsymbol{\beta}\in\Theta} \hat{Q}(\boldsymbol{\beta}) - \sup_{\boldsymbol{\beta}\in\Theta, \tilde{\boldsymbol{\beta}}=0} \hat{Q}(\boldsymbol{\beta})\Big\}, \tag{2.12}
$$

where $\hat{Q}(\boldsymbol{\beta}) = \sum_{i=1}^{n} Q[\mu\{\hat{g}(\boldsymbol{\beta}^\top \mathbf{X}_i)\}, Y_i]$, $Q[\mu, y] = \int_{\mu}^{y} \frac{s-y}{V\{\mu^{-1}(s)\}}\, ds$ and $\mu^{-1}(\cdot)$ is the inverse function of $\mu(\cdot)$. The following Wilks type theorem shows that the distribution of T_n is asymptotically chi-squared and independent of nuisance parameters.

THEOREM 2.2. *Under the assumptions of Theorem 2.1, if* $\beta_k = 0, k = r+1, \ldots, d$, *then*

$$
T_n \xrightarrow{\mathcal{L}} \chi^2(d-r). \tag{2.13}
$$

3. Numerical studies.

3.1. *Computation of the estimates.* Solving the joint estimating equations (2.1) and (2.6) poses some interesting challenges, since the functions $\hat{g}(\boldsymbol{\beta}^\top \mathbf{X})$ and $\hat{g}'(\boldsymbol{\beta}^\top \mathbf{X})$ depend on $\boldsymbol{\beta}$ implicitly. Treating $\boldsymbol{\beta}^\top X$ as a new predictor (with given $\boldsymbol{\beta}$), (2.1) gives us $\hat{g}, \hat{g}'$ as in Fan, Heckman and Wand (1995). We therefore focus on (2.6), as estimating equations. It cannot be solved explicitly, and hence one needs to find solutions using numerical methods. The Newton–Raphson algorithm is one of the popular and successful methods for finding roots. However, the computational speed of this algorithm crucially depends on the initial value. We propose therefore a fixed-point iterative algorithm that is not very sensitive to starting values and is adaptive to larger dimension. It is worth noting that this algorithm can be implemented in the case that d is slightly larger than n, because the resultant procedure only involves one-dimensional nonparametric smoothers, thereby avoiding the data sparsity problem caused by high dimensionality.

Rewrite the estimating functions as $\hat{\mathbf{G}}(\boldsymbol{\beta}) = \mathbf{J}^\top \hat{\mathbf{F}}(\boldsymbol{\beta})$ with

$$
\hat{\mathbf{F}}(\boldsymbol{\beta}) = (\hat{F}_1(\boldsymbol{\beta}), \ldots, \hat{F}_d(\boldsymbol{\beta}))^\top
$$

and

$$\hat{F}_s(\boldsymbol{\beta}) = \sum_{i=1}^{n} \{X_{si} - \hat{h}_s(\boldsymbol{\beta}^\top \mathbf{X}_i)\} \mu'\{\hat{g}(\boldsymbol{\beta}^\top \mathbf{X}_i)\} \hat{g}'(\boldsymbol{\beta}^\top \mathbf{X}_i) V^{-1} \{\hat{g}(\boldsymbol{\beta}^\top \mathbf{X}_i)\} \times [Y_i - \mu\{\hat{g}(\boldsymbol{\beta}^\top \mathbf{X}_i)\}].$$

Setting $\hat{\mathbf{G}}(\boldsymbol{\beta}) = 0$, we have that

$$\begin{cases} -\beta_2 \hat{F}_1(\boldsymbol{\beta}) / \sqrt{1 - \|\boldsymbol{\beta}^{(1)}\|^2} + \hat{F}_2(\boldsymbol{\beta}) = 0, \\ -\beta_3 \hat{F}_1(\boldsymbol{\beta}) / \sqrt{1 - \|\boldsymbol{\beta}^{(1)}\|^2} + \hat{F}_3(\boldsymbol{\beta}) = 0, \\ \cdots \\ -\beta_d \hat{F}_1(\boldsymbol{\beta}) / \sqrt{1 - \|\boldsymbol{\beta}^{(1)}\|^2} + \hat{F}_d(\boldsymbol{\beta}) = 0. \end{cases} \tag{3.1}$$

Note that $\|\boldsymbol{\beta}^{(1)}\|^2 = \sum_{r=2}^{d} \beta_r^2$, $\beta_1 = \sqrt{1 - \|\boldsymbol{\beta}^{(1)}\|^2}$ and after some simple calculations, we can get that

$$\begin{cases} \beta_1 = |\hat{F}_1(\boldsymbol{\beta})| / \|\hat{\mathbf{F}}(\boldsymbol{\beta})\|, & s = 1, \\ \beta_s^2 = \hat{F}_s^2(\boldsymbol{\beta}) / \|\hat{\mathbf{F}}(\boldsymbol{\beta})\|^2, & s \geq 2, \end{cases}$$

and $\operatorname{sign}\{\beta_s \hat{F}_1(\boldsymbol{\beta})\} = \operatorname{sign}\{\hat{F}_s(\boldsymbol{\beta})\}$, $s \geq 2$. The above equation can also be rewritten as

$$\boldsymbol{\beta} \frac{\hat{F}_1(\boldsymbol{\beta})}{\|\hat{\mathbf{F}}(\boldsymbol{\beta})\|} = \frac{|\hat{F}_1(\boldsymbol{\beta})|}{\|\hat{\mathbf{F}}(\boldsymbol{\beta})\|} \times \frac{\hat{\mathbf{F}}(\boldsymbol{\beta})}{\|\hat{\mathbf{F}}(\boldsymbol{\beta})\|}. \tag{3.2}$$

Then solving the equation (2.6) is equivalent to finding a fixed point for (3.2). Though $\|\boldsymbol{\beta}^{(1)}\| < 1$ holds almost surely in (3.2) and always $\|\boldsymbol{\beta}\| = 1$, there will be some trouble if (3.2) is directly used as iterative equations. Note that the value of $\|\hat{\mathbf{F}}(\boldsymbol{\beta})\|$ is used as denominator that may sometimes be small, which potentially makes the algorithm unstable. On the other hand, the convergence rate of the fixed-point iterative algorithm derived from (3.2) depends on L, where $\|\frac{\partial\{\hat{\mathbf{F}}(\boldsymbol{\beta})|/\|\hat{\mathbf{F}}(\boldsymbol{\beta})\|\}}{\partial \boldsymbol{\beta}}\| \leq L$. For a fast convergence rate, it technically needs a shrinkage value L. An ad hoc fix introduces a constant M, adding $M\boldsymbol{\beta}$ on both sides of (3.2) and dividing by $\hat{F}_1(\boldsymbol{\beta})/\|\hat{\mathbf{F}}(\boldsymbol{\beta})\| + M$:

$$\boldsymbol{\beta} = \frac{M}{\hat{F}_1(\boldsymbol{\beta})/\|\hat{\mathbf{F}}(\boldsymbol{\beta})\| + M} \boldsymbol{\beta} + \frac{|\hat{F}_1(\boldsymbol{\beta})|/\|\hat{\mathbf{F}}(\boldsymbol{\beta})\|^2}{\hat{F}_1(\boldsymbol{\beta})/\|\hat{\mathbf{F}}(\boldsymbol{\beta})\| + M} \hat{\mathbf{F}}(\boldsymbol{\beta}),$$

where M is chosen such that $\hat{F}_1(\boldsymbol{\beta})/\|\hat{\mathbf{F}}(\boldsymbol{\beta})\| + M \neq 0$. In addition, to accelerate the rate of convergence, we reduce the derivative of the term on the right-hand side of the above equality, which can be achieved by choosing some appropriate M. This is the iteration formulation in Step 2. Here the norm of $\boldsymbol{\beta}_{new}$ is not equal to 1 and we have to normalize it again. Since the iteration in Step 2 makes $\boldsymbol{\beta}_{new}$

to violate the identifiability constraint with norm 1, we design (3.2) to include the whole $\boldsymbol{\beta}$ vector. The possibility of renormalization for $\boldsymbol{\beta}_{new}$ avoids the difficulty of controlling $\|\boldsymbol{\beta}_{new}^{(1)}\| < 1$ in each iteration in Step 2.

Based on these observations, the fixed-point iterative algorithm is summarized as:

Step 0. Choose initial values for $\boldsymbol{\beta}$, denoted by $\boldsymbol{\beta}_{old}$.

Step 1. Solve the estimating equation (2.1) with respect to $\boldsymbol{\alpha}$, which yields $\hat{g}(\boldsymbol{\beta}_{old}^{\top}\mathbf{x}_i)$ and $\hat{g}'(\boldsymbol{\beta}_{old}^{\top}\mathbf{x}_i)$, $1 \le i \le n$.

Step 2. Update $\boldsymbol{\beta}_{old}$ with $\boldsymbol{\beta}_{old} = \boldsymbol{\beta}_{new}/\|\boldsymbol{\beta}_{new}\|$ by solving the equation (2.6) in the fixed-point iteration

$$\boldsymbol{\beta}_{new} = \frac{M}{\hat{F}_1(\boldsymbol{\beta}_{old})/\|\hat{F}(\boldsymbol{\beta}_{old})\| + M}\boldsymbol{\beta}_{old} + \frac{|\hat{F}_1(\boldsymbol{\beta}_{old})|/\|\hat{F}(\boldsymbol{\beta}_{old})\|^2}{\hat{F}_1(\boldsymbol{\beta}_{old})/\|\hat{F}(\boldsymbol{\beta}_{old})\| + M}\hat{\mathbf{F}}(\boldsymbol{\beta}_{old}),$$

where M is a constant satisfying $\hat{F}_1(\boldsymbol{\beta})/\|\hat{F}(\boldsymbol{\beta})\| + M \neq 0$ for any $\boldsymbol{\beta}$.

Step 3. Repeat Steps 1 and 2 until $\max_{1\le s\le d} |\beta_{new,s} - \beta_{old,s}| \le tol$ is met with *tol* being a prescribed tolerance.

The final vector $\boldsymbol{\beta}_{new}/\|\boldsymbol{\beta}_{new}\|$ is the estimator of $\boldsymbol{\beta}^0$. Similarly to other direct estimation methods [Horowitz and Härdle (1996)], the preceding calculation is easy to implement. Empirically the initial value for $\boldsymbol{\beta}$, $(1, 1, \ldots, 1)^{\top}/\sqrt{d}$ can be used in the calculations. The Epanechnikov kernel function $K(t) = 3/4(1 - t^2)I(|t| \le 1)$ is used. The bandwidth involved in Step 1 can be chosen to be optimal for estimation of $\hat{g}(t)$ and $\hat{g}'(t)$ based on the observations $\{\boldsymbol{\beta}_{old}^{\top}\mathbf{X}_i, Y_i\}$. So the standard bandwidth selection methods, such as K-fold cross-validation, generalized cross-validation (GCV) and the rule of thumb, can be adopted. In this step, we recommend K-fold cross-validation to determine the optimal bandwidth using the quasi-likelihood as a criterion function. The K-fold cross-validation is not too computationally intensive while making K not take too large values (e.g., $K = 5$). Here we recommend trying a number of smoothing parameters that smooth the data and picking the one that seems most reasonable. As an adjustment factor, M will increase the stability of iteration. Ideally, in each iteration an optimum value for M should be chosen guaranteeing that the derivative on the right-hand side of the iteration formulation in Step 2 is close to zero. Following this idea, M will be depending the changes of $\boldsymbol{\beta}$ and $\hat{\mathbf{F}}(\boldsymbol{\beta})/\|\hat{\mathbf{F}}(\boldsymbol{\beta})\|$. This will be an expensive task due to the computation for the derivative on the right-hand side of the iteration formulation in Step 2. We therefore consider M as constant nonvarying in each iteration, and select M by the K-fold cross-validation method, according to minimizing the model prediction error. When the dimension d gets larger, M will get smaller. In our simulation runs, we empirically search M in the interval $[2/\sqrt{d}, d/2]$. This choice gives pretty good practical performance.

3.2. *Simulation results.*

EXAMPLE 1 (Continuous response). We report a simulation study to investigate the finite-sample performance of the proposed estimator and compare it with the rMAVE [refined MAVE; for details see Xia et al. (2002)] estimator and the EDR estimator [see Hristache et al. (2001), Polzehl and Sperlich (2009)]. We consider the following model similar to that used in Xia (2006):

$$\begin{aligned} &E(Y|\boldsymbol{\beta}^\top\mathbf{X}) = g(\boldsymbol{\beta}^\top\mathbf{X}), \qquad g(\boldsymbol{\beta}^\top\mathbf{X}) = (\boldsymbol{\beta}^\top\mathbf{X})^2\exp(\boldsymbol{\beta}^\top\mathbf{X}); \\ &\mathrm{Var}(Y|\boldsymbol{\beta}^\top\mathbf{X}) = \sigma^2, \qquad \sigma = 0.1. \end{aligned} \tag{3.3}$$

Let the true parameter $\boldsymbol{\beta} = (2, 1, 0, \ldots, 0)^\top/\sqrt{5}$. Two sets of designs for $\mathbf{X}$ are considered: Design (A) and Design (B). In Design (A), $(X_s + 1)/2 \sim \mathrm{Beta}(\tau, 1)$, $1 \le s \le d$ and, in Design (B), $(X_1 + 1)/2 \sim \mathrm{Beta}(\tau, 1)$ and $P(X_s = \pm 0.5) = 0.5$, $s = 2, 3, 4, \ldots, d$. The data generated in Design (A) are not elliptically symmetric. All the components of Design (B) are discrete except for the first component X_1. Y is generated from a normal distribution. This simulation data set consists of 400 observations with 250 replications. The results are shown in Table 1. All rMAVE, EDR and EFM estimates are close to the true parameter vector for $d = 10$. However, the average estimation errors from rMAVE and EDR estimates for $d = 50$ are about 2 and 1.5 times as large as those of the EFM estimates, respectively. This indicates that the fixed-point algorithm is more adaptive to high dimension.

EXAMPLE 2 (Binary response). This simulation design assumes an underlying single-index model for binary responses with

$$\begin{aligned} P(Y = 1|\mathbf{X}) &= \mu\{g(\boldsymbol{\beta}^\top\mathbf{X})\} = \exp\{g(\boldsymbol{\beta}^\top\mathbf{X})\}/[1 + \exp\{g(\boldsymbol{\beta}^\top\mathbf{X})\}], \\ g(\boldsymbol{\beta}^\top\mathbf{X}) &= \exp(5\boldsymbol{\beta}^\top\mathbf{X} - 2)/\{1 + \exp(5\boldsymbol{\beta}^\top\mathbf{X} - 3)\} - 1.5. \end{aligned} \tag{3.4}$$

The underlying coefficients are assumed to be $\boldsymbol{\beta} = (2, 1, 0, \ldots, 0)^\top/\sqrt{5}$. We consider two sets of designs: Design (C) and Design (D). In Design (C), X_1 and X_2

TABLE 1
Average estimation errors $\sum_{s=1}^{d} |\hat{\beta}_s - \beta_s|$ *for model* (3.3)

		Design (A)			Design (B)		
d	τ	rMAVE	EDR	EFM	rMAVE	EDR	EFM
10	0.75	0.0559*	0.0520	0.0792	0.0522*	0.0662	0.0690
10	1.5	0.0323*	0.0316	0.0298	0.0417*	0.0593	0.0457
50	0.75	0.9900	0.7271	0.5425	0.9780	0.7712	0.4515
50	1.5	0.3776	0.3062	0.1796	0.4693	0.4103	0.2211

*The values are adopted from Xia (2006).

TABLE 2
Average estimation errors $\sum_{s=1}^{d} |\hat{\beta}_s - \beta_s|$ *for model* (3.4)

	Design (C)			Design (D)		
d	**rMAVE**	**EDR**	**EFM**	**rMAVE**	**EDR**	**EFM**
10	0.5017	0.5281	0.4564	0.9614	0.9574	0.7415
50	2.0991	1.2695	1.1744	2.5040	2.4846	1.9908

follow the uniform distribution $U(-2, 2)$. In Design (D), X_1 is also assumed to be uniformly distributed in interval $(-2, 2)$ and $(X_2 + 1)/2 \sim \text{Beta}(1, 1)$. Similar designs for generalized partially linear single-index models are assumed in Kane, Holt and Allen (2004). Here a sample size of 700 is used for the case $d = 10$ and 3,000 is used for $d = 50$. Different sample sizes from Example 1 are used due to varying complexity of the two examples. For this example, 250 replications are simulated and the results are displayed in Table 2. In this set of simulations, the average estimation errors from rMAVE estimates and EDR estimates are about 1.5 and 1.2 times as large as EFM estimates, under both Design (C) and Design (D) for $d = 10$ or $d = 50$. The values in the row marked by $d = 50$ look a little bigger. However, it is reasonable because the number of summands in the average estimate error for $d = 50$ is five times as large as that for $d = 10$. Again it appears that the EFM procedure achieves more precise estimators.

EXAMPLE 3 (A simple model). To illustrate the adaptivity of our algorithm to high dimension, we consider the following simple single-index model:

$$Y = (\boldsymbol{\beta}^\top \mathbf{X})^2 + \varepsilon. \tag{3.5}$$

The true parameter is $\boldsymbol{\beta} = (2, 1, 0, \ldots, 0)^\top/\sqrt{5}$; $\mathbf{X}$ is generated from $N_d(2, \mathbf{I})$. Both homogeneous errors and heterogeneous ones are considered. In the former case, $\varepsilon \sim N(0, 0.2^2)$ and in the latter case, $\varepsilon = \exp(\sqrt{5}\boldsymbol{\beta}^\top \mathbf{X}/14)\widetilde{\varepsilon}$ with $\widetilde{\varepsilon} \sim N(0, 1)$. The latter case is designed to show whether our method can handle heteroscedasticity. A similar modeling setup was also used in Wang and Xia (2008), Example 5. The simulated results given in Table 3 are based on 250 replicates with a sample of $n = 100$ observations. An important observation from this simulation is that the proposed EFM approach still works even when the dimension of the parameter is equal to or slightly larger than the number of observations. It can be seen from Table 3 that our approach also performs well under the heteroscedasticity setup.

EXAMPLE 4 (An oscillating function model). A single-index model is designed as

$$Y = \sin(a\boldsymbol{\beta}^\top \mathbf{X}) + \varepsilon, \tag{3.6}$$

TABLE 3
Average estimation errors $\sum_{s=1}^{d} |\hat{\beta}_s - \beta_s|$ *for model* (3.5)

ε		$d = 10$	$d = 50$	$d = 100$	$d = 120$
$\varepsilon \sim N(0, 0.2^2)$	rMAVE	0.0318	0.3484	—	—
	EDR	0.0363	0.5020	—	—
	EFM	0.0272	0.2302	2.9409	5.0010
$\varepsilon \sim N(0, \exp(\frac{2X_1+X_2}{7}))$	rMAVE	0.3427	4.6190	—	—
	EDR	0.2542	2.1112	—	—
	EFM	0.2201	1.7937	4.1435	6.4973

— means that the values cannot be calculated by rMAVE and EDR because of high dimension.

where $\boldsymbol{\beta} = (2, 1, 0, \ldots, 0)^\top/\sqrt{5}$, $\mathbf{X}$ is generated from $N_d(2, \mathbf{I})$ and $\varepsilon \sim N(0, 0.2^2)$. The number of replications is 250 and the sample size $n = 400$. The simulation results are shown in Table 4. In these chosen values for a, we see that EFM performs better than rMAVE and EDR. But as is understood, more oscillating functions are more difficult to handle than those less oscillating functions.

EXAMPLE 5 (Comparison of variance). To make our simulation results comparable with those of Carroll et al. (1997), we mimic their simulation setup. Data of size 200 are generated according to the following model:

$$Y_i = \sin\{\pi(\boldsymbol{\beta}^\top \mathbf{X}_i - A)/(B - A)\} + \alpha Z_i + \varepsilon_i, \tag{3.7}$$

where $\mathbf{X}_i$ are trivariate with independent $U(0, 1)$ components, Z_i are independent of $\mathbf{X}_i$ and $Z_i = 0$ are for i odd and $Z_i = 1$ for i even, and ε_i follow a normal distribution $N(0, 0.01)$ independent of both $\mathbf{X}_i$ and Z_i. The parameters are taken to be $\boldsymbol{\beta} = (1, 1, 1)^\top/\sqrt{3}$, $\alpha = 0.3$, $A = \sqrt{3}/2 - 1.645/\sqrt{12}$ and $B = \sqrt{3}/2 + 1.645/\sqrt{12}$. Note that the EFM approach can still be applicable for this model as the conditionally centered response Y given Z has the model as, because of the independence between $\mathbf{X}$ and Z,

$$Y_i - E(Y_i|Z_i) = a + \sin\{\pi(\boldsymbol{\beta}^\top \mathbf{X}_i - A)/(B - A)\} + \varepsilon_i.$$

TABLE 4
Average estimation errors $\sum_{s=1}^{d} |\hat{\beta}_s - \beta_s|$ *for model* (3.6)

	$a = \pi/2$			$a = 3\pi/4$		
d	rMAVE	EDR	EFM	rMAVE	EDR	EFM
10	0.0981	0.0918	0.0737	0.0970	0.0745	0.0725
50	0.5247	0.6934	0.4355	0.6350	1.8484	0.5407

TABLE 5
Estimation for $\boldsymbol{\beta}$ of model (3.7) based on two randomly chosen samples

	One group of sample			Another group of sample		
	X_1	X_2	X_3	X_1	X_2	X_3
GPLSIM est.	0.595*	0.568*	0.569*	0.563*	0.574*	0.595*
GPLSIM s.e.	0.013*	0.013*	0.013*	0.010*	0.010*	0.010*
EFM est.	0.579	0.575	0.577	0.573	0.577	0.580
EFM s.e.	0.011	0.011	0.011	0.010	0.010	0.010

*The values are adopted from Carroll et al. (1997). We abbreviate "estimator" to "est." and "standard error" to "s.e.," which are computed from the sample version of $\boldsymbol{\Sigma}_{\hat{\boldsymbol{\beta}}}$ defined in (2.10).

As Z_i are dummy variables, estimating $E(Y_i|Z_i)$ is simple. Thus, when we regard $Y_i - E(Y_i|Z_i)$ as response, the model is still a single-index model. Here the number of replications is 100. The method derived from Carroll et al. (1997) is referred to be the GLPSIM approach. The numerical results are reported in Table 5. It shows that compared with the GPLSIM estimates, the EFM estimates have smaller bias and smaller (or equal) variance. Also in this example both EFM and GPLSIM can provide reasonably accurate estimates.

Performance of profile quasi-likelihood ratio test. To illustrate how the profile quasi-likelihood ratio performs for linear hypothesis problems, we simulate the same data as above, except that we allow some components of the index to follow the null hypothesis:

$$H_0 : \beta_4 = \beta_5 = \cdots = \beta_d = 0.$$

We examine the power of the test under a sequence of the alternative hypotheses indexed by parameter δ as follows:

$$H_1 : \beta_4 = \delta, \qquad \beta_s = 0 \qquad \text{for } s \geq 5.$$

When $\delta = 0$, the alternative hypothesis becomes the null hypothesis.

We examine the profile quasi-likelihood ratio test under a sequence of alternative models, progressively deviating from the null hypothesis, namely, as δ increases. The power functions are calculated at the significance level: 0.05, using the asymptotic distribution. We calculate test statistics from 250 simulations by employing the fixed-point algorithm and find the percentage of test statistics greater than or equal to the associated quantile of the asymptotic distribution. The pictures in Figures 1, 2 and 3 illustrate the power function curves for two models under the given significance levels. The power curves increase rapidly with δ, which shows the profile quasi-likelihood ratio test is powerful. When δ is close to 0, the test sizes are all approximately the significance levels.

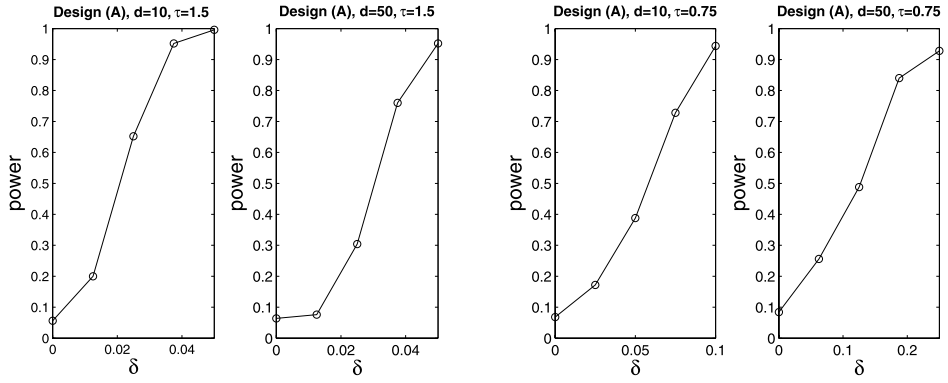


FIG. 1. *Simulation results for Design* (A) *in Example* 1. *The left graphs depict the case* $\tau = 1.5$ *with* τ *the first parameter in* Beta$(\tau, 1)$. *The right graphs are for* $\tau = 0.75$.

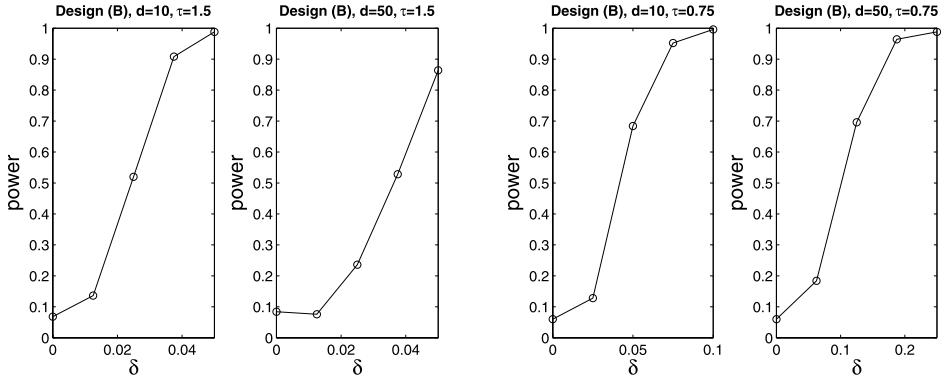


FIG. 2. *Simulation results for Design* (B) *in Example* 1. *The left graphs depict the case* $\tau = 1.5$ *with* τ *the first parameter in* Beta$(\tau, 1)$. *The right graphs are for* $\tau = 0.75$.

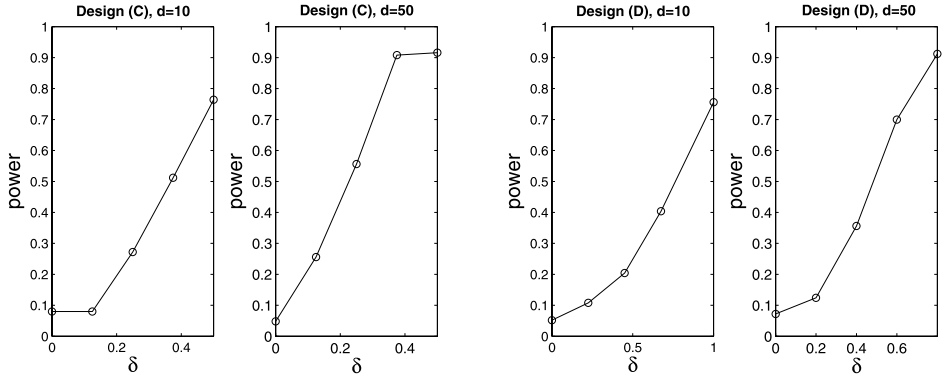


FIG. 3. *Simulation results for Example* 2. *The left graphs depict the case of Design* (C) *with parameter dimension being* 10 *and* 50. *The right graphs are for Design* (D).

3.3. *A real data example*. Income, to some extent, is considered as an index of a successful life. It is generally believed that demographic information, such as education level, relationship in the household, marital status, the fertility rate and gender, among others, has effects on amounts of income. For example, Murray (1997) illustrated that adults with higher intelligence have higher income. Kohavi (1996) predicted income using a Bayesian classifier offered by a machine learning algorithm. Madalozzo (2008) examined income differentials between married women and those who remain single or cohabit by using multivariate linear regression. Here we will use the single-index model to explore the relationship between income and some of its possible determinants.

We use the "Adult" database, which was extracted from the Census Bureau database and is available on website: http://archive.ics.uci.edu/ml/datasets/Adult. It was originally used to model income exceeds over USD 50,000/year based on census data. The purpose of using this example is to understand the personal income patterns and demonstrate the performance of the EFM method in real data analysis. After excluding a few missing data, the data set in our study includes 30,162 subjects. The selected explanatory variables are:

- *sex* (categorical): 1 = Male, 0 = Female.
- *native-country* (categorical): 1 = United-States, 0 = others.
- *work-class* (categorical): 1 = Federal-gov, 2 = Local-gov, 3 = Private, 4 = Self-emp-inc (self-employed, incorporated), 5 = Self-emp-not-inc (self-employed, not incorporated), 6 = State-gov.
- *marital-status* (categorical): 1 = Divorced, 2 = Married-AF-spouse (married, armed forces spouse present), 3 = Married-civ-spouse (married, civilian spouse present), 4 = Married-spouse-absent [married, spouse absent (exc. separated)], 5 = Never-married, 6 = Separated, 7 = Widowed.
- *occupation* (categorical): 1 = Adm-clerical (administrative support and clerical), 2 = Armed-Forces, 3 = Craft-repair, 4 = Exec-managerial (executive-managerial), 5 = Farming-fishing, 6 = Handlers-cleaners, 7 = Machine-op-inspct (machine operator inspection), 8 = Other-service, 9 = Priv-house-serv (private household services), 10 = Prof-specialty (professional specialty), 11 = Protective-serv, 12 = Sales, 13 = Tech-support, 14 = Transport-moving.
- *relationship* (categorical): 1 = Husband, 2 = Not-in-family, 3 = Other-relative, 4 = Own-child, 5 = Unmarried, 6 = Wife.
- *race* (categorical): 1 = Amer-Indian-Eskimo, 2 = Asian-Pac-Islander, 3 = Black, 4 = Other, 5 = White.
- *age* (integer): number of years of age and greater than or equal to 17.
- *fnlwgt* (continuous): The final sampling weights on the CPS files are controlled to independent estimates of the civilian noninstitutional population of the United States.
- *education* (ordinal): 1 = Preschool (less than 1st Grade), 2 = 1st–4th, 3 = 5th–6th, 4 = 7th–8th, 5 = 9th, 6 = 10th, 7 = 11th, 8 = 12th (12th Grade no

Diploma), 9 = HS-grad (high school Grad-Diploma or Equiv), 10 = Some-college (some college but no degree), 11 = Assoc-voc (associate degree-occupational/vocational), 12 = Assoc-acdm (associate degree-academic program), 13 = Bachelors, 14 = Masters, 15 = Prof-school (professional school), 16 = Doctorate.

- *education-num* (continuous): Number of years of education.
- *capital-gain* (continuous): A profit that results from investments into a capital asset.
- *capital-loss* (continuous): A loss that results from investments into a capital asset.
- *hours-per-week* (continuous): Usual number of hours worked per week.

Note that all the explanatory variables up to "age" are categorical with more than two categories. As such, we use dummy variables to link up the corresponding categories. Specifically, for every original explanatory variable up to "age," we use dummy variables to indicate it in which the number of dummy variables is equal to the number of categories minus one. By doing so, we then have 41 explanatory variables, where the first 35 ones are dummy and the remaining ones are continuous. After a preliminary data check, we find that the explanatory variables X_{37} = "fnlwgt," X_{39} = "capital-gain" and X_{40} = "capital-loss" are very skewed to the left and the latter two often take zero value. Before fitting (3.8) we first make a logarithm transformation for these three variables to have log("fnlwgt"), log(1 + "capital-gain") and log(1 + "capital-loss"). To make the explanatory variables comparable in scale, we standardize each of them individually to obtain mean 0 and variance 1. Since "education" and "education-num" are correlated, "education" is dropped from the model and it results in a significantly smaller mean residual deviance.

The single-index model will be used to model the relationship between income and the relevant 43 predictors $\mathbf{X} = (X_1, \ldots, X_{43})^\top$:

$$P(\text{"income"} > 50{,}000|\mathbf{X}) = \exp\{g(\boldsymbol{\beta}^\top \mathbf{X})\}/[1 + \exp\{g(\boldsymbol{\beta}^\top \mathbf{X})\}], \tag{3.8}$$

where $Y = I(\text{"income"} > 50{,}000)$ and $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_{43})^\top$ and β_s represents the effect of the sth predictor. Formally, we are testing the effect of gender, that is,

$$H_0 : \beta_1 = 0 \quad \longleftrightarrow \quad H_1 : \beta_1 \neq 0. \tag{3.9}$$

The fixed-point iterative algorithm is employed to compute the estimate for $\boldsymbol{\beta}$. To illustrate further the practical implications of this approach, we compare our results to those obtained by using an ordinary logistic regression (LR). The coefficients of the two models are given in Table 6. To make the analyses presented in the table comparable, we consider two standardizations. First, we standardize every explanatory variable with mean 0 and variance 1 so that the coefficients can be used to compare the relative influence from different explanatory variables. However, such a standardization does not allow us to compare between the single-index

TABLE 6
Fitted coefficients for model (3.8) (*estimated standard errors in parentheses*)

Variables	$\hat{\boldsymbol{\beta}}$ **of SIM**	$\hat{\boldsymbol{\beta}}$ **of LR**
Sex	0.1102 (0.0028)	0.1975 (0.0181)
Native-country	0.0412 (0.0027)	0.0354 (0.0116)
Work-class		
Federal-gov	0.1237 (0.0059)	0.0739 (0.0108)
Local-gov	0.2044 (0.0065)	0.0155 (0.0135)
Private	−0.2603 (0.0075)	0.0775 (0.0200)
Self-em-inc	0.1252 (0.0068)	0.0520 (0.0112)
Self-emp-not-inc	0.1449 (0.0066)	−0.0157 (0.0147)
Marital-Status		
Divorced	−0.0353 (0.0061)	−0.0304 (0.0264)
Married-AF-spouse	0.0195 (0.0036)	0.0333 (0.0079)
Married-civ-spouse	0.3257 (0.0150)	0.4545 (0.0754)
Married-spouse-absent	−0.0115 (0.0029)	−0.0095 (0.0146)
Never-married	−0.1876 (0.0085)	−0.1452 (0.0370)
Separated	−0.0412 (0.0050)	−0.0221 (0.0179)
Occupation		
Adm-clerical	−0.0302 (0.0050)	0.0131 (0.0164)
Armed-Forces	−0.0086 (0.0031)	−0.0091 (0.0131)
Craft-repair	−0.0913 (0.0050)	0.0263 (0.0146)
Exec-managerial	0.1813 (0.0061)	0.1554 (0.0148)
Farming-fishing	−0.0370 (0.0036)	−0.0772 (0.0125)
Handlers-cleaners	−0.0947 (0.0033)	−0.0662 (0.0153)
Machine-op-inspct	−0.1067 (0.0038)	−0.0290 (0.0133)
Other-service	−0.1227 (0.0045)	−0.1192 (0.0195)
Priv-house-serv	−0.0501 (0.0020)	−0.0833 (0.0379)
Prof-specialty	0.2502 (0.0065)	0.1153 (0.0160)
Protective-serv	0.1954 (0.0061)	0.0508 (0.0095)
Sales	0.0316 (0.0050)	0.0615 (0.0147)
Tech-support	0.0181 (0.0037)	0.0619 (0.0102)
Relationship		
Husband	−0.1249 (0.0093)	−0.3264 (0.0254)
Not-in-family	−0.0932 (0.0093)	−0.2074 (0.0612)
Other-relative	−0.0958 (0.0038)	−0.1498 (0.0219)
Own-child	−0.2218 (0.0076)	−0.3769 (0.0498)
Unmarried	−0.1124 (0.0067)	−0.1739 (0.0446)
Race		
Amer-Indian-Eskimo	−0.0252 (0.0024)	−0.0226 (0.0109)
Asian-Pac-Islander	0.0114 (0.0030)	0.0062 (0.0101)
Black	−0.0300 (0.0024)	−0.0182 (0.0111)
Other	−0.0335 (0.0021)	−0.0286 (0.0129)

TABLE 6
(*Continued*)

Variables	$\hat{\boldsymbol{\beta}}$ **of SIM**	$\hat{\boldsymbol{\beta}}$ **of LR**
Age	0.2272 (0.0042)	0.1798 (0.0111)
Fnlwgt	0.0099 (0.0028)	0.0414 (0.0092)
Education-num	0.4485 (0.0045)	0.3732 (0.0122)
Capital-gain	0.2859 (0.0055)	0.2582 (0.0084)
Capital-loss	0.1401 (0.0042)	0.1210 (0.0078)
Hours-per-week	0.2097 (0.0035)	0.1823 (0.0101)

model and the ordinary logistic regression model. We then further normalize the coefficients to be with Euclidean norm 1, and then the estimates of their standard errors are also adjusted accordingly. The single-index model provides more reasonable results: X_{38} = "education-num" has its strongest positive effect on income; those who got a bachelor's degree or higher seem to have much higher income than those with lower education level. In contrast, results derived from a logistic regression show that "married-civ-spouse" is the largest positive contributor.

Some other interesting conclusions could be obtained by looking at the output. Both "sex" and "native-country" have a positive effect. Persons who worked without pay in a family business, unpaid childcare and others earn a lower income than persons who worked for wages or for themselves. The "fnlwgt" attribute has a positive relation to income. Males are likely to make much more money than females. The expected sign for marital status except the *married* (married-AF-spouse, married-civ-spouse) is negative, given that the household production theory affirms that division of work is efficient when each member of a family dedicates his or her time to the more productive job. Men usually receive relatively better compensation for their time in the labor market than in home production. Thus, the expectation is that married women dedicate more time to home tasks and less to the labor market, and this would imply a different probability of working given the marital status choice.

Also "race" influences the income and Asian or Pacific Islanders seem to make more money than other races. And also, one's income significantly increases as working hours increase. Both "capital-gain" and "capital-loss" have positive effects, so we think that people make more money who can use more money to invest. The presence of young children has a negative influence on the income. "age" accounts for the experience effect and has a positive effect. Hence the conclusion based on the single-index model is consistent with what we expect.

To help with interpretation of the model, plots of $\boldsymbol{\beta}^\top \mathbf{X}$ versus predicted response probability and $\hat{g}(\boldsymbol{\beta}^\top \mathbf{X})$ are generated, respectively, and can be found on the right column in Figure 4. When the estimated single-index is greater than 0, $\hat{g}(\hat{\boldsymbol{\beta}}\mathbf{X})$ shows some degree of curvature. An alternative choice is to fit the data

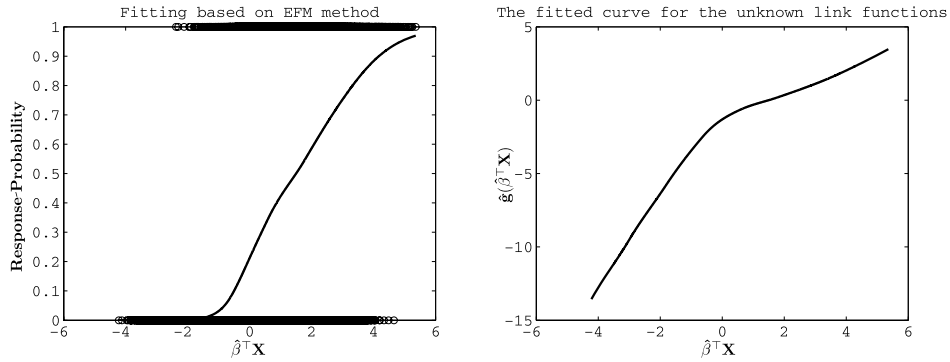


FIG. 4. *Adult data*: *The left graph is a plot of predicted response probability based on the single-index model. The right graph is the fitted curve for the unknown link function* $g(\cdot)$.

using generalized partially linear additive models (GPLAM) with nonparametric components of continuous explanatory variables. The relationships among "age," "fnlwgt," "capital-gain," "capital-loss" and "hours-per-week" all show nonlinearity. The mean residual deviances of SIM, LR and GPLAM are 0.7811, 0.6747 and 0.6240, respectively. SIM under study provides a slightly worse fit than the others. However, we note that LR is, up to a link function, linear about $\mathbf{X}$, and, according to the results of GPLAM, which is a more general model than LR, the actual relationship cannot have such a structure. SIM can reveal nonlinear structure. On the other hand, although the minimum mean residual deviance can be not surprisingly attained by GPLAM, this model has, respectively, ≈ 34 and 41 more degrees of freedom than SIM and LR have.

We now employ the quasi-likelihood ratio test to the test problem (3.9). The QLR test statistic is 166.52 with one degree of freedom, resulting in a P-value of $< 10^{-5}$. Hence this result provides strong evidence that gender has a significant influence on high income.

The Adult data set used in this paper is a rich data set. Existing work mainly focused on the prediction accuracy based on machine learning methods. We make an attempt to explore the semiparametric regression pattern suitable for the data. Model specification and variable selection merit further study.

APPENDIX: OUTLINE OF PROOFS

We first introduce some regularity conditions.

Regularity Conditions:

(a) $\mu(\cdot), V(\cdot), g(\cdot), \mathbf{h}(\cdot) = E(\mathbf{X}|\boldsymbol{\beta}^\top \mathbf{X} = \cdot)$ have two bounded and continuous derivatives. $V(\cdot)$ is uniformly bounded and bounded away from 0.

(b) Let $q(z, y) = \mu'(z)V^{-1}(z)\{y - \mu(z)\}$. Assume that $\partial q(z, y)/\partial z < 0$ for $z \in \mathbb{R}$ and y in the range of the response variable.

(c) The largest eigenvalue of $\mathbf{\Omega}_{22}$ is bounded away from infinity.

(d) The density function $f_{\boldsymbol{\beta}^\top \mathbf{x}}(\boldsymbol{\beta}^\top \mathbf{x})$ of random variable $\boldsymbol{\beta}^\top \mathbf{X}$ is bounded away from 0 on $T_{\boldsymbol{\beta}}$ and satisfies the Lipschitz condition of order 1 on $T_{\boldsymbol{\beta}}$, where $T_{\boldsymbol{\beta}} = \{\boldsymbol{\beta}^\top \mathbf{x} : \mathbf{x} \in T\}$ and T is a compact support set of $\mathbf{X}$.

(e) Let $Q^*[\boldsymbol{\beta}] = \int Q[\mu\{g(\boldsymbol{\beta}^\top \mathbf{x})\}, y] f(y|\boldsymbol{\beta}^{0\top}\mathbf{x}) f(\boldsymbol{\beta}^{0\top}\mathbf{x})\, dy\, d(\boldsymbol{\beta}^{0\top}\mathbf{x})$ with $\boldsymbol{\beta}^0$ denoting the true parameter value and $Q[\mu, y] = \int_\mu^y \frac{s-y}{V\{\mu^{-1}(s)\}}\, ds$. Assume that $Q^*[\boldsymbol{\beta}]$ has a unique maximum at $\boldsymbol{\beta} = \boldsymbol{\beta}^0$, and

$$E\Big[\sup_{\boldsymbol{\beta}^{(1)}} \sup_{\boldsymbol{\beta}^\top \mathbf{X}} |\mu'\{g(\boldsymbol{\beta}^\top \mathbf{X})\} V^{-1}\{g(\boldsymbol{\beta}^\top \mathbf{X})\}[Y - \mu\{g(\boldsymbol{\beta}^\top \mathbf{X})\}]|^2\Big] < \infty$$

and $E\|\mathbf{X}\|^2 < \infty$.

(f) The kernel K is a bounded and symmetric density function with a bounded derivative, and satisfies

$$\int_{-\infty}^{\infty} t^2 K(t)\, dt \neq 0 \quad \text{and} \quad \int_{-\infty}^{\infty} |t|^j K(t)\, dt < \infty, \qquad j = 1, 2, \ldots.$$

Condition (a) is some mild smoothness conditions on the involved functions of the model. We impose condition (b) to guarantee that the solutions of (2.1), $\hat{g}(t)$ and $\hat{g}'(t)$, lie in a compact set. Condition (c) implies that the second moment of estimating equation (2.7), $\operatorname{tr}(\mathbf{J}^\top \mathbf{\Omega} \mathbf{J})$, is bounded. Then the CLT can be applied to $G(\boldsymbol{\beta})$. Condition (d) means that $\mathbf{X}$ may have discrete components and the density function of $\boldsymbol{\beta}^\top \mathbf{X}$ is positive, which ensures that the denominators involved in the nonparametric estimators, with high probability, are bounded away from 0. The uniqueness condition in condition (e) can be checked in the following case for example. Assume that Y is a Poisson variable with mean $\mu\{g(\boldsymbol{\beta}^\top \mathbf{x})\} = \exp\{g(\boldsymbol{\beta}^\top \mathbf{x})\}$. The maximizer β_0 of $Q^*[\boldsymbol{\beta}]$ is equal to the solution of the equation $E[E\{[\exp\{g(\boldsymbol{\beta}^{0\top}\mathbf{X})\} - \exp\{g(\boldsymbol{\beta}^\top \mathbf{X})\}] g'(\boldsymbol{\beta}^\top \mathbf{X}) \mathbf{J}^\top \mathbf{X} | \boldsymbol{\beta}^{0\top}\mathbf{X}\}] = 0$. $\boldsymbol{\beta}_0$ is unique when $g'(\cdot)$ is not a zero-valued constant function and the matrix $\mathbf{J}^\top E(\mathbf{X}\mathbf{X}^\top)\mathbf{J}$ is not singular. Under the second part of condition (e), it is permissible to interchange differentiation and integration when differentiating $E[Q[\mu\{g(\boldsymbol{\beta}^\top \mathbf{X})\}, Y]]$. Condition (f) is a commonly used smoothness condition, including the Gaussian kernel and the quadratic kernel. All of the conditions can be relaxed at the expense of longer proofs.

Throughout the Appendix, $Z_n = \mathcal{O}_P(a_n)$ denotes that $a_n^{-1} Z_n$ is bounded in probability and the derivation for the order of Z_n is based on the fact that $Z_n = \mathcal{O}_P\{\sqrt{E(Z_n^2)}\}$. Therefore, it allows to apply the Cauchy–Schwarz inequality to the quantity having stochastic order a_n.

A.1. Proof of Proposition 1. We outline the proof here, while the details are given in the supplementary materials [Cui, Härdle and Zhu (2010)].

(i) Conditions (a), (b), (d) and (f) are essentially equivalent conditions given by Carroll, Ruppert and Welsh (1998), and as a consequence the derivation of bias and variance for $\hat{g}(\boldsymbol{\beta}^\top \mathbf{x})$ and $\hat{g}'(\boldsymbol{\beta}^\top \mathbf{x})$ is similar to that of Carroll, Ruppert and Welsh (1998).

(ii) The first equation of (2.1) is

$$0 = \sum_{j=1}^{n} K_h(\boldsymbol{\beta}^\top \mathbf{X}_j - \boldsymbol{\beta}^\top \mathbf{x}) \mu'\{\hat{\alpha}_0 + \hat{\alpha}_1(\boldsymbol{\beta}^\top \mathbf{X}_j - \boldsymbol{\beta}^\top \mathbf{x})\}$$
$$\times V^{-1}\{\hat{\alpha}_0 + \hat{\alpha}_1(\boldsymbol{\beta}^\top \mathbf{X}_j - \boldsymbol{\beta}^\top \mathbf{x})\}[Y_j - \mu\{\hat{\alpha}_0 + \hat{\alpha}_1(\boldsymbol{\beta}^\top \mathbf{X}_j - \boldsymbol{\beta}^\top \mathbf{x})\}].$$

Taking derivatives with respect to $\boldsymbol{\beta}^{(1)}$ on both sides, direct observations lead to

$$\frac{\partial \hat{\alpha}_0}{\partial \boldsymbol{\beta}^{(1)}} = \{B(\boldsymbol{\beta}^\top \mathbf{x})\}^{-1}\{A_1(\boldsymbol{\beta}^\top \mathbf{x}) + A_2(\boldsymbol{\beta}^\top \mathbf{x}) + A_3(\boldsymbol{\beta}^\top \mathbf{x})\},$$

where

$$B(\boldsymbol{\beta}^\top \mathbf{x}) = -\sum_{j=1}^{n} K_h(\boldsymbol{\beta}^\top \mathbf{X}_j - \boldsymbol{\beta}^\top \mathbf{x}) q_z'\{\hat{\alpha}_0 + \hat{\alpha}_1(\boldsymbol{\beta}^\top \mathbf{X}_j - \boldsymbol{\beta}^\top \mathbf{x}), Y_j\},$$

$$A_1(\boldsymbol{\beta}^\top \mathbf{x}) = \sum_{j=1}^{n} K_h(\boldsymbol{\beta}^\top \mathbf{X}_j - \boldsymbol{\beta}^\top \mathbf{x}) \mathbf{J}^\top (\mathbf{X}_j - \mathbf{x}) q_z'\{\hat{\alpha}_0 + \hat{\alpha}_1(\boldsymbol{\beta}^\top \mathbf{X}_j - \boldsymbol{\beta}^\top \mathbf{x}), Y_j\}\hat{\alpha}_1,$$

$$A_2(\boldsymbol{\beta}^\top \mathbf{x}) = \sum_{j=1}^{n} K_h(\boldsymbol{\beta}^\top \mathbf{X}_j - \boldsymbol{\beta}^\top \mathbf{x}) q_z'\{\hat{\alpha}_0 + \hat{\alpha}_1(\boldsymbol{\beta}^\top \mathbf{X}_j - \boldsymbol{\beta}^\top \mathbf{x}), Y_j\}$$
$$\times (\boldsymbol{\beta}^\top \mathbf{X}_j - \boldsymbol{\beta}^\top \mathbf{x}) \frac{\partial \hat{\alpha}_1}{\partial \boldsymbol{\beta}^{(1)}},$$

$$A_3(\boldsymbol{\beta}^\top \mathbf{x}) = \sum_{j=1}^{n} h^{-1} K_h'(\boldsymbol{\beta}^\top \mathbf{X}_j - \boldsymbol{\beta}^\top \mathbf{x}) \mathbf{J}^\top (\mathbf{X}_j - \mathbf{x}) q\{\hat{\alpha}_0 + \hat{\alpha}_1(\boldsymbol{\beta}^\top \mathbf{X}_j - \boldsymbol{\beta}^\top \mathbf{x}), Y_j\}$$

with $K_h'(\cdot) = h^{-1}K'(\cdot/h)$. Note that $\partial \hat{\alpha}_0 / \partial \boldsymbol{\beta}^{(1)} = \partial \hat{g}(\boldsymbol{\beta}^\top \mathbf{x}) / \partial \boldsymbol{\beta}^{(1)}$; then we have

$$\frac{\partial \hat{g}(\boldsymbol{\beta}^\top \mathbf{x})}{\partial \boldsymbol{\beta}^{(1)}} = \{B(\boldsymbol{\beta}^\top \mathbf{x})\}^{-1} A_1(\boldsymbol{\beta}^\top \mathbf{x})$$
$$+ \{B(\boldsymbol{\beta}^\top \mathbf{x})\}^{-1} A_2(\boldsymbol{\beta}^\top \mathbf{x}) + \{B(\boldsymbol{\beta}^\top \mathbf{x})\}^{-1} A_3(\boldsymbol{\beta}^\top \mathbf{x}). \tag{A.1}$$

We will prove that

$$E\|\{B(\boldsymbol{\beta}^\top \mathbf{x})\}^{-1} A_1(\boldsymbol{\beta}^\top \mathbf{x}) - g'(\boldsymbol{\beta}^\top \mathbf{x}) \mathbf{J}^\top \{\mathbf{x} - \mathbf{h}(\boldsymbol{\beta}^\top \mathbf{x})\}\|^2$$
$$= \mathcal{O}_P(h^4 + n^{-1}h^{-3}), \tag{A.2}$$

the second term in (A.1) is of order $\mathcal{O}_P(h^4 + n^{-1}h)$, and the third term is of order $\mathcal{O}_P(h^4 + n^{-1}h^{-3})$. The combination of (A.1) and these three results can directly

lead to result (ii) of Proposition 1. The detailed proof is summarized in three steps and is given in the supplementary materials [Cui, Härdle and Zhu (2010)].

(iii) By mimicking the proof of (ii), we can show that (iii) holds. See supplementary materials for details.

A.2. Proofs of (2.6) and (2.7). It is proved in the supplementary materials [Cui, Härdle and Zhu (2010)].

A.3. Proof of Theorem 2.1. (i) Note that the estimating equation defined in (2.6) is just the gradient of the following quasi-likelihood:

$$\hat{Q}(\boldsymbol{\beta}) = \sum_{i=1}^{n} Q[\mu\{\hat{g}(\boldsymbol{\beta}^\top \mathbf{X}_i)\}, Y_i]$$

with $Q[\mu, y] = \int^{\mu} \frac{y-s}{V\{\mu^{-1}(s)\}}\, ds$ and $\mu^{-1}(\cdot)$ is the inverse function of $\mu(\cdot)$. Then for $\boldsymbol{\beta}^{(1)}$ satisfying $(\sqrt{1 - \|\boldsymbol{\beta}^{(1)}\|^2}, \boldsymbol{\beta}^{(1)\top})^\top \in \Theta$, we have

$$\hat{\boldsymbol{\beta}}^{(1)} = \arg\max_{\boldsymbol{\beta}^{(1)}} \hat{Q}(\boldsymbol{\beta}).$$

The proof is based on Theorem 5.1 in Ichimura (1993). In that theorem the consistency of $\boldsymbol{\beta}^{(1)}$ is proved by means of proving that

$$\sup_{\boldsymbol{\beta}^{(1)}} \left| \frac{1}{n} \sum_{i=1}^{n} Q[\mu\{\hat{g}(\boldsymbol{\beta}^\top \mathbf{X}_i)\}, Y_i] - \frac{1}{n} \sum_{i=1}^{n} Q[\mu\{g(\boldsymbol{\beta}^\top \mathbf{X}_i)\}, Y_i] \right| = o_P(1), \tag{A.3}$$

$$\sup_{\boldsymbol{\beta}^{(1)}} \left| \frac{1}{n} \sum_{i=1}^{n} Q[\mu\{g(\boldsymbol{\beta}^\top \mathbf{X}_i)\}, Y_i] - \frac{1}{n} \sum_{i=1}^{n} E[Q[\mu\{g(\boldsymbol{\beta}^\top \mathbf{X}_i)\}, Y_i]] \right| = o_P(1) \tag{A.4}$$

and

$$\left| \frac{1}{n} \sum_{i=1}^{n} Q[\mu\{\hat{g}(\boldsymbol{\beta}_0^\top \mathbf{X}_i)\}, Y_i] - \frac{1}{n} \sum_{i=1}^{n} E[Q[\mu\{g(\boldsymbol{\beta}_0^\top \mathbf{X}_i)\}, Y_i]] \right| = o_P(1). \tag{A.5}$$

Regarding the validity of (A.5), this directly follows from (A.3) and (A.4). The type of uniform convergence result such as (A.4) has been well established in the literature; see, for example, Andrews (1987). We now verify the validity of (A.3), which reduces to showing the uniform convergence of the estimator $\hat{g}(t)$ under condition (e) [see Ichimura (1993)]. This can be obtained in a similar way as in Kong, Linton and Xia (2010), taking into account that the regularity conditions imposed in Theorem 2.1 are stronger than the corresponding ones in that paper.

(ii) Recall the notation $\mathbf{J}$, $\boldsymbol{\Omega}$ and $\mathbf{G}(\boldsymbol{\beta})$ introduced in Section 2. By (2.7), we have shown that

$$\sqrt{n}(\hat{\boldsymbol{\beta}}^{(1)} - \boldsymbol{\beta}^{(1)0}) = \frac{1}{\sqrt{n}} \{\mathbf{J}^\top \boldsymbol{\Omega} \mathbf{J}\}^{+} \mathbf{G}(\boldsymbol{\beta}) + o_P(1). \tag{A.6}$$

Theorem 2.1 follows directly from the above asymptotic expansion and the fact that $E\{\mathbf{G}(\boldsymbol{\beta})\mathbf{G}^\top(\boldsymbol{\beta})\} = n\mathbf{J}^\top \boldsymbol{\Omega} \mathbf{J}$. □

A.4. Proof of Corollary 1. The asymptotic covariance of $\hat{\boldsymbol{\beta}}$ can be obtained by adjusting the asymptotic covariance of $\hat{\boldsymbol{\beta}}^{(1)}$ via the multivariate delta method, and is of form $\mathbf{J}(\mathbf{J}^\top\boldsymbol{\Omega}\mathbf{J})^+\mathbf{J}^\top$. Next we will compare this asymptotic covariance with that (denoted by $\boldsymbol{\Omega}^+$) given in Carroll et al. (1997). Write $\boldsymbol{\Omega}$ as

$$\boldsymbol{\Omega} = \begin{pmatrix} \boldsymbol{\Omega}_{11} & \boldsymbol{\Omega}_{12} \\ \boldsymbol{\Omega}_{21} & \boldsymbol{\Omega}_{22} \end{pmatrix},$$

where $\boldsymbol{\Omega}_{22}$ is a $(d-1)\times(d-1)$ matrix. We will next investigate two cases, respectively: $\det(\boldsymbol{\Omega}_{22}) \neq 0$ and $\det(\boldsymbol{\Omega}_{22}) = 0$. Let $\boldsymbol{\alpha} = -\boldsymbol{\beta}^{(1)}/\sqrt{1-\|\boldsymbol{\beta}^{(1)}\|^2} = -\boldsymbol{\beta}^{(1)}/\beta_1$.

Consider the case that $\det(\boldsymbol{\Omega}_{22}) \neq 0$. Because $\operatorname{rank}(\boldsymbol{\Omega}) = d-1$, $\det(\boldsymbol{\Omega}_{11}\boldsymbol{\Omega}_{22} - \boldsymbol{\Omega}_{21}\boldsymbol{\Omega}_{12}) = 0$. Note that $\boldsymbol{\Omega}_{22}$ is nondegenerate; it can be easily shown that $\boldsymbol{\Omega}_{11} = \boldsymbol{\Omega}_{12}\boldsymbol{\Omega}_{22}^{-1}\boldsymbol{\Omega}_{21}$. Combining this with the following fact:

$$\begin{aligned}\mathbf{J}^\top\boldsymbol{\Omega}\mathbf{J} &= (\boldsymbol{\alpha} \quad \mathbf{I}_{d-1}) \begin{pmatrix} \boldsymbol{\Omega}_{11} & \boldsymbol{\Omega}_{12} \\ \boldsymbol{\Omega}_{21} & \boldsymbol{\Omega}_{22} \end{pmatrix} \begin{pmatrix} \boldsymbol{\alpha}^\tau \\ \mathbf{I}_{d-1} \end{pmatrix} \\ &= \boldsymbol{\Omega}_{22} + (\boldsymbol{\Omega}_{21}/\sqrt{\boldsymbol{\Omega}_{11}} + \sqrt{\boldsymbol{\Omega}_{11}}\boldsymbol{\alpha})(\boldsymbol{\Omega}_{12}/\sqrt{\boldsymbol{\Omega}_{11}} + \sqrt{\boldsymbol{\Omega}_{11}}\boldsymbol{\alpha}^\top) - \boldsymbol{\Omega}_{21}\boldsymbol{\Omega}_{12}/\boldsymbol{\Omega}_{11},\end{aligned}$$

we can get that $\mathbf{J}^\top\boldsymbol{\Omega}\mathbf{J}$ is nondegenerate. In this situation, its inverse $(\mathbf{J}^\top\boldsymbol{\Omega}\mathbf{J})^+$ is just the ordinary inverse $(\mathbf{J}^\top\boldsymbol{\Omega}\mathbf{J})^{-1}$. Then $\mathbf{J}(\mathbf{J}^\top\boldsymbol{\Omega}\mathbf{J})^+\mathbf{J}^\top = \{\mathbf{J}(\mathbf{J}^\top\boldsymbol{\Omega}\mathbf{J})^{-1/2}\}\{(\mathbf{J}^\top \times \boldsymbol{\Omega}\mathbf{J})^{-1/2}\mathbf{J}^\top\}$, a full-rank decomposition. Then

$$\begin{aligned}\{\mathbf{J}(\mathbf{J}^\top\boldsymbol{\Omega}\mathbf{J})^+\mathbf{J}^\top\}^+ &= \{\mathbf{J}(\mathbf{J}^\top\boldsymbol{\Omega}\mathbf{J})^{-1/2}\} \\ &\quad \times \{(\mathbf{J}^\top\boldsymbol{\Omega}\mathbf{J})^{-1/2}\mathbf{J}^\top\mathbf{J}(\mathbf{J}^\top\boldsymbol{\Omega}\mathbf{J})^{-1}\mathbf{J}^\top\mathbf{J}(\mathbf{J}^\top\boldsymbol{\Omega}\mathbf{J})^{-1/2}\}^{-1} \\ &\quad \times \{(\mathbf{J}^\top\boldsymbol{\Omega}\mathbf{J})^{-1/2}\mathbf{J}^\top\} \\ &= \mathbf{J}(\mathbf{J}^\top\mathbf{J})^{-1}\mathbf{J}^\top\boldsymbol{\Omega}\mathbf{J}(\mathbf{J}^\top\mathbf{J})^{-1}\mathbf{J}^\top \\ &= \boldsymbol{\Omega}.\end{aligned}$$

This means that $\mathbf{J}(\mathbf{J}^\top\boldsymbol{\Omega}\mathbf{J})^+\mathbf{J}^\top = \boldsymbol{\Omega}^+$.

When $\det(\boldsymbol{\Omega}_{22}) = 0$, we can obtain that

$$\boldsymbol{\Omega}^+ = \begin{pmatrix} 1/\boldsymbol{\Omega}_{11} + \boldsymbol{\Omega}_{12}\boldsymbol{\Omega}_{22.1}^+\boldsymbol{\Omega}_{21}/\boldsymbol{\Omega}_{11}^2 & -\boldsymbol{\Omega}_{12}\boldsymbol{\Omega}_{22.1}^+/\boldsymbol{\Omega}_{11} \\ -\boldsymbol{\Omega}_{22.1}^+\boldsymbol{\Omega}_{21}/\boldsymbol{\Omega}_{11} & \boldsymbol{\Omega}_{22.1}^+ \end{pmatrix}$$

with $\boldsymbol{\Omega}_{22.1} = \boldsymbol{\Omega}_{22} - \boldsymbol{\Omega}_{21}\boldsymbol{\Omega}_{12}/\boldsymbol{\Omega}_{11}$. Write $\mathbf{J}(\mathbf{J}^\top\boldsymbol{\Omega}\mathbf{J})^+\mathbf{J}^\top$ as

$$\begin{pmatrix} \boldsymbol{\alpha}^\top(\mathbf{J}^\top\boldsymbol{\Omega}\mathbf{J})^+\boldsymbol{\alpha} & \boldsymbol{\alpha}^\top(\mathbf{J}^\top\boldsymbol{\Omega}\mathbf{J})^+ \\ (\mathbf{J}^\top\boldsymbol{\Omega}\mathbf{J})^+\boldsymbol{\alpha} & (\mathbf{J}^\top\boldsymbol{\Omega}\mathbf{J})^+ \end{pmatrix}.$$

Note that $\mathbf{J}^\top\boldsymbol{\Omega}\mathbf{J} = \boldsymbol{\Omega}_{22.1} + (\boldsymbol{\Omega}_{21}/\sqrt{\boldsymbol{\Omega}_{11}} + \sqrt{\boldsymbol{\Omega}_{11}}\boldsymbol{\alpha})(\boldsymbol{\Omega}_{12}/\sqrt{\boldsymbol{\Omega}_{11}} + \sqrt{\boldsymbol{\Omega}_{11}}\boldsymbol{\alpha}^\top)$, so $\mathbf{J}^\top\boldsymbol{\Omega}\mathbf{J} \geq \boldsymbol{\Omega}_{22.1}$. Combining this with $\operatorname{rank}(\boldsymbol{\Omega}_{22}) = d-2$, we have that $(\mathbf{J}^\top\boldsymbol{\Omega}\mathbf{J})^+ \leq \boldsymbol{\Omega}_{22.1}^+$. It is easy to check that $\boldsymbol{\alpha}^\top\boldsymbol{\Omega}_{22.1} = 0$, so $\boldsymbol{\alpha} \perp \operatorname{span}(\boldsymbol{\Omega}_{22.1})$ and $\boldsymbol{\alpha}^\top\boldsymbol{\Omega}_{22.1}^+\boldsymbol{\alpha} = 0$, and then $\boldsymbol{\alpha}^\top(\mathbf{J}^\top\boldsymbol{\Omega}\mathbf{J})^+ = 0$. In this situation, $\mathbf{J}(\mathbf{J}^\top\boldsymbol{\Omega}\mathbf{J})^+\mathbf{J}^\top \leq \boldsymbol{\Omega}^+$ and the stick less-than sign holds since $\mathbf{J}^\top\boldsymbol{\Omega}\mathbf{J} \neq \boldsymbol{\Omega}_{22.1}$ and $1/\boldsymbol{\Omega}_{11} > 0$. □

A.5. Proof of Theorem 2.2. Under H_0, we can rewrite the index vector as $\boldsymbol{\beta} = [\mathbf{e}\ \mathbf{B}]^\top(\sqrt{1-\|\boldsymbol{\omega}^{(1)}\|^2}, \boldsymbol{\omega}^{(1)\tau})^\top$ where $\mathbf{e} = (1, 0, \ldots, 0)^\top$ is an r-dimensional vector,

$$\mathbf{B} = \begin{pmatrix} \mathbf{0}^\top & 0 \\ \mathbf{I}_{r-1} & \mathbf{0} \end{pmatrix}$$

is an $r \times (d-1)$ matrix and $\boldsymbol{\omega}^{(1)} = (\beta_2, \ldots, \beta_r)^\top$ is an $(r-1)\times 1$ vector. Let $\boldsymbol{\omega} = (\sqrt{1-\|\boldsymbol{\omega}^{(1)}\|^2}, \boldsymbol{\omega}^{(1)\top})^\top$. So under H_0 the estimator is also the local maximizer $\hat{\boldsymbol{\omega}}$ of the problem

$$\hat{Q}([\mathbf{e}\quad \mathbf{B}]^\top\hat{\boldsymbol{\omega}}) = \sup_{\|\boldsymbol{\omega}^{(1)}\|<1} \hat{Q}([\mathbf{e}\quad \mathbf{B}]^\top\boldsymbol{\omega}).$$

Expanding $\hat{Q}(\mathbf{B}^\top\hat{\boldsymbol{\omega}})$ at $\hat{\boldsymbol{\beta}}^{(1)}$ by a Taylor's expansion and noting that $\partial\hat{Q}(\boldsymbol{\beta})/\partial\boldsymbol{\beta}^{(1)}|_{\boldsymbol{\beta}^{(1)}=\hat{\boldsymbol{\beta}}^{(1)}} = 0$, then $\hat{Q}(\hat{\boldsymbol{\beta}}) - \hat{Q}(\mathbf{B}^\top\hat{\boldsymbol{\omega}}) = T_1 + T_2 + \mathcal{O}_P(1)$, where

$$T_1 = -\frac{1}{2}(\hat{\boldsymbol{\beta}}^{(1)} - \mathbf{B}^\top\hat{\boldsymbol{\omega}})^\top \frac{\partial^2\hat{Q}(\boldsymbol{\beta})}{\partial\boldsymbol{\beta}^{(1)}\partial\boldsymbol{\beta}^{(1)\tau}}\bigg|_{\boldsymbol{\beta}^{(1)}=\hat{\boldsymbol{\beta}}^{(1)}} (\hat{\boldsymbol{\beta}}^{(1)} - \mathbf{B}^\top\hat{\boldsymbol{\omega}}),$$

$$T_2 = \frac{1}{6}(\hat{\boldsymbol{\beta}}^{(1)} - \mathbf{B}^\top\hat{\boldsymbol{\omega}})^\top \times \frac{\partial\{(\hat{\boldsymbol{\beta}}^{(1)} - \mathbf{B}^\top\hat{\boldsymbol{\omega}})^\top\partial^2\hat{Q}(\boldsymbol{\beta})/(\partial\boldsymbol{\beta}^{(1)}\,\partial\boldsymbol{\beta}^{(1)\tau})|_{\boldsymbol{\beta}^{(1)}=\hat{\boldsymbol{\beta}}^{(1)}}(\hat{\boldsymbol{\beta}}^{(1)} - \mathbf{B}^\top\hat{\boldsymbol{\omega}})\}}{\partial\boldsymbol{\beta}^{(1)}}.$$

Assuming the conditions in Theorem 2.1 and under the null hypothesis H_0, it is easy to show that

$$\sqrt{n}(\mathbf{B}^\top\hat{\boldsymbol{\omega}} - \mathbf{B}^\top\boldsymbol{\omega}) = \frac{1}{\sqrt{n}}\mathbf{B}^\top\mathbf{B}(\mathbf{J}^\top\boldsymbol{\Omega}\mathbf{J})^+\mathbf{G}(\boldsymbol{\beta}) + \mathcal{O}_P(1).$$

Combining this with (A.6), under the null hypothesis H_0,

$$\begin{aligned}\sqrt{n}(\hat{\boldsymbol{\beta}}^{(1)} - \mathbf{B}^\top\hat{\boldsymbol{\omega}}^{(1)}) &= \frac{1}{\sqrt{n}}(\mathbf{J}^\top\boldsymbol{\Omega}\mathbf{J})^{1/2+}\{\mathbf{I}_{d-1} - (\mathbf{J}^\top\boldsymbol{\Omega}\mathbf{J})^{1/2}\mathbf{B}^\top\mathbf{B}(\mathbf{J}^\top\boldsymbol{\Omega}\mathbf{J})^{1/2+}\} \\ &\quad\times (\mathbf{J}^\top\boldsymbol{\Omega}\mathbf{J})^{1/2+}\mathbf{G}(\boldsymbol{\beta}) + o_P(1).\end{aligned} \tag{A.7}$$

Since $\frac{1}{\sqrt{n}}\mathbf{G}(\boldsymbol{\beta}) = \mathcal{O}_P(1)$, $\frac{\partial^2\hat{Q}(\boldsymbol{\beta})}{\partial\boldsymbol{\beta}^{(1)}\partial\boldsymbol{\beta}^{(1)\tau}}|_{\boldsymbol{\beta}^{(1)}} = -n\mathbf{J}^\top\boldsymbol{\Omega}\mathbf{J} + \mathcal{O}_P(n)$ and matrix $\mathbf{J}^\top\boldsymbol{\Omega}\mathbf{J}$ has eigenvalues uniformly bounded away from 0 and infinity, we have $\|\hat{\boldsymbol{\beta}}^{(1)} - \mathbf{B}^\top\hat{\boldsymbol{\omega}}^{(1)}\| = \mathcal{O}_P(n^{-1/2})$ and then $|T_2| = \mathcal{O}_P(1)$. Combining this and (A.7), we have

$$\begin{aligned}\hat{Q}(\hat{\boldsymbol{\beta}}) - \hat{Q}(\mathbf{B}^\top\hat{\boldsymbol{\omega}}) &= \frac{n}{2}(\hat{\boldsymbol{\beta}}^{(1)} - \mathbf{B}^\top\hat{\boldsymbol{\omega}}^{(1)})^\top\mathbf{J}^\top\boldsymbol{\Omega}\mathbf{J}(\hat{\boldsymbol{\beta}}^{(1)} - \mathbf{B}^\top\hat{\boldsymbol{\omega}}^{(1)}) \\ &= \frac{n}{2}\mathbf{G}^\top(\boldsymbol{\beta})(\mathbf{J}^\top\boldsymbol{\Omega}\mathbf{J})^{1/2+}\mathbf{P}(\mathbf{J}^\top\boldsymbol{\Omega}\mathbf{J})^{1/2+}\mathbf{G}(\boldsymbol{\beta})\end{aligned}$$

with $\mathbf{P} = \mathbf{I}_{d-1} - (\mathbf{J}^\top \boldsymbol{\Omega} \mathbf{J})^{1/2} \mathbf{B}^\top \mathbf{B} (\mathbf{J}^\top \boldsymbol{\Omega} \mathbf{J})^{1/2+}$. Here $\mathbf{P}$ is idempotent having rank $d - r$, so it can be written as $\mathbf{P} = \mathbf{S}^\top \mathbf{S}$ where $\mathbf{S}$ ia a $(d - r) \times (d - 1)$ matrix satisfying $\mathbf{S}\mathbf{S}^\top = \mathbf{I}_{d-r}$. Consequently,

$$\begin{aligned} 2\{\hat{Q}(\hat{\boldsymbol{\beta}}) - \hat{Q}(\mathbf{B}^\top \hat{\boldsymbol{\omega}})\} &= \big(\sqrt{n}\mathbf{S}(\mathbf{J}^\top \boldsymbol{\Omega} \mathbf{J})^{1/2+} \mathbf{G}(\boldsymbol{\beta})\big)^\top \big(\sqrt{n}\mathbf{S}(\mathbf{J}^\top \boldsymbol{\Omega} \mathbf{J})^{1/2+} \mathbf{G}(\boldsymbol{\beta})\big) \\ &\xrightarrow{\mathcal{L}} \chi^2(d - r). \end{aligned}$$

Acknowledgments. The authors thank the Associate Editor and two referees for their constructive comments and suggestions which led to a great improvement over an early manuscript.

SUPPLEMENTARY MATERIAL

Supplementary materials (DOI: 10.1214/10-AOS871SUPP; .pdf). Complete proofs of Proposition 1, (2.6) and (2.7).

REFERENCES

ANDREWS, D. W. K. (1987). Consistency in nonlinear econometric models: A genetic uniform law of large numbers. *Econometrica* **55** 1465–1471. MR0923471

CARROLL, R. J., RUPPERT, D. and WELSH, A. H. (1998). Local estimating equations. *J. Amer. Statist. Assoc.* **93** 214–227. MR1614624

CARROLL, R. J., FAN, J., GIJBELS, I. and WAND, M. P. (1997). Generalized partially linear single-index models. *J. Amer. Statist. Assoc.* **92** 447–489. MR1467842

CHANG, Z. Q., XUE, L. G. and ZHU, L. X. (2010). On an asymptotically more efficient estimation of the single-index model. *J. Multivariate Anal.* **101** 1898–1901. MR2651964

CUI, X., HÄRDLE, W. and ZHU, L. (2010). Supplementary materials for "The EFM approach for single-index models." DOI:10.1214/10-AOS871SUPP.

FAN, J. and GIJBELS, I. (1996). *Local Polynomial Modeling and Its Applications*. Chapman & Hall, London. MR1383587

FAN, J., HECKMAN, N. E. and WAND, M. P. (1995). Local polynomial kernel regression for generalized linear models and quasi-likelihood functions. *J. Amer. Statist. Assoc.* **90** 141–150. MR1325121

FAN, J. and JIANG, J. (2007). Nonparametric inference with generalized likelihood ratio test. *Test* **16** 409–478. MR2365172

HÄRDLE, W., HALL, P. and ICHIMURA, H. (1993). Optimal smoothing in single-index models. *Ann. Statist.* **21** 157–178. MR1212171

HÄRDLE, W. and MAMMEN, E. (1993). Testing parametric versus nonparametric regression. *Ann. Statist.* **21** 1926–1947. MR1245774

HÄRDLE, W., MAMMEN, E. and MÜLLER, M. (1998). Testing parametric versus semiparametric modelling in generalized linear models. *J. Amer. Statist. Assoc.* **93** 1461–1474. MR1666641

HÄRDLE, W., MAMMEN, E. and PROENCA, I. (2001). A bootstrap test for single index models. *Statistics* **35** 427–452. MR1880174

HÄRDLE, W. and STOKER, T. M. (1989). Investigating smooth multiple regression by method of average derivatives. *J. Amer. Statist. Assoc.* **84** 986–995. MR1134488

HEYDE, C. C. (1997). *Quasi-likelihood and Its Application*: *A General Approach to Optimal Parameter Estimation*. Springer, New York. MR1461808

HOROWITZ, J. L. and HÄRDLE, W. (1996). Direct semiparametric estimation of a single-index model with discrete covariates. *J. Amer. Statist. Assoc.* **91** 1632–1640. MR1439104

HRISTACHE, M., JUDITSKI, A. and SPOKOINY, V. (2001). Direct estimation of the index coefficients in a single-index model. *Ann. Statist.* **29** 595–623. MR1865333

HRISTACHE, M., JUDITSKY, A., POLZEHL, J. and SPOKOINY, V. (2001). Structure adaptive approach for dimension reduction. *Ann. Statist.* **29** 1537–1566. MR1891738

HUH, J. and PARK, B. U. (2002). Likelihood-based local polynomial fitting for single-index models. *J. Multivariate Anal.* **80** 302–321. MR1889778

ICHIMURA, H. (1993). Semiparametric least squares (SLS) and weighted SLS estimation of single-index models. *J. Econometrics* **58** 71–120. MR1230981

KANE, M., HOLT, J. and ALLEN, B. (2004). Results concerning the generalized partially linear single-index model. *J. Stat. Comput. Simul.* **72** 897–912. MR2100843

KOHAVI, R. (1996). Scaling up the accuracy of naive-Bayes classifiers: A decision-tree hybrid. In *Proceedings of the Second International Conference on Knowledge Discovery and Data Mining* 202–207. AAAI Press, Menlo Park, CA.

KONG, E., LINTON, O. and XIA, Y. (2010). Uniform Bahadur representation for local polynomial estimates of M-regression and its application to the additive model. *Econometric Theory* **26** 1529–1564. MR2684794

LIN, W. and KULASEKERA, K. B. (2007). Identifiability of single-index models and additive-index models. *Biometrika* **94** 496–501. MR2380574

MADALOZZO, R. C. (2008). An analysis of income differentials by marital status. *Estudos Econômicos* **38** 267–292.

MCCULLAGH, P. and NELDER, J. A. (1989). *Generalized Linear Models*, 2nd ed. Champman & Hall, London.

MURRAY, C. (1997). IQ and economic success. *The Public Interest* **128** 21–35.

POLZEHL, J. and SPERLICH, S. (2009). A note on structural adaptive dimension reduction. *J. Stat. Comput. Simul.* **79** 805–818. MR2751594

POWELL, J. L., STOCK, J. H. and STOKER, T. M. (1989). Semiparametric estimation of index coefficients. *Econometrica* **57** 1403–1430. MR1035117

WANG, H. and XIA, Y. (2008). Sliced regression for dimension reduction. *J. Amer. Statist. Assoc.* **103** 811–821. MR2524332

WANG, J. L., XUE, L. G., ZHU, L. X. and CHONG, Y. S. (2010). Estimation for a partial-linear single-index model. *Ann. Statist.* **38** 246–274. MR2589322

XIA, Y. (2006). Asymptotic distributions for two estimators of the single-index model. *Econometric Theory* **22** 1112–1137. MR2328530

XIA, Y., TONG, H., LI, W. K. and ZHU, L. (2002). An adaptive estimation of dimension reduction space (with discussions). *J. R. Stat. Soc. Ser. B Stat. Methodol.* **64** 363–410. MR1924297

YU, Y. and RUPPERT, D. (2002). Penalized spline estimation for partially linear single index models. *J. Amer. Statist. Assoc.* **97** 1042–1054. MR1951258

ZHOU, J. and HE, X. (2008). Dimension reduction based on constrained canonical correlation and variable filtering. *Ann. Statist.* **36** 1649–1668. MR2435451

ZHU, L. X. and XUE, L. G. (2006). Empirical likelihood confidence regions in a partially linear single-index model. *J. R. Stat. Soc. Ser. B Stat. Methodol.* **68** 549–570. MR2278341

ZHU, L. P. and ZHU, L. X. (2009a). Nonconcave penalized inverse regression in single-index models with high dimensional predictors. *J. Multivariate Anal.* **100** 862–875. MR2498719

ZHU, L. P. and ZHU, L. X. (2009b). On distribution weighted partial least squares with diverging number of highly correlated predictors. *J. R. Stat. Soc. Ser. B Stat. Methodol.* **71** 525–548. MR2649607

X. CUI
SCHOOL OF MATHEMATICS
AND COMPUTATIONAL SCIENCE
SUN YAT-SEN UNIVERSITY
GUANGZHOU
GUANGDONG PROVINCE, 510275
P.R. CHINA
E-MAIL: cuixia@mail.sysu.edu.cn

W. K. HÄRDLE
CASE-CENTER FOR APPLIED STATISTICS
AND ECONOMICS
HUMBOLDT-UNIVERSITÄT ZU BERLIN
WIRTSCHAFTSWISSENSCHAFTLICHE FAKULTÄT
SPANDAUER STR. 1
10178 BERLIN
GERMANY
E-MAIL: haerdle@wiwi.hu-berlin.de

L. ZHU
FSC1207, FONG SHU CHUEN BUILDING
DEPARTMENT OF MATHEMATICS
HONG KONG BAPTIST UNIVERSITY
KOWLOON TONG
HONG KONG
P.R. CHINA
E-MAIL: lzhu@hkbu.edu.hk

Econometric Theory, **28**, 2012, 861–887.
doi:10.1017/S0266466611000685

A CONSISTENT NONPARAMETRIC TEST FOR CAUSALITY IN QUANTILE

KIHO JEONG
Kyungpook National University

WOLFGANG K. HÄRDLE
Humboldt-Universität zu Berlin

SONG SONG
Humboldt-Universität zu Berlin
and
University of California, Berkeley

This paper proposes a nonparametric test of Granger causality in quantile. Zheng (1998, *Econometric Theory* 14, 123–138) studied the idea to reduce the problem of testing a quantile restriction to a problem of testing a particular type of mean restriction in independent data. We extend Zheng's approach to the case of dependent data, particularly to the test of Granger causality in quantile. Combining the results of Zheng (1998) and Fan and Li (1999, *Journal of Nonparametric Statistics* 10, 245–271), we establish the asymptotic normal distribution of the test statistic under a β-mixing process. The test is consistent against all fixed alternatives and detects local alternatives approaching the null at proper rates. Simulations are carried out to illustrate the behavior of the test under the null and also the power of the test under plausible alternatives. An economic application considers the causal relations between the crude oil price, the USD/GBP exchange rate, and the gold price in the gold market.

1. INTRODUCTION

Whether movements in one economic variable cause reactions in another variable is an important issue in economic policy and also for financial investment decisions. A framework for investigating causality between economic indicators has been developed by Granger (1969). Testing for Granger causality between

The research was conducted while Jeong was visiting C.A.S.E.—Center for Applied Statistics and Economics—Humboldt-Universität zu Berlin in the summers of 2005 and 2007. Jeong is grateful for their hospitality during the visits. Jeong's work was supported by a Korean Research Foundation grant funded by the Korean government (MOEHRD) (KRF-2006-B00002), and Härdle and Song's work was supported by the Deutsche Forschungsgemeinschaft through the SFB 649 "Economic Risk." We thank the editor, two anonymous referees, and Holger Dette for concrete suggestions on improving the manuscript and restructuring the paper. Their valuable comments and suggestions are gratefully acknowledged. Address correspondence to Kiho Jeong, Kyungpook National University, Korea; e-mail: khjeong@knu.ac.kr.

economic time series has been since studied intensively in empirical macroeconomics and empirical finance. The majority of research results were obtained in the context of Granger causality in the conditional mean. The conditional mean, though, is a questionable element of analysis if the distributions of the variables involved are nonelliptic or fat tailed as is to be expected with, for example, financial returns. The focus of a causality analysis on the mean might result in unclear news. The conditional mean is only one element of an overall summary for the conditional distribution. A tail area causal relation may be quite different from a causality based on the center of the distribution. Lee and Yang (2007) explore money-income Granger causality in the conditional quantile and find that Granger causality is significant in tail quantiles, whereas it is not significant in the center of the distribution.

An illustrating motivation for the research presented here is from labor market analysis where one tries to find out how income depends on the age of the employee for different education levels, genders, and nationalities, and so on (discrimination effects); see, for example, Buchinsky (1995). In particular, the effect of education on income is summarized by the basic claim of Day and Newburger (2007): At most ages, more education equates with higher earnings, and the payoff is most notable at the highest educational level, which is actually from the point of view of mean regression. However, whether this difference is significant or not is still questionable, especially for different ends of the (conditional) income distribution. Härdle, Ritov, and Song (2009) show that for the 0.20 quantile confidence bands for income given "university," "apprenticeship," and "low education" status do not differ significantly from one another although they become progressively lower, which indicates that high education does not equate to higher earnings significantly for the lower tails of income, whereas increasing age seems to be the main driving force. For the conditional median, the bands for "university" and "low education" differ significantly. For the 0.80 quantiles, all conditional quantiles differ, which indicates that higher education is associated with higher earnings. However, these findings do not necessarily indicate causalities. To answer the question "Does education Granger cause income in various conditional quantiles?" the concept of Granger causality in means cannot be used to estimate or test for these effects. Hence the need for the concept of Granger causality in quantiles and the need to develop tests for these effects emerge.

