Journal of the American Statistical Association
Publication details, including instructions for authors and subscription information:
http://amstat.tandfonline.com/loi/uasa20

Localized Realized Volatility Modeling

Ying Chen, Wolfgang Karl Härdle and Uta Pigorsch
Ying Chen is Assistant Professor, Department of Statistics & Applied Probability, National University of Singapore, 6 Science Drive 2, Singapore 117546 . Wolfgang Karl Härdle is Ladislaus von Bortkiewicz Chair of Statistics, C.A.S.E.—Center for Applied Statistics and Economics, School of Business and Economics, Humboldt-Universität zu Berlin, Spandauerstr. 1, 10178 Berlin, Germany. Uta Pigorsch is Junior Professor, Department of Economics, Universität Mannheim, L7, 3-5, 68131 Mannheim, Germany. We are grateful to two editors, the associate editor, and three anonymous referees for their valuable comments. This research was supported by the Deutsche Forschungsgemeinschaft through the SFB 649 "Economic Risk" and the SFB 884 "Political Economy of Reforms," and by the Berkeley-NUS Risk Management Institute at the National University of Singapore.

Available online: 01 Jan 2012

To cite this article: Ying Chen, Wolfgang Karl Härdle and Uta Pigorsch (2010): Localized Realized Volatility Modeling, Journal of the American Statistical Association, 105:492, 1376-1393

To link to this article: http://dx.doi.org/10.1198/jasa.2010.ap09039

Localized Realized Volatility Modeling

Ying CHEN, Wolfgang Karl HÄRDLE, and Uta PIGORSCH

With the recent availability of high-frequency financial data the long-range dependence of volatility regained researchers' interest and has led to the consideration of long-memory models for volatility. The long-range diagnosis of volatility, however, is usually stated for long sample periods, while for small sample sizes, such as one year, the volatility dynamics appears to be better described by short-memory processes. The ensemble of these seemingly contradictory phenomena point towards short-memory models of volatility with nonstationarities, such as structural breaks or regime switches, that spuriously generate a long memory pattern. In this paper we adopt this view on the dependence structure of volatility and propose a localized procedure for modeling realized volatility. That is at each point in time we determine a past interval over which volatility is approximated by a local linear process. A simulation study shows that long memory processes as well as short memory processes with structural breaks can be well approximated by this local approach. Furthermore, using S&P500 data we find that our local modeling approach outperforms long-memory type models and models with structural breaks in terms of predictability.

KEY WORDS: Adaptive procedure; Localized autoregressive modeling.

1. INTRODUCTION

Volatility is one of the key elements in modeling the stochastic dynamic behavior of financial assets. It is not only a measure of uncertainty about returns but also an important input parameter in derivative pricing, hedging, and portfolio selection. Accurate volatility modeling is therefore in the focus of financial econometrics and quantitative finance research. With the availability of high-frequency data, so-called realized volatility estimators (sums of squared high-frequency returns) have been proposed and have been shown to provide better volatility forecasts than the concurrent volatility estimators based on a coarser (e.g., daily) sampling frequency; see, for example, Andersen et al. (2001b).

Realized volatility together with other volatility measures exhibit significant autocorrelation which is the basis for the statistical predictability of volatility. In fact, the sample autocorrelation function has typically a hyperbolically-like decaying shape, also known as "long memory." A strand of literature focused on this kind of correlation phenomenon. The long memory "diagnosis," however, is usually stated for long sample periods such as three to 10 years. Over shorter sample periods, however, the autocorrelation function usually exhibits less persistence. This is also illustrated in Figure 1, which depicts the daily sample autocorrelation functions of daily logarithmic realized volatility of the S&P500 index futures for a long sample period (1985–2005) and for a short sample period (1995). The different degrees of persistence suggest that the diagnosis can also be generated by a simple model with structural change inside such a rather long interval; the possibility of such intermediate changes provides an alternative view on the described phenomenon. Like in the physical sciences, where one uses wave and particle theory to explain the emission of light, we have here a duality of theories for the emission of volatility. It is the objective of our study to investigate this dual view on volatility phenomenon.

