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

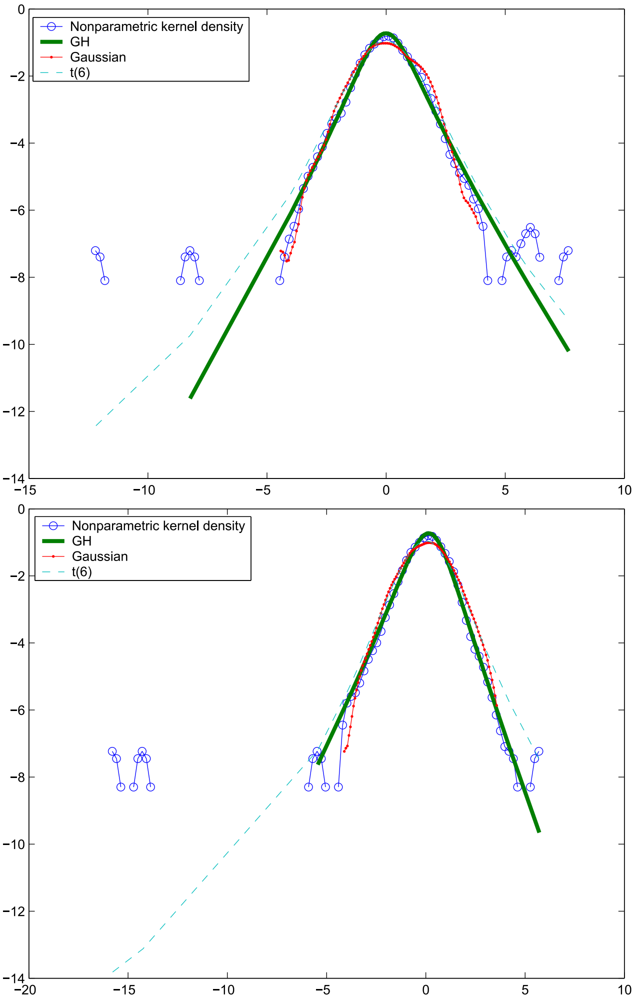


Fig. 1: Density comparisons of the standardized returns in log scale based on the Allianz stock (top) and the DAX portfolio (bottom) with static weights $b(t) = \text{unit}(1/20)$. Time interval: 1988/01/04 - 1996/12/30. The nonparametric kernel density is considered as benchmark. The GH distributional 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 $\tilde{}$ 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:

$$\begin{aligned} I(y) &= \sum_{j=1}^{d} H(y_j) - H(y) \qquad (2) \\ \text{where } H(y_j) &= -\int f_{y_j}(u) \log f_{y_j}(u) du \end{aligned}$$

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).

$$\hat{w}_j = \text{argmin} H(y_j) = \text{argmax} J(w_j, y_j).$$

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:

$$\begin{aligned} \hat{w}_j &= \text{argmin} H(y_j) = \text{argmax} J(y_j) \\ J(y_j) &\approx \mathit{const.}\{\mathsf{E}[G(y)] - \mathsf{E}[G(y_0)]\}^2 \\ &= \mathit{const.}\{\mathsf{E}[G(w_j^\top x)] - \mathsf{E}[G(y_0)]\}^2 \\ G(y_j) &= \log\cosh(y_j) \end{aligned} \tag{3}$$

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:

$$\begin{aligned} D_y(t) &= \operatorname{diag}\{\sigma^2_{y_1}(t), \cdots, \sigma^2_{y_d}(t)\} \\ \Sigma_x(t) &= W^{-1} D_y(t) W^{-1\top} \end{aligned} \tag{4}$$

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:

$$\begin{aligned} y_p(t) &= \operatorname{sign}\{y(t)\}|y(t)|^p \\ \theta(t) &= Var\{y_p(t)|\mathcal{F}_{t-1}\} = \mathsf{E}\{y_p^2(t)|\mathcal{F}_{t-1}\} = \mathsf{E}\{|y(t)|^{2p}|\mathcal{F}_{t-1}\} \\ &= \sigma^{2p}(t)\,\mathsf{E}\,|\varepsilon(t)|^{2p} = \sigma^{2p}(t)C_p \end{aligned} \tag{6}$$

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 $\mathfrak{z}_k$:

$$\gamma_k = K_{ag}\left\{ L\left(\eta_k, \tilde{\theta}^{(k)}(t), \hat{\theta}^{(k-1)}(t)\right) / \mathfrak{z}_k \right\}$$

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 $\sigma^* \overset{\text{def}}{=} 1$. 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 $|\beta| \le \alpha$. 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

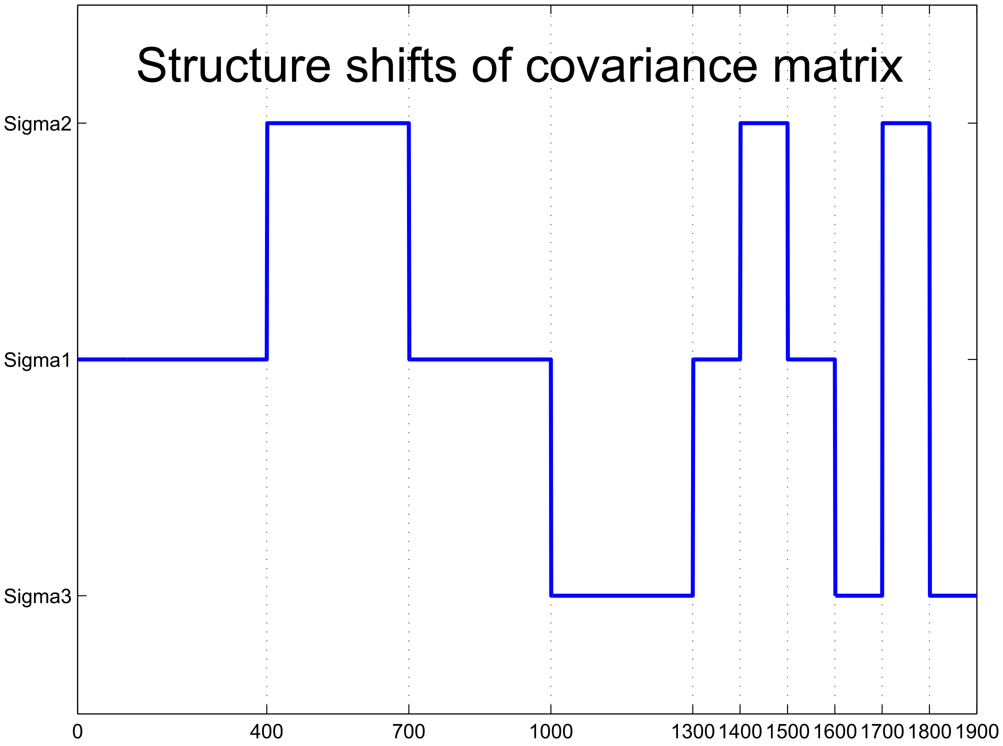


Fig. 2: Structure shifts of the generated covariance through time. Notice that there are shifts among matrices not up-and-down movements.

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:

$$\begin{aligned} \mathsf{E}(\varepsilon_x) &= \mu + \frac{\beta\delta}{\sqrt{\alpha^2-\beta^2}} = 0 \\ \operatorname{Var}(\varepsilon_x) &= \frac{\delta}{\sqrt{\alpha^2-\beta^2}} + \frac{\beta^2}{\delta^3\sqrt{\alpha^2-\beta^2}} = \frac{\delta}{\alpha} = 1 \end{aligned}$$

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:

$$\Sigma_x(t) = D_x(t)R_x(t)D_x(t)^\top.$$

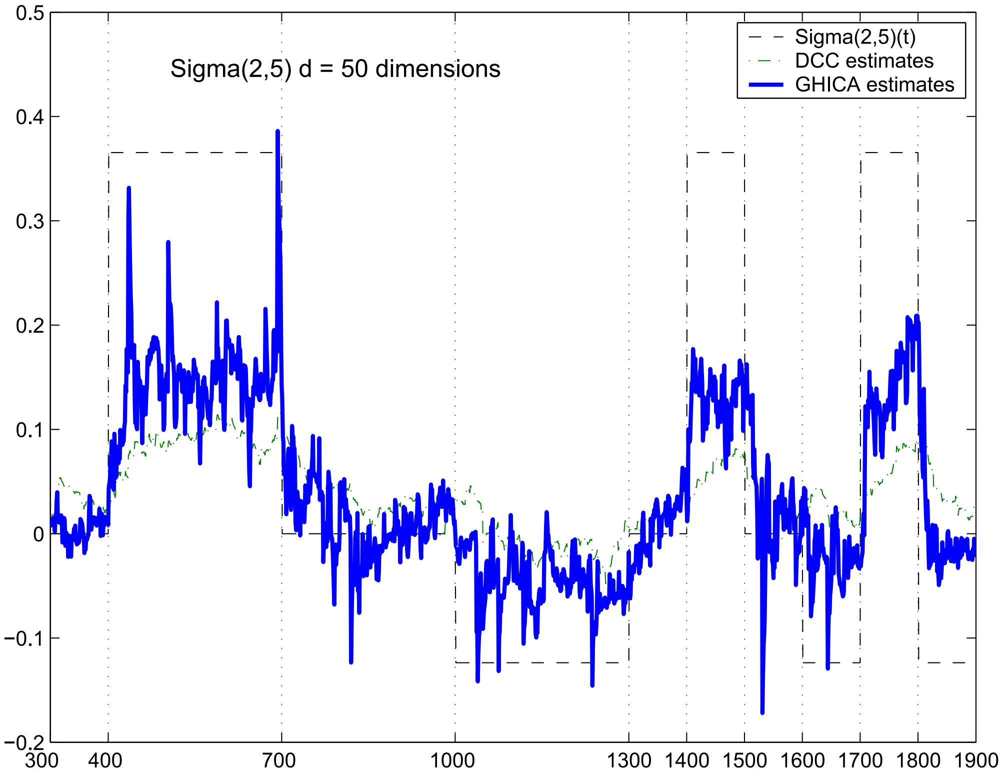


Fig. 3: Realized estimates of $\Sigma(2,5)$ based on the GHICA and DCC methods. The generated data consists of 50 NIG distributed components.

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.

$$\text{RAE}(i,j) \quad = \quad \frac{\sum_{t=301}^{T} |\hat{\Sigma}_{(i,j)}^{\text{GHICA}}(t) - \Sigma_{(i,j)}(t)|}{\sum_{t=301}^{T} |\hat{\Sigma}_{(i,j)}^{\text{DCC}}(t) - \Sigma_{(i,j)}(t)|}$$

If $\text{RAE}(i,j) \leq 1$, it means that the GHICA method reaches higher accuracy in the estima-

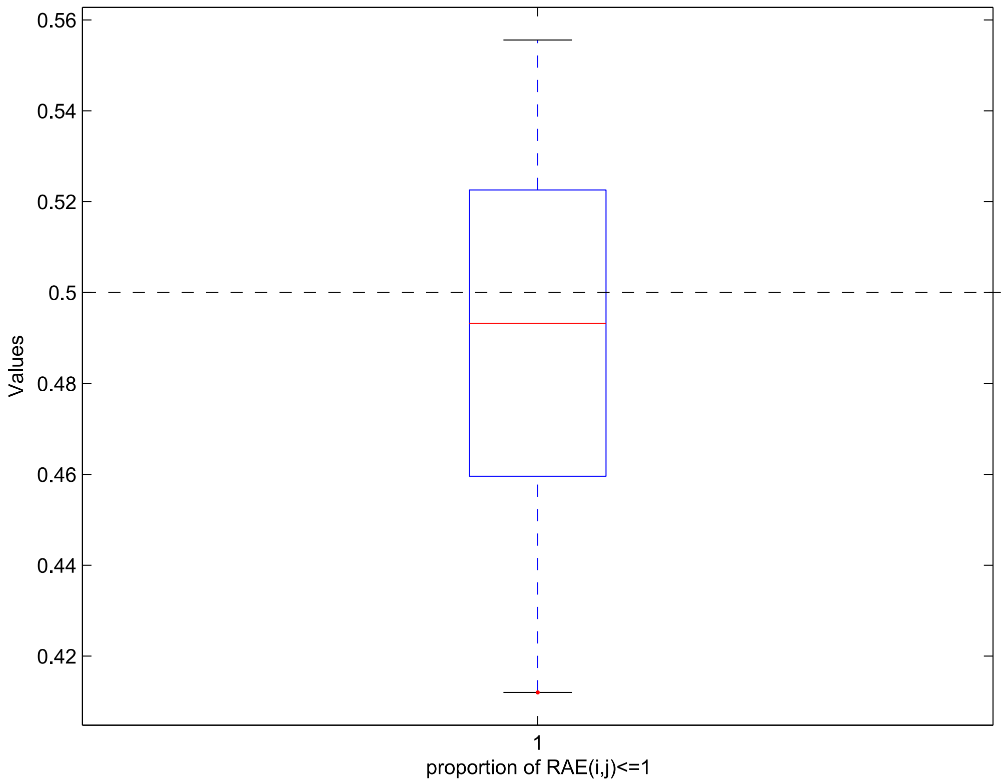


Fig. 4: Boxplot of the proportion $\frac{\sum_i \sum_j \mathbf{1}(\mathrm{RAE}(i,j) \leq 1)}{d \times d}$ for $i, j = 1, \cdots, d$. Here $d = 50$ and the proportions on the base of 100 simulations are considered.

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. $\frac{\sum_i \sum_j \mathbf{1}(\mathrm{RAE}(i,j) \leq 1)}{d \times d}$ 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

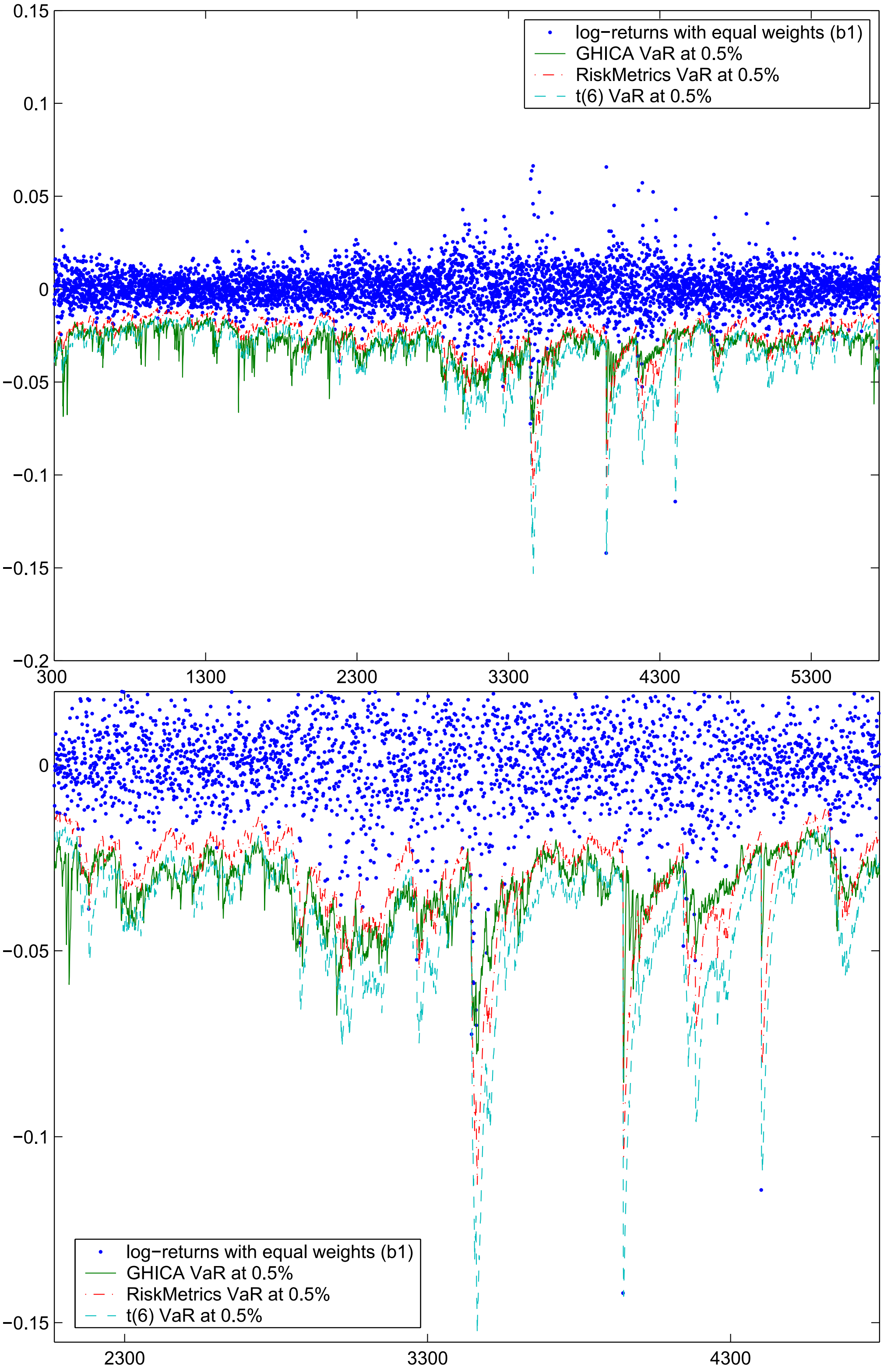


Fig. 5: One day log-returns of the DAX portfolio with the static trading strategy $b(t) = b^{(1)}$. The VaRs are from 1975/03/17 to 1996/12/30 at pr $= 0.5\%$ w.r.t. three methods, the GHICA, the RiskMetrics and the $t(6)$. Part of the VaR time plot is enlarged and displayed on the bottom.