Another motivation comes from controlling and monitoring downside market risk and investigating large comovements between financial markets. These are important for risk management and portfolio/investment diversification (Hong, Liu, and Wang, 2009). Various other risk management tasks are described in Bollerslev (2001) and Campbell and Cochrane (1999) indicating the importance of Granger causality in quantile. Yet another motivation comes from the well-known robustness properties of the conditional quantile: like the parallel boxplot—calculated across an explanatory variable—the set of conditional quantiles characterizes the entire distribution in more detail.

Based on the kernel method, we propose a nonparametric test for Granger causality in quantile. Testing conditional quantile restrictions by nonparametric estimation techniques in dependent data situations has not been considered in the literature before. This paper intends to fill this literature gap. In an unpublished working paper that has been independently carried out from ours, Lee and Yang (2007) also propose a test for Granger causality in the conditional quantile. Their test, however, relies on linear quantile regression and thus is subject to possible functional misspecification of quantile regression. Recently, Hong et al. (2009) investigated Granger causality in value at risk (VaR) with a corresponding (kernel-based) test. Their method, however, offers two possible improvements. The first is that it needs a parametric specification of VaR, again subject to misspecification errors. The second is that their test does not directly check causality but rather a necessary condition for causality.

The problem of testing conditional mean restrictions using nonparametric estimation techniques has been actively studied for dependent data. Among the related work, the testing procedures of Fan and Li (1999) and Li (1999) use the general hypothesis of the form $\mathrm{E}(\varepsilon|z) = 0$, where ε and z are the regression error term and the vector of regressors, respectively. They consider the distance measure of $J = \mathrm{E}[\varepsilon \mathrm{E}(\varepsilon|z) f(z)]$ to construct kernel-based procedures. For the advantages of using this distance measure in kernel-based testing procedures, see Li and Wang (1998) and Hsiao and Li (2001). A feasible test statistic based on J has a second-order degenerate U-statistic as the leading term under the null hypothesis. Generalizing the result of Hall (1984) for independent data, Fan and Li (1999) establish the asymptotic normal distribution for a general second-order degenerate U-statistic with dependent data.

All the results stated previously on testing mean restrictions are however irrelevant when testing quantile restrictions. Zheng (1998) proposed an idea to transform quantile restrictions to mean restrictions in independent data. Following his idea, one can use the existing technical results on testing mean restrictions in testing quantile restrictions. In this paper, by combining Zheng's idea and the results of Fan and Li (1999) and Li (1999), we derive a test statistic for Granger causality in quantile and establish the asymptotic normal distribution of the proposed test statistic under a β-mixing process. Our testing procedure can be extended to several hypothesis testing problems with conditional quantile in dependent data; for example, testing a parametric regression functional form, testing the insignificance of a subset of regressors, and testing semiparametric versus nonparametric regression models.

The paper is organized as follows. Section 2 presents the test statistic. Section 3 establishes the asymptotic normal distribution under the null hypothesis of no causality in quantile. Section 4 displays a fairly extensive simulation study to illustrate the behavior of the test under the null, in addition to the power of the test under plausible alternatives. Section 5 considers the causal relations between the crude oil and gold prices as an economic application. Section 6 concludes the paper. Technical proofs are given in the Appendix.

2. NONPARAMETRIC TEST FOR GRANGER CAUSALITY IN QUANTILE

To simplify the exposition, we assume a bivariate case, or that only $\{y_t, w_t\}$ are observable. Granger causality in mean (Granger, 1988) is defined as follows.

1. w_t does not cause y_t in mean with respect to $\{y_{t-1}, \ldots, y_{t-p}, w_{t-1}, \ldots, w_{t-q}\}$ if

$$\mathrm{E}(y_t | y_{t-1}, \ldots, y_{t-p}, w_{t-1}, \ldots, w_{t-q}) = \mathrm{E}(y_t | y_{t-1}, \ldots, y_{t-p}) \quad \text{and}$$

2. w_t is a prima facie cause in mean of y_t with respect to $\{y_{t-1}, \ldots, y_{t-p}, w_{t-1}, \ldots, w_{t-q}\}$ if

$$\mathrm{E}(y_t | y_{t-1}, \ldots, y_{t-p}, w_{t-1}, \ldots, w_{t-q}) \neq \mathrm{E}(y_t | y_{t-1}, \ldots, y_{t-p}).$$

Motivated by the definition of Granger causality in mean, we define Granger causality in quantile as follows.

1. w_t does not cause y_t in the θ-quantile with respect to $\{y_{t-1}, \ldots, y_{t-p}, w_{t-1}, \ldots, w_{t-q}\}$ if

$$Q_\theta(y_t | y_{t-1}, \ldots, y_{t-p}, w_{t-1}, \ldots, w_{t-q}) = Q_\theta(y_t | y_{t-1}, \ldots, y_{t-p}). \quad \textbf{(1)}$$

2. w_t is a prima facie cause in the θ-quantile of y_t with respect to $\{y_{t-1}, \ldots, y_{t-p}, w_{t-1}, \ldots, w_{t-q}\}$ if

$$Q_\theta(y_t | y_{t-1}, \ldots, y_{t-p}, w_{t-1}, \ldots, w_{t-q}) \neq Q_\theta(y_t | y_{t-1}, \ldots, y_{t-p}), \quad \textbf{(2)}$$

where $Q_\theta(y_t|\cdot)$ is the θth $(0 < \theta < 1)$ conditional quantile of y_t given $\cdot$, which depends on t.

Denote $x_t \equiv (y_{t-1}, \ldots, y_{t-p})$, $z_t \equiv (y_{t-1}, \ldots, y_{t-p}, w_{t-1}, \ldots, w_{t-q})$, and the conditional distribution function y_t given $z_t(x_t)$ by $F_{y_t|z_t}(y_t|z_t)(F_{y_t|z_t}(y_t|x_t))$, which is abbreviated as $F_{y|z}(y|z)$ $(F_{y|x}(y|x))$ later, and $v_t = (x_t, z_t)$. In this paper, $F_{y|z}(y|z)$ is assumed to be absolutely continuous in y for almost all $v = (x, z)$. Denote $Q_\theta(z_t) \equiv Q_\theta(y_t|z_t)$ and $Q_\theta(x_t) \equiv Q_\theta(y_t|x_t)$. Then we have, with probability 1,

$$F_{y|z}\{Q_\theta(z_t)|z_t\} = \theta, \qquad v = (x, z) \quad \text{and}$$

from the definitions (1) and (2), the hypotheses to be tested are

$$\mathrm{H}_0 : \mathrm{P}\left\{F_{y|z}(Q_\theta(x_t)|z_t) = \theta\right\} = 1 \quad \text{a.s.} \quad \textbf{(3)}$$

$$\mathrm{H}_1 : \mathrm{P}\left\{F_{y|z}(Q_\theta(x_t)|z_t) = \theta\right\} < 1 \quad \text{a.s.} \quad \textbf{(4)}$$

Zheng (1998) proposed an idea to reduce the problem of testing a quantile restriction to a problem of testing a particular type of mean restriction. The null hypothesis (3) is true if and only if $\mathrm{E}[1\{y_t \leqslant Q_\theta(x_t)|z_t\}] = \theta$ or $1\{y_t \leqslant Q_\theta(x_t)\} = \theta + \varepsilon_t$ where $\mathrm{E}(\varepsilon_t|z_t) = 0$ and $1(\cdot)$ is the indicator function. For a list of related literature we refer to Li and Wang (1998) and Zheng (1998). Although various distance measures can be used to consistently test the hypothesis (3), we consider the following distance measure:

$$J \equiv \mathrm{E}\left[\left\{F_{y|z}(Q_\theta(x_t)|z_t) - \theta\right\}^2 f_z(z_t)\right], \tag{5}$$

with $f_{z_t}(z_t)$ being the marginal density function of z_t, which is sometimes abbreviated as $f_z(z_t)$. Note that $J \geqslant 0$ and the equality holds if, and only if, H_0 is true, with strict inequality holding under H_1. Thus J can be used as a proper candidate for consistent testing of H_0 (Li, 1999, p. 104). Because $\mathrm{E}(\varepsilon_t|z_t) = F_{y|z}\{Q_\theta(x_t)|z_t\} - \theta$ we have

$$J = \mathrm{E}\{\varepsilon_t \mathrm{E}(\varepsilon_t|z_t) f_z(z_t)\}. \tag{6}$$

The test is based on a sample analogue of $\mathrm{E}\{\varepsilon\, \mathrm{E}(\varepsilon|z) f_z(z)\}$. Including the density function $f_z(z)$ avoids the problem of trimming on the boundary of the density function; see Powell, Stock, and Stoker (1989) for an analogue approach. The density-weighted conditional expectation $\mathrm{E}(\varepsilon|z) f_z(z)$ can be estimated by kernel methods

$$\hat{\mathrm{E}}(\varepsilon_t|z_t)\hat{f}_z(z_t) = \frac{1}{(T-1)h^m}\sum_{s\neq t}^{T} K_{ts}\varepsilon_s, \tag{7}$$

where $m = p + q$ is the dimension of z, $K_{ts} = K\{(z_t - z_s)/h\}$ is the kernel function, and h is a bandwidth. Then we have a sample analogue of J as

$$\begin{aligned} J_T &\equiv \frac{1}{T(T-1)h^m}\sum_{t=1}^{T}\sum_{s\neq t}^{T} K_{ts}\varepsilon_t\varepsilon_s \\ &= \frac{1}{T(T-1)h^m}\sum_{t=1}^{T}\sum_{s\neq t}^{T} K_{ts}[1\{y_t \leqslant Q_\theta(x_t)\} - \theta][1\{y_s \leqslant Q_\theta(x_s)\} - \theta]. \end{aligned} \tag{8}$$

The θth conditional quantile of y_t given x_t, $Q_\theta(x_t)$, can also be estimated by the nonparametric kernel method

$$\hat{Q}_\theta(x_t) = \hat{F}^{-1}_{y|x}(\theta|x_t), \tag{9}$$

where

$$\hat{F}_{y|x}(y_t|x_t) = \frac{\sum_{s\neq t} L_{ts} 1(y_s \leqslant y_t)}{\sum_{s\neq t} L_{ts}} \tag{10}$$

is the Nadaraya–Watson kernel estimator of $F_{y|x}(y_t|x_t)$ with the kernel function of $L_{ts} = L(x_t - x_s)/a$ and the bandwidth parameter of α. The unknown error ε can be estimated as

$$\hat{\varepsilon}_t \equiv I\left\{y_t \leqslant \hat{Q}_\theta(x_t)\right\} - \theta. \tag{11}$$

Replacing ε by $\hat{\varepsilon}$, we have a feasible kernel-based test statistic of J,

$$\begin{aligned}\hat{J}_T &\equiv \frac{1}{T(T-1)h^m}\sum_{t=1}^{T}\sum_{s\neq t}^{T} K_{ts}\hat{\varepsilon}_t\hat{\varepsilon}_s \\ &= \frac{1}{T(T-1)h^m}\sum_{t=1}^{T}\sum_{s\neq t}^{T} K_{ts}\left[1\left\{y_t \leqslant \hat{Q}_\theta(x_t)\right\} - \theta\right]\left[1\left\{y_s \leqslant \hat{Q}_\theta(x_s)\right\} - \theta\right]. \end{aligned} \tag{12}$$

3. THE LIMITING DISTRIBUTIONS OF THE TEST STATISTIC

Two existing works are useful in deriving the limiting distribution of the test statistic; one is Theorem 2.3 of Franke and Mwita (2003) on the uniform convergence rate of a nonparametric quantile estimator; another is Lemma 2.1 of Li (1999) on the asymptotic distribution of a second-order degenerate U-statistic, which is derived from Theorem 2.1 of Fan and Li (1999). We restate these results in lemmas subsequently for ease of reference. We collect the assumptions needed for Theorem 3.1.

(A1)

(a) $\{y_t, w_t\}_{t=1}^{T}$ is strictly stationary and absolutely regular with mixing coefficients $\beta(\tau) = \mathcal{O}(\rho^\tau)$ for some $0 < \rho < 1$.

(b) For some integer $\upsilon \geqslant 2$, f_y, f_z, and f_x all are bounded and belong to $\mathfrak{A}_\upsilon^\infty$ (see (D2) later in this section).

(c) Use $\mu_s^t(z)$ ($\mu_s^t(\varepsilon)$) to denote the σ algebra generated by $(z_s, ..., z_t)$ $((\varepsilon_s, ..., \varepsilon_t))$ for $s \leq t$. With probability 1, $\mathrm{E}\left[\varepsilon_t | \mu_{-\infty}^t(z), \mu_{-\infty}^{t-1}(\varepsilon)\right] = 0$, that is, the error ε_t is a martingale difference process. The terms $\mathrm{E}\left[\left|\varepsilon_t^{4+\eta}\right|\right] < \infty$ and $\mathrm{E}\left[\left|\varepsilon_{t_1}^{i_1}\varepsilon_{t_2}^{i_2}\ldots\varepsilon_{t_l}^{i_l}\right|^{1+\xi}\right] < \infty$ for some arbitrarily small $\eta > 0$ and $\xi > 0$, where $2 \leqslant l \leqslant 4$ is an integer, $0 \leqslant i_j \leqslant 4$, and $\sum_{j=1}^{l} i_j \leqslant 8$. The terms $\sigma_\varepsilon^2(z) = \mathrm{E}(\varepsilon_t^2 | z_t = z)$ and $\mu_{\varepsilon 4}(z) = \mathrm{E}\left[\varepsilon_t^4 | z_t = z\right]$ all satisfy some Lipschitz conditions: $|a(u+\upsilon) - a(u)| \leqslant D(u)\|\upsilon\|$ with $\mathrm{E}\left[|D(z)|^{2+\eta'}\right] < \infty$ for some small $\eta' > 0$, where $a(\cdot) = \sigma_\varepsilon^2(\cdot)$, $\mu_{\varepsilon 4}(\cdot)$.

(d) Let $f_{\tau_1,...,\tau_l}(\,)$ be the joint probability density function of $\left(z_{\tau_1}, \ldots, z_{\tau_l}\right)$ $(1 \leqslant l \leqslant 3)$. Then $f_{\tau_1,...,\tau_l}(\,)$ is bounded and satisfies a Lipschitz condition: $\left|f_{\tau_1,...,\tau_l}(z_1+u_1, z_2+u_2, \ldots, z_l+u_l) - f_{\tau_1,...,\tau_l}(z_1, z_2, \ldots, z_l)\right| \leqslant D_{\tau_1,...,\tau_l}(z_1, \ldots, z_l)\|u\|$, where $u = (u_1, ..., u_l)$, $z = (z_1, ..., z_l)$, and $D_{\tau_1,...,\tau_l}(\,)$ is integrable and satisfies the condition that $\int\int\int D_{\tau_1,...,\tau_l}(z_1, \ldots, z_l)\|z\|^{2\xi} dz_1, \ldots, dz_l < M < \infty$ and $\int\int\int D_{\tau_1,...,\tau_l}(z_1, \ldots, z_l) f_{\tau_1,...,\tau_l}(z_1, \ldots, z_l) dz_1, ..., dz_l < M < \infty$ for some $\xi > 1$.

(e) For any y and x satisfying $0 < F_{y|x}(y|x) < 1$ and $f_x(x) > 0$, $F_{y|x}$ and $f_x(x)$ are continuous and bounded in x and y; for fixed y, the conditional distribution function $F_{y|x}$ and the conditional density function $f_{y|x}$ belong to $\mathfrak{A}_3^\infty$; $f_{y|x}(Q_\theta(x)|x) > 0$ for all x; $f_{y|x}$ is uniformly bounded in x and y by, say, c_f.

(f) For some compact set G, there are $\varepsilon > 0$ and $\gamma > 0$ such that $f_x \geqslant \gamma$ for all x in the ε-neighborhood $\{x \,|\, \|x-u\| < \varepsilon,\ u \in G\}$ of G. For the compact set G and some compact neighborhood Θ_0 of 0, the set $\Theta = \{v = Q_\theta(x) + \mu | x \in G, \mu \in \Theta_0\}$ is compact, and for some constant $c_0 > 0$, $f_{y|x}(y|x) \geqslant c_0$ for all $x \in G, v \in \Theta$.

(g) There is an increasing sequence s_T of positive integers such that for some finite A,

$$\tfrac{T}{s_T}\beta^{2s_T/(3T)}(s_T) \leqslant A, \qquad 1 \leqslant s_T \leqslant \tfrac{T}{2} \quad \text{for all } T \geqslant 1.$$

(A2)

(a) We use product kernels for both $L(\cdot)$ and $K(\cdot)$. Let l and k be their corresponding univariate kernel which is bounded and symmetric. Then $l(\cdot)$ is nonnegative, $l(\cdot) \in \Upsilon_v$, $k(\cdot)$ is nonnegative, and $k(\cdot) \in \Upsilon_2$.

(b) $h = \mathcal{O}(T^{-\alpha'})$ for some $0 < \alpha' < (7/8)m$.

(c) $a = \mathcal{O}(1)$ and $\tilde{S}_T = Ta^p(s_T \log T)^{-1} \to \infty$ for some $s_T \to \infty$.

(d) A positive number δ exists such that for $r = 2+\delta$ and a generic number M_0

$$\int\int \left|\frac{1}{h^m} K\left(\frac{z_1 - z_2}{h}\right)\right|^r dF_z(z_1) dF_z(z_2) \leqslant M_0 < \infty \quad \text{and}$$

$$\mathrm{E}\left|\frac{1}{h^m} K\left(\frac{z_1 - z_2}{h}\right)\right|^r \leqslant M_0 < \infty.$$

(e) For some δ' $(0 < \delta' < \delta)$, $\beta(T) = \mathcal{O}(T^{-(2+\delta')/\delta'})$.

The following definitions are due to Robinson (1988).

DEFINITION (D1). Υ_λ, $\lambda \geqslant 1$ *is the class of even functions* $k : R \to R$ *satisfying* $\int_R u^i k(u)du = \delta_{i0}$ $(i = 0, 1, \ldots, \lambda - 1)$,

$$k(u) = \mathcal{O}\left((1 + |u|^{\lambda+1+\varepsilon})^{-1}\right), \quad \textit{for some } \varepsilon > 0,$$

where δ_{ij} *is the Kronecker's delta.*

DEFINITION (D2). $\mathfrak{A}_\mu^\alpha$, $\alpha > 0$, $\mu > 0$ *is the class of functions* $g : R^m \to R$ *satisfying that* g is $(d-1)$*-times partially differentiable for* $d - 1 \leqslant \mu \leqslant d$; *for some* $\rho > 0$, $\sup_{y \in \phi_{z\rho}} |g(y) - g(z) - G_g(y,z)| / |y - z|^\mu \leqslant D_g(z)$ *for all* z, *where* $\phi_{z\rho} = \{y \,|\, |y - z| < \rho\}$; $G_g = 0$ *when* $d = 1$; G_g *is a* $(d-1)$*th degree*

homogeneous polynomial in $y - z$ with coefficients being the partial derivatives of g at z of orders 1 through $d-1$ when $d > 1$; and $g(z)$, its partial derivatives of order $d-1$ and less, and $D_g(z)$ have finite αth moments.

The functions in $\mathfrak{A}_\mu^\alpha$ are thus expanded in a Taylor series with a local Lipschitz condition on the remainder, (α, μ) depending simultaneously on smoothness and moment properties. Bounded functions in Lip(μ) (the Lipschitz class of degree μ) for $0 < \mu \leq 1$ are in $\mathfrak{A}_\mu^\alpha$; for $\mu > 1$, $\mathfrak{A}_\mu^\alpha$ contains the bounded and $(d-1)$-times boundedly differentiable functions whose $(d-1)$th partial derivatives are in Lip $(\mu - d + 1))$. In applying $\mathfrak{A}_\mu^\alpha$ to f and F, we take $\alpha = \infty$.

Conditions (A1)(a)–(d) and (A2)(a) and (b) are adopted from conditions (D1) and (D2) of Li (1999), which are used to derive the asymptotic normal distribution of a second-order degenerate U-statistic. Assumption (A1)(a) requires $\{y_t, w_t\}_{t=1}^T$ to be a stationary absolutely regular process with geometric decay rate. Assumptions (A1)(b)–(d) are mainly some smoothness and moment conditions; these conditions are quite weak in the sense that they are similar to those used in Fan and Li (1996) for the independent data case. However, for autoregressive conditionally skedastic (ARCH) or generalized autoregressive conditionally heteroskedastic (GARCH) type error processes as considered in Engle (1982) and Bollerslev (1986), the error term ε_t may not have finite fourth moments in some situations. For example, let $\varepsilon_t | \varepsilon_{t-1} \sim N(0, \alpha_0 + \alpha_1 \varepsilon_{t-1}^2)$. Engle (1982) showed that ε_t does not have a finite fourth moment if $\alpha_1 > 1/\sqrt{3}$. Thus, Assumption (A1)(c) will be violated in such a case.

Assumption (A2)(a) requires $L(\cdot)$ to be a vth- ($v \geqslant 2$) order kernel. This condition together with (A1)(b) ensures that the bias in the kernel estimation (of the null model) is bounded. The requirement that k is a nonnegative second-order kernel function in (A2)(b) is a quite weak and standard assumption.

Conditions (A1)(e)–(g) and (A2)(c) are technical conditions (A1), (A2), (B1), (B2), (C1), and (C2) of Theorem 2.3 of Franke and Mwita (2003), which are required to get the uniform convergence rate of the nonparametric kernel estimator of the conditional distribution function and corresponding conditional quantile with mixing data. Because the simple ARCH models (Engle, 1982; Masry and Tjøstheim, 1995, 1997), their extensions (Diebolt and Guegan, 1993), and the bilinear Markovian models are geometrically strongly mixing under some general ergodicity conditions, Assumption (A1)(g) is usually satisfied. There also exist simple methods to determine the mixing rates for various classes of random processes, for example, Gaussian, Markov, autoregressive moving average, ARCH, and GARCH. Hence the assumption of a known mixing rate is reasonable and has been adopted in many studies, for example, Györfi, Härdle, Sarda, and Vieu (1989), Irle (1997), Meir (2000), Modha and Masry (1998), Roussas (1988), and Yu (1993). Auestad and Tjøstheim (1990) provided excellent discussions on the role of mixing for model identification in nonlinear time series analysis. But since the restriction of Assumption (A1)(c) as discussed before, ARCH or GARCH type processes may not satisfy all assumptions here. Finally conditions (A2)(d)

and (e) are adopted from conditions of Lemma 3.2 of Yoshihara (1976), which are required to get the asymptotic equivalence of the nondegenerate U-statistic and its projection under the β-mixing process. They are technical assumptions and are quite standard.

LEMMA 3.1 (Franke and Mwita, 2003). *Suppose conditions (A1)(e)–(g) and (A2)(c) hold. The bandwidth sequence is such that* $a = \mathcal{O}(1)$ *and* $\tilde{S}_T = Ta^p(s_T \log T)^{-1} \to \infty$ *for some* $s_T \to \infty$. *Let* $S_T = a^2 + \tilde{S}_T^{-1/2}$. *Then for the nonparametric kernel estimator of the conditional quantile of* $\hat{Q}_\theta(x_t)$, *equation (9), we have*

$$\sup_{\|x\| \in G} \left| \hat{Q}_\theta(x) - Q_\theta(x) \right| = \mathcal{O}(S_T) + \mathcal{O}\left(\frac{1}{Ta^p}\right) \quad a.s. \tag{13}$$

LEMMA 3.2 (Li, 1999). *Let* $L_t = (\varepsilon_t, z_t)^T$ *be a stochastic process that satisfies conditions (A1)(a)–(d).* $\varepsilon_t \in R$, $z_t \in R^m$, *and* $K(\cdot)$ *be the kernel function with* h *being the smoothing parameter that satisfies conditions (A2)(a) and (b). Define*

$$\sigma_\varepsilon^2(z) = \mathrm{E}[\varepsilon_t^2 |\, z_t = z] \quad \textit{and} \tag{14}$$

$$J_T \equiv \frac{1}{T(T-1)h^m} \sum_{t=1}^{T} \sum_{s \neq t}^{T} K_{ts} \varepsilon_t \varepsilon_s. \tag{15}$$

Then

$$Th^{m/2} J_T \;\to\; N(0, \sigma_0^2) \quad \textit{in distribution,} \tag{16}$$

where $\sigma_0^2 = 2\mathrm{E}\left\{\sigma_\varepsilon^4(z_t) f_z(z_t)\right\} \left\{\int K^2(u)du\right\}$ *and* $f_z(\cdot)$ *is the marginal density function of* z_t.

We consider testing for local departures from the null that converge to the null at the rate $T^{-1/2}h^{-m/4}$. More precisely we consider the sequence of local alternatives

$$H_{1T} \;:\; F_{y|z}\{Q_\theta(x_t) + d_T l(z_t) | z_t\} = \theta, \tag{17}$$

where $d_T = T^{-1/2}h^{-m/4}$ and the function $l(\cdot)$ and its first-order derivatives are bounded.

THEOREM 3.1. *Assume the conditions (A1) and (A2). Then*

(i) Under the null hypothesis (3), $Th^{m/2}\hat{J}_T \xrightarrow{L} N(0, \sigma_0^2)$ *in distribution, where*

$$\sigma_0^2 = 2\mathrm{E}\left\{\sigma_\varepsilon^4(z_t) f_z(z_t)\right\} \left\{\int K^2(u)du\right\} \quad \textit{and}$$

$$\sigma_\varepsilon^2(z_t) = \mathrm{E}(\varepsilon_t^2 | z_t) = \theta(1-\theta).$$

(ii) Under the null hypothesis (3), $\hat{\sigma}_0^2 \equiv 2\theta^2(1-\theta)^2 1/(T(T-1)h^m) \sum_{s \neq t} K_{ts}^2$ *is a consistent estimator of* $\sigma_0^2 = 2\mathrm{E}\left\{\sigma_\varepsilon^4(z_t) f_z(z_t)\right\} \int K^2(u)du$. *Thus*

$$Th^{m/2}\hat{J}_T/\hat{\sigma}_0 = \sqrt{\frac{T}{T-1}}\frac{\sum_{t=1}^{T}\sum_{s\neq t}^{T} K_{ts}\left[1\left\{y_t \leqslant \hat{Q}_\theta(x_t)\right\}-\theta\right]\left[1\left\{y_s \leqslant \hat{Q}_\theta(x_s)\right\}-\theta\right]}{\sqrt{2}\theta(1-\theta)\sqrt{\sum_{s\neq t} K_{ts}^2}}.$$

(iii) Under the alternative hypothesis (4),

$$\hat{J}_T \to \mathrm{E}\{[F_{y|z}(Q_\theta(x_t)|z_t)-\theta]^2 f_z(z_t)\} > 0 \quad \textit{in probability.}$$

(iv) Under the local alternatives (A.2) in the Appendix, $Th^{m/2}\hat{J}_T \to N(\mu,\sigma_1^2)$ *in distribution, where*

$$\mu = \mathrm{E}\left[f_{y|z}^2\{Q_\theta(z_t)|z_t\}l^2(z_t)f_z(z_t)\right],$$

$$\sigma_1^2 = 2\mathrm{E}\left\{\sigma_v^4(z_t)f_z(z_t)\right\}\int K^2(u)du, \quad \textit{and}$$

$$\sigma_v^2(z_t) = \mathrm{E}(v_t^2|z_t) \quad \textit{with } v_t \equiv I\{y_t \leqslant Q_\theta(x_t)\} - F(Q_\theta(x_t)|z_t).$$

Theorem 3.1 generalizes the results of Zheng (1998) for independent data to the weakly dependent data case. A detailed proof of Theorem 3.1 is given in the Appendix. The main difficulty in deriving the asymptotic distribution of the statistic defined in equation (12) is that a nonparametric quantile estimator is included in the indicator function that is not differentiable with respect to the quantile argument and thus prevents taking a Taylor expansion around the true conditional quantile $Q_\theta(x_t)$. To circumvent the problem, Zheng (1998) made use of the work of Sherman (1994) on uniform convergence of U-statistics indexed by parameters. In this paper, we bound the test statistic by the statistics in which the nonparametric quantile estimator in the indicator function is replaced with sums of the true conditional quantile and upper and lower bounds consistent with the uniform convergence rate of the nonparametric quantile estimator, $1(y_t \leqslant Q_\theta(x_t) - C_T)$ and $1(y_t \leqslant Q_\theta(x_t) + C_T)$.

An important further step is to show that the differences of the ideal test statistic J_T given in equation (8) and the statistics having the indicator functions obtained from the first step stated previously are asymptotically negligible. We may directly show that the second moments of the differences are asymptotically negligible by using the result of Yoshihara (1976) on the bound of moments of U-statistics for absolutely regular processes. However, it is tedious to get bounds on the second moments with dependent data. In the proof we use instead the fact that the differences are second-order degenerate U-statistics. Thus by using the result on the asymptotic normal distribution of the second-order degenerate U-statistic of Fan and Li (1999), we can derive the asymptotic variance that is based on the independent and identically distributed (i.i.d.) sequence having the same marginal distributions as weakly dependent variables in the test statistic. With this little

trick we only need to show that the asymptotic variance is $\mathcal{O}(1)$ in an i.i.d. situation. For details refer to the Appendix.

4. SIMULATION

We generate bivariate data $\{y_t, w_t\}_{t=1}^{T}$ according to the following model:

$$y_t = \frac{1}{2} y_{t-1} + c w_{t-1}^2 + \varepsilon_{1t},$$
$$w_t = 1 + \frac{1}{2} w_{t-1} + \varepsilon_{2t},$$

where ε_{1t} and ε_{2t} are independent standard normal random variables. Here $c = 0$ corresponds to the hypothetical model; that is, w_t does not cause y_t in the θ quantile with respect to $\{y_{t-1}, w_{t-1}\}$. All the coefficients are set such that the corresponding time series is stationary and β-mixing with corresponding densities bounded to satisfy the assumptions discussed before. We use different values of $c \in [0, 1]$ to investigate the power of the test, such that the higher c is, the stronger the causality of w_t on y_t is. Without loss of generality, we choose $\theta = 0.1, 0.5, 0.9$ and $T = 500, 1{,}000, 5{,}000$ here with the bandwidth h and a as in (7) and (10) as for a typical Nadaraya–Watson type estimator. We consider the nominal 0.05 significance level and repeat the test 500 times to generate the power.

Table 1 displays the power performance of the test for different combinations of T, c, and θ. First, obviously the power is very sensitive to the choice of T; that is, the larger T is, for the same c and θ, the larger the power is. From a technical point of view, this makes sense, because the more data we have, the more evidence we can draw from to detect the "causality" effect. Our asymptotic result, Theorem 3.1, needs the plug-in estimation of the asymptotic covariance matrix that is used to normalize the test statistic. Note that such an estimator is model-dependent and under the alternative is consistent with a different value than the one under the null. As a result, the power deteriorates for small T. On the other hand, whether the causality effect exists or not is the nature of the series, which is independent of the sample size used in this technical test. Enhancing the power performance for small-sample data using the simulation-based method deserves further research. Second, as discussed before, the higher c is, the stronger the causality of w_t on y_t is, which is confirmed by the larger and larger power values. Third, for different quantiles θ, we find that the powers with respect to $\theta = 0.5$ are usually larger than the powers with respect to $\theta = 0.1$ and 0.9.

5. APPLICATION TO COMMODITY PRICES

In financial and commodity markets, it has been argued that the covariation of the tails may be different from that of the rest of the distribution. The gold market is one of the most important markets in the world, where trading takes place 24 hours a day around the globe and transactions involving billions of dollars of

TABLE 1. Power performance for different combinations of T, c, and θ

c	Power (θ 0.1)	c	Power (θ 0.5)	c	Power (θ 0.9)
			$T = 500$		
0.00	0.024	0.00	0.108	0.00	0.010
0.03	0.030	0.03	0.288	0.03	0.020
0.06	0.058	0.06	0.796	0.06	0.108
0.09	0.190	0.09	0.991	0.09	0.585
0.12	0.414	0.12	1.000	0.12	0.950
0.15	0.696	0.15	1.000	0.15	0.994
0.18	0.888	0.18	1.000	0.18	1.000
0.21	0.962	0.21	1.000	0.21	1.000
0.24	0.988	0.24	1.000	0.24	1.000
0.27	1.000	0.27	1.000	0.27	1.000
0.30	1.000	0.30	1.000	0.30	1.000
			$T = 1{,}000$		
0.00	0.014	0.00	0.130	0.00	0.018
0.01	0.022	0.01	0.144	0.01	0.024
0.02	0.038	0.02	0.296	0.02	0.024
0.03	0.026	0.03	0.564	0.03	0.040
0.04	0.060	0.04	0.788	0.04	0.108
0.05	0.110	0.05	0.946	0.05	0.284
0.06	0.196	0.06	0.990	0.06	0.506
0.07	0.356	0.07	1.000	0.07	0.838
0.08	0.530	0.08	1.000	0.08	0.950
0.09	0.676	0.09	1.000	0.09	0.994
0.10	0.816	0.10	1.000	0.10	0.996
0.11	0.906	0.11	1.000	0.11	1.000
0.12	0.958	0.12	1.000	0.12	1.000
0.13	0.972	0.13	1.000	0.13	1.000
0.14	0.994	0.14	1.000	0.14	1.000
0.15	0.998	0.15	1.000	0.15	1.000
0.16	1.000	0.16	1.000	0.16	1.000
			$T = 5{,}000$		
0.00	0.020	0.00	0.116	0.00	0.026
0.01	0.028	0.01	0.328	0.01	0.046
0.02	0.124	0.02	0.904	0.02	0.142
0.03	0.490	0.03	1.000	0.03	0.728
0.04	0.924	0.04	1.000	0.04	0.988
0.05	1.000	0.05	1.000	0.05	1.000
0.06	1.000	0.06	1.000	0.06	1.000
0.07	1.000	0.07	1.000	0.07	1.000
0.08	1.000	0.08	1.000	0.08	1.000
0.09	1.000	0.09	1.000	0.09	1.000
0.10	1.000	0.10	1.000	0.10	1.000

TABLE 2. Unit root tests

Variable	Test type	Time trend term	Test statistics	CR value 5%	Unit root	Unit root after differencing
LN Oil	DF	no	0.86955	−1.94160	yes	no
	ADF	no	0.72255	−1.94160	yes	no
	PP	no	0.73107	−1.94160	yes	no
	KPSS	no	2.16221	0.14600	yes	no
	DF	include	−0.81819	−2.86386	yes	no
	ADF	include	−1.03287	−2.86386	yes	no
	PP	include	−0.94355	−2.86386	yes	no
	KPSS	include	2.16221	0.14600	yes	no
GBP	DF	no	−0.12461	−1.94160	yes	no
	ADF	no	−0.16186	−1.94160	yes	no
	PP	no	−0.12506	−1.94160	yes	no
	KPSS	no	5.26720	0.14600	yes	no
	DF	include	−1.53295	−2.86386	yes	no
	ADF	include	−1.51000	−2.86386	yes	no
	PP	include	−1.53853	−2.86386	yes	no
	KPSS	include	5.26720	0.14600	yes	no
LN Gold	DF	no	0.45931	−1.94160	yes	no
	ADF	no	1.03139	−1.94160	yes	no
	PP	no	0.69975	−1.94160	yes	no
	KPSS	no	3.50910	0.14600	yes	no
	DF	include	−1.98422	−2.86386	yes	no
	ADF	include	−1.36627	−2.86386	yes	no
	PP	include	−1.66336	−2.86386	yes	no
	KPSS	include	3.50910	0.14600	yes	no

Note: "LN Oil", "GBP", and "LN Gold" refer to the logarithmic Brent crude oil price, USD/GBP exchange rate, and logarithmic NYMEX spot gold price, respectively. The "Test types" DF, ADF, PP, and KPSS refer to unit root tests of, respectively, Dickey–Fuller (Fuller, 1976), augmented Dickey–Fuller (Fuller, 1976), Phillips–Perron (Phillips & Perron, 1988), and (Kwaitkowski et al., 1992).

gold are carried out each day. Understanding the mechanism of gold price changes is important for many outstanding issues in international economics and finance. Market participants are increasingly concerned with their exposure to large gold price fluctuations with special interest in which factors drive the changes. In this section, we apply the quantile causality test to investigate relations between the Brent crude oil, USD/GBP exchange rate and NYMEX spot gold prices (in USD per barrel and per ounce, respectively). The data, as seen in Figure 1, obtained from Datastream, are daily observations from 20 February 1997 to 17 July 2009 ($T = 3{,}237$). We use the USD/GBP instead of USD/EUR because the euro was only introduced as a new currency from 1 January 1999. As indicated by Table 2, we assume differenced logarithmic data are stationary and β-mixing with corresponding densities bounded. Because a long memory effect is not expected, we choose $p = q = 1$ and $m = 2$.

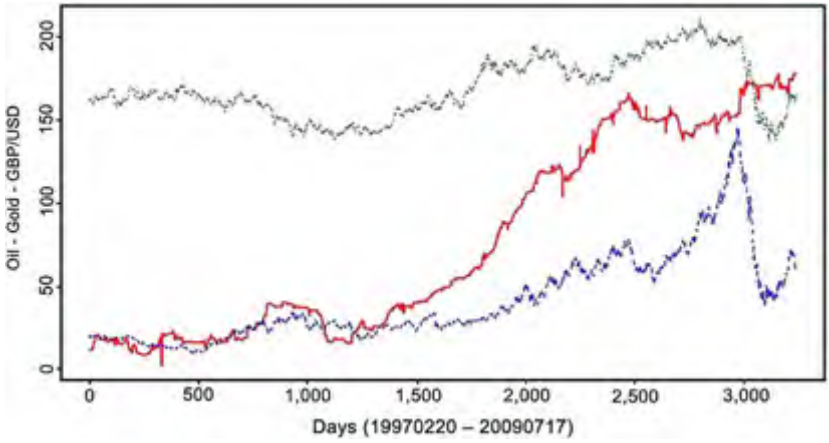


FIGURE 1. Plot of the gold prices, oil price, and exchange rate time series from 20 February 1997 to 17 July 2009.

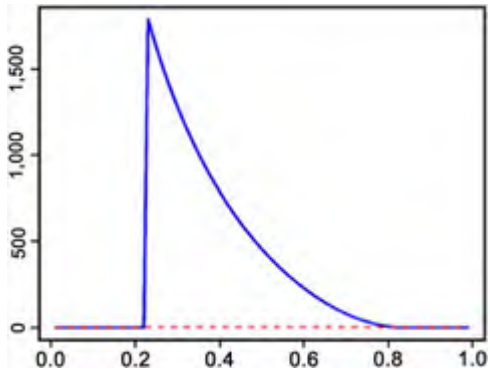


FIGURE 2. Test statistics with respect to different quantiles for the oil-gold prices causality test.

Figures 2 and 3 present results of testing whether oil prices Granger cause gold prices and whether the USD/GBP exchange rate Granger causes gold prices at the various quantiles, respectively, where logarithmic returns instead of the raw observations are used. The solid line and dotted line represent the standardized test statistics with respect to different quantiles (x-axis) and the critical value 1.96, respectively. In Figures 2 and 3, because the test statistic exceeds the critical value when $0.22 \leq \theta \leq 0.80$, we conclude that the oil price and exchange rate changes do not cause the gold price change in $\theta < 0.22$ or $\theta > 0.80$, whereas it is a prima facie cause in the $0.22 \leq \theta \leq 0.80$ quantile, respectively. For example, the oil price and USD/GBP exchange rate increases suggest that investors are wary of the U.S. dollar's weakness and inflation. Because gold is typically bought as an

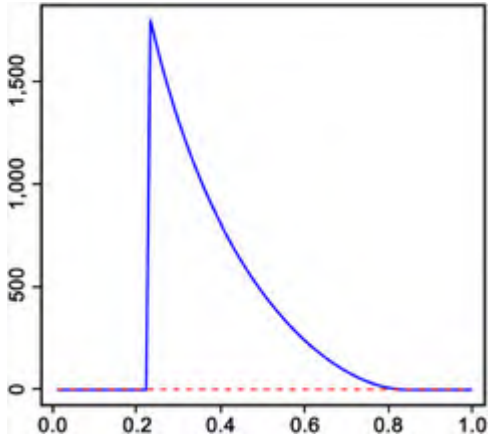


FIGURE 3. Test statistics with respect to different quantiles for the exchange rate-gold prices causality test.

alternative to the dollar among safe-haven assets, investors seeking safety from inflation risk and currency devaluation will cause the gold price to rise. However, the extreme low and high changes of the gold market may be caused by speculation. This is consistent with most of the empirical findings in the literature that the codependency may be stronger in the center than in the tails. By combining results from Figures 2 and 3, we find that the oil price and exchange rate changes have a significant predictive power for nonextreme gold price changes, which is, however, not significant for extreme changes. This finding could help to make it possible to use the gold price and GBP to hedge oil price changes in a more precise way with more careful investigation of their relations, which deserves further research.

6. CONCLUSION

By extending the Zheng (1998) idea to dependent data, we propose a consistent test for Granger causality in conditional quantile. The appealing feature of our proposed test is that it can investigate Granger causality in various conditional quantiles. The benefit of this is illustrated in the commodity market application where the causal relationships among the oil price, USD/GBP exchange rate, and gold price were shown to be different between a tail area and in the center of the distribution. We also illustrate that oil price and USD/GBP changes have significant predictive power on nonextreme gold price changes.

The test can be extended in a number of ways to test conditional quantile restrictions with dependent data: First, it can be extended to test functional misspecification, or the insignificance of a subset of regressors in quantile regression function, and second, it can also be used to test some semiparametric versus nonparametric models in quantile regression models.

REFERENCES

Auestad, B. & D. Tjøstheim (1990) Identification of nonlinear time series: First order characterisation and order determination. *Biometrica* 77, 669–687.

Bollerslev, T. (1986) Generalized autoregressive conditional heteroskedasticity. *Journal of Econometrics* 31, 307–327.

Bollerslev, T. (2001) Financial econometrics: Past developments and future challenges. *Journal of Econometrics* 100, 41–51.

Buchinsky, M. (1995) Quantile regression, Box-Cox transformation model, and the U.S. wage structure. 1963–1987. *Journal of Econometrics* 65, 65–154.

Campbell, J. & J. Cochrane (1999) By force of habit: A consumption-based explanation of aggregate stock market behaviour. *Journal of Political Economy* 107, 205–251.

Day, J.C. & E.C. Newburger (2002) The big payoff: Educational attainment and synthetic estimates of work-life earnings. Special studies. Current population reports. Statistical report p23–210, U.S. Department of Commerce, U.S. Census Bureau.

Diebolt, J. & D. Guégan (1993) Tail behavior of the stationary density of general nonlinear autoregressive processes of order 1. *Journal of Applied Probability* 30, 315–329.

Engle, R. (1982) Autoregressive conditional heteroskedasticity with estimates of the variance of United Kingdom inflation. *Econometrica* 50, 987–1007.

Fan, Y. & Q. Li (1996) Consistent model specification tests: Omitted variables, parametric and semiparametric functional forms. *Econometrica* 64, 865–890.

Fan, Y. & Q. Li (1999) Central limit theorem for degenerate U-statistics of absolutely regular processes with applications to model specification tests. *Journal of Nonparametric Statistics* 10, 245–271.

Franke, J. & P. Mwita (2003) Nonparametric Estimates for Conditional Quantiles of Time Series. Wirtschaftsmathematik 87, University of Kaiserslautern.

Fuller, W. (1976) *Introduction to Statistical Time Series*. Wiley.

Granger, C. (1969) Investigating causal relations by econometric models and cross-spectral methods. *Econometrica* 37, 424–438.

Granger, C. (1988) Some recent developments in a concept of causality. *Journal of Econometrics* 39, 199–211.

Györfi, L., W. Härdle, P. Sarda, & P. Vieu (1989) *Nonparametric Curve Estimation from Time Series*. Springer-Verlag.

Hall, P. (1984) Central limit theorem for integrated square error of multivariate nonparametric density estimators. *Journal of Multivariate Analysis* 14, 1–16.

Härdle, W., Y. Ritov, & S. Song (2009) Bootstrap Partial Linear Quantile Regression and Confidence Bands. SFB649 Discussion paper 2010-002, Humboldt Universität zu Berlin.

Härdle, W. & T. Stoker (1989) Investigating smooth multiple regression by the method of average derivatives. *Journal of the American Statistical Association* 84, 986–995.

Hong, Y., Y. Liu, & S. Wang (2009) Granger causality in risk and detection of extreme risk spillover between financial markets. *Journal of Econometrics* 150, 271–287.

Hsiao, C. & Q. Li (2001) A consistent test for conditional heteroskedasticity in time-series regression models. *Econometric Theory* 17, 188–221.

Irle, A. (1997) On the consistency in nonparametric estimation under mixing assumptions. *Multivariate Analysis* 60, 123–147.

Kwiatkowski, D., P.C.B. Phillips, P. Schmidt, & Y. Shin (1992) Testing the null hypothesis of stationarity against the alternative of a unit root. *Journal of Econometrics* 54, 159–178.

Lee, T. & W. Yang (2007) Money-Income Granger-Causality in Quantiles. Manuscript, University of California, Riverside.

Li, Q. (1999) Consistent model specification tests for time series econometric models. *Journal of Econometrics* 92, 101–147.

Li, Q. & S. Wang (1998) A simple consistent bootstrap test for a parametric regression functional form. *Journal of Econometrics* 87, 145–165.

Masry, E. & D. Tjøstheim (1995) Nonparametric estimation and identification of nonlinear ARCH time series: Strong convergence and asymptotic normality. *Econometric Theory* 11, 258–289.

Masry, E. & D. Tjøstheim (1997) Additive nonlinear ARX time series and projection estimates. *Econometric Theory* 13, 214–252.

Meir, R. (2000) Nonparametric time series prediction through adaptive model selection. *Machine Learning* 39, 5–34.

Modha, D. & E. Masry (1998) Memory-universal prediction of stationary random processes. *IEEE Transactions on Information Theory* 44, 117–133.

Phillips, P.C.B. & P. Perron (1988) Testing for a unit root in time series regression. *Biometricka* 75, 335–346.

Powell, J., J. Stock, & T. Stoker (1989) Semiparametric estimation of index coefficients. *Econometrica* 57, 1403–1430.

Robinson, P.M. (1988) Root-n-consistent semiparametric regression. *Econometrica* 56, 931–954.

Roussas, G.G. (1988) Nonparametric estimation in mixing sequences of random variables. *Journal of Statistical Planning and Inference* 18, 135–149.

Sherman, R. (1994) Maximal inequalities for degenerate U-processes with applications to optimization estimators. *Annals of Statistics* 22, 439–459.

Yoshihara, K. (1976) Limiting behavior of u-statistics for stationary absolutely regular processes. *Zeitschrift für Wahrscheinlichkeitstheorie und Verwandte Gebiete* 35, 237–252.

Yu, B. (1993) Density estimation in the l1 norm for dependent data with applications. *Annals of Statistics* 21, 711–735.

Zheng, J. (1998) A consistent nonparametric test of parametric regression models under conditional quantile restrictions. *Econometric Theory* 14, 123–138.

APPENDIX

Proof of Theorem 3.1(i). In the proof, we use several approximations to $\hat{J}_T$. We define them now and recall a few already defined statistics for convenience of reference.

$$\hat{J}_T \equiv \frac{1}{T(T-1)h^m} \sum_{t=1}^{T} \sum_{s \neq t}^{T} K_{ts} \hat{\varepsilon}_t \hat{\varepsilon}_s, \tag{A.1}$$

$$J_T \equiv \frac{1}{T(T-1)h^m} \sum_{t=1}^{T} \sum_{s \neq t}^{T} K_{ts} \varepsilon_t \varepsilon_s, \tag{A.2}$$

$$J_{TU} \equiv \frac{1}{T(T-1)h^m} \sum_{t=1}^{T} \sum_{s \neq t}^{T} K_{ts} \varepsilon_{tU} \varepsilon_{sU}, \tag{A.3}$$

$$J_{TL} \equiv \frac{1}{T(T-1)h^m} \sum_{t=1}^{T} \sum_{s \neq t}^{T} K_{ts} \varepsilon_{tL} \varepsilon_{sL}, \tag{A.4}$$

where
$$\begin{aligned} \hat{\varepsilon}_t &= I\left\{ y_t \leqslant \hat{Q}_\theta(x_t) \right\} - \theta, \\ \varepsilon_t &= I\{ y_t \leqslant Q_\theta(x_t) \} - \theta, \\ \varepsilon_{tU} &= I\{ y_t + C_T \leqslant Q_\theta(x_t) \} - \theta, \\ \varepsilon_{tL} &= I\{ y_t - C_T \leqslant Q_\theta(x_t) \} - \theta, \end{aligned}$$

and C_T is an upper bound consistent with the uniform convergence rate of the nonparametric estimator of conditional quantile given in equation (13). The proof of Theorem 3.1(i) consists of three steps.

Step 1. Asymptotic normality.

$$Th^{m/2}J_T \to N(0, \sigma_0^2), \tag{A.5}$$

where $\sigma_0^2 = 2\mathrm{E}\left\{\theta^2(1-\theta)^2 f(z_t)\right\}\left\{\int K^2(u)du\right\}$ under the null.

Step 2. Conditional asymptotic equivalence. Suppose that both $Th^{m/2}(J_T - J_{TU}) = \mathcal{O}_p(1)$ and $Th^{m/2}(J_T - J_{TL}) = \mathcal{O}_p(1)$.

$$\text{Then} \quad Th^{m/2}(\hat{J}_T - J_T) = \mathcal{O}_p(1). \tag{A.6}$$

Step 3. Asymptotic equivalence.

$$Th^{m/2}(J_T - J_{TU}) = \mathcal{O}_p(1) \quad \text{and} \quad Th^{m/2}(J_T - J_{TL}) = \mathcal{O}_p(1). \tag{A.7}$$

The combination of steps 1–3 yields Theorem 3.1(i).

Proof of Step 1. Because J_T is a degenerate U-statistic of order 2, the result follows from Lemma 3.2. ■

Proof of Step 2. The proof of step 2 is motivated by the technique of Härdle and Stoker (1989) that was used in treating trimming an indicator function asymptotically. Suppose that the following two statements hold:

$$Th^{m/2}(J_T - J_{TU}) = \mathcal{O}_p(1) \quad \text{and} \tag{A.8}$$

$$Th^{m/2}(J_T - J_{TL}) = \mathcal{O}_p(1). \tag{A.9}$$

Use C_T to denote an upper bound consistent with the uniform convergence rate of the nonparametric estimator of conditional quantile given in equation (13). Suppose that

$$\sup|\hat{Q}_\theta(x) - Q_\theta(x)| \leqslant C_T. \tag{A.10}$$

If inequality (A.10) holds, then the following statements also hold:

$$\{Q_\theta(x) > y_t + C_T\} \subset \{\hat{Q}_\theta(x) > y_t\} \subset \{Q_\theta(x) > y_t - C_T\}, \tag{A.11}$$

$$1(Q_\theta(x) > y_t + C_T) \leqslant 1(\hat{Q}_\theta(x) > y_t) \leqslant 1(Q_\theta(x) > y_t - C_T), \tag{A.12}$$

$$J_{TU} \leqslant \hat{J}_T \leqslant J_{TL}, \tag{A.13}$$

$$|Th^{m/2}(J_T - \hat{J}_T)| \leqslant \max\{|Th^{m/2}(J_T - J_{TU})|, |Th^{m/2}(J_T - J_{TL})|\}. \tag{A.14}$$

Using (A.10) and (A.14), we have the following inequality:

$$\begin{aligned}&\mathrm{P}\left\{|Th^{m/2}(J_T - \hat{J}_T)| > \delta \,|\, \sup\left|\hat{Q}_\theta(x) - Q_\theta(x)\right| \leqslant C_T\right\} \\ &\leqslant \mathrm{P}\left\{\max\{|Th^{m/2}(J_T - J_{TU})|, |Th^{m/2}(J_T - J_{TL})|\} > \delta \,\middle|\, \sup|\hat{Q}_\theta(x) - Q_\theta(x)| \leqslant C_T\right\} \\ &\quad \text{for all}\, \delta > 0.\end{aligned} \tag{A.15}$$

Invoking Lemma 3.1 and condition (A2)(c), we have

$$\mathrm{P}\left\{\sup|\hat{Q}_\theta(x)-Q_\theta(x)|\leqslant C_T\right\}\ \to\ 1\quad \text{as } T\ \to\ \infty. \tag{A.16}$$

By (A.8) and (A.9), as $T\ \to\ \infty$, we have

$$\mathrm{P}\left\{\max\{|Th^{m/2}(J_T-J_{TU})|,|Th^{m/2}(J_T-J_{TL})|\}>\delta\right\}\to\ 0\quad \text{for all } \delta>0. \tag{A.17}$$

Therefore, as $T\to\infty$,

the right-hand side of the inequality (A.15) $\times$ $\mathrm{P}\left\{\sup|\hat{Q}_\theta(x)-Q_\theta(x)|\leqslant C_T\right\}\to 0$;

the left-hand side of the inequality (A.15) $\times$ $\mathrm{P}\left\{\sup|\hat{Q}_\theta(x)-Q_\theta(x)|\leqslant C_T\right\}$

$$=\mathrm{P}\left\{|Th^{m/2}(J_T-\hat{J}_T)|>\delta\right\}\to\ 0.$$

In summary, we have that if both $Th^{m/2}(J_T-J_{TU})\ =\mathcal{O}_p(1)$ and $Th^{m/2}(J_T-J_{TL})\ =\mathcal{O}_p(1)$, then $Th^{m/2}(\hat{J}_T-J_T)\ =\mathcal{O}_p(1)$. ■

Proof of Step 3. In the remaining proof, we focus on showing that $Th^{m/2}(J_T-J_{TU})\ =\mathcal{O}_p(1)$, with the proof of $Th^{m/2}(J_T-J_{TL})\ =\mathcal{O}_p(1)$ being treated similarly. The proof of step 3 is close in line with the proof in Zheng (1998). Denote

$$H_T(s,t,g)\equiv K_{ts}\{1(y_t\leqslant g(x_t))-\theta\}\{1(y_s\leqslant g(x_s))-\theta\}\quad \text{and} \tag{A.18}$$

$$J[g]\equiv\frac{1}{T(T-1)h^m}\sum_{t=1}^{T}\sum_{s\neq t}^{T}H_T(s,t,g). \tag{A.19}$$

Then we have $J_T\equiv J[Q_\theta]$ and $J_{TU}\equiv J[Q_\theta-C_T]$. We decompose $H_T(s,t,g)$ into three parts:

$$\begin{aligned}H_T(s,t,g)&=K_{ts}\{1(y_t\leqslant g(x_t))-F(g(x_t)|z_t)\}\{1(y_s\leqslant g(x_s))-F(g(x_s)|z_s)\}\\&\quad+2\times K_{ts}\{1(y_t\leqslant g(x_t))-F(g(x_t)|z_t)\}\{F(g(x_s)|z_s)-\theta\}\\&\quad+K_{ts}\{F(g(x_t)|z_t)-\theta\}\{F(g(x_s)|z_s)-\theta\}\\&=H_{1T}(s,t,g)+2H_{2T}(s,t,g)+H_{3T}(s,t,g).\end{aligned} \tag{A.20}$$

Then let $J_j[g]=1/(T(T-1)h^m)\sum_{t=1}^{T}\sum_{s\neq t}^{T}H_{jT}(s,t,g)$, $i=1,2,3$. We will treat $J_j[Q_\theta]-J_j[Q_\theta-C_T]$ for $j=1,2,3$ separately.

(1) $Th^{m/2}\left[J_1(Q_\theta)-J_1(Q_\theta-C_T)\right]=\mathcal{O}_p(1)$. By simple manipulation, we have

$$\begin{aligned}&J_1(Q_\theta)-J_1(Q_\theta-C_T)\\&\quad=\frac{1}{T(T-1)h^m}\sum_{t=1}^{T}\sum_{s\neq t}^{T}\left[H_{1T}(s,t,Q_\theta)-H_{1T}(s,t,Q_\theta-C_T)\right]\\&\quad=\frac{1}{T(T-1)h^m}\sum_{t=1}^{T}\sum_{s\neq t}^{T}K_{ts}\Big\{[1(y_t\leqslant Q_\theta(x_t))-F(Q_\theta(x_t)|z_t)]\\&\qquad\qquad\times[1(y_s\leqslant Q_\theta(x_s))-F(Q_\theta(x_s)|z_s)]\end{aligned}$$

$$-[1(y_t \leqslant (Q_\theta(x_t) - C_T)) - F((Q_\theta(x_t) - C_T)|z_t)]$$
$$\times [1(y_s \leqslant (Q_\theta(x_s) - C_T)) - F((Q_\theta(x_s) - C_T)|z_s)]\Big\}. \quad \textbf{(A.21)}$$

To avoid tedious work to get bounds on the second moment of $J_1(Q_\theta) - J_1(Q_\theta - C_T)$ with dependent data, we note that the right-hand side of (A.21) is a degenerate U-statistic of order 2. Thus we can apply Lemma 3.2 and have

$$Th^{m/2}\left[J_1(Q_\theta) - J_1(Q_\theta - C_T)\right] \to N(0, \sigma_2^2) \quad \text{in distribution,} \quad \textbf{(A.22)}$$

where the definition of the asymptotic variance σ_2^2 is based on the i.i.d. sequence having the same marginal distributions as weakly dependent variables in (A.21). That is,

$$\sigma_2^2 = 2h^{-m}\tilde{\mathrm{E}}\big[H_{1T}(s,t,Q_\theta) - H_{1T}(s,t,Q_\theta - C_T)\big]^2,$$

where the notation $\tilde{\mathrm{E}}$ is an expectation evaluated at an i.i.d. sequence having the same marginal distribution as the mixing sequences in (A.21) (Fan and Li, 1999, p. 248). Now, to show that $Th^{m/2}\left[J_1(Q_\theta) - J_1(Q_\theta - C_T)\right] = \mathcal{O}_p(1)$, we only need to show that the asymptotic variance $\sigma_2^2(z)$ is $\mathcal{O}(1)$ with i.i.d. data. Use Λ_T to denote an upper bound consistent with the integral over K_{ts} being of the order $\mathcal{O}(h^m)$. We have

$$\begin{aligned}
&\tilde{\mathrm{E}}\big[H_{1T}(s,t,Q_\theta) - H_{1T}(s,t,Q_\theta - C_T)\big]^2 \\
&\quad \leqslant \Lambda_T\tilde{\mathrm{E}}\{[1_t(Q_\theta) - F_t(Q_\theta)][1_s(Q_\theta) - F_s(Q_\theta)] \\
&\qquad\quad -[1_t(Q_\theta - C_T) - F_t(Q_\theta - C_T)][1_s(Q_\theta - C_T) - F_s(Q_\theta - C_T)]\}^2 \\
&\quad \leqslant \Lambda_T\tilde{\mathrm{E}}\{F_t(Q_\theta)[1 - F_t(Q_\theta)]F_s(Q_\theta)[1 - F_s(Q_\theta)]\} \\
&\qquad + \tilde{\mathrm{E}}\{F_t(Q_\theta - C_T)[1 - F_t(Q_\theta - C_T)]F_s(Q_\theta - C_T)[1 - F_s(Q_\theta - C_T)]\} \\
&\qquad - 2\mathrm{E}\{[F_t(\min(Q_\theta, Q_\theta - C_T)) - F_t(Q_\theta)F_t(Q_\theta - C_T)] \\
&\qquad\qquad \times\ [F_s(\min(Q_\theta, Q_\theta - C_T)) - F_s(Q_\theta)F_s(Q_\theta - C_T)]\} \\
&\quad = \Lambda_T\tilde{\mathrm{E}}\{[F_t(Q_\theta) - F_t(Q_\theta)F_t(Q_\theta)][F_s(Q_\theta) - F_s(Q_\theta)F_s(Q_\theta)]\} \\
&\qquad - \Lambda_T\tilde{\mathrm{E}}\{[F_t(\min(Q_\theta, Q_\theta - C_T)) - F_t(Q_\theta)F_t(Q_\theta - C_T)] \\
&\qquad\qquad \times [F_s(\min(Q_\theta, Q_\theta - C_T)) - F_s(Q_\theta)F_s(Q_\theta - C_T)]\} \\
&\qquad + \Lambda_T\tilde{\mathrm{E}}\{[F_t(Q_\theta - C_T) - F_t(Q_\theta - C_T)F_t(Q_\theta - C_T)] \\
&\qquad\qquad \times [F_s(Q_\theta - C_T) - F_s(Q_\theta - C_T)F_s(Q_\theta - C_T)]\} \\
&\qquad - \Lambda_T\tilde{\mathrm{E}}\{[F_t(\min(Q_\theta, Q_\theta - C_T)) - F_t(Q_\theta)F_t(Q_\theta - C_T)] \\
&\qquad\qquad \times [F_s(\min(Q_\theta, Q_\theta - C_T)) - F_s(Q_\theta)F_s(Q_\theta - C_T)]\} \\
&\quad \leqslant \Lambda_T C_T. \qquad \textbf{(A.23)}
\end{aligned}$$

Thus we have that $\sigma_2^2 = \mathcal{O}(C_T) = \mathcal{O}(1)$, and so

$$Th^{m/2}\left[J_1(Q_\theta) - J_1(Q_\theta - C_T)\right] = \mathcal{O}_p(1). \quad \textbf{(A.24)}$$

(2) $Th^{m/2}\left[J_2(Q_\theta) - J_2(Q_\theta - C_T)\right] = \mathcal{O}_p(1)$. Note that $H_{2T}(s,t,Q_\theta) = 0$ because of $F_{y|z}(Q_\theta(x_s)|z_s) - \theta = 0$. Then we have

$$\begin{aligned} J_2(Q_\theta) - J_2(Q_\theta - C_T) &= -J_2(Q_\theta - C_T) \\ &= -\frac{1}{T(T-1)} \sum_{t=1}^{T} \sum_{s \neq t}^{T} \frac{1}{h^m} K\left(\frac{z_t - z_s}{h}\right) \\ &\quad \times \{1(y_t \leqslant Q_\theta(x_t) - C_T) - F_{y|z}(Q_\theta(x_t) - C_T|z_t)\} \\ &\quad \times \{F_{y|z}(Q_\theta(x_s) - C_T|z_s) - \theta\}. \end{aligned} \tag{A.25}$$

By taking a Taylor expansion of $F_{y|z}(Q_\theta(x_s) - C_T|z_s)$ around $Q_\theta(x_s)$, it equals

$$\begin{aligned} &-\frac{1}{T(T-1)} \sum_{t=1}^{T} \sum_{s \neq t}^{T} \frac{1}{h^m} K\left(\frac{z_t - z_s}{h}\right) \\ &\quad \times \{1(y_t \leqslant Q_\theta(x_t) - C_T) - F_{y|z}(Q_\theta(x_t) - C_T|z_t)\} \\ &\quad \times (-C_T) f_{y|z}(\bar{Q}_\theta(x_s)|z_s), \end{aligned} \tag{A.26}$$

where $\bar{Q}_\theta$ is between Q_θ and $Q_\theta - C_T$. Thus we have

$$\begin{aligned} &(J_2(Q_\theta) - J_2(Q_\theta - C_T))^2 \\ &\quad \leqslant \left[\frac{1}{T(T-1)} \sum_{t=1}^{T} \sum_{s \neq t}^{T} \frac{1}{h^m} K\left(\frac{z_t - z_s}{h}\right) \right. \\ &\qquad \left. \times \left\{1(y_t \leqslant Q_\theta(x_t) - C_T) - F_{y|z}(Q_\theta(x_t) - C_T|z_t)\right\}\right]^2 \Lambda^2 C_T^2 \\ &\quad = \Lambda^2 C_T^2 \left[\frac{1}{T} \sum_{t=1}^{T} \left\{1(y_t \leqslant Q_\theta(x_t) - C_T) - F_{y|z}(Q_\theta(x_t) - C_T)\right\} \hat{f}_z(z_t)\right]^2 \\ &\quad \equiv \Lambda^2 C_T^2 \left\{\frac{1}{T} \sum_{t=1}^{T} u_t \hat{f}_z(z_t)\right\}^2 \\ &\quad = \Lambda^2 C_T^2 T^{-2} \sum_{t=1}^{T} u_t^2 \hat{f}_z^2(z_t) + \Lambda^2 C_T^2 T^{-2} \sum_{t=1}^{T} \sum_{s \neq t}^{T} u_t u_s \hat{f}_z(z_t) \hat{f}_z(z_s) \\ &\quad \equiv J_{21} + J_{22}, \end{aligned} \tag{A.27}$$

where the inequality holds because of Assumption (A.1)(e).

$$\begin{aligned} \mathrm{E}|J_{21}| &= \Lambda^2 C_T^2 T^{-1} \left[T^{-1} \sum_{t=1}^{T} \mathrm{E}\left\{u_t^2 \hat{f}_z^2(z_t)\right\}\right] \\ &= O\left(C_T^2 T^{-2} h^{-m}\right), \end{aligned} \tag{A.28}$$

where the second equality is derived by using Lemma C.3(iii) of Li (1999).

$$J_{22} = \Lambda^2 C_T^2 \left[T^{-2} \sum_{t=1}^{T} \sum_{s \neq t}^{T} u_t u_s f_z(z_t) f_z(z_s) \right.$$
$$+ 2T^{-2} \sum_{t=1}^{T} \sum_{s \neq t}^{T} u_t u_s f_z(z_t) \left\{ \hat{f}_z(z_s) - f_z(z_s) \right\}$$
$$\left. + T^{-2} \sum_{t=1}^{T} \sum_{s \neq t}^{T} u_t u_s \left\{ \hat{f}_z(z_t) - f_z(z_t) \right\} \left\{ \hat{f}_z(z_s) - f_z(z_s) \right\} \right]$$
$$\equiv \Lambda C_T^2 \left(J_{221} + J_{222} + J_{223} \right). \tag{A.29}$$

Following the line of the proof of Lemma A.2(i) of Li (1999) we have that

$$J_{221} = \mathcal{O}_p\left(T^{-2}\right), \qquad J_{222} = \mathcal{O}_p\left(T^{-1}\right), \quad \text{and} \quad J_{223} = \mathcal{O}_p\left(T^{-1}\right); \quad \text{thus}$$
$$J_{22} = \mathcal{O}_p\left(C_T^2 T^{-1}\right). \tag{A.30}$$

Thus, combining (A.28) and (A.30), we have

$$Th^{m/2}\left[J_2(Q_\theta) - J_2(Q_\theta - C_T)\right] = \mathcal{O}_p\left(C_T\right) + \mathcal{O}_p\left(C_T T^{1/2} h^{m/2}\right)$$
$$= \mathcal{O}_p(1). \tag{A.31}$$

(3) $Th^{m/2}\left[J_3(Q_\theta) - J_3(Q_\theta - C_T)\right] = \mathcal{O}_p(1)$. Noting that $H_{3T}(s,t,Q_\theta) = 0$ because of $F(Q_\theta(x_j)|z_j) - \theta = 0$ for $j = t, s$, we have

$$J_3(Q_\theta) - J_3(Q_\theta - C_T)$$
$$= -\frac{1}{T(T-1)} \sum_{t=1}^{T} \sum_{s \neq t}^{T} \frac{1}{h^m} K\left(\frac{z_t - z_s}{h}\right)$$
$$\times \{F(Q_\theta(x_t) - C_T|z_t) - \theta\}\{F(Q_\theta(x_s) - C_T|z_s) - \theta\}$$
$$= \frac{1}{T(T-1)} \sum_{t=1}^{T} \sum_{s \neq t}^{T} \frac{1}{h^m} K\left(\frac{z_t - z_s}{h}\right) C_T^2 f_{y|z}(\bar{Q}_\theta(x_t)|z_t) f_{y|z}(\bar{Q}_\theta(x_s)|z_s)$$
$$= C_T^2 \frac{1}{T} \sum_{t=1}^{T} f_{y|z}(\bar{Q}_\theta(x_t)|z_t) f_{y|z}(\bar{Q}_\theta(x_s)|z_s) \hat{f}_z(z_t). \tag{A.32}$$

Thus, we have

$$\mathrm{E}|J_3(Q_\theta) - J_3(Q_\theta - C_T)|$$
$$\leqslant \Lambda C_T^2 \frac{1}{T} \sum_{t=1}^{T} \mathrm{E}\left|\hat{f}_z(z_t)\right|$$
$$\leqslant \Lambda C_T^2 \frac{1}{T} \sum_{t=1}^{T} \mathrm{E}|f_z(z_t)| + \Lambda C_T^2 \frac{1}{T} \sum_{t=1}^{T} \mathrm{E}\left|\hat{f}_z(z_t) - f_z(z_t)\right|$$
$$= \mathcal{O}\left(C_T^2\right). \tag{A.33}$$

Finally, we have

$$Th^{m/2}\left[J_3(Q_\theta) - J_3(Q_\theta - C_T)\right] = \mathcal{O}_p\left(Th^{m/2}C_T^2\right)$$
$$= \mathcal{O}_p(1). \tag{A.34}$$

By combining (A.24), (A.31), and (A.34), we have the result of step 3. ■

Proof of Theorem 3.1(ii). Because

$$\sigma_0^2 = 2\theta^2(1-\theta)^2 \mathrm{E}\{f_z(z_t)\}\int K^2(u)du \quad \text{and}$$
$$\hat{\sigma}_0^2 \equiv 2\theta^2(1-\theta)^2 \frac{1}{T(T-1)h^m}\sum_{s\neq t} K_{ts}^2,$$

it is enough to show that

$$\sigma_T^2 \equiv \frac{1}{T(T-1)h^m}\sum_{s\neq t} K_{ts}^2$$
$$= \mathrm{E}\{f_z(z_t)\}\int K^2(u)du + \mathcal{O}_p(1). \tag{A.35}$$

Note that σ_T^2 is a nondegenerate U-statistic of order 2 with kernel

$$H_T(z_t, z_s) = \frac{1}{h^m}K^2\left(\frac{z_t - z_s}{h}\right). \tag{A.36}$$

Because Assumptions (A2)(d) and (e) satisfy the conditions of Lemma 3.2 of Yoshihara (1976) on the asymptotic equivalence of the U-statistic and its projection under β-mixing, we have for $\gamma = 2(\delta - \delta')/\delta'(2+\delta) > 0$

$$\begin{aligned}
\sigma_T^2 &\equiv \frac{1}{T(T-1)}\sum_{s\neq t} H_T(z_t, z_s) \\
&= \int\int H_T(z_1, z_2)dF_z(z_1)dF_z(z_2) \\
&\quad + 2T^{-1}\sum_{t=1}^{T}\left[\int H_T(z_t, z_2)dF_z(z_2) - \int\int H_T(z_1, z_2)dF_z(z_1)dF_z(z_2)\right] \\
&\quad + \mathcal{O}_p(T^{-1-\gamma}) \\
&= \int\int H_T(z_1, z_2)dF_z(z_1)dF_z(z_2) + \mathcal{O}_p(1) \\
&= \int\int \frac{1}{h^m}K^2\left(\frac{z_1 - z_2}{h}\right)dF_z(z_1)dF_z(z_2) + \mathcal{O}_p(1) \\
&= \int K^2(u)\,du \int f_z^2(z)dz + \mathcal{O}_p(1).
\end{aligned} \tag{A.37}$$

The result of Theorem 3.1(ii) follows from (A.37). ■

Proof of Theorem 3.1(iii). The proof of Theorem 3.1(iii) consists of two steps.

Step 1. Show that $\hat{J}_T = J_T + \mathcal{O}_p(1)$ under the alternative hypothesis (4).

Step 2. Show that $J_T = J + \mathcal{O}_p(1)$ under the alternative hypothesis (4),

where $J = \mathrm{E}\{[F_{y|z}(Q_\theta(x_t)|z_t) - \theta]^2 f_z(z_t)\}$. The combination of steps 1 and 2 yields Theorem 3.1(iii).

Proof of Step 1. We note that the results of step 2 and $Th^{m/2}\left[J_1(Q_\theta) - J_1(Q_\theta - C_T)\right] = \mathcal{O}_p(1)$ of step 3 in the proof of Theorem 3.1(i) still hold under the alternative hypothesis (4). Thus we focus on showing that $J_2(Q_\theta) - J_2(Q_\theta - C_T) = \mathcal{O}_p(1)$ and $J_3(Q_\theta) - J_3(Q_\theta - C_T) = \mathcal{O}_p(1)$.

We begin with showing that $J_2(Q_\theta) - J_2(Q_\theta - C_T) = \mathcal{O}_p(1)$. By the same procedures as in (A.27), we can show that $J_2(Q_\theta - C_T) = \mathcal{O}_p(T^{-1}h^{-m/2})$. Thus it remains to show that $J_2(Q_\theta) = \mathcal{O}_p(1)$. By taking a Taylor expansion of $F_{y|z}(Q_\theta(x_s)|z_s)$ around $Q_\theta(x_s)$, we have

$$\begin{aligned} J_2(Q_\theta) &= -\frac{1}{T(T-1)}\sum_{t=1}^{T}\sum_{s\neq t}^{T}\frac{1}{h^m}K\left(\frac{z_t - z_s}{h}\right) \\ &\quad \times\{1(y_t \leqslant Q_\theta(x_t)) - F_{y|z}(Q_\theta(x_t)|z_t)\} \times f_{y|z}(\bar{Q}_\theta(x_s)|z_s) \\ &= \frac{1}{T}\sum_{t=1}^{T}\{1(y_t \leqslant Q_\theta(x_t)) - F_{y|z}(Q_\theta(x_t))\} f_{y|z}(\bar{Q}_\theta(x_s)|z_s)\hat{f}_z(z_t) \\ &\equiv \frac{1}{T}\sum_{t=1}^{T} u_t\, f_{y|z}(\bar{Q}_\theta(x_s)|z_s)\hat{f}_z(z_t). \end{aligned} \tag{A.38}$$

By similar arguments as in (A.26) and (A.31), we have

$$J_2(Q_\theta) = \mathcal{O}\left(T^{-1}h^{-m}\right). \tag{A.39}$$

Next, we show that $Th^{m/2}\left[J_3(Q_\theta) - J_3(Q_\theta - C_T)\right] = \mathcal{O}_p(1)$ under the alternative hypothesis (4). Because $F(Q_\theta(x_j)|z_j) - \theta \neq 0$ for $j = t, s$ under the alternative hypothesis, we have

$$\begin{aligned} &J_3(Q_\theta) - J_3(Q_\theta - C_T) \\ &\quad = \frac{1}{T(T-1)}\sum_{t=1}^{T}\sum_{s\neq t}^{T}\frac{1}{h^m}K\left(\frac{z_t - z_s}{h}\right)\{F(Q_\theta(x_t)|z_t) - \theta\}\{F(Q_\theta(x_s)|z_s) - \theta\} \\ &\qquad - \frac{1}{T(T-1)}\sum_{t=1}^{T}\sum_{s\neq t}^{T}\frac{1}{h^m}K\left(\frac{z_t - z_s}{h}\right) \\ &\qquad \times \{F(Q_\theta(x_t) - C_T|z_t) - \theta\}\{F(Q_\theta(x_s) - C_T|z_s) - \theta\} \\ &\quad = \frac{1}{T}\sum_{t=1}^{T}\{F(Q_\theta(x_t)|z_t) - \theta\}\{F(Q_\theta(x_s)|z_s) - \theta\}\hat{f}_z(z_t) \\ &\qquad - \frac{1}{T}\sum_{t=1}^{T}\{F(Q_\theta(x_t) - C_T|z_t) - \theta\}\{F(Q_\theta(x_s) - C_T|z_s) - \theta\}\hat{f}_z(z_t). \end{aligned} \tag{A.40}$$

By taking a Taylor expansion of $F_{y|z}(Q_\theta(x_j) - C_T|z_j)$ around $Q_\theta(z_j)$ for $j = t, s$, we have

$$\begin{aligned} J_3(Q_\theta) - J_3(Q_\theta - C_T) = & \frac{1}{T}\sum_{t=1}^{T}\{F(Q_\theta(x_t)|z_t) - \theta\} C_T f_{y|z}(\bar{Q}_\theta(x_t)|z_t)\hat{f}_z(z_t) \\ & + \frac{1}{T}\sum_{t=1}^{T} C_T f_{y|z}(\bar{Q}_\theta(x_t)|z_t)\{F(Q_\theta(x_s)|z_s) - \theta\}\hat{f}_z(z_t) \\ & - \frac{1}{T}\sum_{t=1}^{T} C_T^2 f_{y|z}(\bar{Q}_\theta(x_t)|z_t) f_{y|z}(\bar{Q}_\theta(x_s)|z_s)\hat{f}_z(z_t). \end{aligned} \tag{A.41}$$

We further take a Taylor expansion of $F_{y|z}(Q_\theta(x_j)|z_j)$ around $Q_\theta(z_j)$ for $j = t, s$ and have

$$\begin{aligned} J_3(Q_\theta) - J_3(Q_\theta - C_T) = & \frac{1}{T}\sum_{t=1}^{T} f_{y|z}(\bar{Q}_\theta(x_t, z_t)|z_t) C_T f_{y|z}(\bar{Q}_\theta(x_s)|z_s)\hat{f}_z(z_t) \\ & + \frac{1}{T}\sum_{t=1}^{T} C_T f_{y|z}(\bar{Q}_\theta(x_t)|z_t) f_{y|z}(\bar{Q}_\theta(x_s, z_s)|z_s)\hat{f}_z(z_t) \\ & - \frac{1}{T}\sum_{t=1}^{T} C_T^2 f_{y|z}(\bar{Q}_\theta(x_t)|z_t) f_{y|z}(\bar{Q}_\theta(x_s)|z_s)\hat{f}_z(z_t), \end{aligned} \tag{A.42}$$

where $\bar{Q}_\theta(x_s, z_s)$ is between $Q_\theta(x_s)$ and $Q_\theta(z_s)$. Then by using the same procedures as in (A.30), we have

$$J_3(Q_\theta) - J_3(Q_\theta - C_T) = \mathcal{O}(C_T). \tag{A.43}$$

Now we have the result of step 1 for the proof of Theorem 3.1(iii). ■

Proof of Step 2. Using (7) and the uniform convergence rate of the kernel regression estimator under a β-mixing process, we have

$$\begin{aligned} J_T &= \frac{1}{T(T-1)h^m}\sum_{t=1}^{T}\sum_{s\neq t}^{T} K_{ts}\varepsilon_t\varepsilon_s \\ &= \frac{1}{T}\sum_{t=1}\hat{\mathrm{E}}(\varepsilon_t|z_t)\hat{f}_z(z_t)\varepsilon_t \\ &= \frac{1}{T}\sum_{t=1}\mathrm{E}(\varepsilon_t|z_t) f_z(z_t)\varepsilon_t + \frac{1}{T}\sum_{t=1}\left\{\hat{\mathrm{E}}(\varepsilon_t|z_t)\hat{f}_z(z_t) - \mathrm{E}(\varepsilon_t|z_t) f_z(z_t)\right\}\varepsilon_t \\ &= \frac{1}{T}\sum_{t=1}\mathrm{E}(\varepsilon_t|z_t) f_z(z_t)\varepsilon_t + \mathcal{O}_p(1) \\ &= \mathrm{E}\left[\mathrm{E}(\varepsilon_t|z_t) f_z(z_t)\varepsilon_t\right] + \mathcal{O}_p(1) \\ &= J + \mathcal{O}_p(1). \end{aligned} \tag{A.44}$$

■

Proof of Theorem 3.1(iv). The proof of Theorem 3.1(iv) is close in line with the proof in Zheng (1998). The proof of Theorem 3.1(iv) consists of two steps.

Step 1. Show that $\hat{J}_T = J_T + \mathcal{O}_p(T^{-1}h^{-m/2})$ under the alternative hypothesis (A.2).

Step 2. Show that $Th^{m/2}J_T \to N(\mu, \sigma_1^2)$ under the alternative hypothesis (A.2), where $\mu = \mathrm{E}\left[f_{y|z}^2\{Q_\theta(z_t)|z_t\}l^2(z_t)f_z(z_t)\right]$, $\sigma_1^2 = 2\mathrm{E}\left\{\sigma_v^4(z_t)f_z(z_t)\right\}$ $\int K^2(u)du$, and $\sigma_v^2(z_t) = \mathrm{E}(v_t^2|z_t)$ with $v_t \equiv I\{y_t \leqslant Q_\theta(x_t)\} - F(Q_\theta(x_t)|z_t)$.

Proof of Step 1. The results of step 1 in the proof of Theorem 3.1(iii) show that, under the general alternative hypothesis (4), the elements consisting of $\hat{J}_T - J_T$ are all $\mathcal{O}_p(T^{-1}h^{-m/2})$ except for $J_2(Q_\theta(x))$, the order of which is $\mathcal{O}\left(T^{-1}h^{-m}\right)$ as in (A.39). Thus we need to show that $J_2(Q_\theta(x)) = \mathcal{O}_p(T^{-1}h^{-m/2})$ under the local alternative hypothesis (A.2). Taking a Taylor expansion of $F_{y|z}\{Q_\theta(z_t) + d_T l(z_t)|z_t\}$ around $d_T = 0$, we have

$$F_{y|z}\{Q_\theta(z_t) + d_T l(z_t)|z_t\} = \theta + d_T f_{y|z}\{Q_\theta(z_t)|z_t\}l(z_t) + \mathcal{O}_p(d_T^2). \tag{A.45}$$

By similar procedures as in (A.38) and (A.39), we have

$$\begin{aligned}
J_2(Q_\theta(x)) &= -\frac{1}{T(T-1)}\sum_{t=1}^{T}\sum_{s\neq t}^{T}\frac{1}{h^m}K\left(\frac{z_t - z_s}{h}\right)\{1(y_t \leqslant Q_\theta(x_t)) - F_{y|z}(Q_\theta(x_t)|z_t)\} \\
&\quad \times d_T f_{y|z}\{Q_\theta(z_t)|z_t\}l(z_t) + \mathcal{O}_p\left(d_T^2\right) \\
&= -d_T\frac{1}{T}\sum_{t=1}^{T}\{1(y_t \leqslant Q_\theta(x_t)) - F_{y|z}(Q_\theta(x_t)|z_t)\} \\
&\quad \times f_{y|z}\{Q_\theta(z_t)|z_t\}l(z_t)\hat{f}_z(z_t) + \mathcal{O}_p\left(d_T^2\right) \\
&\equiv -d_T\frac{1}{T}\sum_{t=1}^{T}u_t f_{y|z}\{Q_\theta(z_t)|z_t\}l(z_t)\hat{f}_z(z_t) + \mathcal{O}_p\left(d_T^2\right) \\
&= \mathcal{O}_p\left(d_T^2\right).
\end{aligned} \tag{A.46}$$

■

Proof of Step 2. Taking a Taylor expansion of $F_{y|z}\{Q_\theta(z_t) + d_T l(z_t)|z_t\}$ around $d_T = 0$, we have

$$\begin{aligned}
J_T(Q_\theta(x)) &= \frac{1}{T(T-1)h^m}\sum_{t=1}^{T}\sum_{s\neq t}^{T}K_{ts}\{1(y_t \leqslant Q_\theta(x_t))F(Q_\theta(x_t)|z_t)\} \\
&\quad \times \{1(y_s \leqslant Q_\theta(x_s)) - F(Q_\theta(x_s)|z_s)\} \\
&\quad -\frac{2d_T}{T(T-1)h^m}\sum_{t=1}^{T}\sum_{s\neq t}^{T}K_{ts}\{1(y_t \leqslant Q_\theta(x_t)) - F(Q_\theta(x_t)|z_t)\} \\
&\quad \times f_{y|z}\{Q_\theta(z_s)|z_s\}l(z_s)
\end{aligned}$$

$$+\frac{d_T^2}{T(T-1)h^m}\sum_{t=1}^{T}\sum_{s\neq t}^{T}K_{ts}f_{y|z}\{Q_\theta(z_t)|z_t\}l(z_t)f_{y|z}\{Q_\theta(z_s)|z_s\}l(z_s)$$

$$+\mathcal{O}_p\left(d_T^2\right)$$

$$=T_{1T}-2d_TT_{2T}+d_T^2T_{3T}+\mathcal{O}_p\left(d_T^2\right). \tag{A.47}$$

Noting that T_{1T} is a degenerate U-statistic of order 2, by Lemma 3.2, we have

$$Th^{m/2}T_{1T}\ \to N\left(0,\sigma_1^2\right)\quad \text{in distribution,} \tag{A.48}$$

Similarly to the proof for (A.31), we can show that $T_{2T}=\mathcal{O}\left\{(Th^m)^{-1}\right\}$, and so $d_TT_{2T}=\mathcal{O}\left\{(Th^{m/2})^{-1}\right\}$. And by the same procedures as in (A.44), we have

$$T_{3T}\ \to\ \mathrm{E}\left[f_{y|z}^2\{Q_\theta(z_t)|z_t\}l^2(z_t)f_z(z_t)\right]\quad \text{in probability.} \tag{A.49}$$

Thus,

$$Th^{m/2}J_T\ \to N\left(\mu,\sigma_1^2\right), \tag{A.50}$$

where $\mu=\mathrm{E}\left[f_{y|z}^2\{Q_\theta(z_t)|z_t\}l^2(z_t)f_z(z_t)\right]$. ■

Journal of Multivariate Analysis 107 (2012) 244–262

Contents lists available at SciVerse ScienceDirect

Journal of Multivariate Analysis

journal homepage: www.elsevier.com/locate/jmva

Bootstrap confidence bands and partial linear quantile regression☆

Song Song [a,b,*], Ya'acov Ritov [c], Wolfgang K. Härdle [a]

[a] Humboldt-Universität zu Berlin, Germany
[b] The University of Texas at Austin, United States
[c] The Hebrew University of Jerusalem, Israel

ARTICLE INFO

Article history:
Received 13 June 2011
Available online 31 January 2012

JEL classification:
C14
C21
C31
J01
J31
J71

AMS subject classifications:
62F40
62G08
62G86

Keywords:
Bootstrap
Quantile regression
Confidence bands
Nonparametric fitting
Kernel smoothing
Partial linear model

ABSTRACT

In this paper bootstrap confidence bands are constructed for nonparametric quantile estimates of regression functions, where resampling is done from a suitably estimated empirical distribution function (edf) for residuals. It is known that the approximation error for the confidence band by the asymptotic Gumbel distribution is logarithmically slow. It is proved that the bootstrap approximation provides an improvement. The case of multidimensional and discrete regressor variables is dealt with using a partial linear model. An economic application considers the labor market differential effect with respect to different education levels.