In the literature of the long memory view of volatility, fractionally integrated [$I(d)$] processes have frequently been under consideration due to their hyperbolically decaying shock propagation for $0 < d < 1$. These processes have been proposed by, for example, Granger (1980), Granger and Joyeux (1980), and Hosking (1981). When applied to volatility they seem to provide a better description and predictability than short-memory models estimated over (the same) long sample periods. A typical example is the empirically better performance of the fractional integrated generalized autoregressive conditional heteroscedaticity (FIGARCH) model of Baillie, Bollerslev, and Mikkelsen (1996) as opposed to a standard GARCH model. For realized volatility, the autoregressive fractional integrated moving average (ARFIMA) process emerged as a standard model; see, for example, Andersen et al. (2003) and Pong et al. (2004). An alternative and quite popular model that does not belong to the class of fractionally integrated processes but approximates the long-range dependence by a sum of several multiperiod volatility components is the heterogenous autoregressive (HAR) model proposed by Corsi (2009).

The question on the true source of the long-memory diagnosis, however, still remains. Long memory in realized volatility may in fact be due to its construction, that is, by the aggregation over squared intraday returns, which are well known to exhibit also long-range dependence. Liebermann and Phillips (2008) therefore develop refined methods for conducting inference on long memory. Their empirical results, however, support the general finding on long memory in realized volatility.

Moreover, the presence of structural breaks may result in misleading inference on the long memory diagnosis, as has already been noted in Diebold (1986) and Lamoureux and Lastrapes (1990). In fact, the theoretical results provided in Diebold and Inoue (2001) and Granger and Hyung (2004) show that this phenomenon can also be spuriously generated by a short-memory model with structural breaks or regime shifts. More

Ying Chen is Assistant Professor, Department of Statistics & Applied Probability, National University of Singapore, 6 Science Drive 2, Singapore 117546 (E-mail: *stacheny@nus.edu.sg*). Wolfgang Karl Härdle is Ladislaus von Bortkiewicz Chair of Statistics, C.A.S.E.—Center for Applied Statistics and Economics, School of Business and Economics, Humboldt-Universität zu Berlin, Spandauerstr. 1, 10178 Berlin, Germany. Uta Pigorsch is Junior Professor, Department of Economics, Universität Mannheim, L7, 3-5, 68131 Mannheim, Germany. We are grateful to two editors, the associate editor, and three anonymous referees for their valuable comments. This research was supported by the Deutsche Forschungsgemeinschaft through the SFB 649 "Economic Risk" and the SFB 884 "Political Economy of Reforms," and by the Berkeley–NUS Risk Management Institute at the National University of Singapore.

DOI: 10.1198/jasa.2010.ap09039

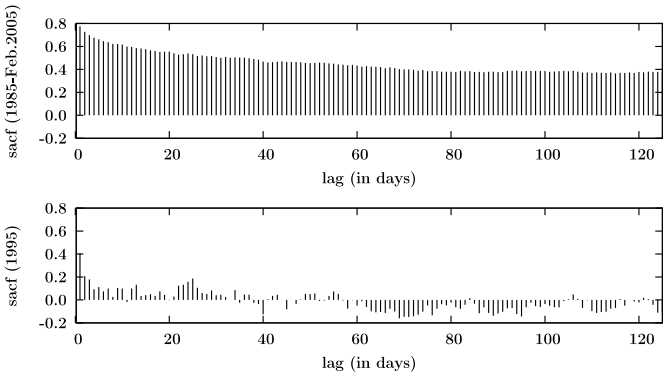

Figure 1. Sample ACF plots of daily logarithmic realized volatility of the S&P500 index futures for the sample from 1985–2005 (upper panel) and for the year 1995 (lower panel).

generally, Mikosch and Stărică (2004b) even argue independently of any particular model assumptions that nonstationarities in the data, such as changes in the unconditional mean or variance, can lead to the diagnosis of long-range dependencies. Such findings have led to the development of structural break detection methods and their application to financial volatility, where breaks are found in volatility processes using real data; see, for example, Chen and Gupta (1997), Mikosch and Stărică (2004a), Liu and Maheu (2008), and Čížek, Härdle, and Spokoiny (2009). Similarly, volatility models with time-varying coefficients have been proposed, which allow some or all of the model parameters to vary over time either in an abrupt fashion—for example, via Markov-Switching (see, e.g., Hamilton and Susmel 1994 and So, Lam, and Li 1998) and mixture multiplicative error specifications (see Lanne 2006)—or via a smooth function of time or other transition variables; see, for example, Baillie and Morana (2009b) and Scharth and Medeiros (2009), who show that nonlinearities, such as structural breaks and regimes induced by asymmetries like the leverage effect, may generate the observed long-range dependence. Such methods are also applied to long-memory models, addressing the possibility of the coexistence of long memory and structural breaks; see, for example, Baillie and Morana (2009a), Hillebrand and Medeiros (2008), and McAleer and Medeiros (2008b). The number of breaks in long memory realized volatility models is usually found to be one or two. Most of these studies, however, focus on sample periods covering at least 10 years. Given such a long time span of data, the presence of breaks even in long-memory models may be expected. Noteworthy, when it comes to forecasting, the more complicated models with breaks are often unable to significantly outperform the no-break long memory alternatives; see Hillebrand and Medeiros (2008), McAleer and Medeiros (2008b), and Martens, Dijk, and de Pooter (2009). Moreover, in some cases short memory models with breaks have been found to provide superior realized volatility forecasts than alternative long-memory models and regime switching ARFIMA models; see, for example, Lanne (2006) and Morana and Beltratti (2004).