			GHICA			RiskMetrics $\mathrm{N}(\mu, \sigma^2)$			Exponential smoothing $t(6)$		
h	$b(t)$	pr	$\hat{\text{pr}}$	RC	ES	$\hat{\text{pr}}$	RC	ES	$\hat{\text{pr}}$	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	$\hat{\text{pr}}$	RC	ES	$\hat{\text{pr}}$	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

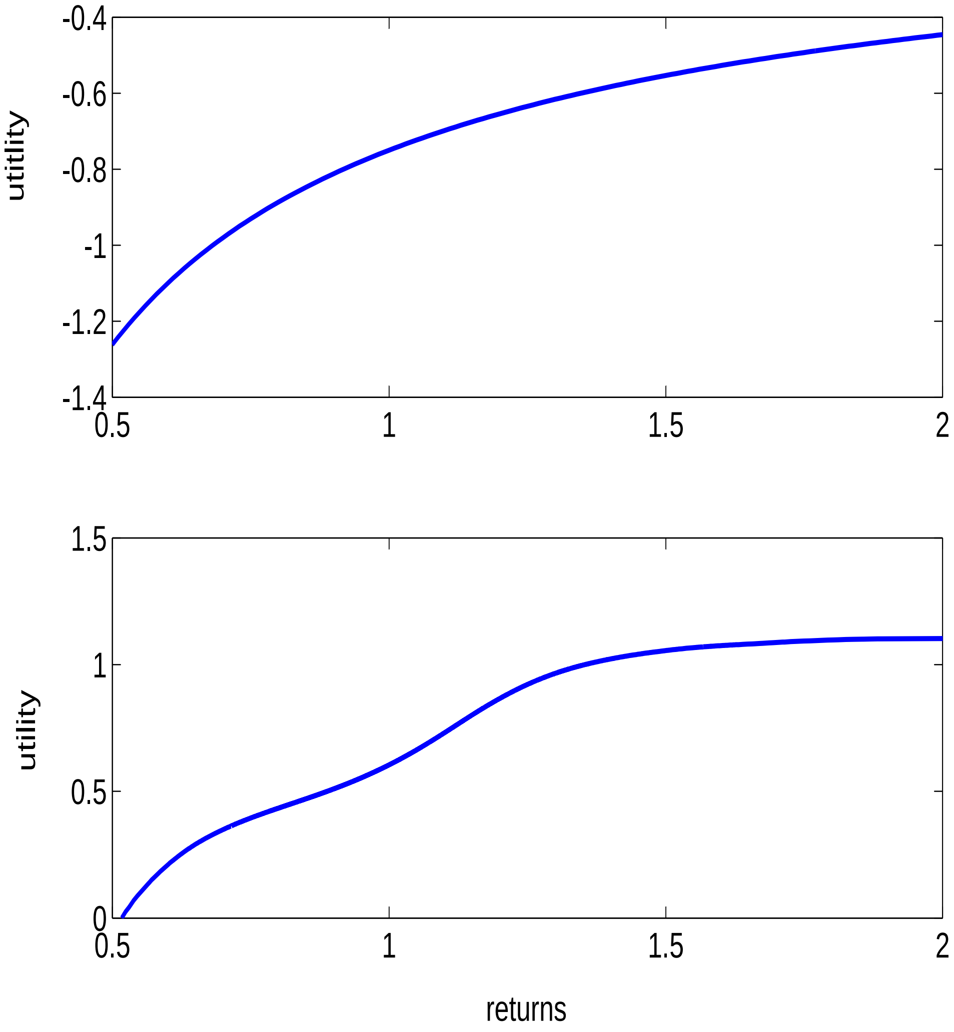


Figure 1: up: Utility function in the Black Scholes model for $T = 0.5$ years ahead and drift $\mu = 0.1$, volatility $\sigma = 0.2$ and interest rate $r = 0.03$. down: Market utility function on 06/30/2000 for $T = 0.5$ years ahead.

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 kernel 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 because 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

$$\begin{aligned} \phi_T(z) &= \exp\{\frac{-(z^2+\mathbf{i}z)V_0}{\gamma(z)\coth\frac{\gamma(z)T}{2}+\xi-\mathbf{i}\rho\theta z}\} \\ &\times \frac{\exp\{\frac{\xi\eta T(\xi-\mathbf{i}\rho\theta z)}{\theta^2}+\mathbf{i}zTr+\mathbf{i}z\log(S_0)\}}{(\cosh\frac{\gamma(z)T}{2}+\frac{\xi-\mathbf{i}\rho\theta z}{\gamma(z)}\sinh\frac{\gamma(z)T}{2})^{\frac{2\xi\eta}{\theta^2}}} \end{aligned} \tag{4}$$

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

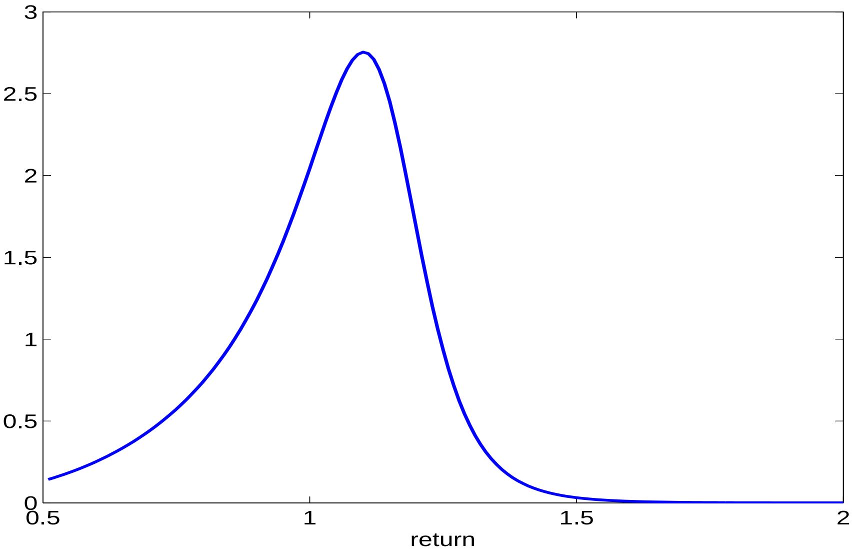


Figure 2: Risk neutral density on 24/03/2000 half a year ahead.

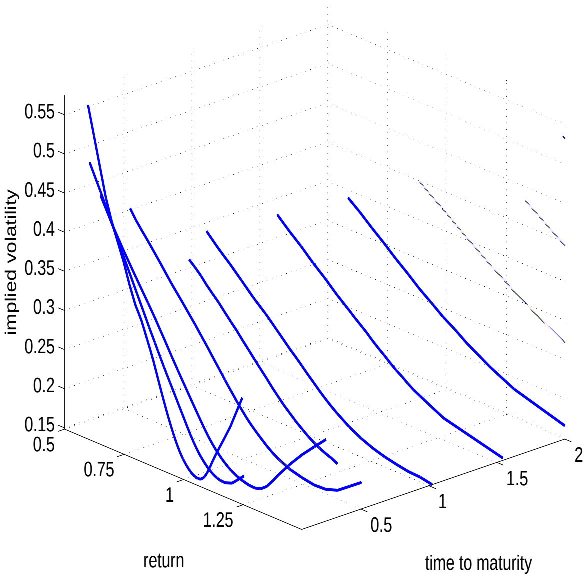


Figure 3: Implied volatility surface on 24/03/00.

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

$$R_i = \mu - \frac{1}{2} V_i + \sqrt{V_i} Z_i$$

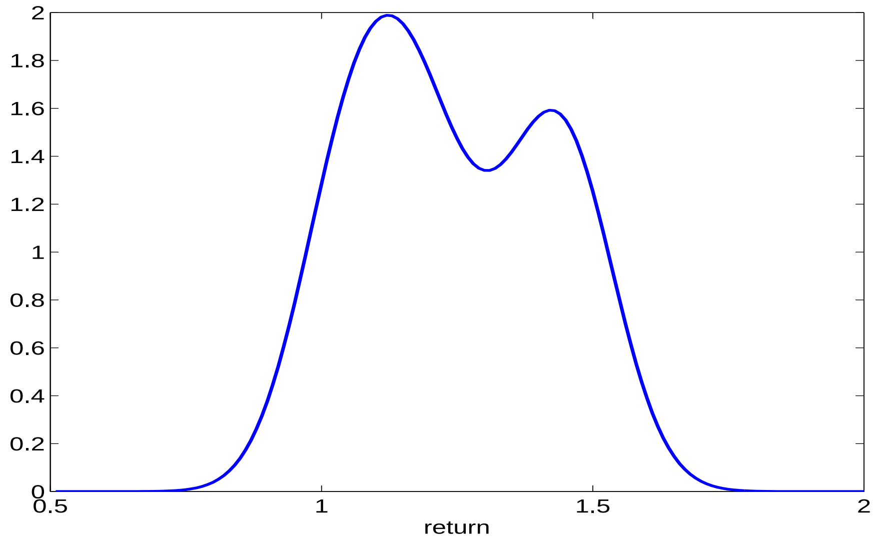


Figure 4: Historical density on 24/03/2000 half a year ahead.

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.

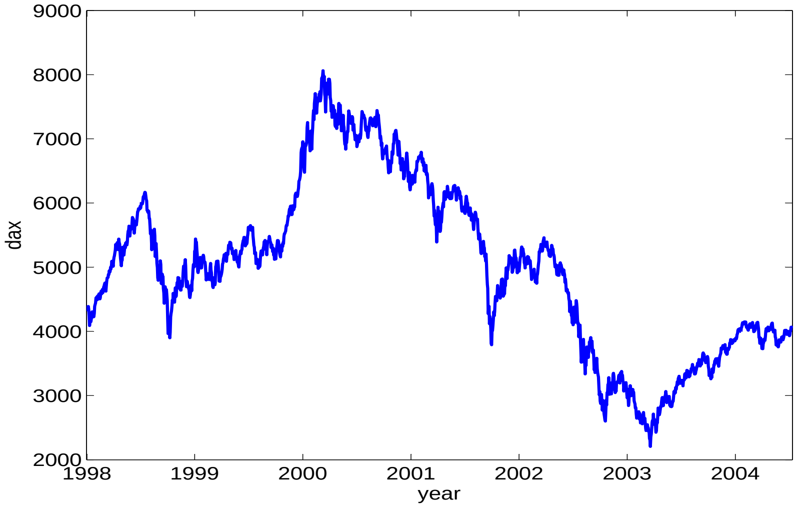


Figure 5: DAX, 1998 - 2004.

	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 construction 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

$$RRA(R) = -R\frac{q'(R)p(R) - q(R)p'(R)}{p^2(R)} / \frac{q(R)}{p(R)} = R\left(\frac{p'(R)}{p(R)} - \frac{q'(R)}{q(R)}\right).$$

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

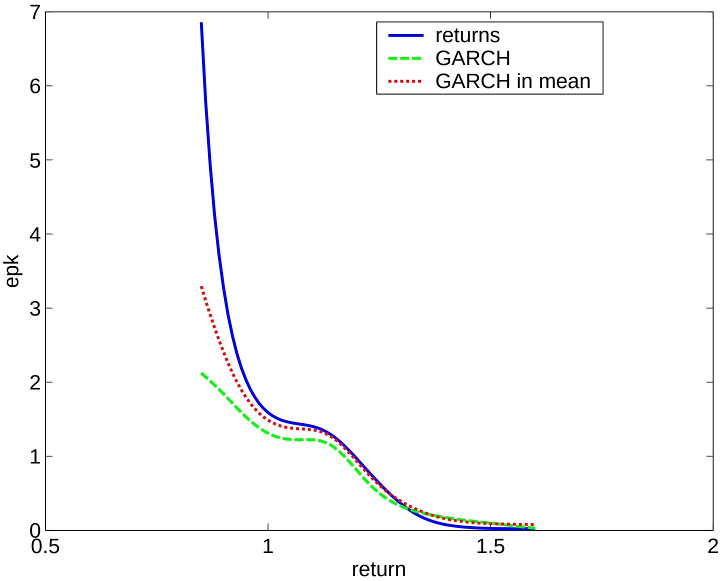


Figure 6: Empirical pricing kernel on 24/03/2000 (bullish market).

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

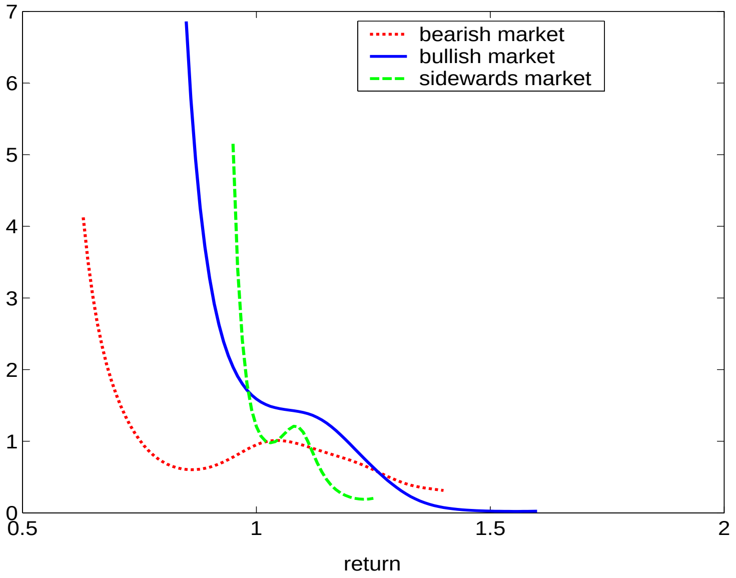


Figure 7: Empirical pricing kernel on 24/03/2000 (bullish), 30/07/2002 (bearish) and 30/06/2004 (unsettled or sidewards market).

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

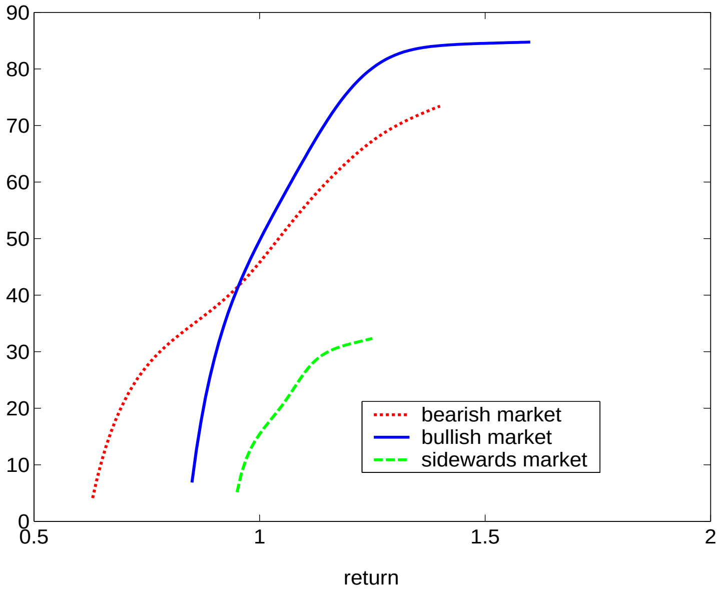


Figure 8: Market utility functions on 24/03/2000 (bullish), 30/07/2002 (bearish) and 30/06/2004 (unsettled or sidewards market).