1. Introduction

Quantile regression, as first introduced by Koenker and Bassett [25], is "gradually developing into a comprehensive strategy for completing the regression prediction" as claimed by Koenker and Hallock [26]. Quantile smoothing is an effective method to estimate quantile curves in a flexible nonparametric way. Since this technique makes no structural assumptions on the underlying curve, it is very important to have a device for understanding when observed features are significant and deciding between functional forms. For example, a question often asked in this context is whether or not an observed peak or valley is actually a feature of the underlying regression function or is only an artifact of the observational noise. For such issues, confidence bands (i.e., uniform over location) give an idea about the global variability of the estimate.

The nonparametric quantile estimate could be obtained either using a check function such as a robustified local linear smoother [10,35,36], or through estimating the conditional distribution function using the double-kernel local linear

☆ The financial support from the Deutsche Forschungsgemeinschaft via SFB 649 "Ökonomisches Risiko", Humboldt-Universität zu Berlin is gratefully acknowledged. Ya'acov Ritov's research is supported by an ISF grant and a Humboldt Award. We thank Thorsten Vogel and Alexandra Spitz-Oener for sharing their data, comments and suggestions.

* Correspondence to: The University of Texas at Austin, 78751 Austin, United States.
E-mail address: ssoonngg123@gmail.com (S. Song).

0047-259X/$ – see front matter © 2012 Elsevier Inc. All rights reserved.
doi:10.1016/j.jmva.2012.01.020

technique [11,35,36]. Besides these, [17] proposed a weighted version of the Nadaraya–Watson estimator, which was further studied by Cai [5]. In the previous work the theoretical focus has mainly been on obtaining consistency and asymptotic normality of the quantile smoother, and thereby providing the necessary ingredients to construct its pointwise confidence intervals. This, however, is not sufficient to get an idea about the global variability of the estimate; neither can it be used to correctly answer questions about the curve's shape, which contains the lack of fit test as an immediate application. This motivates us to construct the confidence bands.

To this end, [22] used strong approximations of the empirical process and extreme value theory. However, the very poor convergence rate of extremes of a sequence of n independent normal random variables is well documented and was first noticed and investigated by Fisher and Tippett [12], and discussed in greater detail by Hall [16]. In the latter paper it was shown that the rate of the convergence to its limit (the suprema of a stationary Gaussian process) can be no faster than $(\log n)^{-1}$. For example, the supremum of a nonparametric quantile estimate can converge to its limit no faster than $(\log n)^{-1}$. These results may make extreme value approximation of the distributions of suprema somewhat doubtful, for example in the context of the uniform confidence band construction for a nonparametric quantile estimate.

This paper proposes and analyzes a bootstrap-based method of obtaining the confidence bands for nonparametric quantile estimates. The method is simple to implement, does not rely on the evaluation of quantities which appear in asymptotic distributions, and takes the bias properly into account (at least asymptotically). Additionally, we show that the bootstrap distribution can approximate the true one (w.r.t. the $\|\cdot\|_\infty$ norm, details in Theorem 2.1) up to $n^{-2/5}$, which represents a significant improvement relative to $(\log n)^{-1}$, which is based on the asymptotic Gumbel distribution, as studied by Härdle and Song [22]. Previous research by Hahn [15] showed consistency of a bootstrap approximation to the cumulative distribution function (cdf) without assuming independence of the error and regressor terms. Ref. [23] showed bootstrap methods for median regression models based on a smoothed least-absolute-deviations (SLAD) estimate.

Let $(X_1, Y_1), (X_2, Y_2), \ldots, (X_n, Y_n)$ be a sequence of independent identically distributed bivariate random variables with joint pdf $f(x, y)$, joint cdf $F(x, y)$, conditional pdf $f(y|x), f(x|y)$, conditional cdf $F(y|x), F(x|y)$ for Y given X and X given Y respectively, and marginal pdf $f_X(x)$ for X, $f_Y(y)$ for Y. With some abuse of notation we use the letters f and F to denote different pdfs and cdfs respectively. The exact distribution will be clear from the context. At the first stage we assume that $x \in J^* = (a, b)$ for some $0 < a < b < 1$. Let $l(x)$ denote the p-quantile curve, i.e. $l(x) = F^{-1}_{Y|x}(p)$.

In economics, discrete or categorial regressors are very common. An example is from labor market analysis where one tries to find out how revenues depend on the age of the employee w.r.t. different education levels, labor union statuses, genders and nationalities, i.e. in econometric analysis one targets the differential effects. For example, [4] examined the US wage structure by quantile regression techniques. This motivates the extension to multivariate covariables by partial linear modelling (PLM). This is convenient especially when we have categorial elements of the X vector. Partial linear models, which were first considered by Green and Yandell [14,8,34,32], are gradually developing into a class of commonly used and studied semiparametric regression models, which can retain the flexibility of nonparametric models and ease the interpretation of linear regression models while avoiding the "curse of dimensionality". Recently [29] used penalized quantile regression for variable selection of partially linear models with measurement errors.

In this paper, we propose an extension of the quantile regression model to $x = (u, v)^\top \in \mathbb{R}^d$ with $u \in \mathbb{R}^{d-1}$ and $v \in J^* \subset \mathbb{R}$. The quantile regression curve we consider is $\tilde{l}(x) = F^{-1}_{Y|x}(p) = u^\top\beta + l(v)$. The multivariate confidence band can then be constructed, based on the univariate uniform confidence band, plus the estimated linear part which we will prove is more accurately ($\sqrt{n}$ consistency) estimated. This makes various tasks in economics, e.g. labor market differential effect investigation, multivariate model specification tests and the investigation of the distribution of income and wealth across regions or countries or the distribution across households possible. Additionally, since the natural link between quantile and expectile regression was developed by Newey and Powell [30], we can further extend our result into expectile regression for various tasks, e.g. demography risk research or expectile-based Value at Risk (EVAR) as in [28]. For high-dimensional modelling, [2] recently investigated high-dimensional sparse models with L_1 penalty. Additionally, our result might also be further extended to intersection bounds (one side confidence bands), which is similar to the work of Chernozhukov et al. [6].

The rest of this article is organized as follows. To keep the main idea transparent, in Section 2, as an introduction to the more complicated situation, the bootstrap approximation rate for the (univariate) confidence band is presented through a coupling argument. An extension to multivariate covariance X with partial linear modelling is shown in Section 3 with the actual type of confidence bands and their properties. In Section 4, we compare via a Monte Carlo study the bootstrap uniform confidence band with the one based on the asymptotic theory and investigate the behavior of partial linear estimates with the corresponding confidence band. In Section 5, an application considers the labor market differential effect. The discussion is restricted to the semiparametric extension. We do not discuss the general nonparametric regression. We conjecture that this extension is possible under appropriate conditions. Section 6 contains concluding remarks. All proofs are sketched in the Appendix.

2. Bootstrap confidence bands in the univariate case

Suppose $Y_i = l(X_i) + \varepsilon_i$, $i = 1, \ldots, n$, where ε_i has the (conditional) distribution function $F(\cdot|X_i)$. For simplicity, but without any loss of generality, we assume that $F(0|X_i) = p$. $F(\xi|x)$ is smooth as a function of x and ξ for any x, and for any ξ in the neighborhood of 0. We assume:

(A1) $X_1, \ldots, X_n$ are an i.i.d. sample, and $\inf_x f_X(x) = \lambda_0 > 0$. The quantile function satisfies $\sup_x |l^{(j)}(x)| \le \lambda_j < \infty$, $j = 1, 2$.

(A2) The distribution of Y given X has a density and $\inf_{x,t} f(t|x) \ge \lambda_3 > 0$, continuous at all $x \in J^*$, and at t only in a neighborhood of 0. More exactly, we have the following Taylor expansion at $x' = x$, $t = 0$, for some $A(\cdot)$ and $f_0(\cdot)$:

$$F(t|x') = F(0|x) + \left.\frac{\partial F(t|x')}{\partial x'}\right|_{x'=x,t=0} t + \left.\frac{\partial F(t|x')}{\partial t}\right|_{x'=x,t=0} (x' - x) + R(t, x'; x)$$
$$\stackrel{\text{def}}{=} p + f_0(x)t + A(x)(x' - x) + R(t, x'; x), \tag{1}$$

where

$$\sup_{t,x,x'} \frac{|R(t, x'; x)|}{t^2 + |x' - x|^2} < \infty.$$

Let K be a symmetric density function with compact support and $d_K = \int u^2 K(u)du < \infty$. Let $l_h(\cdot) = l_{n,h}(\cdot)$ be the nonparametric p-quantile estimate of $Y_1, \ldots, Y_n$ with weight function $K\{(X_i - \cdot)/h\}$ for some global bandwidth $h = h_n$ ($K_h(u) = h^{-1}K(u/h)$), that is, a solution of

$$\frac{\sum_{i=1}^n K_h(x - X_i)\mathbf{1}\{Y_i < l_h(x)\}}{\sum_{i=1}^n K_h(x - X_i)} < p \le \frac{\sum_{i=1}^n K_h(x - X_i)\mathbf{1}\{Y_i \le l_h(x)\}}{\sum_{i=1}^n K_h(x - X_i)}. \tag{2}$$

Generally, the bandwidth may also depend on x. A local (adaptive) bandwidth selection though deserves future research.

Note that by assumption (A1), $l_h(x)$ is the quantile of a discrete distribution, which is equivalent to a sample of size $\mathcal{O}_p(nh)$ from a distribution with p-quantile whose bias is $\mathcal{O}(h^2)$ relative to the true value. Let δ_n be the local rate of convergence of the function l_h, essentially $\delta_n = h^2 + (nh)^{-1/2} = \mathcal{O}(n^{-2/5})$ with optimal bandwidth choice $h = h_n = \mathcal{O}(n^{-1/5})$ as in [36]. We employ also an auxiliary estimate $l_g \stackrel{\text{def}}{=} l_{n,g}$, essentially one similar to $l_{n,h}$ but with a slightly larger bandwidth $g = g_n = h_n n^{\zeta}$ (a heuristic explanation of why it is essential to oversmooth g is given later), where ζ is some small number. The asymptotically optimal choice of ζ as shown later is 4/45.

(A3) The estimate l_g satisfies

$$\sup_{x \in J^*} |l_g''(x) - l''(x)| = \mathcal{O}_p(1),$$
$$\sup_{x \in J^*} |l_g'(x) - l'(x)| = \mathcal{O}_p(\delta_n/h). \tag{3}$$

Assumption (A3) is only stated to overwrite the issue here. It actually follows from the assumptions on (g, h). A sequence $\{a_n\}$ is slowly varying if $n^{-\alpha}a_n \to 0$ for any $\alpha > 0$. With some abuse of notation we will use S_n to denote *any* slowly varying function which may change from place to place, e.g. $S_n^2 = S_n$ is a valid expression (since if S_n is a slowly varying function, then S_n^2 is slowly varying as well). λ_i and C_i are generic constants throughout this paper and the subscripts have no specific meaning. Note that there is no S_n term in (3) exactly because the bandwidth g_n used to calculate l_g is slightly larger than that used for l_h. We want to smooth it such that l_g, as an estimate of the quantile function, has a slightly worse rate of convergence, but its derivatives converge faster.

We also consider a family of estimates $\hat{F}(\cdot|X_i)$, $i = 1, \ldots, n$, estimating respectively $F(\cdot|X_i)$ and satisfying $\hat{F}(0|X_i) = p$. For example we can take the distribution with a point mass $[\sum_{j=1}^n K\{\alpha_n(X_j - X_i)\}]^{-1}K\{(X_j - X_i)/h\}$ on $Y_j - l_h(X_i)$, $j = 1, \ldots, n$, i.e.

$$\hat{F}(\cdot|X_i) = \frac{\sum_{j=1}^n K_h(X_j - X_i)\mathbf{1}\{Y_j - l_h(X_i) \le \cdot\}}{\sum_{j=1}^n K_h(X_j - X_i)}. \tag{4}$$

We additionally assume:

(A4) $f_X(x)$ is twice continuously differentiable and $f(t|x)$ is continuous in x, Hölder-continuous in t and uniformly bounded in x and t by, say, λ_4.

For the precision of $\hat{F}(\cdot|X_i)$'s approximation around 0, we employ the following lemma from Franke and Mwita [13]:

Lemma 2.1 (*[13, Lemma A.3-5]*). *If assumptions* (A1, A2, A4) *hold, then for* $|t| < S_n\delta_n$, $\delta_n \to 0$, $i = 1, \ldots, n$, $X_i \in J^*$,

$$\sup_{|t| < S_n\delta_n, i=1,\ldots,n, X_i \in J^*} |\hat{F}(t|X_i) - F(t|X_i)| = \mathcal{O}_p\{S_n\delta_n\}. \tag{5}$$

Let $F^{-1}(\cdot|\cdot)$ and $\hat{F}^{-1}(\cdot|\cdot)$ be the inverse function of the conditional cdf and its estimate. We consider the following bootstrap procedure. Let $U_1, \ldots, U_n$ be i.i.d. uniform [0, 1] variables. Let

$$Y_i^* = l_g(X_i) + \hat{F}^{-1}(U_i|X_i), \quad i = 1, \ldots, n \tag{6}$$

be the bootstrap sample. We couple this sample to an unobserved hypothetical sample from the true conditional distribution

$$Y_i^\# = l(X_i) + F^{-1}(U_i|X_i), \quad i = 1, \ldots, n. \tag{7}$$

Note that the vectors $(Y_1, \ldots, Y_n)$ and $(Y_1^\#, \ldots, Y_n^\#)$ are equally distributed given $X_1, \ldots, X_n$. We are really interested in the exact values of $Y_i^\#$ and Y_i^* only when they are near the appropriate quantile, that is, only if $|U_i - p| < S_n\delta_n$. But then, by Eq. (1), Lemma 2.1 and the inverse function theorem, we have

$$\max_{i:|F^{-1}(U_i|X_i)-F^{-1}(p)|<S_n\delta_n} |F^{-1}(U_i|X_i) - \widehat{F}^{-1}(U_i|X_i)| = \max_{i:|Y_i^\#-l(X_i)|<S_n\delta_n} |Y_i^\# - l(X_i) - Y_i^* + l_g(X_i)| = \mathcal{O}_p\{S_n\delta_n\}. \tag{8}$$

Let now $q_{hi}(Y_1, \ldots, Y_n)$ be the solution of the local quantile as given by (2) at X_i, with bandwidth h, i.e. $q_{hi}(Y_1, \ldots, Y_n) \stackrel{\text{def}}{=} l_h(X_i)$ for data set $\{(X_i, Y_i)\}_{i=1}^n$. Note that by (3), if $|X_i - X_j| = \mathcal{O}(h)$, then

$$\max_{|X_i-X_j|<ch} |l_g(X_i) - l_g(X_j) - l(X_i) + l(X_j)| = \mathcal{O}_p(\delta_n). \tag{9}$$

Let l_h^* and $l_h^\#$ be the local bootstrap quantile and its coupled sample analogue. Then

$$\begin{aligned} l_h^*(X_i) - l_g(X_i) &= q_{hi}[\{Y_j^* - l_g(X_i)\}_{j=1}^n] \\ &= q_{hi}[\{Y_j^* - l_g(X_j) + l_g(X_j) - l_g(X_i)\}_{j=1}^n], \end{aligned} \tag{10}$$

while

$$l_h^\#(X_i) - l(X_i) = q_{hi}[\{Y_j^\# - l(X_j) + l(X_j) - l(X_i)\}_{j=1}^n]. \tag{11}$$

From (8)–(11) we conclude that

$$\max_i |l_h^*(X_i) - l_g(X_i) - l_h^\#(X_i) + l(X_i)| = \mathcal{O}_p(\delta_n). \tag{12}$$

Based on (12), we obtain the following theorem (the proof is given in the Appendix):

Theorem 2.1. *If assumptions* (A1–A4) *hold, then*

$$\sup_{x\in J^*} |l_h^*(x) - l_g(x) - l_h^\#(x) + l(x)| = \mathcal{O}_p(\delta_n) = \mathcal{O}_p(n^{-2/5}).$$

Remark. Theorem 2.1 indicates that the r.v. $l_h^*(x) - l_g(x)$ approximates the one of $l_h^*(x)$ up to $n^{-2/5}$ (w.r.t. the $\|\cdot\|_\infty$ norm). Thus a number of replications of $l_h^*(x)$ can be used as the basis for simultaneous error bars.

Although Theorem 2.1 is stated with a fixed bandwidth, in practice, to take care of the heteroscedasticity effect, we construct confidence bands with the width depending on the densities, which is motivated by the counterpart based on the asymptotic theory as in [22]. Thus we have the following corollary.

Corollary 2.1. *Let* $d*_\alpha$ *be defined by* $P^*(|l_h^*(x) - l_g(x)| > d_\alpha^*) = \alpha$, *where* P^* *is the bootstrap distribution conditioned on the sample. If* (A1)–(A4) *hold, then the confidence interval* $l_h(x) \pm d_\alpha^*$ *has an asymptotic uniform coverage of* $1 - \alpha$, *in the sense that* $P(\sup_{x\in J^*} |l_h(x) - l(x)| > d_\alpha^*) \to \alpha$.

In practice we would use the approximate $(1 - \alpha) \times 100\%$ confidence band over $\mathbb{R}$ given by

$$l_h(x) \pm \left[\hat{f}\{l_h(x)|x\}\sqrt{\hat{f}_X(x)}\right]^{-1} d_\alpha^*, \tag{13}$$

where d_α^* is based on the bootstrap sample (defined later) and $\hat{f}\{l_h(x)|x\}$, $\hat{f}_X(x)$ are consistent estimators of $f\{l(x)|x\}$, $f_X(x)$ with use of $f(y|x) = f(x, y)/f_X(x)$.

Below is the summary of the basic steps for the bootstrap procedure.

(1) Given $(X_i, Y_i), i = 1, \ldots, n$, compute the local quantile smoother $l_h(x)$ of $Y_1, \ldots, Y_n$ with bandwidth h and obtain residuals $\hat{\varepsilon}_i = Y_i - l_h(X_i)$, $i = 1, \ldots, n$.

(2) Compute the conditional edf:

$$\hat{F}(t|x) = \frac{\sum_{i=1}^{n} K_h(x - X_i)\mathbf{1}\{\hat{\varepsilon}_i \leqslant t\}}{\sum_{i=1}^{n} K_h(x - X_i)}.$$

(3) For each $i = 1, \ldots, n$, generate random variables $\varepsilon^*_{i,b} \sim \hat{F}(t|X_i)$, $b = 1, \ldots, B$ and construct the bootstrap sample $Y^*_{i,b}$, $i = 1, \ldots, n$, $b = 1, \ldots, B$ as follows:

$$Y^*_{i,b} = l_g(X_i) + \varepsilon^*_{i,b}.$$

(4) For each bootstrap sample $\{(X_i, Y^*_{i,b})\}_{i=1}^n$, compute l^*_h and the random variable

$$d_b \stackrel{\text{def}}{=} \sup_{x \in J^*} \left[\hat{f}\{l^*_h(x)|x\} \sqrt{\hat{f}_X(x)} |l^*_h(x) - l_g(x)| \right] \tag{14}$$

where $\hat{f}\{l(x)|x\}$, $\hat{f}_X(x)$ are consistent estimators of $f\{l(x)|x\}$, $f_X(x)$.

(5) Calculate the $(1 - \alpha)$ quantile d^*_α of $d_1, \ldots, d_B$.

(6) Construct the bootstrap uniform confidence band centered around $l_h(x)$, i.e. $l_h(x) \pm \left[\hat{f}\{l_h(x)|x\} \sqrt{\hat{f}_X(x)} \right]^{-1} d^*_\alpha$.

While bootstrap methods are well-known tools for assessing variability, more care must be taken to properly account for the type of bias encountered in nonparametric curve estimation. The choice of bandwidth is crucial here. In our experience the bootstrap works well with a rather crude choice of g; one may, however, specify g more precisely. Since the main role of the pilot bandwidth is to provide a correct adjustment for the bias, we use the goal of bias estimation as a criterion. Recall that the bias in the estimation of $l(x)$ by $l^{\#}_h(x)$ is given by

$$b_h(x) = \mathsf{E} l^{\#}_h(x) - l(x).$$

The bootstrap bias of the estimate constructed from the resampled data is

$$\hat{b}_{h,g}(x) = \mathsf{E} l^*_h(x) - l_g(x). \tag{15}$$

Note that in (15) the expected value is computed under the bootstrap estimation. The following theorem gives an asymptotic representation of the mean squared error for the problem of estimating $b_h(x)$ by $\hat{b}_{h,g}(x)$. It is then straightforward to find g to minimize this representation. Such a choice of g will make the quantiles of the original and coupled bootstrap distributions close to each other. In addition to the technical assumptions before, we also need:

(A5) l and f are four times continuously differentiable.
(A6) K is twice continuously differentiable.

Theorem 2.2. *Under assumptions (*A1–A6*), for any $x \in J^*$*

$$\mathsf{E}\left[\left\{ \hat{b}_{h,g}(x) - b_h(x) \right\}^2 \big| X_1, \ldots, X_n \right] \sim h^4(C_1 g^4 + C_2 n^{-1} g^{-5}) \tag{16}$$

in the sense that the ratio between the RHS and the LHS tends in probability to 1 *for some constants C_1, C_2.*

An immediate consequence of Theorem 2.2 is that the rate of convergence of g should be $n^{-1/9}$, see also [20]. This makes precise the previous intuition which indicated that g should slightly oversmooth. Under our assumptions, reasonable choices of h will be of the order $n^{-1/5}$ as in [36]. Hence, (16) shows once again that g should tend to zero more slowly than h. Note that Theorem 2.2 is not stated uniformly over h. The reason is that we are only trying to give some indication of how the pilot bandwidth g should be selected.

We summarize how to select the bandwidth h for the local quantile smoother and g for the oversmoothed estimate as below.

1 Select h as in [36] which is also quoted below.
   - Use ready-made and sophisticated methods to select h_{mean}, the optimal bandwidth choice for regresion mean estimation; we use the technique of Ruppert et al. [33].
   - Use $h = h_{\text{mean}}\{p(l-p)/\phi(\Phi^{-1}(p))^2\}^{1/5}$ to obtain all other h's (w.r.t. different p's) from h_{mean}. ϕ and Ψ are the PDF and CDF of standard normal distributions respectively.
2 According to Theorem 2.2, select g as $g = n^{4/45}h$.

3. Bootstrap confidence bands in PLMs

The case of multivariate regressors may be handled via a semiparametric specification of the quantile regression curve. More specifically we assume that with $x = (u, v)^\top \in \mathbb{R}^d,\ v \in \mathbb{R}$:

$$\tilde{l}(x) = u^\top \beta + l(v).$$

In this section we show how to proceed in this multivariate setting and how — based on Theorem 2.1 — a multivariate confidence band may be constructed. We first describe the numerical procedure for obtaining estimates of β and l, where l denotes — as in the earlier sections — the one-dimensional conditional quantile curve. We then move on to the theoretical properties. First note that the PLM quantile estimation problem can be seen as estimating (β, l) in

$$\begin{aligned} y &= u^\top \beta + l(v) + \varepsilon \\ &= \tilde{l}(x) + \varepsilon \end{aligned} \tag{17}$$

where the p-quantile of ε conditional on both u and v is 0.

In order to estimate β, let a_n denote an increasing sequence of positive integers and set $b_n = a_n^{-1}$. For each $n = 1, 2, \ldots,$ partition the unit interval $[0, 1]$ for v in a_n intervals $I_{ni}, i = 1, \ldots, a_n$, of equal length b_n and let m_{ni} denote the midpoint of I_{ni}. In each of these small intervals $I_{ni}, i = 1, \ldots, a_n$, $l(v)$ can be considered as being approximately constant, and hence (17) can be considered as a linear model. This observation motivates the following two stage estimation procedure.

(1) A linear quantile regression inside each partition is used to estimate $\hat{\beta}_i, i = 1, \ldots, a_n$. Their weighted mean yields $\hat{\beta}$. More exactly, consider the parametric quantile regression of y on u, $\mathbf{1}\big(v \in [0, b_n)\big), \mathbf{1}\big(v \in [b_n, 2b_n)\big), \ldots, \mathbf{1}\big(v \in [1 - b_n, 1]\big)$. That is, let

$$\psi(t) \stackrel{\text{def}}{=} (p-1)t\mathbf{1}(t < 0) + pt\mathbf{1}(t \geqslant 0).$$

Then let

$$\hat{\beta} = \arg\min_{\beta} \min_{l_1, \ldots, l_{a_n}} \sum_{i=1}^{n} \psi \left\{ Y_i - \beta^{\mathsf{T}} U_i - \sum_{j=1}^{a_n} l_j \mathbf{1}\big(V_i \in I_{ni}\big) \right\}.$$

(2) Calculate the smooth quantile estimate as in (2) from $(V_i, Y_i - U_i^\top \hat{\beta})_{i=1}^n$, and name it as $\tilde{\tilde{l}}_h(v)$.

The following theorem states the asymptotic distribution of $\hat{\beta}$.

Theorem 3.1. *If assumption* (A1) *holds, for the above two stage estimation procedure, there exist positive definite matrices D, C, such that*

$$\sqrt{n}(\hat{\beta} - \beta) \xrightarrow{\mathcal{L}} N\{0, p(1-p)D^{-1}CD^{-1}\} \quad \text{as } n \to \infty,$$

where $C = \operatorname{plim}_{n\to\infty} C_n$ and $D = \operatorname{plim}_{n\to\infty} D_n$ with $C_n = \frac{1}{n}\sum_{i=1}^n U_i^\top U_i$ and $D_n = \frac{1}{n}\sum_{j=1}^n f\{l(V_j)|V_i\} U_j^\top U_j$ respectively.

Note that $l(v), \tilde{l}_h(v)$ (quantile smoother based on $(v, y - u^\top\beta)$) and $\tilde{\tilde{l}}_h(v)$ can be treated as zeros (w.r.t. $\theta, \theta \in I$ where I is a possibly infinite, or possibly degenerate, interval in $\mathbb{R}$) of the functions

$$\widetilde{H}(\theta, v) \stackrel{\text{def}}{=} \int_{\mathbb{R}} f(v, \tilde{y})\psi(\tilde{y} - \theta)d\tilde{y}, \tag{18}$$

$$\widetilde{H}_n(\theta, v) \stackrel{\text{def}}{=} n^{-1} \sum_{i=1}^{n} K_h(v - V_i)\psi(\widetilde{Y}_i - \theta), \tag{19}$$

$$\widetilde{\widetilde{H}}_n(\theta, v) \stackrel{\text{def}}{=} n^{-1} \sum_{i=1}^{n} K_h(v - V_i)\psi(\widetilde{\widetilde{Y}}_i - \theta), \tag{20}$$

where

$$\widetilde{Y}_i \stackrel{\text{def}}{=} Y_i - U_i^\top \beta,$$
$$\widetilde{\widetilde{Y}}_i \stackrel{\text{def}}{=} Y_i - U_i^\top \hat{\beta} = Y_i - U_i^\top \beta + U_i^\top(\beta - \hat{\beta}) \stackrel{\text{def}}{=} \widetilde{Y}_i + Z_i.$$

From Theorem 3.1 we know that $\hat{\beta} - \beta = \mathcal{O}_p(1/\sqrt{n})$ and $\|Z_i\|_\infty = \mathcal{O}_p(1/\sqrt{n})$. Under the following assumption, which is satisfied by exponential and generalized hyperbolic distributions, also used in [18]:

(A7) The conditional densities $f(\cdot|\tilde{y}), \tilde{y} \in \mathbb{R}$, are uniformly local Lipschitz continuous of order $\tilde{\alpha}$ (ulL-$\tilde{\alpha}$) on J, uniformly in $\tilde{y} \in \mathbb{R}$, with $0 < \tilde{\alpha} \leqslant 1$, and $(nh)/\log n \to \infty$,

for some constant C_3 not depending on n, Lemma 2.1 in [22] shows a.s. as $n \to \infty$:

$$\sup_{\theta\in I}\sup_{v\in J^*}|\widetilde{H}_n(\theta, v) - \widetilde{H}(\theta, v)| \le C_3 \max\{(nh/\log n)^{-1/2}, h^{\tilde{\alpha}}\}.$$

Observing that $\sqrt{h/\log n} = \mathcal{O}(1)$, we then have

$$\sup_{\theta\in I}\sup_{v\in J^*}|\widetilde{\widetilde{H}}_n(\theta, v) - \widetilde{H}(\theta, v)| \le \sup_{\theta\in I}\sup_{v\in J^*}|\widetilde{H}_n(\theta, v) - \widetilde{H}(\theta, v)| + \underbrace{\sup_{\theta\in I}\sup_{v\in J^*}|\widetilde{H}_n(\theta, v) - \widetilde{\widetilde{H}}_n(\theta, v)|}_{\le \mathcal{O}_p(1/\sqrt{n})\sup_{v\in J}|n^{-1}\sum K_h|}$$

$$\le C_4 \max\{(nh/\log n)^{-1/2}, h^{\tilde{\alpha}}\} \tag{21}$$

for a constant C_4 which can be different from C_3. To show the uniform consistency of the quantile smoother, we shall reduce the problem of strong convergence of $\tilde{\tilde{l}}_h(v) - l(v)$, uniformly in v, to an application of the strong convergence of $\widetilde{\widetilde{H}}_n(\theta, v)$ to $\widetilde{H}(\theta, v)$, uniformly in v and θ. For our result on $\tilde{\tilde{l}}_h(\cdot)$, we shall also require

(A8) $\inf_{v\in J^*}\left|\int \psi\{y - l(v) + \varepsilon\}dF(y|v)\right| \geqslant \tilde{q}|\varepsilon|, \quad$ for $|\varepsilon| \leqslant \delta_1$,

where δ_1 and $\tilde{q}$ are some positive constants, see also [19]. This assumption is satisfied if a constant $\tilde{q}$ exists giving $f\{l(v)|v\} > \tilde{q}/p$, $x \in J$. Ref. [22] showed:

Lemma 3.1. *Under assumptions* (A7) *and* (A8) *, we have a.s. as* $n \to \infty$

$$\sup_{v\in J^*}|\tilde{\tilde{l}}_h(v) - l(v)| \le C_5 \max\{(nh/\log n)^{-1/2}, h^{\tilde{\alpha}}\} \tag{22}$$

with another constant C_5 *not depending on n. If we consider the bandwidth* $h = \mathcal{O}(n^{-1/5})$ *and then skip the slow varying function* $\log n$*, then* $(nh/\log n)^{-1/2} = \mathcal{O}(n^{-2/5}) < \mathcal{O}(n^{-1/5}) \leqslant h^{\tilde{\alpha}}$*,* (22) *can be further simplified to*

$$\sup_{v\in J^*}|\tilde{\tilde{l}}_h(v) - l(v)| \le C_5\{h^{\tilde{\alpha}}\}.$$

Since the proof is essentially the same as Theorem 2.1 of the above mentioned reference, it is omitted here.

The convergence rate for the parametric part $\mathcal{O}_p(n^{-1/2})$ (Theorem 3.1) is smaller than the bootstrap approximation error for the nonparametric part $\mathcal{O}_p(n^{-2/5})$ as shown in Theorem 2.1. This makes the construction of uniform confidence bands for multivariate $x \in \mathbb{R}^d$ with a partial linear model possible.

Proposition 3.1. *Under the assumptions* (A1)–(A8)*, an approximate* $(1 - \alpha) \times 100\%$ *confidence band over* $\mathbb{R}^{d-1} \times [0, 1]$ *is*

$$u^\top\hat{\beta} + \tilde{\tilde{l}}_h(v) \pm \left[\hat{f}\{\tilde{\tilde{l}}_h(x)|x\}\sqrt{\hat{f}_X(x)}\right]^{-1} d^*_\alpha,$$

where $\hat{f}\{\tilde{\tilde{l}}_h(x)|x\}$*,* $\hat{f}_X(x)$ *are consistent estimators of* $f\{l(x)|x\}$*,* $f_X(x)$*.*

Note that here we actually only require that the convergence rate of the parametric part, which is typically $\mathcal{O}_p(n^{-1/2})$, is smaller than the bootstrap approximation error for the nonparametric part $\mathcal{O}_p(n^{-2/5})$. This makes construction for the uniform confidence bands of more general semiparametric models possible instead of just the partial linear model shown here and similar results could be obtained easily.

4. A Monte Carlo study

This section is divided into two parts. First we concentrate on a univariate regressor variable x, check the validity of the bootstrap procedure together with settings in the specific example, and compare it with asymptotic uniform bands. Secondly we incorporate the partial linear model to handle the multivariate case of $x \in \mathbb{R}^d$.

Below is the summary of the simulation procedure.

(1) Simulate (X_i, Y_i), $i = 1, \ldots, n$ according to their joint pdf $f(x, y)$.
In order to compare with earlier results in the literature, we choose the joint pdf of bivariate data $\{(X_i, Y_i)\}_{i=1}^n$, $n = 1000$ as

$$f(x, y) = f_{y|x}(y - \sin x)\mathbf{1}(x \in [0, 1]), \tag{23}$$

where $f_{y|x}(x)$ is the pdf of $N(0, x)$ with an increasing heteroscedastic structure. Thus the theoretical quantile is $l(x) = \sin(x) + \sqrt{x}\Phi^{-1}(p)$. Based on this normality property, all the assumptions can be seen to be satisfied.

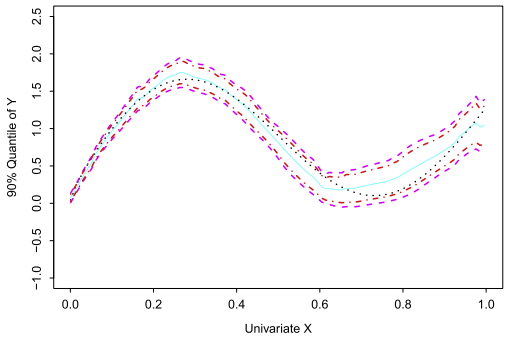


Fig. 1. The real 0.9 quantile curve (black dotted line), 0.9 quantile estimate (cyan solid line) with corresponding 95% uniform confidence band from asymptotic theory (magenta dashed lines) and confidence band from bootstrapping (red dashed–dot lines). (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

(2) Compute the local quantile smoother $l_h(x)$ of $Y_1, \ldots, Y_n$ with bandwidth h and obtain residuals $\hat{\varepsilon}_i = Y_i - l_h(X_i)$, $i = 1, \ldots, n$.
If we choose $p = 0.9$, then $\Phi^{-1}(p) = 1.2816$, $l(x) = \sin(x) + 1.2816\sqrt{x}$. Set $h = 0.05$.

(3) Compute the conditional edf:

$$\hat{F}(t|x) = \frac{\sum_{i=1}^{n} K_h(x - X_i)\mathbf{1}\{\hat{\varepsilon}_i \leqslant t\}}{\sum_{i=1}^{n} K_h(x - X_i)}.$$

The choice of kernel functions plays a minor role here. Section 3.4.3 and Table 3.3 of Härdle et al. [21] discuss the efficiencies of different kernels. The Epanechnikov kernel would be the optimal one; however, the differences among various kernels are small. Thus, we just use the Gaussian kernel to assure numerical stability. This is also convenient because the optimal bandwidth suggested by Yu and Jones [36] is also calculated based on the Gaussian kernel.

(4) For each $i = 1, \ldots, n$, generate random variables $\varepsilon^*_{i,b} \sim \hat{F}(t|x), b = 1, \ldots, B$ and construct the bootstrap sample $Y^*_{i,b}$, $i = 1, \ldots, n$, $b = 1, \ldots, B$ as follows:

$$Y^*_{i,b} = l_g(X_i) + \varepsilon^*_{i,b},$$

with $g = 0.2$.

(5) For each bootstrap sample $\{(X_i, Y^*_{i,b})\}_{i=1}^n$, compute l^*_h and the random variable

$$d_b \stackrel{\text{def}}{=} \sup_{x \in I^*} \left[\hat{f}\{l^*_h(x)|x\} \sqrt{\hat{f}_X(x)} |l^*_h(x) - l_g(x)| \right], \tag{24}$$

where $\hat{f}\{l(x)|x\}$, $\hat{f}_X(x)$ are consistent estimators of $f\{l(x)|x\}$, $f_X(x)$ with use of $f(y|x) = f(x, y)/f_X(x)$.

(6) Calculate the $(1 - \alpha)$ quantile d^*_α of $d_1, \ldots, d_B$.

(7) Construct the bootstrap uniform confidence band centered around $l_h(x)$, i.e. $l_h(x) \pm \left[\hat{f}\{l_h(x)|x\} \sqrt{\hat{f}_X(x)} \right]^{-1} d^*_\alpha$.

Fig. 1 shows the theoretical 0.9 quantile curve, 0.9 quantile estimate with corresponding 95% uniform confidence band from the asymptotic theory and the confidence band from the bootstrap. The real 0.9 quantile curve is marked as the black dotted line. We then compute the classic local quantile estimate $l_h(x)$ (cyan solid) with its corresponding 95% uniform confidence band (magenta dashed) based on asymptotic theory according to Härdle and Song [22]. The 95% confidence band from the bootstrap is displayed as red dashed–dot lines. At first sight, the quantile smoother, together with two corresponding bands, all capture the heteroscedastic structure quite well, and the width of the bootstrap confidence band is similar to the one based on asymptotic theory in [22]. Fig. 2 presents the bootstrap confidence bands constructed using different oversmoothing bandwidths w.r.t. the same (but different from the one used for Fig. 1) randomly generated data set, namely, 1/2, 1 and 2 times (from left to right) of the oversmoothing bandwidth $g = n^{4/45}h$ used before. As we can see, when we deviate from $g = n^{4/45}h$, the bootstrap confidence bands get wider.

We now extend x to the multivariate case and use a different quantile function to verify our method. Choose $x = (u, v)^\top \in \mathbb{R}^d$, $v \in \mathbb{R}$, and generate the data $\{(U_i, V_i, Y_i)\}_{i=1}^n$, $n = 1000$ with

$$y = 2u + v^2 + \varepsilon - 1.2816, \tag{25}$$

where u and v are uniformly distributed random variables in $[0, 2]$ and $[0, 1]$ respectively. ε has a standard normal distribution. The theoretical 0.9-quantile curve is $\tilde{l}(x) = 2u + v^2$. Since the choice of a_n is uncertain here, we test different

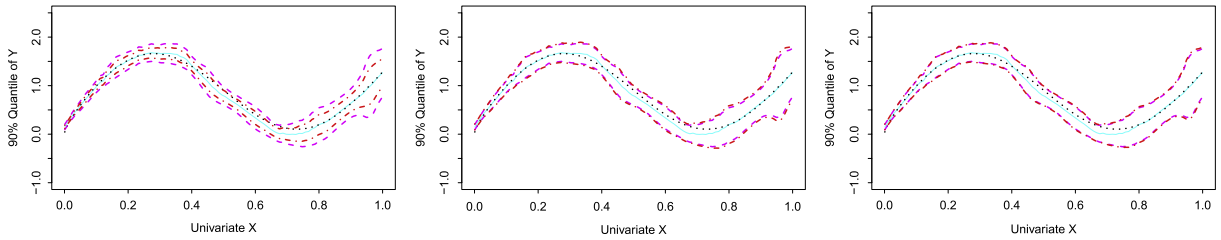


Fig. 2. The real 0.9 quantile curve (black dotted line), 0.9 quantile estimate (cyan solid line) with corresponding 95% uniform confidence band from asymptotic theory (magenta dashed lines) and confidence band from bootstrapping (red dashed–dot lines). The left, middle and right plots correspond to the oversmoothing bandwidth set as $n^{4/45}h/2$, $n^{4/45}h$ and $2n^{4/45}h$ respectively. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

Table 1
SSE of $\hat{\beta}$ with respect to a_n for different numbers of observations.

a_n	$n = 1000$	$n = 8000$	$n = 261148$
$n^{1/3}/8$			3.6×10^{-3}
$n^{1/3}/4$	5.4×10^{-1}	4.0×10^{-2}	3.3×10^{-3}
$n^{1/3}/2$	6.1×10^{-1}	3.5×10^{-2}	3.2×10^{-3}
$n^{1/3}$	6.2×10^{-1}	3.6×10^{-2}	3.1×10^{-3}
$n^{1/3} \cdot 2$	8.0×10^{-1}	3.9×10^{-2}	2.9×10^{-3}
$n^{1/3} \cdot 4$	4.9×10^{-1}	3.6×10^{-2}	2.8×10^{-3}
$n^{1/3} \cdot 8$			3.4×10^{-3}

choices of a_n for different n by simulation. To this end, we modify the theoretical model as follows:

$$y = 2u + v^2 + \varepsilon - \Phi^{-1}(p)$$

such that the real β is always equal to 2 no matter if p is 0.01 or 0.99. The result is displayed in Fig. 3 for $n = 1000$, $n = 8000$, $n = 261148$ (number of observations for the data set used in the following application part including both uncensored and censored observations). Different lines correspond to different a_n, i.e. $n^{1/3}/8$, $n^{1/3}/4$, $n^{1/3}/2$, $n^{1/3}$, $n^{1/3} \cdot 2$, $n^{1/3} \cdot 4$ and $n^{1/3} \cdot 8$. At first, it seems that the choice of a_n does not matter too much. To further investigate this, we calculate the SSE ($\sum_1^{99}\{\hat{\beta}(i/100) - \beta\}$) where $\hat{\beta}(i/100)$ denotes the estimate corresponding to the $i/100$ quantile. The results are displayed in Table 1. Obviously a_n has much less effect than n on SSE. Considering the computational cost, which increases with a_n, and the estimation performance, empirically we suggest $a_n = n^{1/3}$. Certainly this issue is far from settled and needs further investigation.

Thus for the specific model (25), we have $a_n = 10$, $\hat{\beta} = 1.997$, $h = 0.2$ and $g = 0.7$. In Fig. 4 the theoretical 0.9 quantile curve with respect to v, and the 0.9 quantile estimate with corresponding uniform confidence band are displayed. The real 0.9 quantile curve is marked as the black dotted line. We then compute the quantile smoother $l_h(x)$ (magenta solid). The 95% bootstrap uniform confidence band is displayed as red dashed lines and covers the true quantile curve quite well.

5. A labor market application

Our intuition of the effect of education on income is summarized by Day and Newburger's basic claim [7]: "At most ages, more education equates with higher earnings, and the payoff is most notable at the highest educational levels", which is actually from the point of view of mean regression. However, whether this difference is significant or not is still questionable, especially for different ends of the (conditional) income distribution. To this end, a careful investigation of quantile regression is necessary. Since different education levels may reflect different productivity, which is unobservable and may also results from different ages, abilities etc., to study the labor market differential effect with respect to different education levels, a semiparametric partial linear quantile model is preferred, which can retain the flexibility of the nonparametric models for the age and other unobservable factors and ease the interpretation of the education factor.

We use the administrative data from the German National Pension Office (Deutsche Rentenversicherung Bund) for the following group: West German part, males, born between 1939 and 1942 who began receiving a pension in 2004 or 2005 (when they were 62–66 years old) with at least 30 yearly uncensored observations. Since different people entered into the pension system and stopped receiving job earnings at different ages, we only consider those earnings recorded by the pension system when they were between 25 and 59 years old. For example, we consider person A's yearly earnings when he was 25–59 (entering into the pension system at 25), person B's when he was 27–59 (entering into the pension system at 27), and person C's when he was 30–59 (entering into the pension system at 30). In total, $n = 128429$ observations are available. We have the following three education categories: "low education", "apprenticeship" and "university" for the variable u (we assign them the numerical values 1, 2 and 3 respectively); the variable v is the age of the employee. "Low education" means without post-secondary education in Germany. "Apprenticeship" means part of Germany's dual education

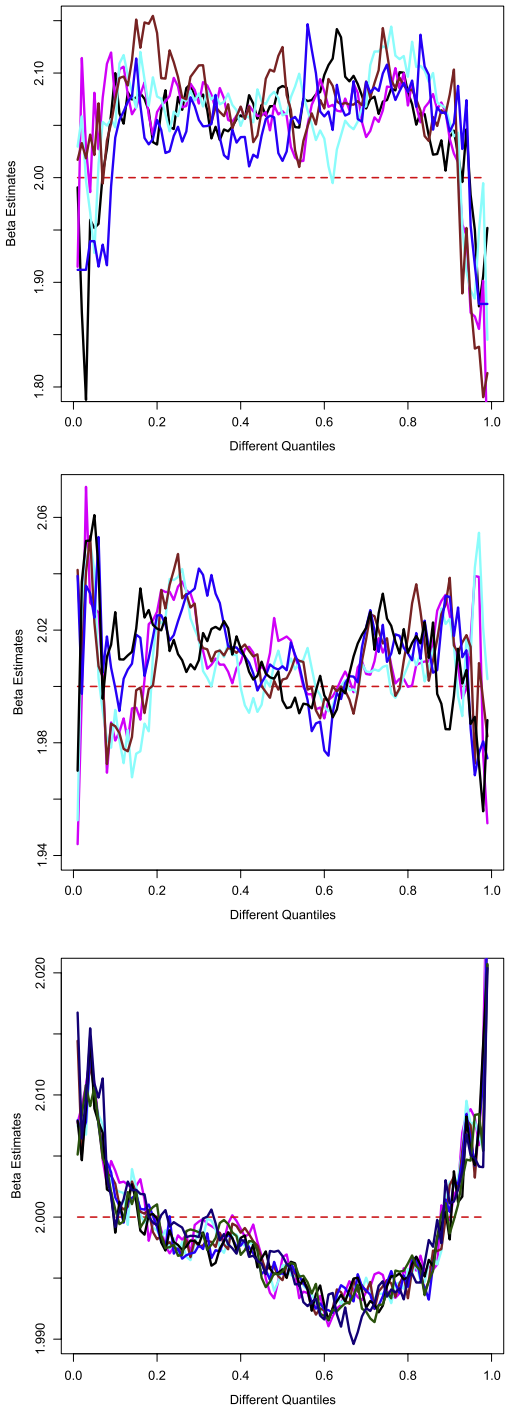


Fig. 3. $\hat{\beta}$ with respect to different quantiles for different numbers of observations, i.e. $n = 1000$ (top), $n = 8000$ (middle), $n = 261\,148$ (bottom). Different lines in the same plot correspond to different a_n, i.e. $n^{1/3}/8$, $n^{1/3}/4$, $n^{1/3}/2$, $n^{1/3}$, $n^{1/3} \cdot 2$, $n^{1/3} \cdot 4$ and $n^{1/3} \cdot 8$.

system. Depending on the profession, a person may work for three to four days a week in the company and then spend one or two days at a vocational school (Berufsschule). "University" in Germany also includes technical colleges (applied universities). Since the level and structure of wages differ substantially between East and West Germany, we concentrate on West Germany only here (which we usually refer to simply as Germany). Our data have several advantages over the most often used German Socio-Economics Panel (GSOEP) data for analyzing wages in Germany. Firstly, they are available for a much longer period, as opposed to from 1984 only for the GSOEP data. Secondly, and more importantly, they have a much larger sample size. Thirdly, wages are likely to be measured much more precisely. Fourthly, we observe a complete earnings

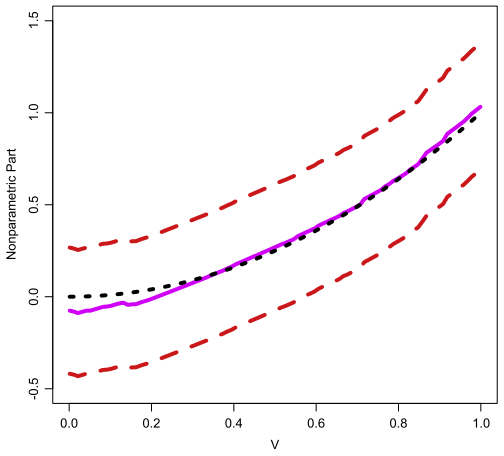


Fig. 4. Nonparametric part smoothing, real 0.9 quantile curve (black dotted line) with respect to v, 0.9 quantile smoother (magenta solid line) with corresponding 95% bootstrap uniform confidence band (red dashed lines). (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

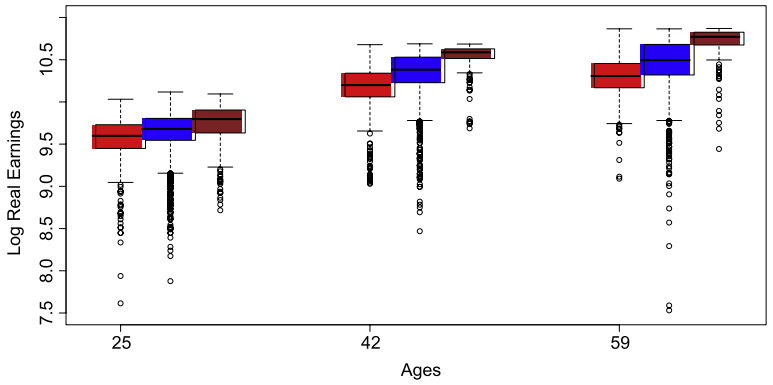


Fig. 5. Boxplots for "low education" (red), "apprenticeship" (blue) and "university" (brown) groups corresponding to different ages. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

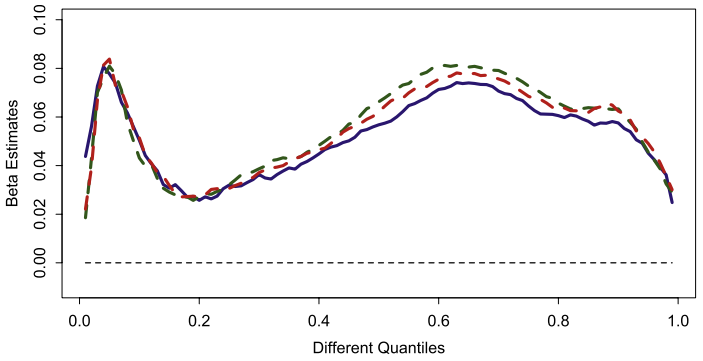


Fig. 6. $\hat{\beta}$ corresponding to different quantiles with 6, 13, 25 partitions.

history from the individual's first job until his retirement, therefore this is a true panel, not a pseudo-panel. There are also several drawbacks. For example, some very wealthy individuals are not registered in the German pension system, e.g. if their monthly income is more than some threshold (which may vary for different years due to the inflation effect), the individual has the right not to be included in the public pension system, and thus is not recorded. Besides this, it is also right-censored at the highest level of earnings that is subject to social security contributions, so the censored observations in the data are only for those who actually decided to stay within the public system. Because of the combination of truncation and censoring, this paper focuses on the uncensored data only, and we should not draw inferences from the very high quantile, i.e. we only consider the 0.80 quantiles here. Recently, similar data were also used to investigate the German wage structure as in [9].

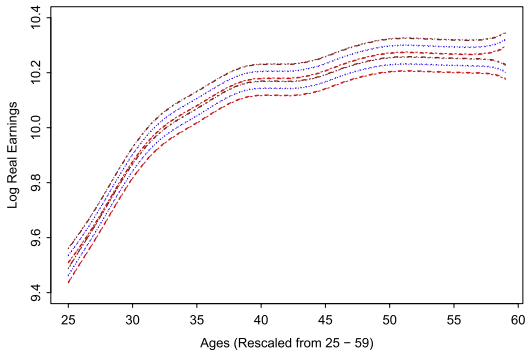


Fig. 7. 95% bootstrap (thick) and asymptotic (thin) uniform confidence bands for 0.20-quantile smoothers w.r.t. 3 different education levels. The "low education", "apprenticeship" and "university" levels are marked as red dashed, blue dotted and brown dashed–dot lines respectively. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

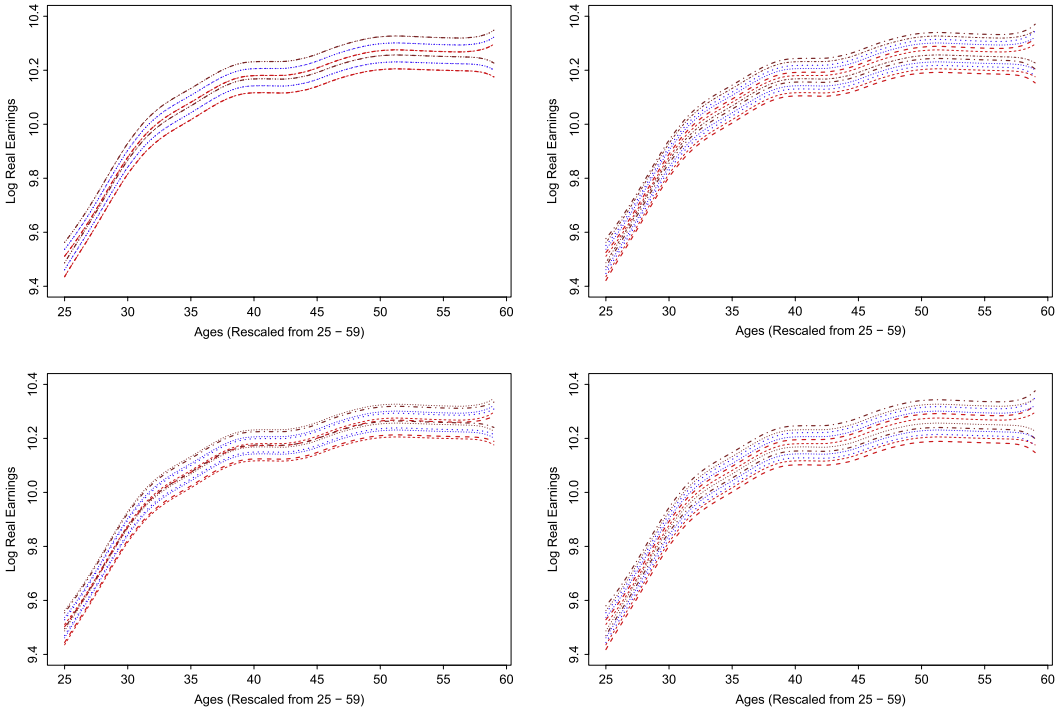


Fig. 8. 95% bootstrap (thick) and asymptotic (thin) uniform confidence bands for 0.20-quantile smoothers w.r.t. 3 different education levels with the oversmoothing bandwidth set as $g/2$, $g/4$, $2g$ and $4g$ (from left to right, up to down) respectively. The "low education", "apprenticeship" and "university" levels are marked as red dashed, blue dotted and brown dashed–dot lines respectively. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

Following from Becker's [1] human capital model, a log transformation is performed first on the hourly real wages (unit: EUR, at year 2000 prices). Fig. 5 displays the boxplots for the "low education", "apprenticeship" and "university" groups corresponding to different ages. In the data all ages (25–59) are reported as integers and are categorized in one-year groups. We rescaled them to the interval [0, 1] by dividing by 40, with corresponding bandwidths h of 0.041, 0.039, 0.041 for the 0.20, 0.50, 0.80 nonparametric quantile smoothers respectively. Correspondingly, as discussed before, we choose $g = n^{4/45}h$, thus 0.12, 0.11, 0.12 for the corresponding oversmoothers respectively. To detect whether a differential effect for different education levels exists, we compare the corresponding uniform confidence bands, i.e. differences indicate that the differential effect may exist for different education levels in the German labor market for that specific labor group.

Following an application of the partial linear model in Section 3, Fig. 6 displays $\hat{\beta}$ with respect to different quantiles for 6, 13, and 25 partitions, respectively. At first, the $\hat{\beta}$ curve is quite surprising, since it is not, as in mean regression, a positive constant, but rather varies a lot, e.g. $\hat{\beta}(0.20) = 0.026$, $\hat{\beta}(0.50) = 0.057$ and $\hat{\beta}(0.80) = 0.061$. Furthermore, it is robust to different numbers of partitions. It seems that the differences between the "low education" and "university" groups are different for different tails of the wage distribution. To judge whether these differences are significant, we use the uniform

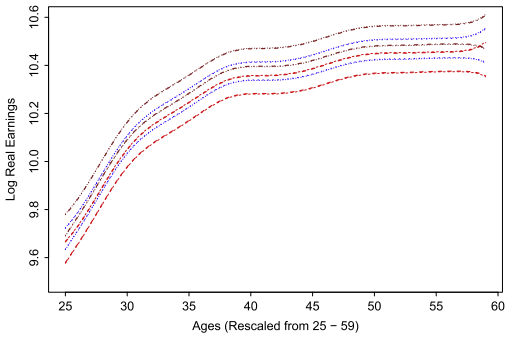


Fig. 9. 95% bootstrap (thick) and asymptotic (thin) uniform confidence bands for 0.50-quantile smoothers w.r.t. 3 different education levels. The "low education", "apprenticeship" and "university" levels are marked as red dashed, blue dotted and brown dashed–dot lines respectively. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

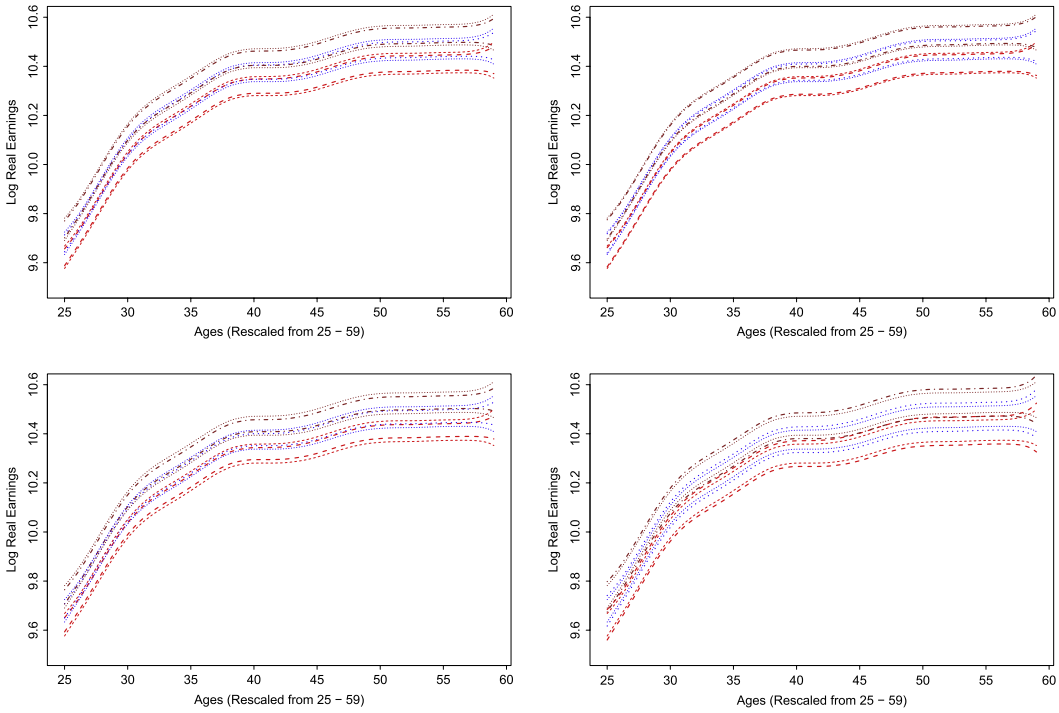


Fig. 10. 95% bootstrap (thick) and asymptotic (thin) uniform confidence bands for 0.50-quantile smoothers w.r.t. 3 different education levels with the oversmoothing bandwidth set as $g/2$, $g/4$, $2g$ and $4g$ (from left to right, up to down) respectively. The "low education", "apprenticeship" and "university" levels are marked as red dashed, blue dotted and brown dashed–dot lines respectively. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

confidence band techniques discussed in Section 2 which are displayed in Figs. 7–11 corresponding to the 0.20, 0.50 and 0.80 quantiles respectively.

The 95% uniform confidence bands from bootstrapping for the "low education" group are marked as red dashed lines, while the ones for "apprenticeship" and "university" are displayed as blue dotted and brown dashed–dot lines, respectively. The corresponding asymptotic bands studied in [22] are also added for reference (thin lines with the same style and color), which overlap with the bootstrap bands for large samples as here. For the 0.20 quantile in Fig. 7, the bands for "university", "apprenticeship" and "low education" do not differ significantly from one another although they become progressively lower, which indicates that high education does not equate to higher earnings significantly for the lower tails of wages, while increasing age seems to be the main driving force. For the 0.50 quantile in Fig. 9, the bands for "university" and "low education" differ significantly from one another although not from that for "apprenticeship". However, for the 0.80 quantiles in Fig. 11, all the bands differ significantly (except on the right boundary because of the nonparametric method's boundary effect) resulting from the relatively large $\hat{\beta}(0.80) = 0.061$, which indicates that high education is significantly associated with higher earnings for the upper tails of wages.

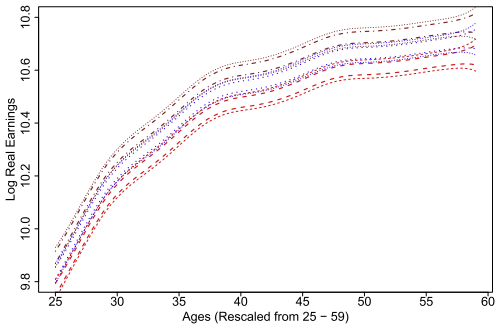


Fig. 11. 95% bootstrap (thick) and asymptotic (thin) uniform confidence bands for 0.80-quantile smoothers w.r.t. 3 different education levels. The "low education", "apprenticeship" and "university" levels are marked as red dashed, blue dotted and brown dashed–dot lines respectively. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

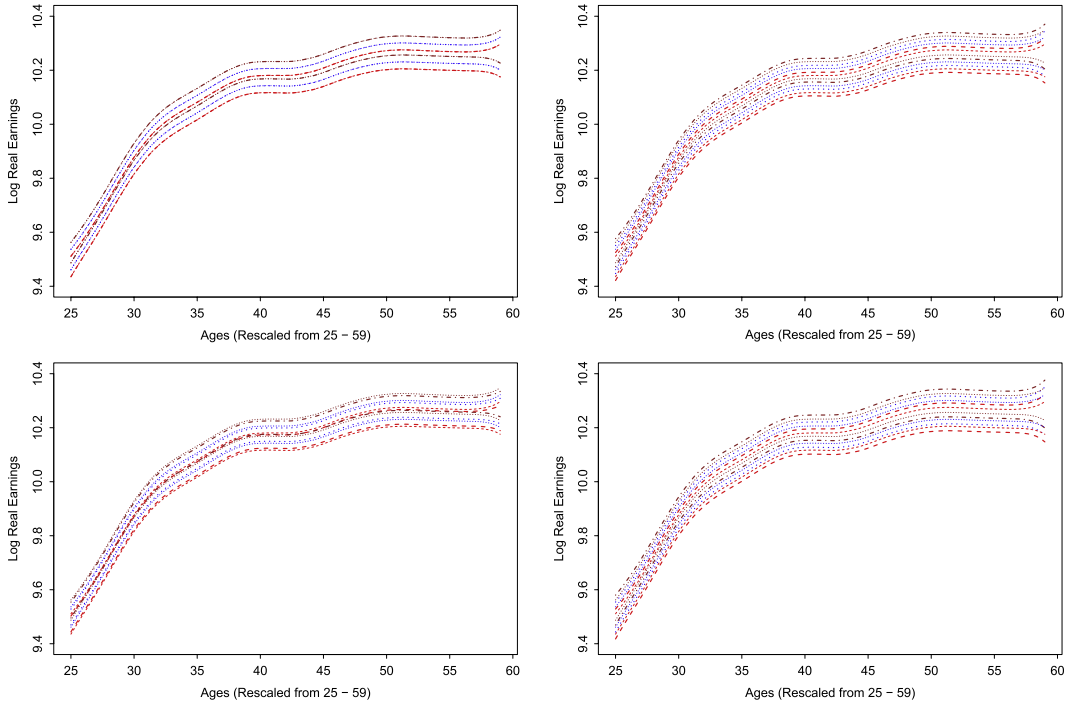


Fig. 12. 95% bootstrap (thick) and asymptotic (thin) uniform confidence bands for 0.80-quantile smoothers w.r.t. 3 different education levels with the oversmoothing bandwidth set as $g/2$, $g/4$, $2g$ and $4g$ (from left to right, up to down) respectively. The corresponding line styles and colors are the same as in Fig. 7. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

Coupled with Figs. 7, 9 and 11, Figs. 8, 10 and 12 present the corresponding bootstrap confidence bands constructed using different oversmoothing bandwidths, namely, half, quarter, twice and quadruple (from left to right, up to down) of the oversmoothing bandwidth $g = n^{4/45}h$ used before. The corresponding asymptotic bands are also added for reference (thin lines with the same style and color). As we can see, in practice, for the typically large labor economic data set, the bootstrap confidence bands are quite robust to the choice of the oversmoothing bandwidth.

If we investigate the explanations for the differences in different tails of the income distribution, maybe the most prominent reason is the rapid development of technology, which has been extensively studied. The point is that technology does not simply increase the demand for upper-end labor relative to that of lower-end labor, but instead asymmetrically affects the bottom and the top of the wage distribution, resulting in its strong asymmetry.

6. Conclusions

In this paper we construct confidence bands for nonparametric quantile estimates of regression functions. The method is based on bootstrapping, where resampling is done from a suitably estimated empirical distribution function (edf) for residuals. It is proven that the bootstrap approximation provides an improvement over the confidence bands constructed

via the asymptotic Gumbel distribution. We also propose a partial linear model to handle the case of multidimensional and discrete regressor variables. An economic application considering the labor market differential effect with respect to various education levels is studied. The conclusions from the point of view of quantile regression are consistent with those of the (grouped) mean regression, but in a more careful way in the sense that we provide formal statistical tools to judge these uniformly. The partial linear quantile regression techniques, together with confidence bands, developed in this paper display very interesting findings compared with classic (mean) methods and will bring in more contributions to the differential analysis of the labor market.

Appendix

Proof of Theorem 2.1. We start by proving Eq. (8). Write first $\hat{F}^{-1}(U_i|X_i) = F^{-1}(U_i|X_i) + \Delta_i$. Fix any i such that $|F^{-1}(U_i|X_i) - F^{-1}(p)| \leq S_n\delta_n$, which, by Eq. (1), implies that $|U_i - p| < S_n\delta_n$. Lemma 2.1 gives

$$\max_i |\hat{F}(S_n^2\delta_n|X_i) - F(S_n^2\delta_n|X_i)| = \mathcal{O}_p(S_n\delta_n). \tag{26}$$

Together with $F(\pm S_n^2\delta_n|X_i) = p \pm \mathcal{O}(S_n^2\delta_n)$, again by Eq. (1), we have $\hat{F}(\pm S_n^2\delta_n|X_i) = p \pm \mathcal{O}_p(S_n^2\delta_n)$ and thus

$$\begin{aligned}\hat{F}(-S_n^2\delta_n|X_i) = p - \mathcal{O}_p(S_n^2\delta_n) &\leqslant p - S_n\delta_n < U_i < p + S_n\delta_n \\ &< p + \mathcal{O}_p(S_n^2\delta_n) = \hat{F}(S_n^2\delta_n|X_i).\end{aligned}$$

Since $\hat{F}(\cdot|X_i)$ is monotone non-decreasing, $|\hat{F}^{-1}(U_i|X_i)| \leq S_n^2\delta_n$, which means, by $S_n^2 = S_n$,

$$|\hat{F}^{-1}(U_i|X_i)| \leq S_n\delta_n. \tag{27}$$

Apply now Lemma 2.1 again to Eq. (27), and obtain

$$\begin{aligned}S_n\delta_n &\geq |\hat{F}\{\hat{F}^{-1}(U_i|X_i)|X_i\} - F\{\hat{F}^{-1}(U_i|X_i)|X_i\}| \\ &= |U_i - F\{F^{-1}(U_i|X_i) + \Delta_i|X_i\}| \\ &= |F\{F^{-1}(U_i|X_i)|X_i\} - F\{F^{-1}(U_i|X_i) + \Delta_i|X_i\}| \\ &\geq f_0(X_i)|\Delta_i|.\end{aligned} \tag{28}$$

Hence $|\Delta_i| < S_n\delta_n$, and we summarize it as

$$\max_{i:|F^{-1}(U_i|X_i) - F^{-1}(p)| < S_n\delta_n} |F^{-1}(U_i|X_i) - \widehat{F}^{-1}(U_i|X_i)| = \mathcal{O}_p\{S_n\delta_n\}.$$

To show Eq. (12), define

$$\begin{aligned}Z_{1j} &\stackrel{\text{def}}{=} Y_j^* - l_g(X_j) + l_g(X_j) - l_g(X_i), \\ Z_{2j} &\stackrel{\text{def}}{=} Y_j^{\#} - l(X_j) + l(X_j) - l(X_i).\end{aligned}$$

Thus $q_{hi}[\{(Y_j^* - l_g(X_j) + l_g(X_j) - l_g(X_i))\}_{j=1}^n]$ and $q_{hi}[\{Y_j^{\#} - l(X_j) + l(X_j) - l(X_i)\}_{j=1}^n]$ can be seen as $l_h(X_i)$ for data sets $\{(X_i, Z_{1i})\}_{i=1}^n$ and $\{(X_i, Z_{2i})\}_{i=1}^n$ respectively. Similarly to Härdle and Song [22], they can be treated as zeros (w.r.t. $\theta, \theta \in I$ where I is a possibly infinite, or possibly degenerate, interval in $\mathbb{R}$) of the functions

$$\widetilde{G}_n(\theta, X_i) \stackrel{\text{def}}{=} n^{-1}\sum_{j=1}^n K_h(X_i - X_j)\psi(Z_{1j} - \theta), \tag{29}$$

$$\widetilde{\widetilde{G}}_n(\theta, X_i) \stackrel{\text{def}}{=} n^{-1}\sum_{j=1}^n K_h(X_i - X_j)\psi(Z_{2j} - \theta). \tag{30}$$

From (8) and (9), we have

$$\begin{aligned}&\max_i \left|[\{Y_j^* - l_g(X_j) + l_g(X_j) - l_g(X_i)\}_{j=1}^n] - [\{Y_j^{\#} - l(X_j) + l(X_j) - l(X_i)\}_{j=1}^n]\right| \\ &= \mathcal{O}_p\{S_n\delta_n\} + \mathcal{O}_p(\delta_n) = \mathcal{O}_p(\delta_n).\end{aligned} \tag{31}$$

Thus

$$\sup_{\theta\in I}\max_i |\widetilde{G}_n(\theta, X_i) - \widetilde{\widetilde{G}}_n(\theta, X_i)| \leq \mathcal{O}_p(\delta_n)\max\left|n^{-1}\sum K_h\right| = \mathcal{O}_p(\delta_n).$$

To show the difference of the two quantile smoothers, we shall reduce the strong convergence of $q_{hi}[\{Y_j^* - l_g(X_j) + l_g(X_j) - l_g(X_i)\}_{j=1}^n] - q_{hi}[\{Y_j^\# - l(X_j) + l(X_j) - l(X_i)\}_{j=1}^n]$, for any i, to an application of the strong convergence of $\widetilde{G}(\theta, X_i)$ to $\widetilde{\widetilde{G}}_n(\theta, X_i)$, uniformly in θ, for any i. Under assumptions (A7) and (A8), in a similar spirit to Härdle and Song [22], we get

$$\max_i |l_h^*(X_i) - l_g(X_i) - l_h^\#(X_i) - l(X_i)| = \mathcal{O}_p(\delta_n).$$

To show the supremum of the bootstrap approximation error, without loss of generality, based on assumption (A1), we reorder the original observations $\{X_i, Y_i\}_{i=1}^n$, such that $X_1 \leqslant X_2 \leqslant, \ldots, \leqslant X_n$. First decompose:

$$\begin{aligned}\sup_{x\in J^*} |l_h^*(x) - l_g(x) - l_h^\#(x) - l(x)| &= \max_i |l_h^*(X_i) - l_g(X_i) - l_h^\#(X_i) + l(X_i)| \\ &\quad + \max_i \sup_{x\in[X_i, X_{i+1}]} |l_h^*(x) - l_g(x) - l_h^\#(x) + l(x)|. \end{aligned} \tag{32}$$

From assumption (A1) we know $l'(\cdot) \leq \lambda_1$ and $\max_i(X_{i+1} - X_i) = \mathcal{O}_p(S_n/n)$. By the mean value theorem, we conclude that the second term of (32) is of a lower order than the first term. Together with Eq. (12) we have

$$\sup_{x\in J^*} |l_h^*(x) - l_g(x) - l_h^\#(x) - l(x)| = \mathcal{O}\{\max_i |l_h^*(X_i) - l_g(X_i) - l_h^\#(X_i) - l(X_i)|\} = \mathcal{O}_p(\delta_n),$$

which means that the supremum of the approximation error over all x is of the same order of the maximum over the discrete observed X_i. □

Proof of Theorem 2.2. The proof of (16) uses methods related to those in the proof of Theorem 3 of Härdle and Marron [20], so only the main steps are explicitly given. The first step is a bias-variance decomposition,

$$\mathsf{E}\left[\left\{\hat{b}_{h,g}(x) - b_h(x)\right\}^2 | X_1, \ldots, X_n\right] = \mathcal{V}_n + \mathcal{B}_n^2, \tag{33}$$

where

$$\begin{aligned}\mathcal{V}_n &= \mathsf{Var}\big[\hat{b}_{h,g}(x)|X_1, \ldots, X_n\big], \\ \mathcal{B}_n &= \mathsf{E}\big[\hat{b}_{h,g}(x) - b_h(x)|X_1, \ldots, X_n\big].\end{aligned}$$

Following the uniform Bahadur representation techniques for quantile regression as in Theorem 3.2 of Kong et al. [27], we have the following linear approximation for the quantile smoother as a local polynomial smoother corresponding to a specific loss function:

$$l_h^\#(x) - l(x) = L_n + \mathcal{O}_p(L_n),$$

where

$$L_n = \frac{n^{-1}\sum K_h(x - X_i)\psi\{Y_i - l(x)\}}{f\{l(x)|x\} f_X(x)}$$

for

$$\begin{aligned}\psi(u) &= p\mathbf{1}\{u \in (0, \infty)\} - (1-p)\mathbf{1}\{u \in (-\infty, 0)\} \\ &= p - \mathbf{1}\{u \in (-\infty, 0)\}, \\ l(x-t) - l(x) &= l'(x)(-t) + l''(x)t^2 + \mathcal{O}(t^2), \\ \{l(x-t) - l(x)\}' &= l''(x)(-t) + l'''(x)t^2 + \mathcal{O}(t^2), \\ f(x-t) &= f(x) + f'(x)(-t) + f''(x)(t^2) + \mathcal{O}(t^2), \\ f'(x-t) &= f'(x) + f''(x)(-t) + f'''(x)t^2 + \mathcal{O}(t^2), \\ \int K_h(t)tdt &= 0, \\ \int K_h(t)t^2dt &= h^2 d_K, \\ \int K_h(t)\mathcal{O}(t^2)dt &= \mathcal{O}(h^2).\end{aligned}$$

Then we have

$$\mathcal{B}_n = \mathcal{B}_{n1} + \mathcal{O}(\mathcal{B}_{n1}),$$

where

$$\mathcal{B}_{n1} = \frac{\int K_g(x-t)\mathcal{U}_h(t)dt - \mathcal{U}_h(x)}{f_X(x)f\{l(x)|x\}}$$

for

$$\begin{aligned}\mathcal{U}_h(x) &= \int K_h(x-s)\psi\{l(s)-l(x)\}f(s)ds \\ &= \int K_h(t)\psi\{l(x-t)-l(x)\}f(x-t)dt.\end{aligned}$$

By differentiation, a Taylor expansion and properties of the kernel K (see assumption (A2)),

$$\mathcal{U}_h'(x) = \int K_h(t)[\psi'\{l(x-t)-l(x)\}'f(x-t) + \psi\{l(x-t)-l(x)\}f'(x-t)]dt.$$

Here ψ' is the derivative of ψ except the 0 point, which actually does not matter since there is integration afterwards. Collecting terms, we get

$$\begin{aligned}\mathcal{U}_h'(x) &= \int K_h(t)\{\psi' l''(x)f_X'(x)t^2 + \psi' l''' f_X(x)t^2 + af'''(x)t^2 + \mathcal{O}(t^2)\}dt \\ &= \int K_h(t)\left\{C_0 t^2 + o(t^2)\right\}dt = h^2 d_K \cdot C_0 + \mathcal{O}(h^2),\end{aligned}$$

where a is a constant with $|a| < 1$ and $C_0 = \psi' l''(x)f_X'(x) + \psi' l''' f_X(x) + af'''(x)$.

Hence, by another substitution and Taylor expansion, for the first term in the numerator of $\mathcal{B}_{n1}$, we have

$$\mathcal{B}_{n2} = g^2h^2(d_K)^2 \cdot C_0 + \mathcal{O}(g^2h^2).$$

Thus, along almost all sample sequences,

$$\mathcal{B}_n^2 = C_1 g^4 h^4 + \mathcal{O}(g^4h^4) \tag{34}$$

for $C_1 = (d_K)^4 C_0^2/[f_X^2(x)f^2\{l(x)|x\}]$.

For the variance term, calculation in a similar spirit shows that

$$\mathcal{V}_n = \mathcal{V}_{n1} + \mathcal{O}(\mathcal{V}_{n1}),$$

where

$$\mathcal{V}_{n1} = \frac{\int K_g^2(x-t)\mathcal{W}_h(t)dt - \left\{\int K_g(x-t)\mathcal{U}_h(t)dt\right\}^2 f_X(x)f\{l(x)|x\}}{f_X(x)f\{l(x)|x\}}$$

for

$$\begin{aligned}\mathcal{W}_h(x) &= \int K_h^2(x-s)\psi\{l(s)-l(x)\}^2 f(s)ds \\ &= \int K_h^2(t)\psi\{l(x-t)-l(x)\}^2 f(x-t)dt.\end{aligned}$$

Hence, by Taylor expansion, collecting items and similar calculation, we have

$$\mathcal{V}_n = n^{-1}h^4g^{-5}C_2 + \mathcal{O}(n^{-1}h^4g^{-5}) \tag{35}$$

for a constant C_2. This, together with (33) and (34), completes the proof of Theorem 2.2. □

Proof of Theorem 3.1. In the case where the function l is known, the estimate $\hat{\beta}_I$ is

$$\hat{\beta}_I = \underset{\beta}{\operatorname{argmin}} \sum_{i=1}^{n} \psi\{Y_i - l(V_i) - U_i^\top \beta\}.$$

Since l is unknown, in each of these small intervals I_{ni}, $l(V_i)$ could be regarded as a constant $\alpha = l(m_{ni})$ for some i whose corresponding interval I_{ni} covers V_i. From assumption (A1), we know that $|l(V_i) - \alpha_i| \le \lambda_1 b_n < \infty$. If we define our first step estimate $\hat{\beta}_i$ inside each small interval as

$$(\hat{\alpha}_i, \hat{\beta}_i) = \underset{\alpha,\,\beta}{\arg\min} \sum \psi(Y_i - \alpha - U_i^\top \beta),$$

$|\{Y_i - l(V_i) - U_i^\top \beta\} - (Y_i - \alpha - U_i^\top \beta)| \le \lambda_1 b_n < \infty$ indicates that we could treat $\hat\beta_i$ as $\hat\beta_I$ inside each partition. If we use d_i to denote the number of observations inside partition I_{ni} (based on the i.i.d. assumption as in assumption (A1), on average $d_i = n/a_n$). For each of the $\hat\beta_i$'s inside interval I_{ni}, various parametric quantile regression works, e.g. the convex function rule in [31,24], yield

$$\sqrt{d_i}(\hat\beta_i - \beta) \xrightarrow{\mathcal{L}} \mathrm{N}\{0, p(1-p)D_i'^{-1}(p)C_i'D_i'^{-1}(p)\} \tag{36}$$

with the matrices $C_i' = d_i^{-1}\sum_{i=1}^{d_i} U_i^\top U_i$ and $D_i'(p) = d_i^{-1}\sum_{i=1}^{d_i} f\{l(V_i)|V_i\}U_i^\top U_i$.

To get $\hat\beta$, our second step is to take the weighted mean of $\hat\beta_1, \ldots, \hat\beta_{a_n}$ as

$$\hat\beta = \arg\min_\beta \sum_{i=1}^{a_n} d_i(\hat\beta_i - \beta)^2 = \sum_{i=1}^{a_n} d_i\hat\beta_i/n.$$

Note that under this construction, $\hat\beta_1, \ldots, \hat\beta_{a_n}$ are independent but not identical. Thus we intend to use the Lindeberg condition for the central limit theorem. To this end, we use s_n^2 to denote $\mathrm{Var}(\sum_{i=1}^{a_n} d_i\hat\beta_i/n)$, and we need to further check whether the following "Lindeberg condition" holds:

$$\lim_{a_n\to\infty} \frac{1}{s_n^2}\sum_{i=1}^{a_n}\int_{(|d_i\hat\beta_i/n-\beta|>\varepsilon s_n)} (\hat\beta_i - \beta)^2\, dF = 0, \quad \text{for all } \varepsilon > 0. \tag{37}$$

Since

$$\begin{aligned}
\mathrm{Var}\left\{\sum_{i=1}^{a_n} d_i(\hat\beta_i - \beta)/n\right\} &= \sum_i^{a_n} p(1-p)\left\{\left[n/d_i\sum_{j=1}^{d_i} f\{l(V_j)|v\}U_j^\top U_j\right]^{-1}\right. \\
&\quad \left.\times \sum_{i=1}^{d_i} U_i^\top U_i\left[n/d_i\sum_{j=1}^{d_i} f\{l(V_j)|v\}U_j^\top U_j\right]^{-1}\right\} \\
&\approx p(1-p)\left[\sum_{j=1}^{n} f\{l(V_j)|v\}U_j^\top U_j\right]^{-1}\sum_{i=1}^{n} U_i^\top U_i\left[\sum_{j=1}^{n} f\{l(V_j)|v\}U_j^\top U_j\right]^{-1} \\
&\stackrel{\mathrm{def}}{=} \frac{1}{n}p(1-p)D_n^{-1}C_nD_n^{-1},
\end{aligned}$$

where $D_n = \frac{1}{n}\sum_{j=1}^{n} f\{l(V_j)|V_i\}U_j^\top U_j$ and $C_n = \frac{1}{n}\sum_{i=1}^{n} U_i^\top U_i$, together with the normality of $\hat\beta_i$ as in (36) and properties of the tail of the normal distribution, e.g. Exe. 14.3–14.4 of Borak et al. [3], (37) follows.