In this paper we introduce the *localized realized volatility modeling* approach to describe realized volatility. In this approach the time-varying (local) structure of volatility is conveniently determined via adaptive statistical techniques, that allow us to find for each time point a past time interval, over which a local volatility model is a good approximator. Thus, in contrast to the previously cited literature our approach is local rather than global. The parameters of the local model as well as the length of the past time interval are determined at each point in time and may, therefore, differ from period to period. The method basically tries to adapt to local volatility. In doing so, it does not require any prior information or modeling assumptions on the number of break points, the potential (economic) sources of the break, its magnitude nor on its type (e.g., abrupt or smooth). This makes it very appealing. Moreover, it also allows to straightforwardly account for time-varying volatility of volatility, a feature that currently attracts researcher's interest, like Barndorff-Nielsen and Veraart (2009), and has been recognized to be important also for realized volatility; see Corsi et al. (2008) and Allen, McAleer, and Scharth (2010).

Although localized realized volatility modeling is a quite general concept that can be applied to various types of local parametric volatility models, we investigate it here based on autoregressive processes. In particular, we push here the alternative view on long memory to its limit by assuming a local linear short-memory model. Estimation and forecasting based on our approach is thus computationally straightforward.

The flexibility of our procedure is demonstrated within a simulation study, which shows that both, short-memory processes with breaks as well as long-memory processes, can be well described by the local approach. We additionally apply localized realized volatility modeling to S&P500 data and compare it to (approximate) long-memory techniques, such as the ARFIMA and HAR models, and to models with breaks. We find that our technique provides improved volatility forecasts.

The remainder of the paper is structured as follows. The next section reviews the concept of realized volatility, its construction, and the empirical properties of realized volatility of the S&P500 index futures. Section 3 presents in detail the localized realized volatility modeling approach along with a simulation study. Section 4 briefly reviews the alternative models considered in this paper, and Section 5 empirically compares the various models within a forecasting exercise. Section 6 concludes.

2. REALIZED VOLATILITY

Measuring the volatility of a financial asset based on high-frequency data has been one of the major focuses in the recent financial econometrics literature. The idea is to measure ex post the variation of asset prices over a lower frequency, commonly a day, by summing over products of high frequency, that is, intradaily returns. The approach is motivated by the theory of quadratic variation of semimartingales. For the ease of exposition, consider the case where the log price of a financial asset, p, follows a Brownian semimartingale—an assumption that is very popular in the asset pricing literature, that is,

$$p_t = \int_0^t \mu(s)\,ds + \int_0^t \sigma(s)\,dW(s) \qquad \forall t \geq 0, \tag{1}$$

where the instantaneous mean process $\{\mu(t)\}_{t\geq 0}$ is continuous and of finite variation, $\{\sigma(t)\}_{t\geq 0}$ with $\sigma(t) > 0$ $\forall t$ denotes the càdlàg instantaneous volatility, and $\{W(t)\}_{t\geq 0}$ is a standard Brownian Motion. Then the quadratic variation process of (1),

$$[p]_t = \operatorname{plim} \sum_{j=0}^{l-1} \left(p_{\tau_{j+1}} - p_{\tau_j}\right)^2, \tag{2}$$

where $\tau_0 = 0 \le \tau_1 \le \cdots \le \tau_l = t$ denotes a sequence of partitions with $\sup_j\{\tau_{j+1} - \tau_j\} \to 0$ for $l \to \infty$, is given by

$$[p]_t = \int_0^t \sigma^2(s)\, ds, \tag{3}$$

that is, as the integrated variance $\int_0^t \sigma^2(s)\, ds$ of the price process.