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

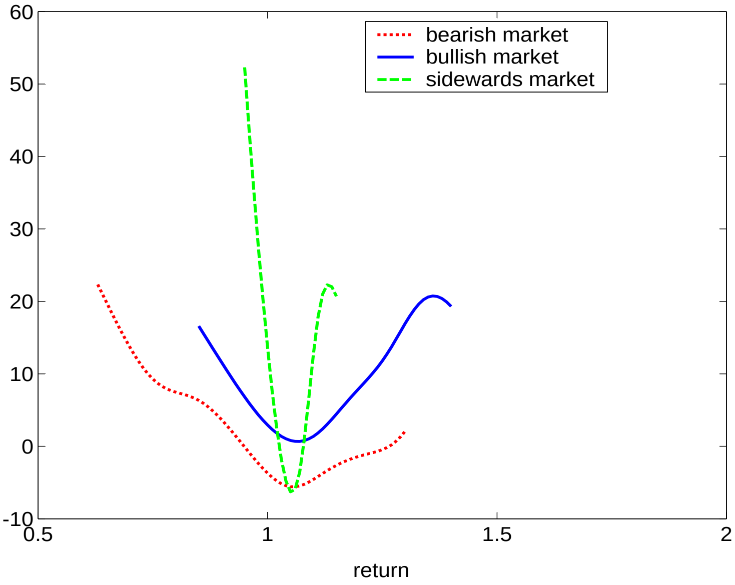


Figure 9: Relative risk aversions on 24/03/2000 (bullish), 30/07/2002 (bearish) and 30/06/2004 (unsettled or sidewards market).

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

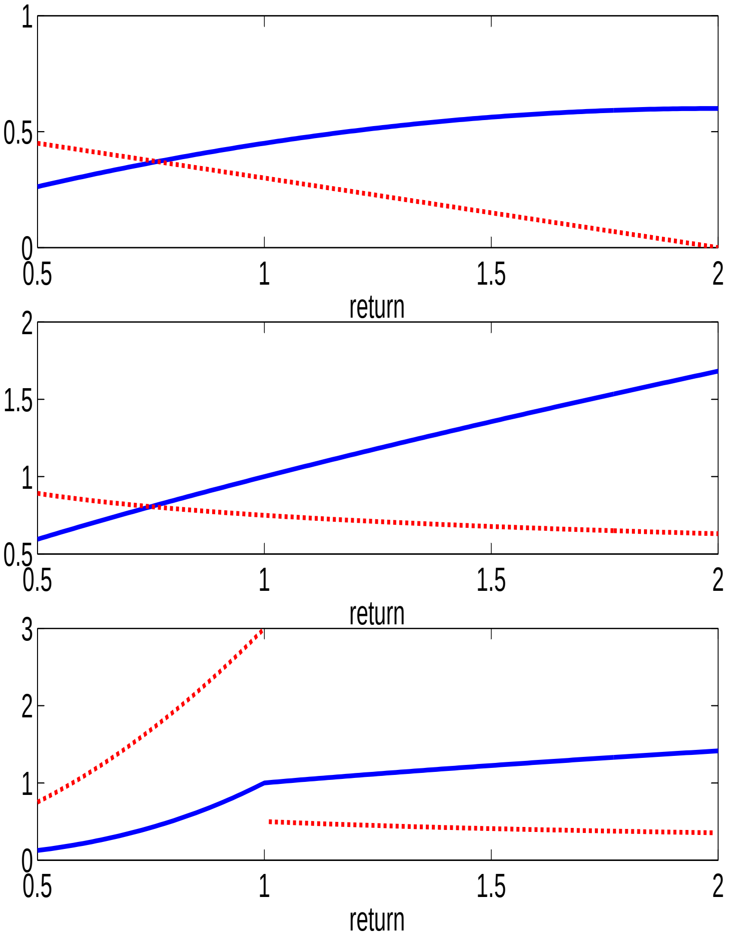


Figure 10: Common utility functions (solid) and their pricing kernels (dotted) (upper: quadratic, middle: power, lower panel: Kahneman and Tversky utility function).

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:

$$U_i(R) = \begin{cases} \max\left\{U(R, \theta_1, c_1); U(R, \theta_2, c_{2,i})\right\}, & \text{if} \quad R > c_1 \\ -\infty, \quad \text{if} \quad R \leq c_1 \end{cases}$$

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:

$$U_i(R) = \begin{cases} \max\left\{U_{bear}(R); U_{bull}(R, c_i)\right\}, & \text{if} \quad \mathrm{R} > \mathrm{c}_1; \\ -\infty, \quad \text{if} \quad \mathrm{R} \leq \mathrm{c}_1. \end{cases} \tag{5}$$

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}

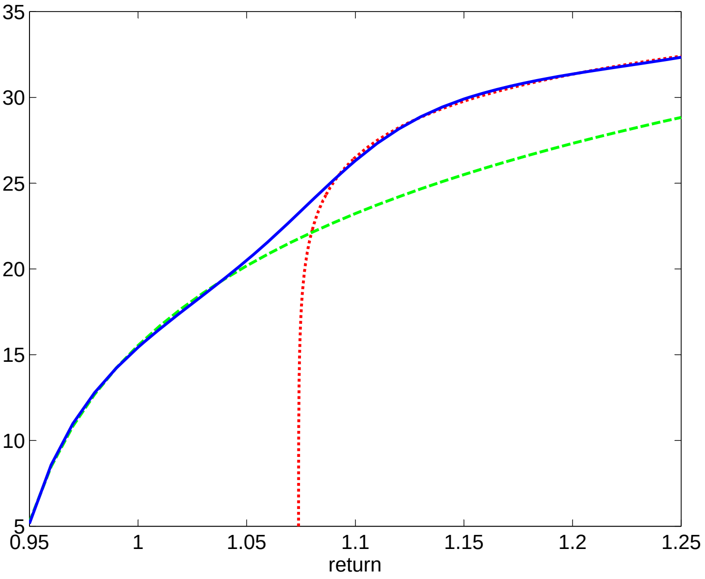


Figure 11: Market utility function (solid) with bearish (dashed) and bullish (dotted) part of an individual utility function 5 estimated in the unsettled market of 30/06/2004.

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 $x_1 \dots, x_n$ 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.

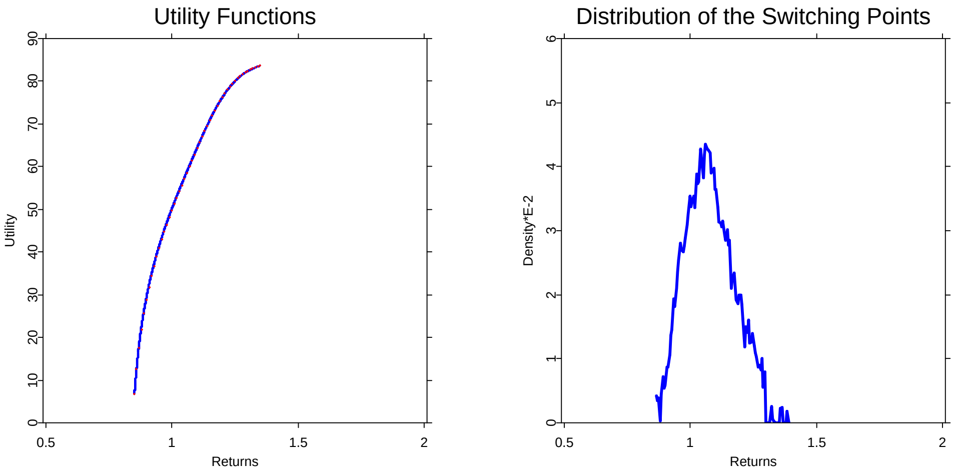


Figure 12: Left panel: the market utility function (red) and the fitted utility function (blue). Right panel: the distribution of the reference points. 24 March 2000, a bullish market.

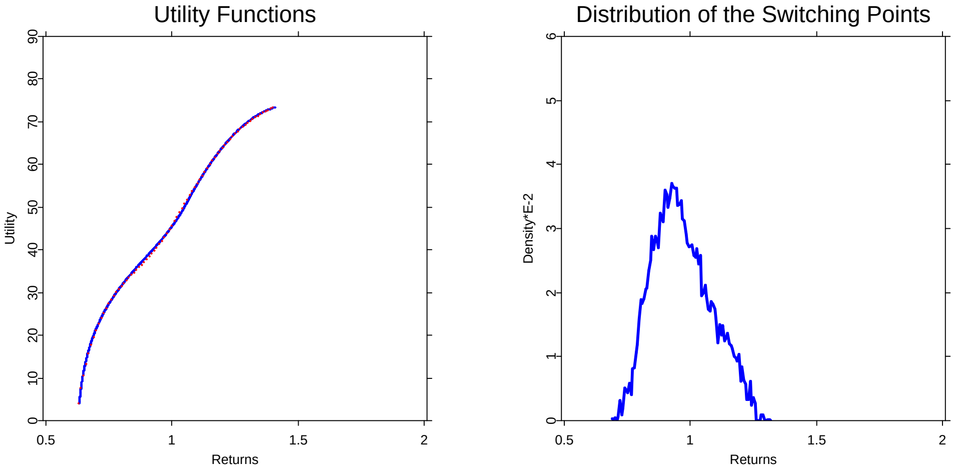


Figure 13: Left panel: the market utility function (red) and the fitted utility function (blue). Right panel: the distribution of the reference points. 30 July 2002, a bearish market.

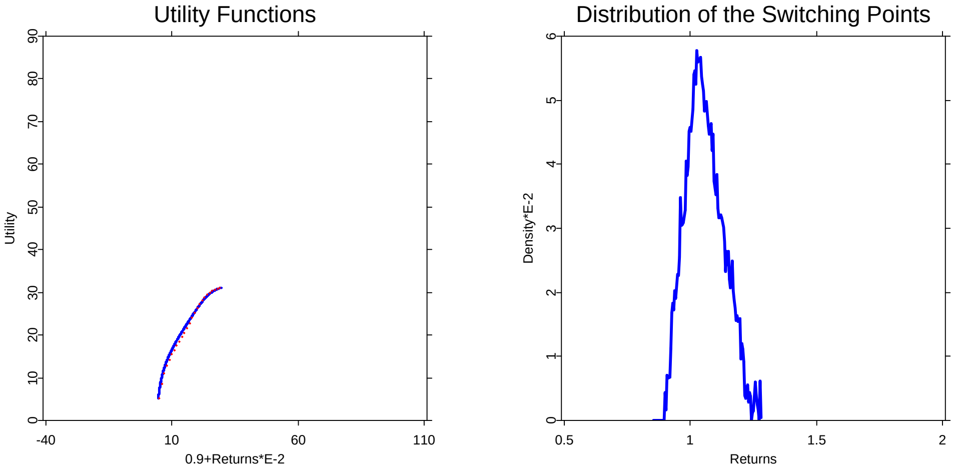


Figure 14: Left panel: the market utility function (red) and the fitted utility function (blue). Right panel: the distribution of the reference points. 30 June 2004, an unsettled market.

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:

$$F(x_1 \ldots, x_d) \quad = \quad C\{F_1(x_1), \ldots, F_d(x_d)\}.$$

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 $X_1, \dots, X_d$ are stochastically independent from Theorem 1, then the structure of such a relationship is given by the product copula

$$\Pi(u_1, \dots, u_d) = \prod_{j=1}^{d} u_j. \tag{2}$$

Copulae are bounded, this means that for all $u = (u_1, \dots, u_d)^\top \in [0,1]^d$:

$$W(u_1, \dots, u_d) \le C(u_1, \dots, u_d) \le M(u_1, \dots, u_d)$$

where

$$M(u_1, \dots, u_d) = \min(u_1, \dots, u_d)$$

is called the *Fréchet-Hoeffding lower bound* and

$$W(u_1, \dots, u_d) = \max\left(\sum_{i=1}^{d} u_i - d + 1, 0\right)$$

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 $X = (X, \dots, X_d)^\top \sim \mathrm{N}_d(\mu, \Sigma)$ with $X_j \sim \mathrm{N}(\mu_j, \sigma_j)$ for $j = 1, \dots, d$. A copula C exists:

$$F(x_1, \dots, x_d) = C\{F_1(x_1), \dots, F_d(x_d)\},$$

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 $Y = (Y_1, \dots, Y_d)^\top \sim \mathrm{N}_d(0, \Psi)$ where Ψ is the correlation matrix associated with Σ. A copula C_Ψ^{Ga}, called *Gaussian copula* exists as follows:

$$F_Y(y_1, \dots, y_d) = C_\Psi^{Ga}\{\Phi(y_1), \dots, \Phi(y_d)\}. \tag{3}$$

An explicit expression for the Gaussian copula is obtained by rewriting (3) with $u_j = \Phi(y_j)$:

$$\begin{aligned} C_\Psi^{Ga}(u_1, \dots, u_d) &= F_Y\{\Phi^{-1}(u_1), \dots, \Phi^{-1}(u_d)\} \\ &= \int_{-\infty}^{\Phi^{-1}(u_1)} \dots \int_{-\infty}^{\Phi^{-1}(u_d)} (2\pi)^{-\frac{d}{2}} \mid \Psi \mid^{-\frac{1}{2}} \exp{(-\frac{1}{2} r^\top \Psi^{-1} r)} dr_1 \dots dr_d. \end{aligned}$$

The density of the Gaussian copula is given by

$$c_\Psi^{Ga}(u_1, \dots, u_d) = \mid \Psi \mid^{-\frac{1}{2}} \exp\left\{ -\frac{1}{2} \zeta^\top (\Psi^{-1} - I_d) \zeta \right\}. \tag{4}$$

Student's t-Copula

The t-copula, containing the dependence structure from the multivariate t-distribution, may be obtained in a similar way.