Thus as n, $a_n \to \infty$ (although at a lower rate than n), together with $C = \mathrm{plim}_{n\to\infty} C_n$, $D = \mathrm{plim}_{n\to\infty} D_n$, we have

$$\sqrt{n}(\hat\beta - \beta) \xrightarrow{\mathcal{L}} \mathrm{N}\{0, p(1-p)D^{-1}CD^{-1}\}. \quad \square \tag{38}$$

References

[1] G. Becker, Human Capital: A Theoretical and Empirical Analysis with Special Reference to Education, third ed., The University of Chicago Press, 1994.
[2] A. Belloni, V. Chernozhukov, L_1-penalized quantile regression in high-dimensional sparse models, Annals of Statistics 39 (1) (2011) 82–130.
[3] S. Borak, W. Härdle, B. Lopez, Statistics of Financial Markets Exercises and Solutions, Springer-Verlag, Heidelberg, 2010.
[4] M. Buchinsky, Quantile regression, Box–Cox transformation model, and the US wage structure, 1963–1987, Journal of Econometrics 65 (1995) 109–154.
[5] Z.W. Cai, Regression quantiles for time series, Econometric Theory 18 (2002) 169–192.
[6] V. Chernozhukov, S. Lee, A.M. Rosen, Intersection bounds: estimation and inference, CEMMAP Working Paper, 19/09, 2009.
[7] J.C. Day, E.C. Newburger, The big payoff: educational attainment and synthetic estimates of work-life earnings. Special studies, current population reports, Statistical Report, US Department of Commerce Economics and Statistics Administration, US CENSUS BUREAU, 2002, pp. 23–210.
[8] L. Denby, Smooth regression functions. Statistical Report 26, AT&T Bell Laboratories, 1986.
[9] C. Dustmann, J. Ludsteck, U. Schönberg, Revisitng the German wage structure, The Quarterly Journal of Economics 124 (2) (2009) 843–881.
[10] J. Fan, T.C. Hu, Y.K. Troung, Robust nonparametric function estimation, Scandinavian Journal of Statistics 21 (1994) 433–446.
[11] J. Fan, Q. Yao, H. Tong, Estimation of conditional densities and sensitivity measures in nonlinear dynamical systems, Biometrika 83 (1996) 189–206.
[12] R.A. Fisher, L.H.C. Tippett', Limiting forms of the frequency distribution of the largest or smallest member of a sample, Proceedings of the Cambridge Philosophical Society 24 (1928) 180–190.
[13] J. Franke, P. Mwita, Nonparametric estimates for conditional quantiles of time series, Report in Wirtschaftsmathematik 87, University of Kaiserslautern, 2003.
[14] P.J. Green, B.S. Yandell, Semi-parametric generalized linear models, in: Proceedings 2nd International GLIM Conference, in: Lecture Notes in Statistics 32, vol. 32, Springer, New York, 1985, pp. 44–55.
[15] J. Hahn, Bootstrapping quantile regression estimators, Econometric Theory 11 (1) (1995) 105–121.
[16] P. Hall, On convergence rates of suprema, Probability Theory and Related Fields 89 (1991) 447–455.
[17] P. Hall, R. Wolff, Q. Yao, Methods for estimating a conditional distribution function, Journal of the American Statistical Association 94 (1999) 154–163.
[18] W. Härdle, P. Janssen, R. Serfling, Strong uniform consistency rates for estimators of conditional functionals, Annals of Statistics 16 (1988) 1428–1429.

[19] W. Härdle, S. Luckhaus, Uniform consistency of a class of regression function estimators, Annals of Statistics 12 (1984) 612–623.
[20] W. Härdle, J. Marron, Bootstrap simultaneous error bars for nonparametric regression, Annals of Statistics 19 (1991) 778–796.
[21] W. Härdle, M. Müller, S. Sperlich, A. Werwatz, Nonparametric and Semiparametric Models, Springer Verlag, Heidelberg, 2004.
[22] W. Härdle, S. Song, Confidence bands in quantile regression, Econometric Theory 26 (2010) 1–22. with corrections forthcoming or obtained from the authors.
[23] J.L. Horowitz, Bootstrap methods for median regression models, Econometrica 66 (6) (1998) 1327–1351.
[24] K. Knight, Comparing conditional quantile estimators: first and second order considerations, Unpublished Manuscript, 2001.
[25] R. Koenker, G.W. Bassett, Regression quantiles, Econometrica 46 (1978) 33–50.
[26] R. Koenker, K.F. Hallock, Quantile regression, The Journal of Economic Perspectives 15 (4) (2001) 143–156.
[27] E. Kong, O. Linton, Y. Xia, Uniform Bahadur representation for local polynomial estimates of M-regression and its application to the additive model, Econometric Theory 26 (2010) 1529–1564.
[28] C. Kuan, J. Yeh, Y. Hsu, Assessing value at risk with care, the conditional autoregressive expectile models, Journal of Econometrics 150 (2009) 261–270.
[29] H. Liang, R. Li, Variable selection for partially linear models with measurement errors, Journal of the American Statistical Association 104 (485) (2009) 234–248.
[30] W. Newey, J. Powell, Asymmetric least squares estimation and testing, Econometrica 55 (1987) 816–847.
[31] D. Pollard, Asymptotics for least absolute deviation estimators, Econometric Theory 7 (1991) 186–199.
[32] P.M. Robinson, Semiparametric econometrics: a survey, Journal of Applied Econometrics 3 (1988) 35–51.
[33] D. Ruppert, S.J. Sheather, M.P. Wand, An effective bandwidth selector for local least squares regression, Journal of the American Statistical Association 90 (432) (1995) 1257–1270.
[34] P.E. Speckman, Regression analysis for partially linear models, Journal of the Royal Statistical Society, Series B 50 (1988) 413–436.
[35] K. Yu, M.C. Jones, A comparison of local constant and local linear regression quantile estimation, Computational Statistics and Data Analysis 25 (1997) 159–166.
[36] K. Yu, M.C. Jones, Local linear quantile regression, Journal of the American Statistical Association 93 (1998) 228–237.

Journal of Multivariate Analysis 105 (2012) 164–175

Contents lists available at SciVerse ScienceDirect

Journal of Multivariate Analysis

journal homepage: www.elsevier.com/locate/jmva

Difference based ridge and Liu type estimators in semiparametric regression models☆

Esra Akdeniz Duran [a,1], Wolfgang Karl Härdle [b], Maria Osipenko [c,*]

[a] Department of Statistics, Gazi University, Turkey

[b] Center for Applied Statistics & Economics, Humboldt-Universität zu Berlin, Germany

[c] CASE, School of Business and Economics, Humboldt-Universiät zu Berlin, Unter den Linden 6, 10099, Germany

ARTICLE INFO

Article history:
Received 4 March 2011
Available online 7 September 2011

AMS subject classifications:
62G08
62J07

Keywords:
Difference based estimator
Differencing estimator
Differencing matrix
Liu estimator
Liu type estimator
Multicollinearity
Ridge regression estimator
Semiparametric model

ABSTRACT

We consider a difference based ridge regression estimator and a Liu type estimator of the regression parameters in the partial linear semiparametric regression model, $y = X\beta + f + \varepsilon$. Both estimators are analyzed and compared in the sense of mean-squared error. We consider the case of independent errors with equal variance and give conditions under which the proposed estimators are superior to the unbiased difference based estimation technique. We extend the results to account for heteroscedasticity and autocovariance in the error terms. Finally, we illustrate the performance of these estimators with an application to the determinants of electricity consumption in Germany.

1. Introduction

Semiparametric partial linear models have received considerable attention in statistics and econometrics. They have a wide range of applications, from biomedical studies to economics. In these models, some explanatory variables have a linear effect on the response while others are entering nonparametrically. Consider the semiparametric regression model:

$$y_i = x_i^\top \beta + f(t_i) + \varepsilon_i, \quad i = 1, \ldots, n \tag{1}$$

where y_i's are observations at t_i, $0 \le t_1 \le t_2 \le \cdots \le t_n \le 1$ and $x_i^\top = (x_{i1}, x_{i2}, \ldots, x_{ip})$ are known p-dimensional vectors with $p \le n$. In many applications, t_i's are values of an extra univariate "time" variable at which responses y_i are observed. In the case $t_i \in \mathbb{R}^k$, $t_i = (t_{1i}, \ldots, t_{ki})^\top$, the triples $(y_1, x_1, t_1), \ldots, (y_n, x_n, t_n)$ should be ordered using one of the algorithms mentioned in [30], Appendix A, or in [8, Section 2.2].

☆ This research was supported by Deutsche Forschungsgemeinschaft through the SFB 649 'Economic Risk'.

* Corresponding author.
E-mail addresses: esraakdeniz@gmail.com (E. Akdeniz Duran), maria.osipenko@wiwi.hu-berlin.de (M. Osipenko).

[1] Dr. Esra Akdeniz Duran was a research associate at Humboldt-Universität zu Berlin, Germany during this research.

0047-259X/$ – see front matter © 2011 Elsevier Inc. All rights reserved.
doi:10.1016/j.jmva.2011.08.018

In Eq. (1), $\beta = (\beta_1, \ldots, \beta_p)^\top$ is an unknown p-dimensional parameter vector, $f(\cdot)$ is an unknown smooth function and ε's are independent and identically distributed random errors with $\mathrm{E}(\varepsilon|x, t) = 0$ and $\mathrm{Var}(\varepsilon|x, t) = \sigma^2$. We shall call $f(t)$ the smooth part of the model and assume that it represents a smooth unparameterized functional relationship.

The goal is to estimate the unknown parameter vector β and the nonparametric function $f(t)$ from the data $\{y_i, x_i, t_i\}_{i=1}^n$. In vector/matrix notation, (1) is written as

$$y = X\beta + f + \varepsilon \tag{2}$$

where $y = (y_1, \ldots, y_n)^\top$, $X = (x_1, \ldots, x_n)$, $f = \{f(t_1), \ldots, f(t_n)\}^\top$, $\varepsilon = (\varepsilon_1, \ldots, \varepsilon_n)^\top$.

Semiparametric models are by design more flexible than standard linear regression models since they combine both parametric and nonparametric components. There exist various goodness-of-fit tests to identify the nonparametric part in this kind of models; see [8] and the references therein. Estimation techniques for semiparametric partially linear models are based on different nonparametric regression procedures. The most important approaches to estimate β and f are given in [12,4,7,6,5,14,24,15,33].

In practice, researchers often encounter the problem of multicollinearity. In case of multicollinearity, we know that the $(p \times p)$ matrix $X^\top X$ has one or more small eigenvalues; the estimates of the regression coefficients can therefore have large variances: the least squares estimator performs poorly in this case. Hoerl and Kennard [17] proposed the ridge regression estimator and it has become the most common method to overcome this particular weakness of the least squares estimator. For the purpose of this paper, we will employ the biased estimator that was proposed by Liu [20] to combat the multicollinearity. The Liu estimator combines the Stein [26] estimator with the ridge regression estimator; see also [1,13].

The condition number is a measure of multicollinearity. If $X^\top X$ is ill-conditioned with a large condition number, the ridge regression estimator or Liu estimator can be used to estimate β, [21]. We consider difference based ridge and Liu type estimators in comparison to the unbiased difference based approach. We give theoretical conditions that determine superiority among the estimation techniques in the mean squared error matrix sense.

We use data on monthly electricity consumption and its determinants (income, electricity and gas prices, temperature) for Germany. The purpose is to understand electricity consumption as a linear function of income and price and a nonlinear function of temperature: semiparametric approach is therefore necessary here. The data reveal a high condition number of 20.5; we therefore expect a more precise estimation with Ridge or Liu type estimators. We show how our theoretically derived conditions can be implemented for a given data set and be used to determine the appropriate biased estimation technique.

The paper is organized as follows. In Section 2, the model and the differencing estimator is defined. We introduce difference based ridge and Liu type estimators in Section 3. In Section 4, the differencing estimator proposed by Yatchew [30] and the difference based Liu type estimator are compared in terms of the mean squared error. In Section 5, both biased regression methodologies in semiparametric regression models are compared in terms of the mean squared error. Section 6 relaxes the assumption of i.i.d. errors and replicates the results of the previous sections in the presence of heteroscedasticity and autocorrelation. Section 7 gives a real data example to show the performance of the proposed estimators.

2. The model and differencing estimator

In this section, we introduce a difference based technique for the estimation of the linear coefficient vector in a semiparametric regression. This technique has been used to remove the nonparametric component in the partially linear model by various authors (e.g. [30,32,19,3]).

Consider the semiparametric regression model (2). Let $d = (d_0, d_1, \ldots, d_m)^\top$ be an $m + 1$ vector where m is the order of differencing and $d_0, d_1, \ldots, d_m$ are differencing weights that minimize

$$\sum_{k=1}^{m} \left(\sum_{j=1}^{m-k} d_j d_{k+j} \right)^2,$$

such that

$$\sum_{j=0}^{m} d_j = 0 \quad \text{and} \quad \sum_{j=0}^{m} d_j^2 = 1 \tag{3}$$

are satisfied.

Let us define the $(n - m) \times n$ differencing matrix D to have first and last rows $(d^\top, 0_{n-m-1}^\top)$, $(0_{n-m-1}^\top, d^\top)$ respectively, with i-th row $(0_i, d^\top, 0_{n-m-i-1}^\top)$, $i = 1, \ldots, (n - m - 1)$, where 0_r indicates an r-vector of all zero elements

$$D = \begin{pmatrix} d_0 & d_1 & d_2 & \cdots & d_m & 0 & \cdots & \cdots & 0 \\ 0 & d_0 & d_1 & d_2 & \cdots & d_m & 0 & \cdots & 0 \\ \vdots & \vdots & & & & & & & \\ 0 & \cdots & \cdots & d_0 & d_1 & d_2 & \cdots & d_m & 0 \\ 0 & 0 & \cdots & \cdots & d_0 & d_1 & d_2 & \cdots & d_m \end{pmatrix}.$$

Applying the differencing matrix to (2) permits direct estimation of the parametric effect. Eubank et al. [6] showed that the parameter vector in (2) can be estimated with parametric efficiency. If f is an unknown function with bounded first derivative, then Df is essentially 0, so that applying the differencing matrix we have

$$\begin{aligned} Dy &= DX\beta + Df + D\varepsilon \approx DX\beta + D\varepsilon \\ \widetilde{y} &\approx \widetilde{X}\beta + \widetilde{\varepsilon} \end{aligned} \tag{4}$$

where $\widetilde{y} = Dy$, $\widetilde{X} = DX$ and $\widetilde{\varepsilon} = D\varepsilon$. Constraints (3) ensure that the nonparametric effect is removed and $\mathrm{Var}(\widetilde{\varepsilon}) = \mathrm{Var}(\varepsilon) = \sigma^2$. With (4), a simple differencing estimator of the parameter β in the semiparametric regression model results:

$$\begin{aligned} \widehat{\beta}_{(0)} &= \{(DX)^\top (DX)\}^{-1} (DX)^\top Dy \\ &= \left(\widetilde{X}^\top \widetilde{X}\right)^{-1} \widetilde{X}^\top \widetilde{y}. \end{aligned} \tag{5}$$

Thus, differencing allows one to perform inferences on β as if there were no nonparametric component f in model (2), [9].

We will also use the modified estimator of σ^2 proposed by Eubank et al. [7]

$$\widehat{\sigma}^2 = \frac{\widetilde{y}^\top (I - P^\perp)\widetilde{y}}{\mathrm{tr}\{D^\top (I - P^\perp) D\}} \tag{6}$$

with $P^\perp = \widetilde{X}(\widetilde{X}^\top \widetilde{X})^{-1}\widetilde{X}^\top$, I $(p \times p)$ identity matrix and $\mathrm{tr}(\cdot)$ denoting the trace function for a square matrix.

3. Difference based ridge and Liu type estimator

As an alternative to $\widehat{\beta}_{(0)}$ in (5), [27] propose:

$$\widehat{\beta}_{(1)}(k) = (\widetilde{X}^\top \widetilde{X} + kI)^{-1} \widetilde{X}^\top \widetilde{y}, \quad k \geq 0;$$

here k is the ridge-biasing parameter selected by the researcher. We call $\widehat{\beta}_{(1)}(k)$ a difference based ridge regression estimator of the semiparametric regression model.

From the least squares perspective, the coefficients β are chosen to minimize

$$(\widetilde{y} - \widetilde{X}\beta)^\top (\widetilde{y} - \widetilde{X}\beta). \tag{7}$$

Adding to the least squares objective (7) a penalizing function of the squared norm $\|\eta\widehat{\beta}_{(0)} - \beta\|^2$ for the vector of regression coefficients, yields a conditional objective:

$$L = (\widetilde{y} - \widetilde{X}\beta)^\top (\widetilde{y} - \widetilde{X}\beta) + (\eta\widehat{\beta}_{(0)} - \beta)^\top (\eta\widehat{\beta}_{(0)} - \beta). \tag{8}$$

Minimizing (8) with respect to β, we obtain the estimator $\widehat{\beta}_{(2)}(\eta)$ an alternative to $\widehat{\beta}_{(0)}$ in (5):

$$\widehat{\beta}_{(2)}(\eta) = (\widetilde{X}^\top \widetilde{X} + I)^{-1} (\widetilde{X}^\top \widetilde{y} + \eta\widehat{\beta}_{(0)}), \tag{9}$$

where η, $0 \leq \eta \leq 1$, is a biasing parameter and when $\eta = 1$, $\widehat{\beta}_{(2)}(\eta) = \widehat{\beta}_{(0)}$. The formal resemblance between (9) and the Liu estimator motivated [1,18,29] to call it the difference based Liu type estimator of the semiparametric regression model.

4. Mean squared error matrix (MSEM) comparison of $\widehat{\beta}_{(0)}$ with $\widehat{\beta}_{(2)}(\eta)$

In this section, the objective is to examine the difference of the mean square error matrices of $\widehat{\beta}_{(0)}$ and $\widehat{\beta}_{(2)}(\eta)$. We note that for any estimator $\widetilde{\beta}$ of β, its mean squared error matrix (MSEM) is defined as $\mathrm{MSEM}(\widetilde{\beta}) = \mathrm{Cov}(\widetilde{\beta}) + \mathrm{Bias}(\widetilde{\beta})\,\mathrm{Bias}(\widetilde{\beta})^\top$, where $\mathrm{Cov}(\widetilde{\beta})$ denotes the variance–covariance matrix and $\mathrm{Bias}(\widetilde{\beta}) = \mathrm{E}(\widetilde{\beta}) - \beta$ is the bias vector. The expected value of $\widehat{\beta}_{(2)}(\eta)$ can be written as

$$\mathrm{E}\{\widehat{\beta}_{(2)}(\eta)\} = \beta - (1 - \eta)(\widetilde{X}^\top \widetilde{X} + I)^{-1}\beta.$$

The bias of the $\widehat{\beta}_{(2)}(\eta)$ is given as

$$\mathrm{Bias}\{\widehat{\beta}_{(2)}(\eta)\} = -(1 - \eta)(\widetilde{X}^\top \widetilde{X} + I)^{-1}\beta. \tag{10}$$

Denoting $F_\eta = (\widetilde{X}^\top \widetilde{X} + I)^{-1}(\widetilde{X}^\top \widetilde{X} + \eta I)$ and observing F_η and $(\widetilde{X}^\top \widetilde{X})^{-1}$ are commutative, we may write $\widehat{\beta}_{(2)}(\eta)$ as

$$\begin{aligned} \widehat{\beta}_{(2)}(\eta) &= F_\eta \widehat{\beta}_{(0)} = F_\eta (\widetilde{X}^\top \widetilde{X})^{-1} \widetilde{X}^\top \widetilde{y} \\ &= (\widetilde{X}^\top \widetilde{X})^{-1} F_\eta \widetilde{X}^\top \widetilde{y}. \end{aligned}$$

Setting $S = (D^\top \widetilde{X})^\top (D^\top \widetilde{X})$ and $U = (\widetilde{X}^\top \widetilde{X})^{-1}$ we may write $\mathrm{Cov}\{\widehat{\beta}_{(2)}(\eta)\}$ as

$$\mathrm{Cov}\{\widehat{\beta}_{(2)}(\eta)\} = \sigma^2 F_\eta USUF_\eta^\top, \tag{11}$$

$$\mathrm{Cov}(\widehat{\beta}_{(0)}) = \sigma^2 USU. \tag{12}$$

Using (11) and (12), the difference $\Delta_1 = \mathrm{Cov}(\widehat{\beta}_{(0)}) - \mathrm{Cov}\{\widehat{\beta}_{(2)}(\eta)\}$ can be expressed as

$$\begin{aligned}\Delta_1 &= \sigma^2 \left(USU - F_\eta USUF_\eta^\top\right)\\ &= \sigma^2 F_\eta \{F_\eta^{-1} USU (F_\eta^\top)^{-1} - USU\} F_\eta^\top\\ &= \sigma^2 (1-\eta^2)(U^{-1}+I)^{-1} \left\{ \frac{1}{1+\eta}(US+SU) + USU \right\} (U^{-1}+I)^{-1}. \end{aligned} \tag{13}$$

Let $\tau = \frac{1}{1+\eta} > 0$, $M = USU$, $N = US + SU$. Since $M = L^\top L$ and $\mathrm{rank}(L) = p < n - m$, then M is a $(p \times p)$ positive definite matrix, where $L = D^\top \widetilde{X}(\widetilde{X}^\top \widetilde{X})^{-1}$ and $N = US + SU$ is a symmetric matrix. Thus, we may write (13) as

$$\begin{aligned}\Delta_1 &= \sigma^2 (1-\eta^2) H(M + \tau N) H\\ &= \sigma^2 (1-\eta^2) H (Q^\top)^{-1} (Q^\top M Q + \tau Q^\top N Q) Q^{-1} H\\ &= \sigma^2 (1-\eta^2) H (Q^\top)^{-1} (I + \tau E) Q^{-1} H, \end{aligned}$$

where $I + \tau E = \mathrm{diag}(1+\tau e_{11}, \ldots, 1+\tau e_{pp})$ and $H = (U^{-1}+I)^{-1}$. Since M is a positive definite and N is a symmetric matrix, a nonsingular matrix Q exists such that $Q^\top M Q = I$ and $Q^\top N Q = E$; here E is a diagonal matrix and its diagonal elements are the roots of the polynomial equation $|M^{-1}N - eI| = 0$ (see [11, pp. 408] and [16, pp. 563]) and since $N = US + SU \neq 0$, there is at least one diagonal element of E that is nonzero. Let $e_{ii} < 0$ for at least one i; then positive definiteness of $I + \tau E$ is guaranteed by

$$0 < \tau < \min_{e_{ii}<0} \left| \frac{1}{e_{ii}} \right|. \tag{14}$$

Hence $1 + \tau e_{ii} > 0$ for all $i = 1, \ldots, p$ and therefore $I + \tau E$ is a positive definite matrix. Consequently, Δ_1 becomes a positive definite matrix, as well. It is now evident that the estimator $\widehat{\beta}_{(2)}(\eta)$ has a smaller variance compared with the estimator $\widehat{\beta}_{(0)}$ if and only if (14) is satisfied.

Next, we give necessary and sufficient conditions for the difference based Liu type estimator $\widehat{\beta}_{(2)}(\eta)$ to be superior to $\widehat{\beta}_{(0)}$ in the mean squared error matrix (MSEM) sense.

The proof of the next theorem requires the following lemma.

Lemma 4.1 (*Farebrother [10]*). *Let A be a positive definite $(p \times p)$ matrix, b a $(p \times 1)$ nonzero vector and δ a positive scalar. Then $\delta A - bb^\top$ is non-negative if and only if $b^\top A^{-1} b \leq \delta$.*

Let us compare the performance of $\widehat{\beta}_{(2)}(\eta)$ with the differencing estimator $\widehat{\beta}_{(0)}$ with respect to the MSEM criterion. In order to do that, define $\Delta_2 = \mathrm{MSEM}(\widehat{\beta}_{(0)}) - \mathrm{MSEM}\{\widehat{\beta}_{(2)}(\eta)\}$. Observe that

$$\mathrm{MSEM}(\widehat{\beta}_{(0)}) = \mathrm{Cov}(\widehat{\beta}_{(0)}) = \sigma^2 USU \tag{15}$$

and

$$\mathrm{MSEM}\{\widehat{\beta}_{(2)}(\eta)\} = \sigma^2 F_\eta USUF_\eta^\top + (1-\eta)^2 (U^{-1}+I)^{-1} \beta\beta^\top (U^{-1}+I)^{-1}. \tag{16}$$

Then from (15) and (16) one derives

$$\begin{aligned}\Delta_2 &= \sigma^2 F_\eta \{F_\eta^{-1} USU (F_\eta^\top)^{-1} - USU\} F_\eta^\top - (1-\eta)^2 (U^{-1}+I)^{-1} \beta\beta^\top (U^{-1}+I)^{-1},\\ &= H \left\{ \sigma^2 (1-\eta^2)(M + \tau N) - (1-\eta)^2 \beta\beta^\top \right\} H,\\ &= (1-\eta)^2 H \left\{ \sigma^2 \frac{1+\eta}{1-\eta} (M + \tau N) - \beta\beta^\top \right\} H. \end{aligned}$$

Applying Lemma 4.1 and assuming condition (14) to be satisfied, we see Δ_2 is positive definite if and only if

$$\beta^\top (M + \tau N)^{-1} \beta \leq \sigma^2 \frac{1+\eta}{1-\eta}, \quad 0 < \eta < 1.$$

Now we may state the following theorem.

Theorem 4.1. *Consider the two estimators $\widehat{\beta}_{(2)}(\eta)$ and $\widehat{\beta}_{(0)}$ of β. Let $W = \frac{1+\eta}{1-\eta}(M + \tau N)$ be a positive definite matrix. Then the biased estimator $\widehat{\beta}_{(2)}(\eta)$ is MSEM superior to $\widehat{\beta}_{(0)}$ if and only if*

$$\beta^\top W^{-1} \beta \leq \sigma^2.$$

5. MSEM comparison of $\widehat{\beta}_{(1)}(k)$ and $\widehat{\beta}_{(2)}(\eta)$

Let us now compare the MSEM performance of

$$\begin{aligned}\widehat{\beta}_{(1)}(k) &= (\widetilde{X}^\top\widetilde{X}+kI)^{-1}\widetilde{X}^\top\widetilde{y}\\ &= S_k\widetilde{X}^\top Dy\\ &= A_1y \end{aligned} \tag{17}$$

with

$$\begin{aligned}\widehat{\beta}_{(2)}(\eta) &= (\widetilde{X}^\top\widetilde{X}+I)^{-1}(\widetilde{X}^\top y+\eta\widehat{\beta}_{(0)})\\ &= (\widetilde{X}^\top\widetilde{X})^{-1}(\widetilde{X}^\top\widetilde{X}+I)^{-1}(\widetilde{X}^\top\widetilde{X}+\eta I)\widetilde{X}^\top\widetilde{y}\\ &= UF_\eta\widetilde{X}^\top Dy\\ &= A_2y. \end{aligned} \tag{18}$$

The MSEM of the difference based ridge regression estimator $\widehat{\beta}_{(1)}(k)$ is given by

$$\begin{aligned}\mathrm{MSEM}\{\widehat{\beta}_{(1)}(k)\} &= \mathrm{Cov}\{\widehat{\beta}_{(1)}(k)\}+\mathrm{Bias}\{\widehat{\beta}_{(1)}(k)\}\,\mathrm{Bias}\{\widehat{\beta}_{(1)}(k)\}^\top\\ &= S_k(\sigma^2S+k^2\beta\beta^\top)S_k^\top\\ &= \sigma^2(A_1A_1^\top)+d_1d_1^\top,\end{aligned}$$

where $S_k=(\widetilde{X}^\top\widetilde{X}+kI)^{-1}$ and $d_1=\mathrm{Bias}\{\widehat{\beta}_{(1)}(k)\}=-kS_k\beta$; see [27]. The MSEM in (16) may be written as

$$\mathrm{MSEM}\{\widehat{\beta}_{(2)}(\eta)\}=\sigma^2(A_2A_2^\top)+d_2d_2^\top,$$

with $d_2=\mathrm{Bias}\{\widehat{\beta}_{(2)}(\eta)\}=-(1-\eta)(U^{-1}+I)^{-1}\beta$.

Define

$$\Delta_3=\mathrm{MSEM}\{\widehat{\beta}_{(1)}(k)\}-\mathrm{MSEM}\{\widehat{\beta}_{(2)}(\eta)\}=\sigma^2(A_1A_1^\top-A_2A_2^\top)+(d_1d_1^\top-d_2d_2^\top). \tag{19}$$

For the following proofs we employ the following lemma.

Lemma 5.1 (*Trenkler and Toutenburg [28]*). *Let $\widetilde{\beta}_{(j)}=A_jy, j=1,2$ be the two linear estimators of β. Suppose the difference $\mathrm{Cov}(\widetilde{\beta}_{(1)})-\mathrm{Cov}(\widetilde{\beta}_{(2)})$ of the covariance matrices of the estimators $\widetilde{\beta}_{(1)}$ and $\widetilde{\beta}_{(2)}$ is positive definite. Then $\mathrm{MSEM}(\widetilde{\beta}_{(1)})-\mathrm{MSEM}(\widetilde{\beta}_{(2)})$ is positive definite if and only if $d_2^\top\{\mathrm{Cov}(\widetilde{\beta}_{(1)})-\mathrm{Cov}(\widetilde{\beta}_{(2)})+d_1d_1^\top\}^{-1}d_2<1$.*

Theorem 5.1. *The sampling variance of $\widehat{\beta}_{(2)}(\eta)$ is smaller than that of $\widehat{\beta}_{(1)}(k)$, if and only if $\lambda_{\min}(G_2^{-1}G_1)>1$, where $\lambda_{\min}$ is the minimum eigenvalue of $G_2^{-1}G_1$ and $G_j=\sigma^2A_jA_j^\top,\ j=1,2$.*

Proof. Consider the difference

$$\begin{aligned}\Delta^* &= \mathrm{Cov}\{\widehat{\beta}_{(1)}(k)\}-\mathrm{Cov}\{\widehat{\beta}_{(2)}(\eta)\}\\ &= \sigma^2(A_1A_1^\top-A_2A_2^\top),\\ &= G_1-G_2\end{aligned}$$

with $G_1=(D^\top\widetilde{X}W_kU)^\top=V^\top V$, $W_k=I+kU$ and $G_2=(\widetilde{X}F_\eta^\top U)^\top(\widetilde{X}F_\eta^\top U)$. Since $\mathrm{rank}(V)=p<n-m$, G_1 is a $(p\times p)$ positive definite matrix and G_2 is a symmetric matrix. Hence, a nonsingular matrix O exists such that $O^\top G_1O=I$ and $O^\top G_2O=\Lambda$, with Λ diagonal matrix with diagonal elements roots λ of the polynomial equation $|G_1-\lambda G_2|=0$ (see [16, p. 563] or [25, p. 160]). Thus, we may write $\Delta^*=(O^\top)^{-1}(O^\top G_1O-O^\top G_2O)O^{-1}=(O^\top)^{-1}(\Lambda-I)O^{-1}$ or $O^\top\Delta^*O=\Lambda-I$. If G_1-G_2 is positive definite, then $O^\top G_1O-O^\top G_2O=\Psi-I$ is positive definite. Hence $\lambda_i-1>0$, $i=1,2,\ldots,p$, so we get $\lambda_{\min}(G_2^{-1}G_1)>1$.

Now let $\lambda_{\min}(G_2^{-1}G_1)>1$ hold. Furthermore, with G_2 positive definite and G_1 symmetric, we have $\lambda_{\min}<\frac{\nu^\top G_1\nu}{\nu^\top G_2\nu}<\lambda_{\max}$ for all nonzero $(p\times 1)$ vectors ν, so G_1-G_2 is positive definite; see [23, p. 74]. It is obvious that $\mathrm{Cov}\{\widehat{\beta}_{(2)}(\eta)\}-\mathrm{Cov}\{\widehat{\beta}_{(1)}(k)\}$ is positive definite for $0\le\eta\le1,\ k\ge0$ if and only if $\lambda_{\min}(G_2^{-1}G_1)>1$. □

Theorem 5.2. *Consider $\widehat{\beta}_{(1)}(k)=A_1y$ and $\widehat{\beta}_{(2)}(\eta)=A_2y$ of β. Suppose that the difference $\mathrm{Cov}\{\widehat{\beta}_{(1)}(k)\}-\mathrm{Cov}\{\widehat{\beta}_{(2)}(\eta)\}$ is positive definite. Then*

$$\Delta_3=\mathrm{MSEM}\{\widehat{\beta}_{(1)}(k)\}-\mathrm{MSEM}\{\widehat{\beta}_{(2)}(\eta)\}$$

is positive definite if and only if

$$d_2^\top\{\sigma^2(A_1A_1^\top-A_2A_2^\top)+d_1d_1^\top\}^{-1}d_2<1$$

with $A_1=S_k\widetilde{X}^\top D,\ A_2=UF_\eta\widetilde{X}^\top D$.

Proof. The difference between the MSEMs of $\widehat{\beta}_{(2)}(\eta)$ and $\widehat{\beta}_{(1)}(k)$ is given by

$$\begin{aligned}\Delta_3 &= \mathrm{MSEM}\{\widehat{\beta}_{(1)}(k)\} - \mathrm{MSEM}\{\widehat{\beta}_{(2)}(\eta)\}\\ &= \sigma^2(A_1A_1^\top - A_2A_2^\top) + (d_1d_1^\top - d_2d_2^\top)\\ &= \mathrm{Cov}\{\widehat{\beta}_{(1)}(k)\} - \mathrm{Cov}\{\widehat{\beta}_{(2)}(\eta)\} + (d_1d_1^\top - d_2d_2^\top).\end{aligned}$$

Applying Lemma 5.1 yields the desired result. □

It should be noted that all results reported above are based on the assumption that k and η are non-stochastic. The theoretical results indicate that the $\widehat{\beta}_{(2)}(\eta)$ is not always better than the $\widehat{\beta}_{(1)}(k)$, and vice versa. For practical purposes, we have to replace these unknown parameters by some suitable estimators.

6. The heteroscedasticity and correlated error case

Up to this point, independent errors with equal variance were assumed. The error term might also exhibit autocorrelation. To account for these effects, we extend the results in this section and consider the more general case of heteroscedasticity and autocovariance in the error terms.

Consider now observations $\{y_t, x_t, t_t\}_{t=1}^T$ and the semiparametric partial linear model $y_t = x_t^\top\beta + f(t_t) + \varepsilon_t,\ t = 1, \ldots, T$. Let $\mathsf{E}(\varepsilon\varepsilon^\top|x, t) = \Omega$ not necessarily diagonal. To keep the structure of the errors for later inference, we define an $(n \times n)$ permutation matrix P as in [32]. Consider a permutation:

$$\begin{pmatrix} 1 & t_{(1)} \\ \cdots & \cdots \\ i & t_{(i)} \\ \cdots & \cdots \\ n & t_{(n)} \end{pmatrix}$$

where $i = 1, \ldots, n$ is the index of the ordered nonparametric variable and $t_{(i)} = 1, \ldots, T$ corresponding time index of the observations. Then P is defined for $i, j = 1, \ldots, n$:

$$P_{ij} = \begin{cases} 1, & j = t_{(i)} \\ 0, & \text{otherwise.} \end{cases}$$

We can now rewrite the model after reordering and differencing:

$$DPy = DPX\beta + DPf(x) + DP\varepsilon, \qquad \mathsf{E}(\varepsilon\varepsilon^\top|x, t) = \Omega. \tag{20}$$

Then, with $\widetilde{X} = DPX$ and $\widetilde{y} = DPy$ from (20), $\widehat{\beta}_{(0)}$ is given:

$$\widehat{\beta}_{(0)} = (\widetilde{X}^\top\widetilde{X})^{-1}\widetilde{X}^\top\widetilde{y} \tag{21}$$

with

$$\begin{aligned}\mathrm{Cov}(\widehat{\beta}_{(0)}) &= (\widetilde{X}^\top\widetilde{X})^{-1}\widetilde{X}^\top DP\Omega D^\top P^\top\widetilde{X}(\widetilde{X}^\top\widetilde{X})^{-1}\\ &= U\widetilde{X}^\top DP\Omega D^\top P^\top\widetilde{X}U.\end{aligned} \tag{22}$$

We will use a heteroscedasticity and autocovariance consistent estimator described in [22] for the interior matrix of (22), which is in our case:

$$\widehat{DP\Omega D^\top P^\top} = \{\widehat{DP\varepsilon}(\widehat{DP\varepsilon})^\top\} \odot \left\{\sum_{\ell=0}^{\mathcal{L}}\left(1 - \frac{\ell}{\mathcal{L}+1}\right)H^\ell\right\} \tag{23}$$

with $\widehat{DP\varepsilon} = \widetilde{y} - \widetilde{X}\widehat{\beta}_{(0)}$, $\odot$ denoting the elementwise matrix product, $\mathcal{L}$ the maximum lag of nonzero autocorrelation in the errors and H^0 the identity matrix. Let L_ℓ be a matrix with ones on the ℓth diagonal; then $H^\ell,\ \ell = 1, \ldots \mathcal{L}$ are such that:

$$H_{ij}^\ell = \begin{cases} 0, & \text{if } \{DP(L_\ell + L_\ell^\top)D^\top P^\top\}_{ij} = 0, \\ 1, & \text{otherwise and } i, j = 1, \ldots, p. \end{cases}$$

Plugging (23) in (22), we obtain a consistent estimator for $\mathrm{Cov}(\widehat{\beta}_{(0)})$; see [31] for details.

Denoting $\widetilde{S} = \widetilde{X}^\top DP\Omega D^\top P^\top\widetilde{X}$, we can write down $\mathrm{Cov}\{\widehat{\beta}_{(1)}(k)\}$ and $\mathrm{Cov}\{\widehat{\beta}_{(2)}(\eta)\}$ in model (20).

$$\mathrm{Cov}\{\widehat{\beta}_{(1)}(k)\} = S_k\widetilde{S}S_k \tag{24}$$

$$\mathrm{Cov}\{\widehat{\beta}_{(2)}(\eta)\} = F_\eta U\widetilde{S}UF_\eta. \tag{25}$$

Using (22) and (25), the difference $\Delta_1 = \mathrm{Cov}(\widehat{\beta}_{(0)}) - \mathrm{Cov}\{\widehat{\beta}_{(2)}(\eta)\}$ can be expressed as

$$\begin{aligned}\Delta_1 &= \left(U\widetilde{S}U - F_\eta U\widetilde{S}UF_\eta^\top\right)\\ &= F_\eta\{F_\eta^{-1}U\widetilde{S}U(F_\eta^\top)^{-1} - U\widetilde{S}U\}F_\eta^\top\\ &= (1-\eta^2)(U^{-1}+I)^{-1}\left\{\frac{1}{1+\eta}(U\widetilde{S}+\widetilde{S}U)+U\widetilde{S}U\right\}(U^{-1}+I)^{-1},\end{aligned} \tag{26}$$

with $\tau = \frac{1}{1+\eta} > 0$, $\widetilde{M} = U\widetilde{S}U$, $\widetilde{N} = U\widetilde{S} + \widetilde{S}U$. Since $\widetilde{M}$ is a $(p \times p)$ positive definite matrix and $\widetilde{N}$ is a symmetric matrix, a nonsingular matrix T exists such that $T^\top\widetilde{M}T = I$ and $T^\top\widetilde{N}T = \widetilde{E}$; here $\widetilde{E}$ is a diagonal matrix and its diagonal elements are the roots of the polynomial equation $|\widetilde{M}^{-1}\widetilde{N} - \widetilde{e}I| = 0$ (see [11, pp. 408] and [16, pp. 563]) and we may write (26) as

$$\begin{aligned}\Delta_1 &= (1-\eta^2)H(\widetilde{M}+\tau\widetilde{N})H\\ &= (1-\eta^2)H(T^\top)^{-1}(T^\top\widetilde{M}T + \tau T^\top\widetilde{N}T)T^{-1}H\\ &= (1-\eta^2)H(T^\top)^{-1}(I+\tau\widetilde{E})T^{-1}H,\end{aligned}$$

where $I + \widetilde{\tau}\widetilde{E} = \mathrm{diag}(1+\tau\widetilde{e}_{11}, \ldots, 1+\tau\widetilde{e}_{pp})$ and $H = (U^{-1}+I)^{-1}$. Since $\widetilde{N} = U\widetilde{S} + \widetilde{S}U \neq 0$, there is at least one diagonal element of $\widetilde{E}$ that is nonzero.

Let $\widetilde{e}_{ii} < 0$ for at least one i; then positive definiteness of $I + \tau\widetilde{E}$ is guaranteed by

$$0 < \tau < \min_{\widetilde{e}_{ii}<0}\left|\frac{1}{\widetilde{e}_{ii}}\right|. \tag{27}$$

Hence $1+\tau\widetilde{e}_{ii} > 0$ for all $i = 1, \ldots, p$ and therefore $I+\tau\widetilde{E}$ is a positive definite matrix. Consequently, Δ_1 becomes a positive definite matrix, as well. It is now evident that the estimator $\widehat{\beta}_{(2)}(\eta)$ has a smaller variance compared with the estimator $\widehat{\beta}_{(0)}$ if and only if (27) is satisfied.

With

$$\begin{aligned}\Delta_1' &= \mathrm{Cov}(\widehat{\beta}_{(0)}) - \mathrm{Cov}\{\widehat{\beta}_{(1)}(k)\}\\ &= k^2S_k\left\{\frac{1}{k}(U\widetilde{S}+\widetilde{S}U)+U\widetilde{S}U\right\}S_k\\ &= k^2S_k\left(\frac{1}{k}\widetilde{N}+\widetilde{M}\right)S_k\end{aligned}$$

and analogous argumentation as above obtained for $\widehat{\beta}_{(1)}(k)$:

$$0 < \frac{1}{k} < \min_{\widetilde{e}_{ii}<0}\left|\frac{1}{\widetilde{e}_{ii}}\right|. \tag{28}$$

The next theorem extends the results of Theorem 3.1 in [27] and Theorem 4.1 of Section 4 to the more general case of (20).

Theorem 6.1. *Consider the estimators $\widehat{\beta}_{(i)}(x)$, $i = \{1, 2\}$; $x = \{k, \eta\}$ and $\widehat{\beta}_{(0)}$ of β. Let $W_1 = \widetilde{M} + \tau\widetilde{N}$, $W_2 = \frac{1+\eta}{1-\eta}(\widetilde{M}+\tau\widetilde{N})$ be positive definite (alternative: assume that (27) and (28) hold). Then the biased estimator $\widehat{\beta}_{(i)}(x)$ is MSEM superior to $\widehat{\beta}_{(0)}$ if and only if*

$$\beta^\top W_i^{-1}\beta \leq 1.$$

Proof. Consider the differences

$$\begin{aligned}\Delta_2 &= \mathrm{MSEM}(\widehat{\beta}_{(0)}) - \mathrm{MSEM}\{\widehat{\beta}_{(2)}(\eta)\}\\ &= \mathrm{Cov}(\widehat{\beta}_{(0)}) - \mathrm{Cov}\{\widehat{\beta}_{(2)}(\eta)\} - \mathrm{Bias}\{\widehat{\beta}_{(2)}(\eta)\}\,\mathrm{Bias}\{\widehat{\beta}_{(2)}(\eta)\}^\top\\ &= F_\eta\{F_\eta^{-1}U\widetilde{S}U(F_\eta^\top)^{-1} - U\widetilde{S}U\}F_\eta^\top - (1-\eta)^2(U^{-1}+I)^{-1}\beta\beta^\top(U^{-1}+I)^{-1}\\ &= (1-\eta)^2H\left\{\frac{1+\eta}{1-\eta}(\widetilde{M}+\tau\widetilde{N}) - \beta\beta^\top\right\}H\\ &= (1-\eta)^2H\left(W_2 - \beta\beta^\top\right)H.\end{aligned}$$

$$\begin{aligned}\Delta_2' &= \mathrm{MSEM}(\widehat{\beta}_{(0)}) - \mathrm{MSEM}\{\widehat{\beta}_{(1)}(k)\}\\ &= \mathrm{Cov}(\widehat{\beta}_{(0)}) - \mathrm{Cov}\{\widehat{\beta}_{(1)}(k)\} - \mathrm{Bias}\{\widehat{\beta}_{(1)}(k)\}\,\mathrm{Bias}\{\widehat{\beta}_{(1)}(k)\}^\top\\ &= S_k\{k(\widetilde{S}U + U\widetilde{S}) + k^2U\widetilde{S}U - k^2\beta\beta^\top\}S_k\\ &= k^2S_k\left(\frac{1}{k}\widetilde{N} + \widetilde{M} - \beta\beta^\top\right)S_k\\ &= k^2S_k(W_1 - \beta\beta^\top)S_k.\end{aligned}$$

With Lemma 4.1, the assertion follows. □

Theorem 6.1 gives conditions under which the biased estimator $\widehat{\beta}_{(i)}(x),\ i = \{1, 2\};\ x = \{k, \eta\}$ is superior to $\widehat{\beta}_{(0)}$ in the presence of heteroscedasticity and autocorrelation in the data.

Note that for comparison of the biased estimators Theorem 5.1 can be extended straight forwardly to the general case by exchanging G_1 and G_2 by $\widetilde{G}_1 = \widetilde{A}_1\Omega\widetilde{A}_1^\top$ and $\widetilde{G}_2 = \widetilde{A}_2\Omega\widetilde{A}_2^\top$ correspondingly, with $\widetilde{A}_1 = S_k\widetilde{X}^\top DP,\ \widetilde{A}_2 = UF_\eta\widetilde{X}^\top DP$. Hence, the sampling variance of $\widehat{\beta}_{(2)}(\eta)$ is always smaller than that of $\widehat{\beta}_{(1)}(k)$, if and only if $\lambda_{\min}(\widetilde{G}_2^{-1}\widetilde{G}_1) > 1$, where $\lambda_{\min}$ is the minimum eigenvalue of $\widetilde{G}_2^{-1}\widetilde{G}_1$.

Now, we give a generalized version of Theorem 5.2.

Theorem 6.2. *Consider $\widehat{\beta}_{(1)} = \widetilde{A}_1y$ and $\widehat{\beta}_{(2)} = \widetilde{A}_2y$ of β. Suppose that the difference $\mathrm{Cov}\{\widehat{\beta}_{(1)}\} - \mathrm{Cov}\{\widehat{\beta}_{(2)}\}$ is positive definite. Then*

$$\Delta_3 = \mathrm{MSEM}(\widehat{\beta}_{(1)}) - \mathrm{MSEM}(\widehat{\beta}_{(2)})$$

is positive definite if and only if

$$d_2^\top(\widetilde{A}_1\Omega\widetilde{A}_1^\top - \widetilde{A}_2\Omega\widetilde{A}_2^\top + d_1d_1^\top)^{-1}d_2 < 1.$$

Proof. The difference between the MSEMs of $\widehat{\beta}_{(2)}(\eta)$ and $\widehat{\beta}_{(1)}(k)$ is given by

$$\begin{aligned}\Delta_3 &= \mathrm{MSEM}(\widehat{\beta}_{(1)}) - \mathrm{MSEM}(\widehat{\beta}_{(2)})\\ &= \widetilde{A}_1\Omega\widetilde{A}_1^\top - \widetilde{A}_2\Omega\widetilde{A}_2^\top + d_1d_1^\top - d_2d_2^\top\\ &= \mathrm{Cov}(\widehat{\beta}_{(1)}) - \mathrm{Cov}(\widehat{\beta}_{(2)}) + d_1d_1^\top - d_2d_2^\top.\end{aligned}$$

Applying Lemma 5.1 yields the desired result. □

We note that in order to use the criteria above, one has to estimate the parameters. The estimation of Ω is thereby the most challenging. However, as long as estimator (23) is available, all considered criteria can be evaluated on the real data and can be used for practical purposes.

7. Determinants of electricity demand

The empirical study example is motivated by the importance of explaining variation in electricity consumption. Since electricity is a non-storable good, electricity providers are interested in understanding and hedging demand fluctuations.

Electricity consumption is known to be influenced negatively by the price of electricity and positively by the income of the consumers. As electricity is frequently used for heating and cooling, the effect of the air temperature must also be present. Both heating by low temperatures and cooling by high temperatures result in higher electricity consumption and motivate the use of a nonparametric specification for the temperature effect. Thus we consider the semiparametric regression model defined in (1)

$$y = f(t) + \beta_1x_1 + \beta_2x_2 + \beta_3x_3 + \cdots + \beta_{13}x_{13} + \varepsilon, \tag{29}$$

where y is the log monthly electricity consumption per person (aggregated electricity consumption was divided by population interpolated linearly from quarterly data), t is cumulated average temperature index for the corresponding month taken as average of 20 German cities computed from the data of German weather service (Deutscher Wetterdienst), x_1 is the log GDP per person interpolated linearly from quarterly data, detrended and deseasonalized and x_2 is the log rate of electricity price to the gas price, detrended. The data for 199601-201009 comes from EUROSTAT. Reference prices for electricity were computed as an average of electricity tariffs for consumer groups IND-Ie and HH-Dc, for gas—IND-I3-2 and HH-D3 with reference period 2005S1. Time series of prices were obtained by scaling with electricity price or correspondingly gas price indices. $x_3, x_4, \ldots, x_{13}$ are dummy variables for the monthly effects.

The model in (29) includes both parametric effects and a nonparametric effect. The only nonparametric effect is implied by the temperature variable. From Fig. 1, we can see that the effect of t on y is likely to be nonlinear, while the effects of other variables are roughly linear. The dummy variables enter into the linear part in the specification of the semiparametric regression as well.

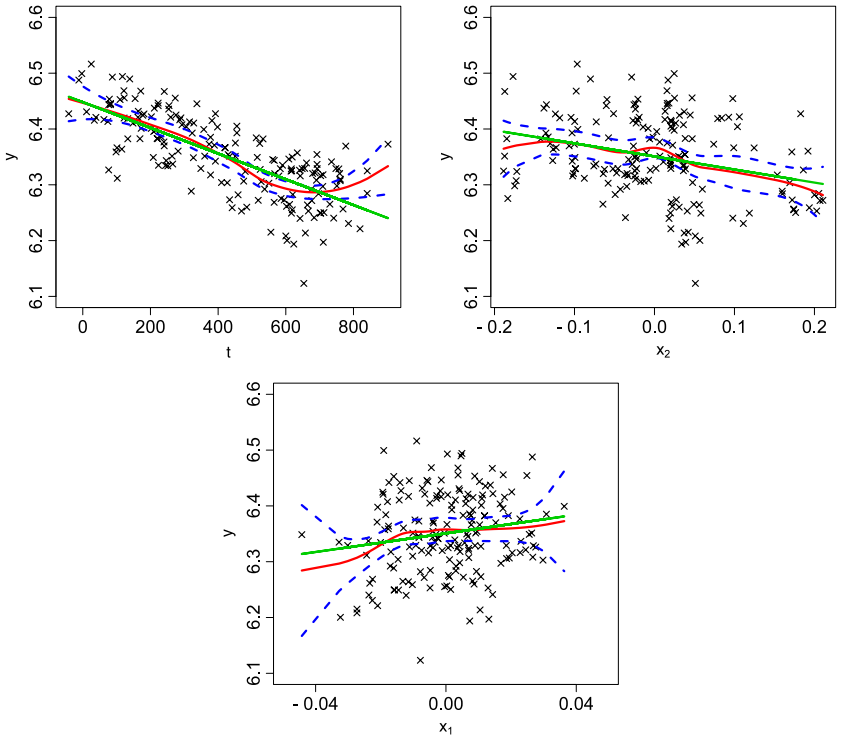


Fig. 1. Plots of individual exp. variables vs. dependent variable, linear fit (green), local polynomial fit (red), 95% confidence bands (black). (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

We note that the condition number of $X^\top X$ of these explanatory variables is 20.5, which justifies the use of $\widehat{\beta}_{(1)}(k)$ and $\widehat{\beta}_{(2)}(\eta)$; see [2].

Throughout the paper, we use fifth-order differencing ($m = 5$). Results for other orders of differencing were similar.

The admissible regions for the biasing parameters η and k for MSEM superiority were $\eta \geq 0.923$ and $k \leq 0.0085$. These bounds were determined using the estimated parameters and the inequalities from Theorem 4.1 and Theorem 3.1 in [27], respectively. Under more general assumptions on Ω and resulting heteroscedasticity and autocovariance consistent Newey–West covariance estimator, defined in (23), the admissible region for η (Theorem 6.1 and restriction (27)) was shrinked to $\eta \geq 0.927$. For $\widehat{\beta}_{(1)}(k)$, no admissible values of k were found, since admissible $k \geq 1.57$ of (28) do not satisfy the condition of Theorem 6.1 (see Table 2).

Alternatively, we used a scalar mean squared error (SMSE), defined as the trace of the corresponding MSEM, to compare the estimators. The bounds for k and η can then be calculated only numerically using a grid on $[0, 1]$ for the biasing parameters and determining the regions where SMSEs of the proposed estimators are lower. SMSE superiority of $\widehat{\beta}_{(1)}(k)$ and $\widehat{\beta}_{(2)}(\eta)$ over $\widehat{\beta}_{(0)}$ under general Ω is given for $k \leq 0.0267$ and $\eta \geq 0.384$ compared to $k \leq 0.0123$ and $\eta \geq 0.708$ by standard assumptions; see Fig. 2 which depicts SMSE of the estimators and the corresponding η and k under standard and general assumptions. Thus the SMSE superiority intervals for η and k become even larger in the case of the general form of Ω.

Our computations here are performed with R 2.10.1 and the codes are available on www.quantlet.org.

Results of different estimation procedures can be found in Table 1. We note that regardless of the estimator type, the effect of income is positive and the effect of relative price is negative as expected from an economic perspective, as in [4]. However, the R^2 obtained by difference based methods is higher and SMSE lower for Liu type and ridge difference based estimator. The values of biasing parameters for which conditions of Theorems 5.1 and 5.2 are satisfied are given in Table 3. The superiority of $\widehat{\beta}_{(2)}(\eta)$ over $\widehat{\beta}_{(1)}(k)$ is assured for the zone of values marked by plus.

Returning to our semiparametric specification, we may now remove the estimated parametric effect from the dependent variable and analyze the nonparametric effect. We use a local linear estimator of f to model the nonparametric effect of temperature. The resulting plots are presented in Fig. 3 where we also include the linear effect. We notice that all differencing procedures result in similar estimators of f, regardless of notable differences in the coefficients of the linear part. The estimator of f is consistent with findings e.g. of [4] for US electricity data.

In both specifications, f is different from the linear effect and therefore including temperature as a linear effect is misleading.

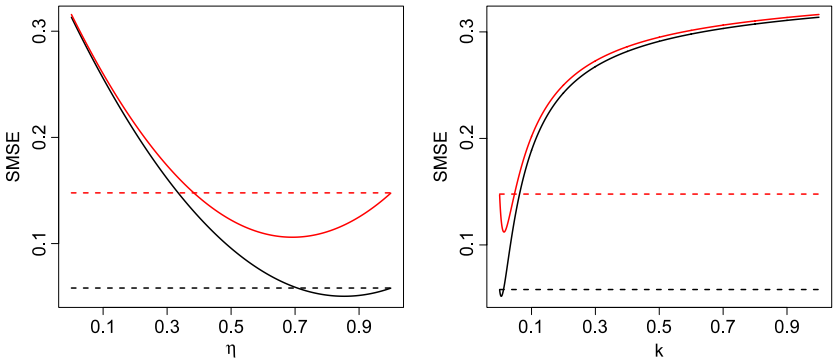


Fig. 2. SMSE of $\widehat{\beta}_{(2)}(\eta)$ in dependence of η (left) and $\widehat{\beta}_{(1)}(k)$ in dependence of k (right) against that of $\widehat{\beta}_{(0)}$ (dashed) under standard assumptions (black) and under generalized assumptions (red). (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

Table 1
Results of OLS, difference based and Liu type difference based estimations.

	$\widehat{\beta}_{OLS}$	$\widehat{\beta}_{(0)}$	$\widehat{\beta}_{(1)}(10^{-3})$	$\widehat{\beta}_{(2)}(0.95)$
x_1	0.634	0.578*	0.550*	0.562*
x_2	−0.152***	−0.160***	−0.158***	−0.161***
x_3	0.030***	0.030*	0.030*	0.030*
x_4	−0.043***	−0.040**	−0.040**	−0.040**
x_5	0.011	0.031	0.031	0.031
x_6	−0.051**	−0.014	−0.013	−0.014
x_7	−0.054*	−0.014	−0.013	−0.014
x_8	−0.079**	−0.065	−0.064	−0.065
x_9	−0.036	−0.037	−0.036	−0.037
x_{10}	−0.052	−0.044	−0.043	−0.044
x_{11}	−0.049	−0.013	−0.012	−0.013
x_{12}	−0.000	0.040	0.040	0.040
x_{13}	−0.001	0.016	0.016	0.016
t	$-13 \cdot 10^{-5}$***	–	–	–
R^2	0.729	0.749	0.749	0.749

* Indicates significance on 10%.
** Indicates significance on 5%.
*** Indicates significance on 1%.

Table 2
Standard errors of the estimators in comparison to Newey–West standard errors for the effects of x_1 (income) and x_2 (relative price).

$\widehat{\Omega}$	$\widehat{\beta}_{(0)}$		$\widehat{\beta}_{(1)}(10^{-3})$		$\widehat{\beta}_{(2)}(0.95)$	
	$\widehat{\sigma}^2 I$	$\widehat{\Omega}_{NW}$	$\widehat{\sigma}^2 I$	$\widehat{\Omega}_{NW}$	$\widehat{\sigma}^2 I$	$\widehat{\Omega}_{NW}$
x_1	0.215	**0.347**	0.209	**0.337**	0.205	**0.215**
x_2	0.034	**0.047**	0.034	**0.047**	0.034	**0.034**
SMSE	0.058	**0.148**	0.056	**0.141**	0.054	**0.058**

8. Conclusion

We proposed a difference based Liu type estimator and a difference based ridge regression estimator for the partial linear semiparametric regression model.

The results show that in case of multicollinearity, the proposed estimator, $\widehat{\beta}_{(2)}(\eta)$ is superior to the difference based estimator $\widehat{\beta}_{(0)}$. We gave bounds on the value of η which ensure the superiority of the proposed estimator. The two biased estimators $\widehat{\beta}_{(2)}(\eta)$ and $\widehat{\beta}_{(1)}(k)$ for different values of η and k can be compared in terms of MSEM with the theoretical results above.

Finally, an application to electricity consumption has been provided to show properties of the proposed estimator based on the mean square error criterion. We could estimate the linear effects of the linear determinants as well as the nonparametric effect f of a cumulated average temperature index.

Thus, the theoretical results obtained allow us to tackle the problem of multicollinearity in real applications of semiparametric models. Moreover, we are able to get estimators of the linear effects with lower standard errors by tuning parameters k and η accordingly.

Table 3
Admissible biasing parameters η and k marked by plus if they satisfy conditions of Theorems 5.1 and 5.2, i.e. $\widehat{\beta}_{(2)}(\eta)$ is superior to $\widehat{\beta}_{(1)}(k)$.

$\eta \cdot 10^2$	$k \cdot 10^4$												
	1	2	3	4	5	6	7	8	9	10	11	12	13
9.23–9.23	−	−	−	−	−	−	−	−	−	−	−	−	−
9.24–9.24	+	−	−	−	−	−	−	−	−	−	−	−	−
9.25–9.25	+	+	−	−	−	−	−	−	−	−	−	−	−
9.26–9.26	+	+	+	−	−	−	−	−	−	−	−	−	−
9.27–9.27	+	+	+	+	−	−	−	−	−	−	−	−	−
9.28–9.28	+	+	+	+	+	−	−	−	−	−	−	−	−
9.29–9.30	+	+	+	+	+	+	−	−	−	−	−	−	−
9.31–9.31	+	+	+	+	+	+	+	−	−	−	−	−	−
9.32–9.32	+	+	+	+	+	+	+	+	−	−	−	−	−
9.34–9.35	+	+	+	+	+	+	+	+	+	−	−	−	−
9.36–9.37	+	+	+	+	+	+	+	+	+	+	−	−	−
9.38–9.39	+	+	+	+	+	+	+	+	+	+	+	−	−
9.40–9.43	+	+	+	+	+	+	+	+	+	+	+	+	−
9.44–9.56	+	+	+	+	+	+	+	+	+	+	+	+	+
9.57–9.61	+	+	+	+	+	+	+	+	+	+	+	+	−
9.62–9.65	+	+	+	+	+	+	+	+	+	+	+	−	−
9.66–9.69	+	+	+	+	+	+	+	+	+	+	−	−	−
9.70–9.72	+	+	+	+	+	+	+	+	+	−	−	−	−
9.73–9.76	+	+	+	+	+	+	+	+	−	−	−	−	−
9.77–9.79	+	+	+	+	+	+	+	−	−	−	−	−	−
9.80–9.82	+	+	+	+	+	+	−	−	−	−	−	−	−
9.83–9.85	+	+	+	+	+	−	−	−	−	−	−	−	−
9.86–9.88	+	+	+	+	−	−	−	−	−	−	−	−	−
9.89–9.91	+	+	+	−	−	−	−	−	−	−	−	−	−
9.92–9.94	+	+	−	−	−	−	−	−	−	−	−	−	−
9.95–9.97	+	−	−	−	−	−	−	−	−	−	−	−	−
9.98–9.99	−	−	−	−	−	−	−	−	−	−	−	−	

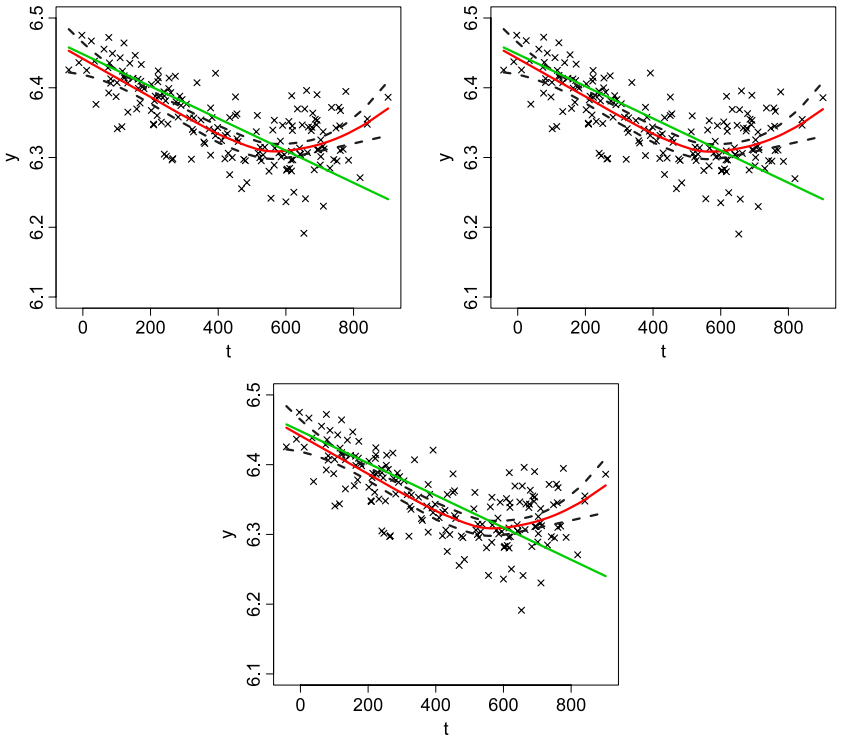


Fig. 3. Estimated f nonlinear effect of t on y via differenced based (left), Liu-type differenced based (right) and difference-based ridge (center) approaches.

References

[1] F. Akdeniz, S. Kaçiranlar, On the almost unbiased generalized Liu estimator and unbiased estimation of the bias and MSE, Communications in Statistics Theory and Methods 24 (7) (1995) 1789–1797.
[2] D. Belsley, E. Kuh, R. Welsch, Regression Diagnostics, Wiley, New York, 1980.

[3] L. Brown, M. Levine, Variance estimation in nonparametric regression via the difference sequence method, Annals of Statistics 35 (2007) 2219–2232.
[4] R.F. Engle, C. Granger, J. Rice, A. Weiss, Semiparametric estimates of the relation between weather and electricity sales, Journal of American Statistical Association 81 (1986) 310–320.
[5] R. Eubank, Nonparametric Regression and Spline Smoothing, Marcel Dekker, New York, 1999.
[6] R. Eubank, E. Kambour, J. Kim, K. Klipple, C. Reese, M. Schimek, Kernel smoothing in partial linear models, Journal of the Royal Statistical Society Series B 50 (3) (1988) 413–436.
[7] R. Eubank, E. Kambour, J. Kim, K. Klipple, C. Reese, M. Schimek, Estimation in partially linear models, Computational Statistics and Data Analysis 29 (1998) 27–34.
[8] J. Fan, L. Huang, Goodness-of-fit tests for parametric regression models, Journal of American Statistical Association 96 (2001) 640–652.
[9] J. Fan, Y. Wu, Semiparametric estimation of covariance matrices for longitudinal data, Journal of American Statistical Association 103 (2008) 1520–1533.
[10] R. Farebrother, Further results on the mean square error of ridge regression, Journal of the Royal Statistical Society Series B 38 (1976) 248–250.
[11] F. Graybill, Matrices with Applications in Statistics, Duxbury Classic, 1983.
[12] P. Green, C. Jennison, A. Seheult, Analysis of field experiments by least squares smoothing, Journal of the Royal Statistical Society Series B 47 (1985) 299–315.
[13] M. Gruber, Improving Efficiency by Shrinkage: The James–Stein and Ridge Regression Estimators, Marcell Dekker, Inc., New York, 1985
[14] W. Härdle, H. Liang, J. Gao, Partially Linear Models, Physika Verlag, Heidelberg, 2000.
[15] W. Härdle, M. Müller, S. Sperlich, A. Werwatz, Nonparametric and Semiparametric Models, Springer Verlag, Heidelberg, 2004.
[16] D. Harville, Matrix Algebra from a Statistician's Perspective, Springer Verlag, New York, 1997.
[17] A. Hoerl, R. Kennard, Ridge regression:biased estimation for orthogonal problems, Technometrics 12 (1970) 55–67.
[18] M. Hubert, P. Wijekoon, Improvement of the Liu estimation in linear regression model, Statistical Papers 47 (3) (2006) 471–479.
[19] K. Klipple, R. Eubank, Difference-based variance estimators for partially linear models, Festschrift in honor of Distinguished Professor Mir Masoom Ali on the occasion of his retirement, 2007, pp. 313–323.
[20] K. Liu, A new class of biased estimate in linear regression, Communications in Statistics Theory and Methods 22 (1993) 393–402.
[21] K. Liu, Using Liu type estimator to combat multicollinearity, Communications in Statistics Theory and Methods 32 (5) (2003) 1009–1020.
[22] W. Newey, K. West, A simple, positive semi-definite, heteroskedasticity and autocorrelation consistent covariance matrix, Econometrica 55 (3) (1987) 703–708.
[23] C. Rao, Linear Statistical Inference and its Applications, Wiley, New York, 1973.
[24] D. Ruppert, M. Wand, R. Carroll, R. Gill, Semiparametric Regression, Cambridge University Press, 2003.
[25] J. Schott, Matrix Analysis for Statistics, second ed., Wiley Inc., New Jersey, 2005.
[26] C. Stein, Inadmissibility of the usual estimator for the mean of a multivariate normal distribution, in: Proc. Third Berkeley Symp. Math. Statist. Prob., vol. 1, 1956, pp. 197–206.
[27] G. Tabakan, F. Akdeniz, Difference-based ridge estimator of parameters in partial linear model, Statistical Papers 51 (2010) 357–368.
[28] G. Trenkler, H. Toutenburg, Mean square matrix comparisons between two biased estimators-an overview of recent results, Statistical Papers 31 (1990) 165–179.
[29] H. Yang, J. Xu, An alternative stochastic restricted Liu estimator in linear regression, Statistical Papers 50 (2009) 639–647.
[30] A. Yatchew, An elementary estimator of the partial linear model, Economics Letters 57 (1997) 135–143.
[31] A. Yatchew, Differencing methods in nonparametric regression: simple techniques for the applied econometrician, 1999. http://www.economics.utoronto.ca/yatchew/.
[32] A. Yatchew, Semiparametric Regression for the Applied Econometrician, Cambridge University Press, 2003.
[33] J. You, G. Chen, Y. Zhou, Statistical inference of partially linear regression models with heteroscedastic errors, Journal of Multivariate Analysis 98 (8) (2007) 1539–1557.

Journal of Empirical Finance 19 (2012) 610–625

Contents lists available at SciVerse ScienceDirect

Journal of Empirical Finance

journal homepage: www.elsevier.com/locate/jempfin

Modelling and forecasting liquidity supply using semiparametric factor dynamics ☆

Wolfgang Karl Härdle, Nikolaus Hautsch, Andrija Mihoci *

Humboldt Universität zu Berlin and C.A.S.E. – Center for Applied Statistics and Economics, Spandauer Str. 1, D-10178 Berlin, Germany
Center for Financial Studies (CFS), Frankfurt, Germany

ARTICLE INFO

Article history:
Received 21 April 2010
Received in revised form 21 March 2012
Accepted 2 April 2012
Available online 10 April 2012

JEL classification:
C14
C32
C53
G11

Keywords:
Limit order book
Liquidity risk
Semiparametric model
Factor structure
Prediction

ABSTRACT

We model the dynamics of ask and bid curves in a limit order book market using a dynamic semiparametric factor model. The shape of the curves is captured by a factor structure which is estimated nonparametrically. Corresponding factor loadings are modelled jointly with best bid and best ask quotes using a vector error correction specification. Applying the framework to four stocks traded at the Australian Stock Exchange (ASX) in 2002, we show that the suggested model captures the spatial and temporal dependencies of the limit order book. We find spill-over effects between both sides of the market and provide evidence for short-term quote predictability. Relating the shape of the curves to variables reflecting the current state of the market, we show that the recent liquidity demand has the strongest impact. In an extensive forecasting analysis we show that the model is successful in forecasting the liquidity supply over various time horizons during a trading day. Moreover, it is shown that the model's forecasting power can be used to improve optimal order execution strategies.

1. Introduction

Due to technological progress in the organization of trading systems and exchanges, electronic limit order book trading has become the dominant trading form for equities. Open limit order books provide important information on the current liquidity supply as reflected by the offered price-quantity relationships on both sides of the market. These supply and demand schedules provide valuable information on traders' price expectations in the spirit of the seminal paper by Glosten (1994), reflect the current implied costs of trading as well as demand and supply elasticities. However, while the dynamic behavior of liquidity demand, as reflected by trading intensities and trade sizes, has been already studied in various papers (see, e.g., (Hautsch and Huang (2012) and Brownlees et al. (2009)), the stochastic properties of liquidity supply is still widely unknown. An obvious reason is that liquidity supply is reflected by high-dimensional bid and ask schedules which are not straightforwardly modelled in a dynamic setting. Consequently, it is a widely open question whether and to which extent liquidity supply might be predictable.

☆ This work was supported by the Deutsche Forschungsgemeinschaft via Collaborative Research Center 649 "Ökonomisches Risiko", Humboldt-Universitt zu Berlin, Germany.

* Corresponding author. Tel.: +49 30 2093 5623.
E-mail address: mihociax@cms.hu-berlin.de (A. Mihoci).

0927-5398/$ – see front matter © 2012 Elsevier B.V. All rights reserved.
doi:10.1016/j.jempfin.2012.04.002

The paper's major idea is to capture the shape of high-dimensional ask and bid curves by a lower-dimensional factor structure which is estimated non-parametrically. We propose a dynamic semiparametric factor model where the shape of order schedules is captured by a non-parametric factor structure while the curves' dynamic behavior is driven by time-varying factor loadings. The latter are modelled parametrically employing a vector error correction model (VECM). We show that the model captures the dynamics of high-dimensional order curves very well and is sufficiently parsimonious to produce valuable out-of-sample predictions. Moreover, the schedule of market depth posted around best quotes reveals strong serial dependence and thus is predictable. This structure is induced by the inventory character of order volume which is strongly persistent over time.

By providing empirical evidence on the dynamics and predictability of order book schedules, this paper fills a gap in empirical literature and complements recent (mostly theoretical) work on order splitting and dynamic order submission strategies. For instance, the question of how to reduce the costs of trading by optimally splitting a large order over time (e.g., over the course of a trading day) is of high relevance in financial practice. Obizhaeva and Wang (2005) and Engle and Ferstenberg (2007) analyze optimal splitting strategies whose implementations ultimately require predictions of future liquidity demand and supply. Bertsimas and Lo (1998) and Almgren and Chriss (2000) derive optimal execution strategies by minimizing expected costs of executing, an order in the context of static price impact functions. Optimal execution in a limit order book market is analyzed by Alfonsi et al. (2010). They allow for general shapes of order book curves and derive explicit optimal execution strategies in discrete time. By providing insights into the actual form of order book curves and their dynamic behavior, our results can be used as valuable inputs in theoretical frameworks.

While to the best of our knowledge our study is the first which models the shapes and dynamics of a complete (high-dimensional) order book, there is a substantial body of empirical literature on the dynamics of limit order books and the analysis of traders' order submission strategies, such as, e.g., Biais et al. (1995), Griffiths et al. (2000), Ahn et al. (2001), Ranaldo (2004), Hollifield et al. (2004), Bloomfield et al. (2005), Degryse et al. (2005), Hall and Hautsch (2006, 2007), Large (2007), Hasbrouck and Saar (2009) or Cao et al. (2009).

An important aspect in this literature is to analyze the question of how to optimally balance risks and gains of a trader's decision whether to post a market order or a limit order. As recently illustrated by Chacko et al. (2008), a limit order can be ultimately seen as an American option and transaction costs are rents that a monopolistic market maker extracts from impatient investors who trade via aggressive limit orders or market orders. Consequently, the analysis of liquidity risks (see, e.g., Johnson, (2008), Liu (2009), Garvey and Wu (2009), Goyenko et al. (2009)) and transaction costs (see, e.g. Chacko et al. (2008), Hasbrouck (2009)) are in the central focus of recent literature.

Given the objective to capture not only the volume around the best quotes but also pending quantities more deeply in the book, the underlying problem becomes inherently high-dimensional. A typical graphical snapshot of ask and bid curves for four stocks traded at the Australian Securities Exchange (ASX) in 2002, is given by Fig. 1. The curse of dimensionality applies immediately as soon as time variations of the order curve shapes have to be taken into account. As shown by Fig. 1 and as illustrated in more detail in the sequel of the paper, order volume is not necessarily only concentrated around the best quotes but can be substantially dispersed over a wider range of price levels. This is a typical scenario for moderately liquid markets as that of the ASX. In such a context, the dynamic modelling of all volume levels individually becomes complicate and intractable.

We suggest reducing the high dimensionality of the order book by means of a factor decomposition using the so-called Dynamic Semiparametric Factor Model (DSFM) proposed by Fengler et al. (2007), Brüggemann et al. (2008), Park et al. (2009) and Cao et al. (2009). Accordingly, we model the shape of the book in terms of underlying latent factors which are defined on a grid space around the best ask or bid quotes and can depend on additional explanatory variables capturing, e.g., the state of the market. In order to avoid specific functional forms for the shape of the curves, the factors as well as the corresponding loadings are estimated nonparametrically using B-splines. Then, in a second step, we model the multivariate dynamics of the factor loadings together with the best bid and the best ask price using a VEC specification.

Using this framework we aim answering the following research questions: (i) How many factors are required to model order book curves reasonably well? (ii) What does the shape of the factors look like? (iii) What do the dynamics of the estimated factor loadings look like? (iv) Does there exist evidence for a strong cross-dependence between both sides of the order book? (v) Can quotes be predictable in the short run? (vi) Does the shape of the order book curves depend on past price movements, past trading volume as well past volatility? (vii) How successful is the model in predicting future liquidity supply and can it be used to improve order execution strategies?

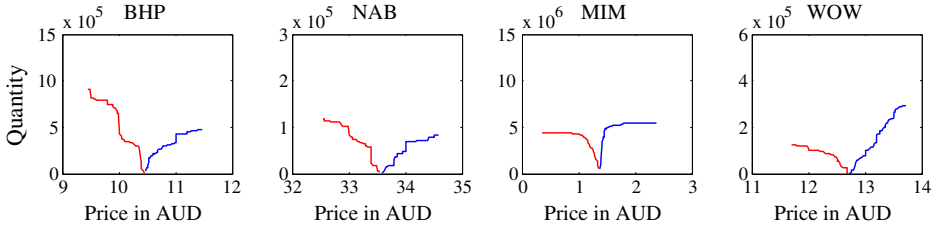


Fig. 1. Limit order books for selected stocks traded at the ASX on July 8, 2002 at 10:15. Red: bid curve, blue: ask curve.

Using limit order book data from four stocks traded at the ASX covering two months in 2002, we show that approximately 95% of the order book variations observed on 5-min intervals can be explained by two underlying time-varying factors. While the first factor captures the overall slope of the curves, the second one is associated with its curvature. Knowing the shape of the order book can help us to predict quotes in the short run. Further empirical results show relatively weak spill-over effects between the bid and the ask side of the market. It turns out that recent liquidity demand represented by the cumulative buy/sell trading observed over the past 5 min has an effect on the shape of the curve but does not induce a higher explanatory power. Similar evidence is shown for the impact of past returns and corresponding (realized) volatility. Moreover, we find that factor loadings follow highly persistent though stationary dynamics.

To evaluate the model's forecasting power, we perform an extensive out-of-sample forecasting analysis which is in line with a typical scenario in financial practice. In particular, at every 5-min interval during a trading day, the model is re-estimated and used to produce forecasts for the pending volume on each price level for all future 5-min intervals during the remainder of the trading day. We show that our approach is able to outperform a naive prediction, where the current order book is used as a predictor for the remaining day. These results can be used to improve intra-day order execution strategies by reducing implied transaction costs.

The remainder of the paper is structured as follows: After the data description in Section 2, the Dynamic Semiparametric Factor Model (DSFM) is introduced in Section 3. Empirical results regarding the modelling and forecasting of liquidity supply are provided in Sections 4 and 5, respectively. Section 6 concludes.

2. Data

2.1. Trading at the ASX and descriptive statistics

The Australian Stock Exchange (ASX) is a continuous double auction electronic market, where the continuous auction trading period is preceded and followed by a call auction. Normal trading takes place continuously on all stocks between 10:09 a.m. and 4:00 p.m. from Monday to Friday. During continuous trading, any buy (sell) order entered that has a price that is greater than (less than) or equal to existing queued buy (sell) orders, will be executed immediately. If an order cannot be executed completely, the remaining volume enters the queues as a limit order. Limit orders are queued in the buy and sell queues according to a strict price-time priority order. Orders can be entered, deleted and modified without restriction.

For order prices below 10 cents, the minimum tick size is 0.1 cents, for order prices above 10 cents and below 50 cents it is 0.5 cents, whereas for orders priced 50 cents and above it is 1 cent. Note that there might be orders which are entered with an undisclosed or hidden volume if the total value of the order exceeds AUD 200,000. Since this applies only to a small fraction of the posted volumes, we can safely neglect the occurrence of hidden volume in our empirical study. For more details on the data, see Hall and Hautsch (2007) using the same data base as well as the official description of the trading rules of the Stock Exchange Automated Trading System (SEATS) on the ASX on www.asxonline.com.

We select four companies traded at the ASX covering the period from July 8 to August 16, 2002 (30 trading days), namely Broken Hill Proprietary Limited (BHP), National Australia Bank Limited (NAB), MIM and Woolworths (WOW). The number of market and limit orders for the selected stocks is given in Table 1.

We observe more buy orders than sell orders implying that the bid side of the limit order book was changing more frequently than the ask side. BHP and NAB are significantly more actively traded than MIM and WOW shares. Aggregated over all stocks, 20.08% (23.98%) of all bid (ask) limit orders have been changed (after posting), whereas 13.70% (14.89%) have been cancelled. Furthermore, for both traded as well as posted quantities we find that on average sell volumes are higher than buy volumes (not reported here). Hence, confirming the result above, liquidity variations on the bid side are higher than that of the ask side. This finding might be explained by the fact that during the analyzed period the market generally went down creating more sell activities than buy activities.

The original dataset contains all limit order book records as well as the corresponding order curves represented by the underlying price-volume combinations. The latter is the particular object of interest for the remainder of the analysis.

Table 1
Total number of market and limit orders for selected stocks traded at the ASX from July 8 to August 16, 2002.

Orders	BHP	NAB	MIM	WOW
Market orders				
(i) Buy	28,030	16,304	4115	7260
(ii) Sell	16,755	15,142	2789	6464
Limit orders				
(i) Buy (bid side)	50,012	28,850	9551	13,234
– Changed	8009	7561	1637	3203
– Cancelled	5202	4725	2044	1951
(ii) Sell (ask side)	32,053	25,953	6474	11,318
– Changed	6891	6261	1862	3164
– Cancelled	4692	3863	1178	1554

2.2. Notation and data preprocessing

The underlying limit order book data contains identification attributes regarding $r=1,\ldots,R$ different orders as well as quantities demanded and offered for different price levels $j=1,\ldots,J$, at any time point $t=1,\ldots,T$. Particularly, at any t, we observe $J=101$ price levels on a fixed minimum tick size grid originating from the best bid and ask quote.

Since the order book dynamics are found to be very persistent, we choose a sampling frequency of 5 min without losing too much information on the liquidity supply. To remove effects due to market opening and closure, the first 15 min and last 5 min are discarded. Hence, at each trading day, starting at 10:15 and ending at 15:55, we select per stock 69 price-quantity vectors, in total $T=2070$ vectors over the whole sample period. Denote $\tilde{Y}^b_{t,j}$ and $\tilde{Y}^a_{t,j}$ as the pending bid and ask volumes at bid and ask limit prices $\tilde{S}^b_{t,j}$ and $\tilde{S}^a_{t,j}$, respectively at time point t.

We define the best bid price at time t as the highest buy price $\tilde{S}^b_{t,101}$, and similarly, the best ask price at t as the lowest sell price $\tilde{S}^a_{t,1}$. The corresponding quantities at best bid and ask prices are then $\tilde{Y}^b_{t,101}$ and $\tilde{Y}^a_{t,1}$, respectively, yielding the mid-quote price to be defined as $\tilde{S}^*_t=\left(\tilde{S}^b_{t,101}+\tilde{S}^a_{t,1}\right)/2$. The absolute price deviations from the best bid and ask price at level j and time t are given by $\breve{S}^b_{t,j}=\tilde{S}^b_{t,j}-\tilde{S}^b_{t,101}$ and $\breve{S}^a_{t,j}=\tilde{S}^a_{t,j}-\tilde{S}^a_{t,1}$, respectively and constitute a fixed price grid. To measure spreads between individual price levels in *relative* terms, i.e., in relation to the prevailing best bid and ask price, we define so-called 'relative price levels' as $S^b_{t,j}=\breve{S}^b_{t,j}/\tilde{S}^b_{t,101}$ and $S^a_{t,j}=\breve{S}^a_{t,j}/\tilde{S}^a_{t,1}$, respectively.

In order to investigate to which extent order book information might reveal information to predict high-frequency returns, we regress 1 min and 5 min mid-quote returns, respectively, on lagged order imbalances

$$\tilde{Y}^b_{t-1,j}/\left(\tilde{Y}^b_{t-1,j}+\tilde{Y}^a_{t-1,j}\right)$$

and

$$\tilde{Y}^a_{t-1,j}/\left(\tilde{Y}^b_{t-1,j}+\tilde{Y}^a_{t-1,j}\right),$$

respectively, for $j=1,\ldots,101$. Fig. 2 shows the implied R^2 values in dependence of the number of included imbalance levels. It turns out that order book imbalances indeed reveal short-term predictability. Interestingly, even levels far apart from the market have still distinct prediction power pushing the R^2 to values of approximately 10%. These findings show that the order book itself reveals predictive content for future price movements which could be exploited in trading strategies.

In order to account for intra-day seasonality effects, we adjust the order volumes correspondingly. To avoid to seasonally adjust all individual volume series separately, we assume that the seasonality impact on quoted volumes at all levels is identical and is well captured by the seasonalities in market depth on the best bid and ask levels $\tilde{Y}^b_{t,101}$ and $\tilde{Y}^a_{t,1}$, respectively. Assuming a multiplicative impact of the seasonality factor, the seasonally adjusted quantities are computed for both sides of the market at price level j, and time t as

$$Y^b_{t,j}=\frac{\tilde{Y}^b_{t,j}}{s^b_t} \tag{1}$$

$$Y^a_{t,j}=\frac{\tilde{Y}^a_{t,j}}{s^a_t}, \tag{2}$$

with s^b_t and s^a_t representing the seasonality components at time t for the bid and the ask side, respectively.