The theory of quadratic variation, thus, suggests that the sum over squared high-frequency returns may provide an ex post measure of the integrated variance and this is what is, oftentimes interchangeably, referred to as realized variance or realized volatility. Suppose we are interested in measuring volatility over a day t using $M+1$ intraday prices observed at time points $n_0, \ldots, n_M$. Furthermore, let p_{t,n_j} denote the logarithmic price observed at time point n_j of trading day t. The continuously compounded jth within-day return of day t is therefore given by

$$r_{t,j} = p_{t,n_j} - p_{t,n_{j-1}}, \qquad j = 1, \ldots, M. \tag{4}$$

Then *daily realized volatility* is defined as

$$\widetilde{RV}_t = \sum_{j=1}^{M} r_{t,j}^2. \tag{5}$$

Now, if $M \to \infty$, that is, the intraday sampling frequency goes to infinity, realized volatility converges to the quadratic variation of the price process; see, for example, Andersen and Bollerslev (1998) and Barndorff-Nielsen and Shephard (2002b). This implies that if the price follows a pure diffusion process as given in (1), realized volatility converges to the daily integrated variance, that is, $\widetilde{RV}_t \to IV_t$ for $M \to \infty$ with $IV_t = \int_{t-1}^t \sigma^2(s)\, ds$, which is oftentimes the main object of interest. Consistency and asymptotic distribution of realized volatility as an estimator of the integrated variance are derived in Barndorff-Nielsen and Shephard (2002a).

The theoretical results on realized volatility obviously build on the notion of an infinite sampling frequency. In practice, however, the sampling frequency is invariably limited by the actual quotation, or transaction frequency. Moreover, the observed high-frequency prices are further contaminated by market microstructure effects, such as the bid-and-ask bounce effect and price discreteness, which are due to the particular design and trading mechanism of financial markets; see, for example, Hasbrouck (2007). These effects introduce biases into realized volatility; see, for example, Andersen et al. (2001a) and Barndorff-Nielsen and Shephard (2002a). A common approach to reduce their impact is to simply construct realized volatility based on lower frequency returns (e.g., 10 to 30 minutes), at which market microstructure effects are negligible. However, such a procedure comes at the cost of a less precise volatility estimate, as it makes no use of all available data. Various alternative methods have therefore been proposed to solve this bias-variance trade-off. For a review, see McAleer and Medeiros (2008a) and Pigorsch, Pigorsch, and Popov (2010).

In this paper we compute a market microstructure noise robust version of realized volatility based on the approach of Barndorff-Nielsen et al. (2008). The reason for our choice is that their class of so-called *realized kernel estimators* of quadratic variation have very attractive properties. In particular, they are consistent and efficient and they are robust to a host of different market microstructure effects.

2.1 Noise-Corrected Realized Volatility

The idea of the realized kernel estimators is similar to that of autocorrelation and heteroscedasticity robust variance and covariance estimators, like the Newey–West estimator, that is, the correction is based on the sum of weighted autocovariances. Define the hth realized autocovariance for day t by $\gamma_{t,h} = \sum_{j=1}^{M} r_{t,j} r_{t,j-h}$. In the realized kernel estimators, realized volatility is then corrected by the weighted sum of those realized autocovariances. In particular, the flat-top realized kernel estimator, that we employ in this paper, provides a noise-corrected realized volatility RV_t by

$$RV_t = \widetilde{RV}_t + \sum_{h=1}^{H_t^*} k\left(\frac{h-1}{H_t^*}\right)(\gamma_{t,h} + \gamma_{t,-h}), \tag{6}$$

where the weights are given by the kernel function k being twice continuously differentiable on $[0, 1]$ and satisfying $k(0) = 1$, and $k(1) = k'(0) = k'(1) = 0$. The bandwith parameter H_t^* denotes the optimal number of lags to be considered for day t. It is optimal in the sense that it minimizes the asymptotic variance of the noise-corrected realized volatility. Barndorff-Nielsen et al. (2008) show that H_t^* depends on the chosen kernel weight function and on the noise-to-signal ratio $\xi_t = \omega_t^2/IV_t$, that relates the (daily) variance of the market microstructure noise, ω_t^2, to the (daily) integrated variance. In particular, $H_t^* = c^* \xi_t \sqrt{M}$, where c^* is a constant that depends, inter alia, on the specific kernel weight function. Its value is chosen such that it minimizes the asymptotic variance. The bandwith selection H_t^* and the computation of the noise-corrected realized volatility, thus, involve the precise specification of the kernel weight function and the estimation of the noise-to-signal ratio. We now turn to these issues.