Let $X = (X_1, \dots, X_d)^\top \sim t_d(\nu, \mu, \Sigma)$ and $Y = (Y_1, \dots, Y_d)^\top \sim t_d(\nu, 0, \Psi)$ 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 $u = (u_1, \dots, u_d)^\top \in [0, 1]^d$, the *Student's t-copula* is given by

$$C_{\nu,\Psi}^t(u_1, \dots, u_d) = t_{\nu,\Psi}\{t_\nu^{-1}(u_1), \dots, t_\nu^{-1}(u_d)\}$$

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

$$\begin{aligned} c_{\nu,\Psi}^t(u_1, \dots, u_d) &= \frac{t_{\nu,\Psi}\{t_\nu^{-1}(u_1), \dots, t_\nu^{-1}(u_d)\}}{\prod_{j=1}^d t_{\nu,\Psi}\{t_\nu^{-1}(u_j)\}}. \\ &= \mid \Psi \mid^{-\frac{1}{2}} \frac{\Gamma(\frac{\nu+d}{2}) \left\{\Gamma(\frac{\nu}{2})\right\}^{d-1} \left(1 + \frac{1}{\nu} \zeta^\top \Psi^{-1} \zeta\right)^{-\frac{\nu+d}{2}}}{\left\{\Gamma(\frac{\nu+1}{2})\right\}^d \prod_{j=1}^d \left(1 + \frac{1}{\nu} \zeta_j^2\right)^{-\frac{\nu+1}{2}}}. \end{aligned}$$

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.

$$\mathbb{L} = \{\phi : [0;\infty) \to [0,1] \,|\, \phi(0) = 1,\, \phi(\infty) = 0;\, (-1)^j \phi^{(j)} \geq 0;\, j = 1,\dots,\infty\}.$$

The function $C : [0,1]^d \to [0,1]$ defined as

$$C(u_1,\dots,u_d) = \phi\{\phi^{-1}(u_1) + \dots + \phi^{-1}(u_d)\}, \quad u_1,\dots,u_d \in [0,1]$$

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 $C(u_1,\dots,u_d)$ 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

$$\begin{aligned} \phi(x,\theta) &= \theta^{-1}\log\{1-(1-e^{-\theta})e^{-x}\}, \quad 0 \leq \theta < \infty,\ x \in [0,\infty). \\ C_\theta(u_1,\dots,u_d) &= -\frac{1}{\theta}\log\left[1 + \frac{\prod_{j=1}^{d}\{\exp(-\theta u_j)-1\}}{\{\exp(-\theta)-1\}^{d-1}}\right]. \end{aligned}$$

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

$$\begin{aligned} \phi(x,\theta) &= \exp\{-x^{1/\theta}\}, \quad 1 \leq \theta < \infty,\ x \in [0,\infty) \\ C_\theta(u_1,\dots,u_d) &= \exp\left[-\left\{\sum_{j=1}^{d}(-\log u_j)^\theta\right\}^{\theta^{-1}}\right]. \end{aligned}$$

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

$$\begin{aligned}
&\mathbb{L}^* = \{\omega : [0;\infty) \to [0,\infty)\,|\,\omega(0) = 0,\\
&\qquad \omega(\infty) = \infty;\ (-1)^{j-1}\omega^{(j)} \geq 0;\ j = 1,\ldots,\infty\}.
\end{aligned}$$

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

$$c^j_{j-1}(u_1, \ldots, u_j) = \frac{\partial^{j-1} C_j(u_1, \ldots, u_j)}{\partial u_1, \ldots, \partial u_{j-1}}$$

the probability $P(U_j \leq u_j, U_1 = u_1, \ldots, U_{j-1} = u_{j-1})$ can be written as

$$\lim_{\Delta u_1,\ldots,\Delta u_{j-1}\to 0} \frac{C_j(u_1 + \Delta u_1, \ldots, u_{j-1} + \Delta u_{j-1}, u_j) - C_j(u_1, \ldots, u_j)}{\Delta u_1, \ldots, \Delta u_{j-1}}$$
$$= c^j_{j-1}(u_1, \ldots, u_j).$$

Thus, the conditional distribution $\Lambda(u_j)$ (given fixed $u_1, \dots, u_{j-1}$) is a function of the ratio of derivatives:

$$\begin{aligned}\Lambda(u_j) &= P(U_j \le u_j \mid U_1 = u_1, \dots, U_{j-1} = u_{j-1}) \\ &= \frac{c_{j-1}^{j}(u_1, \dots, u_j)}{c_{j-1}^{j-1}(u_1, \dots, u_{j-1})}.\end{aligned}$$

The generation of d pseudo random numbers with given marginal distributions F_j, $j = 1, \dots, d$ and dependence structure given by the copula C follows the steps:

1. generate iid $v_1, \dots, v_d \sim U[0,1]$.
2. for $j = 1, \dots, d$ 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.

$$\phi(s) = \int_0^\infty \exp\{-sw\}\, dM(w),\ s \ge 0.$$

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 $F_1, \dots, F_d$, we obtain

$$C(u_1, \dots, u_d) = \int_0^\infty \prod_{i=1}^d G_i^\alpha\, dM(\alpha) = \phi\left[\sum_{i=1}^d \phi^{-1}\{F_i(u_i)\}\right]$$

which is a multivariate distribution function. By replacing the product of univariate distributions G_i for $i = 1, \dots, d$ with an arbitrary copula function R we get:

$$C(u_1, \dots, u_d) = \int_0^\infty \dots \int_0^\infty R(G_1^\alpha, \dots, G_d^\alpha)\, dM(\alpha).$$

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 $(v_1, \dots, v_d)$ 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 $\alpha = (\delta_1, \dots, \delta_d, \theta)^\top$ 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 $\tilde{\alpha}_{FML} = (\tilde{\delta}_1, \dots, \tilde{\delta}_d, \tilde{\theta})^\top$ solve

$$(\partial\ell/\partial\delta_1, \dots, \partial\ell/\partial\delta_d, \partial\ell/\partial\theta) = 0.$$

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 $j = 1, \dots, d$ 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

   $$\ell(\theta, \hat{\delta}_1, \dots, \hat{\delta}_d) = \sum_{t=1}^{T} \log c\{F_1(x_{1,t}; \hat{\delta}_1), \dots, F_d(x_{d,t}; \hat{\delta}_d); \theta\}$$

   is maximised over θ to get the dependence parameter estimate $\hat{\theta}$.

The estimates $\hat{\alpha}_{IFM} = (\hat{\delta}_1, \dots, \hat{\delta}_d, \hat{\theta})^\top$ solve

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

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 $F = C(F_1, \dots, F_d)$. Thus (7) is replaced by:

$$\Sigma_{d,t}^{-1/2}(X_t - \mu_t) \sim C(F_1, \dots, F_d) \tag{7}$$

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 $i = 1, \dots, d$ 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 $i = 1, \dots, d$. 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 $i = 1, \dots, d$ and $u_1, \dots, u_d \sim C$. The copula C defines the dependency structure between the asset returns. This implies that

$$F_{R_p}(\xi) = \mathrm{P}(R_p \leq \xi) = \int_{\mathcal{U}} c(u_1, \dots, u_d) du_1 \dots du_d, \tag{10}$$

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 $\widehat{\Sigma} = \{\widehat{\Sigma}_{ij}\}_{i,j=1,\dots,d}$, mean vector $\widehat{\mu} = \{\widehat{\mu}_i\}_{i=1,\dots,d}$ 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

$$F_{R_p}(\xi) = \int \cdots \int_{[0,1]^{d-1} \times [0, u_d^*]} c(u_1, \ldots, u_d) du_1 \ldots du_d.$$

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

$$P(\widehat{R_p \leq \xi}) = \frac{1}{n_s} \sum_{i=1}^{n_s} c(u_{1i}, \ldots, u_{di})$$

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 $\varepsilon_t = (\varepsilon_{1,t}, \dots, \varepsilon_{d,t})^\top$ are standardised *i.i.d.* innovations with $\mathsf{E}[\varepsilon_{j,t}] = 0$ and $\mathsf{E}[\varepsilon_{j,t}^2] = 1$ for $j = 1, \dots, d$; $\mathcal{F}_t$ 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 $\varepsilon = (\varepsilon_1, \dots, \varepsilon_d)^\top$ have joint distribution

$$F_\varepsilon(\varepsilon_1, \dots, \varepsilon_d) = C_\theta\{F_1(\varepsilon_1), \dots, F_d(\varepsilon_d)\}, \tag{15}$$

where C_θ is a copula belonging to a parametric family $\mathcal{C} = \{C_\theta, \theta \in \Theta\}$, and F_j, $j = 1, \dots, d$ 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 $\{x_{j,t}\}_{t=1}^T$, $j = 1, \dots, d$ 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

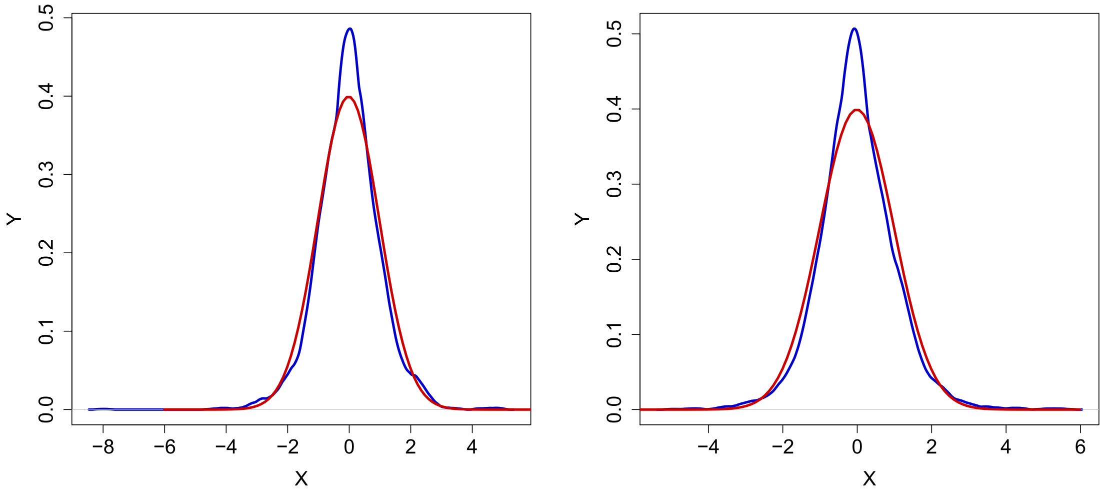


Fig. 1: Kernel density estimator of the residuals and of the normal density from TKA (left) and VOW (right). Quartic kernel, $\hat{h} = 2.78\hat{\sigma}n^{-0.2}$.

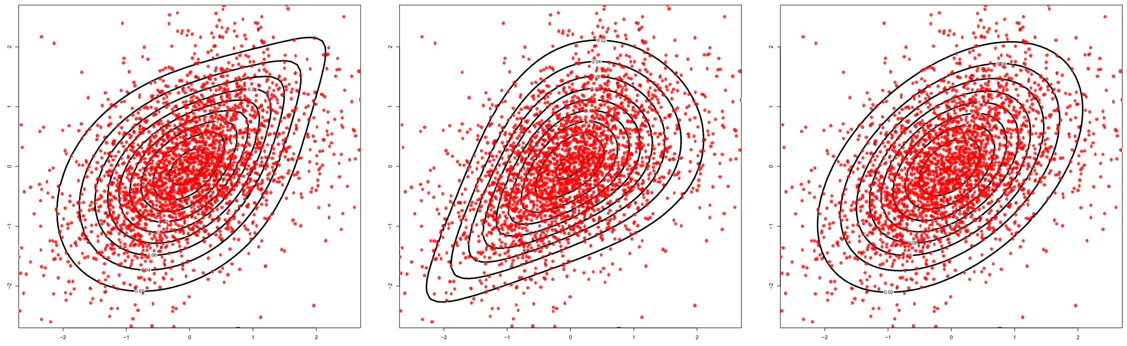


Fig. 2: Residuals $\hat{\varepsilon}$ and fitted copulae: Gaussian ($\hat{\rho} = 0.462$), Clayton ($\hat{\theta} = 0.880$), Gumbel ($\hat{\theta} = 1.439$).

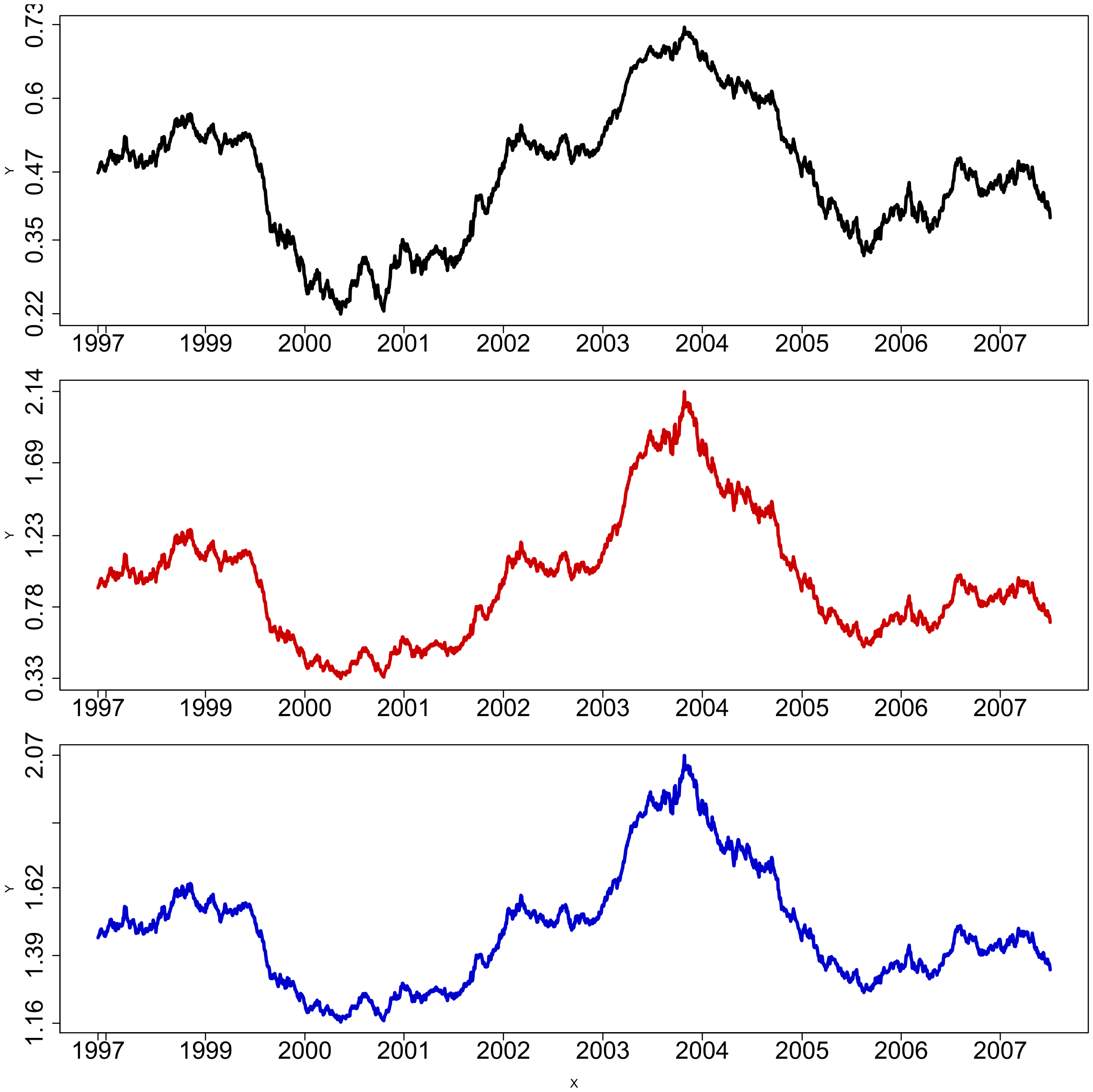


Fig. 3: Dependence parameter $\hat{\theta}$, estimated using the IFM method, Gaussian (upper panel), Gumbel (middle panel) and Clayton (lower panel) copulae, moving window ($w = 250$).

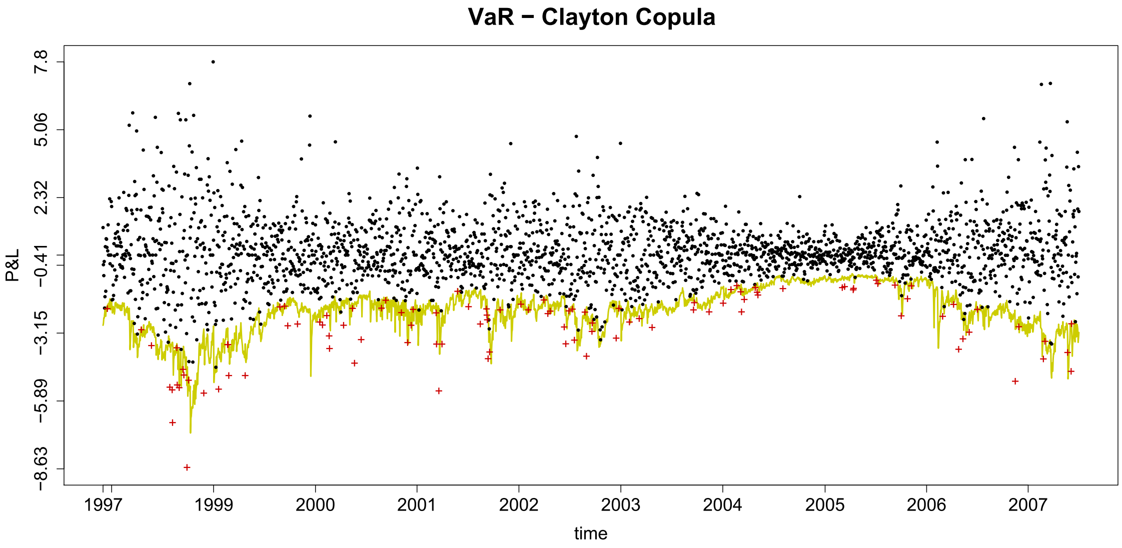


Fig. 4: $\widehat{VaR}(\alpha)$ (solid line), P&L (dots) and exceedances (crosses), $\alpha = 0.05$, $\hat{\alpha} = 0.0424$. P&L samples generated with Clayton copula.