The non-stochastic seasonal trend factors s^b_t and s^a_t are specified parametrically using a flexible Fourier series approximation as proposed by Gallant (1981) and are given by

$$s^b_t=\delta^b\cdot\bar{t}+\sum_{m=1}^{M^b}\left\{\delta^b_{c,m}cos(\bar{t}\cdot 2\pi m)+\delta^b_{s,m}sin(\bar{t}\cdot 2\pi m)\right\} \tag{3}$$

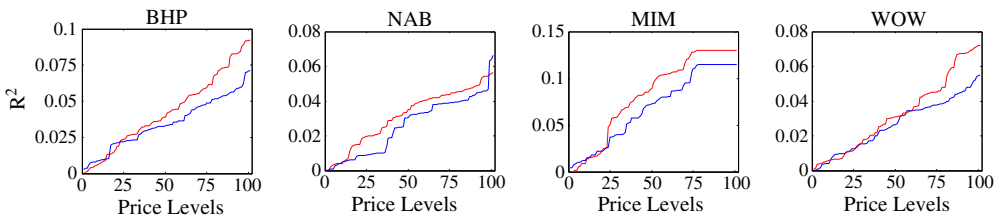


Fig. 2. Coefficients of determination (R^2) implied by linear regression of 1 min (red) and 5 min (blue) mid-quote returns on lagged order imbalances for selected stocks traded at the ASX from July 8 to August 16, 2002 (30 trading days). The horizontal axis depicts the number of included imbalance levels.

$$s_t^a = \delta^a \cdot \bar{t} + \sum_{m=1}^{M^a} \left\{ \delta_{c,m}^a cos(\bar{t} \cdot 2\pi m) + \delta_{s,m}^a sin(\bar{t} \cdot 2\pi m) \right\}. \tag{4}$$

Here δ^b, δ^a, $\delta_{c,m}^b$, $\delta_{c,m}^a$ and $\delta_{s,m}^b$ and $\delta_{s,m}^a$ are coefficients to be estimated, and $\bar{t}$ denotes a normalized time trend mapping the time of the day on a (0,1] intervals. The polynomial orders M^b and M^a are selected according to the Bayes information criterion (BIC). For all stocks we select $M^b = M^a = 1$, except for the bid side for BHP ($M^b = 2$). The resulting intra-day seasonality patterns for both sides of all order book markets are plotted in Fig. 3.

For all stocks, we observe that the liquidity supply increases before market closure. We attribute this finding to traders' pressure and willingness to close positions overnight. Posting aggressive limit orders on the best levels (or even within the spread) maximizes the execution probability and avoids crossing the spread. Moreover, weak evidence for a 'lunch time dip' is presented which, however, is only observed for the more liquid stocks (NAB and BHP). In contrast, for the less liquid stocks, the amount of posted volume nearly monotonically increases over the course of the day.

3. The dynamic semiparametric factor model

Recall that the object of interest is the high-dimensional object of seasonally adjusted level-dependent order volume inventories $\left(Y_{t,j}^b, Y_{t,j}^a\right) \in \mathbb{R}^{202}$, observed on a 5-min frequency. Proposing a suitable statistical model requires finding an appropriate way of reducing the high dimension without losing too much information on the spatial and dynamic structure of the process. Moreover, applicability of the model requires computational tractability as well as numerical stability.

A common way to reduce the dimensionality of multivariate processes is to apply a factor decomposition. The underlying idea is that the high-dimensional process is ideally driven by only a few common factors which contain most underlying information. Factor models are often applied in the asset pricing literature to extract underlying common risk factors. In this spirit, a successful parametric factor model has been proposed, for instance, by Nelson and Siegel (1987) to model yield curves. In this framework, the shape of the curve is parametrically captured by Laguerre polynomials.

Limit order book curves inherently reflect traders' price expectations and the supply and demand in the market (see, e.g. Glosten (1994) for a theoretical framework). As there is no obvious parametric form for ask and bid curves and we want to avoid imposing assumptions on functional form, we prefer to capture the curve's spatial structure in a nonparametric way. A natural and powerful class of models for these kind of problems is the class of Dynamic Semiparametric Factor Models (DSFMs) proposed by Fengler et al. (2007), Brüggemann et al. (2008), Park et al. (2009) and Cao et al. (2009). The DSFM model successfully combines the advantages of a nonparametric approach for cross-sectionally ('spatially') fitting a curve and that of a parametric time series model for modelling persistent multivariate dynamics.

Assume that that the observable J-dimensional random vector, $Y_{t,j}$, can be modelled based on the following orthogonal L-factor model,

$$Y_{t,j} = m_{0,j} + Z_{t,1} m_{1,j} + \cdots + Z_{t,L} m_{L,j} + \varepsilon_{t,j}, \tag{5}$$

where $m(\cdot) = (m_0, m_1, \ldots, m_L)^\top$ denotes the time-invariant factors, a tuple of functions with the property $m_l : \mathbb{R}^d \to \mathbb{R}, l = 0, \ldots, L, Z_t = (1_T, Z_{t,1}, \ldots, Z_{t,L})^\top$ denotes the time series of factor loadings, and $\varepsilon_{t,j}$ represents a white noise error term. The time index is denoted by $t = 1, \ldots, T$, whereas the cross-sectional index is $j = 1, \ldots, J$. Note that this type of factor model is rather restrictive, because it does not take explanatory variables into account.

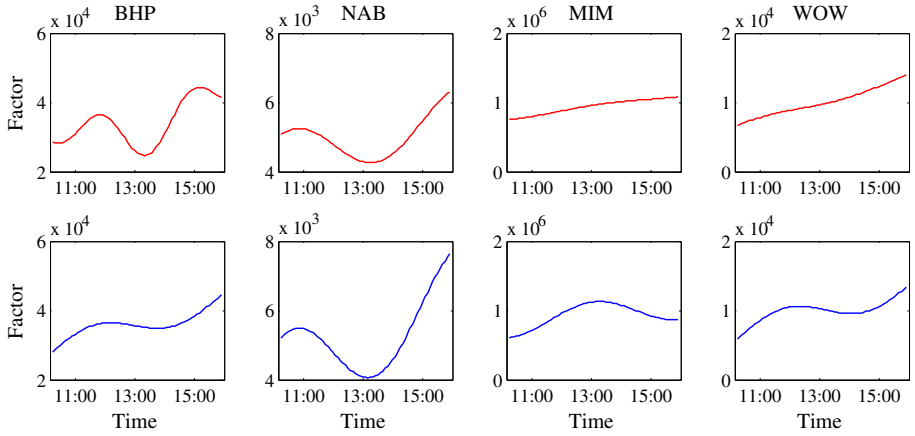


Fig. 3. Estimated intra-day seasonality factors for quantities offered at best bid prices (red) and for quantities supplied at best ask prices (blue) across selected stocks traded at the ASX from July 8 to August 16, 2002 (30 trading days).

The DSFM is a generalization of the factor model given in Eq. (5) and allows the factors m_l to depend upon explanatory variables, $X_{t,j}$. Its analytical form is given by

$$Y_{t,j} = \sum_{l=0}^{L} Z_{t,l} m_l\left(X_{t,j}\right) + \varepsilon_{t,j} = Z_t^{\top} m\left(X_{t,j}\right) + \varepsilon_{t,j}, \tag{6}$$

where the processes $X_{t,j}$, $\varepsilon_{t,j}$ and Z_t are assumed to be independent. Moreover, the number of underlying factors L should not exceed the dimension of the object, J. The main idea of the DSFM is that L is significantly smaller than J resulting in a severe dimension reduction of the process.

As suggested by Park et al. (2009), the estimation of the factors m_l is performed using a series estimator. For $K \geq 1$, appropriate functions $\psi_k : [0,1]^d \to \mathbb{R}, k = 1,\ldots,K$, which are normalized such that $\int \psi_k^2(x)dx = 1$ holds, are selected. Park et al. (2009) select tensor B-spline basis functions for ψ_k, whereas Fengler et al. (2007) use a kernel smoothing approach. In the present study, we follow the former strategy and employ tensor B-spline basis functions.

After selecting the functions ψ_k, the factors $m(\cdot) = (m_0, m_1, \ldots, m_L)^{\top}$ are approximated by $A\psi$, where $A = (a_{l,k}) \in \mathbb{R}^{(L+1)K}$ is a coefficient matrix, and $\psi(\cdot) = (\psi_1, \ldots, \psi_K)^{\top}$ denotes a vector of selected functions. Here, K denotes the number of knots used for the tensor B-spline functions and is interpretable as a bandwidth parameter. Thus, the first part in the right-hand side of (6), which incorporates all factors and factor loadings, can be rewritten as

$$Z_t^{\top} m\left(X_{t,j}\right) = \sum_{l=0}^{L} Z_{t,l} m_l\left(X_{t,j}\right) = \sum_{l=0}^{L} Z_{t,l} \sum_{k=1}^{K} a_{l,k}\psi_k\left(X_{t,j}\right) = Z_t^{\top} A\psi\left(X_{t,j}\right). \tag{7}$$

In modelling liquidity supply we use either the 'relative price levels' on the bid side $S_{t,j}^b$ or those on the ask side $S_{t,j}^a$ as the most important explanatory variable $X_{t,j}$. When focusing on the LOB shape predictability, we add key (weakly exogenous) trading variables, namely the past 5-min aggregated trading volume on both sides of the market, the past 5-min log mid-quote return as well as the past 5-min volatility, see Section 3.

The coefficient matrix A and time series of factor loadings Z_t can be estimated using least squares. Hence, the estimated matrix $\hat{A}$ and factor loadings $\hat{Z}_t = \left(1_T, \hat{Z}_{t,1}, \ldots, \hat{Z}_{t,L}\right)^{\top}$ are defined as minimizers of the sum of squared residuals, $S(A, Z_t)$

$$\left(\hat{Z}_t, \hat{A}\right) = arg \min_{Z_t, A} S(A, Z_t) \tag{8}$$

$$= arg \min_{Z_t, A} \sum_{t=1}^{T} \sum_{j=1}^{J} \left\{Y_{t,j} - Z_t^{\top} A\psi\left(X_{t,j}\right)\right\}^2. \tag{9}$$

To find a solution of the minimization problem stated in Eq. (9), a Newton–Raphson algorithm is used. As shown by Park et al. (2009), this algorithm is shown to converge to a solution at a geometric rate under some weak conditions on the initial choice $\left\{vec(A)^{(0)}, Z_t^{(0)}\right\}$. Moreover, Park et al. (2009) prove that the difference between the estimated loadings $\hat{Z}_t$ and the true loadings Z_t are asymptotically negligible. Consequently, it is justified to use in a second step multivariate time series specifications in order to model the dynamics of the factor loadings. Note that due to the estimation complexity, the coefficients of the seasonal trend factors in Eqs. (1) and (2) are not estimated jointly with the unknown parameters (matrix A) and the factor loadings.

The selection of the number of time-invariant factors (L) and the number of knots K is performed by evaluating the proportion of explained variance (EV) given by

$$EV(L) = 1 - RV(L) = 1 - \frac{\sum_{t=1}^{T} \sum_{j=1}^{J} \left\{Y_{t,j} - \sum_{l=0}^{L} \hat{Z}_{t,l} \hat{m}_l\left(X_{t,j}\right)\right\}^2}{\sum_{t=1}^{T} \sum_{j=1}^{J} \left\{Y_{t,j} - \bar{Y}\right\}^2}. \tag{10}$$

Moreover, the knots used in the tensor B-spline functions should be specified in advance. We choose linearly spaced knots, with a starting point determined by the minimal value of the explanatory variable (corrected by -5%), and the end point corresponding to the maximal value (corrected by 5%). Sensitivity analysis shows that the results are quite stable regarding the choice of grid points.

Because of the use of tensor B-spline functions for the demand and supply curves, which are monotonous in the price levels, our estimated first factor $\hat{m}_1$ and the estimated quantities $\hat{Y}_{t,j}$ are adjusted for extreme price levels. Correspondingly, for the bid side we keep constant the first (lowest) ten level values, and analogously, for the ask side we fix the last (highest) ten level values.

The model's goodness-of-fit is evaluated using the root mean squared error (RMSE) criterion,

$$RMSE = \sqrt{\frac{1}{TJ}\sum_{t=1}^{T}\sum_{j=1}^{J}\left\{Y_{t,j} - \sum_{l=0}^{L}\hat{Z}_{t,l}\hat{m}_l\left(X_{t,j}\right)\right\}^2}. \tag{11}$$

4. Modelling limit order book dynamics

To model order book dynamics we follow a two step procedure for each stock individually. Employing the DSFM approach in the first step, we model the shape of order book curves in dependence of relative price levels. In the following step, the dynamics of the estimated factor loadings is analysed jointly with the best bid quotes, best ask quotes and the bid-ask spread in a parametric multivariate time series context. This procedure allows us to study the cross-dependency between both sides of the market, the interactions between the limit order book and the quotes, as well as the impact of the bid-ask spread on liquidity supply. Moreover, we investigate whether the order book shape itself is predictable by additional covariates, particularly, the past trading volume, past (realized) volatility as well as past log returns.

4.1. Limit order book modelling using the DSFM

We distinguish between two implementation methods of the DSFM:

(i) Separated approach: Separate analysis of both sides of the limit order book, i.e., the bid side $Y^b_{t,j}\in\mathbb{R}^{101}$, and the ask side, $Y^a_{t,j}\in\mathbb{R}^{101}$.
(ii) Combined approach: Simultaneous modelling of both sides of the limit order book with the bid side reversed, i.e. $\left(-Y^b_{t,j}, Y^a_{t,j}\right)\in\mathbb{R}^{202}$.

To model the limit order book in dependence of the relative price levels using the DSFM, i.e., the relative price deviations from the best bid price and best ask price, $S^b_{t,j}$ and $S^a_{t,j}$, respectively, we impose $K=20$ knots for the B-spline functions in case of the separated approach and $K=40$ knots in case of the combined approach. Using more knots does not result in significant improvements of the explained variance or in the corresponding RMSE, as defined in Eqs. (10) and (11).

Empirical results, available from the authors upon request, show that up to approximately 95% of the explained variation in order curves can be explained using $L=2$ factors, whereas the marginal contribution of a potentially third factor is only very small. Consequently, a two-factor DSFM specification is sufficient to capture the curve dynamics and is used in the sequel of the analysis. Furthermore, comparing the performance of the two alternative DSFM specifications, it turns out that in almost all cases the DSFM-separated approach outperforms the DSFM-combined approach in terms of a higher proportion of explained variance and lower values of the root mean squared error. We observe that at almost every price level the DSFM-separated approach outperforms the DSFM-combined approach. Therefore, the remainder of the analysis will rely on the DSFM-separated approach with two factors.

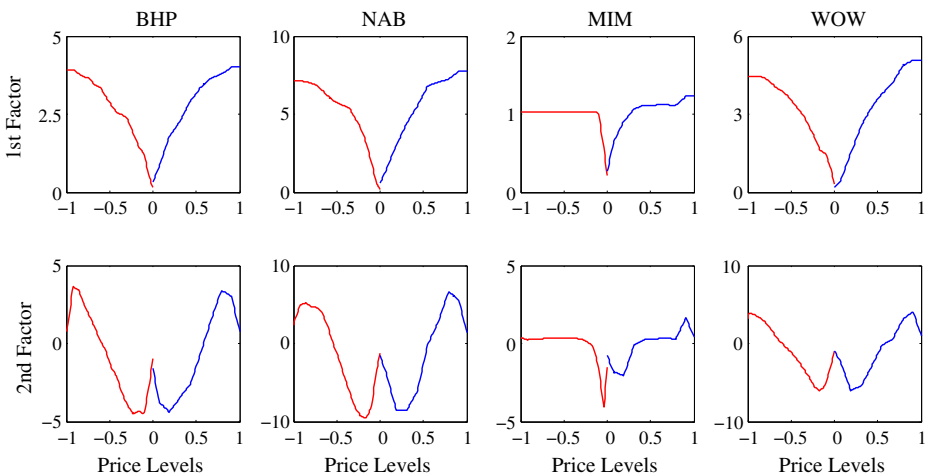


Fig. 4. Estimated first and second factor of the limit order book depending on relative price levels using the DSFM-separated approach with two factors for selected stocks traded at the ASX from July 8 to August 16, 2002 (30 trading days). Red: bid curve, blue: ask curve.

Fig. 4 depicts the nonparametric estimates of the first and second factor $\hat{m}_1$ and $\hat{m}_2$ in dependence of the relative price grids. The first factor obviously captures the overall slope of the curve which is associated with the average trading costs for all volume levels on the corresponding sides of the market. In contrast, the second factor captures order curve fluctuations around the overall slope and thus can be interpreted as a 'curvature' factor in the spirit of Nelson and Siegel (1987). The shape of this factor reveals that the curve's curvature is particularly distinct for levels close to the best quotes and for levels very deep in the book where the curve seems to spread out. The shapes of the estimated factors are remarkably similar for all stocks except for MIM. For the latter stock, the shapes of both factors are quite similar and significantly deviate from those reported for the other stocks. This finding is explained by the peculiarities of MIM for which the relative tick size is larger than for the other stocks. This implies that liquidity is concentrated on relatively few price levels around the best ask and bid quotes whereas the book flattens out for higher levels. This pattern is clearly revealed by the corresponding factors shown in Fig. 4.

However, a priori it is unclear whether modelling order book curves based on all 101 price levels is most appropriate in a prediction context. Besides the well-known trade-off between in-sample fit and out-of-sample prediction performance, we also face the difficulty that the predictive information revealed by order book volume might depend on the distance to the best quotes. For instance, if price levels far away from the market may contain information that help predicting books in the future, this information should be taken into account. However, if they contain virtually only noise (e.g., because of stale orders) it would be more optimal to ignore this information in order to extract a more precise factor structure on lower price levels only. Since optimizing this choice in an (out-of-sample) prediction context is tedious and computationally cumbersome, we restrict ourselves to the quite common proceeding of performing model selection based on in-sample information. Accordingly, we evaluate the model implied explained variance when not the full grid of 101 levels but just 25, 50 and 75 levels are employed. It turns out that the explained variance remains widely unchanged with the model fit increasing with the number of incorporated levels. This is particularly important in the context of order books of less liquid stocks. Therefore, we proceed by extracting the factor structure employing the entire book.

Time series plots of the corresponding factor loadings $\hat{Z}_t^b$ and $\hat{Z}_t^a$ are shown in Fig. 5. We observe that the loadings strongly vary over time reflecting time variations in the shape of the book. The series reveal clustering structures indicating a relatively high persistence in the processes. This result is not very surprising given the fact that order book inventories do not change too severely during short time horizons. Observing order book volumes on even higher frequencies than 5 min further increases this persistence, ultimately driving the processes toward unit root processes. Naturally, this behavior is particularly distinct for less frequently traded stocks and less severe for highly active stocks (cf. Hautsch and Huang (2012) for corresponding results for more liquid assets).

The high persistence is confirmed by autocorrelation functions of $\hat{Z}_t^b$ and $\hat{Z}_t^a$ (not shown in the paper) and corresponding unit root and stationarity tests. According to the Schmidt-Phillips test (see Schmidt and Phillips (1992), H_0: unit root) for all processes the null hypothesis of an unit root can be rejected at the 5% significance level (test statistics for all estimated factor loadings are in the range $[-201.53, -53.88]$, whereas the critical value equals -25.20). Conversely, testing the null hypothesis of stationarity using the KPSS test (see Kwiatkowski et al. (1992), H_0: weak stationarity) implies no rejections for the majority of the processes. Nevertheless, in five cases we have to reject stationarity. Finally, to test for possible cointegration between the factor loadings, we perform Johansen (1991) trace test (not shown in the paper) but do not find significant evidence for common stochastic trends underlying the order book.

A graphical illustration for the goodness-of-fit of the model, depicting the estimated vs. the actually observed limit order book curve, would suggest that the model fits the observed curves very well (no illustrations provided here). This is particularly true

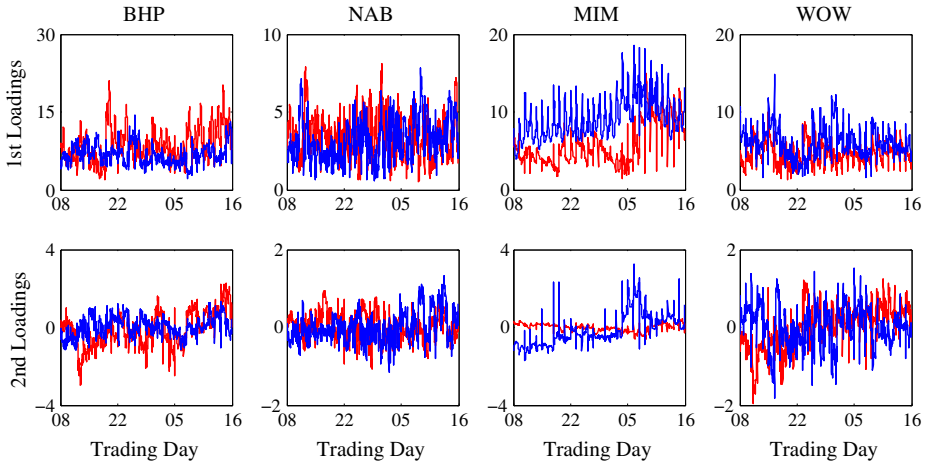


Fig. 5. Estimated first and second factor loadings of the limit order book depending on relative price levels using the DSFM-separated approach with two factors for selected stocks traded at the ASX from July 8 to August 16, 2002 (30 trading days). Red: bid curve, blue: ask curve.

for price levels close to the best ask and bid quotes, at any chosen trading day and stock. Slight deviations are observed for price levels deeply in the book. However, the latter case is less relevant for most applications in practice.

4.2. Modelling limit order book dynamics

Our approach, stipulated under the philosophy *smooth in space and parametric in time*, allows us to investigate the limit order book dynamics in a multivariate time series modelling context, as well as to relate the order book dynamics to the time evolution of additional covariates. Formally, for each stock we focus on the dynamics of the four estimated stationary factor loadings. Including the best bid and the best ask price returns, we consider a (six dimensional) vector of endogenous variables

$$z_t = \left(\hat{Z}^b_{1,t}, \hat{Z}^b_{2,t}, \hat{Z}^a_{1,t}, \hat{Z}^a_{2,t}, \Delta log\tilde{S}^b_{t,101}, \Delta log\tilde{S}^a_{t,1}\right)^\top,$$

where $\hat{Z}^b_{1,t}$, $\hat{Z}^b_{2,t}$, $\hat{Z}^a_{1,t}$ and $\hat{Z}^a_{2,t}$ denote the estimated first (1) and second (2) factor loadings for the bid (b) and ask side (a), respectively. We denote by $\Delta log\tilde{S}^b_{t,101}$ the best bid price return, and similarly, by $\Delta log\tilde{S}^a_{t,1}$ the best ask price return.

Following Engle and Patton (2004) and Hautsch and Huang (2012), the bid-ask spread $\left(log\tilde{S}^b_{t-1,101} - log\tilde{S}^a_{t-1,1}\right)$ serves as a natural cointegration relationship between the two integrated ask and bid series. As all other endogenous variables are shown to be stationary, we obtain a vector error correction (VEC) specification of order q with the spread as the only cointegration relationship, i.e.,

$$z_t = c + \Gamma_1 z_{t-1} + \ldots + \Gamma_q z_{t-q} + \gamma\left(log\tilde{S}^b_{t-1,101} - log\tilde{S}^a_{t-1,1}\right) + \varepsilon_t. \tag{12}$$

Here c denotes a vector with constants, vector $\gamma = (\gamma_1, \ldots, \gamma_6)^\top$ collects parameters associated with the lagged bid-ask spread and ε_t represents a white noise error term. The matrices $\Gamma_1, \Gamma_2, \ldots, \Gamma_q$ are parameter matrices associated with lagged endogenous variables. Technically, we determine the order q according to the BIC.

Estimation results show that in all cases, a maximum lag order of $q = 4$ is sufficient. In particular, the following model orders are selected: BHP and WOW ($q = 3$), NAB ($q = 2$), MIM ($q = 4$). For sake of brevity we refrain from showing all parameter estimates here, but just report the estimates of matrix Γ_1 and vector γ for BHP, NAB, MIM and WOW, respectively, which contain the most relevant information for an economic interpretation (5% significance is denoted by an asterisk (*)):

$$\begin{bmatrix} 0.95^* & 0.63^* & -0.05 & -0.26^* & 3.03 & 18.08 \\ 0.02^* & 0.79 & 0.00 & 0.04 & 10.68 & -16.12 \\ 0.04^* & 0.00 & 0.75^* & 0.02 & -59.60 & 67.60 \\ -0.00 & 0.04 & 0.02^* & 0.77^* & -13.99 & 13.55 \\ 0.00^* & 0.00 & -0.00^* & 0.00^* & -0.59 & 0.29 \\ 0.00^* & 0.00 & -0.00^* & 0.00^* & -0.26 & -0.04 \end{bmatrix}, \begin{bmatrix} -95.70 \\ -34.13 \\ 86.83 \\ -13.21 \\ -0.42 \\ 0.02 \end{bmatrix},$$

$$\begin{bmatrix} 0.71^* & 0.16 & -0.04 & -0.21 & 123.78^* & -124.07^* \\ 0.04^* & 0.78^* & -0.00 & 0.07^* & -22.39^* & 21.91 \\ 0.04 & 0.13 & 0.73^* & 0.18 & -88.56^* & 86.91^* \\ -0.03^* & -0.03 & 0.03^* & 0.71^* & 26.03^* & -25.46^* \\ 0.00^* & 0.00^* & -0.00^* & -0.00 & 0.21 & -0.34 \\ 0.00 & 0.00 & -0.00^* & 0.00 & 0.29 & -0.41 \end{bmatrix}, \begin{bmatrix} -174.41^* \\ 9.26 \\ 47.60 \\ -20.59 \\ -0.85 \\ -0.04 \end{bmatrix},$$

$$\begin{bmatrix} 0.90^* & 1.29^* & -0.00 & 0.55^* & -46.92 & 50.79 \\ 0.00 & 0.93^* & -0.01^* & -0.01 & 1.12 & -1.49 \\ -0.02 & 1.23^* & 0.99^* & 0.48^* & 31.56 & -44.50^* \\ 0.00 & 0.04 & 0.03^* & 0.84^* & 6.73 & -5.89 \\ 0.00 & 0.00 & -0.00 & -0.00 & 0.40 & -0.58 \\ 0.00 & 0.00 & -0.00 & -0.00 & 0.90 & -1.09 \end{bmatrix}, \begin{bmatrix} 62.01^* \\ 0.25 \\ -44.50^* \\ -21.66^* \\ -0.28 \\ -0.18 \end{bmatrix} \textit{and}$$

$$\begin{bmatrix} 0.74^* & -0.02 & 0.12^* & 0.38^* & 28.87 & -37.11 \\ 0.04^* & 0.82^* & -0.02^* & -0.04 & 2.53 & -3.58 \\ 0.04 & 0.03 & 0.87^* & 0.19^* & -70.61^* & 72.84^* \\ -0.03^* & 0.02 & 0.02^* & 0.83^* & 12.81 & -13.70 \\ -0.00^* & -0.00^* & 0.00 & 0.00^* & 0.02 & -0.15 \\ -0.00^* & -0.00^* & 0.00 & 0.00^* & 0.21 & -0.34 \end{bmatrix}, \begin{bmatrix} -27.14 \\ -6.33 \\ 59.98 \\ -4.04 \\ -0.51 \\ 0.05 \end{bmatrix}.$$

The estimation results can be summarised as follows:

Firstly, we observe strong own-process dynamics, but only relatively weak (mostly insignificant) cross-dependencies between the endogenous variables. The latter are most pronounced for less frequently traded stocks (MIM and WOW). Overall, the quite weak inter-dependencies between the processes on the ask and bid side indicate that time variations in the liquidity schedule on the one side is almost unaffected by that on the other side.

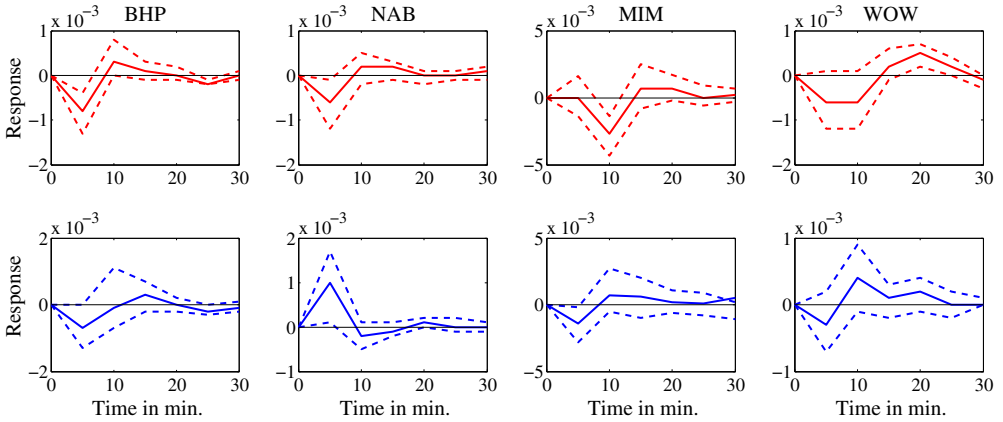


Fig. 6. Orthogonalized impulse-response analysis: responses of the best bid quote return to a one standard deviation shock in the estimated first bid factor loadings (upper panel) and response of the best ask quote return to a one standard deviation shock in the estimated first ask factor loadings (lower panel). We employ the DSFM-separated approach with two factors and a VEC specification for selected stocks traded at the ASX from July 8 to August 16, 2002 (30 trading days). The response variable always enters the VEC specification in the first position. 95% confidence intervals are shown with dashed lines.

Secondly, the major finding is that quote changes are short-run predictable given the shape of the book. More precisely, changes in the factor loading have a short term impact on the quote changes, up to, say, 5–10 min. The impact is significant for the frequently traded stocks, and less severe for less liquid stocks. In particular, a shock on the bid side resulting in upward rotation of the bid curve (inducing a higher sell pressure) leads to an instantaneous decrease in the best bid quote followed by a significant increase of the price within the next few minutes, see, e.g. Fig. 6. This is driven by a growing buy pressure reflected by an increase of bid depth at and behind the market. Fig. 6 depicts the impulse responses of ask and bid quotes driven by a shock in the order book slope. While these effects are quite distinct on the bid side, they are, however, less pronounced on the ask side. A shock on the ask side, however, has a more neutral effect on the price, see, e.g. Fig. 6. However, note that the predictability of quotes only holds over comparably short horizons. Therefore, for daily order execution strategies, as discussed in Section 5, these effects are only of limited use.

Thirdly, we find slight evidence for asymmetric reactions of slope factor loadings on changes of the bid-ask spread. In particular, we observe that rising spreads tend to reduce the order aggressiveness on the bid side while the converse is true on the ask side. Hence, we conclude that as the bid and ask curves move apart, the price is (on average) decreasing. Similarly, as the bid-ask spread shrinks, the price is expected to increase. This re-confirms our finding in Chapter 2, that liquidity variations on the bid side are higher than that of the ask side with more sell activities than buy activities.

4.3. Drivers of the order book shape

In this section, we analyse whether the shape of order book curves is predictable based on key (weakly exogenous) trading variables. We select three variables for which we expect to observe the strongest impact on the book's shape, namely the past 5-min aggregated trading volume on both sides of the market representing the recent liquidity demand, the past 5-min log mid-quote return as well as the past 5-min volatility.

The buy and sell trading volumes at time t are given by the sum of traded quantities from all market orders r, $\tilde{Q}_r^b$ and $\tilde{Q}_r^s$, over 5 min interval, namely, $\tilde{Q}_t^b = \sum_{r=1}^{R_t^b} \tilde{Q}_r^b$ and $\tilde{Q}_t^s = \sum_{r=1}^{R_t^s} \tilde{Q}_r^s$, where R_t^b and R_t^s denote the number of buy and sell orders over the interval $(t-1,t]$, respectively. Correspondingly, log returns r_t and volatility V_t are computed as

$$r_t = log \frac{\tilde{S}_t^*}{\tilde{S}_{t-1}^*} \tag{13}$$

$$V_t = r_t^2, \tag{14}$$

where $\tilde{S}_t^*$ and $\tilde{S}_{t-1}^*$ denote the mid-quotes observed at t and t-1, respectively. Note that the trading volumes as well as the volatility are seasonally adjusted following the procedure explained above. Moreover, the used nonparametric procedure requires the variables to be standardized between -1 and 1. This standardization is performed based on the minimum and maximum observations of the corresponding variables. Finally, as commonly known, nonparametric regression becomes computationally cumbersome for a high number of regressors. To keep our approach computationally tractable and to avoid problems due to the curse of dimensionality, we include the regressors only individually (together with the relative price distances). This ultimately yields a three-dimensional problem.

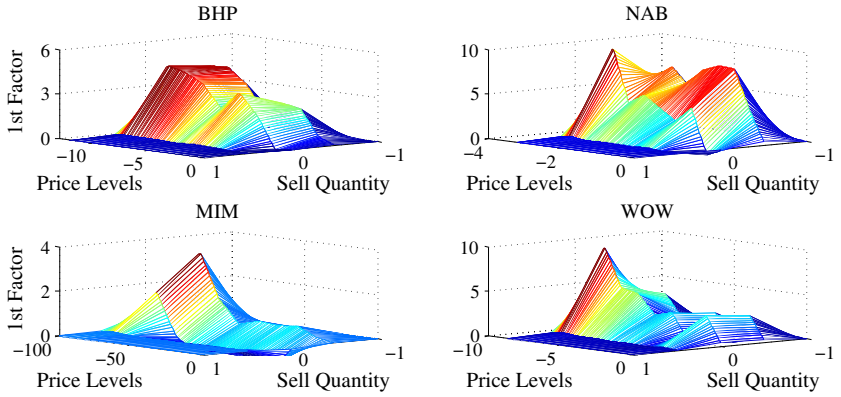


Fig. 7. Estimated first factors of the bid side with respect to relative price levels and the past log traded sell volume using the DSFM-separated approach with two factors for selected stocks traded at the ASX from July 8 to August 16, 2002 (30 trading days).

Figs. 7 and 8 show the estimated first factors for the bid and the ask side in dependence of the past 5-min sell and buy trading volumes, respectively. As expected, we observe that the past liquidity demand influences the order book curve. A high trading volume implies that a non-trivial part of the pending volume in the book is removed. Thus, most of the observed variation of the factor's shape is induced by the fact that either quoted price levels close to the best quotes have been completely absorbed and the remaining volume is correspondingly 'shifted down' in relation to the new best quote or, alternatively, only a part of the pending volume on the best quotes is removed changing the distribution of the pending volumes across the (relative) price levels.

As expected, the curve flattens in the area of high volumes. Strikingly, we also observe a decaying pattern if the volume sizes decline. Actually, in all pictures, the maximum slope (and thus the highest level of liquidity supply) is observed for magnitudes of the standardized volume between −1 and 0, i.e., comparably small (though not zero) trading volumes. This pattern might be technically explained by the standardization procedure based on extreme values or by the usual boundary problems of non-parametric regression. On the other hand, note that due the curse-of-dimensionality problem we cannot simultaneously control for other variables. For instance, very small market-side-specific trading volumes can indicate the occurrence of market imbalances or, alternatively, might be associated with wide spreads. Both scenarios could force investors to post rather limit orders than market orders which might explain the decaying shape of the figures after having observed small trading volumes.

To evaluate whether the inclusion of past trading volume further increases the model's goodness-of-fit, we calculated the corresponding RMSEs. Comparing the results (range from 4.37 to 10.42) with that for the basis model (range from 0.18 to 3.49) shows that the included regressors yield higher estimation errors. Hence, obviously the inclusion of additional regressors ultimately generates more noise overcompensating a possibly higher explanatory power. Similar results are also found for the past log returns and past volatility serving as regressors. The inclusion of log returns yields smaller estimation errors than the inclusion of volatility. However, the overall performance is lower than in the cases above. Because of this reason, we refrain from showing corresponding graphs of the estimated factors.

A possible reason for the declining model performance in case of included regressors might be the lower dimensionality of the regressors in comparison with that of the limit order book. Note that the included regressors do not reveal any variation across

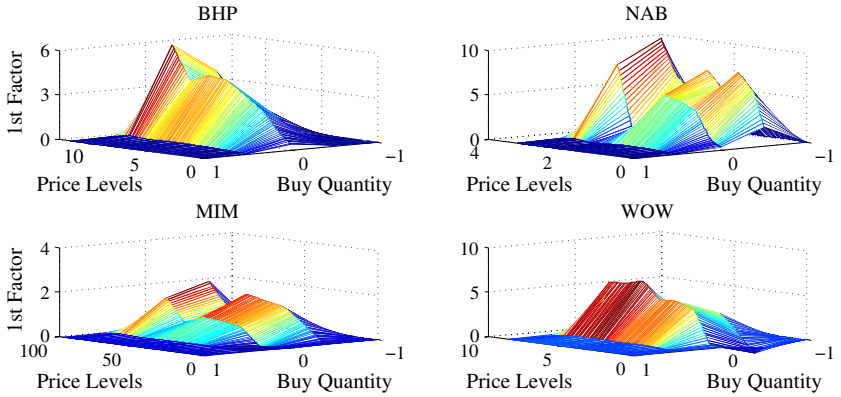


Fig. 8. Estimated first factors of the ask side with respect to relative price levels and the past log traded buy volume using the DSFM-separated approach with two factors for selected stocks traded at the ASX from July 8 to August 16, 2002 (30 trading days).

the levels of the book. Consequently, the explanatory variables cannot improve the model's spatial fit but just its dynamic fit. Obviously, the latter is not sufficient to obtain an overall reduction of estimation errors.

5. Forecasting liquidity supply

5.1. Setup

The aim of this section is to analyse the model's forecasting performance in a realistic setting mimicking the situation in financial applications. We consider an investor observing the limit order book at 5-minute snapshots together with the history over the past 10 trading days. It is assumed that during a trading day an investor updates limit order book every 5 min and requires producing forecasts for all (5 min) intervals of the remainder of the day. Such information might be useful in order to optimally balance order execution during the course of a day. Since we do not exceed beyond the end of the trading day (in order to avoid overnight effects), the forecasting horizon h subsequently declines if we approach market closure. Hence, starting at 10:30, we produce multi-step forecasts for all remaining $h=66$ intervals during the day. Correspondingly, at 15:50, we are left with a horizon of $h=1$. Since quotes themselves – according to our results above – are only predictable over short horizons which are virtually irrelevant for the present analysis, we do not explicitly incorporate this information here.

Consequently, the model is re-estimated every 5 min exploiting past information over a fixed window of 10 trading days (including the recent observation). Due to the length of the estimation period, we do not produce forecasts for the first two weeks of our sample but focus on the period between July 22 and August 16, 2002, thereby covering the period of 20 trading days. In accordance with our in-sample results reported in the previous section, we choose the DSFM-Separated approach based on two factors without additional regressors as underlying specification.

A natural benchmark to evaluate our model is the naive forecast. In this context, we assume that the investor has no appropriate prediction model but just uses the current liquidity supply as a forecast for the remainder of the day. More formally, we suppose that our investor can use the following two approaches in order to forecast liquidity supply $\hat{Y}_{t'+h,j}$ at a given time point t' from July 22 at 10:25 until August 16, 2002, at 15:50, $t'=693,\ldots,2069=T-1$, over a forecasting horizon $1\le h\le 66$, and over the absolute price level j:

(i) DSFM approach: Firstly, the factors and factor loadings are estimated using the DSFM-Separated approach with two factors, $K=20$ knots used for the B-spline basis functions, and with past 690 observed (de-seasonalized) limit order book curves. More precisely, at time point t', relative price levels $S^b_{t'-691:t',j}$ and $S^a_{t'-691:t',j}$ and de-seasonalized observed bid and ask sides $Y^b_{t'-691:t',j}$ and $Y^a_{t'-691:t',j}$ enter the estimation procedures. This yields estimates for the bid (ask) side, 66 times per day for each stock, in total 1320 times over 20 days.
Secondly, since we do not account for (short-term) quote return predictability but only forecast the liquidity supply, we employ a simple 4-dimensional VAR(p) model for the four time-varying factor loadings. When fitted to the entire time series (30 trading days) and according to the BIC, a maximum lag order $p=4$ is sufficient. In particular, the following VAR(p) models are selected: BHP and MIM – VAR(4), NAB – VAR(2), WOW – VAR(3). Using this specifications, we forecast the factor loadings over the forecasting period $\hat{Z}_{t'+h}$. Then, the predicted factor loadings together with the estimated time-invariant factors $\hat{m}_l$ are used to predict the order book.

(ii) Naive approach: Among all historical 690 limit order book curves, only the last one at time t', $(Y^b_{t',j}, Y^a_{t',j})$, is selected as the h-step ahead forecast.

The predictions are evaluated using the root mean squared prediction error (RMSPE), i.e., a version of the in-sample RMSE (11) where the sum over the sampling periods t and the sample size T are replaced by the forecasting horizons h and H, respectively. Since future quotes and relative price grids are not predicted by the model, we assume that quotes themselves follow random walk processes and the spread remains constant. Future quotes are therefore predicted using the current one. Consequently, the predicted future relative price grid remains constant.

5.2. Forecasting results

Fig. 9 shows the RMSPEs for each required forecasting horizon h during a trading day implied by the DSFM as well as the naive model. The following results can be summarized: First, overall the DSFM forecasts outperform the naive ones. Nevertheless, the naive forecast is a serious competitor which is hard to beat. This result is not surprising given the high persistence in liquidity supply. Secondly, the model's forecasting performance is obviously higher on the bid side than on the ask side. This result might be explained by the fact that during the sample period we observe a downward market inducing higher activities on the bid side than on the ask side. This is confirmed by the descriptive statistics shown above. Thirdly, the DSFM outperforms the naive model particularly over horizons up to 1 to 2 hours. For longer horizons, the picture is less clear.

Analyzing average RMSPEs (averaged over all forecasting horizons and both sides of the market) as reported by Table 2 indicate that the overall prediction performance of the DSFM approach is significantly higher than that of the benchmark.

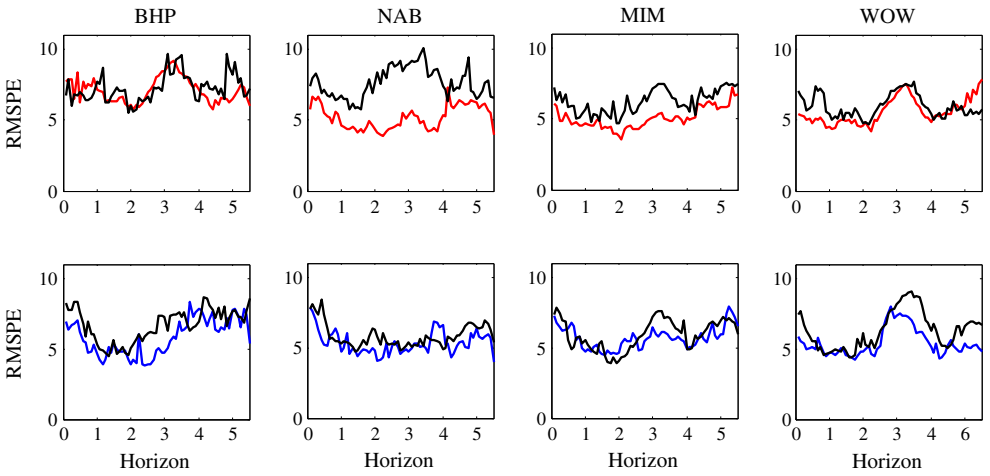


Fig. 9. Root mean squared prediction errors (RMSPEs) implied by the DSFM-separated approach with two factors for the bid side (red) as well as the ask side (blue) and by the naive approach (black) for all intra-day forecasting horizons (in hours) for selected stocks traded at the ASX. Prediction period: July 22 to August 16, 2002 (20 trading days).

5.3. Financial and economic applications

The results in the previous section show that the DSFM is able to successfully predict liquidity supply over various forecasting horizons during a day. In this subsection, we apply these results to two practical examples. The first one is devoted to an order execution strategy, whereas the second one deals with forecasts of demand and supply elasticities.

Example 1. (Trading Strategy)

Suppose an institutional investor decides to buy (sell) a certain number of shares v over the course of a trading day, starting from 10:30 until 15:40. The size of the traded quantity for BHP, NAB and WOW is chosen as to be 5 or 10 times the average pending volume at the best bid (ask) level. In case of MIM, where liquidity supply is much more concentrated at the first level and the book is very thin for higher levels (see the empirical results in the previous sections), we choose the traded volume as being 2 and 5 times the average first level depth. This yields to the following quantities in the respective two cases of high (a) and very high (b) liquidity demand:

(a) BHP – 175,000 shares; NAB – 25,000 shares; WOW – 50,000 shares; MIM – 1,860,000 shares
(b) BHP – 350,000 shares; NAB – 50,000 shares; WOW – 100,000 shares; MIM – 4,650,000 shares.

To reduce the computational burden, we assume that trading is only performed on a 5 min grid throughout the day corresponding to 63 possible trading time points. Moreover, suppose that the investor makes her trading decision at 10:30 but does not monitor the market anymore during the day. Consequently, her forecasting horizon covers $h = 63$ periods at each trading day. Then, she has to decide between two execution strategies:

(i) Splitting the buy (sell) order of size v in a 5 minute frequency proportionally over the trading day resulting into 63 trades of size $v/63$ each.
(ii) Placing orders not proportionally but at those m (5 minute interval) time points throughout the day where the DSFM-based predicted implied trading costs c of the volume v are smallest (among all 63 possible periods). Then, the volume v is split over the m time points according to the relative proportions of expected trading costs. Hence, at interval i, $w_i \cdot v$ shares are traded, with $w_i = c_i / \sum_{j=1}^{m} c_j$ for $i = 1,\ldots,m$.

Table 2
Average root mean squared prediction errors (RMPSEs) of both limit order book sides implied by the DSFM-separated approach with two factors and the naive model for selected stocks traded at the ASX in the period from July 22 to August 6, 2002 (20 forecasting days).

Approach	BID				ASK			
	BHP	NAB	MIM	WOW	BHP	NAB	MIM	WOW
Naive	7.11	7.59	6.03	6.08	6.50	5.96	5.96	6.19
DSFM	7.18	5.10	4.84	5.33	5.56	5.46	5.63	5.45

Strategy (i) can be seen as a special case of strategy (ii) if m is chosen as $m=63$ and the volume v is just equally split. Conversely, for $m=1$, we obtain the extreme case, where the entire quantity is traded once requiring to severely 'walk up' the book. The DSFM predictions of trading costs are computed based on the predicted order book shape at each point and the effective costs to buy or to sell the quantity v while using the ask and bid quotes prevailing at 10:25 (in accordance with the assumption of a random walk). Note that we do not optimize over the quantity underlying the predicted trading costs but just fix it at v corresponding to the maximally possible trade size per time point. Consequently, our strategy selects those trading points where the execution of the entire quantity v is expected to be cheapest and thus covers also the hypothetical (limiting) case of putting all weight w_i on a single point implying a 'one-shot' execution. Of course, an even more sophisticated (and optimized) strategy would require the prediction of trading costs for relative proportions of v which are themselves simultaneously optimized. However, this would substantially increase the numerical and computational burden and is beyond the scope of the current study.

To implement these strategies, we consider 20 forecasting days covering the period from July 22 to August 16, 2002. Fig. 10 shows the average percentage reduction in trading costs of strategy (ii) in excess of the equal-splitting ('naive') strategy (i) for various choices of $m\in[1,63]$. In most cases we observe that a strategic placing of orders according to DSFM predictions yield excess gains of approximately 10 basis points on average. Overall, the selling strategies are more beneficial than the buying strategies confirming the findings on prediction errors above. This is most striking for BHP where we observe a significant difference between sell-based and buy-based profits if the number of trading points are low. Apart from this observation we find a generally non-monotonic behavior of the curves implying losses if m is small, increasing (and positive) gains for a higher number of trading points and a convergence to zero for m reaching the upper limit of 63. This pattern indicates that trading the daily position using only a few large market orders is inferior to an equal-splitting strategy as the underlying transactions have to walk up the book too severely and cause huge price impacts. For higher values of m, the strategic placement according to DSFM predictions become profitable where in the limit of $m=63$, relative benefits are only due to a strategic (non-equal) weighting scheme. However, for MIM we observe a significantly different pattern implying the highest gains for m being small and nearly monotonically declining profits if m is increasing. This pattern is obviously induced by the peculiarities of the MIM order book which is extremely deep on the first level and makes 'one-shot' executions of large volumes quite beneficial. Overall, the patterns are very similar for the two classes of daily quantities, where as expected the relative gains become smaller with higher traded daily volume.

Overall, our findings indicate that the model is successful in predicting times where the market is sufficiently deep in order to execute a large orders. The fact that the model performs reasonably well is promising for more elaborate practical applications of the DSFM. Moreover, note that the reported results are valid under the assumption that there are no transaction fees. Actually, in practice, a proportional splitting strategy induces higher transaction costs as a complete execution via a market order. This component is not taken into account here and would even increase the performance of the DSFM-based execution strategy. Finally, predictions of trading costs could be further improved by exploiting possible predictive information of the limit order book for future returns. Our descriptive statistics reported above show that order book imbalances have indeed (slight) prediction power. We will leave these issues, however, for future research.

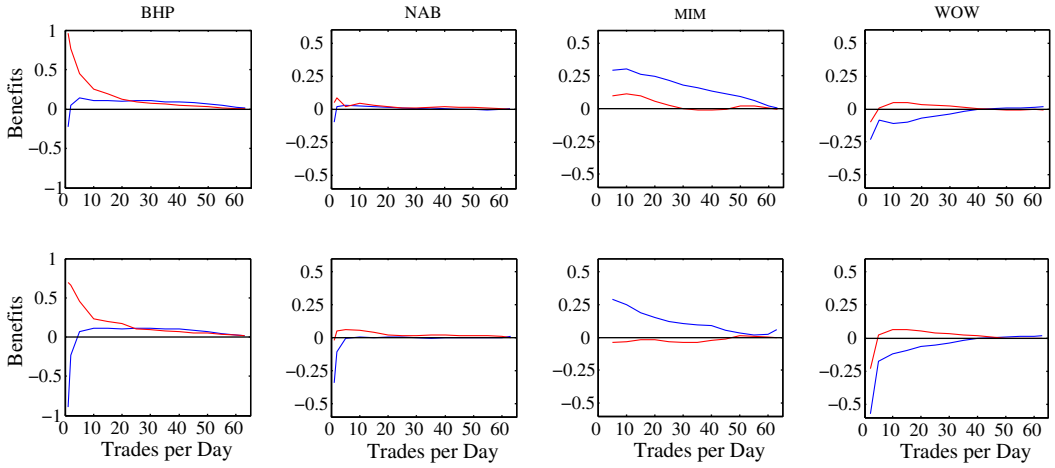


Fig. 10. Average percentage gains by reduced transaction costs compared to an equal-splitting strategy when buying (blue) and selling (red) shares based on m DSFM-predicted time points per day. Upper panel: Daily volumes corresponding to 5 (2) times the average first level market depth for BHP, NAB, WOW (MIM). Lower panel: Daily volumes corresponding to 10 (5) times the average first level market depth for BHP, NAB, WOW (MIM). Prediction period: July 22 to August 16, 2002 (20 trading days).

Example 2. (Demand and Supply Elasticity)

A straightforward dimension-less measure for the order book slope is the curve's elasticity which we compute at best bid ($\tilde{S}^b_{t',101}$) and best ask prices ($\tilde{S}^a_{t',1}$) as

$$\hat{E}^d_{t'+h} = \frac{\hat{Y}^b_{t'+h,1} - \hat{Y}^b_{t'+h,101}}{\hat{Y}^b_{t'+h,101}} \Big/ \frac{\tilde{S}^b_{t',1} - \tilde{S}^b_{t',101}}{\tilde{S}^b_{t',101}}, \tag{15}$$

$$\hat{E}^s_{t'+h} = \frac{\hat{Y}^a_{t'+h,101} - \hat{Y}^a_{t'+h,1}}{\hat{Y}^a_{t'+h,1}} \Big/ \frac{\tilde{S}^a_{t',101} - \tilde{S}^a_{t',1}}{\tilde{S}^a_{t',1}}, \tag{16}$$

for the demand (bid) and supply (ask) side, respectively. The elasticity is a measure for the marginal trading costs reflecting the curve's curvature.

Suppose at 10:30 an investor aims to predict the demand and supply elasticity at best bid and best ask prices for all 5-min intervals during the trading day covering the forecast horizons $h = 1,\ldots,66$. As above, the forecasts are computed using the last 10 trading days. Since we are not forecasting the price process, the last observed ask and bid quotes are used for prediction. Fig. 11 shows the 10:30 predictions of demand and supply elasticities at best bid and best ask prices during all trading days. We observe that marginal trading costs exhibit significant variations over time. The fact that predicted elasticities reveal temporarily trending patterns might be used for improving trading strategies.

Consider the case of NAB on July 24 and July 30, 2002. We observe that the demand elasticities (in absolute terms) are increasing on the first day, and decreasing on the second day. Practically, it would be better to sell shares late on July 24, and early on July 30, under the assumption that the price does not change significantly over both trading days. The supply elasticities show converse patterns across the days. Consequently, it would be advisable to buy shares early on July 24, and late on July 30, provided that the prices remain unchanged. While this section aims to illustrate possible applications of the DSFM approach, more detailed elaborations of dynamic strategies are beyond the scope of the paper.

6. Conclusions

In this paper, we propose a dynamic semiparametric factor model (DSFM) for modelling and forecasting liquidity supply. The main idea of the DSFM as proposed by Brüggemann et al. (2008), Cao et al. (2009), Fengler et al. (2007) and Park et al. (2009) is to capture the order curve's spatial structure across various relative distances to the best quotes using a factor decomposition which is estimated nonparametrically. To capture the order book's time variations, the corresponding factor loadings are modelled using a multivariate time series model. The framework is flexible though parsimonious and turns out to provide a powerful way to reduce the high dimension of the book and to extract the relevant underlying information regarding order book dynamics.

The model is applied to four stocks traded at the Australian Stock Exchange (ASX). It is shown that two underlying factors can explain up to 95% of in-sample variations of ask and bid liquidity supply. While the first factor captures the overall order curve's slope, the second factor is associated with the curve's curvature. The extracted factor loadings reveal highly persistent though weakly stationary dynamics which are successfully captured by a vector error correction specification. We find relatively weak spill-over effects between both sides of the limit order book sides that are more pronounced for less liquid stocks compared to high frequently traded ones. It is shown that order book shapes have short-term prediction power for quote changes. Furthermore, we show that the

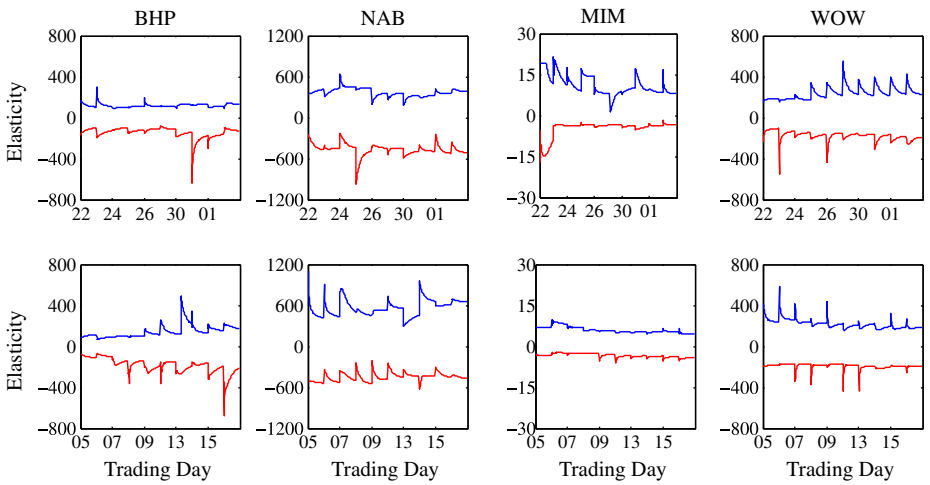


Fig. 11. Predicted demand and supply elasticities at best bid (red) and best ask prices (blue) using the DSFM-separated approach with two factors for selected stocks traded at the ASX from July 22 to August 2, 2002 (upper panels, 10 trading days) and from August 5 to August 16, 2002 (lower panels, 10 trading days).

order curves' shapes are driven by explanatory variables reflecting the recent liquidity demand, volatility as well as mid-quote returns.

Employing the DSFM approach in an extensive and realistic out-of-sample forecasting exercise we show that the model successfully predicts the liquidity supply over various forecasting horizons during a trading day and outperforms a naive approach. Using the forecasting results in a trading strategy it is shown that order execution costs can be reduced if orders are optimally placed according to predictions of liquidity supply. In particular, it turns out that optimal order placement in periods of high liquidity results in smaller transaction costs than in the case of a proportional splitting over time. Finally, our flexible approach allows us to estimate and to predict future (excess) demand and supply elasticities.

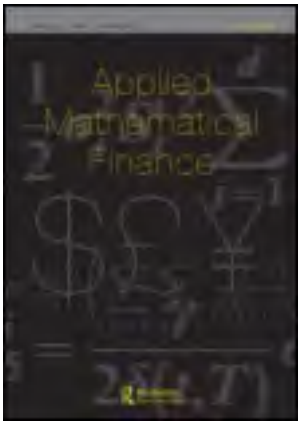
These results show that the DSFM approach is suitable for modelling and forecasting liquidity supply. Since it is computationally tractable, it can serve as a valuable building block for automated trading models.

Acknowledgements

We are very grateful to Anthony Hall for providing us the data. For helpful comments and discussions we thank Jean-Philippe Bouchaud, Joachim Grammig, Jeffrey Russell and the participants of the 2009 Humboldt-Copenhagen Conference on Financial Econometrics in Berlin, the 2009 annual conference of the Society for Financial Econometrics (SoFiE) in Geneva, the 2009 European Meeting of the Econometric Society in Barcelona, the International Conference on Price, Liquidity and Credit Risk in Konstanz, 2008, as well as the 4th World Congress of the International Association for Statistical Computing in Yokohama, 2008. Furthermore, we are grateful to Szymon Borak for helping us with the implementation of the Dynamic Semiparametric Factor Model in MATLAB. Finally, we thank the reviewers for their constructive comments and helpful suggestions.

References

Ahn, H.J., Bae, K.H., Chan, K., 2001. Limit orders, depth, and volatility: evidence from the stock exchange of Hong Kong. Journal of Finance 56, 767–788.
Alfonsi, A., Fruth, A., Schied, A., 2010. Optimal execution strategies in limit order books with general shape functions. Quantitative Finance 10, 143–157.
Almgren, R., Chriss, N., 2000. Optimal execution of portfolio transactions. Journal of Risk 3, 5–39.
Bertsimas, D., Lo, A.W., 1998. Optimal control of execution costs. Journal of Financial Markets 1, 1–50.
Biais, B., Hillion, P., Spatt, C., 1995. An empirical analysis of the limit order book and the order flow in the Paris bourse. Journal of Finance 50, 1655–1689.
Bloomfield, R., O'Hara, M., Saar, G., 2005. The "make or take" decision in an electronic market: evidence on the evolution of liquidity. Journal of Financial Economics 75, 165–200.
Brownlees, C.T., Cipollini, F., Giampiero, M.G., 2009. Intra-daily Volume Modeling and Prediction for Algorithmic Trading. Discussion paper, Stern School of Business. .
Brüggemann, R., Härdle, W., Mungo, J., Trenkler, C., 2008. VAR modelling for dynamic semiparametric factors of volatility strings. Journal of Financial Econometrics 5 (2), 189–218.
Cao, J., Härdle, W., Mungo, J., 2009. A Joint Analysis of the KOSPI 200 Option and ODAX Option Markets Dynamics. Discussion Paper 019, Collaborative Research Center 649 "Economic Risk". Humboldt-Universität zu Berlin.
Chacko, G.C., Jurek, J.W., Stafford, E., 2008. The price of immediacy. Journal of Finance 63, 1253–1290.
Degryse, H., Jong, F., Ravenswaaij, M., Wuyts, G., 2005. Aggressive orders and the resiliency of a limit order market. Review of Finance 9, 201–242.
Engle, R.F., Ferstenberg, R., 2007. Execution risk. Journal of Portfolio Management 33, 34–45.
Engle, R.F., Patton, A.J., 2004. Impacts of trades in an error-correction model of quote prices. Journal of Financial Markets 7, 1–25.
Fengler, M.R., Härdle, W., Mammen, E., 2007. A dynamic semiparametric factor model for implied volatility string dynamics. Journal of Financial Econometrics 5 (2), 189–218.
Gallant, A.R., 1981. On the bias of flexible functional forms and an essentially unbiased form. Journal of Econometrics 15, 211–245.
Garvey, R., Wu, F., 2009. Intraday time and order execution quality dimensions. Journal of Financial Markets 12, 203–228.
Glosten, L., 1994. Is the electronic Limit Order Book inevitable. Journal of Finance 49 (4), 1127–1161.
Goyenko, R.Y., Holden, C.W., Trzcinka, C.A., 2009. Do liquidity measures measure liquidity? Journal of Financial Economics 92, 153–181.
Griffiths, M.D., Smith, B.F., Turnbull, D.A.S., White, R.W., 2000. The costs and determinants of order aggressiveness. Journal of Financial Economics 56, 65–88.
Hall, A.D., Hautsch, N., 2006. Order aggressiveness and order book dynamics. Empirical Economics 30, 973–1005.
Hall, A.D., Hautsch, N., 2007. Modelling the buy and sell intensity in a limit order book market. Journal of Financial Markets 10 (3), 249–286.
Hasbrouck, J., 2009. Trading costs and returns for U.S. equities: estimating effective costs from daily data. Journal of Finance 64, 1445–1477.
Hasbrouck, J., Saar, G., 2009. Technology and liquidity provision: the blurring of traditional definitions. Journal of Financial Markets 12, 143–172.
Hautsch, N., Huang, R., 2012. The market impact of a limit order. Journal of Economic Dynamics & Control 36, 501–522.
Hollifield, B., Miller, R.A., Sandås, P., 2004. Empirical analysis of limit order markets. Review of Economic Studies 71, 1027–1063.
Johansen, S., 1991. Estimation and hypothesis testing of cointegration vectors in gaussian vector autoregressive models. Econometrica 59 (6), 1551–1580.
Johnson, T.C., 2008. Volume, liquidity and liquidity risk. Journal of Financial Economics 87, 388–417.
Kwiatkowski, D., Phillips, P.C.B., Schmidt, P., Shin, Y., 1992. Testing the null hypothesis of stationarity against the alternative of a unit root: how sure are we that economic time series have a unit root? Journal of Econometrics 54, 159–178.
Large, J., 2007. Measuring the resiliency of an electronic limit order book. Journal of Financial Markets 10, 1–25.
Liu, W.-M., 2009. Monitoring and limit order submission risks. Journal of Financial Markets 12, 107–141.
Nelson, C.R., Siegel, A.F., 1987. Parsimonious modelling of yield curves. Journal of Business 60, 473–489.
Obizhaeva, A., Wang, J., 2005. Optimal Trading Strategy and Supply/Demand Dynamics. Working Paper 11444, NBER.
Park, B., Mammen, E., Härdle, W., Borak, S., 2009. Time series modelling with semiparametric factor dynamics. Journal of the American Statistical Association 104 (485), 284–298.
Ranaldo, A., 2004. Order aggressiveness in limit order book markets. Journal of Financial Markets 7, 53–74.
Schmidt, P., Phillips, P.C.B., 1992. LM tests for a unit root in the presence of deterministic trends. Oxford Bulletin of Economics and Statistics 54, 257–287.



Applied Mathematical Finance
Publication details, including instructions for authors and subscription information:
http://www.tandfonline.com/loi/ramf20

The Implied Market Price of Weather Risk

Wolfgang Karl Härdle [a] & Brenda López Cabrera [a]
[a] Ladislaus von Bortkiewicz Chair of Statistics, Humboldt University of Berlin, Berlin, Germany

Available online: 17 Oct 2011

To cite this article: Wolfgang Karl Härdle & Brenda López Cabrera (2012): The Implied Market Price of Weather Risk, Applied Mathematical Finance, 19:1, 59-95

To link to this article: http://dx.doi.org/10.1080/1350486X.2011.591170

Applied Mathematical Finance,
Vol. 19, No. 1, 59–95, February 2012

The Implied Market Price of Weather Risk

WOLFGANG KARL HÄRDLE & BRENDA LÓPEZ CABRERA

Ladislaus von Bortkiewicz Chair of Statistics, Humboldt University of Berlin, Berlin, Germany

(Received 7 May 2010; in revised form 28 February 2011)

ABSTRACT *Weather derivatives (WD) are end-products of a process known as securitization that transforms non-tradable risk factors (weather) into tradable financial assets. For pricing and hedging non-tradable assets, one essentially needs to incorporate the market price of risk (MPR), which is an important parameter of the associated equivalent martingale measure (EMM). The majority of papers so far has priced non-tradable assets assuming zero or constant MPR, but this assumption yields biased prices and has never been quantified earlier under the EMM framework. Given that liquid-derivative contracts based on daily temperature are traded on the Chicago Mercantile Exchange (CME), we infer the MPR from traded futures-type contracts (CAT, CDD, HDD and AAT). The results show how the MPR significantly differs from 0, how it varies in time and changes in sign. It can be parameterized, given its dependencies on time and temperature seasonal variation. We establish connections between the market risk premium (RP) and the MPR.*

KEY WORDS: CAR process, CME, HDD, seasonal volatitity, risk premium

1. Introduction

In the 1990s weather derivatives (WD) were developed to hedge against the random nature of temperature variations that constitute weather risk. WD are financial contracts with payments based on weather-related measurements. WD cover against volatility caused by temperature, rainfall wind, snow, and frost. The key factor in efficient usage of WD is a reliable valuation procedure. However, due to their specific nature one encounters several difficulties. First, because the underlying weather (and indices) is not tradable and second, the WD market is incomplete, meaning that the WD cannot be cost-efficiently replicated by other WD.

Since the largest portion of WD traded at Chicago Mercantile Exchange (CME) is written on temperature indices, we concentrate our research on temperature derivatives. There have been basically four branches of temperature derivative pricing: actuarial approach, indifference pricing, general equilibrium theory or pricing via no arbitrage arguments. While the actuarial approach considers the conditional expectation of the pay-off under the real physical measure discounted at the riskless rate (Brix *et al*., 2005), the indifference pricing relies on the equivalent utility principle (Barrieu and El Karoui, 2002; Brockett *et al*., 2010) and the general equilibrium theory assumes

Correspondence Address: Brenda López Cabrera, Ladislaus von Bortkiewicz Chair of Statistics, Humboldt University of Berlin, Berlin, Germany. Tel: +49(0)30 2093 1457 Email: lopezcab@wiwi.hu-berlin.de

1350-486X Print/1466-4313 Online/12/010059–37
http://dx.doi.org/10.1080/1350486X.2011.591170

investors' preferences and rules of Pareto optimal risk allocation (Cao and Wei, 2004; Horst and Mueller, 2007; Richards *et al*., 2004). The Martingale approach, although less demanding in terms of assumptions, concentrates on the econometric modelling of the underlying dynamics and requires the selection of an adequate equivalent martingale measure (EMM) to value the pay-offs by taking expectations (Alaton *et al*., 2002; Benth, 2003; Benth and Saltyte-Benth, 2007; Benth *et al*., 2007; Brody *et al*., 2002; Huang-Hsi *et al*., 2008; Mraoua and Bari, 2007).

Here we prefer the latter approach. First, since the underlying (temperature) we consider is of local nature, our analysis aims at understanding the pricing at different locations around the world. Second, the EMM approach helps identify the market price of risk (MPR), which is an important parameter of the associated EMM, and it is indispensable for pricing and hedging non-tradable assets. The MPR can be extracted from traded securities and one can use this value to price other derivatives, though any inference about the MPR requires an assumption about its specification.

The MPR is of high scientific interest, not only for financial risk analysis, but also for better economic modelling of fair valuation of risk. Constantinides (1987) and Landskroner (1977) studied the MPR of tradable assets in the Capital Asset Pricing Model (CAPM) framework. For pricing interest rate derivatives, Vasicek (1977) assumed a constant market price of interest rate, while Hull and White (1990) used the specification of Cox *et al*. (1985). In the oil market, Gibson and Schwartz (1990) supposed an intertemporal constant market price of crude oil conveniences yield risk. Benth *et al*. (2008) introduced a parameterization of the MPR to price electricity derivatives. In the WD framework, Cao and Wei (2004) and Richards *et al*. (2004) studied the MPR as an implicit parameter in a generalization of the Lucas' (1978) equilibrium framework. They showed that the MPR is not only statistically significant for temperature derivatives, but also economically large as well. However, calibration problems arise with the methodology suggested by Cao and Wei (2004), since it deals with a global model like the Lucas' (1978) approach while weather is locally specified. Benth and Saltyte-Benth (2007) introduced theoretical ideas of equivalent changes of measure to get no arbitrage futures/option prices written on different temperature indices. Huang-Hsi *et al*. (2008) examined the MPR of the Taiwan Stock Exchange Capitalization-Weighted Stock Index ((the mean of stock returns – risk-free interest rate)/SD of stock returns) and used it as a proxy for the MPR on temperature option prices. The majority of temperature pricing papers so far has priced temperature derivatives assuming 0 or constant MPR (Alaton *et al*., 2002; Cao and Wei, 2004; Dorfleitner and Wimmer, 2010; Huang-Hsi *et al*., 2008; Richards *et al*., 2004), but this assumption yields biased prices and has never been quantified earlier using the EMM framework. This article deals exactly with the differences between 'historical' and 'risk neutral' behaviours of temperature.

The contribution of this article is threefold. First, in contrast to Campbell and Diebold (2005), Benth and Saltyte-Benth (2007) and Benth *et al*. (2007), we correct for seasonality and seasonal variation of temperature with a local smoothing approach to get, independently of the chosen location, the driving stochastics close to a Gaussian Process and with that being able to apply pricing technique tools of financial mathematics (Karatzas and Shreve, 2001). Second and the main contribution, using statistical modelling and given that liquid derivative contracts based on daily

temperature are traded on the CME, we imply the MPR from traded futures-type contracts (CAT/HDD/CDD/AAT) based on a well-known pricing model developed by Benth *et al*. (2007). We have chosen this methodology because it is a stationary model that fits the stylized characteristics of temperature well; it nests a number of previous models (Alaton *et al*., 2002; Benth, 2003; Benth and Saltyte-Benth, 2005, 2007; Brody *et al.*, 2002; Dornier and Querel, 2007); it provides closed-form pricing formulas; and it computes, after deriving the MPR, non-arbitrage prices based on a continuous-time hedging strategy. Moreover, the price dynamics of futures are easy to compute and require only a one-time estimation. Our implied MPR approach is a calibration procedure for financial engineering purposes. In the calibration exercise, a single date (but different time horizons and calibrated instruments are used) is required, since the model is recalibrated daily to detect intertemporal effects. Moreover, we use an economic and statistical testing approach, where we start from a specification of the MPR and check consistency with the data. The aim of this analysis is to study the effect of different MPR specifications (a constant, a (two) piecewise linear function, a time-deterministic function and a 'financial-bootstrapping') on the temperature futures prices. The statistical point of view is to beat this as an inverse problem with different degrees of smoothness expressed through the penalty parameter of a smoothing spline. The degrees of smoothness will allow for a term structure of risk. Since smoothing estimates are fundamentally different from estimating a deterministic function, we also assure our results by fitting a parametric function to all available contract prices (calendar year estimation). The economic point of view is to detect possible time dependencies that can be explained by investor's preferences in order to hedge weather risk. Our findings that the MPR differs significantly from 0 confirm the results found in Cao and Wei (2004), Huang-Hsi *et al*. (2008), Richards *et al*. (2004) and Alaton *et al*. (2002), but we differ from them, by showing that it varies in time and changes in sign. It is not a reflection of bad model specification, but data-extracted MPR. This contradicts the assumption made earlier in the literature that the MPR is 0 or constant and rules out the 'burn-in' analysis, which is popular among practitioners since it uses the historical average index value as the price for the futures (Brix *et al*., 2005). This brings significant challenges to the statistical branch of the pricing literature. We also establish connections between the market risk premium (RP) (a Girsanov-type change of probability) and the MPR. As a third contribution, we discuss how to price over-the-counter (OTC) temperature derivatives with the information extracted.

Our article is structured as follows. Section 2 presents the fundamentals of temperature derivatives (futures and options) and describes the temperature data and the temperature futures traded at CME, the biggest market offering this kind of product. Section 3 is devoted to explaining the dynamics of temperature data – the econometric part. The temperature model captures linear trend, seasonality, mean reversion, intertemporal correlations and seasonal volatility effects. Section 4 – the financial mathematics part – connects the weather dynamics with the pricing methodology. Section 5 solves the inverse problem of determining the MPR of CME temperature futures using different specifications. Section 1 introduces the estimation results and test procedures of our specifications applied into temperature-derivative data. Here we give (statistical and economic) interpretations of the estimated MPR. The pricing of

OTC temperature products is discussed at the end of this section. Section 6 concludes the article. All computations in this article were carried out in Matlab version 7.6 (The MathWorks, Inc., Natick, MA, USA). To simplify notation, dates are denoted with yyyymmdd format.

2. Temperature Derivatives

The largest portion of futures and options written on temperature indices is traded on the CME. Most of the temperature derivatives are written on daily average temperature indices, rather than on the underlying temperature by itself. A call option written on futures $F_{(t,\tau_1,\tau_2)}$ with exercise time $t \leq \tau_1$ and delivery over a period $[\tau_1, \tau_2]$ will pay $\max\left\{F_{(t,\tau_1,\tau_2)} - K, 0\right\}$ at the end of the measurement period $[\tau_1, \tau_2]$. The most common weather indices on temperature are Heating Degree Day (HDD), Cooling Degree Day (CDD) and Cumulative Averages (CAT). The HDD index measures the temperature over a period $[\tau_1, \tau_2]$, usually between October and April:

$$HDD(\tau_1, \tau_2) = \int_{\tau_1}^{\tau_2} \max(c - T_u, 0)\, du, \tag{1}$$

where c is the baseline temperature (typically 18°C or 65°F) and $T_u = (T_{u,max} + T_{u,min})/2$ is the average temperature on day u. Similarly, the CDD index measures the temperature over a period $[\tau_1, \tau_2]$, usually between April and October:

$$CDD(\tau_1, \tau_2) = \int_{\tau_1}^{\tau_2} \max(T_u - c, 0)\, du. \tag{2}$$

The HDD and the CDD index are used to trade futures and options in 24 US cities, 6 Canadian cities and 3 Australian cities. The CAT index accounts the accumulated average temperature over $[\tau_1, \tau_2]$:

$$CAT(\tau_1, \tau_2) = \int_{\tau_1}^{\tau_2} T_u du. \tag{3}$$

The CAT index is the substitution of the CDD index for 11 European cities. Since $\max(T_u - c, 0) - \max(c - T_u, 0) = T_u - c$, we get the HDD–CDD parity:

$$CDD(\tau_1, \tau_2) - HDD(\tau_1, \tau_2) = CAT(\tau_1, \tau_2) - c(\tau_2 - \tau_1). \tag{4}$$

Therefore, it is sufficient to analyse only HDD and CAT indices. An index similar to the CAT index is the Pacific Rim Index, which measures the accumulated total of 24-hr average temperature (C24AT) over a period $[\tau_1, \tau_2]$ days for Japanese cities:

$$\text{C24AT}(\tau_1, \tau_2) = \int_{\tau_1}^{\tau_2} \widetilde{T}_u du, \tag{5}$$

where $\widetilde{T}_u = \frac{1}{24} \int_1^{24} T_{u_i} du_i$ and T_{u_i} denotes the temperature of hour u_i. A difference of the CAT and the C24AT index is that the latter is traded over the whole year. Note that

temperature is a continuous-time process even though the indices used as underlying for temperature futures contracts are discretely monitored.

As temperature is not a marketable asset, the replication arguments for any temperature futures contract do not hold and incompleteness of the market follows. In this context, any probability measure Q equivalent to the objective P is also an EMM and a risk neutral probability turns all the tradable assets into martingales after discounting. However, since temperature futures/option prices dynamics are indeed tradable assets, they must be free of arbitrage. Thanks to the Girsanov theorem, equivalent changes of measures are simply associated with changes of drift. Hence, under a probability space $(\Omega, \mathcal{F}, Q)$ with a filtration $\{\mathcal{F}_t\}_{0\leq t\leq \tau_{\max}}$, where $\tau_{\max}$ denotes a maximal time covering all times of interest in the market, we choose a parameterized equivalent pricing measure $Q = Q_\theta$ that completes the market and pin it down to compute the arbitrage-free temperature futures price:

$$F_{(t,\tau_1,\tau_2)} = E^{Q_\theta}[Y|\mathcal{F}_t], \tag{6}$$

where Y refers to the pay-off from the temperature index in Equations (2)–(5). The MPR θ is assumed to be a real-valued, bounded and piecewise continuous function. We later relax that assumption, by considering the time-dependent MPR θ_t. In fact, the MPR can depend on anything that can affect investors' attitudes. The MPR can be inferred from market data.

The choice of Q determines the RP demanded for investors for holding the temperature derivative, and opposite, having knowledge of the RP determines the choice of the risk-neutral probability. The RP is defined as a drift of the spot dynamics or a Girsanov-type change of probability. In Equation (6), the futures price is set under a risk-neutral probability $Q = Q_\theta$, thereby the RP measures exactly the differences between the risk-neutral $F_{(t,\tau_1^i,\tau_2^i,Q)}$ (market prices) and the temperature market probability predictions $\hat{F}_{(t,\tau_1^i,\tau_2^i,P)}$ (under P):

$$RP = F_{(t,\tau_1^i,\tau_2^i,Q)} - \hat{F}_{(t,\tau_1^i,\tau_2^i,P)}. \tag{7}$$

Using the 'burn-in' approach of Brix *et al*. (2005), the futures price is only the historical average index value, therefore there is no RP since $Q = P$.

2.1 Data

We have temperature data available from 35 US, 30 German, 159 Chinese and 9 European weather stations. The temperature data were obtained from the National Climatic Data Center (NCDC), the Deutscher Wetterdienst (DWD), Bloomberg Professional Service, the Japanese Meteorological Agency (JMA) and the China Meteorological Administration (CMA). The temperature data contain the minimum, maximum and average daily temperatures measured in degree Fahrenheit for US cities and degree Celsius for other cities. The data set period is, in most of the cities, from 19470101 to 20091231.

The WD data traded at CME were provided by Bloomberg Professional Service. We use daily closing prices from 20000101 to 20091231. The measurement periods for the

different temperature indices are standardized to be as each month of the year or as seasonal strips (minimum of 2 and maximum of 7 consecutive calendar months). The futures and options at the CME are cash settled, that is, the owner of a futures contract receives 20 times the index at the end of the measurement period, in return for a fixed price. The currency is British pounds for the European futures contracts, US dollars for the US contracts and Japanese Yen for the Asian cities. The minimum price increment is 1 index point. The degree day metric is Celsius or Fahrenheit and the termination of the trading is two calendar days following the expiration of the contract month. The accumulation period of each CAT/CDD/HDD/C24AT index futures contract begins with the first calendar day of the contract month and ends with the calendar day of the contract month. Earth Satellite Corporation (ESC) reports to CME the daily average temperature. Traders bet that the temperature will not exceed the estimates from ESC.

3. Temperature Dynamics

In order to derive explicitly no arbitrage prices for temperature derivatives, we need first to describe the dynamics of the underlying under the physical measure. This article studies the average daily temperature data (because most of the temperature derivative trading is based on this quantity) for US, European and Asian cities. In particular, we analyse the weather dynamics for Atlanta, Portland, Houston, New York, Berlin, Essen, Tokyo, Osaka, Beijing and Taipei (Table 1). Our interest in these cities is because all of them with the exception of the latter two are traded at CME and because a casual examination of the trading statistics on the CME website reveals that the Atlanta HDD, Houston CDD and Portland CDD temperature contracts have relatively more liquidity.

Most of the literature that discuss models for daily average temperature and capture a linear trend (due to global warming and urbanization), seasonality (peaks in cooler winter and warmer summers), mean reversion, seasonal volatility (a variation that varies seasonally) and strong correlations (long memory); see, for example, Alaton *et al*. (2002), Cao and Wei (2004), Campbell and Diebold (2005) and Benth *et al*. (2007). They differ from their definition of temperature variations, which is exactly the component that characterizes weather risk. Here we show that an autoregressive (AR) model AR of high order (p) for the detrended daily average temperatures (rather than the underlying temperature itself) is enough to capture the stylized facts of temperature.

We first need to remove the seasonality in mean Λ_t from the daily temperature series T_t, check intertemporal correlations and remove the seasonality in variance to deal with a stationary process. The deterministic seasonal mean component can be approximated with Fourier-truncated series (FTS):

$$\Lambda_t = a + bt + \sum_{l=1}^{L} c_l \cos\left\{\frac{2\pi(t - d_l)}{l \cdot 365}\right\}, \tag{8}$$

where the coefficients a and b indicate the average daily temperature and global warming, respectively. We observe low temperatures in winter times and high temperatures in summer for different locations. The temperature data sets do not deviate from its

Table 1. Coefficients of the Fourier-truncated seasonal series of average daily temperatures in different cities.

City	Period	$\hat{a}$ (CI)	$\hat{b}$ (CI)	$\hat{c}_1$ (CI)	$\hat{d}_1$ (CI)
Atlanta	19480101–20081204	61.95 (61.95, 61.96)	−0.0025 (−0.0081, 0.0031)	18.32 (18.31, 18.33)	−165.02 (−165.03, −165.02)
Beijing	19730101–20090831	12.72 (12.71, 12.73)	0.0001 (−0.0070, 0.0073)	14.93 (14.92, 14.94)	−169.59 (−169.59, −169.58)
Berlin	19480101–20080527	9.72 (9.71, 9.74)	−0.0004 (−0.0147, 0.0139)	9.75 (9.74, 9.77)	−164.79 (−164.81, −164.78)
Essen	19700101–20090731	10.80 (10.79, 10.81)	−0.0020 (−0.0134, 0.0093)	8.02 (8.01, 8.03)	−161.72 (−161.73, −161.71)
Houston	19700101–20081204	68.52 (68.51, 68.52)	−0.0006 (−0.0052, 0.0039)	15.62 (15.62, 15.63)	−165.78 (−165.79, −165.78)
New York	19490101–20081204	53.86 (53.86, 53.87)	−0.0004 (−0.0079, 0.0071)	21.43 (21.42, 21.44)	−156.27 (−156.27, −156.26)
Osaka	19730101–20090604	16.78 (16.77, 16.79)	−0.0021 (−0.0109, 0.0067)	11.61 (11.60, 11.62)	−153.57 (−153.58, −153.56)
Portland	19480101–20081204	55.35 (55.35, 55.36)	−0.0116 (−0.0166, −0.0065)	14.36 (14.36, 14.37)	−155.58 (−155.58, −155.57)
Taipei	19920101–20090806	23.32 (23.31, 23.33)	0.0023 (−0.0086, 0.0133)	6.67 (6.66, 6.68)	−158.67 (−158.68, −158.66)
Tokyo	19730101–20090831	16.32 (16.31, 16.33)	−0.0003 (−0.0085, 0.0079)	10.38 (10.37, 10.38)	−153.52 (−153.53, −153.52)

Notes: CI, Confidence interval.
All coefficients are non-zero at 1% significance level. CIs are given in parentheses. Dates given in yyyymmdd format. The daily temperature is measured in degree Celsius, except for American cities measured in degree Fahrenheit.

mean level and in most of the cases a linear trend at 1% significance level is detectable as it is displayed in Table 1.

Our findings are similar to Alaton *et al*. (2002) and Benth *et al*. (2007) for Sweden; Benth *et al*. (2007) for Lithuania; Campbell and Diebold (2005) for the United States; and Papazian and Skiadopoulos (2010) for Barcelona, London, Paris and Rome. In our empirical results, the number of periodic terms of the FTS series varies from city to city, sometimes from 4 to 21 or more terms. We notice that the series expansion in Equation (8) with more and more periodic terms provides a fine tuning, but this will increase the number of parameters. Here we propose a different way to correct for seasonality. We show that a local smoothing approach does that job instead, but with less technical expression. Asymptotically they can be approximated by FTS estimators. For a fixed time point $s \in [1, 365]$, we smooth Λ_s with a Local Linear Regression (LLR) estimator:

$$\Lambda_s = \underset{e,f}{\arg\min} \sum_{t=1}^{365} \left\{ \bar{T}_t - e_s - f_s(t-s) \right\}^2 K\left(\frac{t-s}{h} \right), \tag{9}$$

where $\bar{T}_t$ is the mean of average daily temperature in J years, h is the smoothing parameter and $K(\cdot)$ denotes a kernel. This estimator, using Epanechnikov Kernel, incorporates an asymmetry term since high temperatures in winter are more pronounced than in summer as Figure 1 displays in a stretch of 8 years plot of the average daily temperatures over the FTS estimates.

After removing the LLR-seasonal mean (Equation (9)) from the daily average temperatures ($X_t = T_t - \Lambda_t$), we apply the Augmented Dickey–Fuller (ADF) and the Kwaitkowski–Phillips–Schmidt–Shin (KPSS) tests to check whether X_t is a stationary process. We then plot the Partial Autocorrelation Function (PACF) of X_t to detect possible intertemporal correlations. This suggests that persistence of daily average is captured by AR processes of higher order p:

$$X_{t+p} = \sum_{i=1}^{p} \beta_i X_{t-i} + \varepsilon_t, \varepsilon_t = \sigma_t e_t, e_t \sim N(0,1), \tag{10}$$

with residuals ε_t. Under the stationarity hypothesis of the coefficients βs and the mean zero of residuals ε_t, the mean temperature $E[T_t] = \Lambda_t$. This is different to the approach of Campbell and Diebold (2005), who suggested to regress deseasonalized temperatures on original temperatures. The analysis of the PACFs and Akaike's information criterion (AIC) suggests that the $AR(3)$ model in Benth *et al*. (2007) explains the temperature evolution well and holds for many cities. The results of the stationarity tests and the coefficients of the fitted $AR(3)$ are given in Table 2. Figure 2 illustrates that the ACFs of the residuals ε_t are close to 0 and according to Box-Ljung statistic the first few lags are insignificant, whereas the ACFs of the squared residuals ε_t^2 show a high seasonal pattern.

We calibrate the deterministic seasonal variance function σ_t^2 with FTS and an additional generalized autoregressive conditional heteroskedasticity (GARCH) (p, q) term:

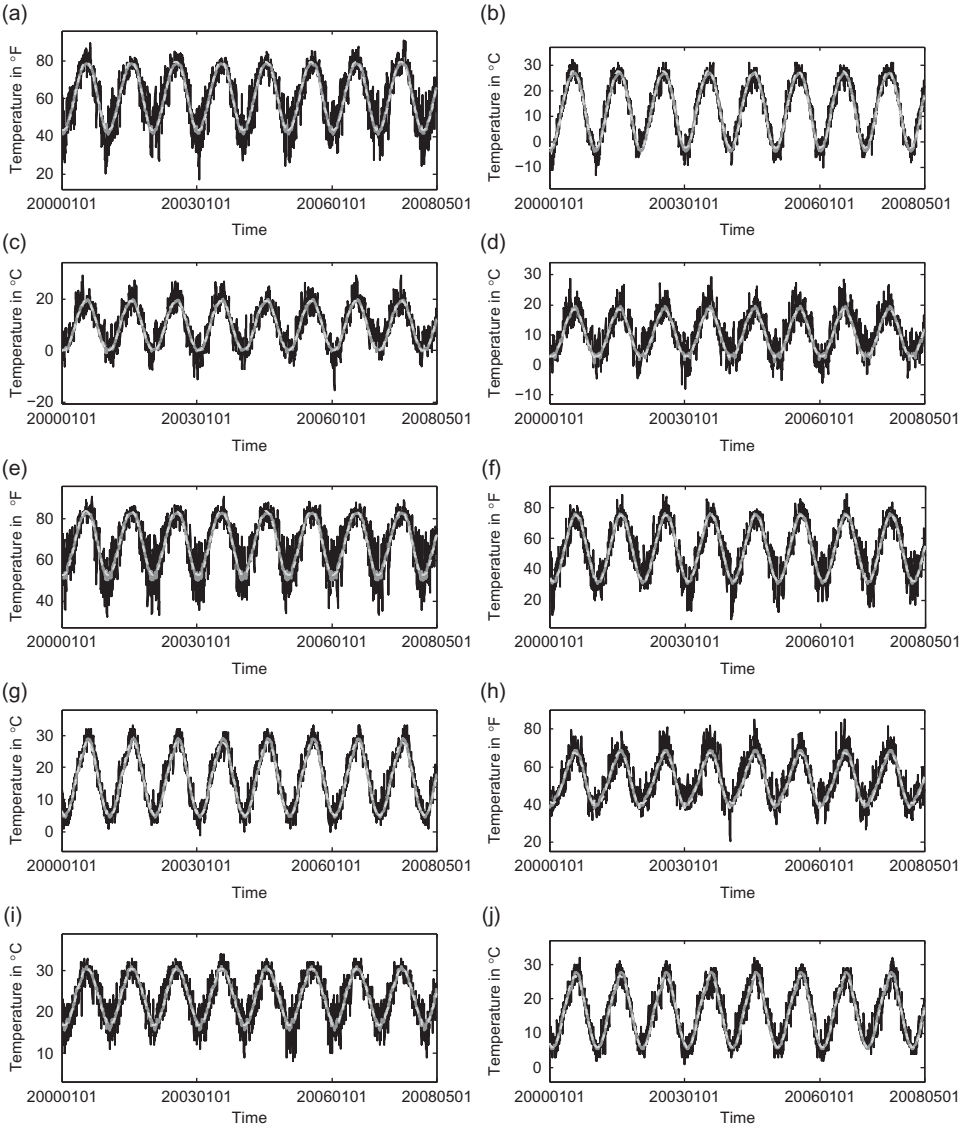


Figure 1. A stretch of 8 years plot of the average daily temperatures (black line), the seasonal component modelled with a Fourier-truncated series (dashed line) and the local linear regression (grey line) using Epanechnikov Kernel. (a) Atlanta, (b) Beijing, (c) Berlin, (d) Essen, (e) Houston, (f) New York, (g) Osaka, (h) Portland, (i) Taipei and (j) Tokyo.

$$\hat{\sigma}_t^2 = c + \sum_{l=1}^{L} \left\{ c_{2l} \cos\left(\frac{2l\pi t}{365}\right) + c_{2l+1} \sin\left(\frac{2l\pi t}{365}\right) \right\} + \alpha_1(\sigma_{t-1}\eta_{t-1})^2 + \beta_1\sigma_{t-1}^2, \eta_t \sim iid\, N(0,1). \tag{11}$$

Table 2. Result of the stationary tests and the coefficients of the fitted $AR(3)$.

	ADF–KPSS		$AR(3)$			$CAR(3)$				
City	$\hat{\tau}$	$\hat{k}$	β_1	β_2	β_3	α_1	α_2	α_3	λ_1	$\lambda_{2,3}$
Atlanta	-55.55^{+}	0.21^{***}	0.96	−0.38	0.13	2.03	1.46	0.28	−0.30	−0.86
Beijing	-30.75^{+}	0.16^{***}	0.72	−0.07	0.05	2.27	1.63	0.29	−0.27	−1.00
Berlin	-40.94^{+}	0.13^{**}	0.91	−0.20	0.07	2.08	1.37	0.20	−0.21	−0.93
Essen	-23.87^{+}	0.11^{*}	0.93	−0.21	0.11	2.06	1.34	0.16	−0.16	−0.95
Houston	-38.17^{+}	0.05^{*}	0.90	−0.39	0.15	2.09	1.57	0.33	−0.33	−0.87
New York	-56.88^{+}	0.08^{*}	0.76	−0.23	0.11	2.23	1.69	0.34	−0.32	−0.95
Osaka	-18.65^{+}	0.09^{*}	0.73	−0.14	0.06	2.26	1.68	0.34	−0.33	−0.96
Portland	-45.13^{+}	0.05^{*}	0.86	−0.22	0.08	2.13	1.48	0.26	−0.27	−0.93
Taipei	-32.82^{+}	0.09^{*}	0.79	−0.22	0.06	2.20	1.63	0.36	−0.40	−0.90
Tokyo	-25.93^{+}	0.06^{*}	0.64	−0.07	0.06	2.35	1.79	0.37	−0.33	−1.01

Notes: ADF, augmented Dickey–Fuller; KPSS, Kwiatkowski–Phillips–Schmidt–Shin; AR, autoregressive process; CAR, continuous autoregressive model.
ADF and KPSS statistics, coefficients of the $AR(3)$, $CAR(3)$ and eigenvalues $\lambda_{1,2,3}$, of the matrix $\mathbf{A}$ of the $CAR(3)$ model for the detrended daily average temperatures for different cities.
$^{+}$0.01 critical values, *0.1 critical value, **0.05 critical value, ***0.01 critical value.

The Fourier part in Equation (11) captures the seasonality in volatility, whereas the GARCH part captures the remaining non-seasonal volatility. Note again that more and more periodic terms in Equation (11) provide a good fitting but this will increase the number of parameters. To avoid this and in order to achieve positivity of the variance, Gaussian risk factors and volatility model flexibility in a continuous time, we propose the calibration of the seasonal variance in terms of an LLR:

$$\arg\min_{g,h} \sum_{t=1}^{365} \left\{\hat{\varepsilon}_t^2 - g_s - h_s(t-s)\right\}^2 K\left(\frac{t-s}{h}\right), \tag{12}$$

where $\hat{\varepsilon}_t^2$ is the mean of squared residuals in J years and $K(\cdot)$ is a kernel. Figure 3 shows the daily empirical variance (the average of squared residuals for each day of the year), the fittings using the FTS-GARCH(1,1) and the LLR (with Epanechnikov kernel) estimators. Here we obtain the Campbell and Diebold's (2005) effect for different temperature data, high variance in winter to earlier summer and low variance in spring to late summer. The effects of non-seasonal GARCH volatility component are small.

Figure 4 displays the ACFs of temperature residuals ε_t and squared residuals ε_t^2 after dividing out the deterministic LLR seasonal variance. The ACF plots of the standardized residuals remain unchanged but now the squared residuals presents a non-seasonal pattern. The LLR seasonal variance creates almost normal residuals and captures the peak seasons as Figure 5 in a log Kernel smoothing density plot shows against a Normal Kernel evaluated at 100 equally spaced points. Table 3 presents the calibrated coefficients of the FTS-GARCH seasonal variance estimates and the

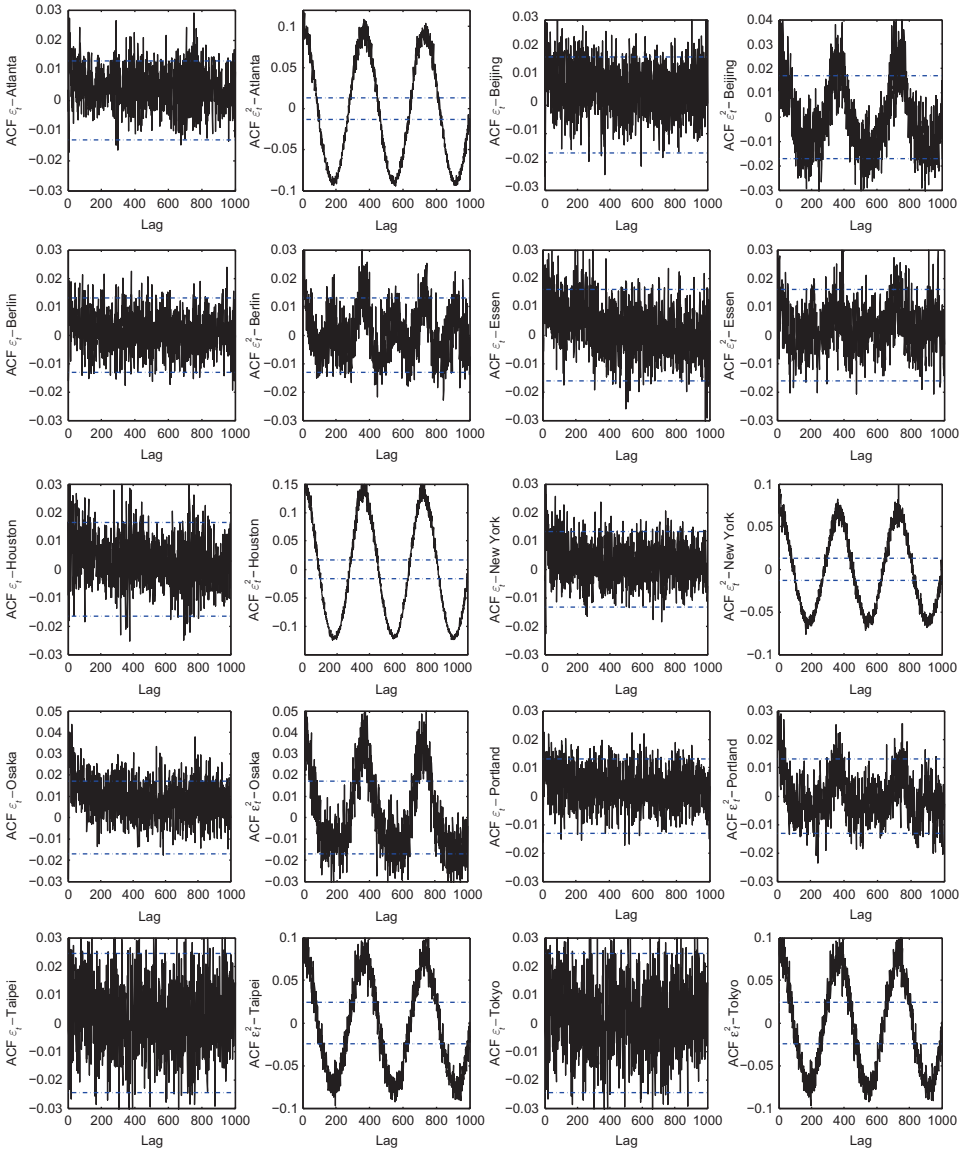


Figure 2. The ACF of residuals ε_t (left panels) and squared residuals ε_t^2 (right panels) of detrended daily temperatures for different cities.

descriptive statistics for the residuals after correcting by the FTS-GARCH and LLR seasonal variance. We observe that independently of the chosen location, the driving stochastics are close to a Wiener process. This will allow us to apply the pricing tools of financial mathematics.

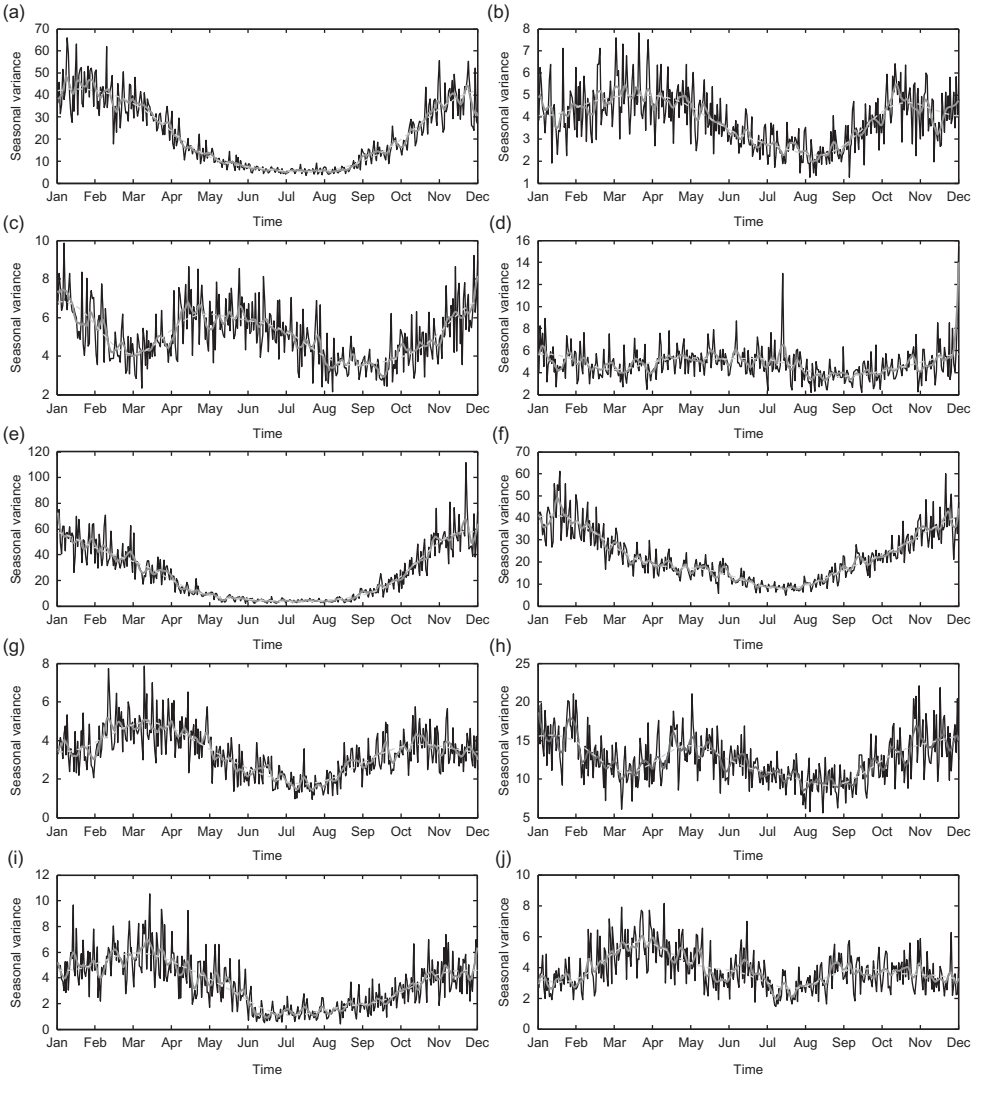


Figure 3. The daily empirical variance (black line), the Fourier-truncated (dashed line) and the local linear smoother seasonal variation using Epanechnikov kernel (grey line) for different cities. (a) Atlanta, (b) Beijing, (c) Berlin, (d) Essen, (e) Houston, (f) New York, (g) Osaka, (h) Portland, (i) Taipei and (j) Tokyo.

4. Stochastic Pricing Model

Temperatures are naturally evolving continuously over time, so it is very convenient to model the dynamics of temperature with continuous-time stochastic processes, although the data may be on a daily scale. We therefore need the reformulation of the underlying process in continuous time to be more convenient with market definitions.

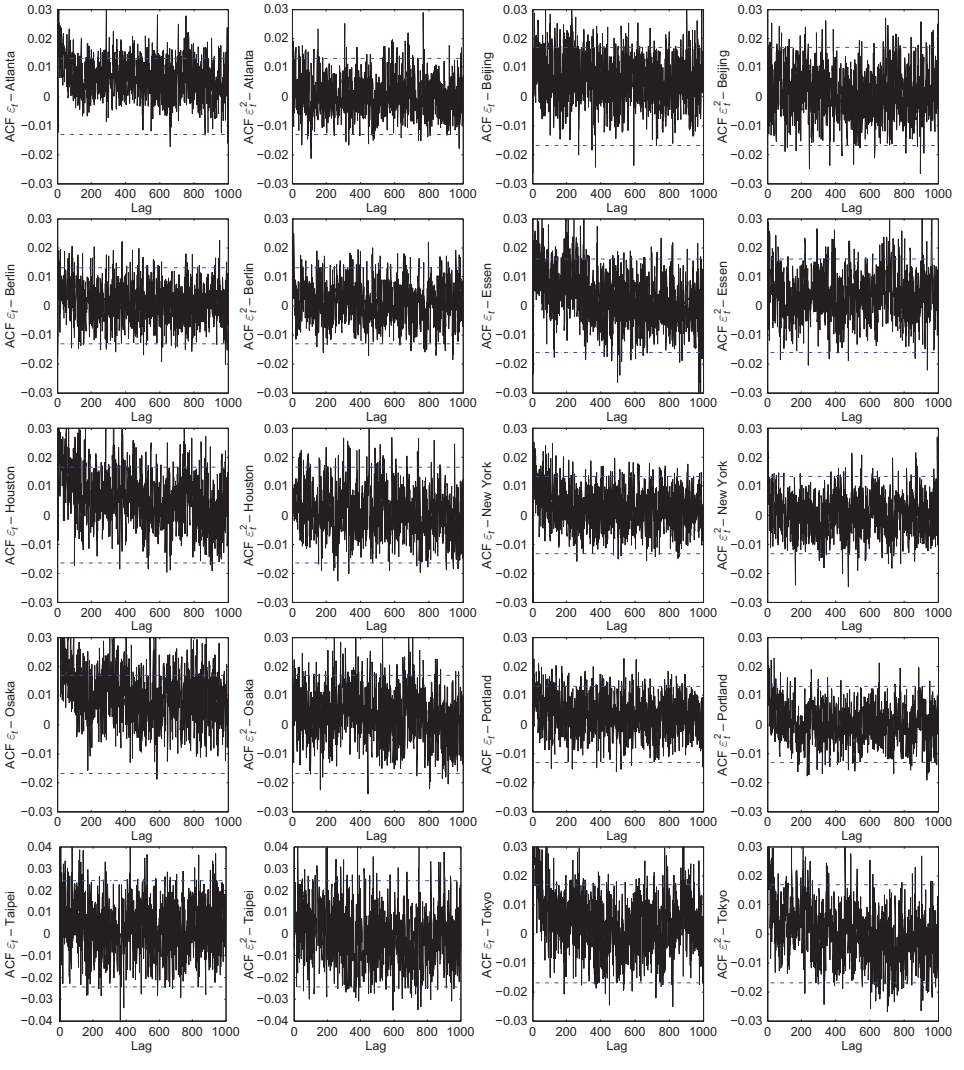


Figure 4. The ACF of residuals ε_t (left panels) and squared residuals ε_t^2 (right panels) of detrended daily temperatures after dividing out the local linear seasonal variance for different cities.

We show that the $AR(p)$ (Equation (10)) estimated in Section 3 for the detrended temperature can be therefore seen as a discretely sampled continuous-time autoregressive (CAR) process ($CAR(p)$) driven by a one-dimensional Brownian motion B_t (though the continuous-time process is Markov in higher dimension) (Benth *et al*., 2007):

$$d\boldsymbol{X}_t = \mathbf{A}\boldsymbol{X}_t dt + \mathbf{e}_p \sigma_t dB_t, \tag{13}$$

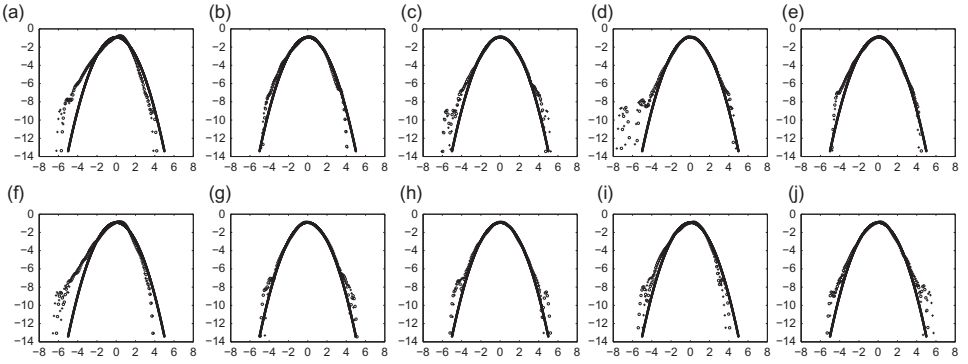


Figure 5. The log of Normal Kernel (*) and log of Kernel smoothing density estimate of residuals after correcting FTS (+) and locar linear (o) seasonal variance for different cities (a) Atlanta, (b) Beijing, (c) Berlin, (d) Essen, (e) Houston, (f) New York, (g) Osaka, (h) Portland, (i) Taipei and (j) Tokyo.

where the state vector $\boldsymbol{X}_t \in \mathbb{R}^p$ for $p \geq 1$ is a vectorial Ornstein–Uhlenbeck process, namely, the temperatures after removing seasonality at times t, $t-1$, $t-2$, $t-3, \ldots$; e_k denotes the kth unit vector in $\mathbb{R}^p$ for $k = 1, \ldots, p$; $\sigma_t > 0$ is a deterministic volatility (real-valued and square integrable function); and $\mathbf{A}$ is a $p \times p$ matrix:

$$\mathbf{A} = \begin{pmatrix} 0 & 1 & 0 & \ldots & 0 & 0 \\ 0 & 0 & 1 & \ldots & 0 & 0 \\ \vdots & & \ddots & & 0 & \vdots \\ 0 & \ldots & \ldots & 0 & 0 & 1 \\ -\alpha_p & -\alpha_{p-1} & \ldots & & 0 & -\alpha_1 \end{pmatrix}, \tag{14}$$

with positive constants α_k. Following this nomenclature, $X_{q(t)}$ with $q = 1, \ldots, p$ is the qth coordinate of X_t and by setting $q = 1$ is equivalent to the detrended temperature time series $X_{1(t)} = T_t - \Lambda_t$. The proof is derived by an analytical link between $X_{1(t)}, X_{2(t)}$ and $X_{3(t)}$ and the lagged temperatures up to time $t-3$. $X_{1(t+3)}$ is approximated by Euler discretization. Thus for $p = 1, \boldsymbol{X}_t = X_{1(t)}$ and Equation (13) becomes

$$dX_{1(t)} = -\alpha_1 X_{1(t)} dt + \sigma_t dB_t, \tag{15}$$

which is the continuous version of an $AR(1)$ process. Similarly for $p = 2$, assume a time step of length 1 $dt = 1$ and substitute $X_{2(t)}$ iteratively to get

$$X_{1(t+2)} \approx (2 - \alpha_1) X_{1(t+1)} + (\alpha_1 - \alpha_2 - 1) X_{1(t)} + \sigma_t e_t, \tag{16}$$

Table 3. Coefficients of the FTS-GARCH seasonal variance.

	Coefficients of the FTS									Corrected residuals $\frac{\hat{\varepsilon}_t}{\hat{\sigma}_t}$ with FTS			Corrected residuals $\frac{\hat{\varepsilon}_t}{\hat{\sigma}_t}$ with LLR		
City	$\hat{c}_1$	$\hat{c}_2$	$\hat{c}_3$	$\hat{c}_4$	$\hat{c}_5$	$\hat{c}_6$	$\hat{c}_7$	$\hat{c}_8$	$\hat{c}_9$	JB	Kurt	Skew	JB	Kurt	Skew
Atlanta	21.51	18.10	7.09	2.35	1.69	−0.39	−0.68	0.24	−0.45	272.01	3.98	−0.70	253.24	3.91	−0.68
Beijing	3.89	0.70	0.84	−0.22	−0.49	−0.20	−0.14	−0.11	0.08	219.67	3.27	−0.28	212.46	3.24	−0.28
Berlin	5.07	0.10	0.72	0.98	−0.43	0.45	0.06	0.16	0.22	224.55	3.48	−0.05	274.83	3.51	−0.08
Essen	4.78	0.00	0.42	0.63	−0.20	0.17	−0.06	0.05	0.17	273.90	3.65	−0.05	251.89	3.61	−0.08
Houston	23.61	25.47	4.49	6.65	−0.38	1.00	−2.67	0.68	−1.56	140.97	3.96	−0.60	122.83	3.87	−0.57
New York	22.29	13.80	3.16	3.30	−0.47	0.80	2.04	0.11	0.01	367.38	3.43	−0.23	355.03	3.43	−0.22
Osaka	3.34	0.80	0.80	−0.57	−0.27	−0.18	−0.07	0.01	−0.03	105.32	3.37	−0.11	101.50	3.36	−0.11
Portland	12.48	1.55	1.05	1.42	−1.19	0.46	0.34	−0.40	0.45	67.10	3.24	0.06	75.01	3.27	0.02
Taipei	3.50	1.49	1.59	−0.38	−0.16	0.03	−0.17	−0.09	−0.18	181.90	3.26	−0.39	169.41	3.24	−0.37
Tokyo	3.80	0.01	0.73	−0.69	−0.33	−0.14	−0.14	0.26	−0.13	137.93	3.45	−0.10	156.58	3.46	−0.13

Notes: FTS, Fourier-truncated series; JB, Jarque Bera; LLR, local linear regression; GARCH, generalized autoregressive conditional heteroskedasticity.
Seasonal variance estimate of $\{c_i\}_{i=1}^{9}$ fitted with an FTS and statistics – Skewness (Skew), kurtosis (Kurt) and JB test statistics – of the standardized residuals with seasonal variances fitted with FTS-GARCH and with LLR.
Coefficients are significant at 1% level.

where $e_t = B_{t+1} - B_t$. For $p = 3$, we have:

$$\begin{aligned}
X_{1(t+1)} - X_{1(t)} &= X_{2(t)}dt, \\
X_{2(t+1)} - X_{2(t)} &= X_{3(t)}dt, \\
X_{3(t+1)} - X_{3(t)} &= -\alpha_3 X_{1(t)}dt - \alpha_2 X_{2(t)}dt - \alpha_1 X_{3(t)}dt + \sigma_t e_t, \\
&\ldots, \\
X_{1(t+3)} - X_{1(t+2)} &= X_{2(t+2)}dt, \\
X_{2(t+3)} - X_{2(t+2)} &= X_{3(t+2)}dt, \\
X_{3(t+3)} - X_{3(t+2)} &= -\alpha_3 X_{1(t+2)}dt - \alpha_2 X_{2(t+2)}dt - \alpha_1 X_{3(t+2)}dt + \sigma_t e_t,
\end{aligned} \tag{17}$$

substituting into the $X_{1(t+3)}$ dynamics and setting $dt = 1$:

$$X_{1(t+3)} \approx \underbrace{(3 - \alpha_1)}_{\beta_1} X_{1(t+2)} + \underbrace{(2\alpha_1 - \alpha_2 - 3)}_{\beta_2} X_{1(t+1)} + \underbrace{(-\alpha_1 + \alpha_2 - \alpha_3 + 1)}_{\beta_3} X_{1(t)}. \tag{18}$$

Please note that this corrects the derivation in Benth *et al*. (2007) and Equation (18) leads to Equation (10) (with $p = 3$). The approximation of Equation (18) is required to compute the eigenvalues of matrix $\mathbf{A}$. The last columns of Table 2 display the $CAR(3)$-parameters and the eigenvalues of the matrix $\mathbf{A}$ for the studied temperature data. The stationarity condition is fulfilled since the eigenvalues of $\mathbf{A}$ have negative real parts and the variance matrix $\int_0^t \sigma_{t-s}^2 \exp\{\mathbf{A}(s)\}\, \mathbf{e}_p \mathbf{e}_p^\top \exp\left\{\mathbf{A}^\top(s)\right\} ds$ converges as $t \to \infty$.

By applying the multidimensional *Itô Formula*, the process in Equation (13) with $\boldsymbol{X}_t = x \in \mathbb{R}^p$ has the explicit form $\boldsymbol{X}_s = \exp\{\mathbf{A}(s-t)\}\, x + \int_t^s \exp\{\mathbf{A}(s-u)\}\, \mathbf{e}_p \sigma_u dB_u$ for $s \geq t \geq 0$.

Since dynamics of temperature futures prices must be free of arbitrage under the pricing equivalent measure Q_θ, the temperature dynamics of Equation (13) becomes for $s \geq t \geq 0$:

$$\begin{aligned}
d\boldsymbol{X}_t &= (\mathbf{A}\boldsymbol{X}_t + \mathbf{e}_p \sigma_t \theta_t)dt + \mathbf{e}_p \sigma_t dB_t^\theta, \\
\boldsymbol{X}_s &= \exp\{\mathbf{A}(s-t)\}\, \mathbf{x} + \int_t^s \exp\{\mathbf{A}(s-u)\}\, \mathbf{e}_p \sigma_u \theta_u du + \int_t^s \exp\{\mathbf{A}(s-u)\}\, \mathbf{e}_p \sigma_u dB_u^\theta.
\end{aligned} \tag{19}$$

By inserting Equations (1)–(3) into Equation (6), Benths *et al*. (2007) explicity calculated the risk neutral prices for HDD/CDD/CAT futures (and options) for contracts traded before the temperature measurement period, that is $0 \leq t \leq \tau_1 < \tau_2$:

$$
\begin{aligned}
F_{HDD(t,\tau_1,\tau_2)} &= \int_{\tau_1}^{\tau_2} \upsilon_{t,s}\psi\left[\frac{c - m_{\left\{t,s,\mathbf{e}_1^\top \exp\{\mathbf{A}(s-t)\}\boldsymbol{X}_t\right\}}}{\upsilon_{t,s}}\right] ds,\\
F_{CDD(t,\tau_1,\tau_2)} &= \int_{\tau_1}^{\tau_2} \upsilon_{t,s}\psi\left[\frac{m_{\left\{t,s,\mathbf{e}_1^\top \exp\{\mathbf{A}(s-t)\}\boldsymbol{X}_t\right\}} - c}{\upsilon_{t,s}}\right] ds,\\
F_{CAT(t,\tau_1,\tau_2)} &= \int_{\tau_1}^{\tau_2} \Lambda_u du + \mathbf{a}_{t,\tau_1,\tau_2}\boldsymbol{X}_t + \int_t^{\tau_1} \theta_u\sigma_u\mathbf{a}_{t,\tau_1,\tau_2}\mathbf{e}_p du\\
&\quad + \int_{\tau_1}^{\tau_2} \theta_u\sigma_u\mathbf{e}_1^\top\mathbf{A}^{-1}\left[\exp\{\mathbf{A}(\tau_2-u)\} - I_p\right]\mathbf{e}_p du,
\end{aligned}
\tag{20}
$$

with $\mathbf{a}_{t,\tau_1,\tau_2} = \mathbf{e}_1^\top\mathbf{A}^{-1}\left[\exp\{\mathbf{A}(\tau_2-t)\} - \exp\{\mathbf{A}(\tau_1-t)\}\right]$; I_p is a $p\times p$ identity matrix; $\psi(x) = x\Phi(x) + \varphi(x)$ (Φ denotes the standard normal cumulative distribution function (cdf)) with $x = \mathbf{e}_1^\top \exp\{\mathbf{A}(s-t)\}\boldsymbol{X}_t$; $\upsilon_{t,s}^2 = \int_t^s \sigma_u^2\left[\mathbf{e}_1^\top \exp\{\mathbf{A}(s-t)\}\mathbf{e}_p\right]^2 du$; and $m_{\{t,s,x\}} = \Lambda_s + \int_t^s \sigma_u\theta_u\mathbf{e}_1^\top \exp\{\mathbf{A}(s-t)\}\mathbf{e}_p du + x$. The solution to Equation (20) depends on the assumed specification for the MPR θ. In the next section, it is shown that different assumed risk specifications can lead into different derivative prices.

The model in Benth *et al*. (2007) nests a number of previous models (Alaton *et al*., 2002; Benth, 2003; Benth and Saltyte-Benth, 2005; Brody *et al*., 2002); it generalizes the Benth and Saltyte-Benth (2007) and Dornier and Querel (2007) approaches and is a very well studied methodology in the literature (Benth *et al*., 2011; Papazian and Skiadopoulos, 2010; Zapranis and Alexandridis, 2008). Besides it gives a clear connection between the discrete- and continuous-time versions, it provides closed-form non-arbitrage pricing formulas and it requires only a one-time estimation for the price dynamics. With the time series approach (Campbell and Diebold, 2005), the continuous-time approaches (Alaton *et al*., 2002; Huang-Hsi *et al*., 2008), neural networks (Zapranis and Alexandridis, 2008, 2009) or the principal component analysis approach (Papazian and Skiadopoulos, 2010) are not easy to compute price dynamics of CAT/CDD/HDD futures and one needs to use numerical approaches or simulations in order to calculate conditional expectations in Equation (6). In that case, partial differential equations or Monte Carlo simulations are being used. For option pricing, this would mean to simulate scenarios from futures prices. This translates into intensive computer simulation procedures.

5. The Implied Market Price of Weather Risk

For pricing and hedging non-tradable assets, one essentially needs to incorporate the MPR θ which is an important parameter of the associated EMM and it measures the additional return for bearing more risk. This section deals exactly with the differences between 'historical' (P) and 'risk-neutral' (Q) behaviours of temperature. Using statistical modelling and given that liquid-derivative contracts based on daily temperatures are traded on the CME, one might infer the MPR (the change of drift) from traded (CAT/CDD/HDD/C24AT) futures–options-type contracts.

Our study is a calibration procedure for financial engineering purposes. In the calibration exercise, a single date (but different time horizons and calibrated instruments are used) is required, since the model is recalibrated daily to detect intertemporal effects. Moreover, we use an economic and statistical testing approach, where we start from a specification of the MPR and check consistency with the data. By giving assumptions about the MPR, we implicitly make an assumption about the aggregate risk aversion of the market. The risk parameter θ can then be inferred by finding the value that satisfies Equation (20) for each specification. Once we know the MPR for temperature futures, then we know the MPR for options and thus one can price new 'non-standard maturities' or OTC derivatives. The concept of implied MPR is similar to that used in extracting implied volatilities (Fengler *et al*., 2007) or the market price of oil risk (Gibson and Schwartz, 1990).

To value temperature derivatives, the following specifications of the MPR are investigated: a constant, a piecewise linear function, a two-piecewise linear function, a time-deterministic function and a 'financial-bootstrapping' MPR. The statistical point of view is to beat this as an inverse problem with different degrees of smoothness expressed through the penalty parameter of a smoothing spline. The economic point of view is to detect possible time dependencies that can be explained by investor's preferences in order to hedge weather risk.

In this article we concentrate on contracts with monthly measurement length periods, but similar implications apply for seasonal strip contracts. We observe different temperature futures contracts $i = 1, \ldots, I$ with measurement periods $t \leq \tau_1^i < \tau_2^i$ and $\tau_2^i \leq \tau_1^{i+1}$ traded at time t, meaning that contracts expire at some point in time and roll over to another contract. Therefore, $i = 1$ denotes contract types with measurement period in 30 days, $i = 2$ denotes contract types in 60 days and so on. For example, a contract with $i = 7$ is six months ahead from the trading day t. For United States and Europe, the number of temperature futures contracts is $I = 7$ (April–October or October–April), while for Asia $I = 12$ (January–December). The details of the temperature futures data are displayed in Table 4. To simplify notation, dates are written in yyyymmdd format.

5.1 Constant MPR for Each Contract per Trading Date

Given observed temperature futures market prices and by inverting Equation (20), we imply the MPR θ_u for $i = 1, \ldots, I$ futures contracts with different measurement time horizon periods $[\tau_1^i, \tau_2^i]$, $t \leq \tau_1^i < \tau_2^i$ and $\tau_2^i \leq \tau_1^{i+1}$ traded at date t. Our first assumption is to set, for the ith contract, a constant MPR over $[t, \tau_2^i]$, that is, we have that $\theta_u = \theta_t^i$:

$$\hat{\theta}_{t,CAT}^i = \underset{\theta_t^i}{\arg\min} \left(F_{CAT(t,\tau_1^i,\tau_2^i)} - \int_{\tau_1^i}^{\tau_2^i} \hat{\Lambda}_u du - \mathbf{a}_{t,\tau_1^i,\tau_2^i} \boldsymbol{X}_t - \theta_t^i \left\{ \int_t^{\tau_1} \hat{\sigma}_u \mathbf{a}_{t,\tau_1^i,\tau_2^i} \mathbf{e}_p du \right.\right.$$

$$\left.\left. + \int_{\tau_1^i}^{\tau_2^i} \hat{\sigma}_u \mathbf{e}_1^\top \mathbf{A}^{-1} \left[\exp\left\{ \mathbf{A}(\tau_2^i - u) \right\} - I_p \right] \mathbf{e}_p du \right\} \right)^2,$$

Table 4. Weather futures and futures prices listed on date (yyyymmdd) at CME.

	Trading date	Measurement period		Futures prices $F_{(t,\tau_1,\tau_2,\hat{\theta})}$			Realized T_t
Contract type	t	τ_1	τ_2	CME	MPR = 0	Constant MPR	$I_{(\tau_1,\tau_2)}$
Berlin-CAT	20070316	20070401	20070430	288.00	363.00	291.06	362.90
Berlin-CAT	20070316	20070501	20070531	457.00	502.11	454.91	494.20
Berlin-CAT	20070316	20070601	20070630	529.00	571.78	630.76	574.30
Berlin-CAT	20070316	20070701	20070731	616.00	591.56	626.76	583.00
Berlin-CAT	20070316	20070801	20070831	610.00	566.14	636.22	580.70
Berlin-CAT	20070316	20070901	20070930	472.00	414.33	472.00	414.80
Berlin-CAT	20070427	20070501	20070531	457.00	506.18	457.52	494.20
Berlin-CAT	20070427	20070601	20070630	529.00	571.78	534.76	574.30
Berlin-CAT	20070427	20070701	20070731	616.00	591.56	656.76	583.00
Berlin-CAT	20070427	20070801	20070831	610.00	566.14	636.22	580.70
Berlin-CAT	20070427	20070901	20070930	472.00	414.33	472.00	414.80
Tokyo-C24AT	20081027	20090301	20090331	450.00	118.32	488.90	305.00
Tokyo-C24AT	20081027	20090401	20090430	592.00	283.18	563.27	479.00
Tokyo-C24AT	20081027	20090501	20090531	682.00	511.07	696.31	623.00
Tokyo-C24AT	20081027	20090601	20090630	818.00	628.24	835.50	679.00
Tokyo-C24AT	20081027	20090701	20090731	855.00	731.30	706.14	812.00

Notes: CME, Chicago Mercantile Exchange; MPR, market price of risk.
Weather futures at CME; futures prices $F_{(t,\tau_1,\tau_2,\hat{\theta})}$ from CME; estimated prices with MPR = 0; constant MPR for different controls per trading date (constant MPR).
Source: Bloomberg Professional Service (Weather futures).

$$\hat{\theta}^i_{t,HDD} = \arg\min_{\theta^i_t} \left(F_{HDD(t,\tau^i_1,\tau^i_2)} - \int_{\tau^i_1}^{\tau^i_2} \upsilon_{t,s}\psi \left[\frac{c - \hat{m}^1_{\{t,s,e_1^\top \exp\{\mathbf{A}(s-t)\}\boldsymbol{X}_t\}}}{\upsilon_{t,s}} \right] ds \right)^2, \quad (21)$$

with $\hat{m}^1_{\{t,s,x\}} = \Lambda_s + \theta^i_t \int_t^s \sigma_u \mathbf{e}_1^\top \exp\{\mathbf{A}(s-t)\}\,\mathbf{e}_p du + x$, $\upsilon^2_{t,s}$, $\psi(x)$ and x defined as in Equation (20). The MPR for CDD futures $\hat{\theta}^i_{t,CDD}$ is equivalent to the HDD case in Equation (21) and we will therefore omit CDD parameterizations. Note that this specification can be seen as a deterministic time-varying MPR θ^i_t that varies with date for any given contract i, but it is constant over $[t, \tau^i_2]$.

5.2 One Piecewise Constant MPR

A simpler MPR parameterization is to assume that it is constant across all time horizon contracts priced in a particular date (θ_t). We therefore estimate this constant MPR for all contract types traded at $t \le \tau^i_1 < \tau^i_2$, $i = 1, \ldots, I$ as follows:

$$\hat{\theta}_{t,CAT} = \arg\min_{\theta_t} \sum_{i=1}^{I} \left(F_{CAT(t,\tau^i_1,\tau^i_2)} - \int_{\tau^i_1}^{\tau^i_2} \hat{\Lambda}_u du - \hat{\mathbf{a}}_{t,\tau^i_1,\tau^i_2} \boldsymbol{X}_t - \theta_t \left\{ \int_t^{\tau^i_1} \hat{\sigma}_u \hat{\mathbf{a}}_{t,\tau^i_1,\tau^i_2} \mathbf{e}_p du \right.\right.$$
$$\left.\left. + \int_{\tau^i_1}^{\tau^i_2} \hat{\sigma}_u \mathbf{e}_1^\top \mathbf{A}^{-1} \left[\exp\left\{\mathbf{A}(\tau^i_2 - u)\right\} - I_p \right] \mathbf{e}_p du \right\} \right)^2,$$

$$\hat{\theta}_{t,HDD} = \arg\min_{\theta_t} \sum_{i=1}^{I} \left(F_{HDD(t,\tau^i_1,\tau^i_2)} - \int_{\tau^i_1}^{\tau^i_2} \upsilon_{t,s}\psi \left[\frac{c - \hat{m}^2_{\{t,s,\mathbf{e}_1^\top \exp\{\mathbf{A}(s-t)\}\boldsymbol{X}_t\}}}{\upsilon_{t,s}} \right] ds \right)^2, \quad (22)$$

with $\hat{m}^2_{\{t,s,x\}} = \Lambda_s + \theta_t \int_t^s \sigma_u \mathbf{e}_1^\top \exp\{\boldsymbol{A}(s-t)\}\,\mathbf{e}_p du + x$ and $\upsilon^2_{t,s}$, $\psi(x)$ and x as defined in Equation (20). This 'one piecewise constant' MPR specification (θ_t) is solved by means of the ordinary least squares (OLS) minimization procedure and differs from θ^i_t in Equation (21) because for all traded contracts at date t, we get only one MPR estimate (instead of i estimates) at time t, that is, θ_t is constant over $[t, \tau^I_2]$.

5.3 Two Piecewise Constant MPR

Assuming now that, instead of one constant MPR per trading day, we have a step function with a given jump point ξ (take e.g. the first 150 days before the beginning of the measurement period), so we have that $\hat{\theta}_t = I(u \le \xi)\,\theta^1_t + I(u > \xi)\,\theta^2_t$. The two piecewise constant function $\hat{\theta}_t$ with $t \le \tau^i_1 < \tau^i_2$ is estimated with the OLS minimization procedure as follows:

$$
\begin{aligned}
f_{CAT}(\xi) = \underset{\theta^1_{t,CAT},\theta^2_{t,CAT}}{\arg\min} \sum_{i=1}^{I} \Bigg(& F_{CAT(t,\tau_1^i,\tau_2^i)} - \int_{\tau_1^i}^{\tau_2^i} \hat{\Lambda}_u du - \hat{\mathbf{a}}_{t,\tau_1^i,\tau_2^i} \boldsymbol{X}_t \\
& - \theta^1_{t,CAT} \left\{ \int_t^{\tau_1^i} I\left(u \le \xi\right) \hat{\sigma}_u \hat{\mathbf{a}}_{t,\tau_1^i,\tau_2^i} \mathbf{e}_p du \right. \\
& \left. + \int_{\tau_1^i}^{\tau_2^i} I\left(u \le \xi\right) \hat{\sigma}_u \mathbf{e}_1^\top \mathbf{A}^{-1} \left[\exp\left\{ \mathbf{A}(\tau_2^i - u) \right\} - I_p \right] \mathbf{e}_p du \right\} \\
& - \theta^2_{t,CAT} \left\{ \int_t^{\tau_1^i} I\left(u > \xi\right) \hat{\sigma}_u \hat{\mathbf{a}}_{t,\tau_1^i,\tau_2^i} \mathbf{e}_p du \right. \\
& \left. + \int_{\tau_1^i}^{\tau_2^i} I\left(u > \xi\right) \hat{\sigma}_u \mathbf{e}_1^\top \mathbf{A}^{-1} \left[\exp\left\{ \mathbf{A}(\tau_2^i - u) \right\} - I_p \right] \mathbf{e}_p du \right\} \Bigg)^2,
\end{aligned} \tag{23}
$$

$$
f_{HDD}(\xi) = \underset{\theta^1_{t,HDD},\theta^2_{t,HDD}}{\arg\min} \sum_{i=1}^{I} \left(F_{HDD(t,\tau_1^i,\tau_2^i)} - \int_{\tau_1^i}^{\tau_2^i} \upsilon_{t,s} \psi \left[\frac{c - \hat{m}^3_{\left\{t,s,e_1^\top \exp\{\mathbf{A}(s-t)\} \boldsymbol{X}_t\right\}}}{\upsilon_{t,s}} \right] ds \right)^2,
$$

$$
\begin{aligned}
\hat{m}^3_{\{t,s,x\}} = & \Lambda_s + \theta^1_{t,HDD} \left\{ \int_t^s I\left(u \le \xi\right) \sigma_u \mathbf{e}_1^\top \exp\left\{\mathbf{A}(s-t)\right\} \mathbf{e}_p du + x \right\} \\
& + \theta^2_{t,HDD} \left\{ \int_t^s I\left(u > \xi\right) \sigma_u \mathbf{e}_1^\top \exp\left\{\mathbf{A}(s-t)\right\} e_p du + x \right\},
\end{aligned}
$$

and $\upsilon^2_{t,s}$, $\psi(x)$ and x as defined in Equation (20). In the next step, we optimized the value of ξ such as $f_{CAT}(\xi)$ or $f_{HDD}(\xi)$ is minimized. This MPR specification will vary according to the unknown ξ. This would mean that the market does a risk adjustment for contracts traded close or far from the measurement period.

5.4 General Form of the MPR per Trading Day

Generalizing the piecewise continuous function given in the previous subsection, the (inverse) problem of determining θ_t with $t \le \tau_1^i < \tau_2^i$, $i = 1, \ldots, I$, can be formulated via a series expansion for θ_t:

$$
\begin{aligned}
\underset{\gamma_k}{\arg\min} \sum_{i=1}^{I} \Bigg(& F_{CAT(t,\tau_1^i,\tau_2^i)} - \int_{\tau_1^i}^{\tau_2^i} \hat{\Lambda}_u du - \hat{\mathbf{a}}_{t,\tau_1^i,\tau_2^i} \hat{\mathbf{X}}_t - \int_t^{\tau_1^i} \sum_{k=1}^{K} \gamma_k h_k(u_i) \hat{\sigma}_{u_i} \hat{\mathbf{a}}_{t,\tau_1^i,\tau_2^i} \mathbf{e}_p du_i \\
& - \int_{\tau_1^i}^{\tau_2^i} \sum_{k=1}^{K} \gamma_k h_k(u_i) \hat{\sigma}_{u_i} \mathbf{e}_1^\top \mathbf{A}^{-1} \left[\exp\left\{ \mathbf{A}(\tau_2^i - u_i) \right\} - I_p \right] \mathbf{e}_p du_i \Bigg)^2,
\end{aligned}
$$

$$\arg\min_{a_k} \sum_{i=1}^{I} \left(F_{HDD(t,\tau_1^i,\tau_2^i)} - \int_{\tau_1^i}^{\tau_2^i} \upsilon_{t,s} \psi \left[\frac{c - \hat{m}^4_{\{t,s,\mathbf{e}_1^\top \exp\{\mathbf{A}(s-t)\}\mathbf{X}_t\}}}{\upsilon_{t,s}} \right] ds \right)^2, \quad (24)$$

with $\hat{m}^4_{\{t,s,x\}} = \Lambda_s + \int_t^s \sum_{k=1}^K a_k l_k(u_i) \hat{\sigma}_{u_i} \mathbf{e}_1^\top \exp\{\mathbf{A}(s-t)\} \mathbf{e}_p du_i + x$ and $\upsilon^2_{t,s}$, $\psi(x)$ and x as defined in Equation (20). $h_k(u_i)$ and $l_k(u_i)$ are vectors of known basis functions and may denote a B-spline basis for example. γ_k and a_k define the coefficients and K is the number of knots. This means that the inferred MPR is going to be a solution for an inverse problem with different degrees of smoothness expressed through the penalty parameter of a smoothing spline. The degrees of smoothness will allow for a term structure of risk. In other words, a time-dependent risk factor offers the possibility to have different risk adjustments for different times of the year.

5.5 Bootstrapping the MPR

In this section we propose a bootstrapping technique to detect possible MPR time-dependent paths of temperature futures contracts. More importantly, since these futures contract types have different measurement periods $[\tau_1^i, \tau_2^i]$ with $\tau_1^i < \tau_1^{i+1} \leq \tau_2^i < \tau_2^{i+1}$, $i = 1, \ldots, I$, and they roll over to another contracts when they expire at some point in time, it makes sense to construct MPR estimates from which we can price contracts with any maturity, without the need of external information. This 'financial' bootstrapping idea consists of estimating by forward substitution the MPR θ_t^i of the futures price contracts with the closest measurement period and placing it into the estimation for the next MPR θ_t^{i+1}. We implement the estimation for CAT contracts, but the idea applies also for HDD/CDD contract types. First, for the first contract $i = 1$ and $t \in [\tau_1^1, \tau_2^1]$, $\hat{\theta}^1_{t,CAT}$ is estimated from Equation (21):

$$\hat{\theta}^1_{t,CAT} = \arg\min_{\theta_t^1} \left(F_{CAT(t,\tau_1^1,\tau_2^1)} - \int_{\tau_1^1}^{\tau_2^1} \hat{\Lambda}_u du - \hat{\mathbf{a}}_{t,\tau_1^1,\tau_2^1} \hat{\mathbf{X}}_t - \theta_t^1 \left\{ \int_t^{\tau_1^1} \hat{\sigma}_u \hat{\mathbf{a}}_{t,\tau_1^1,\tau_2^1} \mathbf{e}_p du \right.\right.$$
$$\left.\left. + \int_{\tau_1^1}^{\tau_2^1} \hat{\sigma}_u \mathbf{e}_1^\top \mathbf{A}^{-1} \left[\exp\left\{ \mathbf{A}(\tau_2^1 - u) \right\} - I_p \right] \mathbf{e}_p du \right\} \right)^2. \quad (25)$$

Second, the estimated $\hat{\theta}^1_{t,CAT}$ is substituted in the period $[\tau_1^1, \tau_2^1]$ to get an estimate of $\hat{\theta}^2_{t,CAT}$:

$$\hat{\theta}^2_{t,CAT} = \arg\min_{\theta^2_{t,CAT}} \left(F_{CAT(t,\tau_1^2,\tau_2^2)} - \int_{\tau_1^2}^{\tau_2^2} \hat{\Lambda}_u du - \hat{\mathbf{a}}_{t,\tau_1^2,\tau_2^2} \hat{\mathbf{X}}_t - \int_t^{\tau_1^1} \hat{\theta}^1_{t,CAT} \hat{\sigma}_u \hat{\mathbf{a}}_{t,\tau_1^2,\tau_2^2} \mathbf{e}_p du \right.$$
$$\left. - \int_{\tau_1^2}^{\tau_2^2} \theta^2_{t,CAT} \hat{\sigma}_u \mathbf{e}_1^\top \mathbf{A}^{-1} \left[\exp\left\{ \mathbf{A}(\tau_2^2 - u) \right\} - I_p \right] \mathbf{e}_p du \right)^2. \quad (26)$$

Then substitute $\hat{\theta}^1_{t,CAT}$ in the period $[\tau_1^1, \tau_2^1]$ and $\hat{\theta}^2_{t,CAT}$ in the period $[\tau_1^2, \tau_2^2]$ to estimate $\hat{\theta}^3_{t,CAT}$:

$$\hat{\theta}^3_{t,CAT} = \underset{\hat{\theta}^3_{t,CAT}}{\arg\min} \left(F_{CAT(t,\tau_1^3,\tau_2^3)} - \int_{\tau_1^3}^{\tau_2^3} \hat{\Lambda}_u du - \hat{\mathbf{a}}_{t,\tau_1^3,\tau_2^3} \boldsymbol{X}_t - \int_t^{\tau_1^1} \hat{\theta}^1_{t,CAT} \hat{\sigma}_u \hat{\mathbf{a}}_{t,\tau_1^3,\tau_2^3} \mathbf{e}_p du \right.$$

$$\left. - \int_{\tau_1^2}^{\tau_2^2} \hat{\theta}^2_{t,CAT} \hat{\sigma}_u \hat{\mathbf{a}}_{t,\tau_1^3,\tau_2^3} \mathbf{e}_p du - \int_{\tau_1^3}^{\tau_2^3} \theta^3_{t,CAT} \hat{\sigma}_u \mathbf{e}_1^\top \mathbf{A}^{-1} \left[\exp\left\{ \mathbf{A}(\tau_2^3 - u) \right\} - I_p \right] \mathbf{e}_p du \right)^2.$$

In a similar way, one obtains the estimation of $\hat{\theta}^4_{t,CAT}, \ldots, \hat{\theta}^I_{t,CAT}$.

5.6 Smoothing the MPR over Time

Since smoothing individual estimates is different from estimating a deterministic function, we also assure our results by fitting a parametric function to all available contract prices (calendar year estimation). After computing the MPR $\hat{\theta}_{t,CAT}, \hat{\theta}_{t,HDD}$ and $\hat{\theta}_{t,CDD}$ for each of the previous specification and for each of the nth trading days t for different ith contracts, the MPR time series can be smoothed with the inverse problem points to find an MPR $\hat{\theta}_u$ for every calendar day u and with that being able to price temperature derivatives for any date:

$$\underset{f \in F_j}{\arg\min} \sum_{t=1}^{n} \left\{ \hat{\theta}_t - f(u_t) \right\}^2 = \underset{\alpha_j}{\arg\min} \sum_{t=1}^{n} \left\{ \hat{\theta}_t - \sum_{j=1}^{J} \alpha_j \Psi_j(u_t) \right\}^2, \tag{27}$$

where $\Psi_j(u_t)$ is a vector of known basis functions, α_j defines the coefficients, J is the number of knots, $u_t = t - \Delta + 1$ with increment Δ and n is the number of days to be smoothed. In our case, $u_t = 1$ day and $\Psi_j(u_t)$ is estimated using cubic splines.

Alternatively, one can first do the smoothing with basis functions of all available futures contracts:

$$\underset{\beta_j}{\arg\min} \sum_{t=1}^{n} \sum_{i=1}^{I} \left\{ F_{(t,\tau_1^i,\tau_2^i)} - \sum_{j=1}^{J} \beta_j \Psi_j(u_t) \right\}^2, \tag{28}$$

and then estimate the time series of $\hat{\theta}^s_t$s with the obtained smoothed futures prices $F^s_{(t,\tau_1^1,\tau_2^I)}$.

For example, for a constant MPR for all CAT futures contracts type traded over all ts with $t \leq \tau_1^i < \tau_2^i$ and $\tau_2^i \leq \tau_1^{i+1}$, we have:

$$\hat{\theta}^s_{t,CAT} = \underset{\theta^s_{t,CAT}}{\arg\min} \left(F^s_{CAT(t,\tau_1^1,\tau_2^I)} - \int_{\tau_1^1}^{\tau_2^I} \hat{\Lambda}_u du - \hat{\mathbf{a}}_{t,\tau_1^1,\tau_2^I} \boldsymbol{X}_t - \theta^s_{t,CAT} \left\{ \int_t^{\tau_1^1} \hat{\sigma}_u \hat{\mathbf{a}}_{t,\tau_1^1,\tau_2^I} \mathbf{e}_p du \right.\right.$$

$$\left.\left. + \int_{\tau_1^1}^{\tau_2^I} \hat{\sigma}_u \mathbf{e}_1^\top \mathbf{A}^{-1} \left[\exp\left\{ \mathbf{A}(\tau_2^I - u) \right\} - I_p \right] \mathbf{e}_p du \right\} \right)^2 . \tag{29}$$

5.7 Statistical and Economical Insights of the Implied MPR

In this section, using the previous specifications, we imply the MPR (the change of drift) for CME (CAT/CDD/HDD/C24AT) futures contracts traded for different cities. Note that one might also infer the MPR from options data and compare the findings with prices in the futures market.

Table 5 presents the descriptive statistics of different MPR specifications for Berlin-CAT, Essen-CAT and Tokyo-C24AT daily futures contracts with $t \leq \tau_1^i < \tau_2^i$ traded during 20031006–20080527 (5102 contracts in 1067 trading days with 29 different measurement periods), 20050617–20090731 (3530 contracts in 926 trading dates with 28 measurement periods) and 20040723–20090831 (2611 contracts in 640 trading dates with 27 measurement periods). The MPR ranges vary between [–10.71, 10.25], [31.05, 5.73] and [–82.62, 52.17] for Berlin-CAT, Essen-CAT and Tokyo-C24AT futures contracts, respectively, whereas the MPR averages are 0.04, 0.00 and –3.08 for constant MPR for different contracts; –0.08, –0.38 and 0.73 for one piecewise constant; –0.22, –0.43 and –3.50 for two piecewise constant; 0.04, 0.00 and –3.08 for spline; and 0.07, 0.00 and –0.11 when bootstrapping the MPR. We observe that the two piecewise constant MPR function is a robust least square estimation, since its values are sensitive to the choice of ξ. Figure 6 shows the MPR estimates for Berlin-CAT futures prices traded on 20060530 with $\xi = 62, 93, 123$ and 154 and sum of squared errors equal to 2759, 14,794, 15,191 and 15,526. The line displays a discontinuity indicating that trading was not taking place (CAT futures are only traded from April to November and MPR estimates cannot be computed since there are no market prices). When the jump ξ is getting far from the measurement period, the value of the MPR $\hat{\theta}_t^1$ decreases and $\hat{\theta}_t^2$ increases, yielding a $\hat{\theta}_t$ around 0. Table 5 also displays the estimates of the time-dependent MPR (or spline MPR) from the bootstrapping technique. The spline MPR smooths the estimates over time and it is estimated using cubic polynomials with k equal to the number of traded contracts I at date t. The performance of the boostrapped MPR is similar to the constant MPR for different contracts per trading date estimates, suggesting that the only risk which the statistical model might imply is that the MPR will be equal at any trading date across all temperature contract types.

The first panel in Figure 7 displays the Berlin-CAT, Essen-CAT and Tokyo-C24AT futures contracts traded at 20060530, 20060530 and 20050531, respectively. The second, third and fourth panels of Figure 7 show the MPR when it is assumed to be constant for different contracts per trading date, a two piecewise constant and the spline MPR. In the case of the constant MPR for different contracts per trading date, the lines overlap because the MPR for every contract $i = 1, \ldots, 12$ is supposed to be constant over the period $[t, \tau_2^i]$ at trading date t. The two piecewise constant function adjusts the risk according to the choice of ξ

Table 5. Statistics of MPR specifications for Berlin-CAT, Essen-CAT and Tokyo-C24AAT.

Type	No. of contracts	Statistic	Constant	1 piecewise	2 piecewise	Bootstrap	Spline
Berlin-CAT		WS (Prob)	0.93 (0.66)	0.01 (0.08)	0.00 (0.04)	0.93 (0.04)	0.91 (0.66)
30 days	487	Min (Max)	−0.28 (0.12)	−0.65 (0.11)	−1.00 (0.20)	−0.28 (0.12)	0.02 (0.21)
($i=1$)		Med (SD)	0.05 (0.07)	−0.10 (0.12)	−0.17 (0.31)	0.05 (0.07)	0.14 (0.06)
60 days	874	Min (Max)	−0.54 (0.84)	−4.95 (8.39)	−10.71 (10.25)	−0.54 (0.84)	0.00 (0.21)
($i=2$)		Med (SD)	0.03 (0.13)	−0.09 (1.06)	−0.17 (1.44)	0.03 (0.13)	0.05 (0.05)
90 days	858	Min (Max)	−0.54 (0.82)	−4.95 (8.39)	−10.71 (10.25)	−0.54 (0.82)	0.00 (0.21)
($i=3$)		Med (SD)	0.02 (0.09)	−0.10 (1.07)	−0.17 (1.46)	0.02 (0.09)	0.11 (0.07)
120 days	815	Min (Max)	−0.53 (0.84)	−4.95 (4.56)	−7.71 (6.88)	−0.53 (0.84)	0.00 (0.21)
($i=4$)		Med (SD)	0.03 (0.14)	−0.10 (0.94)	−0.19 (1.38)	0.03 (0.14)	0.11 (0.09)
150 days	752	Min (Max)	−0.54 (0.84)	−4.53 (4.56)	−8.24 (6.88)	−0.54 (0.84)	−0.00 (0.20)
($i=5$)		Med (SD)	0.13 (0.09)	−0.10 (0.95)	−0.20 (1.47)	0.13 (0.09)	0.01 (0.07)
180 days	711	Min (Max)	−0.54 (0.82)	−4.53 (4.56)	−8.24 (6.88)	−0.54 (0.82)	−0.03 (0.12)
($i=6$)		Med (SD)	0.02 (0.12)	−0.10 (0.95)	−0.14 (1.48)	0.02 (0.12)	0.00 (0.03)
Essen-CAT		WS (Prob)	0.02 (0.11)	0.20 (0.34)	0.01 (0.09)	0.02 (0.11)	1.63 (0.79)
30 days	384	Min (Max)	−0.98 (0.52)	−31.05 (5.73)	−1.83 (1.66)	−0.98 (0.52)	−0.00 (0.00)
($i=1$)		Med (SD)	0.01 (0.12)	−0.39 (1.51)	−0.43 (9.62)	0.01 (0.12)	0.00 (0.00)
60 days	796	Min (Max)	−1.35 (0.62)	−31.05 (5.73)	−1.83 (1.66)	−1.35 (0.62)	0.00 (0.00)
($i=2$)		Med (SD)	0.02 (0.14)	−0.40 (1.56)	−0.46 (9.99)	0.02 (0.14)	0.00 (0.00)
90 days	738	Min (Max)	−1.56 (0.59)	−6.68 (5.14)	−5.40 (1.71)	−1.56 (0.59)	0.00 (0.00)
($i=3$)		Med (SD)	−0.02 (0.19)	−0.40 (0.88)	−0.40 (0.85)	−0.02 (0.19)	0.00 (0.00)
120 days	551	Min (Max)	−0.29 (0.51)	−4.61 (1.44)	−6.60 (1.43)	−0.29 (0.51)	0.00 (0.00)
($i=4$)		Med (SD)	0.03 (0.05)	−0.37 (0.51)	−0.41 (0.85)	0.03 (0.05)	0.00 (0.00)
150 days	468	Min (Max)	−0.44 (0.13)	−4.61 (1.44)	−6.60 (1.43)	−0.44 (0.13)	0.00 (0.00)
($i=5$)		Med (SD)	0.00 (0.07)	−0.35 (0.49)	−0.30 (0.81)	0.00 (0.07)	0.00 (0.00)
180 days	405	Min (Max)	−0.10 (0.57)	−4.61 (1.44)	−4.25 (0.52)	−0.10 (0.57)	0.00 (0.00)
($i=6$)		Med (SD)	−0.02 (0.06)	−0.12 (0.45)	−0.21 (0.52)	−0.02 (0.06)	0.00 (0.00)
Tokyo-C24AT		WS (Prob)	0.76 (0.61)	0.02 (0.13)	0.01 (0.11)	0.76 (0.61)	4.34 (0.96)
30 days	419	Min (Max)	−7.55 (0.17)	−69.74 (52.17)	−69.74 (52.17)	−7.55 (0.17)	−0.23 (−0.18)

(*Continued*)

Table 5. (Continued).

Type	No. of contracts	Statistic	Constant	1 piecewise	2 piecewise	Bootstrap	Spline
($i=2$)		Med (SD)	−3.87 (2.37)	−0.33 (19.68)	−0.48 (20.46)	−3.87 (2.37)	−0.20 (0.01)
60 days	416	Min (Max)	−7.56 (0.14)	−69.74 (52.17)	−69.74 (52.17)	−7.56 (0.14)	−0.18 (−0.13)
($i=3$)		Med (SD)	−3.49 (2.47)	−0.23 (21.34)	−0.41 (21.63)	−3.49 (2.47)	−0.15 (0.01)
90 days	393	Min (Max)	−7.55 (1.02)	−69.74 (26.82)	−69.74 (38.53)	−7.55 (1.02)	−0.13 (−0.09)
($i=4$)		Med (SD)	−2.96 (2.65)	0.04 (20.84)	−0.33 (20.22)	−2.96 (2.65)	−0.11 (0.01)
120 days	350	Min (Max)	−7.55 (1.02)	−69.74 (26.82)	−69.74 (48.32)	−7.55 (1.02)	−0.10 (−0.06)
($i=5$)		Med (SD)	−2.08 (2.74)	1.26 (19.54)	−0.11 (19.69)	−2.08 (2.74)	−0.08 (0.01)
150 days	305	Min (Max)	−7.55 (1.02)	−51.18 (26.82)	−51.18 (48.32)	−7.55 (1.02)	−0.06 (−0.04)
($i=6$)		Med (SD)	−2.08 (2.71)	1.26 (16.03)	7.17 (17.02)	−2.08 (2.71)	−0.05 (0.00)
180 days	243	Min (Max)	−7.39 (1.26)	−51.18 (19.10)	−51.18 (48.32)	−7.39 (1.26)	−0.05 (−0.04)
($i=7$)		Med (SD)	−2.08 (2.74)	3.66 (15.50)	7.63 (17.69)	−2.08 (2.74)	−0.04 (0.00)
210 days	184	Min (Max)	−7.39 (1.26)	−24.88 (26.16)	−54.14 (39.65)	−7.39 (1.26)	−0.07 (−0.05)
($i=8$)		Med (SD)	−3.00 (2.86)	13.69 (10.05)	10.61 (14.63)	−3.00 (2.86)	−0.06 (0.00)
240 days	167	Min (Max)	−7.39 (1.26)	−24.88 (21.23)	−82.62 (42.14)	−7.39 (1.26)	−0.07 (−0.07)
($i=9$)		Med (SD)	−3.00 (2.74)	3.46 (11.73)	−4.24 (37.25)	−3.00 (2.74)	−0.07 (0.00)
270 days	134	Min (Max)	−7.39 (0.44)	−24.88 (26.16)	−82.62 (42.14)	−7.39 (0.44)	−0.07 (−0.03)
($i=10$)		Med (SD)	−4.24 (2.39)	8.48 (13.88)	−7.39 (40.76)	−4.24 (2.39)	−0.05 (0.00)

Notes: WS, Wald statistics; SD, standard deviation.
Futures contracts traded during (20031006–20080527), (20050617–20090731), and (20040723–20090630) respectively, with trading date before measurement period $t \leq \tau_1^i < \tau_2^i, i = 1, \ldots, I$ (where $i = 1$ (30 days), $i = 2$ (60 days), $\ldots, i = I$ (210 days)): the WS, the WS probabilities (Prob), Minimun (Min), Maximum (Max), Median (Med) and SD. MPR specifications: Constant for different contracts per trading date (Constant), 1 piecewise constant, 2 piecewise constant ($\xi = 150$ days), bootstrap and spline.

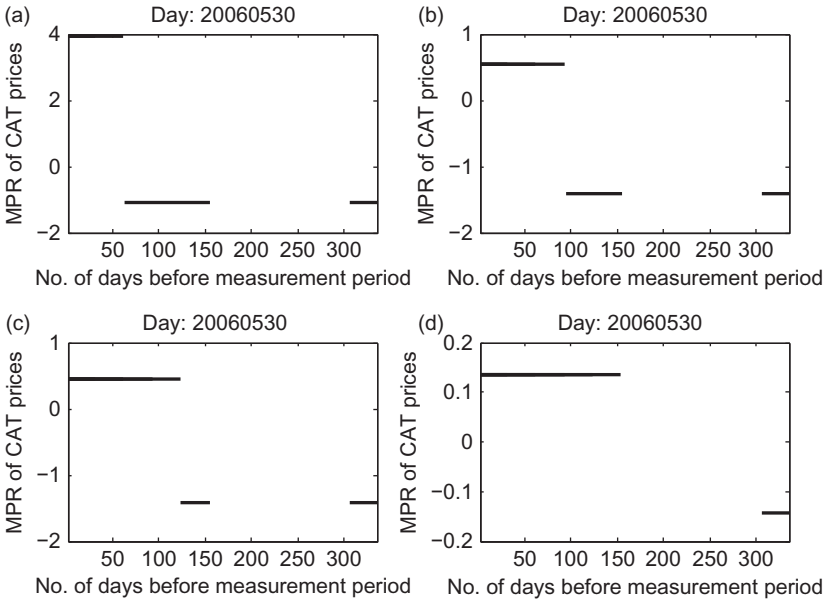


Figure 6. Two piecewise constant MPR with jumps $\xi =$ (a) 62, (b) 93, (c) 123 and (d) 154 days for Berlin-CAT contracts traded on 20060530. The corresponding sum of squared errors are 2759, 14794, 15191 and 15526. When the jump ξ is getting far from the measurement period, the value of the MPR $\hat{\theta}_t^1$ decreases and $\hat{\theta}_t^2$ increases, yielding a $\hat{\theta}_t$ around 0.

(in this case $\xi = 150$ days). The spline MPR smooths over time and for days without trading (see the case of Berlin-CAT or Essen-CAT futures), it displays a maximum, for example, in winter. A penalizing term in Equation (24) might correct for this.

In all the specifications, we verified the discussion that MPR is different from 0 (as Cao and Wei (2004), Huang-Hsi *et al*. (2008), Richards *et al*. (2004) and Alaton *et al*. (2002) do) varies in time and moves from a negative to a positive domain according to the changes in the seasonal variation. The MPR specifications change signs when a contract expires and rolls over to another contract (e.g. from 210 to 180, 150, 120, 90, 60, 30 days before measurement period); they react negatively to the fast changes in seasonal variance σ_t within the measurement period (Figure 3) and to the changes in CAT futures volatility $\sigma_t \mathbf{a}_{t,\tau_1,\tau_2} \mathbf{e}_p$. Figure 8 shows the Berlin-CAT volatility paths for contracts issued before and within the measurement periods 2004–2008. We observed the Samuelson effect for mean-reverting futures: for contracts traded within the measurement period, CAT volatility is close to 0 when the time to measurement is large and it decreases up to the end of the measurement period. For contracts traded before the measurement period, CAT volatility is also close to 0 when the time to measurement is large, but increases up to the start of the measurement period. In Figure 9, two Berlin-CAT contracts issued on 20060517 but with different measurement periods are plotted: the longest the measurement period, the largest the volatility. Besides this, one observes the effect of the $CAR(3)$ in both contracts when the volatility decays just before maturity of the contracts. These two effects are comparable with the study for Stockholm CAT futures in Benth *et al*. (2007); however, the deviations are less smoothed for Berlin.

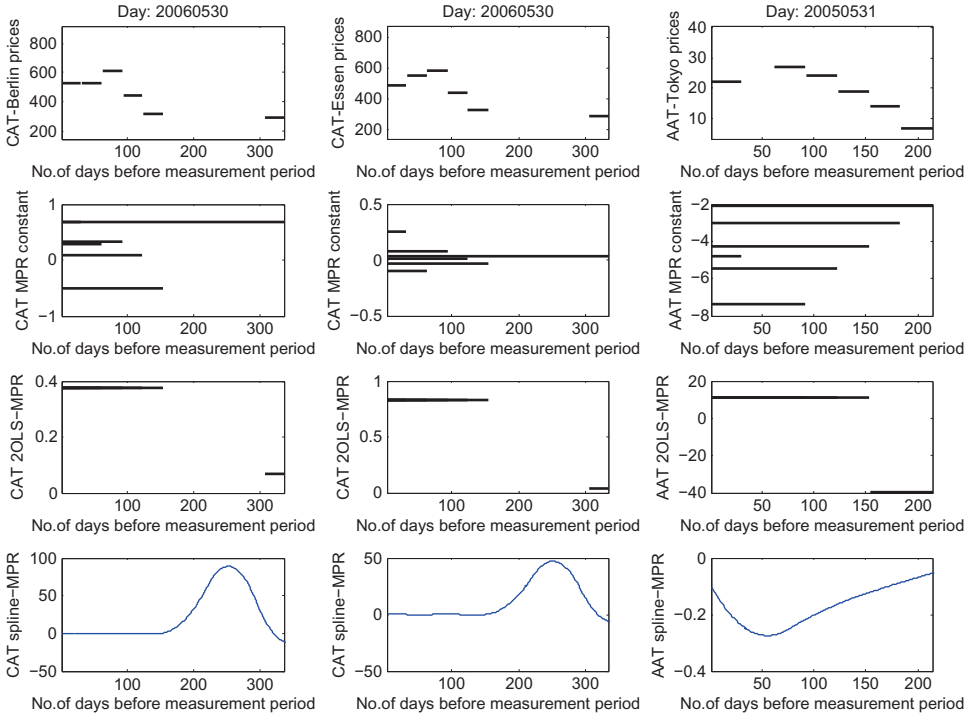


Figure 7. Futures CAT prices (1 row panel) and MPR specifications: constant MPR for different contracts per trading day, two piecewise constant and spline (2, 3 and 4 row panel) for Berlin-CAT (left), Essen-CAT (middle), Tokyo-AAT (right) of futures traded on 20060530, 20060530 and 20060531, respectively.

We investigate the proposition that the MPR derived from CAT/HDD/CDD futures is different from 0. We conduct the Wald statistical test to check whether this effect exists by testing the true value of the parameter based on the sample estimate. In the multivariate case, the Wald statistic for $\left\{\theta_t \in \mathbb{R}^i\right\}_{t=1}^n$ is

$$(\hat{\theta}_t - \theta_0)^\top \Sigma (\hat{\theta}_t - \theta_0) \sim \chi_p^2, \Sigma^{\frac{1}{2}}(\hat{\theta}_t - \theta_0) \sim N\,(0, \mathrm{I}_i),$$

where Σ is the variance matrix and the estimate $\hat{\theta}_t$ is compared with the proposed value $\theta_0 = 0$. Using a sample size of n trading dates of contracts with $t \leq \tau_1^i < \tau_2^i$, i = $1, \ldots, I$, we illustrate in Table 5 the Wald statistics for all previous MPR specifications. We reject $H_0 : \hat{\theta}_t = 0$ under the Wald statistic $\left\{\theta_t \in \mathbb{R}^i\right\}_{t=1}^n$ for all cases. Although the constant per trading day and general MPR specifications smooth deviations over time, the Wald statistic confirms that the MPR differs significantly from 0. Our results are robust to all specifications.

Figure 10 shows the smoothing of MPR individuals (Equation (27)) for different specifications in 1 (20060530), 5 (20060522–20060530) and 30 trading days (20060417–20060530) of Berlin-CAT futures, while the last panel in Figure 10 gives the results

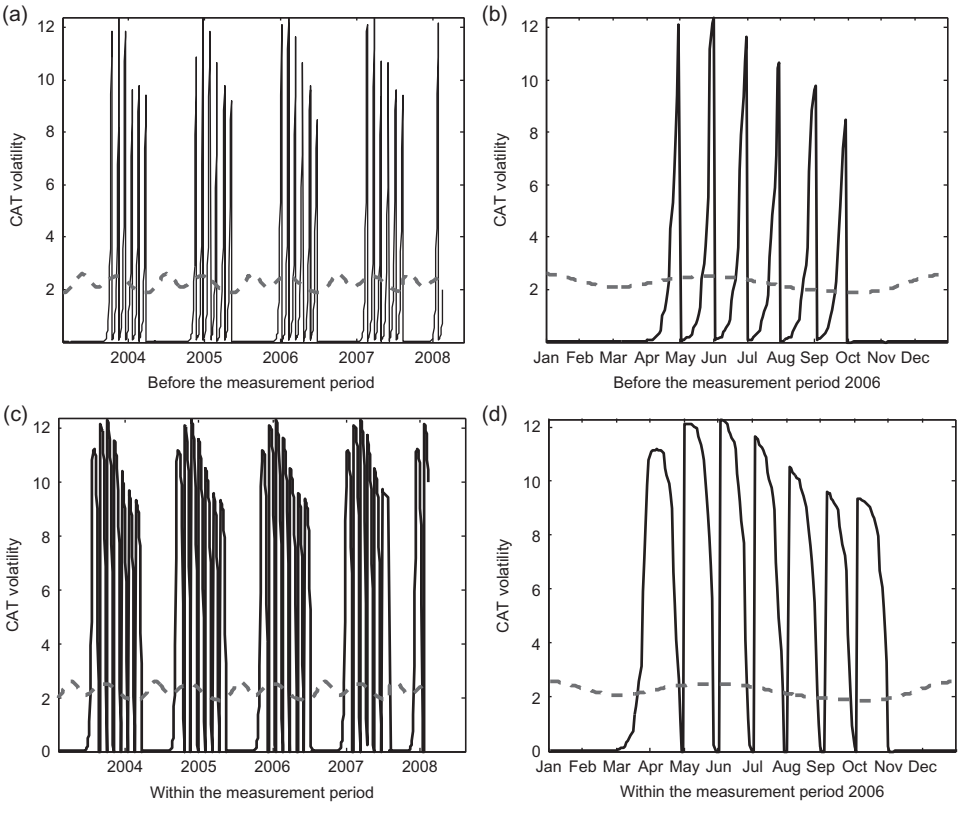


Figure 8. The Samuelson effect for Berlin-CAT futures explained by the CAT volatility $\sigma_t \mathbf{a}_{t,\tau_1,\tau_2} \mathbf{e}_p$ (black line) and the volatility σ_t of Berlin-CAT futures (dash line) from 2004 to 2008 and 2006 for contracts traded before (a) and (b) and within (c) and (d) the measurement period.

when MPR estimates are obtained from smoothed prices using the calendar year estimation (Equation (29)). Both smoothing procedures lead to similar outcomes: notable changes in sign, MPR deviations are smoothed over time and the higher the number of calendar days, the closer the fit of Equations (27) and (29). This indicates that sample size does not influence the stochastic behaviour of the MPR.

To interpret the economic meaning of the previous MPR results, recall, for example, the relationship between the RP (the market price minus the implied futures price with MPR equal to 0) and the MPR for CAT temperature futures:

$$RP_{CAT} = \int_t^{\tau_1^i} \theta_u \sigma_u \mathbf{a}_{t,\tau_1^i,\tau_2^i} \mathbf{e}_p du + \int_{\tau_1^i}^{\tau_2^i} \theta_u \sigma_u \mathbf{e}_1^\top \mathbf{A}^{-1} \left[\exp\left\{ \mathbf{A}(\tau_2^i - u) \right\} - I_p \right] \mathbf{e}_p du, \qquad (30)$$

which can be interpreted as the aggregated MPR times the amount of temperature risk σ_t over $[t, \tau_1^i]$ (first integral) and $[\tau_1^i, \tau_2^i]$ (second integral). By adjusting the MPR value, these two terms contribute to the CAT futures price. For temperature futures with values that are positive related to weather changes in the short term, this implies a

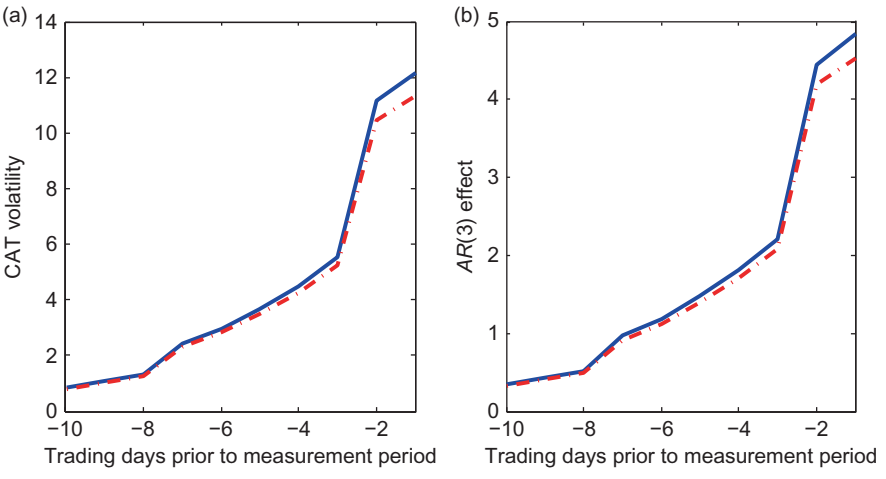


Figure 9. (a) The CAT term structure of volatility and (b) the autoregressive effect of two contracts issued on 20060517: one with whole June as measurement period (straight line) and the other one with only the 1st week of June (dotted line).

negative RP meaning that buyers of temperature derivatives expect to pay lower prices to hedge weather risk (insurance RP). In this case, θ_t must be negative for CAT futures, since σ_t and $\boldsymbol{X}_t$ are both positive. Negative MPRs translate into premiums for bearing risk, implying that investor will accept a reduction in the return of the derivative equal to the right-hand side of Equation (30) in exchange for eliminating the effects of the seasonal variance on pay-offs. On the other side, positive RP indicates the existence of consumers, who consider temperature derivatives for speculation purposes. In this case, θ_t must be positive and implies discounts for taking additional (weather) risk. This rules out the 'burn-in' analysis of Brix *et al*. (2005), which seems to popular among practitioners since it uses the historical average index value as the price for the futures. The sign of MPR–RP reflects the risk attitude and time horizon perspectives of market participants in the diversification process to hedge weather risk in peak seasons. By understanding the MPR, market participants might earn money (by shorting or longing, according to the sign). The investors impute value to the weather products, although they are non-marketable. This might suggest some possible relationships between risk aversion and the MPR.