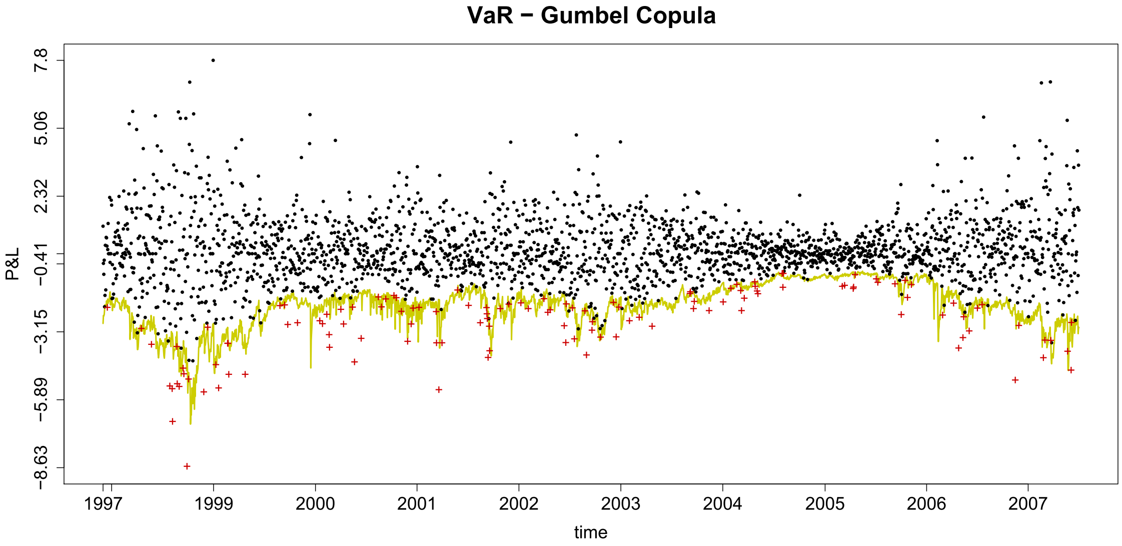


Fig. 5: $\widehat{VaR}(\alpha)$ (solid line), P&L (dots) and exceedances (crosses), $\alpha = 0.05$, $\hat{\alpha} = 0.0508$. P&L samples generated with Gumbel copula.

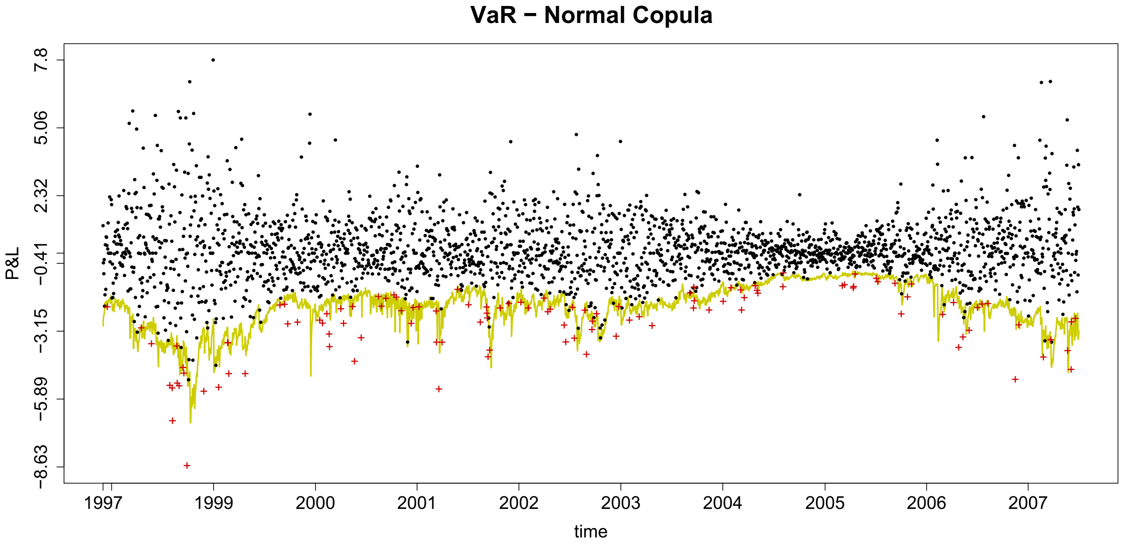


Fig. 6: $\widehat{VaR}(\alpha)$ (solid line), P&L (dots) and exceedances (crosses), $\alpha = 0.05$, $\hat{\alpha} = 0.0464$. P&L samples generated with Gaussian copula.

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.

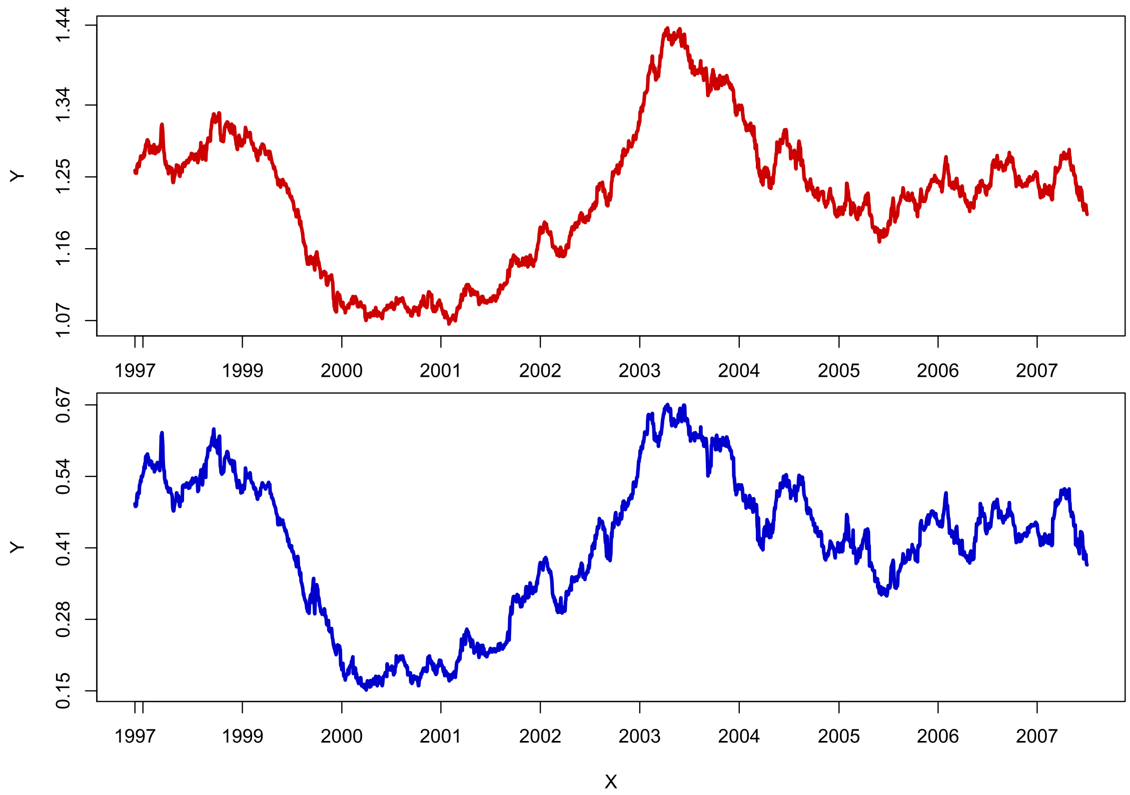


Fig. 7: Dependence parameter $\hat{\theta}$, estimated using the IFM method, Clayton (upper panel) and Gumbel (lower panel) copulae, moving window ($w = 250$).

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.

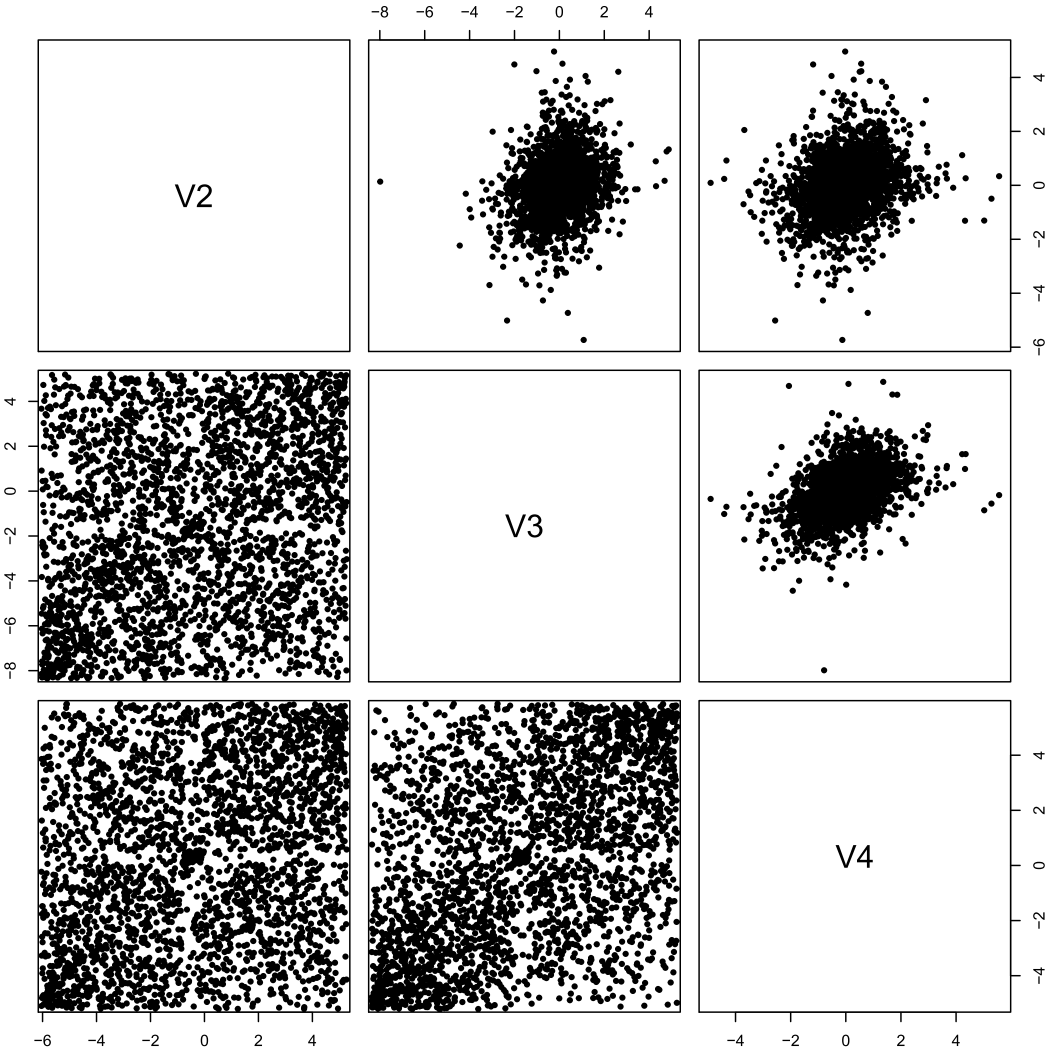


Fig. 8: Scatterplots from GARCH residulas (upper triangular) and from residuals mapped on unit square by the cdf (lower triangular).

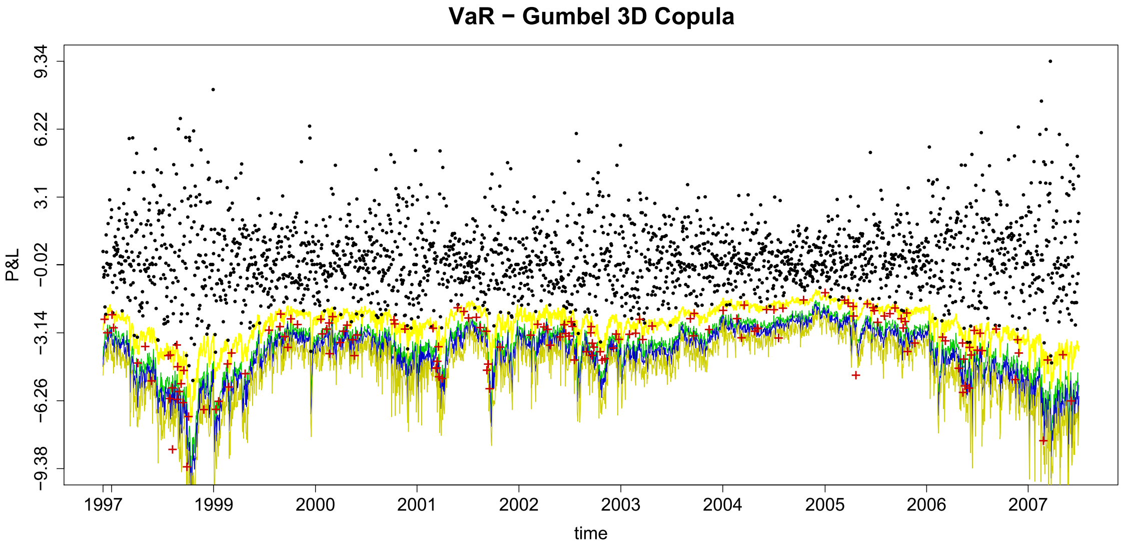


Fig. 9: $\widehat{VaR}(\alpha)$ and P&L (dots), estimated with 3-dimensional simple Gumbel copula, $\alpha_1 = 0.05$ ($\hat{\alpha}_1 = 0.0612$), $\alpha_2 = 0.01$ ($\hat{\alpha}_2 = 0.0232$), $\alpha_3 = 0.005$ ($\hat{\alpha}_3 = 0.016$) and $\alpha_4 = 0.001$ ($\hat{\alpha}_4 = 0.006$).

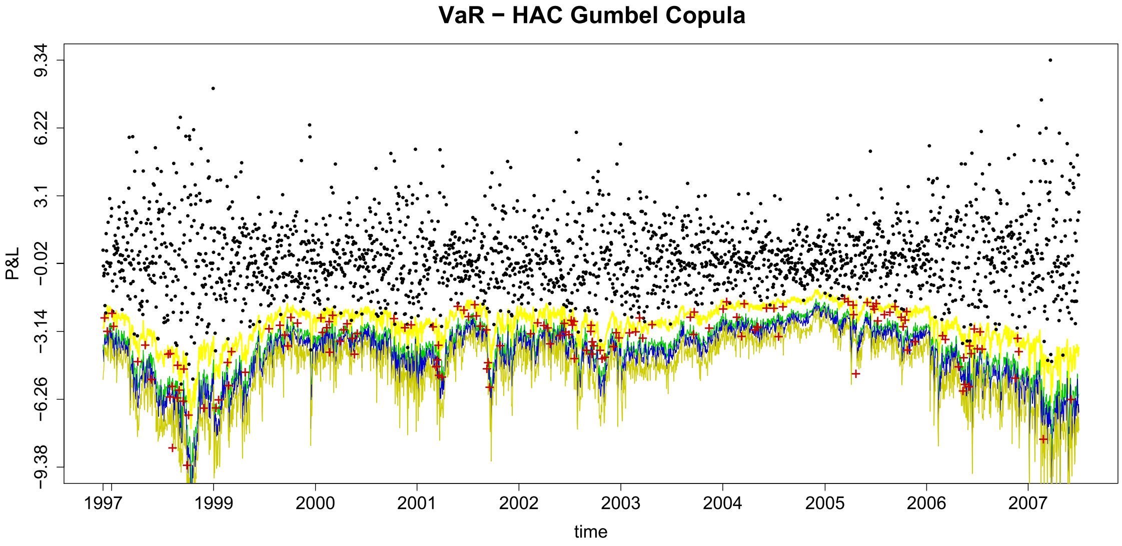


Fig. 10: $\widehat{VaR}(\alpha)$ and P&L (dots), estimated with 3-dimensional HAC with Gumbel generators, $\alpha_1 = 0.05$ ($\hat{\alpha}_1 = 0.0592$), $\alpha_2 = 0.01$ ($\hat{\alpha}_2 = 0.0208$), $\alpha_3 = 0.005$ ($\hat{\alpha}_3 = 0.014$) and $\alpha_4 = 0.001$ ($\hat{\alpha}_4 = 0.004$).

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 $L(\theta) \overset{\text{def}}{=} \mathsf{E}\{\rho_p(y - \theta)|X = x\}$ (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:

$$L_n(\theta) = n^{-1} \sum_{i=1}^{n} \rho_p(Y_i - \theta) K_h(x - X_i) \qquad \textbf{(1)}$$

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

$$\widetilde{H}_n(\theta,x) \overset{\text{def}}{=} n^{-1}\sum_{i=1}^{n} K_h(x-X_i)\psi(Y_i-\theta), \tag{2}$$

$$\widetilde{H}(\theta,x) \overset{\text{def}}{=} \int_{\mathbb{R}} f(x,y)\psi(y-\theta)\,dy, \tag{3}$$

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

$$\mathrm{P}\left((2\delta\log n)^{1/2}\left[\sup_{x\in J} r(x)|\{l_n(x) - l(x)\}|/\lambda(K)^{1/2} - d_n\right] < z\right) \\ \to \exp\{-2\exp(-z)\}, \quad \text{as } n \to \infty \tag{4}$$

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 $n \to \infty$*

$$\sup_{\theta\in I}\sup_{x\in J}\left|\widetilde{H}_n(\theta,x) - \widetilde{H}(\theta,x)\right| \le A^* \max\left\{(nh/\log n)^{-1/2}, h^{\tilde{\alpha}}\right\}. \tag{5}$$

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 $n \to \infty$*

$$\sup_{x\in J}\left|l_n(x) - l(x)\right| \le B^* \max\left\{(nh/\log n)^{-1/2}, h^{\tilde{\alpha}}\right\} \tag{7}$$

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

$$\sup_{x\in J}|l_n(x) - l(x)| \le B^*\{(nh/\log n)^{-1/2}\}.$$

THEOREM 2.2. *Let $h = n^{-\delta}$, $\frac{1}{5} < \delta < \frac{1}{3}$, $\lambda(K) = \int_{-A}^{A} K^2(u)\, du$, and*

$$d_n = (2\delta\log n)^{1/2} + (2\delta\log n)^{-1/2}\left[\log\left\{c_1(K)/\pi^{1/2}\right\} + \frac{1}{2}\{\log\delta + \log\log n\}\right],$$

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*

$$l_n(t) \pm (nh)^{-1/2} \Big\{p(1-p)\lambda(K)/\hat{f}_X(t)\Big\}^{1/2} \hat{f}^{-1}\{l(t)|t\} \Big\{d_n + c(\alpha)(2\delta \log n)^{-1/2}\Big\},$$

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

$$l_n(t) \pm (nh)^{-1/2} \Big\{d_n + c(\alpha)(2\delta \log n)^{-1/2}\Big\} \hat{\tau}(t).$$

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 $\|R_n\|=\sup_{t\in J}|R_n(t)|=\mathcal{O}_p\{(nh\log n)^{-1/2}\}$.

Furthermore, the rescaled linear part

$$Y_n(t)=(nh)^{1/2}\{\sigma^2(t)f_X(t)\}^{-1/2}\{H_n(t)-\mathsf{E}\,H_n(t)\}$$

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

$$T(x_1,x_2,\dots,x_d,y)=\{F_{X_1|y}(x_1|y),F_{X_2|y}(x_2|x_1,y),\dots,F_{X_k|x_{d-1},\dots,x_1,y}$$
$$(x_k|x_{d-1},\dots,x_1,y),F_Y(y)\}. \tag{12}$$

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*

$$\sup_{x\in J,y\in\mathbb{R}}|Z_n(x,y)-B_n\{T(x,y)\}|=\mathcal{O}\Big\{n^{-1/2}(\log n)^2\Big\}\quad a.s.,$$

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.

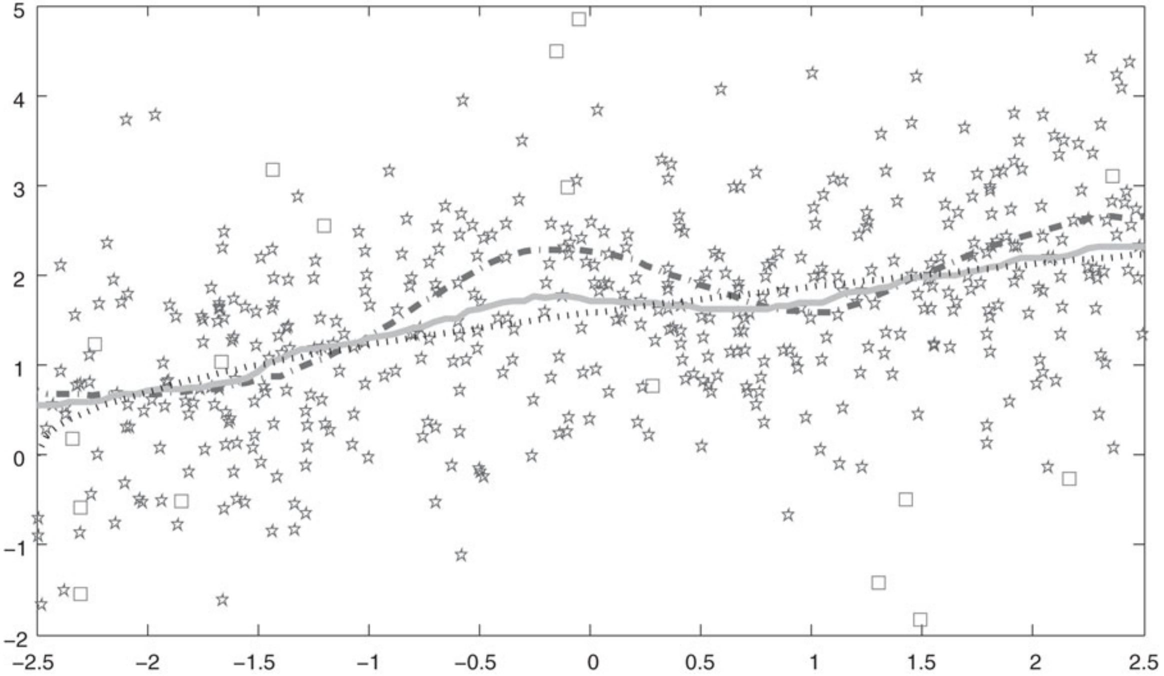


Figure 1. The 0.5-quantile curve, the Nadaraya–Watson estimator $m_n^*(x)$, and the 0.5-quantile smoother $l_n(x)$.

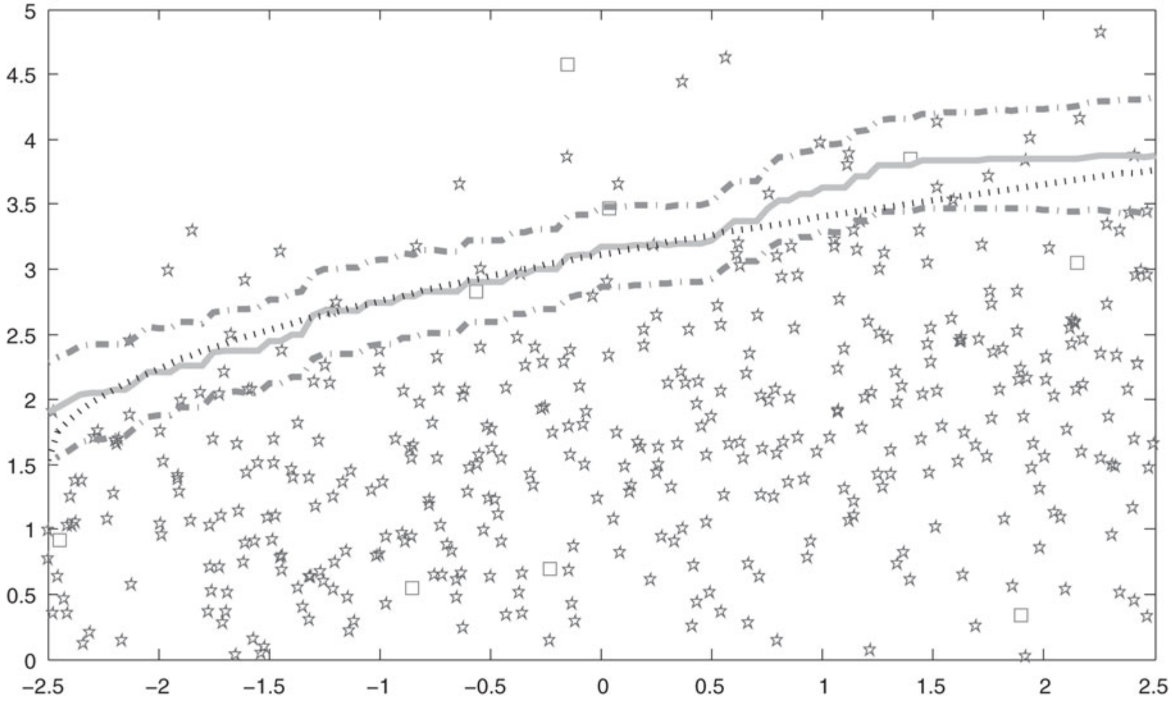


FIGURE 2. The 0.9-quantile curve, the 0.9-quantile smoother, and 95% confidence band.

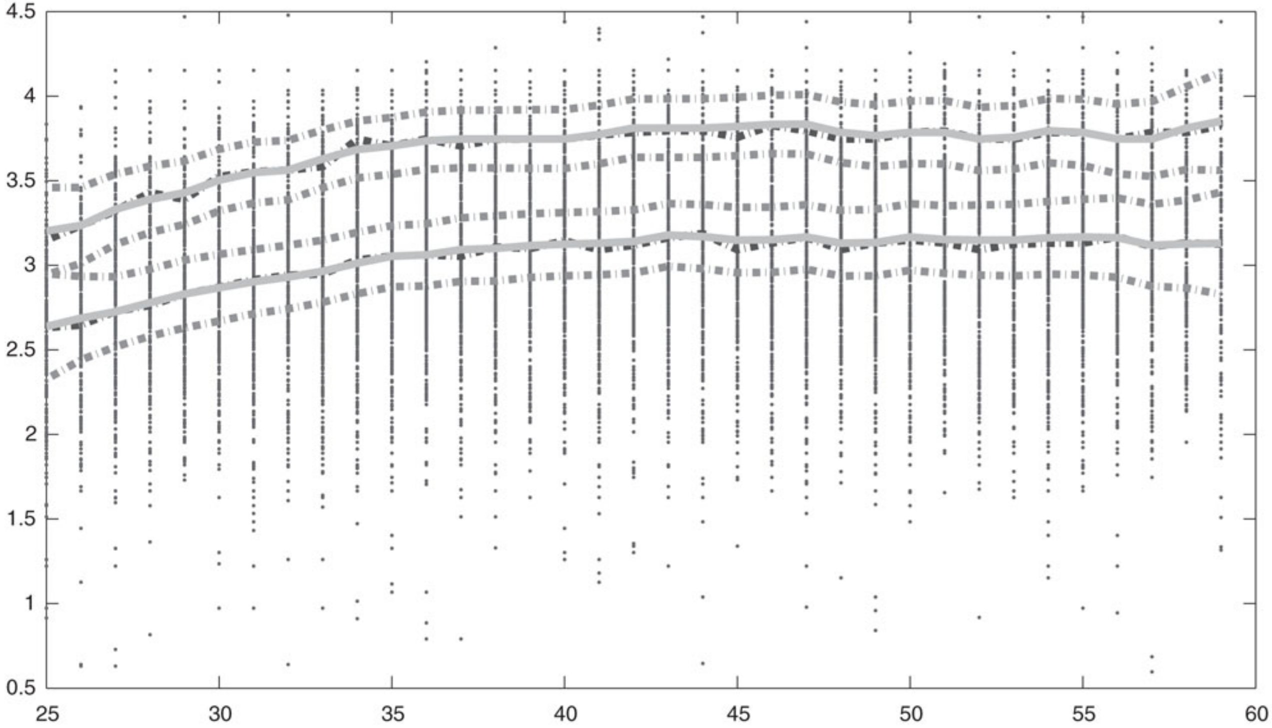


FIGURE 3. The original observations, local quantiles, 0.5- and 0.9-quantile smoothers, and corresponding 95% confidence bands.

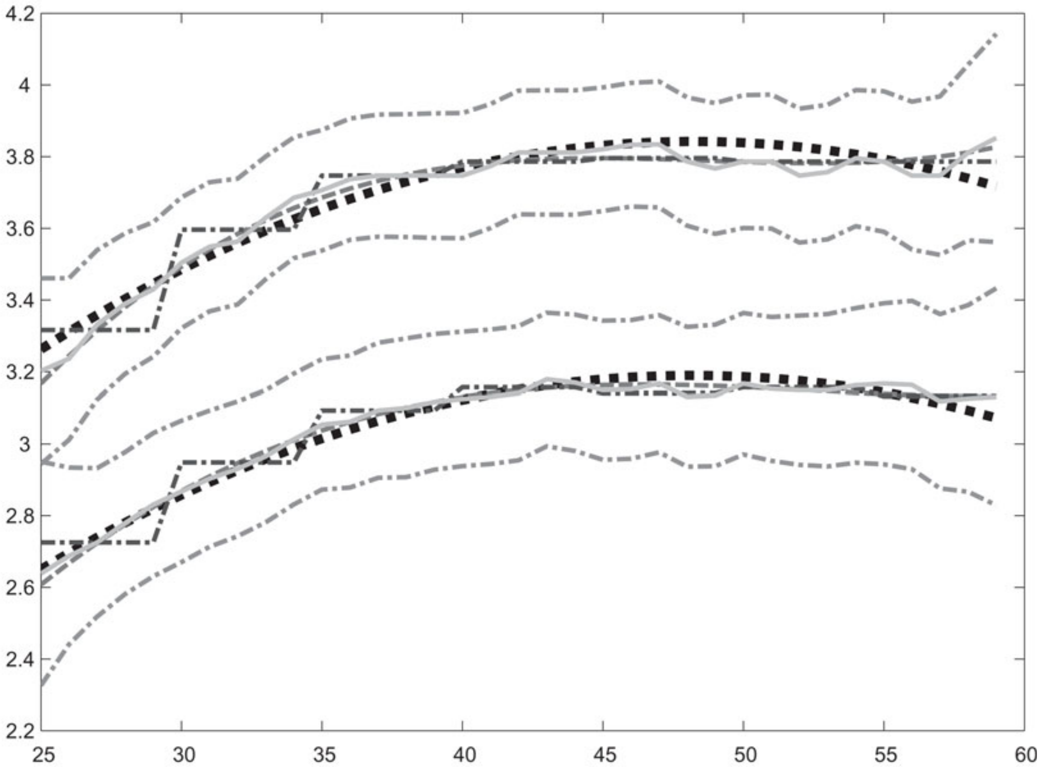


FIGURE 4. Quadratic, quartic, set of dummies (for age groups) estimates, 0.5- and 0.9-quantile smoothers, and their corresponding 95% confidence bands.