The non-stationarity behaviour of the MPR (sign changes) is also possible because it is capturing all the non-fundamental information affecting the futures pricing: investors preferences, transaction costs, market illiquidity or other fractions like effects on the demand function. When the trading is illiquid the observed prices may contain some liquidity premium, which can contaminate the estimation of the MPR.

Figure 11 illustrates the RP of Berlin-CAT futures for monthly contracts traded on 20031006–20080527. We observe RPs different from 0, time dependent, where positive (negative) MPR contributes positively (negatively) to futures prices. The mean for the constant MPR for the $i = 1, \ldots, 7$th Berlin-CAT futures contracts per trading date is of size 0.02, 0.05, 0.02, 0.01, 0.10, 0.02 and 0.04, thus the terms in Equation (30)

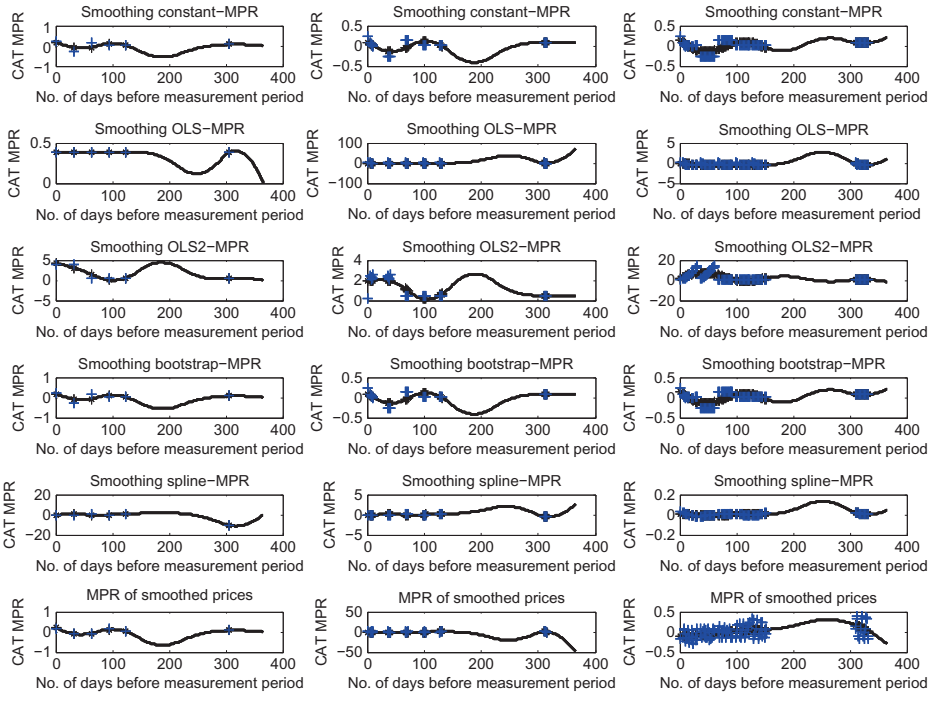


Figure 10. Smoothing the MPR parameterization for Berlin-CAT futures traded on 20060530: the calendar year smoothing (black line) for 1 day (left), 5 days (middle) and 30 days (right). The last row gives MPR estimates obtained from smoothed prices.

contribute little to the prices compared to the seasonal mean Λ_t. The RPs are very small for all contract types, and they behave constant within the measurement month but fluctuate with σ_t and θ_t, leading to higher RPs during volatile months (winters or early summers). This suggests that the temperature market does the risk adjustment according to the seasonal effect, where low levels of mean reversion mean that volatility plays a greater role in determining the prices.

Our data extracted MPR results can be comparable with Cao and Wei (2004), Richards *et al*. (2004) and Huang-Hsi *et al*. (2008), who showed that the MPR is not only different from 0 for temperature derivatives, but also significant and economically large as well. However, the results in Cao and Wei (2004) and Richards *et al*. (2004) rely on the specification of the dividend process and the risk aversion level, while the approach of Huang-Hsi *et al*. (2008) depends on the studied Stock index to compute the proxy estimate of the MPR. Alaton *et al*. (2002) concluded that the MPR impact is likely to be small. Our findings can also be compared with the MPR of other non-tradable assets, for example, in commodities markets; the MPR may be either positive or negative depending on the time horizon considered. In Schwartz (1997), the calibration of futures prices of oil and copper delivered negative MPR in both cases. For electricity, Cartea and Figueroa (2005) estimated a negative MPR. Cartea and Williams (2008) found a positive MPR for gas long-term contracts and for short-term

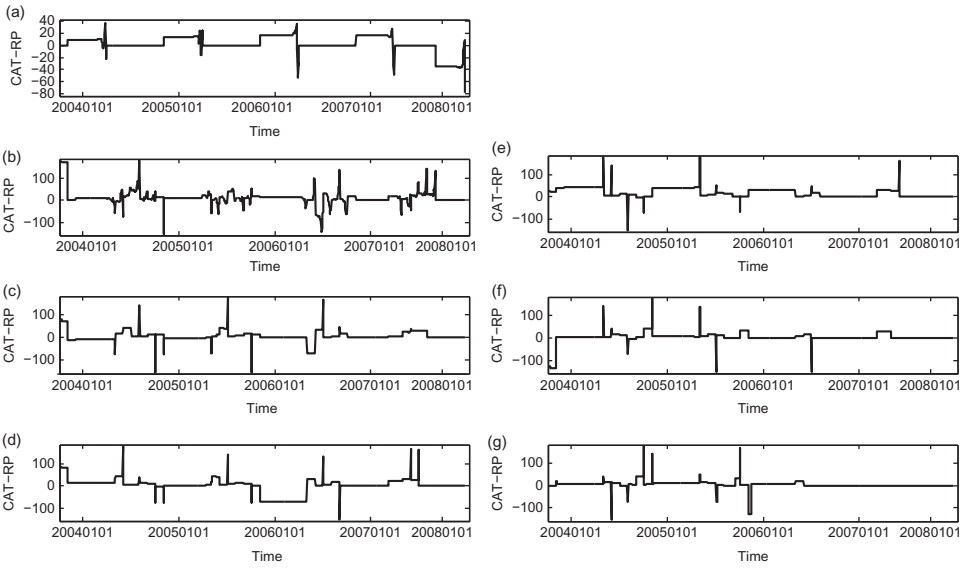


Figure 11. Risk premiums (RPs) of Berlin-CAT monthly futures prices traded during (20031006–20080527) with $t \leq \tau_1^i < \tau_2^i$ and contracts $i = 1$ (30 days), $i = 2$ (60 days),. . ., $i = I$ (210 days) traded before measurement period. RPs of Berlin CAT futures for (a) 30 days, (b) 60 days, (c) 90 days, (d) 120 days, (e) 150 days, (f) 180 days and (g) 210 days.

contracts the MPR changes signs across time. Doran and Ronn (2008) demonstrated the need of a negative market price of volatility risk in both equity and commodity-energy markets (gas, heating oil and crude oil). Similar to weather, electricity, natural gas and heating oil markets show seasonal patterns, where winter months have higher RP. The only difference is that in temperature markets, the spot–futures relation is not clear since the underlying is not storable (Benth *et al*., 2008).

5.8 Pricing CAT–HDD–CDD and OTC Futures

Once that market prices of traded derivatives are used to back out the MPR for temperature futures, the MPR for options is also known and thus one can price other temperature contract types with different maturity (weekly, daily or seasonal contracts) and over the counter OTC derivatives (e.g. Berlin-CDD futures or for cities without formal WD market). This method seems to be popular among practitioners in other markets.

This section tests the MPR specifications to fit market prices in sample. The implied MPR (under multiple specifications) from monthly CAT futures in Section 5.7 are used to calculate theoretical CDD prices Equation (20) for Berlin, Essen and Tokyo. We then compute HDD futures prices from the HDD–CDD parity in Equation (4) and compare them with market data (in sample performance). Table 4 shows the CME futures prices (Column 5), the estimated risk-neutral prices with $P = Q$ (MPR $= 0$), the estimated futures prices with constant MPR for different contracts per trading date and the index values computed from the realized temperature data $I_{(\tau_1,\tau_2)}$. While

the inferred prices with constant MPR replicate market prices, the estimated prices with $P = Q$ are close to the realized temperatures, meaning that the history is likely a good prediction of the future. Table 6 describes the root mean squared errors (RMSEs) of the differences between the market prices and the estimated futures prices, with MPR values implied directly from specific futures contract types and with MPR values extracted from the HDD/CDD/CAT parity method, over different periods and cities. The RMSE is defined as

$$\text{RMSE} = \sqrt{n^{-1}\sum_{t=1}^{n}(F_{t,\tau_1^i,\tau_2^i} - \hat{F}_{t,\tau_1^i,\tau_2^i})^2},$$

where $\hat{F}_{t,\tau_1^i,\tau_2^i}$ are the estimated futures prices and small RMSE values denote good measure of precision. The RMSE estimates in the case of the constant MPR for different CAT futures contracts are statistically significant enough to know CAT futures prices, but fail for HDD futures. Since temperature futures are written on different indices, the implied MPR will be then contract-specific hence requiring a separate estimation procedure. We argue that this inequality in prices results from additional premiums that the market incorporates to the HDD estimation, due to possible temperature market probability predictions operating under a more general equilibrium rather than non-arbitrage conditions (Horst and Mueller, 2007) or due to the incorporation of weather forecast models in the pricing model that influence the risk attitude of market participants in the diversification process of hedging weather risk (Benth and Meyer-Brandis, 2009; Dorfleitner and Wimmer, 2010; Papazian and Skiadopoulos, 2010).

We investigate the pricing algorithm for cities without formal WD market. In this context, the stylized facts of temperature data (Λ_t, σ_t) are the only risk factors. Hence, a natural way to infer the MPR for emerging regions is by knowing the MPR dependency on seasonal variation of the closest geographical location with formal WD market. For example, for pricing Taipei weather futures derivatives, one could take the WD market in Tokyo and learn the dependence structure by simply regressing the average MPR of Tokyo-C24AT futures contracts i over the trading period against the seasonal variation in period $[\tau_1, \tau_2]$:

$$\hat{\theta}^i_{\tau_1,\tau_2} = \frac{1}{\tau_1 - t}\sum_{t}^{\tau_1}\hat{\theta}^i_t,$$

$$\hat{\sigma}^2_{\tau_1,\tau_2} = \frac{1}{\tau_2 - \tau_1}\sum_{t=\tau_1}^{\tau_2}\hat{\sigma}^2_t.$$

In this case, the quadratic function that parameterizes the dependence is $\theta_t = 4.08 - 2.19\hat{\sigma}^2_{\tau_1,\tau_2} + 0.28\hat{\sigma}^4_{\tau_1,\tau_2}$, with $R^2_{\text{adj}} = 0.71$ and MPR increases by increasing the drift and volatility values (Figure 12). The dependencies of the MPR on time and temperature seasonal variation indicate that for regions with homogeneous weather risk there is some common market price of weather risk (as we expect in equilibrium).

Table 6. RMSE of the differences between observed CAT/HDD/CDD.

	Measurement period			RMSE between estimated with MPR (θ_t) and CME prices					
Contract type	τ_1	τ_2	No. of contracts	MPR = 0	Constant	1 piecewise	2 piecewise	Bootstrap	Spline
Atlanta-CDD$^+$	20070401	20070430	230	15.12	20.12	150.54	150.54	20.15	27.34
Atlanta-CDD$^+$	20070501	20070531	228	20.56	53.51	107.86	107.86	53.52	28.56
Atlanta-CDD$^+$	20070601	20070630	230	18.52	43.58	97.86	97.86	44.54	35.56
Atlanta-CDD$^+$	20070701	20070731	229	11.56	39.58	77.78	77.78	39.59	38.56
Atlanta-CDD$^+$	20070801	20070831	229	21.56	33.58	47.86	47.86	33.59	38.56
Atlanta-CDD$^+$	20070901	20070930	230	17.56	53.58	77.86	77.86	53.54	18.56
Berlin-HDD*	20061101	20061130	22	129.94	164.52	199.59	199.59	180.00	169.76
Berlin-HDD*	20061201	20061231	43	147.89	138.45	169.11	169.11	140.00	167.49
Berlin-HDD$^+$	20061101	20061130	22	39.98	74.73	89.59	89.59	74.74	79.86
Berlin-HDD$^+$	20061201	20061231	43	57.89	58.45	99.11	99.11	58.45	88.49
Berlin-CAT$^+$	20070401	20070430	230	18.47	40.26	134.83	134.83	40.26	18.44
Berlin-CAT$^+$	20070501	20070531	38	40.38	47.03	107.342	107.34	47.03	40.38
Berlin-CAT$^+$	20070601	20070630	58	10.02	26.19	78.18	78.18	26.20	10.02
Berlin-CAT$^+$	20070701	20070731	79	26.55	16.41	100.22	100.22	16.41	26.55
Berlin-CAT$^+$	20070801	20070831	101	34.31	12.22	99.59	99.59	12.22	34.31
Berlin-CAT$^+$	20070901	20070930	122	32.48	17.96	70.45	70.45	17.96	32.48
Essen-CAT$^+$	20070401	20070430	230	13.88	33.94	195.98	195.98	33.94	13.87
Essen-CAT$^+$	20070501	20070531	39	52.66	52.95	198.18	198.188	52.95	52.66
Essen-CAT$^+$	20070601	20070630	59	15.86	21.35	189.45	189.45	21.38	15.86
Essen-CAT$^+$	20070701	20070731	80	16.71	44.14	155.82	155.82	44.14	16.71
Essen-CAT$^+$	20070801	20070831	102	31.84	22.66	56.93	56.92	22.66	31.84
Essen-CAT$^+$	20070901	20070930	123	36.93	14.28	111.58	111.58	14.28	33.93
Tokyo-C24AT$^+$	20090301	20090331	57	161.81	148.21	218.99	218.99	148.21	158.16
Tokyo-C24AT$^+$	20090401	20090430	116	112.65	99.55	156.15	156.15	99.55	109.78
Tokyo-C24AT$^+$	20090501	20090531	141	81.64	70.81	111.21	111.21	70.81	79.68
Tokyo-C24AT$^+$	20090601	20090630	141	113.12	92.66	104.75	110.68	92.66	111.20
Tokyo-C24AT$^+$	20090701	20090731	141	78.65	74.95	116.34	3658.39	74.95	77.07

Notes: RMSE, root mean squared error; MPR, market price of risk; CME, Chicago Mercantile Exchange.
Futures prices with $t \leq \tau_1^i < \tau_2^i$ and the estimated futures with implied MPR under different MPR parameterizations (MPR = 0, constant MPR for different contracts (Constant), 1 piecewise constant MPR, 2 piecewise constant MPR, bootstrap MPR and spline MPR).
$^+$Computations with MPR implied directly from specific futures contract types ($^+$) and *through the parity HDD/CDD/CAT parity method(*).

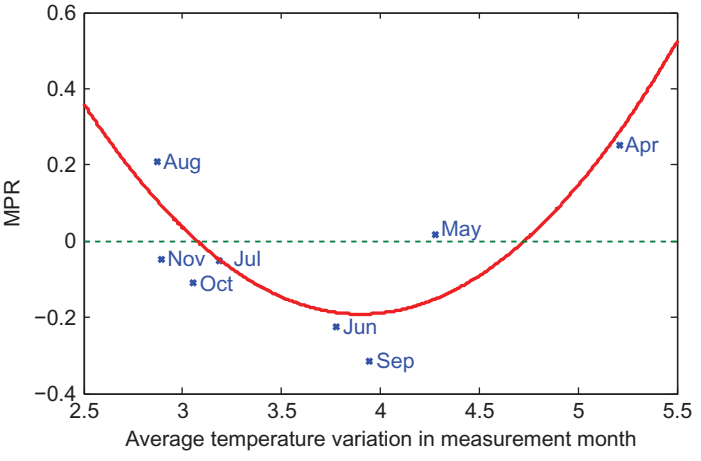


Figure 12. The calibrated MPR as a deterministic function of the monthly temperature variation of Tokyo-C24AT futures from November 2008 to November 2009 (prices for 8 contracts were available).

6. Conclusions and Further Research

This article deals with the differences between 'historical' and 'risk-neutral' behaviours of temperature and gives insights into the MPR, a drift adjustment in the dynamics of the temperature process to reflect how investors are compensated for bearing risk when holding the derivative. Our empirical work shows that independently of the chosen location, the temperature-driving stochastics are close to the Gaussian risk factors that allow us to work under the financial mathematical context.

Using statistical modelling, we imply the MPR from daily temperature futures-type contracts (CAT, CDD, HDD, C24AT) traded at the CME under the EMM framework. Different specifications of the MPR are investigated. It can be parameterized, given its dependencies on time and seasonal variation. We also establish connections between the RP and the MPR. The results show that the MPRs–RPs are significantly different from 0, changing over time. This contradicts with the assumption made earlier in the literature that MPR is 0 or constant and rules out the 'burn-in' analysis, which is popular among practitioners. This brings significant challenges to the statistical branch of the pricing literature, suggesting that for regions with homogeneous weather risk there is a common market price of weather risk. In particular, using a relationship of the MPR with a utility function, one may link the sign changes of the MPR with risk attitude and time horizon perspectives of market participants in the diversification process to hedge weather risk.

A further research on the explicit relationship between the RP and the MPR should be carried out to explain possible connections between modelled futures prices and their deviations from the futures market. An important issue for our results is that the econometric part in Section 2 is carried out with estimates rather than true values. One thus deals with noisy observations, which are likely to alter the subsequent estimations and test procedures. An alternative to this is to use an adaptive local parametric estimation procedure, for example, in Mercurio and Spokoiny (2004) or Härdle *et al*. (2011).

Finally, a different methodology, but related to this article, would be to imply the pricing kernel of option prices.

Acknowledgement

We thank Fred Espen Benth and two anonymous referees for several constructive and insightful suggestions on how to improve the article.

References

Alaton, P., Djehiche, B. and Stillberger, D. (2002) On modelling and pricing weather derivatives, *Applied Mathematical Finance*, 9(1), pp. 1–20.

Barrieu, P. and El Karoui, N. (2002) Optimal design of weather derivatives, *ALGO Research*, 5(1), pp. 79–92.

Benth, F. (2003) On arbitrage-free pricing of weather derivatives based on fractional Brownian motion, *Applied Mathematical Finance*, 10(4), pp. 303–324.

Benth, F., Cartea, A. and Kiesel, R. (2008) Pricing forward contracts in power markets by the certainty equivalence principle: explaining the sign of the market risk premium, *Journal of Finance and Banking*, 32(10), pp. 2006–2021.

Benth, F., Härdle, W. K. and López Cabrera, B. (2011) Pricing Asian temperature risk. In: P. Cizek, W. Härdle and R. Weron (Eds.), *Statistical Tools for Finance and Insurance*, pp. 163–199 (Heidelberg: Springer-Verlag).

Benth, F. and Meyer-Brandis, T. (2009) The information premium for non-storable commodities, *Journal of Energy Markets*, 2(3), pp. 111–140.

Benth, F. and Saltyte-Benth, J. (2005) Stochastic modelling of temperature variations with a view towards weather derivatives, *Applied Mathematical Finance*, 12(1), pp. 53–85.

Benth, F. and Saltyte-Benth, J. (2007) The volatility of temperature and pricing of weather derivatives, *Quantitative Finance*, 7(5), pp. 553–561.

Benth, F., Saltyte-Benth, J. and Jalinska, P. (2007) A spatial-temporal model for temperature with seasonal variance, *Applied Statistics*, 34(7), pp. 823–841.

Benth, F., Saltyte-Benth, J. and Koekebakker, S. (2007) Putting a price on temperature, *Scandinavian Journal of Statistics*, 34, pp. 746–767.

Benth, F., Saltyte-Benth, J. and Koekebakker, S. (2008) *Stochastic Modelling of Electricity and Related Markets*, Advanced Series on Statistical Science and Applied Probability, 2nd ed. (Singapore: World Scientific).

Brix, A., Jewson, S. and Ziehmann, C. (2005) *Weather Derivative Valuation: The Meteorological, Statistical, Financial and Mathematical Foundations* (Cambridge: Cambridge University Press).

Brockett, P., Golden, L. L., Wen, M. and Yang, C. (2010) Pricing weather derivatives using the indifference pricing approach, *North American Actuarial Journal*, 13(3), pp. 303–315.

Brody, D., Syroka, J. and Zervos, M. (2002) Dynamical pricing of weather derivatives, *Quantitative Finance*, 2(3), pp. 189–198.

Campbell, S. and Diebold, F. (2005) Weather forecasting for weather derivatives, *Journal of American Statistical Association*, 100(469), pp. 6–16.

Cao, M. and Wei, J. (2004) Weather derivatives valuation and market price of weather risk, *The Journal of Future Markets*, 24(11), pp. 1065–1089.

Cartea, A. and Figueroa, M. (2005) Pricing in electricity markets: a mean reverting jump diffusion model with seasonality, *Applied Mathematical Finance*, 12(4), pp. 313–335.

Cartea, A. and Williams, T. (2008) UK gas markets: the market price of risk and applications to multiple interruptible supply contracts, *Energy Economics*, 30(3), pp. 829–846.

Constantinides, G. (1987) Market risk adjustment in project valuation, *The Journal of Finance*, 33(2), pp. 603–616.

Cox, J. C., Ingersoll, J. and Ross, S. (1985) A theory of the term structure of interest rates, *Econometrica*, 59, pp. 385–407.

Doran, J. S. and Ronn, E. (2008) Computing the market price of volatility risk in the energy commodity markets, *Journal of Banking and Finance*, 32, pp. 2541–2552.

Dorfleitner, G. and Wimmer, M. (2010) The pricing of temperature futures at the Chicago Mercantile Exchange, *Journal of Banking and Finance*, 34(6), pp. 1360–1370.
Dornier, F. and Querel, M. (2007) Caution to the wind: energy power risk management, *Weather Risk Special Report*, August, pp. 30–32.
Fengler, M., Härdle, W. and Mammen, E. (2007) A dynamic semiparametric factor model for implied volatility string dynamics, *Financial Econometrics*, 5(2), pp. 189–218.
Gibson, R. and Schwartz, E. (1990) Stochastic convenience yield and the pricing of oil contingent claims, *Journal of Finance*, 45, pp. 959–976.
Härdle, W. K., López Cabrera, B., Okhrin, O. and Wang, W. (2011) Localizing temperature risk, Working Paper, Humboldt-Universitât zu Berlin.
Horst, U. and Mueller, M. (2007) On the spanning property of risk bonds priced by equilibrium, *Mathematics of Operation Research*, 32(4), pp. 784–807.
Huang-Hsi, H., Yung-Ming, S. and Pei-Syun, L. (2008) HDD and CDD option pricing with market price of weather risk for Taiwan, *The Journal of Future Markets*, 28(8), pp. 790–814.
Hull, J. and White, A. (1990) Valuing derivative securities using the explicit finite difference method, *Journal of Financial and Quantitative Analysis*, 28, pp. 87–100.
Karatzas, I. and Shreve, S. (2001) *Methods of Mathematical Finance* (New York: Springer-Verlag).
Landskroner, Y. (1977) Intertemporal determination of the market price of risk, *The Journal of Finance*, 32(5), pp. 1671–1681.
Lucas, R. E. (1978) Asset prices in an exchange economy, *Econometrica*, 46(6), pp. 1429–1445.
Mercurio, D. and Spokoiny, V. (2004) Statistical inference for time-inhomogeneous volatility models, *The Annals of Statistics*, 32(2), pp. 577–602.
Mraoua, M. and Bari, D. (2007) Temperature stochastic modelling and weather derivatives pricing: empirical study with Moroccan data, *Afrika Statistika*, 2(1), pp. 22–43.
Papazian, G. and Skiadopoulos, G. (2010) Modeling the dynamics of temperature with a view to weather derivatives, Working Paper, University of Piraeus.
Richards, T., Manfredo, M. and Sanders, D. (2004) Pricing weather derivatives, *American Journal of Agricultural Economics*, 86(4), pp. 1005–1017.
Schwartz, E. (1997) The stochastic behaviour of commodity prices: implications for valuation and hedging, *Journal of Finance*, LII(3), pp. 923–973.
Vasicek, O. (1977) An equilibrium characterization of the term structure, *Journal of Financial Economics*, 5, pp. 177–188.
Zapranis, A. and Alexandridis, A. (2008) Modeling the temperature time-dependent speed of mean reversion in the context of weather derivative pricing, *Applied Mathematical Finance*, 15, pp. 355–386.
Zapranis, A. and Alexandridis, A. (2009) Weather derivatives pricing: modeling the seasonal residual variance of an Ornstein-Uhlenbeck temperature process with neural networks, *Neurocomputing*, 73(1–3), pp. 37–48.

Comput Stat
DOI 10.1007/s00180-012-0312-6

ORIGINAL PAPER

Using wiki to build an e-learning system in statistics in the Arabic language

Taleb Ahmad · Wolfgang Härdle · Sigbert Klinke · Shafiqah Alawadhi

Received: 1 June 2007 / Accepted: 4 February 2012

Abstract E-learning plays an important role in education as it supports online teaching via computer networks and provides educational services by utilising information technologies. This paper presents a case study describing the development of an Arabic language e-learning course in statistics. Under discussion are problems concerning e-learning in Arab countries with special focus on the difficulties of the application of e-learning in the Arabic world as well as designing an Arabic platform with its language and technical challenges. For the platform we have chosen a wiki that supports LaTeX for formulas and R to generate tables and figures as well as some interactivity. Our system, Arabic MM*Stat, can be found at http://mars.wiwi.hu-berlin.de/mediawiki/mmstat_ar.

Keywords E-learning · MM*Stat · Wiki

T. Ahmad (✉)
Tishreen University, Latakia, Syria
e-mail: taleb_ahmad1976@yahoo.de

W. Härdle
Center for Applied Statistics and Economics, Humboldt-Universität zu Berlin,
Unter den Linden 6, 10099 Berlin, Germany
e-mail: haerdle@wiwi.hu-berlin.de

S. Klinke
Ladislaus von Bortkiewicz Chair of Statistics, Humboldt-Universität zu Berlin,
Unter den Linden 6, 10099 Berlin, Germany
e-mail: sigbert@wiwi.hu-berlin.de

S. Alawadhi
Department of Statistics and Operations Research, Faculty of Science,
Kuwait University, P. O. Box 5969, 13060 Safat, Kuwait
e-mail: alawadi@kuc01.kuniv.edu.kw

Published online: 21 February 2012

1 Introduction

Due to the proliferation of the Internet, e-learning has become a significant aspect of education and many universities and educational institutions have created their own web sites and e-learning systems. Future trends predict that e-learning will significantly complement classic learning. Statistics show that the size of the worldwide e-learning market is estimated to be 52.6 *billion* US dollars yearly, with the ratio at 65–75% for the United States and Europe. Statistics also indicate that 30% of the education was delivered electronically. In comparison the e-learning market in Arab countries with a size around 15 *million* US dollars yearly is very weak. The gap between Europe and the United States and the Arab countries is very large.

The reasons for this gap is briefly summarised below:

- According to the latest figures available on Internet World Statistics 2010 (de Argaez 2011), Internet world usage still varies widely across the world and across languages as shown in Table 1. The diffusion of Internet services in the most Arab 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 3.3% of Internet users come from the Arabic region, even though the Arabic population is 5% of world population. Another example for this gap is that the percentage of web users in the Arabic world is 18.8% compared with 58.4% in Europe, 77.4% in the USA and 28.7% on average in the whole world. Arabic users have much less experience with e-learning platforms, telecourses and educational courses.
- English is the most common language in the e-learning platforms, but most Arabic users have difficulties in understanding and speaking English.
- General educational problems: A high level of illiteracy can be found in the Arabic world which varies between 25 and 45% (Clayton 2007; Al-Fadhli 2008).
- There is only a limited number of specialised cadres and scientific expertise in the area of e-learning in Arab countries (Maegaard et al. 2005).

Due to the above mentioned problems Arab countries need more time to acquire the advantage of e-learning. The dissemination of the culture of e-learning in schools and universities needs a new generation of qualified professionals who can deal successfully with modern technology and the experiences of e-learning.

In fact, our Internet research showed that only a few Arabic e-learning platforms exists, especially for statistics we could not find a single one. For this reason we find the creation of a platform that would aid Arabic students in learning statistics highly necessary. The platform should cover the basic statistical topics, and is supported by multiple examples and ease-of-use will be adapted for Arabic students.

From the perspective described above we developed an Arabic e-learning platform in statistics (Arabic MM*Stat), which might become an important reference point in the study of statistics in Arabic through the Internet.

Around about 2000 a system known as MM*Stat was developed at the School for Business and Economics of Humboldt-Universität zu Berlin (Müller et al. 2000).

Table 1 World internet users for 10 languages by June 2010 (de Argaez 2011)

Top 10 languages in the internet	Internet users (Mio.)	Internet penetration, (%)	Growth in internet (2000–2010), (%)	Internet users % of total, (%)	World population (2010 estimate)
English	537	42.0	281	27.3	1,278
Chinese	445	32.6	1,277	22.6	1,278
Spanish	153	36.5	742	7.8	420
Japanese	99	78.2	111	5.0	127
Portuguese	83	33.0	990	4.2	250
German	75	78.6	173	3.8	96
Arabic	65	18.8	2,501	3.3	347
French	60	17.2	389	3.0	348
Russian	60	42.8	1,826	3.0	139
Korean	39	55.2	107	2.0	71
Top 10 languages	1,615	36.4	421	82.2	4,442
Other languages	351	14.6	588	17.8	2,403
World total	1,966	28.7	444	100.0	6,846

MM*Stat is a platform for e-learning statistics and is an HTML based multimedia environment to support teaching and learning statistics via CD or Internet.

A MM*Stat course consists of lectures of specific topics in basic statistics, see Fig. 1 for the *hypergeometric distribution*. Each lecture gives the basic concepts of general statistical theory, definitions, formulae and mathematical proofs. At the bottom is a set of buttons, on the left-hand side three buttons for navigation (go to the previous lecture, jump to the table of contents, go to the next lecture) and on the right-hand side a number of buttons which link to pages with additional information. Four types of additional information are provided, these are:

Explained examples which require only knowledge of the current lecture to understand them.

Enhanced examples which require knowledge from different lectures than the current one to understand them.

Interactive examples which allow the user, via an embedded statistical software, to run them. For example, to plot the probability density function or the cumulative distribution function for different parameters of n and p) or apply tests.

More information which contain for example historical information or mathematical derivations which are not necessary for first-hand understanding.

Each chapter with lectures is finished with a lecture containing multiple-choice questions such that a user can evaluate his/her learning progress.

Students or anyone interested in statistics can interactively learn about the basic concepts of statistics at anytime and anywhere and consequently we based Arabic

Fig. 1 The graphical user interface of MM*Stat, as an example the lecture entitled *hypergeometric distribution*. Note the navigation button (*bottom left*) and buttons to examples and more information (*bottom right*). The tabs at the top reflect the user history and allow for a fast change between lectures

MM*Stat on the existing MM*Stat, which already existed in various languages: Czech, German, English, Spanish, French, Indonesian, Italian, Polish and Portuguese.

2 Difficulties to design Arabic platforms

There are some problems, however, associated with the making of an Arabic platform, these relate to language as well as technology. We summarise these problems below:

Language problems

There are some items related to translation, some words and scientific terms are similar in Arabic and could create a problem when translated. For example, see Table 2. The Arabic language makes no distinction between “administration” and “management” or “calculate” and “compute”. The reader must recognise from the context which meaning is correct. This makes a text more difficult to understand.

Technical problems

1. *User interface*
   The different language versions of MM*Stat were based on two different systems:

Table 2 Some similar words in Arabic

Arabic	English
إدارة	Administration
إدارة	Management
حسب	Calculate
حسب	Compute

- The German version was written in HTML and the user interface was developed with JavaScript for Internet Explorer 5. The problem was that neither later Internet Explorer versions nor browsers other than the Internet Explorer were able to run the JavaScript code.
- The English version was written in LaTeX for a variety of reasons, for example, translating MM*Stat into a new language just required a change to the LaTeX text which is much easier to handle than translating from a HTML page with a lot of embedded JavaScript codes. We used a software based on LaTeX2HTML to create the HTML/JavaScript version of MM*Stat with the same user interface as before (Witzel and Klinke 2002).

2. *Writing from left to right*
   Arabic script runs from right to left as opposed to most other languages, therefore all lists, paragraphs, statistical forms, tables and graphics also 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) such as numbers and Latin texts. Any program that supports the Arabic language should provide the possibility of changing the direction when needed. A solution would be to use ArabTeX (Lagally 2004), but with ArabTeX the Arabic texts are written in English with special character combinations and not in Arabic, see Fig. 2. Obviously this is unfamiliar to most Arabic speaking people. Additionally LaTeX2HTML supports neither text from right to left, Arabic or ArabTeX.
3. *Interactive examples*
   MM*Stat contained a set of interactive examples, which are important since they allow the user to practice repeatedly with various variables or data sets, and with alternate sample sizes or parameters of the statistical methods applied. In this manner, the student obtains a better understanding of how the statistical method works. However, the client-server technology implemented by Lehmann (2004) for MM*Stat worked only with the statistical software XploRe. The development and support of the XploRe software has unfortunately ceased, so the question arises how one should include the interactive examples in Arabic MM*Stat such that they will be runnable in future.

The language problem can only be solved by adapting the texts. To solve the technical problems we decided to use another technology, the so called “wiki technology”.

```
\documentstyle[12pt,arabtex,atrans,nashbf]{article}
\begin{document}
...
\begin{arabtext}
i^starY ^gu.hA ‘a^saraTa .hamIriN.
fari.ha bihA wa-sAqahA ’amAmahu,
_tumma rakiba wA.hidaN minhA.
wa-fI al-.t.tarIqi ‘adda .hamIrahu wa-huwa rAkibuN,
fa-wa^gadahA tis‘aTaN.
_tumma nazala wa-‘addahA fa-ra’AhA ‘a^saraTuN fa-qAla:

’am^sI wa-’aksibu .himAraN,
’af.dalu min ’an ’arkaba wa-’a_hsara .himAraN.
\end{arabtext}
...
\end{document}
```

Fig. 2 Sample Arabtex input in LaTeX, see `examples/guha.tex` in Arabtex

3 Wiki technology

3.1 What is a wiki?

Wiki is a system that allows users to collaborate in forming the content of a web site. The first wiki web site, “WikiWikiWeb”, was designed by Cunningham and Leuf in 1995 (Leuf and Cunnigham 2001). 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 a wiki are:

- Wiki simplifies the process of content editing. Each web page contains a link to change content within the web browser. After saving a modified page it can be viewed immediately.
- It uses simple markup to coordinate content, 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 a record of the page history 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.
- Wiki sites can be publicly open and therefore allow any user to improve the content.
- Wiki simplifies the organisation of a site: Wiki sites create hypertext databases and can regulate the content in any manner desired; many content management systems require the planning of the organization of the content before anything is written. This allows for flexibility which is not available in 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 specialised web sites. For example,

Fig. 3 Entry page of Arabeyes wiki (http://wiki.arabeyes.org)

a teacher could write his course using a wiki and offer it to his students as useful material for study, see e.g., Klinke (2011).

Nowadays, we have many more examples of web sites using wikis as a tool for the development of content, like Wikipedia (2011). The Wikipedia project started 15 January 2001 and today there are more than 10 million articles in the encyclopedia in all languages, more than 3.7 million articles in the English encyclopedia alone. Millions of volunteers around the world modify and add to the contents daily and new articles are created. The Arabic version of the free encyclopedia was launched in July 2003 and currently contains approximately 160 thousand articles as the Arabic encyclopedia is in the content-building phase.

The Arabic wiki platform "Arabeyes" (Afifi et al. 2011) provides a good environment for discussion and exchange of experience and knowledge about the Arabic language. Arabeyes offers the translation into Arabic for free open-source programs. In addition Arabeyes provides a technical dictionary that aims to translate and standardise the technical terms used in translating the software to the Arabic user. Arabeyes is a solution for the *language problem*, see Fig. 3.

3.3 Implementation of Arabic MM*Stat

Arabic MM*Stat is directed at students and Arabic users that serve the e-learning issues in the Arabic region. The content of Arabic MM*Stat is a translation of the content of the former CD's into Arabic.

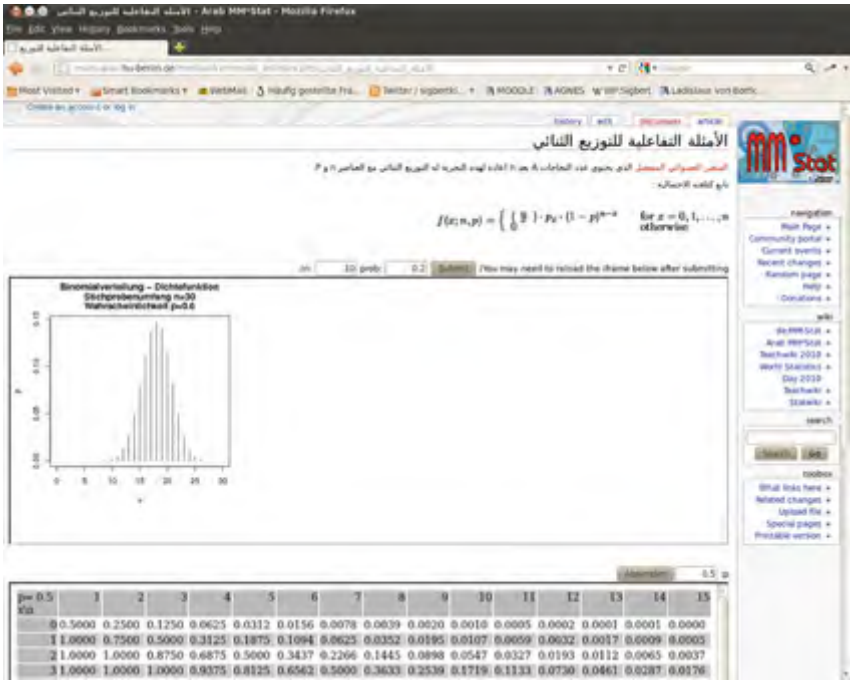
Fig. 4 Graphical user interface (GUI) of Arabic MM*Stat. Note that the interface language (English) is different from the content language (Arabic) due to user settings

Wikimatrix (2011) offers an overview of the available wiki software and their capabilities. A useful wiki should support:

- LaTeX to provide the possibility to write a statistical formula in "mathematical" language rather than integrate it as a graphic, generated for example by LaTeX2HTML.
- Arabic as a language for the content and the interface.
- Integration of statistical software, preferably R, to recreate interactive examples.
- Multiple choice questions to test students knowledge.

As wiki software we finally decided to use the Mediawiki, the software behind the (Arabic) Wikipedia. It solves all possible technical problems (see Figs. 4 and 5):

- *User interface*
  It is able to have the content and the user interface in the Arabic language as the Arabic Wikipedia shows.
- *Writing from left to right*
  To some extent, it can change the writing direction for formulas, list etc.
- *Interactive examples*
  Through Mediawiki extensions we are able to transfer the functionality of the MM*Stat CD to the new system:
  - The R extension allows to embed (interactive) tables and graphics generated by R into wiki page as well as interactive examples.



p= 0.5 / k=	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
0	0.5000	0.2500	0.1250	0.0625	0.0312	0.0156	0.0078	0.0039	0.0020	0.0010	0.0005	0.0002	0.0001	0.0001	0.0000
1	1.0000	0.7500	0.5000	0.3125	0.1875	0.1094	0.0625	0.0352	0.0195	0.0107	0.0059	0.0032	0.0017	0.0009	0.0005
2	1.0000	1.0000	0.8750	0.6875	0.5000	0.3437	0.2266	0.1445	0.0898	0.0547	0.0327	0.0193	0.0112	0.0065	0.0037
3	1.0000	1.0000	1.0000	0.9375	0.8125	0.6562	0.5000	0.3633	0.2539	0.1719	0.1133	0.0730	0.0461	0.0287	0.0176

Fig. 5 An interactive example of a graphic of a probability density function and a table of a cdf of a binomial distribution ($p = 0.6$)

Fig. 6 The wiki source code for the page shown in Fig. 5. On top the Arabic text and within the `RForm` tags the input parameters and within the `R` tags the R program

- The Quiz extension provided multiple choice questions (Babé 2007).
- The Math extension allows formulas written in LaTeX to be embedded into the wiki page (Wegrzanowski and Vibber 2011).

3.4 Integration of R program into Arabic MM*Stat

R is a language and environment for statistical computing and graphics (R Development Core Team 2011). Arabic MM*Stat uses R programs to create tables and graphics which can be incorporated in courses notes. For the Mediawiki software an extension to embed R into the wiki page exists.

They enable the students and learners, for example to visualise statistics distributions and probability tables via the Internet. See Fig. 5 as an example of a graphic of a probability density function and a table of a cdf function of a binomial distribution. Choosing other input values will lead to different tables or graphics.

Figure 6 shows the wiki source code for the example shown in Fig. 5. The interactive example consists of two tags `Rform` und `R` which share a common attribute `name`.

```
<Rform name=''binom''>
... Input parameters...
</Rform>

<R output=''display'' name=''binom''>
... R program...
</R>
```

Between the opening and closing `Rform` tags are the input parameters as defined in an HTML form. The following opening and closing `R` tag contain the R program which produces a graphic. For more detail see Klinke and Zlatkin-Troitschanskaia (2007).

There are in Arabic MM*Stat other examples, e.g., for other distributions like normal, Poisson and exponential distribution.

4 Conclusion

Using E-learning/e-teaching tools to offer 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. Some of the specific characteristics we have discussed earlier for developing an Arabic platform already exist in the wiki. We see that embedding of R is an solution for the interactive examples in Arabic MM*Stat. We hope that the Arabic MM*Stat platform for e-learning of statistics will be a significant help for the Arabic user as it clearly overcomes weaknesses in developing such electronic platforms in Arabic.

This research was supported by the Deutsche Forschungsgemeinschaft through the CRC 649 ‘Economic Risk’.

References

Afifi D, Chahibi Y, Hosny K, Trojette MA (2011) Arabeyes.org—The Arabic Unix Project. http://wiki.arabeyes.org, Accessed 18 Aug 2011

Al-Fadhli S (2008) Students' perceptions of e-learning in Arab society: Kuwait University as a case study. E-Learning 5(4):418–428

Babé LR (2007) Extension: Quiz. http://www.mediawiki.org/wiki/Extension:Quiz, Accessed 26 Aug 2011

CosmoCode (2011) Wikimatrix. http://www.wikimarix.org, Accessed 24 Aug 2011

Clayton PM (2007) Women, violence and the internet. E-Learning 4(1):79–92

de Argaez E (2011) Internet world stats, Columbia 2011. http://www.internetworldstats.com, Accessed 18 Aug 2011

Klinke S (2011) Developing web-based tools for the teaching of statistics: Our Wikis and the German Wikipedia, ISI Conference Proceedings, Dublin/Ireland (forthcoming)

Klinke S, Zlatkin-Troitschanskaia O (2007) Embedding R in the Mediawiki, Sonderforschungsbereich 649 Discussion Papers 61, Humboldt University, Berlin/Germany 2007. http://edoc.hu-berlin.de/docviews/abstract.php?id=28421, Accessed 18 Aug 2011

Lagally K (2004) ArabTeX : a system for typesetting Arabic and Hebrew. Report Nr. 2004/03, User Manual Version 4.00, Fakultät Informatik, University Stuttgart, Germany. http://www2.informatik.uni-stuttgart.de/ivi/bs/research/arab_e.htm, Accessed 18 Aug 2011

Lehmann H (2004) Client/server based statistical computing. Dissertation, Humboldt-Universität zu Berlin, Germany. http://edoc.hu-berlin.de/docviews/abstract.php?id=20803, Accessed 18 Aug 2011

Leuf B, Cunningham W (2001) The wiki way: quick collaboration on the web. Addison-Wesley, Reading, MA

Maegaard B, Choukri K, Mokbel C, Yaseen M (2005) Language technology for Arabic. Report of The NEMLAR project, Center for Sprogteknolo gi. University of Copenhagen, Denmark. http://www.medar.info/The_Nemlar_Project/Publications/Arabic_LT.pdf, Accessed 18 Aug 2011

Müller M, Rönz B, Ziegenhagen U (2006) The Multimedia Project MM*Stat for Teaching Statistics. In: Proceedings of the COMPSTAT 2000. pp 409–414

R Development Core Team (2011) R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna/Austria. http://www.r-project.org, Accessed 18 Aug 2011

Wegrzanowski T, Vibber B et al. (2011) Extension: Math. http://www.mediawiki.org/wiki/Extension:Math, Accessed 26 Aug 2011

Wikimedia Foundation Inc. (2011) Wikipedia: The free encyclopedia. http://www.wikipedia.org, Accessed 18 Aug 2011

Witzel R, Klinke S (2002) MD*Book online & e-stat: Generating e-stat Modules from Latex. In: Proceedings of the COMPSTAT 2002. pp 449–454

Journal of Financial Econometrics, 2013, Vol. 11, No. 2, 370–399

Shape Invariant Modeling of Pricing Kernels and Risk Aversion

Maria Grith
CASE–Center for Applied Statistics and Economics, Humboldt-Universität zu Berlin

Wolfgang Härdle
CASE–Center for Applied Statistics and Economics, Humboldt-Universität zu Berlin

Juhyun Park
Department of Mathematics and Statistics, Lancaster University

ABSTRACT
Several empirical studies reported that pricing kernels exhibit a common pattern across different markets. The main interest in pricing kernels lies in validating the presence of the peaks and their variability in location among curves. Motivated by this observation we investigate the problem of estimating pricing kernels based on the shape invariant model, a semi-parametric approach used for multiple curves with shape-related nonlinear variation. This approach allows us to capture the common features contained in the shape of the functions and at the same time characterize the nonlinear variability with a few interpretable parameters. These parameters provide an informative summary of the curves and can be used to make a further analysis with macroeconomic variables. Implied risk aversion function and utility function can also be derived. The method is demonstrated with the European options and returns values of the German stock index DAX. (*JEL*: C14, C32, G12)

KEYWORDS: pricing kernels, risk aversion, risk neutral density

The financial support from the Deutsche Forschungsgemeinschaft via SFB 649 "Ökonomisches Risiko", Humboldt-Universität zu Berlin is gratefully acknowledged. Address correspondence to Juhyun Park, Department of Mathematics and Statistics, Lancaster University, Lancaster LA1 4YF, UK, or email: juhyun.park@lancaster.ac.uk

doi:10.1093/jjfinec/nbs019 Advanced Access publication November 18, 2012

1 METHODOLOGY

1.1 Pricing Kernel and Risk Aversion

Risk analysis and management drew much attention in quantitative finance recently. Understanding the basic principles of financial economics is a challenging task in particular in a dynamic context. With the formulation of utility maximization theory, individuals' preferences are explained through the shape of the underlying utility functions. Namely, a concave, convex, or linear utility function is associated with risk averse, risk seeking, or risk neutral behavior, respectively. The comparison is often made through the Arrow-Pratt measure of absolute risk aversion (ARA), as a summary of aggregate investor's risk-averseness. The quantity is originated from the expected utility theory and is defined by

$$ARA(u) = -\frac{U''(u)}{U'(u)},$$

where U is the individual utility as a function of wealth.

With an economic consideration that one unit gain and loss does not carry the same value for every individual, understanding state-dependent risk behavior becomes an increasingly important issue. The fundamental problem is that individual agents are not directly observable but it is assumed that the prices of goods traded in the market reflect the dynamics of their risk behavior. Several efforts have been made to relate the price processes of assets and options traded in a market to risk behavior of investors, since options are securities guarding against losses in risky assets.

A standard option pricing model in a complete market assumes a *risk neutral* distribution of returns, which gives the fair price under no arbitrage assumptions. If markets are not complete, there are more risk neutral distributions and the fair price depends on the hedging problem. The *subjective or historical* distribution of observed returns reflects a risk-adaptive behavior of investors based on subjective assessment of the future market. Then the equilibrium price is the arbitrage free price and the transition from risk neutral pricing to subjective rule is achieved through the pricing kernel. Assuming those densities exist, write q for the risk neutral density and p for the historical density. The pricing kernel $\mathcal{K}$ is defined by the ratio of those densities:

$$\mathcal{K}(u) = \frac{q(u)}{p(u)}.$$

Through the intermediation of these densities, there exists a link between the pricing kernel and ARA, see for example Leland (1980)

$$ARA(u) = \frac{p'(u)}{p(u)} - \frac{q'(u)}{q(u)} = -\frac{d\log\mathcal{K}(u)}{du}.$$

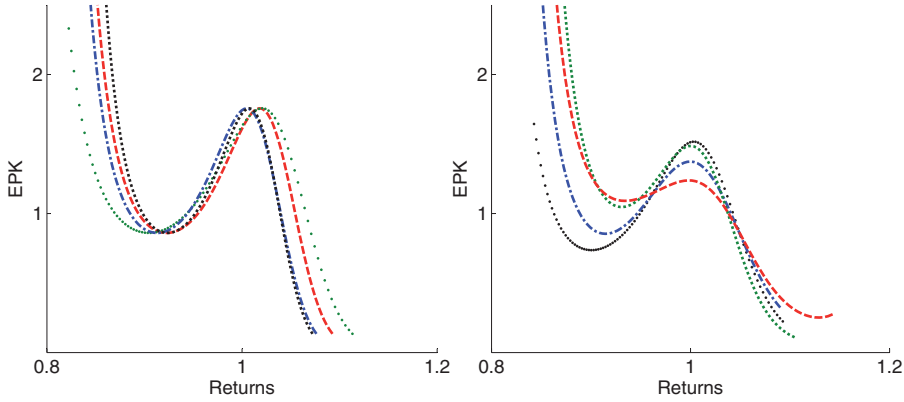


Figure 1 Examples of inter-temporal pricing kernels for various maturities in January–February 2006 (left) and monthly pricing kernels from the first six months in 2006 for maturity one month (right).

In this way, rather than specifying a priori preferences of agents (risk neutral, averse, or risk seeking) and implicitly the monotonicity of the pricing kernel, we can infer the risk patterns from the shape of the pricing kernel.

1.2 Dynamics of Empirical Pricing Kernels (EPKs)

With increasing availability of large market data, several approaches to recovering pricing kernels from empirical data have been proposed. As many of them estimate p and q separately to recover $\mathcal{K}$, potentially relevant are studies focusing on recovering risk neutral density, see e.g. Jackwerth (1999), and Bondarenko (2003) for comparison of different approaches. For the estimation of p nonparametric kernel methods or parametric models such as GARCH or Heston models are popular choices.

Examples of empirical pricing kernels obtained from European options data on the German stock index DAX (Deutscher Aktien index) in 2006 are shown in Figure 1, based on separate estimation of p and q. A detailed account of estimation is given in Section 3.4. To make these comparable, they are shown on a continuously compounded returns scale. Throughout the article, the pricing kernel is considered as a function of this common scale of returns. Figure 1 depicts inter-temporal pricing kernels with various maturities in January–February 2006 (left), and monthly pricing kernels with fixed maturity one month in 2006 (right). The sample of curves appears to have a bump around 1 and has convexity followed by concavity in all cases. The location as well as the magnitude of the bump vary among curves, which reflects individual variability on different dates or under different investment horizons. Some features that are of particular economic interest include

the maximum of the bump, the spread or duration of the bump and the location of the bump.

From a statistical perspective, properties of the pricing kernel are intrinsically related to assumptions about the data generation process. A very restrictive model, with normal marginal distributions, is the Black–Scholes model. This results in an overall decreasing pricing kernel in wealth, which is consistent with overall risk-averse behavior. These preferences are often assumed in the classical economic theory of utility-maximizing agent and correspond to a concave indirect von Neumann and Morgenstern utility function. However, under richer parametric specifications or nonparametric models monotonicity of the pricing kernel has been rejected in practice (Rosenberg and Engle, 2002; Giacomini and Härdle, 2008). The phenomenon of locally nondecreasing pricing kernel is referred to as *the pricing kernel puzzle* in the literature. There have been many attempts to reconcile the underlying economic theory with the empirical findings. A recent solution suggested by Hens and Reichlin (2012) relates the puzzle to the violation of the fundamental assumptions in the equilibrium model framework.

Most of earlier works adopt a static viewpoint, showing a snapshot of markets on selected dates but report that there is a common pattern across different markets. The first dynamic viewpoint appears in Jackwerth (2000), who recovers a series of pricing kernels in consecutive times and claims that these do not correspond to the basic assumptions of asset pricing theory. In a similar framework Giacomini and Härdle (2008) perform a factor analysis based on the so-called dynamic semiparametric factor models, while Giacomini, Härdle and Handel (2008) introduce time series analysis of daily summary measures of pricing kernels to examine variability between curves.

Chabi-Yo, Garcia, and Renault (2008) explain the observed dynamics or the puzzles by means of latent variables in the asset pricing models. Effectively, they propose to build conditional models of the pricing kernels given the state variables reflecting preferences, economic fundamentals, or beliefs. Within this framework they are able to reproduce the puzzles, in conjunction with some joint parametric specifications for the pricing kernel and the asset return processes.

Due to evolution of markets over time under different circumstances, the pricing kernels are intrinsically time varying. Thus, approaches that do not take into account the changing market make limited use of information available in the current data. On the other hand, changes over time may not be completely arbitrary, as there are common rules and underlying laws that assure some consistency across different market system. Moreover, variability observed in pricing kernels, as shown in Figures 1, is not necessarily linear, and thus factors constructed from a linear combination of observations are only meaningful for explaining aggregated effects.

Considering the pricing kernels as an object of curves, we approach the problem of estimating the pricing kernels and implied risk aversion functions from a functional data analysis viewpoint (Ramsay and Silverman, 2002). The main interest in pricing kernels lies in validating the presence of the peaks and their

variability in location among curves. Motivated by this observation we investigate the estimation method based on the *shape invariant model*, which will be formally introduced in Section 2. This is chosen over the commonly adopted functional principal component analysis to accommodate the nonlinear features such as variation of peak locations, which encapsulate quantities amenable to economic interpretation. The shape invariant model allows us to capture the common characteristics, reported across different studies on different markets. We then explain individual variability as a deviation from the common curve or a reference.

Our contribution is three-fold. Firstly, we analyze the phenomenon of pricing kernel puzzle from a dynamic viewpoint using shape invariant modeling approach. The starting question was how to compare the empirical evidence. By taking into account variability among curves, we quantify a *trend of the puzzle* in the series of the pricing kernels by a few interpretable parameters. Secondly, we provide a unified framework for estimation and interpretation of ARA and utility functions consistent with the underlying pricing kernels with the same set of parameters. The ARA corresponding to the reference pricing kernel could be viewed as a typical pattern of risk behavior for the period under consideration. Due to nonlinear transformation involved in deriving ARA from the pricing kernel function, this common ARA function does not necessarily coincide with the simple average ARA functions. Thirdly, the output of the analysis provides a summary measure to study the relationship with macroeconomic variables. Through real data example we have related the changes in risk behavior to some macroeconomic variables of interest and found that local risk loving behavior is procyclical. We acknowledge that we do not provide an economic explanation to the puzzle but rather try to understand the nature of the phenomenon by means of statistical analysis.

The paper is organized as follows. Section 2 motivates common shape modeling approach and Section 3 reviews the shape invariant model and describes it in detail in the context of pricing kernel estimation. This section serves the basis of our analysis. Numerical studies based on simulation are found in Section 4. An application to real data example is summarized in Section 5.

2 COMMON SHAPE MODELING

2.1 Shape Invariant Model for Pricing Kernel

We consider a common shape modeling approach for the series of pricing kernels with explicit components of location and scale. To represent varying pricing kernels, we introduce the time index t in the pricing kernel as $\mathcal{K}_t$ and consider a general regression model:

$$Y_t = \mathcal{K}_t + \varepsilon_t,$$

where ε_t represents an error with mean 0 and variance σ_t^2. We begin with a working assumption of independent error as in Kneip and Engel (1995). The effect of

dependent error is investigated in simulation studies in Section 4.2. The relationship among K_ts is specified as

$$\mathcal{K}_t(u) = \theta_{t1} g\left(\frac{u-\theta_{t3}}{\theta_{t2}}\right) + \theta_{t4}, \tag{1}$$

with some unknown constants $\boldsymbol{\theta}_t = (\theta_{t1}, \theta_{t2}, \theta_{t3}, \theta_{t4})$ and an unknown function g. The common shape function g can be interpreted as a reference curve. Deviation from the reference curve is described by four parameters $\boldsymbol{\theta}_t = (\theta_{t1}, \theta_{t2}, \theta_{t3}, \theta_{t4})$ that capture scale changes and a shift in horizontal and vertical direction. This parametrization in (1) is commonly known as a shape invariant model (SIM), originally introduced by Lawton, Sylvestre and Maggio (1972) and studied by Kneip and Engel (1995). Note that the model includes as a special case complete parametric models with known g.

In contrast to standard applications of SIM as a regression model, the SIM application to pricing kernel estimation does not, strictly speaking, satisfy the model assumption. There is no realization of the pricing kernels available and thus our formulation of regression model should be viewed as an approximation. The original data used would be intraday options data and daily returns data, which are collected from separate sources with sample sizes of different orders of magnitude but estimation of p and q can be effectively done independently of each other. It may be possible to elaborate our approach to incorporate simultaneous estimation with a two-step state-dependent dynamic model formulation whereby the dynamics of the observed return processes are specified and the unobserved pricing kernel processes enter as a state variable. However, with current advancement in the methodology, this is only possible with limited parametric model choices, see for example Chabi-Yo *et al*. (2008), and extension to a flexible shape invariant model is left for future work.

Instead we exploit the fact that preliminary estimates of pricing kernels based on separate estimation of p and q are readily available from market data and this can easily substitute Y. From now on, we treat the estimates as something observable and denote by Y_t, similar to the regression formulation with direct measurements Y_t and state the asymptotic result without further complication of pre-processing steps. After all, these estimates of curves are available from the beginning and the SIM aims to characterize a structural relationship among these curves. This however may impact the parametric rate of convergence attainable (Kneip and Engel, 1995) because our observations are already contaminated by a nonparametric error of estimation. As is shown in Section 3.6, the dominating error comes from the estimation of q, which involves second derivative estimation. The optimal rate of convergence for estimating second derivative is known to be $\mathcal{O}(N^{-2/9})$, where N is the sample size used (Stone, 1982). This implies that $\sigma_t^2 = \alpha_{N,t} v^2$ where $\alpha_{N,t}$ is a constant of order $\mathcal{O}(N^{-2/9})$, which should be understood as the multiplication factor for the parametric rate of convergence.

A particular choice of estimates of individual pricing kernels is not part of the model formulation but affects the starting values for the estimation of shape

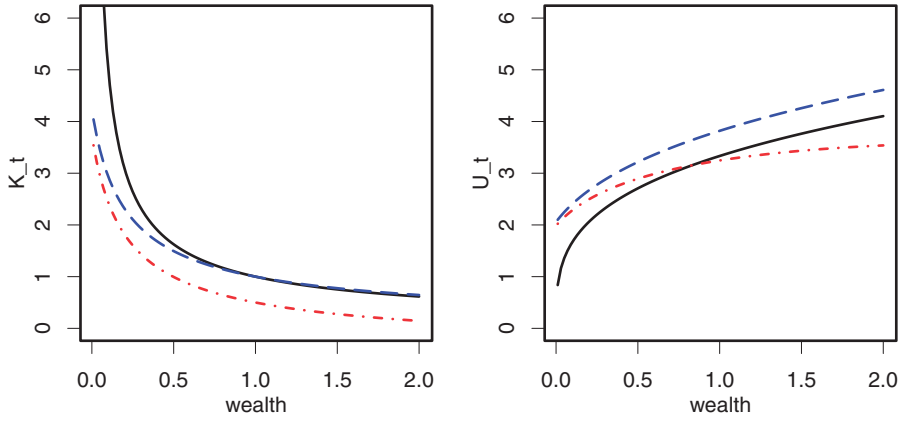


Figure 2 Example of location and scale shift family of pricing kernels (left) and corresponding utility functions (right). Solid line in each plot represents reference curves of $g(u)=u^{-\gamma}$ and $U_0(u)=u^{1-\gamma}/(1-\gamma)$ with $\gamma=0.7$, respectively. Parameters are $\theta_{t1}=1.1, \theta_{t2}=1, \theta_{t3}=1-\theta_{t1}^{(1/\gamma)}$, and $\theta_{t4}=0$ for dot-dashed (red) and $\theta_{t4}=-0.5$ for dashed (blue) lines.

invariant model. Our choice of initial estimates will be explained in Section 3.4. Our main interest lies in quantifying the variation among the pricing kernels given those estimates.

The new message here is an analysis of a sequence of pricing kernels through shape invariant models. Although we start with different motivation, our approach is in line with that of Chabi-Yo *et al*. (2008). In contrast to their approach, we impose a structural constraint that is related to the shape of the function. This way we strike a balance between flexibility much desired in parametric model specification and interpretability of the results lacking in full nonparametric models.

2.2 SIM and Black–Scholes Model

To appreciate the model formulation, given in the Equation (1), it is instructive to consider utility functions implied by this family of pricing kernels together. The utility function can be derived from

$$U_t(u)=\alpha\int_0^u \mathcal{K}_t(x)dx,$$

for a constant α. Figure 2 shows an example based on a power utility function, which corresponds to risk averse behavior. Pricing kernels $\mathcal{K}_t$ are shown on the left and the corresponding utility functions U_t are on the right. The solid lines represent reference curves and the dashed and dot-dashed lines represent $\mathcal{K}_t$ and U_t with appropriate parameters $\boldsymbol{\theta}_t$ in the Equation (1). Depending on the choice of

parameters, the utility function can increase quickly or slowly. As an illustration, we consider the Black–Scholes model with power utility function. The Black–Scholes model assumes that the stock price follows a geometric Brownian motion

$$dS_t/S_t = \mu dt + \sigma dW_t,$$

which gives rise to a log normal distribution for the historical density p. Under the risk neutral measure, the drift μ is replaced by the riskless rate r but the density q is still log normal. The pricing kernel can be written as a power function

$$\mathcal{K}(u) = \lambda u^{-\gamma}, 0 < \gamma < 1,$$

with appropriate constants λ and γ. The corresponding utility function is a power utility

$$U(u) = \lambda \frac{u^{1-\gamma}}{1-\gamma}.$$

Assume that $\lambda = 1$ and suppose that g is a power function, say $u^{-\gamma}$. Then the class of pricing kernels implied by (1) is given by

$$\begin{aligned} \mathcal{K}_t(u) &= \theta_{t1}\left(\frac{u-\theta_{t3}}{\theta_{t2}}\right)^{-\gamma} + \theta_{t4} \\ &= \theta_{t1}^*(u-\theta_{t3})^{-\gamma} + \theta_{t4}, \end{aligned}$$

where $\theta_{t1}^* = \theta_{t1}\theta_{t2}^{\gamma}$. Notice that with this family of functions θ_{t1} and θ_{t2} are not identifiable and $\mathcal{K}_t$ is defined for $u > \theta_{t3}$. For the sake of argument we set $\theta_{t2} = 1$ for the moment. The corresponding utility function is

$$\begin{aligned} U_t(u) &= \int_{\theta_{t3}}^{u} \mathcal{K}_t(x)dx \\ &= \frac{\theta_{t1}}{1-\gamma}(u-\theta_{t3})^{(1-\gamma)} + \theta_{t4}(u-\theta_{t3}) \\ &\stackrel{\text{def}}{=} \theta_{t1}^{**}(u-\theta_{t3})^{(1-\gamma)} + \theta_{t4}(u-\theta_{t3}). \end{aligned}$$

When $\theta_{t4} = 0$, this produces again a transformed power utility. When $\theta_{t4} \neq 0$, there is additional linear term in the function. See Figure 2 for comparison.

2.3 Identifiability Condition for SIM

The previous section illustrates two aspects of applicability of the shape invariant models. The class of functions that can be generated by the relation (1) is rich, but in order to uniquely identify the model parameters, some restriction is necessary.

For example, we have seen that the two scale parameters in the pricing kernel functions corresponding to the Black–Scholes model are not separable. Basically, unless there exist some qualitatively distinct common characteristics for each curve, the model is not identifiable (Kneip and Gasser, 1988). In the case of no prior structural information available as in the case of pricing kernels, it is sufficient to consider a few landmarks such as peaks and inflection points.

Even with a unique g, some translation and scaling of parameters lead to multiple representations of the models. For uniqueness of parameters, we will impose normalizing conditions suggested in Kneip and Engel (1995):

$$T^{-1}\sum_{t=1}^{T}\theta_{t1}=1,\quad T^{-1}\sum_{t=1}^{T}\theta_{t2}=1,\quad T^{-1}\sum_{t=1}^{T}\theta_{t3}=0,\quad T^{-1}\sum_{t=1}^{T}\theta_{t4}=0$$

in the sense that there exists an *average curve*. These conditions are not restriction at all and can be replaced by any appropriate combination of parameters. Alternatively, we could consider the first curve as a reference, as done in Härdle and Marron (1990), which implies the restriction $\boldsymbol{\theta}_1=(1,1,0,0)$. Generally, an application-driven normalization scheme can be devised and examples are found in Lawton, Sylvestre and Maggio (1972).

2.4 SIM Implied Risk Aversion and Utility Function

The utility function corresponding to $\mathcal{K}_t$ is given by

$$\begin{aligned}U_t(u) &= \theta_{t1}\theta_{t2}\left\{G\left(\frac{u-\theta_{t3}}{\theta_{t2}}\right)-G\left(-\frac{\theta_{t3}}{\theta_{t2}}\right)\right\}+\theta_{t4}u\\ &\equiv \theta^*_{t1}G\left(\frac{u-\theta_{t3}}{\theta_{t2}}\right)+\theta^*_{t4}+\theta_{t4}u,\end{aligned}$$

where $G(t)=\int_0^t g(u)du$. The utility function U_t is a combination of a SIM class of the common utility function and a linear utility function.

The ARA measure is given by

$$ARA_t(u)=\frac{-\frac{\theta_{t1}}{\theta_{t2}}g'\left(\frac{u-\theta_{t3}}{\theta_{t2}}\right)}{\theta_{t1}g\left(\frac{u-\theta_{t3}}{\theta_{t2}}\right)+\theta_{t4}}. \tag{2}$$

For example, assuming $g(u)=u^{-\gamma}$ with $\theta_{t2}=1$ gives

$$ARA_t(u)=\gamma\left\{(u-\theta_{t3})+(\theta_{t4}/\theta_{t1})(u-\theta_{t3})^{\gamma+1}\right\}^{-1}.$$

When $\theta_{t4}=0$, this function is monotonically decreasing but in general this is not the case. Note the common ARA function corresponding to g is γu^{-1} compared to the mean ARA function computed by taking the sample average $T^{-1}\sum_{t=1}^{T}ARA_t(u)$.

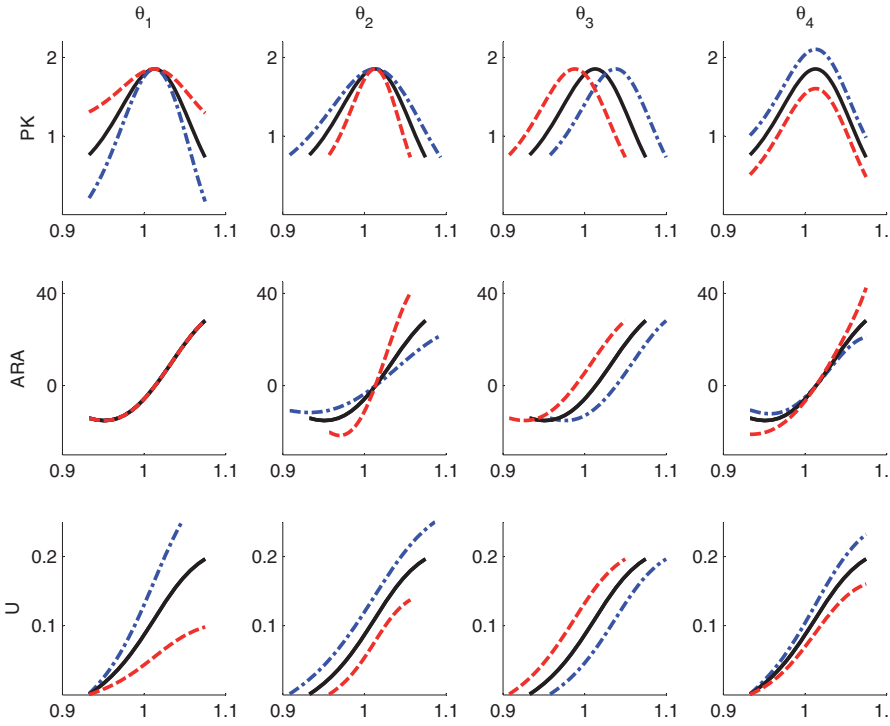


Figure 3 Effect of parameters on pricing kernel (top), ARA (middle), and utility function (bottom) compared to the baseline model $\theta_0 = (1,1,0,0)$ (black). Dot-dashed lines are used for increasing direction and dashed lines for decreasing direction.

In order to gain some insights, we take a closer look at the changes in relation to individual scale and shift parameters. These individual effects are demonstrated in Figure 3. We vary each θ_i with respect to a baseline model and then we show how these modifications translate into changes of the risk attitudes and the corresponding utility functions. The parameters used in Figure 3 are $\boldsymbol{\theta} = (0.5, 0.7, -0.025, -0.25)$ in dashed line and $\boldsymbol{\theta} = (1.5, 1.3, 0.025, 0.25)$ in dot-dashed line.

For this exercise we first standardize the common curve that we have estimated via the shape invariant model so that the peak occurs at the value 0 on the abscissa and the effect of the scale and shift parameters is separately captured. But we added the peak coordinates back for visualization so that they are comparable to other figures shown on returns scale. We observe that an increase in θ_1 marks the bump of the pricing kernel more distinctive while the shape of ARA remains unchanged compared to the baseline model because, as we can see from (2), ARA does not depend on θ_1 when $\theta_4 = 0$. Yet, the effect of θ_1 on ARA can be analyzed by considering two distinct cases: $\theta_4 > 0$ and $\theta_4 < 0$. These specifications are important

because the direction of change in the slope of ARA is dictated by the sign of θ_4. In the present case—after normalization—θ_1 varies around 0 and its effect on ARA is almost nil.

A larger value in the parameter θ_2 as compared to a benchmark value stretches the x-axis, which implies larger spread of the bump. When we vary θ_2 alone the slope of ARA($\theta_2 u$) is $1/\theta_2^2\left[\left\{g'^2(u)-g''(u)/g(u)\right\}/g^2(u)\right]$. The term in brackets does not depend on θ_2; it is equal to the slope of $ARA(u)$. Therefore, there is an inverse relationship between the direction of change in the parameter and that of the absolute value of the slope. These changes in slope occur around an inflection point that corresponds to the peak of the pricing kernel.

A positive increment in θ_3 shifts both curves to the left without any modification in the shape. θ_4 simply translates pricing kernel curves above or below the reference curve following a sign rule. Similarly to θ_2, the shape of ARA modifies around the fixed inflection point that marks the change from risk proclivity (negative ARA) to risk aversion (positive ARA). The effect of θ_4 on the values of ARA is straightforward: since θ_4 adds to the g in the denominator its increase will diminish the absolute ARA level and the other way around. Insulating the effects of a change in θ_4 on the slope of $ARA(u)$ analytically proves to be a more complicated task than in the case of θ_2 because the change in the slope depends jointly on the change in θ_4 and on the pricing kernel values and its first two derivatives. In our case, the slope around the inflection point increases when θ_4 decreases.

As for the utility function, positive changes in θ_1 and θ_4 increases its absolute slope. In the horizontal direction, θ_3 translates the curve to the left or right similarly to the pricing kernel and ARA while θ_2 shrinks or expands its domain.

With this information at hand we can characterize the changes in risk patterns in relation to economic variables of interest, see Section 5.4.

3 FITTING SHAPE INVARIANT MODELS

3.1 Estimation of SIM

The model in (1) is equivalently written as

$$\mathcal{K}_t(\theta_{t2}u+\theta_{t3})=\theta_{t1}g(u)+\theta_{t4}, \qquad \theta_{t1}>0, \quad \theta_{t2}>0. \tag{3}$$

The estimation procedure is developed using the least squares criterion based on nonparametric estimates of individual curves. If there are only two curves, parameter estimates are obtained by minimizing

$$\int \{\hat{\mathcal{K}}_2(\theta_2 u+\theta_3)-\theta_1\hat{\mathcal{K}}_1(u)-\theta_4\}^2 w(u)du, \tag{4}$$

where $\hat{\mathcal{K}}_i$ are nonparametric estimates of the curves. Härdle and Marron (1990) studied comparison of two curves and Kneip and Engel (1995) extended to multiple curves with an iterative algorithm. We consider an adaption of such algorithm here.

The weight function w is introduced to ensure that the functions are compared in a domain where the common features are defined. We assume that there is an interval $[a,b]\in J$ where boundary effects are eliminated and then define

$$w(u)=\prod_t 1_{[a,b]}\{(u-\theta_{t3})/\theta_{t2}\}.$$

The parameter estimates are compared only in the common region defined by w but the individual curve estimates are defined on the whole interval. Weights can be extended to account for additional variability.

The normalization leads to:

$$T^{-1}\sum_{t=1}^{T}\mathcal{K}_t(\theta_{t2}u+\theta_{t3})=g(u). \tag{5}$$

Formula (5) was exploited in the algorithm proposed by Kneip and Engel (1995). We adopt a similar strategy here.

- Initialize
  - Let $\hat{K}_t=Y_t$ and set starting values $\left(\theta_{t2}^{(0)},\theta_{t3}^{(0)}\right)$ for $t=1,2,\cdots,T$.
  - Construct an initial estimate $g^{(0)}$ by

$$g^{(0)}(u)=T^{-1}\sum_{t=1}^{T}\hat{\mathcal{K}}_t\left(\theta_{t2}^{(0)}u+\theta_{t3}^{(0)}\right).$$

- For r-th step, $r=1,2,\cdots,R$,
  - Determine parameters $\boldsymbol{\theta}^{(r)}$ separately for $t=1,2,\cdots,T$ by minimizing

$$\int\left\{\hat{\mathcal{K}}_t(\theta_{t2}u+\theta_{t3})-\theta_{t1}g^{(r-1)}(u)-\theta_{t4}\right\}^2 w(u)du.$$

  - Normalize parameters: for $j=(1,2)$ and $k=(3,4)$

$$\theta_{tj}^{(r)}\leftarrow\frac{\theta_{tj}^{(r)}}{\sum_t\theta_{tj}^{(r)}},\qquad \theta_{tk}^{(r)}\leftarrow\theta_{tk}^{(r)}-T^{-1}\sum_t\theta_{tk}^{(r)}.$$

  - Update $g^{(r-1)}$ to

$$g^{(r)}(u)=T^{-1}\sum_{t=1}^{T}\hat{\mathcal{K}}_t\left(\theta_{t2}^{(r)}u+\theta_{t3}^{(r)}\right).$$

- Determine final estimates:

$$\tilde{\boldsymbol{\theta}}_t = \boldsymbol{\theta}_t^{(R)},$$

$$\tilde{g}(u) = T^{-1}\sum_{t=1}^{T}\hat{\mathcal{K}}_t\left(\tilde{\theta}_{t2}u+\tilde{\theta}_{t3}\right).$$

Kneip and Engel (1995) proved consistency of the estimator. In particular despite nonparametric initial curve estimates, the parameters are shown to be $\sqrt{T}$ consistent. In their analysis it is noted that the initial estimates of the curves are of minor importance compared to the final estimate of g. So the original algorithm includes the final updating of each curve. This improves precision of the estimates because the pooled sample estimate reduces the variance of $\tilde{g}$, which allows undersmoothing at the final stage to reduce bias. However, this final updating step is not practical for our situation with indirect measurements and is not implemented here for pricing kernel estimation. On the other hand, we can take advantage of having smooth curves evaluated at finite grid points as data. It is easier to improve the initialization step, explained in Section 3.2. This leads to simplification of the estimating procedure with little compromise of the quality of the fit. In fact, the number of iterations required is very small and often 3 or 4 is sufficient in practical terms. We found that when the initial estimates are determined sufficiently accurate, the iteration is not necessary.

As a working model we have assumed an independent error. If there is a reasonable dependence structure available, this could be incorporated easily in the estimation algorithm with weighted least squares estimation in (4). The effect of independence assumption mainly appears in the standard error estimation and a correction can be made with a sandwich variance–covariance estimator. To assess the effect of model misspecification, we also carried out some simulation studies with dependent errors and reported the results in Section 4.

3.2 Starting Values

If there is no scale change in horizontal direction, due to prominent peaks in each curve, the parameter θ_3 can be identified easily by the location of the individual peak. If the models hold true, and there are two unique landmarks identifiable for each curve, simple linear regression between the individual mark and the average mark provides an estimate of the slope parameter θ_2. Suppose that the peak is identified by u satisfying $K_t'(u)=0$. Then we have

$$0=\mathcal{K}_t'(u)=\frac{\theta_{t1}}{\theta_{t2}}g'\left(\frac{u-\theta_{t3}}{\theta_{t2}}\right).$$

Writing u_t^* for $\mathcal{K}_t'$ and u_0^* for g' leads to a simple linear relation:

$$u_t^*=\theta_{t2}u_0^*+\theta_{t3}. \tag{6}$$

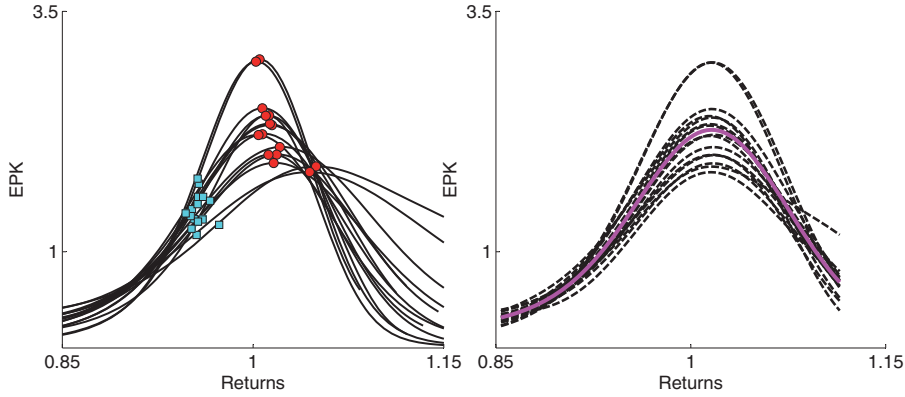


Figure 4 Initial estimates $\mathcal{K}_t(u)$ (left) and final estimates $\mathcal{K}_t(\theta_{t2}u+\theta_{t3})$ from SIM (right) with g overlayed. Marked in the left plot are two landmarks identified for estimation of the starting values of $(\theta_{t2},\theta_{t3})$.

If an inflection point is used, we would have

$$0=\mathcal{K}_t''(u)=\frac{\theta_{t1}}{\theta_{t2}^2}g''\left(\frac{u-\theta_{t3}}{\theta_{t2}}\right),$$

which gives rise to the same relation as (6), with the corresponding u_t^{**} and u_0^{**} substituted. The coefficients of intercept and slope estimates are used for starting values of θ_{t3} and θ_{t2}, respectively.

We used the peak and the inflection points around 1 as landmarks, marked in Figure 4. The location of the landmarks is defined by the zero crossings of the first and second derivatives. Because the initial observations $\mathcal{K}_t$ are a smoothed curve, we find that additional smoothing procedure is not required at this stage: a finite difference operation is sufficient to apply mean value theorem with linear interpolation.

The slope between any two points did not vary much, which is consistent with the model specification. This step is also used as an informal check and should there be any nonlinearity detected, the model needs to be extended to include a nonlinear transformation. With our example, this was not the case.

3.3 Nonlinear Optimization

Given the estimates of $(\theta_{t2},\theta_{t3})$, the nonlinear least squares optimization uses (4), which is approximated by

$$\sum_j\left\{\hat{\mathcal{K}}_t\left(\theta_{t2}u_j+\theta_{t3}\right)-\theta_{t1}\hat{g}\left(u_j\right)-\theta_{t4}\right\}^2 w\left(u_j\right). \tag{7}$$

When the initial values of $(\theta_{t2}, \theta_{t3})$ are sufficiently accurate, this step is simplified to a linear regression. Conditional on θ_{t2}, θ_{t3} and $\hat{g}$, the solutions to the least square regression with response variable $\hat{\mathcal{K}}_t(\theta_{t2}u_j + \theta_{t3})$ and explanatory variable $\hat{g}(u_j)$ provide $(\theta_{t1}, \theta_{t4})$. When a further optimization routine is employed to improve the estimates, these numbers serve as initial values for $(\theta_{t1}, \theta_{t4})$.

3.4 Initial Estimates of $\mathcal{K}$

To start the algorithm the initial estimates of $\mathcal{K}$ should be supplied. An example of initial estimates of $\mathcal{K}$ is shown in Figure 4 on the scale of continuously compounded returns. These are obtained from separate estimation of p and q, which are described below. Individual smoothing parameter choice is discussed in Section 5 with real data example.

3.4.1 Estimation of the historical density p. We use the nonparametric kernel density estimates similar to Ait-Sahalia and Lo (2000) based on the past observations of returns for a fixed maturity τ. With this approach the returns of the stock prices are assumed to vary slowly and thus the process can be assumed stationary for a short period of time. Alternatively, if additional modeling assumption is made for the evolution of the stock price such as GARCH, a simulation-based approach could be employed.

At given time t and $T = t + \tau$ we obtain realizations of future return values from a window of historical return values of length J:

$$r_T^k = \log\left(S_{t-(k-1)}/S_{t-\tau-k+1}\right) \quad \text{and} \quad S_T^k = S_t e^{r_T^k}, \quad k = 1, \ldots, J.$$

The probability density of r_T is obtained by the kernel density estimator

$$\hat{p}_{h_p}(r) = \frac{1}{Jh_p}\sum_{k=1}^{J} K\left(\frac{r_T^k - r}{h_p}\right),$$

where K is a kernel weight function and h_p is the bandwidth. Some variations are also explored such as overlapping and nonoverlapping windows with a real data example in Section 5.

3.5 Estimation of the Risk Neutral Density q

We begin with the call price option formula that links the call prices to the risk neutral density estimation. The European call price option formula is given

by (Ait-Sahalia and Duarte, 2003)

$$C(X,\tau,r_{t,\tau},\delta_{t,\tau},S_t)=e^{-r_{t,\tau}\tau}\int_0^\infty \max(S_T-X,0)q(S_T|\tau,r_{t,\tau},\delta_{t,\tau},S_t)\,dS_T$$

where

- S_t: the underlying asset price at time t,
- X: the strike price,
- τ: the time to maturity,
- $T=t+\tau$: the expiration date,
- $r_{t,\tau}$: the deterministic risk free interest rate for that maturity,
- $\delta_{t,\tau}$: the corresponding dividend yield of the asset.

Write $q(S_T)$ for $q(S_T|\tau,r_{t,\tau},\delta_{t,\tau},S_t)$. For fixed t and τ, assume $r_{t,\tau}=r$ and $\delta_{t,\tau}=\delta$, the risk neutral density is expressed as

$$q(u)=e^{r\tau}\frac{\partial^2 C}{\partial X^2}|_{X=u}.$$

The relation is due to Breeden and Litzenberger (1978) and serves the basis of many current semi-parametric and nonparametric approaches. We employ the semiparametric estimates of Rookley (1997), where the parametric Black–Scholes formula is assumed except that the volatility parameter σ is a function of the option's moneyness and the time to maturity τ. In this work, we fix the maturity and consider it as one dimensional regression problem.