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*

$$\|R_n\| = \mathcal{O}_p\left\{(nh\log n)^{-1/2}\right\}. \tag{A.5}$$

Proof. First we have by the positivity of the kernel K,

$$\|R_n\| \leqslant \left[\inf_{0\leqslant t\leqslant 1}\{|D_n(t)|\cdot q(t)\}\right]^{-1}\{\|H_n\|\cdot\|q - D_n\| + \|D_n\|\cdot\|\mathsf{E}\,H_n\|\}$$

$$+ C_1\cdot\|l_n - l\|^2\cdot\left\{\inf_{0\leqslant t\leqslant 1}|D_n(t)|\right\}^{-1}\cdot\|f_n\|_\infty,$$

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

$$\|H_n\| = \mathcal{O}_p\left\{(nh)^{-1/2}(\log n)^{1/2}\right\}, \tag{A.6}$$

$$\|q - D_n\| = \mathcal{O}_p\left\{(nh)^{-1/4}(\log n)^{-1/2}\right\}, \tag{A.7}$$

$$\|\mathsf{E}\,H_n\| = \mathcal{O}\left(h^2\right), \tag{A.8}$$

$$\|l_n - l\|^2 = \mathcal{O}_p\left\{(nh)^{-1/2}(\log n)^{-1/2}\right\}. \tag{A.9}$$

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),

$$\sup_{t\in J}|H_n(t) - \mathsf{E}\,H_n(t)| = \mathcal{O}\left\{(nh)^{-1/2}(\log n)^{1/2}\right\}$$

and the following inequality

$$\begin{aligned}\|H_n\| &\leqslant \|H_n - \mathsf{E}\,H_n\| + \|\mathsf{E}\,H_n\| \\ &= \mathcal{O}\Big\{(nh)^{-1/2}(\log n)^{1/2}\Big\} + \mathcal{O}_p\Big\{(nh)^{-1/2}(\log n)^{-1/2}\Big\} \\ &= \mathcal{O}\Big\{(nh)^{-1/2}(\log n)^{1/2}\Big\},\end{aligned}$$

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.

$$\|Y_{0,n} - Y_{1,n}\| = \mathcal{O}\Big\{(nh)^{-1/2}(\log n)^2\Big\} \quad a.s.$$

Proof. Let t be fixed and put $L(y) = \psi\{y - l(t)\}$ still depending on t. Using integration by parts, we obtain

$$\begin{aligned}&\iint_{\Gamma_n} L(y)K\{(t-x)/h\}\,dZ_n(x,y) \\ &\quad= \int_{u=-A}^{A}\int_{y=-a_n}^{a_n} L(y)K(u)\,dZ_n(t-h\cdot u, y) \\ &\quad= -\int_{-A}^{A}\int_{-a_n}^{a_n} Z_n(t-h\cdot u, y)\,d\{L(y)K(u)\} \\ &\qquad+ L(a_n)(a_n)\int_{-A}^{A} Z_n(t-h\cdot u, a_n)\,dK(u) \\ &\qquad- L(-a_n)(-a_n)\int_{-A}^{A} Z_n(t-h\cdot u, -a_n)\,dK(u) \\ &\qquad+ K(A)\Bigg\{\int_{-a_n}^{a_n} Z_n(t-h\cdot A, y)\,dL(y) \\ &\qquad\qquad+ L(a_n)(a_n)Z_{n_a}(t-h\cdot A, a_n) - L(-a_n)(-a_n)Z_n(t-h\cdot A, -a_n)\Bigg\}\end{aligned}$$

$$- K(-A)\left\{\int_{-a_n}^{a_n} Z_n(t+h\cdot A,y)\,dL(y) + L(a_n)(a_n)Z_n(t+h\cdot A,a_n)\right.$$
$$\left. - L(-a_n)(-a_n)Z_n(t+h\cdot A,-a_n)\right\}.$$

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

$$\begin{aligned} h^{-1/2}S_{1,n}(t) = \Bigg| & h^{-1}\int_{-A}^{A} W(t-uh)K'(u)\left[\left\{\frac{g(t-uh)}{g(t)}\right\}^{1/2}-1\right] du \\ & \cdot\frac{1}{2}\int_{-A}^{A} W(t-uh)K(u)\left\{\frac{g(t-uh)}{g(t)}\right\}^{1/2}\left\{\frac{g'(t-uh)}{g(t)}\right\} du\Bigg| \\ = {} & |T_{1,n}(t)-T_{2,n}(t)|,\quad \text{say;} \end{aligned}$$

$\|T_{2,n}\| \leqslant C_5\cdot\int_{-A}^{A}|W(t-hu)|du = \mathcal{O}_p(1)$ 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:

$$\begin{aligned} V_n(t) &= (\log n)^{1/2}(nh)^{-1/2} \sum_{i=1}^{n} [\psi\{Y_i - l(t)\}\mathbf{1}(|Y_i| > a_n) K\{(t-X_i)/h\} \\ &\qquad - \mathsf{E}\,\psi\{Y_i - l(t)\}\mathbf{1}(|Y_i| > a_n)K\{(t-X_i)/h\}] \\ &= \sum_{i=1}^{n} X_{n,t}(t), \end{aligned}$$

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

$$\begin{aligned} \mathsf{E}\, X_{n,t}^2(t) &\leqslant (\log n)(nh)^{-1}\, \mathsf{E}\,\psi^2\{Y_i - l(t)\}\mathbf{1}(|Y_i| > a_n)K^2\{(t-X_i)/h\} \\ &\leqslant \sup_{-A\leqslant u\leqslant A} K^2(u)\cdot(\log n)(nh)^{-1}\,\mathsf{E}\,\psi^2\{Y_i - l(t)\}\mathbf{1}(|Y_i| > a_n). \end{aligned}$$

Hence

$$\begin{aligned} \operatorname{Var}\{V_n(t)\} &= \mathsf{E}\left\{\sum_{i=1}^{n} X_{n,t}(t)\right\}^2 = n\cdot \mathsf{E}\, X_{n,t}^2(t) \\ &\leqslant \sup_{-A\leqslant u\leqslant A} K^2(u) h^{-1}(\log n)\int_{\{|y|>a_n\}} f_y(y)\,dy \cdot M_\psi, \end{aligned}$$

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:

$$\mathsf{E}\{|V_n(t) - V_n(t_1)|\cdot|V_n(t_2) - V_n(t)|\} \leqslant C'\cdot(t_2 - t_1)^2$$

$$C' \text{ denoting a constant,} \qquad t \in [t_1, t_2].$$

We again estimate the left-hand side by Schwarz's inequality and estimate each factor separately:

$$\begin{aligned} \mathsf{E}\{V_n(t) - V_n(t_1)\}^2 = (\log n)(nh)^{-1}\,\mathsf{E}\Bigg[&\sum_{i=1}^{n} \Psi_n(t, t_1, X_i, Y_i)\cdot \mathbf{1}(|Y_i| > a_n) \\ &- \mathsf{E}\{\Psi_n(t, t_1, X_i, Y_i)\cdot\mathbf{1}(|Y_i| > a_n)\}\Bigg]^2, \end{aligned}$$

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

$$\begin{aligned} r_3(t_1, t_2) &= \mathsf{E}\, Y_{3,n}(t_1) Y_{3,n}(t_2) \\ &= \{g(t_1) g(t_2)\}^{-1/2} h^{-1} \iint_{\Gamma_n} \psi^2\{y - l(x)\} K\{(t_1 - x)/h\} \\ &\quad \times K\{(t_2 - x)/h\} f(x, y)\, dx\, dy \\ &= \{g(t_1) g(t_2)\}^{-1/2} h^{-1} \iint_{\Gamma_n} \psi^2\{y - l(x)\} f(y|x)\, dy K\{(t_1 - x)/h\} \\ &\quad \times K\{(t_2 - x)/h\} f_X(x)\, dx \\ &= \{g(t_1) g(t_2)\}^{-1/2} h^{-1} \int g(x) K\{(t_1 - x)/h\} K\{(t_2 - x)/h\}\, dx \\ &= r_4(t_1, t_2), \end{aligned}$$

where $r_4(t_1, t_2)$ is the covariance function of the Gaussian process $Y_{4,n}(t)$, which proves the lemma. ■

estimation. *Statist.*
tion of dimension
class of dimension
ssions. *Biometrika*
om estimating the
9, 1733-1757.
series with special
26, 267-298.

U.S.A.

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.

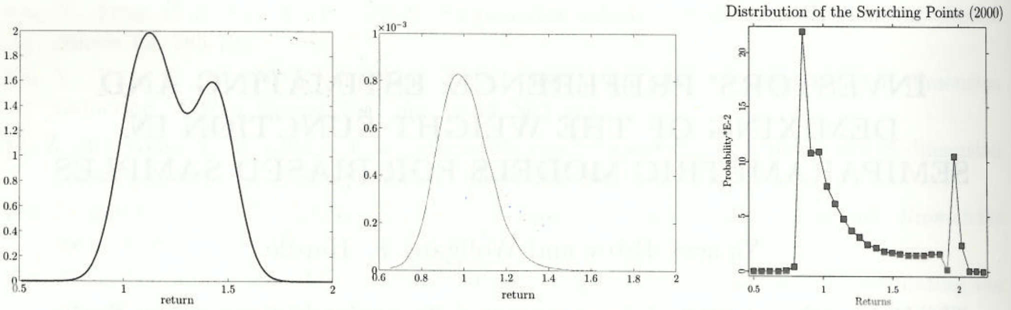
Key words and phrases: Empirical pricing kernel, exponential mixture, inverse problem, mixture distribution, risk aversion.

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).

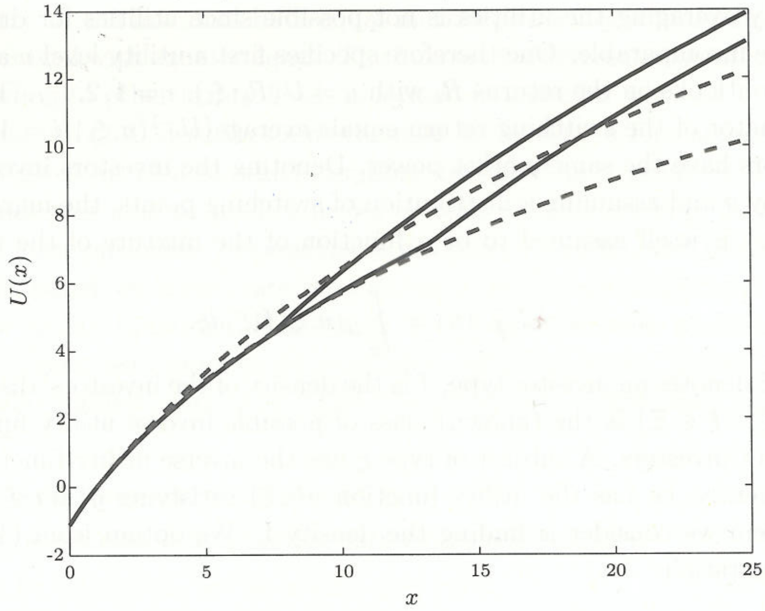


Figure 2. The utility function $U(\cdot;\xi)$ of (3.5) ($\alpha_1 = 2$, $\alpha_2 = 2.25$, $c = 2$) for two different values of ξ (solid lines), and of (3.8) for two values broken lines.

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

$$\begin{aligned} &p\Big\{\int g(u;\xi)f(\xi)d\mu(\xi)\Big\}\int\frac{\partial}{\partial u}g(u;\xi)f(\xi)d\mu(\xi)\\ &\quad=cq\Big\{\int g(u;\xi)f(\xi)d\mu(\xi)\Big\}\Big\{\int\frac{\partial}{\partial u}g(u;\xi)f(\xi)d\mu(\xi)\Big\}^2, \qquad (2.1)\end{aligned}$$

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

$$p\Big\{\int g(u;\xi)f(\xi)d\mu(\xi)\Big\} = \frac{q\{\int g(u;\xi)f(\xi)d\mu(\xi)\}\int \frac{\partial}{\partial u}g(u;\xi)f(\xi)d\mu(\xi)}{\int q\{\int g(v;\xi)f(\xi)d\mu(\xi)\}\{\int \frac{\partial}{\partial u}g(v;\xi)f(\xi)d\mu(\xi)\}^2 dv}.$$

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

$$\begin{aligned}\dot{\ell}_f(\xi) &= \sum_{i=1}^n \frac{1}{\psi'_f\{\psi_f^{-1}(X_i)\}}\Big[\frac{\partial}{\partial u}g\{\psi_f^{-1}(X_i);\xi\} - \frac{\psi''_f}{\psi'_f}\{\psi_f^{-1}(X_i)\}g\{\psi_f^{-1}(X_i);\xi\}\Big] \\ &\quad - nA_f(\xi), \\ &= \sum_{i=1}^n \frac{1}{\psi'_f\{U_i\}}\Big\{\frac{\partial}{\partial u}g\{U_i;\xi\} - \frac{\psi''_f}{\psi'_f}(U_i)g(U_i;\xi)\Big\} - nA_f(\xi),\end{aligned}$$

with $U_i = \psi_f^{-1}(X_i)$, and for all $\xi \in$ 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 discuss 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 $X_i = \psi_f(U_i) + \varepsilon_i$, $i = 1, \dots, n$, where $\varepsilon_1, \dots, \varepsilon_n$ is a random sample from $N(0, \sigma^2)$ independent from the random sample $U_1, \dots, U_n$ 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 $C_f = \log \int ql\{\psi_f(v)\}\{\psi'_f(v)\}^2 dv$. 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:

$$U_i \leftarrow \psi_f^{-1}(X_i), \qquad i = 1, \dots, n, \tag{2.3}$$

The M step:

$$f \leftarrow \text{argmax} \left[\sum_{i=1}^{n} \Big\{ \log q\{\psi_f(U_i)\} + 2\log\{\psi'_f(U_i)\} \Big\} - nC_f \right].$$

Let $\boldsymbol{U} = (U_1, \dots, U_n)$, $\boldsymbol{X} = (X_1, \dots, X_n)$, 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 $\xi \in$ 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 $x, \xi > 0$,

$$U(x;\xi) = \alpha_2(1-c)^{1-1/\alpha_2}\Big\{[x-\xi]_+^{1/\alpha_1} \vee (x-c)^{1/\alpha_2}\Big\} - \alpha_2(1-c), \tag{3.5}$$

where $\alpha_2 > \alpha_1 > 1$ are given, $c < 0$, and $[x]_+ = x\mathbf{1}(x > 0)$. See Figure 2. Then

$$g(u;\xi) = \min\Big\{\beta^{\alpha_2}\{u + \alpha_2(1-c)\}^{\alpha_2} + c,\ \beta^{\alpha_1}\{u + \alpha_2(1-c)\}^{\alpha_1} + \xi\Big\},$$

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 $\Sigma_{ij}(u_1,\ldots,u_k) = \operatorname{Cov}\{\hat{\psi}(u_i), \hat{\psi}(u_j)\}$. 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 $\max_i |\mathrm{E}\,\psi(u_i) - \psi(u_i)|^2 = \mathcal{O}(\log n/n)$.

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,

$$\begin{aligned}
k^{-1}\mathrm{E}\sum_{i=1}^k\Big\{\tilde{\psi}(u_i) - \psi(u_i)\Big\}^2 &= k^{-1}\mathrm{E}\Big\{\varepsilon^\top Z(Z^\top Z)^{-1}Z^\top\varepsilon\Big\} \\
&= k^{-1}\,\mathrm{trace}\Big\{Z(Z^\top Z)^{-1}Z^\top \mathrm{E}\,(\varepsilon\varepsilon^\top)\Big\}.
\end{aligned}$$

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

$$k^{-1}\mathrm{E}\sum_{i=1}^k\Big\{\tilde{\psi}(u_i) - \psi(u_i)\Big\}^2 = \mathcal{O}\left[\frac{m}{k}\left\{\frac{k}{n} + k\frac{\log n}{n} + k\Big(\frac{b^m}{m!}\Big)^2\right\}\right].$$

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.

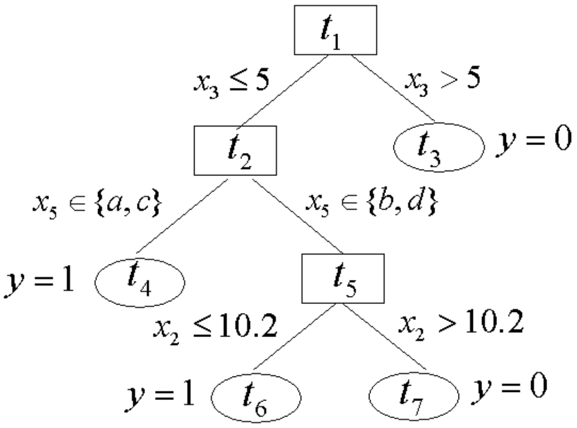


Figure 1: Example of a classification tree.

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

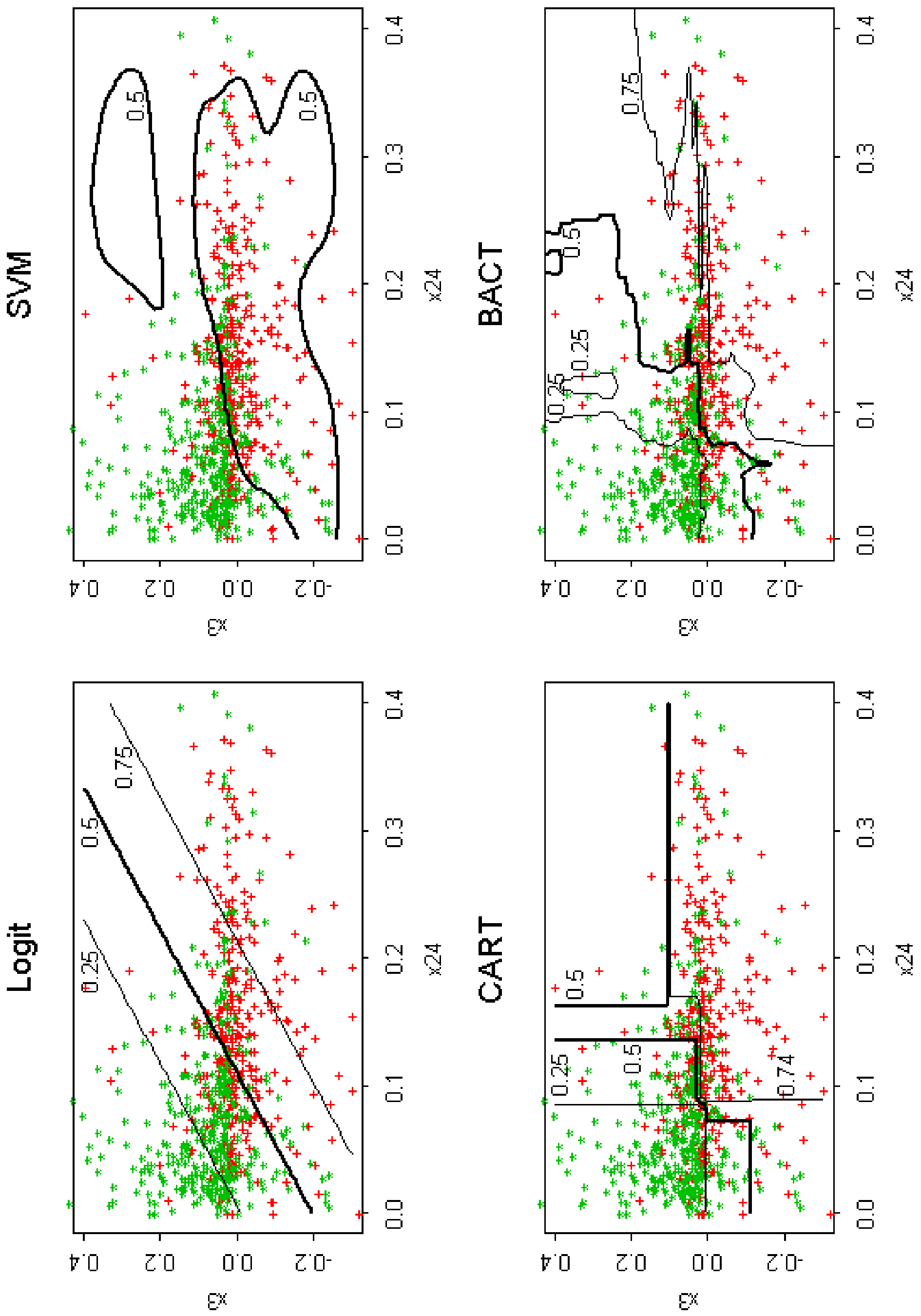


Figure 2: The contour plots for the logit model, SVM, CART, BACT. The pluses and stars represent insolvent firms and solvent firms respectively. The numbers by the contours indicate the probabilities of insolvency.

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 $\sigma_\mu^{2(0)} = \max\{-\min(\boldsymbol{y}^{*(0)})/2\sqrt{m}, \max(\boldsymbol{y}^{*(0)})/2\sqrt{m}\}$. 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., $\sigma_\mu^2 = \max\{-\min(\boldsymbol{y}^{*(1)})/2\sqrt{m}, \max(\boldsymbol{y}^{*(1)})/2\sqrt{m}\}$.

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.

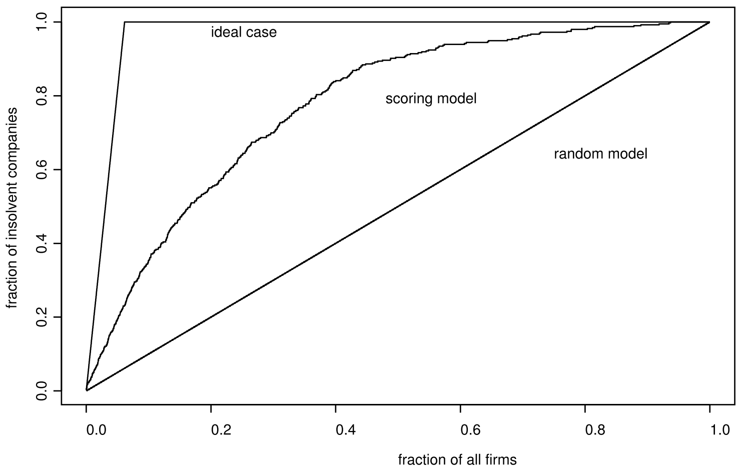


Figure 3: The CAP curve for the test set of the Creditreform data where the scoring model is the BACT model trained using one bootstrap training subset.

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. Copyright © 2009 John Wiley & Sons, Ltd.

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$^{\text{TM}}$ 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 co-workers (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

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 $\xi^{(')}$; 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:

$$\min_{\mathbf{w}\in\mathbb{R}^l,\, \xi^{(')}\in\mathbb{R}^{2T},\, b\in\mathbb{R}} \mathbf{C}(\mathbf{w}, b, \xi_t, \xi_t') = \frac{1}{2}\|\mathbf{w}\|^2 + C\sum_{t=1}^{T}(\xi_t + \xi_t') \tag{3}$$

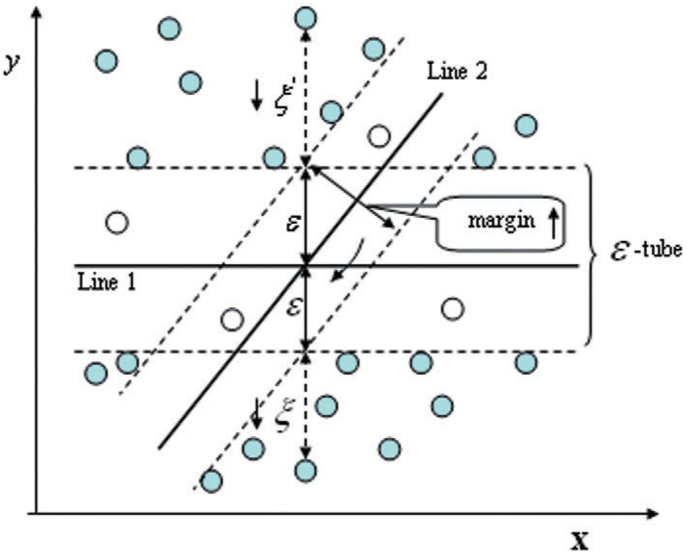


Figure 1. Principle of structural risk minimization (SRM) of SVM

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:

$$\mathbf{w}^*=\sum_{t=1}^{T}(\alpha_t'-\alpha_t)\phi(\mathbf{x_t}) \tag{10}$$

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:

$$\text{Linear:} \quad K(\mathbf{x_t}, \mathbf{x}) = \mathbf{x}_t^T \mathbf{x} \tag{14}$$

$$\text{Polynomial:} \quad K(\mathbf{x_t}, \mathbf{x}) = (\mathbf{x_t}^T \mathbf{x} + 1)^d \tag{15}$$

$$\text{Gaussian:} \quad K(\mathbf{x_t}, \mathbf{x}) = \exp\left(\frac{-\|\mathbf{x} - \mathbf{x_t}\|^2}{2\sigma^2} \right) \tag{16}$$

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)

$$\hat{u}_{t+1}^2 = \frac{1}{5}\sum_{j=t-4}^{t} (y_j - \bar{y}_{5,t})^2 \tag{23}$$

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

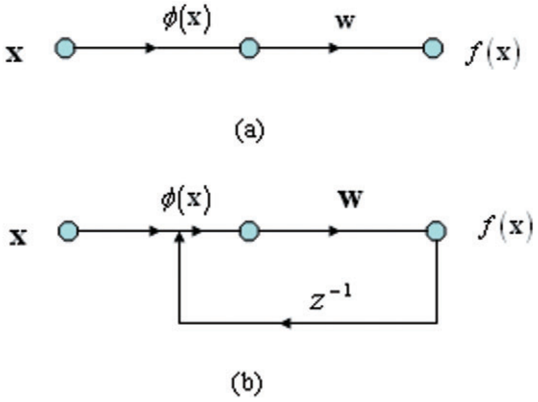


Figure 2. Signal-flow graphs of feedforward and recurrent SVM. (a) Signal-flow graph of a feedforward SVW. (b) Signal-flow graph of a single-loop recurrent SVW

[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, $\hat{u}_{T-59}^2$, $\hat{u}_{T-58}^2$, . . . , $\hat{u}_{T-1}^2$, $\hat{u}_T^2$, 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):

$$\text{MAE} = \frac{1}{n}\sum_{t=T_1}^{T-1}\left|u_{t+1}^2 - \hat{u}_{t+1}^2\right| \tag{24}$$

$$\text{DA}(\%) = \frac{100}{n}\sum_{t=T_1}^{T-1} a_t \tag{25}$$

where

$$a_t = \begin{cases} 1 & (u_{t+1}^2 - u_t^2)(\hat{u}_{t+1}^2 - \hat{u}_t^2) \geqslant 0 \\ 0 & \text{otherwise} \end{cases}$$

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

$$u_t^2 = (y_t - \bar{y})^2 \tag{26}$$

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

$$H_0\colon \text{MAE}_1 - \text{MAE}_0 = 0 \textit{ versus } H_1\colon \text{MAE}_1 - \text{MAE}_0 \neq 0$$

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 $T \rightarrow \infty$, $n \rightarrow \infty$ and $n/T_1 \rightarrow 0$, 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 $(c, \phi_1, \kappa, \delta_1, \alpha_1) = (0, 0.5, 0.0005, 0.8, 0.1)$ 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

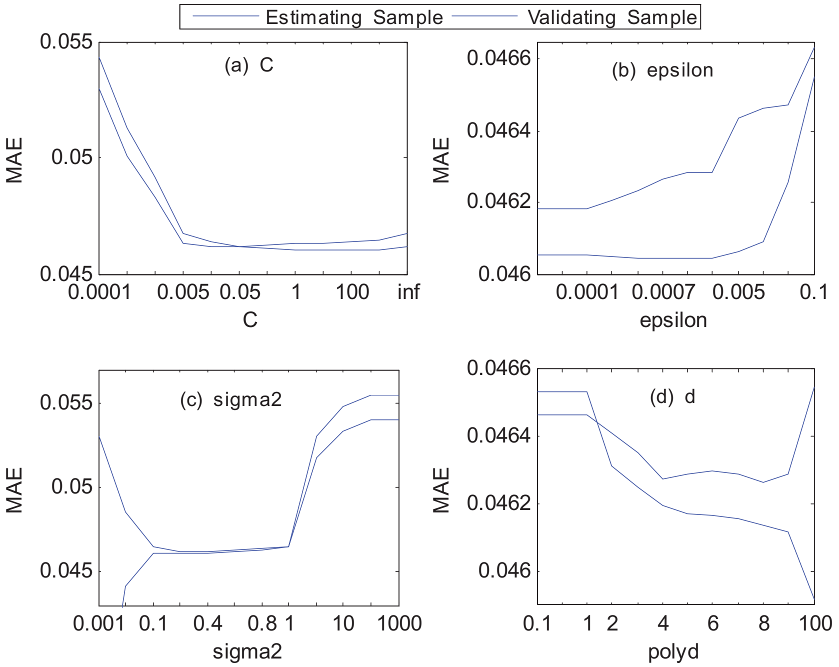


Figure 3. Sensitivity analysis of SVM in conditional mean estimation

$C = 0.05$, 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

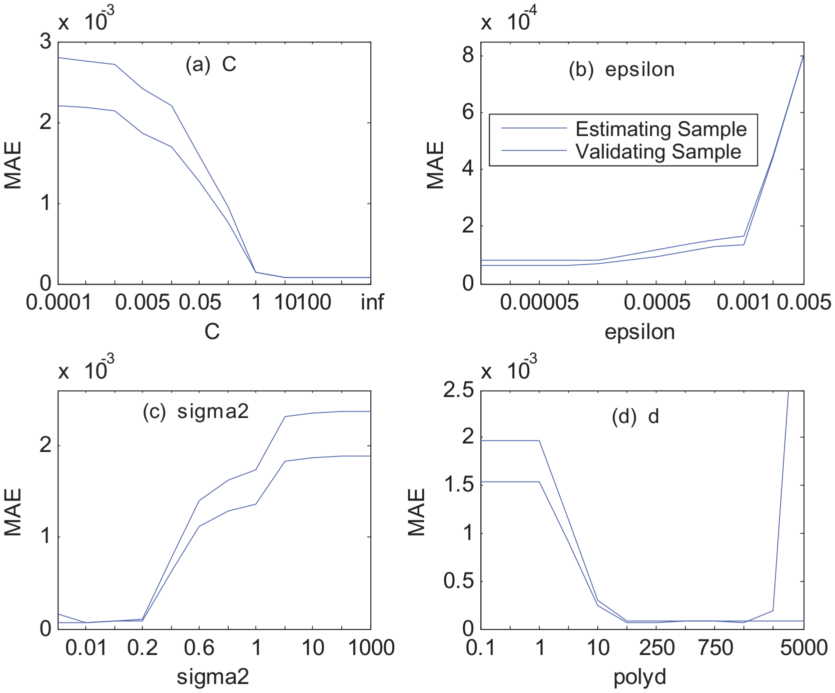


Figure 4. Sensitivity analysis of SVM in conditional variance estimation

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.

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.

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

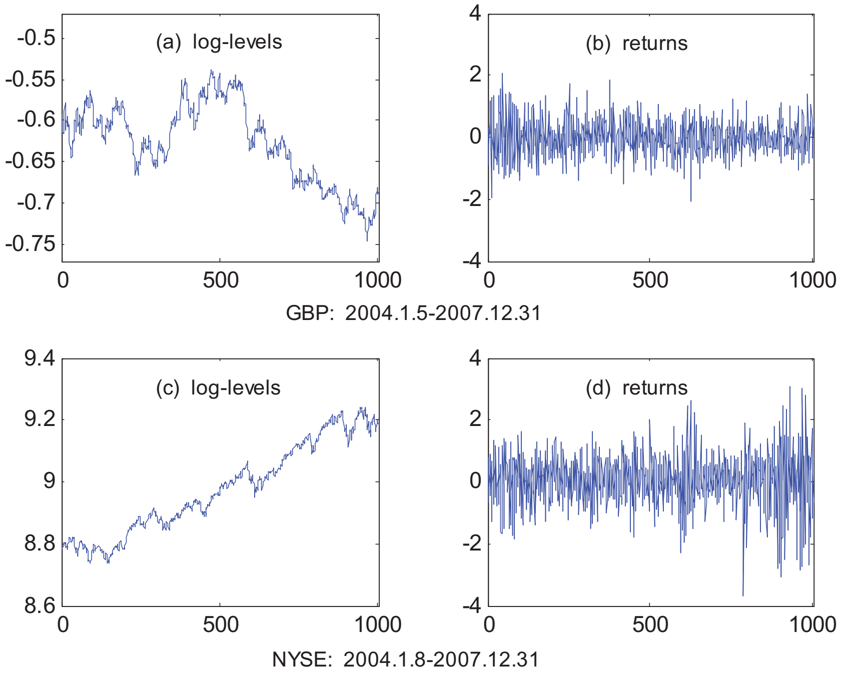


Figure 5. Log levels and returns of GBP exchange rates and NYSE stock index

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.

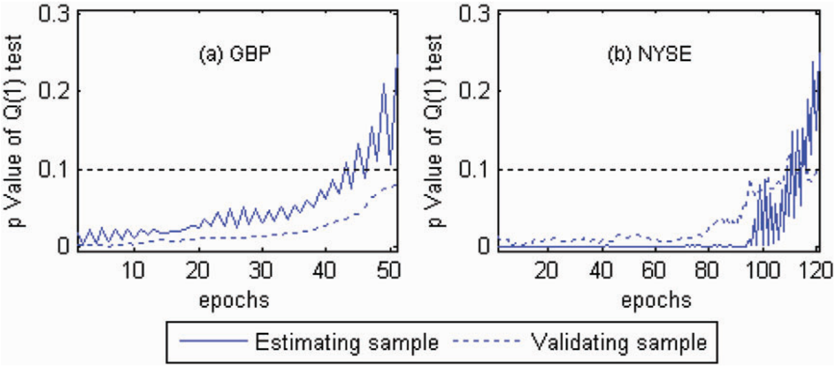


Figure 6. Iterative epochs of recurrent SVR procedure for real data

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.

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 Quantitativ*e *Financ*e 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.

DOI: 10.1002/for