Define $F=S_te^{(r-\delta)\tau}$ and $m=X/F$ is moneyness. Write Φ and ϕ for the cumulative distribution function and its density of standard normal random variable, respectively. The Black–Scholes model assumes

$$\begin{aligned}C_{BS}(X,\tau) &= S_te^{-\delta\tau}\Phi(d_1)-e^{-r\tau}X\Phi(d_2)\\ &= e^{-r\tau}F\{\Phi(d_1)-m\Phi(d_2)\}.\end{aligned}$$

In a semiparametric call price function, the volatility parameter σ is expressed as a function of the option's moneynes and the time to maturity τ:

$$C(X,\tau,r,\delta,S_t)=C_{BS}(X,\tau,F,\sigma(m,\tau)).$$

To derive the second derivative of C, it is simpler to work with a standardized call price function $c(m,\tau)=e^{r\tau}C(X,\tau,r,\delta,\sigma)/F=\Phi(d_1)-m\Phi(d_2)$. The derivatives of C and c are related as

$$\begin{aligned}\frac{\partial C}{\partial X} &= e^{-r\tau}F\frac{\partial c}{\partial m}\frac{\partial m}{\partial X}=e^{-r\tau}\frac{\partial c}{\partial m},\\ \frac{\partial^2 C}{\partial X^2} &= e^{-r\tau}\frac{\partial c^2}{\partial m^2}\frac{\partial m}{\partial X}=\frac{e^{-r\tau}}{F}\frac{\partial c^2}{\partial m^2}.\end{aligned}$$

With some manipulation we obtain the following expressions, which are only functions of $(\sigma, \sigma', \sigma'')$:

$$\frac{\partial c}{\partial m} = \phi(d_1)\frac{\partial d_1}{\partial m} - \Phi(d_2) - m\phi(d_2)\frac{\partial d_2}{\partial m}$$

$$\frac{\partial^2 c}{\partial m^2} = -d_1\phi(d_1)\left(\frac{\partial d_1}{\partial m}\right)^2 + \phi(d_1)\frac{\partial^2 d_1}{\partial m^2} - \phi(d_2)\frac{\partial d_2}{\partial m} - \phi(d_2)\frac{\partial d_2}{\partial m} + md_2\phi(d_2)\left(\frac{\partial d_2}{\partial m}\right)^2 - m\phi(d_2)\frac{\partial^2 d_2}{\partial m^2},$$

where

$$\frac{\partial d_1}{\partial m} = -\frac{1}{\sqrt{\tau}}\frac{1}{m\sigma(m,\tau)} + \frac{1}{\sqrt{\tau}}\ln(m)\frac{\sigma'(m,\tau)}{\sigma^2(m,\tau)} + \frac{\sqrt{\tau}}{2}\sigma'(m,\tau)$$

$$\frac{\partial d_2}{\partial m} = \frac{\partial d_1}{\partial m} - \sqrt{\tau}\sigma'(m,\tau)$$

$$\frac{\partial^2 d_1}{\partial m^2} = \frac{1}{m^2\sqrt{\tau}\sigma(m,\tau)} + \frac{2}{\sqrt{\tau}}\frac{\sigma'(m,\tau)}{\sigma^2(m,\tau)}\left\{\frac{1}{m} - \ln(m)\frac{\sigma'(m,\tau)}{\sigma(m,\tau)}\right\} + \sigma''(m,\tau)\left\{\frac{\ln(m)}{\sigma^2(m,\tau)\sqrt{\tau}} + \frac{\sqrt{\tau}}{2}\right\}$$

$$\frac{\partial^2 d_2}{\partial m^2} = \frac{\partial^2 d_1}{\partial m^2} - \sqrt{\tau}\sigma''(m,\tau).$$

Note that this leads to a slightly different derivation from Rookley (1997), albeit using the same principle.

In order to compute the derivatives of σ, we used the local polynomial smoothing on implied volatility. Let σ_i be the implied volatility corresponding to the call price C_i with moneyness m_i. The local polynomial smoothing estimates are obtained by minimizing

$$\sum_i \left\{\sigma_i - \sum_{j=0}^{3}\beta_j(m)(m_i - m)^j\right\}^2 W\left((m_i - m)/h_q\right),$$

where $W(\cdot)$ is a weight function. The estimates are computed as $\hat{\sigma}(m) = \hat{\beta}_0(m), \hat{\sigma}'(m) = \hat{\beta}_1(m)$ and $\hat{\sigma}''(m) = 2\hat{\beta}_2(m)$. Substituting the estimates to the above expressions gives an estimate of q. The density estimates are defined on the scale of S_T. To define the density on the same returns scale $r_T = \log(S_T/S_t)$ as p, a simple transformation can be applied:

$$q(r_T) = q(S_T)S_T.$$

Notice that all results are shown on a continuously compounded 1-month period returns $R_T = 1 + r_T = 1 + \log(S_T/S_t)$.

3.6 Word on Asymptotics

There are two layers of estimation involved. The first step deals with individual curve estimation and the second step introduces shape invariant modeling. The shape invariant modeling is largely robust to how the data are prepared before entering the iterative algorithm and the resulting estimates are interpreted as conditional on the individual curves. Therefore, the main estimation error arises in the first stage where p and q are separately estimated with possibly different sample sizes and separately chosen bandwidths.

In practical terms, the sample size used in estimating p is normally of smaller order, say n compared to $N=nM$ for q for a constant M. This is due to the difference between the daily observations available for estimating p and the intraday observations available for estimating q. Thus it might be expected that the estimation error will be dominated by the estimation error of p. On the other hand, the underlying function p for which simple kernel estimation is used is much simpler and more stable compared to q for which nonparametric second derivative estimation is required.

Because the estimates of ratios are constructed from the ratio of the estimates, we can decompose the error as

$$\begin{aligned}\hat{\mathcal{K}}(u)-\mathcal{K}(u) &= \frac{\hat{q}(u)}{\hat{p}(u)}-\frac{q(u)}{p(u)}\\ &\simeq \frac{\hat{q}(u)-q(u)}{p(u)}-\frac{q(u)}{p(u)}\frac{\hat{p}(u)-p(u)}{p(u)}.\end{aligned}$$

Numerical instability might occur in the region where $\hat{p}\approx 0$ however this is not of theoretical concern. In fact, the pricing kernel is the Radon-Nikodym derivative of an absolutely continuous measure, and thus p and q are equivalent measures, that is, the null set of p is the same as the null set of q. So we can limit our attention to the case where $p(u)>\epsilon$ for some constant ϵ. Provided that $p(u)>\epsilon$ and $q(u)>\epsilon$, the asymptotic approximation is straightforward and asymptotic bias and variance can be computed from

$$\begin{aligned}\mathrm{E}\left[\hat{\mathcal{K}}(u)-\mathcal{K}(u)\right] &\simeq \frac{\mathrm{E}\left[\hat{q}(u)-q(u)\right]}{p(u)}-\frac{q(u)}{p(u)}\frac{\mathrm{E}\left[\hat{p}(u)-p(u)\right]}{p(u)}\\ &= \mathcal{O}\left(h_q^4\right)+\mathcal{O}\left(h_p^2\right)+o\left(h_p^2+h_q^4\right),\\ \mathrm{Var}\left[\hat{\mathcal{K}}(u)-\mathcal{K}(u)\right] &\simeq \mathcal{K}^2(u)\left\{\frac{\mathrm{Var}\left[\hat{q}(u)\right]}{q^2(u)}+\frac{\mathrm{Var}\left[\hat{p}(u)\right]}{p^2(u)}\right\}\\ &= \mathcal{O}\left\{\left(Nh_q\right)^{-1}\right\}+\mathcal{O}\left\{\left(nh_p\right)^{-1}\right\}+o\left\{\left(Nh_q\right)^{-1}+\left(nh_p\right)^{-1}\right\}.\end{aligned}$$

Since $\hat{q}$ involves estimation of second derivative of a regression function, the error is dominated by the estimation of q. The optimal rate of convergence for q is $\mathcal{O}(N^{-2/9})$

while that for p is $\mathcal{O}(n^{-2/5})$. These will be equivalent when $M=\mathcal{O}(n^{39/15})>\mathcal{O}(n^2)$. In practice M is of much smaller order and therefore the leading error terms come from the estimation of q. Ait-Sahalia and Lo (2000) showed in a similar framework that the error is dominated by the estimation of q and for the purpose of asymptotics p can be regarded as a fixed quantity. For this reason we actually implement a semiparametric estimator for q to stabilize the estimator.

Consistency and asymptotic normality of the parameter estimates are shown in Härdle and Marron (1990) for two curves and in Kneip and Engel (1995) for multiple curves. We write the approximate distribution for $\hat{\boldsymbol{\theta}}_t$ as

$$\hat{\boldsymbol{\theta}}_t \approx \mathrm{N}(\boldsymbol{\theta}_t, \Sigma_t).$$

Due to the iterative algorithm, the asymptotic covariance matrix is more complicated for multiple curves but Kneip and Engel (1995) show that, as the number of curves increases, the additional terms arising in the covariance matrix is of lower order than the standard error term due to nonlinear least square methods. There is no suggested estimate for the asymptotic covariance matrix but a consistent estimate can be constructed as in standard nonlinear least square methods. Define the residual $\hat{e}_{tj}=\hat{\mathcal{K}}_t(u_j)-\tilde{\mathcal{K}}_t(u_j)$ where $\hat{\mathcal{K}}$ is the initial estimates and $\tilde{\mathcal{K}}$ is the SIM estimates and let

$$\hat{\sigma}_t^2=\frac{1}{n}\sum_{j=1}^{n}\hat{e}_{tj}^2.$$

The covariance matrix can be estimated as

$$\hat{\Sigma}_t=\hat{\sigma}_t^2\left[n^{-1}\sum_{j=1}^{n}\left\{\nabla_{\boldsymbol{\theta}}\tilde{\mathcal{K}}_t\left(u_j;\tilde{\boldsymbol{\theta}}\right)\right\}\left\{\nabla_{\boldsymbol{\theta}}\tilde{\mathcal{K}}_t\left(u_j;\tilde{\boldsymbol{\theta}}\right)\right\}^{\top}\right]^{-1},$$

where $\nabla_{\boldsymbol{\theta}}\mathcal{K}(u;\boldsymbol{\theta})$ is the first derivative of the function, given by

$$\begin{aligned}
\frac{\partial\mathcal{K}(u)}{\partial\theta_1} &= g\left(\frac{u-\theta_3}{\theta_2}\right),\\
\frac{\partial\mathcal{K}(u)}{\partial\theta_2} &= -\frac{\theta_1}{\theta_2}\left(\frac{u-\theta_3}{\theta_2}\right)g'\left(\frac{u-\theta_3}{\theta_2}\right),\\
\frac{\partial\mathcal{K}(u)}{\partial\theta_3} &= -\frac{\theta_1}{\theta_2}g'\left(\frac{u-\theta_3}{\theta_2}\right),\\
\frac{\partial\mathcal{K}(u)}{\partial\theta_4} &= 1.
\end{aligned}$$

To see whether the location or scale parameters are different between any pair of curves, we can compute the standard errors of the estimates to make a comparison. A formal hypothesis testing also appears in Härdle and Marron (1990) for kernel-based estimates and in Ke and Wang (2001) for spline-based estimates. For example

Table 1 Parameter values of $\boldsymbol{\theta}$

	Distribution	Mean	Standard deviation
θ_1	Log-normal	1	0.33
θ_2	Log-normal	1	0.28
θ_3	Normal	0	0.27
θ_4	Normal	0	0.35

we might be interested in testing whether a location or a scale parameter can be removed.

Although these results are practically relevant, we note that the methods mentioned all assume direct observations of the underlying function of interest, with one smoothing parameter selection involved. Obtaining comparable rigorous results for our estimator is complicated in the present situation due to the nonstandard nature of the estimator being a ratio of two separate nonparametric estimates with independent bandwidths. We consider this out of scope of this paper and leave it for separate work.

4 NUMERICAL STUDIES OF SIM ESTIMATION

Applying the SIM to pricing kernels involves two rather separate estimation steps, the initial estimation of the pricing kernels and the SIM estimation given the pre-estimates. The former has been studied extensively and in particular the properties of the nonparametric methods that we have used are well established in the literature. This section mainly concerns the latter.

We identify the two main factors that could affect the performance of SIM estimation to be error misspecification and smoothing parameter selection for the individual curves. Their effects are evaluated in the following simulation studies. The effects on pricing kernel estimation are separately studied in Section 5.4, in comparison to the standard nonparametric approach used in Jackwerth (2000).

4.1 Generating Curves

In each simulation 50 curves are generated at 50 (random uniform) grid points. In order to mimic the common shape of the observed pricing kernel, we generated the common curve by a ratio of two densities

$$g(u) = q_0(u)/p_0(u),$$

where p_0 is density of Gamma(0.8,1) distribution and q_0 is density of mixture $w*$ Gamma$(0.2, 1)+(1-w)*N(0.91, 0.3^2)$ distribution with $w=0.3$. In accordance with the normalization scheme, the $\boldsymbol{\theta}$ values are set as in Table 1. The values of the

Table 2 Parameter values for error specification

		Error 1	Error 2	Error3
Case 1	σ	0.02	0.05	0.10
Case 2	ϕ	0.75	0.75	0.75
	σ_u	0.02	0.03	0.09
Case 3	α	−3.69	−2.99	−2.30
	β	0.75	0.52	0.53
	σ_v	0.01	0.02	0.02
Case 4	α	−2.41	−1.89	−1.39
	β	0.45	0.40	0.42
	ϕ	0.75	0.45	0.45
	σ_v	0.10	0.25	0.25

standard deviation were chosen to be similar to the observed ones in the real data example.

4.2 Error Specification

For the error specification, we have included dependent errors in time as well as in moneyness as following.

- Case 1: Independent error: $\varepsilon_{t,j} \sim N\left(0, \sigma^2\right)$
- Case 2: Dependent error in moneyness:

$$\varepsilon_{t,j} = \phi \varepsilon_{t,j-1} + u_{t,j}, \qquad \text{the set of the } u_{t,j} \sim N\left(0, \sigma_u^2\right)$$

- Case 3: Dependent error in time: $\varepsilon_{t,j} \sim N\left(0, \sigma_t^2\right)$

$$\log\left(\sigma_t\right) = \alpha + \beta \log\left(\sigma_{t-1}\right) + v_t, \qquad v_t \sim N\left(0, \sigma_v^2\right)$$

- Case 4: Dependent error in moneyness and time:

$$\varepsilon_{t,j} = \phi \varepsilon_{t,j-1} + u_{t,j}, \qquad u_{t,j} \sim N\left(0, \sigma_{ut}^2\right),$$

$$\log\left(\sigma_{ut}\right) = \alpha + \beta \log\left(\sigma_{u,t-1}\right) + v_t, \qquad v_t \sim N\left(0, \sigma_v^2\right)$$

Cases 1 and 2 are commonly assumed but Cases 3 and 4 were rarely used in the literature with SIM estimation. Table 2 lists the parameter values used for simulation. These values are chosen to be comparable in terms of overall variability among cases.

4.3 Smoothing Parameter Selection

We consider three versions of the least squares cross-validation (CV) based criteria for bandwidth selection:

$$\mathrm{CV}_t(h)=\sum_{i=1}^{n}\left\{Y_{t,i}-\widehat{\mathcal{K}}_{t,h}^{-(i)}(u_i)\right\}^2,$$

where $\widehat{\mathcal{K}}_{t,h}^{-(i)}$ is the local linear fit without using the i-th observation. For each observed curve we find the optimal bandwidth $h_t^*=\mathrm{argmin}\mathrm{CV}_t(h)$. Due to considerable variability in the x-dimension we standardize the optimal bandwidths ($\tilde{h}_t^*=h_t^*/s_t$), where s_t is the empirical standard deviation of u_i, and we choose the common bandwidth as follows:

$$h_{opt,1}=\max(\tilde{h}_t^*)\qquad h_{opt,2}=\mathrm{average}(\tilde{h}_t^*)\qquad \text{or}\qquad h_{opt,3}=\mathrm{argmin}\sum_t \mathrm{CV}_t(h).$$

Finally, we multiply h_{opt} by s_t and use these values to perform smoothing of each curve.

4.4 Results of Simulation

We considered various simulation scenarios based on the combinations of the case of errors and bandwidth selection methods. Table 3 summarizes the results of the goodness of fit measured by MSE for the case $\sigma=0.05$. For comparison we added in the last row the MSE for the standard nonparametric estimates based on individual optimal bandwidths to their advantage. For larger error ($\sigma=0.1$, not shown) we also observed some dramatic deterioration with Case 4. Nevertheless, the simulation studies suggest that the overall error is in the same order of magnitude and we suspect that the impact of these factors is limited. The fit was however best with smoothing parameters selected by h_1.

5 REAL DATA EXAMPLE

We use intraday European options data on the Deutscher Aktien index (DAX), provided by European Exchange EUREX and maintained by the CASE, RDC SFB 649 (http://sfb649.wiwi.hu-berlin.de) in Berlin. We have identified options data with maturity one month (31 working days/23 trading days) from June 2003 to June 2006 from DAX 30 Index European options, which adds up to 37 days.

We obtain the initial estimates for p and q as described in Section 3.4. For the choice of kernel functions, we have used quartic function for both p and q.

Table 3 Comparison of SIM estimation with respect to error misspecification and smoothing parameter selection

		$\sigma = 0.05$			
methods	parms.	case 1	case 2	case 3	case 4
h_1	θ_1	31	32	67	65
	θ_2	60	70	84	77
	θ_3	54	62	81	76
	θ_4	32	32	77	75
	$\mathcal{K}_i$s	1.2	1.6	1.5	1.5
h_2	θ_1	67	68	80	69
	θ_2	115	115	110	99
	θ_3	111	110	105	103
	θ_4	70	72	99	85
	$\mathcal{K}_i$s	1.1	1.6	1.9	1.9
h_3	θ_1	67	71	67	73
	θ_2	115	108	91	82
	θ_3	111	100	88	84
	θ_4	70	74	83	88
	$\mathcal{K}_i$s	1.1	1.6	1.8	1.8
	np$\mathcal{K}$	3.5	2.0	4.2	3.6

Numbers are MSE multiplied by 10000. $\mathcal{K}_i$s computes the average MSE for all curves from SIM and np$\mathcal{K}$ without SIM but using individual optimal bandwidths for each curve.

5.1 Estimation of the Risk Neutral Density *q*

The stock index price varies within one day and we would need to identify the price at which a certain transaction has taken place. However, several authors (e.g. Jackwerth, 2000) report that the intraday change of the index price is stale and we use instead the prices of futures contracts closest to the time of the registered pair of the option and strike prices to derive the corresponding stock price, corrected for dividends and difference in taxation following a methodology described in Fengler (2005).

The data contains the actually traded call prices, the implied stock index price corrected for the dividends from the futures derivatives on the DAX, the strike prices, the interest rates (linearly interpolated based on EURIBOR to approximate a *riskless* interest rate for the specific option's time to maturity), the maturity, the type of the options, calculated moneyness, calculated Black and Scholes implied volatility, the volume, and the date. For each day, we use only at-the-money and out-of-the-money call options and in-the-money puts to compute the Black–Scholes implied volatilities. This guarantees that unreliable observations (high volatility) will be removed from our estimation samples. Since the intraday stock price varies, we use its median (S_t) to compute the risk neutral density and correct the strike

price to preserve the ratio relative to the underlying stock price. For this price, we verify if our observations satisfy the no arbitrage condition:

$$S_t \geq C_i \geq \max(S_t - X_i e^{-r\tau}, 0),$$

where X_i is the adjusted strike price and C_i is the corresponding call price. For the remaining observations (X_i, C_i) we compute the (m_i, σ_i) counterparts for the fixed S_t by implicitly assuming that the volatility does not depend on the changes in the intraday stock price. The estimates are computed based on these pairs (m_i, σ_i).

5.2 Estimation of the Historical Density *p*

We compute the nonparametric kernel density estimates as described earlier. Jackwerth (2000) argues that some discrepancies between the nonparametric estimates are attributed to overlapping and nonoverlapping windows of the past observations selected. For comparison to the earlier works, we also experimented with a choice of time varying equity premium and constant equity premium (we demean the densities and supplant it with the risk free rate on the estimation day plus 8% equity premium per annum as in Jackwerth (2000) adjusted for the corresponding maturity), overlapping and nonoverlapping returns, window lengths of 2, 4, and 6 years, respectively. The estimates for different choices of parameters are then compared subsequently in terms of pricing kernel, implied risk aversion and implied utility function estimation. We find that with varying degrees of assumptions on the model, common characteristics such as peaks and skewness are reportedly observed in a wide range of estimates.

5.3 Smoothing Parameter Selection

In contrast to the simulation studies, the effect of smoothing parameter is less transparent with real data when we estimate p and q separately. At first glance, the bandwidth selection for q seems more influential than that of p in gauging performance of the estimates, as it involves derivative estimation. Figure 5 examines the effect of the bandwidth choices on $\hat{q}$. Top left panel shows the implied volatility estimates overlayed, the top right shows the first derivative estimates and bottom left shows the second derivative estimates, respectively, which are used as inputs to create the estimates of q on bottom right panel. The bandwidths used are (0.05, 0.10, 0.15, 0.20). With the apparent undersmoothing at the smallest bandwidth, there is notable variability in terms of smoothness in estimation of implied volatility and its derivatives, however the resulting density estimates demonstrate robustness. Similar observations are made to other dates. However by smoothing on implied volatility domain, we find that the estimates are stable with relatively a wide range of bandwidth choices.

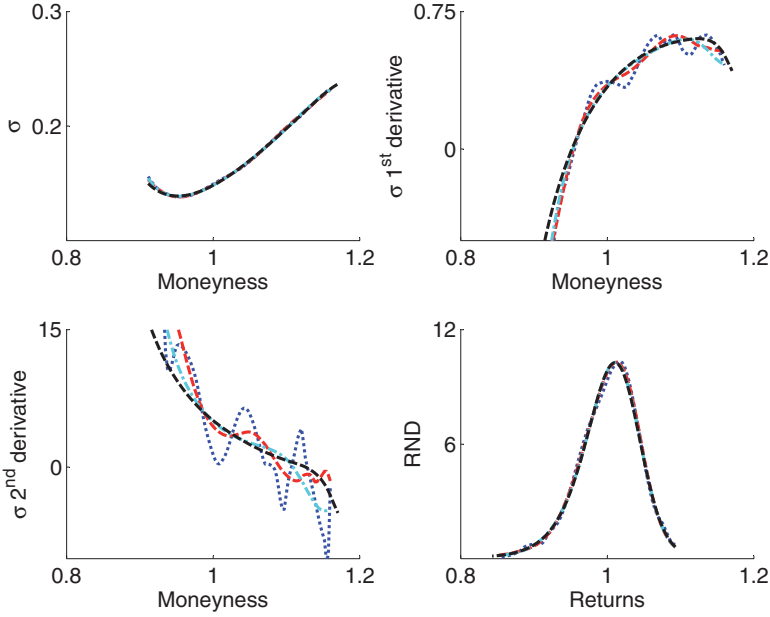


Figure 5 Example of q estimates with varying bandwidths (0.05, 0.1, 0.15, 0.20). The first three panels show estimates of implied volatility, its first and second derivative. The corresponding densities are shown in lower right panel. Estimates are stable for a wide range of bandwidths choices.

For a systematic choice, we employed a version of CV criteria ($h_{opt,1}$ defined in Section 4.3) for p and q estimation. For estimation of q, we have used the least squares CV for local cubic estimation to include the second derivative of σ:

$$\mathrm{CV}(h_q) = \sum_{i=1}^{n}\sum_{j\neq i}^{n}\left\{\sigma_i - \widehat{\sigma}^{(0)}_{h_q,-i}(m_i) - \widehat{\sigma}^{(1)}_{h_q,-i}(m_i)(m_j - m_i) - \frac{1}{2}\widehat{\sigma}^{(2)}_{h_q,-i}(m_i)(m_j - m_i)^2\right\}^2 w(m_i),$$

where $\widehat{\sigma}^{(k)}_{h_q,-i}$ is the k-th derivative estimate without the i-th observation (m_i, σ_i) and $0 \leq w(m_i) \leq 1$ is a weight function. The h_1-optimal bandwidth in implied volatility space turns out to be $h_q = 0.2$.

For estimation of p, we have used the likelihood CV for each curve on returns scale:

$$\log L(h_p) = \sum_{i=1}^{n} \log \hat{p}^{-(i)}_{h_p}(r_i),$$

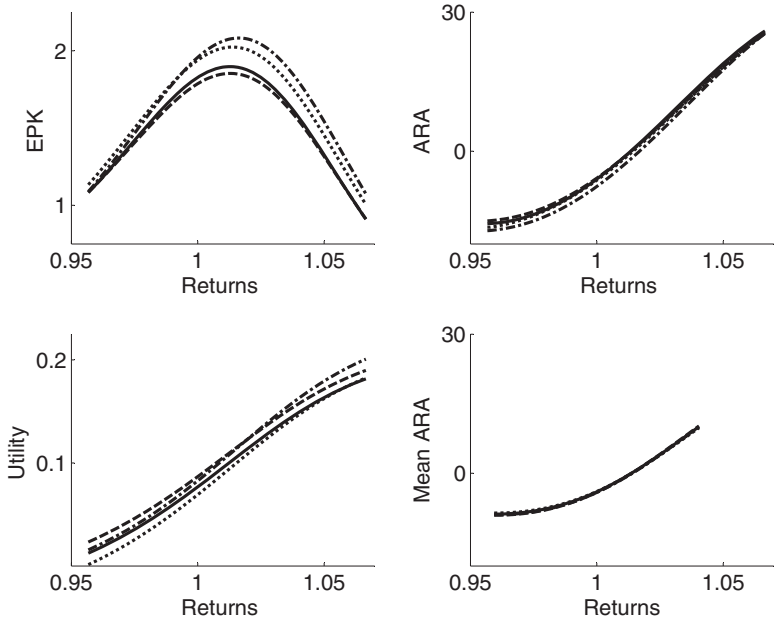


Figure 6 Illustration of SIM with common EPK, ARA, utility function, and mean ARA.

where $\hat{p}_{h_p}^{-(i)}(r_i)$ is the leave-one-out kernel estimator for $p_{h_p}(r_i)$. However, we found that the optimal bandwidth selected tends to systematically oversmooth and thus we chose a smaller value close to the maximum of individually optimal bandwidths, which is in our case $h_p = 0.05$.

5.4 Estimation of Pricing Kernels, ARA and Utility Function

We have considered in Section 5.2 various options for the parameter choice in estimating p and have ended up with 12 series of pre-estimates of pricing kernel. We are interested in seeing how these choices influence the estimated common curves and θ_t parameters by SIM. Since, as it turns out, the results are very similar among specifications we depict graphically only four of them in Figures 6 and 7: those based on nonoverlapping (solid) and overlapping (dashed) returns over the last two years, nonoverlapping returns over the last four (dot-dashed) and six (dotted) years, respectively with varying equity premium. The added lines in Figure 7 are 95% pointwise confidence band for the first series of pre-estimates.

The common curves are represented in Figure 6. All estimates display a *paradoxical* feature: pricing kernel has a bump, ARA has a region of negative values that correspond to the increasing region in the pricing kernel, utility function has a convex region in the domain around the peak of the pricing kernel. The variability

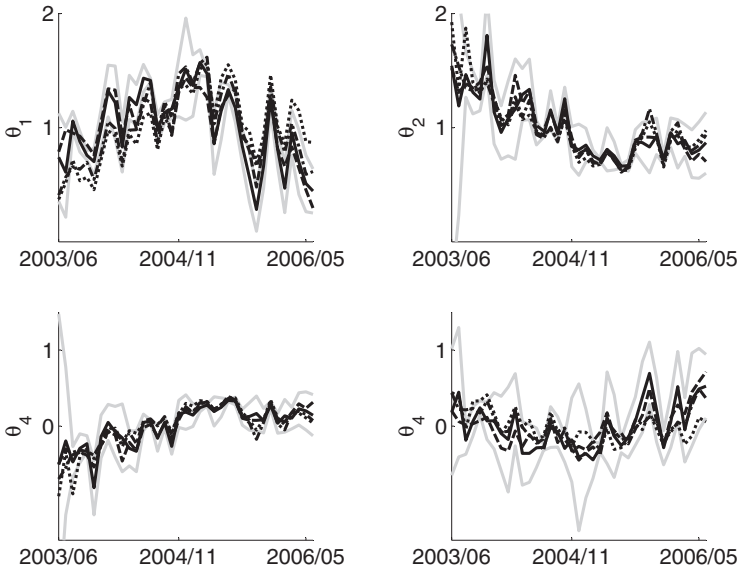


Figure 7 Estimated SIM parameters under variations in the choice of the window lengths of returns values.

among curves is expressed by θ_t-s. In Figure 7, we observe that the main difference in the dynamics of different series has to do with the magnitude but less with the direction of change. In addition, we computed the mean of implied ARA corresponding to our estimation period by computing the sample average and found that it was similar to the the mean ARA for S&P500 appearing in Figure 3C—19 March 1991 to 19 August 1993 in Jackwerth (2000), and to a certain extent to the yearly average from 2003 and 2005 shown in Figure 4 in Chabi-Yo *et al.* (2008). It is worth noting that the mean ARA and the common ARA curves differ a great deal due to the nonlinear transformation involved in deriving ARA from the pricing kernel, e.g. see Equation (2) in Section 2.4. This is not surprising since the interpretation of common curve is different from the average curve, in particular the common curve and the mean curves have different scales of the x-domains—by means of registration.

5.5 Relation to Macroeconomic Variables

With an aid of the SIM model for EPK, we wish to characterize changes in risk patterns in relation to economic variables of interest. Before doing this, we should mention that in the case of nonstandard common curves—in our empirical example the peak does not occur at 0—both θ_1 and θ_2 introduce a shift effect in EPK together with its shape effect. In order to disentangle these effects and improve interpretation we first standardize the EPK curves by the location of the peak before applying SIM.

Table 4 Correlation table for the first difference of SIM parametethers and the selected macro economic variables

	θ_1	θ_2	θ_3	θ_4	CS	DAX	YT
θ_1	1.00	0.55*	0.02	0.78*	−0.25	0.38**	−0.26
θ_2		1.00	0.38*	−0.04	0.06	−0.12	−0.39**
θ_3			1.00	−0.18	0.07	−0.21	−0.28***
θ_4				1.00	−0.37**	0.62*	−0.04

*, **, and *** significant at 1%, 5%, and 10% levels, respectively.

This introduces two more parameters, the horizontal and vertical coordinates of the peaks in the analysis. Since their shift effect is comprised by parameters θ_3 and θ_4 we will not treat them here separately.

Previous studies trying to link the parameters describing risk attitudes to the business conditions include Rosenberg and Engle (2002). Based on power pricing kernel specifications they show that risk aversion is counter-cyclical. Other related work investigates the relation between equity premiums, (e.g., Fama and French, 1989), smile asymmetry of volatility (Bekaert and Wu, 2000; Drechsler and Yaron, 2010), or market efficiency (Marshall, Cahan, and Cahan, 2008). The advantage of our approach over Rosenberg and Engle (2002) is that it allows us to identify how the change in economic variables relates to the shape of a nonparametrically estimated pricing kernel. Due to limited sample size—37 observations—it is impossible to estimate a structural model that correctly deals with the simultaneity of our set of dependent variables. Further research will involve the estimation of a (S)VAR specification, in order to account for the aforementioned endogeneity. We instead evaluate the potential univariate correlations between the estimated θ_t parameters and macroeconomic variables associated with the business cycle and interpret our results from the perspective of local EPK and risk aversion functions. We use the following variables that have a revealed relation with the state of the economy: credit spread (CS) is the difference between the yield on the corporate bond, based on the German CORPTOP Bond maturing in 3–5 years, and the government bond maturing in 5 years; the yield curve slope (YT) refers to the difference between the 30-year government bond yield and three-months interbank rate; short-term interest rate (IR) is the three-months interbank rate; and DAX 30 Performance index as a proxy for consumption. Depending on data availability we collect daily or monthly data. Tests on unit roots failed to reject stationarity in all parameter series and economic variables; we therefore work with their first difference. For conciseness we present only the correlation table for nonoverlapping returns over the past two years with varying equity premium and interpret the results below in relation to Figure 3.

In Table 4, we read significant positive correlation between changes in θ_1 and DAX and negative one with the credit spread, indicating that the EPK becomes more pronounced when the economic indicators suggest an expanding economy; changes in θ_2 and YT are negatively correlated, suggesting that risk aversion slope

becomes locally steeper during economic boom. The same interpretation holds for the negative correlation between changes in θ_3 and YT. The height of the peak varies with the returns on the index, pointing to an increasing local risk proclivity in periods of economic expansion. We have not found any significant correlation between changes in θ_t and in the short-term interest rate. Finally, we observe a positive correlation between the increments in θ_1 and θ_2 that suggests that over periods of concerted negative evolution of the economic indicators the EPK bump will shrink in both horizontal and vertical direction, possibly leading to an overall decreasing EPK.

In summary, the sense of the relations between the indicators of the business cycle and the parameters that summarize risk preferences indicates that locally risk loving behavior is procyclical. These findings are also in line with the results found in Rosenberg and Engle (2002).

REFERENCES

Ait-Sahalia, Y., and J. Duarte. 2003. Nonparametric Option Pricing under Shape Restrictions. *Journal of Econometrics* 16: 9–47.

Ait-Sahalia, Y., and A. Lo. 2000. Nonparametric Risk Management and Implied Risk Aversion. *Journal of Econometrics* 94: 9–51.

Bekaert, G., and G. Wu. 2000. Asymmetric Volatility and Risk in Equity Markets. *Review of Financial Studies* 13: 1–42.

Bondarenko, O. 2003. Estimation of Risk-Neutral Densities using Positive Convolution Approximation. *Journal of Econometrics* 116: 85–112.

Breeden, D., and R. Litzenberger. 1978. Prices of State-Contingent Claims Implicit in Option Prices. *The Journal of Business* 51: 621–651.

Chabi-Yo, Y., R. Garcia, and E. Renault. 2008. State Dependence can Explain the Risk Aversion Puzzle. *Review of Financial Studies* 21: 973–1011.

Drechsler, I., and A. Yaron. 2010. What's Vol Got to Do with it. *Review of Financial Studies* 20: 1–45.

Fama, E. F., and K. R. French. 1989. Business Conditions and Expected Returns on Stocks and Bonds. *Journal of Financial Economics* 25: 23–49.

Fengler, M. R. 2005. *Semiparametric Modeling of Implied Volatility*. Springer-Verlag Berlin and Heidelberg.

Giacomini, E., and W. Härdle. 2008. "Dynamic semiparametric factor models in pricing kernel estimation." In S. Dabo-Niang and F. Ferraty (eds.), *Functional and Operational Statistics*, Contributions to Statistics, pp. 181–187. Physica-Verlag HD.

Giacomini, E., W. Härdle, and M. Handel. 2008. "Time dependent relative risk aversion." In G. Bol, S. Rachev, and R. Wrth (eds.), *Risk Assessment: Decisions in Banking and Finance*, Contributions to Economics, pp. 15–46. Physica-Verlag HD.

Härdle, W., and J. S. Marron. 1990. Semiparametric Comparison of Regression Curves. *The Annals of Statistics* 18: 63–89.

Hens, T., and C. Reichlin. 2012. "Three Solutions to the Pricing Kernel Puzzle." Technical report, Swiss Finance Institute Research Paper No. 10-14. Available at SSRN: http://ssrn.com/abstract=1582888 or http://dx.doi.org/10.2139/ssrn.1582888

Jackwerth, J. C. 1999. Option-Implied Risk-Neutral Distributions and Implied Binomial Trees: a Literature Review. *Journal of Derivatives* 2: 66–82.

Jackwerth, J. C. 2000. Recovering Risk Aversion from Option Prices and Realized Returns. *Review of Financial Studies* 13: 433–451.

Ke, C., and Y. Wang. 2001. Semiparametric Nonlinear Mixed-Effects Models and their Applications. *Journal of the American Statistical Association* 96: 1272–1298.

Kneip, A., and J. Engel. 1995. Model Estimation in Nonlinear Regression under Shape Invariance. *The Annals of Statistics* 23: 551–570.

Kneip, A., and T. Gasser. 1988. Convergence and Consistency Results for Self-Modeling Nonlinear Regression. *The Annals of Statistics* 16: 82–112.

Lawton, W. H., E. A. Sylvestre, and M. S. Maggio. 1972. Self Modeling Nonlinear Regression. *Technometrics* 14: 513–532.

Leland, H. E. 1980. Who Should Buy Portfolio Insurance? *Journal of Finance* 35: 581–594.

Marshall, B. R., R. H. Cahan, and J. Cahan. 2008. "Technical Analysis Around the World: Does it Ever Add Value?" Technical report, SSRN eLibrary.

Ramsay, J. O., and B. W. Silverman. 2002. *Applied Functional Data Analysis*. Springer Series in Statistics. New York: Springer. Methods and case studies.

Rookley, C. 1997. "Fully Exploiting the Information Content of Intra day Option Quotes: Applications in Option Pricing and Risk Management." Technical report, University of Arizona.

Rosenberg, J. V., and R. F. Engle. 2002. Empirical Pricing Kernels. *Journal of Financial Economics* 64: 341–372.

Stone, C. J. 1982. Optimal Global Rates of Convergence for Nonparametric Regression. *The Annals of Statistics* 10: 1040–1053.

Quantitative Finance, 2013
Vol. 13, No. 5, 675–685, http://dx.doi.org/10.1080/14697688.2012.749420

Routledge
Taylor & Francis Group

Variance swap dynamics

K. DETLEFSEN* and W. K. HÄRDLE

Center for Applied Statistics and Economics, Humboldt Universität zu Berlin, Spandauer Straße 1, 10178 Berlin, Germany

(*Received 23 September 2007; in final form 8 November 2012*)

We compare several parametric and non-parametric approaches for modelling variance swap curves by conducting an in-sample and an out-of-sample analysis using market prices. The forecasted Heston model gives the best overall performance. Moreover, the static Heston model highlights some problems of stochastic volatility models in option pricing of forward starting products.

Keywords: Empirical finance; Equities; Volatility modelling; Term structure

JEL Classification: C5, D4, G1

1. Introduction

Variance swap markets have become liquid and this has led to the trading of options on realized variance. In this way, variance swap markets have become important for investments and hedging. Because of this we analyse the dynamics of variance swap curves in order to find a model that gives a good overall performance—in- and out-of-sample. On the one hand, this may help build models for direct investment. On the other hand, we look at a popular stochastic volatility model and analyse its forecasting performance, which is essential for forward starting products.

For implied volatility surfaces, the dynamics have often been analysed and modeled by factor approaches (see, e.g. Cont and da Fonseca (2002) and Fengler (2005)). As variance swap curves are basically the term structure of implied volatility surfaces, these studies also analyse, in principal, the dynamics of variance swaps. General studies on variance swaps that do not focus on the dynamics are the survey of Demeterfi *et al.* (1999) and the hedging and trading analysis of Carr and Madan (1998). More recently, Bühler (2006) considered a modeling approach for the joint plain vanilla and variance swap market. For yield curves, forecasting questions have been analysed by, for example, Duffie and Kan (1996) and Diebold and Li (2003). Amengual (2009) found a significant improvement of the fit of variance swap prices when using higher-dimensional models with two stochastic factors for the variance. We focus on lower-dimensional models as they are easier to use and give more stable fits.

In section 2 we describe the modeling framework for variance swaps and the analysed models, which comprise a Heston model, a Nelson–Siegel parametrization and a semiparametric approach. In section 3 we conduct an empirical analysis, describing the data, estimating the models and forecasting the variance swap curves. In section 4 we summarize our results.

*Corresponding author. Email: kai.detlefsen@gmx.de

2. Modeling the term structure

In this section we introduce variance swaps, explain the construction of variance swap curves and describe the approach that we use for fitting and forecasting the variance swap curves. We start with the stochastic volatility model of Heston (1993) and derive the corresponding model for the variance curves. In addition to this two-parameter model we consider a generalization with three parameters. Moreover, we describe a semiparametric factor model. Finally, we see how good some stylized facts are replicable in these models.

2.1. Constructing variance swap curves

Variance swaps are forward contracts on future realized volatility. They exchange at expiration the realized annualized variance of the log returns of an underlying against a predefined strike. These contracts vary in several respects: they may or may not assume zero mean of the log returns, they differ with respect to the annualization factor and it must be specified when the underlying is sampled. We assume a zero mean of the returns, use $c = 252$ trading days for annualization with daily sampling and focus on zero strikes.

Given an underlying S, the price of such a variance swap for the period $[0, T]$ with business days $0 = t_0 < \cdots < t_n = T$ is given by

$$\sigma_R^2(T) \stackrel{\text{def}}{=} \frac{c}{n} \sum_{i=1}^{n} \left(\log \frac{S_{t_i}}{S_{t_{i-1}}} \right)^2.$$

At time $t \in (0, T)$, the first part of the variance is already realized while the second is still uncertain. Hence, the prices are composed of the value of the realized variance and the price of the uncertain variance. In our analysis, we will focus on the uncertain part and denote the price for the non-annualized variance that still has to be realized by $V_t(T)$.

At a point in time we observe the prices of variance swaps $V(x_i)/x_i$ with times-to-expiry $x_1, \ldots, x_n$. The variance swap curve at this time is then given by the mapping $T \mapsto V(T)/T$. We call V the variance curve and V' the forward variance curve. The variance swap curve quoted in volatility strikes is given by $T \mapsto \sqrt{V(T)/T}$. Several approaches for modeling variance swap curves are based on forward variance curves, but variance curves or forward variance curves are not observed. Instead, they must be estimated from a discrete set of observed variance swap prices which is often done via (piecewise) polynomial functions for interpolation between the observations. Such an approach often makes the forward variance curves vary significantly from day to day.

Hence we choose a non-parametric method for constructing smooth curves (see, e.g. Härdle *et al.* (2004)). We apply a local quadratic regression to the variance prices $V(x_{ji})$, leading to the following minimization problem:

$$\min_{\beta} \sum_{i=1}^{n} \{V(x_i) - \beta_0 - \beta_1(x_i - x) - \beta_2(x_i - x)^2\} K_h(x_i - x),$$

where the vector $\beta = (\beta_0, \beta_1, \beta_2)$ depends on x. The result $\hat{\beta}(x)$ is a weighted least-squares estimator where the variance curve is given by $\hat{\beta}_0$ and its first derivative by $\hat{\beta}_1$. We use the quartic kernel K and choose the bandwidth h by a rule of thumb described by Fan and Gijbels (1996). We do not consider higher-order kernels because of their inferior finite sample bias.

2.2. *Modeling variance swap curves*

On each day, we construct a variance curve to which we fit two parametric models and a semiparametric model. The resulting time series of factor loadings are the basis for the forecasting. We use the functional form derived from the Heston model and, in addition, we consider an extension that leads to the form of Nelson and Siegel (1987). These models are convenient and parsimonious exponential factor approximations. Moreover, we analyse a semiparametric approach.

In the Heston model,

$$\frac{\mathrm{d}S_t}{S_t} = r\mathrm{d}t + \sqrt{\zeta_t}\mathrm{d}W_t^{(1)},$$

$$\mathrm{d}\zeta_t = \kappa(\theta - \zeta_t)\mathrm{d}t + \nu\sqrt{\zeta_t}\mathrm{d}W_t^{(2)}$$

with correlated Wiener processes $W^{(1)}$ and $W^{(2)}$, the prices of (annualized) variance swaps $V(T)/T$ are given by

$$\theta + (\zeta_0 - \theta)\frac{1 - \exp(-\kappa T)}{\kappa T}.$$

Hence, only the short variance ζ_0, the long variance θ and the mean-reversion speed κ determine the variance swap prices. In the Heston model, the smile of the implied volatility surfaces is controlled by the other two parameters: the correlation between the Brownian motions and the volatility of variance ν. These two parameters do not enter the formula for the variance swap price.

The corresponding model for the forward variance curve is given by

$$v(T) \stackrel{\text{def}}{=} V'(T) = \theta + (\zeta_0 - \theta)\exp(-\kappa T).$$

This forward variance curve model implies exactly the above variance swap prices because of the constraint $V(0) = 0$. Reparametrizing this model and writing it in factor notation we obtain for the prices of variance swaps $V(T)/T$

$$z_1 + z_2\frac{1 - \exp(-\kappa T)}{\kappa T}, \tag{1}$$

where $z = (z_1, z_2)$ are the two factor loadings. They correspond to the model parameters $(\theta, \zeta_0 - \theta)$ for the volatility. The reparametrization can be described formally in terms of a reparametrization matrix by

$$\begin{pmatrix} 1 & 0 \\ -1 & 1 \end{pmatrix} \begin{pmatrix} \theta \\ \zeta \end{pmatrix} = \begin{pmatrix} z_1 \\ z_2 \end{pmatrix}.$$

This model for forward variance curves is also called the linearly mean-reverting (forward) variance curve model (Bühler 2006).

As Diebold and Li (2003) obtained good results with the Nelson–Siegel parametrization, we generalize the above model in such a way that the resulting variance swap prices have a Nelson–Siegel parametrization:

$$v(T) = z_1 + z_2\exp(-\kappa T) + z_3\kappa T\exp(-\kappa T).$$

This model is called the double mean-reverting (forward) variance curve model. The variance swap prices $V(T)/T$ are given in this model by

$$z_1 + z_2\frac{1 - \exp(-\kappa T)}{\kappa T} + z_3\left\{\frac{1 - \exp(-\kappa T)}{\kappa T} - \exp(-\kappa T)\right\}. \tag{2}$$

Thus, the generalized Heston model leads exactly to a Nelson–Siegel parametrization for the prices of variance swaps.

While the linearly mean-reverting model is basically the Heston model, the second approach was considered by Bühler (2006), who analysed the conditions for an arbitrage-free joint market of variance swaps and stock. His considerations imply that the mean-reversion speed κ should be constant. In practice, a constant mean-reversion speed is important for the stability of the parameters. For these theoretical and practical reasons we fix this parameter.

We interpret z_{1t}, z_{2t} and z_{3t} as latent dynamic factor loadings for the prices of variances swaps $V(T)/T$. The factor on z_{1t} is the constant 1. As this factor does not decay to zero in the long run it can be interpreted as a long-term factor. The factor on z_{2t} is $\{1-\exp(\kappa T)\}/(\kappa T)$. This function is monotonically decreasing from 1 to 0. As it influences only the short end of the curve it can be interpreted as a short-term factor. Besides these two factors the generalized model also controls the medium term. The factor on z_{3t} is $\{1-\exp(-\kappa T)\}/(\kappa T)-\exp(-\kappa T)$. This mapping increases monotonically from 0 to a peak and then decreases to zero in the long term in a similar way as the second factor. This form explains the interpretation as a medium-term factor. These three factors are presented in figure 1.

The interpretation of these factors corresponds to their meaning in the Heston model: z_1 is the long variance and z_1+z_2 is the short variance. Moreover, these quantities can be recovered from the variance swap curve: from the limits $\lim_{T\to 0}V(T)/T=z_1+z_2$ and $\lim_{T\to\infty}V(T)/T=z_1$ we see that z_1 is the long variance (i.e. θ) and z_1+z_2 is the short variance (i.e. ζ_0). Hence, we have used the parametrization $\zeta_0=z_1+z_2$ and $\theta=z_1$. Thus, the original parameters of the Heston model (θ,ζ) can be recovered by multiplying the inverse of the reparametrization matrix by the factor loadings (z_1,z_2).

Moreover, the parameters have interpretations as level, slope and curvature. As an increase in z_1 increases the whole curve by the same amount, the factor on z_1 represents the level of the curve. An increase of the short-term factor increases the curve more at the short end than at the long end. Hence it controls the slope of the curve. Finally, the third factor moves the middle of the curve while keeping the ends almost fixed. In this way it changes the curvature of the curve. Hence, the difference between the Heston model and its three-factor generalization is the capability to control the curvature.

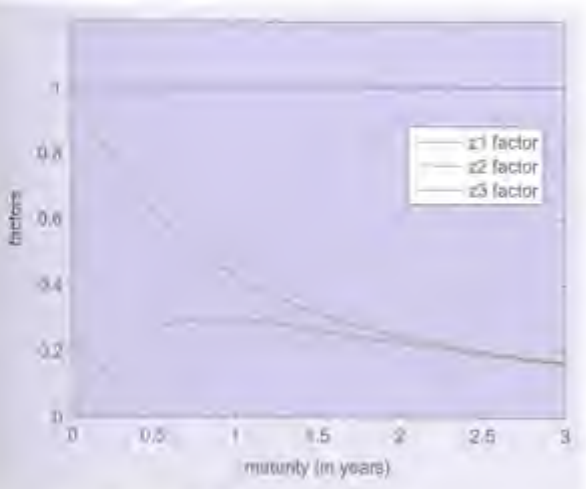


Figure 1. Factors for variance curves in the Heston model and its generalization (*for* $\kappa=2$).

Besides these parametric approaches we analyse a semiparametric model described by Fengler (2005). It offers a low-dimensional representation of variance swap curves that are approximated by basis functions. These basis functions are unknown and have to be estimated from the data. The dynamics of the curves are described by the time series of the corresponding factor loadings.

Let $Y_{i,j}$ be an observed price of a variance swap on day i with maturity $T_j \in \{0.12,0.25,0.5,0.75,1.0,1.5,2.0\}$. Let $X_{i,j}$ be a one-dimensional variable representing the time-to-maturity. Then the model regresses $Y_{i,j}$ on $X_{i,j}$ by

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

where m_0 is an invariant basis function, m_l $(l=1,\ldots,L)$ are the 'dynamic' basis functions and $\beta_{i,j}$ are the factor weights depending on time i. We describe the estimation procedure and the obtained basis functions in section 3.2 where we use actual data.

2.3. *Stylized facts of variance swap curves*

A model of variance swap curves should in principle have a reasonable in-sample fit and be able to reproduce the variety of observed shapes of variance swap curves. A good model should moreover reflect the dynamics of the curves by a reasonable out-of-sample performance.

We consider briefly some stylized facts of variance swap curves and see how the described approaches can model these characteristic shapes. The average variance swap curve is increasing and concave. The slope factor can readily replicate the increasing structure. The concavity can be modeled in the generalized model by the third factor. Variance swap curves show many different shapes in different markets over time. They can be upward- or downward-sloping and some have a hump. These shapes can be replicated, in principle, by varying the three factors accordingly. The short end of the curves is more volatile than the long end. This is reflected in the models because two factors (z_1 and z_2) control the short end while only one factor (z_1) models the long end. The Heston model can replicate many patterns, but not the humps (i.e. the curvature), while the generalized model has, in principle, the capability of replicating all stylized facts. The semiparametric model should be able to replicate these stylized facts because of its more flexible structure.

3. Forecasting the term structure

In this section, we describe the data used, estimate the factor loadings, model them and compare the forecasted variance swap curves.

3.1. The data

The data set studied contains prices of variance swaps on the S&P 500 index between 1 October 2003 and 30 September 2005. These swaps use daily closing prices of the index, have 252 business days as annualization factor and assume a zero mean for the calculation of the variance of the returns. The prices are quoted in volatility strikes and represent the mid-market prices. Hence, we observe on every trading day prices $\sqrt{V(x_i)/x_i}$, $i = 1, \ldots, n$, of variance swaps with times-to-maturity $x_1, \ldots, x_n$. We have around $n = 7$ observations per day.

Our analysis does not require the use of fixed maturities because we always model the entire variance swap curve. But we use fixed maturities in order to simplify the following variance swap curve forecasts. Hence we first create from the discrete data curves by local quadratic smoothing as described in section 2.1. Then we extract the data for the fixed maturities 1.5, 3, 6, 9, 12, 18 and 24 months.

The variance swap prices (not quoted in volatility strikes) and level, slope and curvature of these curves are the basis for the following. In figure 2 we present the smoothed variance swap curves. The figure also shows the variance swap curves quoted in volatility strikes. Although these prices are often quoted in volatility strikes, we estimate and forecast the variance curves because the integrated variance is normally the essential quantity in option pricing. In figure 3 we show the corresponding variance and forward variance curves. The variation of the level is clearly visible for the variance swap curves; the changes in the slope and curvature are less apparent. The changes are more readily observed from the forward variance curves.

We provide some descriptive statistics of the variance swap curves in table 1. Here, we also present the level (defined as the 24 month price), the slope (defined as the 24 month price minus the 1.5 month price) and the curvature (defined as twice the 6 month price minus the 1.5 month price minus the 24 month price). We will see below that these empirical factor loadings are highly correlated with the loadings of the parametric models. In figure 4 we show the median variance swap curve, with pointwise interquartile ranges. The mentioned upward-sloping and concave form is visible.

3.2. Fitting the variance swap curves

We estimate the Heston model and its generalization by minimizing the difference between the observed variance swap curves and the model prices. In the Heston model, these prices are given by

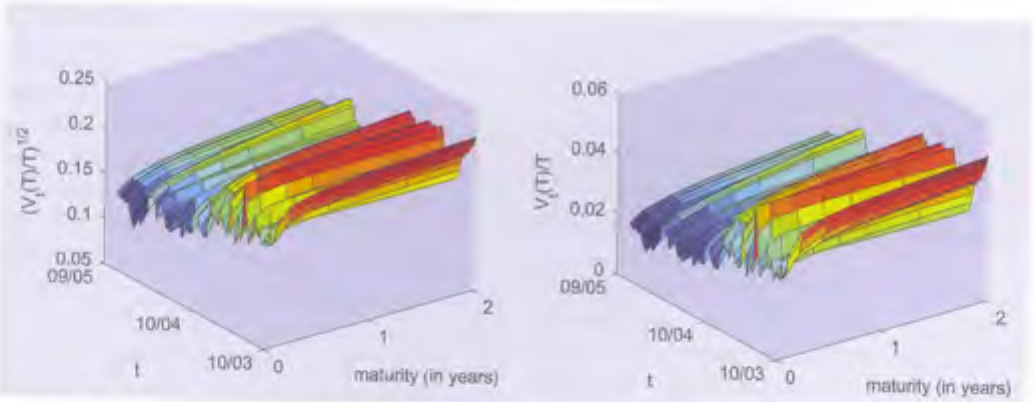


Figure 2. Variance swap curves quoted in volatility strikes (left) and variance swap curves (right), 01/10/03–30/09/05. The sample consists of weekly curves from October 2003 to September 2005 at maturities 1.5, 3, 6, 9, 12, 18 and 24 months.

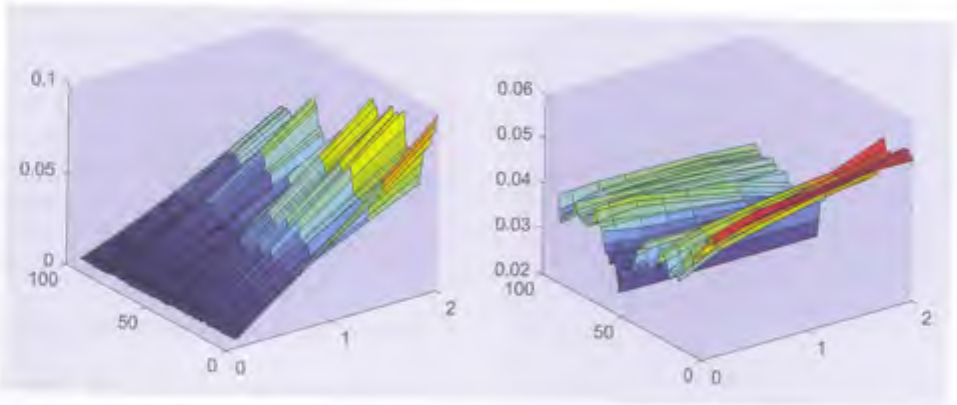


Figure 3. Variance curves (left) and forward variance curves (right), 01/10/03–30/09/05. The sample consists of weekly curves from October 2003 to September 2005 at maturities 1.5, 3, 6, 9, 12, 18 and 24 months.

Table 1. Descriptive statistics of variance swap curves $[E^{-2}]$.

Mat. (months)	Mean	Std. dev.	Min	Max	$\hat{\rho}(1)$	$\hat{\rho}(4)$	$\hat{\rho}(12)$
1.5	2.03	0.65	0.91	4.36	80.1	64.5	51.1
3	2.48	0.70	1.39	4.39	90.1	79.3	65.4
6	2.75	0.73	1.51	4.41	93.3	82.6	64.7
9	2.89	0.73	1.60	4.45	94.1	82.5	61.3
12	2.98	0.72	1.69	4.54	94.5	81.8	57.1
18	3.14	0.70	1.85	4.68	94.9	79.6	47.3
24	3.27	0.69	2.02	4.79	9.50	77.1	36.8
Slope	1.24	0.44	–0.24	2.21	74.1	38.8	–16.6
Curvature	0.20	0.31	–0.50	0.98	83.6	62.0	49.5

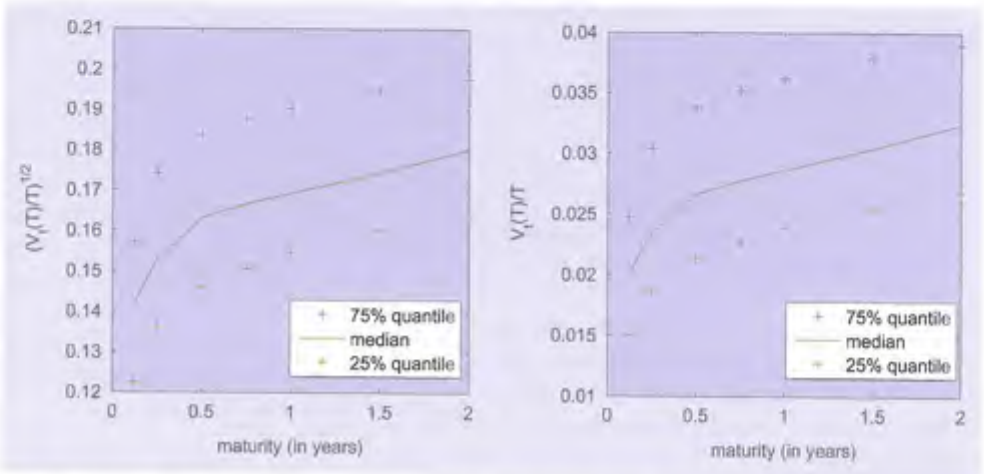


Figure 4. Median data-based forward variance curve quoted in volatility strikes with pointwise interquartile range. For each maturity, we plot the median along with the 25th and 75th quantiles.

$$z_1 + z_2 \frac{1 - \exp(-\kappa T)}{\kappa T},$$

and in the generalized Heston model by

$$z_1 + z_2 \frac{1 - \exp(-\kappa T)}{\kappa T} + z_3 \left\{ \frac{1 - \exp(-\kappa T)}{\kappa T} - \exp(-\kappa T) \right\}.$$

The factor loadings z and the parameter κ can be estimated by nonlinear least squares. In the approach of Nelson and Siegel (1987) for interest rates it is common to fix the parameter κ. As the generalized Heston model leads to variance swap prices in the form of Nelson–Siegel, it makes sense that we also fix the parameter κ. Moreover, it is practical to fix this parameter in the Heston model for the modeling, pricing and hedging of options. Hence, we use $\kappa = 2$ as in Bergomi (2004). Keeping this parameter constant simplifies the numerics considerably because the factor loadings z are given by OLS.

On every day we apply an ordinary least-squares estimation to the variance swap curves. In this way we obtain a time series of estimated factor loadings $(\hat{z}_1, \hat{z}_2, \hat{z}_3)^\top$. As we do not explicitly use weights, the short end is more important for the estimation because we sample more observations with short maturities.

We estimate the factors in the semiparametric model from the first year of our time series. The factors or basis functions $\widehat{m}_l$ and the factor loadings $\widehat{\beta}_{i,l}$ are estimated by minimizing the following least-squares criterion ($\beta_{i,0} = 1$):

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

where K_h denotes a kernel function. The minimization procedure searches through all functions $\widehat{m}_l : \mathbb{R} \to \mathbb{R}$ $l = 0, \ldots, L$ and time series $\widehat{\beta}_{i,l} \in \mathbb{R} (i = 1, \ldots, I, l = 1, \ldots, L)$ by an iterative procedure. The estimates are then orthogonalized and normalized (see Fengler (2005) for details). The estimated factors are plotted in figure 5. They can be interpreted as in the parametric models as level, slope and curvature (see section 2.2). A comparison with the parametric factors in figure 1 reveals that the estimated factors have a different scaling and become negative. The positivity of the parametric factors allows us to ensure the positivity of the resulting curve by imposing simple constraints on the loadings. After the factors have been estimated the factor loadings can be estimated by ordinary least squares as in the parametric models.

Information about the in-sample fit of the models is presented in figure 6. It shows that the Heston and the semiparametric model have problems fitting the short end of the curves. This systematic deviation holds in principle

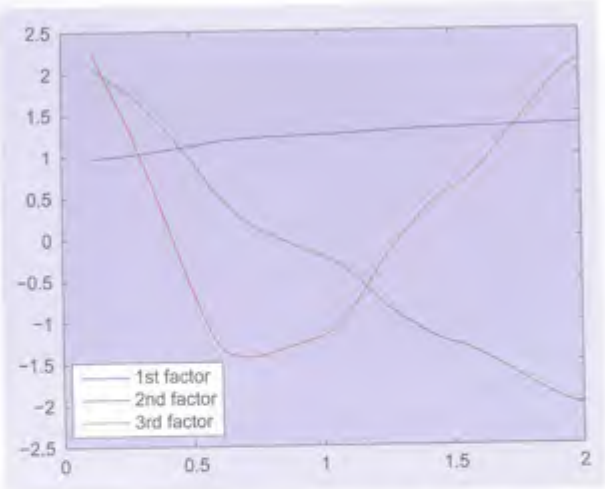


Figure 5. Factors in the semiparametric model estimated from the variance swap curves, 01/10/03–30/09/04.

also for the Nelson–Siegel parametrization. The prices for long maturities show no systematic error for all models and the semiparametric model leads to the smallest errors in this region.

Although the semiparametric model is quite flexible it has problems fitting the short maturities. This can be explained by the fact that the semiparametric model puts less weight on the short maturities than the other models.

In figure 7 we show the time series of the estimated factor loadings for all models. In this figure we have plotted the negative slope loadings of the models and we have scaled the curvature loadings of the models by 0.3. As the factors have interpretations as level, slope and curvature, we compare them with the empirical level, slope and curvature as defined in section 3.1. As the empirical and model quantities are highly correlated we see that the definitions and interpretations of the empirical quantities are appropriate. The empirical level factor is very similar to the level loadings in the generalized Heston model. The corresponding loadings of the Heston model lie above, and the loadings of the semiparametric model lie below the empirical levels. The empirical slope is a good estimator for slope loadings in the Heston and in the generalized Heston model. The empirical curvature differs from the loadings of the models. In all three cases the model and the empirical values are highly correlated.

The loadings of the semiparametric factor model differ the most from the empirical factor loadings. This can partly be explained by the different scaling in the semiparametric factor model. In Table 2 we give some summary statistics of the time series of factor loadings. These statistics confirm the different scaling of the parametric approaches and the semiparametric model. Moreover, we find that the time series of each model are only weakly correlated. The two-factor loadings in the Heston model have an empirical correlation of –0.39. This quantity in the generalized Heston model is –0.33, while the other two correlations of the model are below 0.2. The correlations in the semiparametric factor model are similarly small. Hence, it appears that the number of factors cannot be reduced without sacrificing significant aspects of the model.

3.3. Modeling the factor loadings and forecasting the variance swap curves

Diebold and Li (2003) model the factor loadings of the Nelson–Siegel framework by univariate AR(1) processes for

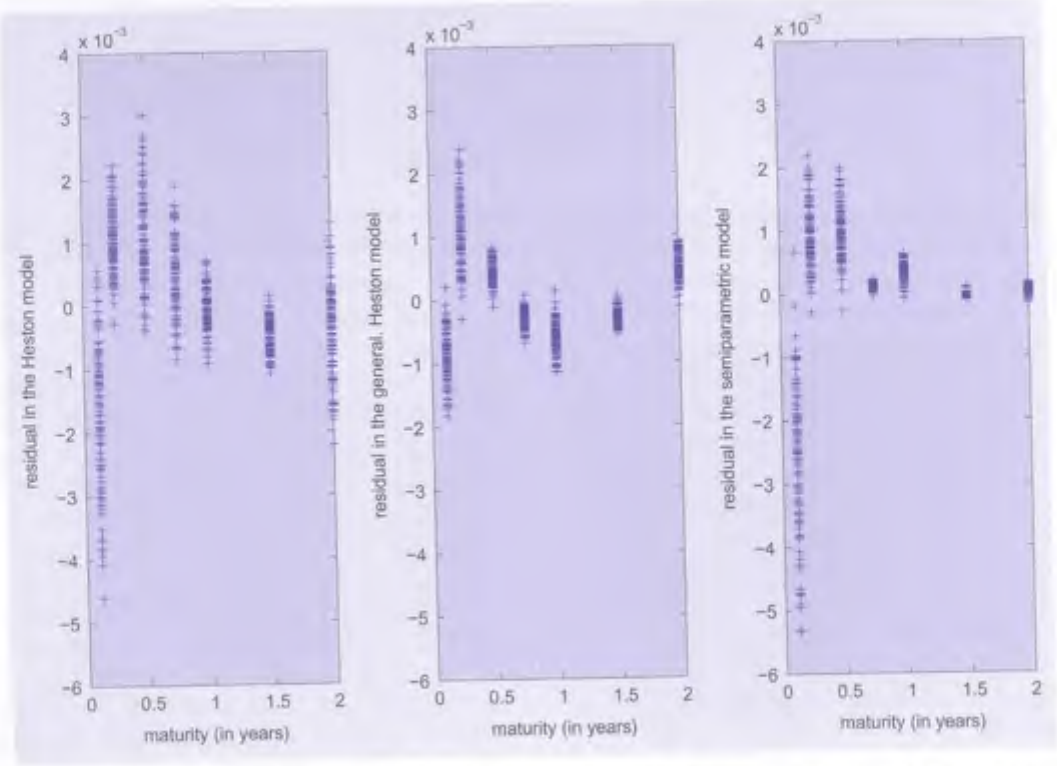


Figure 6. Variance swap curve residuals, 01/10/03–30/09/05. left, Heston; middle, generalized Heston; right, semiparametric model.

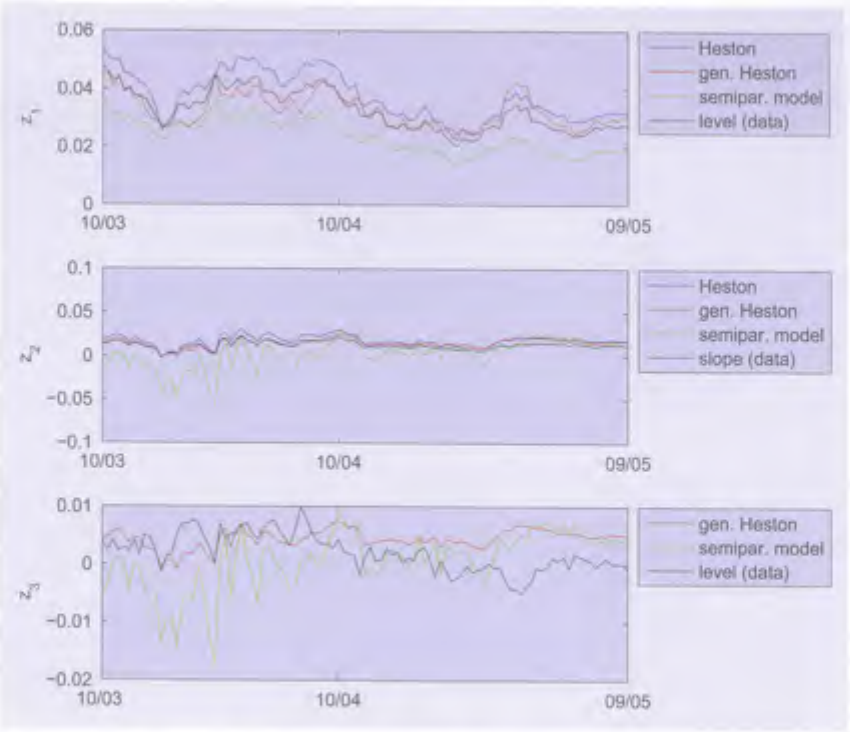


Figure 7. Factor loadings in the models and in the data.

Table 2. Descriptive statistics of the factor loadings $[E^{-2}]$.

Model	Factor	Mean	Std. dev.	Min	Max	$\hat{\rho}(1)$	$\hat{\rho}(4)$	$\hat{\rho}(12)$
Heston	z_1	3.74	0.80	2.21	5.32	95.4	75.2	31.1
	z_2	−1.74	0.61	−3.08	0.33	74.1	38.1	−17.9
Generalized Heston	z_1	3.27	0.58	2.45	4.60	87.9	51.9	−26.6
	z_2	−1.46	0.56	−2.50	0.27	79.8	49.0	−11.2
	z_3	1.34	1.30	−1.37	4.61	81.3	57.6	45.7
Semiparam. model	z_1	2.37	0.58	1.36	3.67	93.5	81.1	58.1
	z_2	−0.01	0.10	−0.17	0.33	73.1	50.0	23.1
	z_3	0.00	0.03	−0.06	0.06	70.0	36.9	24.8

yield curves. Cont and da Fonseca (2002) also use these models for the factor loadings in their principal components analysis of implied volatility surfaces. Hence, we follow this approach. Moreover, more complex ARMA models did not improve the forecasting results in tests. We do not consider multivariate AR processes because there is only little correlation between the factor loadings (see section 3.2). In addition, the use of AR(1) processes allows us to more easily compare our results with the findings of Diebold and Li (2003).

The resulting forecasts of the variance swap curves τ weeks ahead at time t are given by

$$\widehat{V_{t+\tau}(T)}/T = \hat{z}_{1,t/t+\tau} f_1(T) + \hat{z}_{2,t/t+\tau} f_2(T) + \hat{z}_{3,t/t+\tau} f_3(T),$$

where $\hat{z}_{i,t/t+r}$ are the forecasts of the i th factor loading and f_1, f_2 and f_3 are the factors. These factor loading forecasts can be computed by regressing the loadings at $t+h$ on the loadings at t. As we obtained better results using repeated 1-day forecasts we used this second approach instead. In figure 8 we show the autocorrelation functions of the residuals of the estimated AR(1) models. In general, only a few autocorrelations lie slightly outside the 95% confidence interval for all models. Hence, the models seem to be in line with the data.

As the models describe the variance swap curves by the factor loadings z, we forecast the loadings $(\hat{z}_{1t}, \hat{z}_{2t}, \hat{z}_{3t})$. Our data set consists of observations from 10/2003 to 9/2005. We use the first part of the data for the estimation of the factor loadings and forecast the variance swap curves of the second year. In the semiparametric model, the factors are estimated from the data of the first year. Then we keep these factors fixed for the forecasting. Actually, these factors differ only slightly from the factors estimated from the whole data set. If we want to forecast at time t the variance curve at time $t+\tau$, then we use the whole history of the factor loadings up to time t, i.e. $z_1, \ldots, z_t$.

We made forecasts for 1 week, 1 month, 3 months and 6 months. In Tables 3 and 4 we show the results for 1 week and 6 months forecasts. As before, we consider the variance swap curves at maturities 1.5, 3, 6, 9, 12, 18 and 24 months. In addition to the three models described above we

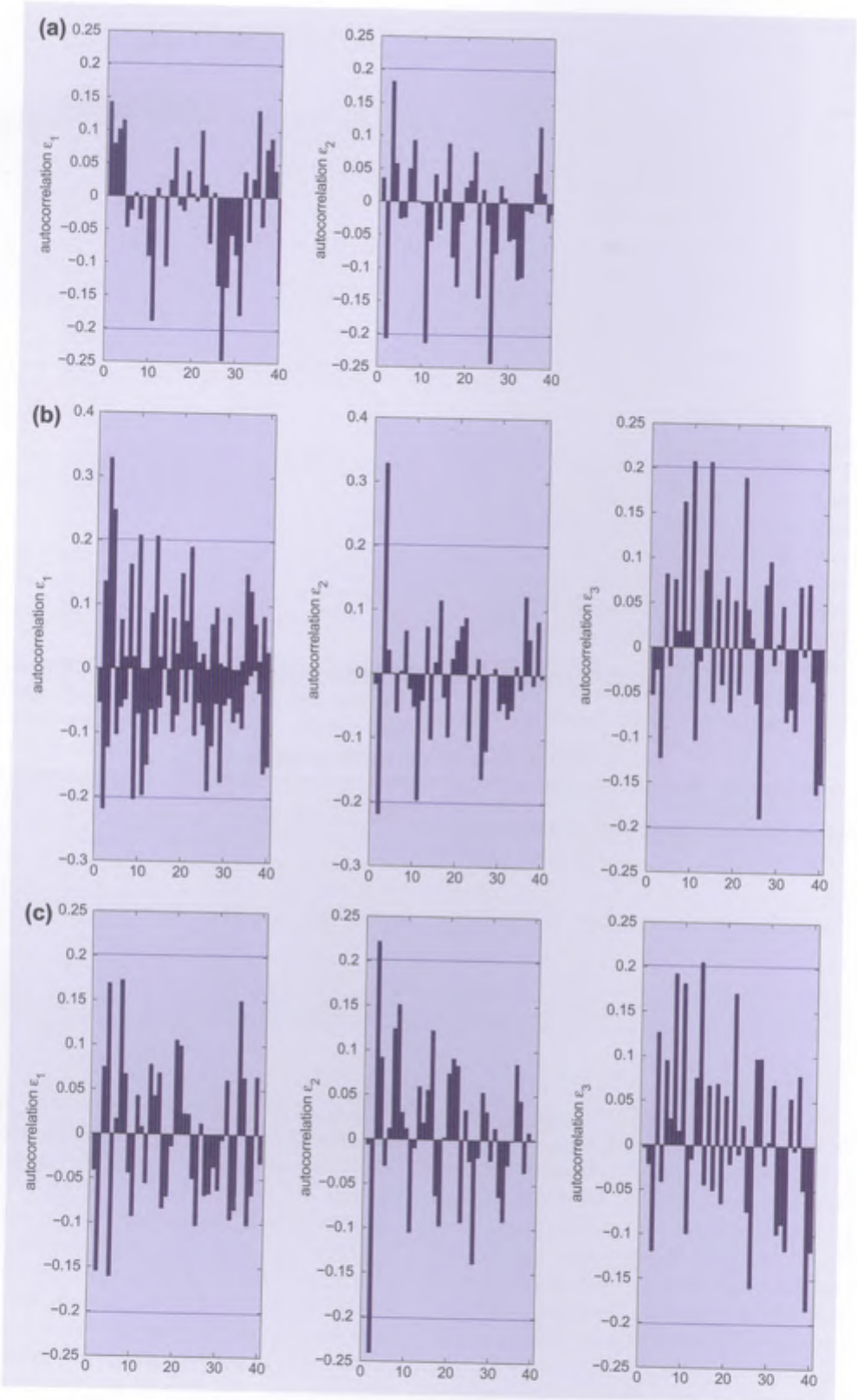


Figure 8. Factor loadings in the Heston model (top), in the generalized Heston model (middle) and in the semiparametric factor model (bottom).

consider two simple alternatives, the static Heston model and the random walk.

- *The static Heston model*: The Heston model (or, in general, stochastic volatility models) is often used when the dynamics of the volatility surfaces are important. For example, forward starting call spreads depend on the skew of the volatility surface at the start date. For such products, the model is normally calibrated at the valuation date and the price is computed without any parameter forecast to the start date. Thus the model is used

Table 3. Out-of-sample 1-week-ahead forecasting results $[E^{-2}]$.

Model	Mat. (months)	Mean	Std. dev.	MAE	MARE
Heston	1.5	0.14	0.27	0.24	16.8
	3	–0.04	0.21	0.16	8.4
	6	–0.02	0.21	0.15	7.2
	9	0.03	0.21	0.15	6.8
	12	0.07	0.21	0.16	7.0
	18	0.08	0.21	0.17	6.6
	24	0.04	0.20	0.16	5.7
Generalized Heston	1.5	0.19	0.23	0.24	17.1
	3	0.06	0.20	0.16	8.6
	6	0.13	0.20	0.18	9.0
	9	0.20	0.20	0.23	10.4
	12	0.22	0.20	0.25	10.7
	18	0.19	0.19	0.22	8.8
	24	0.10	0.19	0.17	6.2
Static Heston	1.5	0.17	0.25	0.25	17.0
	3	–0.01	0.21	0.15	8.1
	6	–0.00	0.20	0.15	6.9
	9	0.04	0.20	0.15	6.7
	12	0.07	0.20	0.16	6.8
	18	0.08	0.20	0.16	6.4
	24	0.03	0.20	0.15	5.5
Semiparam. model	1.5	0.32	0.26	0.35	24.4
	3	0.04	0.20	0.16	8.6
	6	0.03	0.20	0.15	7.3
	9	0.10	0.21	0.18	7.9
	12	0.06	0.20	0.16	6.8
	18	0.07	0.20	0.16	6.3
	24	0.04	0.19	0.15	5.4
Random walk	1.5	0.00	0.24	0.18	12.0
	3	0.01	0.21	0.15	8.2
	6	0.02	0.20	0.15	7.2
	9	0.02	0.20	0.15	6.7
	12	0.02	0.20	0.15	6.3
	18	0.02	0.19	0.15	5.7
	24	0.02	0.19	0.14	5.2

in a static way. For the Heston model, this method yields, for variance swaps that start at time τ, the price

$$\widehat{V_{t+\tau}(T)}/T = \frac{V_t(T+\tau) - V_t(\tau)}{T}, \tag{3}$$

where V_t denotes the variance curve at time t. This price can be interpreted as the forecasted variance swap price in the static Heston model. Considering this model, we can see if this static approach or the forecast of factor loadings gives better price forecasts for variance swaps, i.e. better dynamics.

- *The random walk*: This natural benchmark 'model' forecasts that the variance swap curves do not change:

$$\widehat{V_{t+\tau}(T)}/T = V_t(T)/T.$$

Duffie and Kan (1996) found that the analysed yield curve models seem to have problems in giving better forecasts than this model. Diebold and Li (2003) found that the reparametrized Nelson–Siegel approach seems to outperform the random walk for yield curves. The Nelson–Siegel model corresponds to the generalized Heston model.

The forecast errors at time t are defined as the difference between the variance swap curve observed at $t+\tau$ and the forecasted curve:

$$\widehat{V_{t+\tau}(T)}/T - V_{t+\tau}(T)/T,$$

for $T = 1.5$, 3, 6, 9, 12, 18 or 24 months. We examine a number of descriptive statistics of the errors, including the mean absolute error,

$$MAE \overset{\text{def}}{=} \frac{1}{n}\sum_t \left\| \widehat{V_{t+\tau}(T)}/T - V_{t+\tau}(T)/T \right\|,$$

and the corresponding relative error,

$$MARE \overset{\text{def}}{=} \frac{1}{n}\sum_t \left\| \frac{\widehat{V_{t+\tau}(T)}/T - V_{t+\tau}(T)/T}{V_{t+\tau}(T)/T} \right\|,$$

Table 4. Out-of-sample 6-months-ahead forecasting results $[E^{-2}]$.

Model	Mat. (months)	Mean	Std. dev	MAE	MARE
Heston	1.5	0.52	0.40	0.54	39.8
	3	0.37	0.44	0.44	26.0
	6	0.40	0.48	0.46	24.6
	9	0.45	0.51	0.50	24.9
	12	0.48	0.53	0.53	24.8
	18	0.48	0.56	0.53	23.1
	24	0.42	0.58	0.50	20.6
Generalized Heston	1.5	0.82	0.30	0.82	62.3
	3	0.79	0.27	0.79	46.1
	6	0.92	0.31	0.92	46.9
	9	0.98	0.33	0.98	47.0
	12	0.99	0.35	0.99	44.8
	18	0.87	0.38	0.87	36.9
	24	0.71	0.40	0.71	28.6
Static Heston	1.5	1.16	0.48	1.16	84.4
	3	0.90	0.53	0.90	52.5
	6	0.76	0.59	0.76	39.5
	9	0.68	0.63	0.69	34.0
	12	0.62	0.66	0.64	30.3
	18	0.49	0.71	0.57	25.1
	24	0.36	0.75	0.54	22.2
Semiparam. model	1.5	0.97	0.31	0.97	71.8
	3	0.74	0.31	0.74	43.2
	6	0.77	0.36	0.77	39.7
	9	0.84	0.39	0.84	40.5
	12	0.79	0.41	0.79	36.1
	18	0.74	0.44	0.74	31.3
	24	0.64	0.46	0.64	26.0
Random walk	1.5	0.18	0.35	0.33	24.0
	3	0.21	0.38	0.32	18.9
	6	0.21	0.47	0.36	18.9
	9	0.20	0.52	0.38	18.9
	12	0.18	0.55	0.40	18.8
	18	0.14	0.62	0.44	19.1
	24	0.11	0.67	0.48	19.2

where the index t runs over all forecast days and n is the number of forecast days. All these errors are measured in variance.

The results of the 1-week forecasts are presented in table 3. All models except the generalized Heston model show a similar forecasting performance over this short time horizon with an average absolute error of around 0.15%. This seems reasonably small because the average variance swap curve has a level around 2.7%. We see that the random walk model has the smallest errors of all models. The static Heston model and the Heston model with parameter forecasts have similar errors. The three-factor models have the worst performance, and the Nelson–Siegel framework leads to larger errors for long maturities. Moreover, the errors tend to be larger for short maturities. Hence, the short ends of the variance swap curves are harder to forecast. This corresponds to the in-sample fit problems at the short end.

Forecasts for 1 month and 3 months show qualitatively similar results. The errors in the dynamic Heston model become smaller than in the static Heston model. Both Heston models produce better forecasts than the three-factor model. The semiparametric factor model, as before, outperforms the Nelson–Siegel parametrization.

In table 4, we present the forecast results for half a year ahead. For these long periods the dynamic Heston model gives forecasts of similar quality as the random walk. The static Heston model leads to rather large forecast errors and the semiparametric model produces similar errors as the Nelson–Siegel approach.

Variance swap curve forecasts are hard to forecast over short time horizons with models because the models already exhibit in-sample fit problems. For longer time horizons, the dynamic Heston model performs well and particularly better than the static Heston model.

4. Conclusion

We analyse the modeling and forecasting of variance swap curves in a Heston model, a three-factor Nelson–Siegel parametrization and a three-factor semiparametric model. The in-sample fit gives good results for long maturities, but

all models have problems in fitting the short end of the variance swap curves. The Nelson–Siegel parametrization naturally outperforms the Heston model, and the flexible semiparametric factor model leads to the best fit for long maturities, but also has problems with short maturities. In the forecasting analysis, all models give similar results for short forecasting horizons. But for longer time horizons the Heston model clearly leads to the smallest forecasting errors. In option pricing, model parameters are generally not forecasted. When we consider the Heston model in such a static way its performance is worse than the forecasted Heston model. The semiparametric approach seems problematic for forecasting.

Combining the in-sample and out-of-sample results the Heston model gives the best overall performance due to its good forecast results for long time horizons. This can be interpreted in a way that convexity is either difficult to model or not significant for long time horizons. In yield curve modelling, convexity seems more important, and the Nelson–Siegel model appears to be superior.

In addition, we show that the static Heston model can lead to bad variance swap curve dynamics. Hence the usual way of pricing forward starting products could be problematic in standard stochastic volatility models. Recent stochastic volatility approaches such as, for example, that of Bergomi (2005) appear necessary in order to capture the variance swap dynamics better.

References

Amengual, D., The term structure of variance risk premia. Technical report, Princeton University, March 2009.

Bergomi, L., Smile dynamics. *RISK*, 2004, **17**(9), 117–123.

Bergomi, L., Smile dynamics II. *RISK*, 2005, **18**(10), 67–73.

Bühler, H., Consistent variance curve models. *Finance Stochast.*, 2006, **10**(2), 178–203.

Carr, P. and Madan, D., Towards a theory of volatility trading. Reprinted in *Option Pricing, Interest Rates, and Risk Management*, edited by Musiella, Jouini, Cvitanic, pp. 417–427, 1998 (University Press).

Cont, R. and da Fonseca, J., Dynamics of implied volatility surfaces. *Quant. Finance*, 2002, **2**(1), 45–60.

Demeterfi, K., Derman, E., Kamal, M. and Zou, J., More than you ever wanted to know about volatility swaps. Quantitative Strategies Research Notes, Goldman Sachs, 1999.

Diebold, F.X. and Li, C., Forecasting the term structure of government bond yields. NBER Working Papers 10048, National Bureau of Economic Research, October 2003.

Duffie, D. and Kan, R., A yield factor model of interest rates. *Math. Finance*, 1996, **6**(4), 379–406.

Fan, J. and Gijbels, I., *Local Polynomial Modelling and Its Applications*, 1996 (Chapman and Hall: London).

Fengler, M., *Semiparametric Modeling of Implied Volatility*, 2005 (Springer: Berlin).

Härdle, W., Müller, M., Sperlich, S. and Werwatz, A., *Nonparametric and Semiparametric Models*, 2004 (Springer: Heidelberg).

Heston, S., A closed-form solution for options with stochastic volatility with applications to bond and currency options. *Rev. Financ. Stud.*, 1993, **6**(2), 327–343.

Nelson, C. and Siegel, A., Parsimonious modeling of the yield curve. *J. Business*, 1987, **60**, 473–489.