For our empirical application we consider the modified Tukey–Hanning kernel with weight function $k(x) = \sin^2\{\frac{\pi}{2}(1 - x)^a\}$, as it is most efficient among the finite lag kernels analyzed in Barndorff-Nielsen et al. (2008). Moreover, for increasing a the noise-corrected realized volatility approaches the (parametric) efficiency bound. As such, a large value of a might be preferable. However, an increasing number of a also leads to an increase in the number of autocovariances H_t^* considered in the noise correction (6), as c^* is increasing with a; see Barndorff-Nielsen et al. (2008). In practice, this imposes some limitations as the computation of the autocovariances $\gamma_{t,h}$ then involves an increasing number of returns outside the daily time interval. Note that in our application we make exclusive use of price observations within a day, such that fewer observations are available for the estimation of $\gamma_{t,h}$ as h increases. An increase in a therefore implies the use of less precisely estimated autocovariance terms. We therefore follow Barndorff-Nielsen et al. (2008) and choose $a = 2$ for our empirical application. For this kernel specification $c^* = 5.74$, see Barndorff-Nielsen et al. (2008). Noteworthy, the chosen realized kernel estimator is still close to efficient.

To finally determine H_t^* we estimate the noise-to-signal ratio ξ_t in the following way: we employ the estimator of the noise variance suggested by Bandi and Russell (2005) and compute the (scaled) conventional realized volatility estimator based on one minute returns, that is, $\hat{\omega}_t^2 = \widetilde{RV}^{1\,\mathrm{min}}/2M^{1\,\mathrm{min}}$, where the superscripts indicate the used sampling interval. An estimate

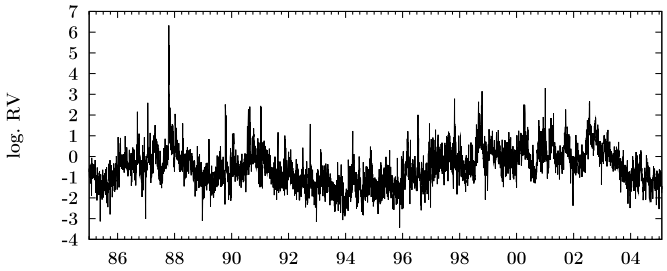

Figure 2. Time evolvement of logarithmic realized volatility of the S&P500 index futures.

of the variance of the "signal" (the integrated variance) is obtained by the realized volatility computed at a low, that is, 15 minutes, sampling interval, at which market microstructure effects should be negligible, thus $\widehat{IV}_t = \widetilde{RV}^{15\,\mathrm{min}}$. The optimal bandwidth is thus based on the estimate

$$\hat{H}_t^* = 5.74 \frac{\hat{\omega}_t^2}{\widehat{IV}_t} \sqrt{M^{1\,\mathrm{min}}}. \tag{7}$$

Rounding $\hat{H}_t^*$ to the nearest integer gives the final value of the bandwith. Given this bandwith, the noise-corrected realized volatility RV_t is then finally computed according to (6). Note that we estimate the realized autocovariances $\gamma_{t,h}$ and the market microstructure noise uncorrected realized volatility, $\widetilde{RV}_t$, based on one minute returns. Moreover, all intraday returns are constructed using the previous-tick method and by excluding overnight returns.

2.2 Data Description

Our empirical analysis focuses on the noise-corrected realized volatility of the S&P500 index futures ranging from January 2, 1985 to February 4, 2005; see Figure 2. From the various S&P500 Index futures with maturity dates in March, June, September, and December, we consider only the most liquid contracts. In addition, we have removed one day, February 18, 1990, from our dataset as there are only two transactions reported.

The descriptive statistics of the resulting realized volatility series are presented in Table 1. In summary, the empirical characteristics of the series are in line with the findings reported in the earlier literature on realized volatility. In particular, realized volatility is strongly skewed and fat-tailed, while its logarithmic version is much closer to Gaussianity. This is also confirmed by the kernel density estimate of logarithmic realized volatility, which is presented in Figure 3 along with the kernel density estimate of iid random variables simulated from the fitted normal distribution (with a sample size corresponding to the empirical one). Moreover, the sample autocorrelation function

Table 1. Descriptive statistics of realized volatility

Series	Mean	Std.Dev.	Skewness	Kurtosis	Ljung–Box(21)[(1)]
RV_t	1.0880	8.6961	55.5857	3412	1204
$\log(RV_t)$	−0.5314	0.8875	0.5343	4.9912	4.6861

[(1)]The critical value of the Ljung–Box test statistic of no autocorrelation up to approximately 1 month is 32.671.

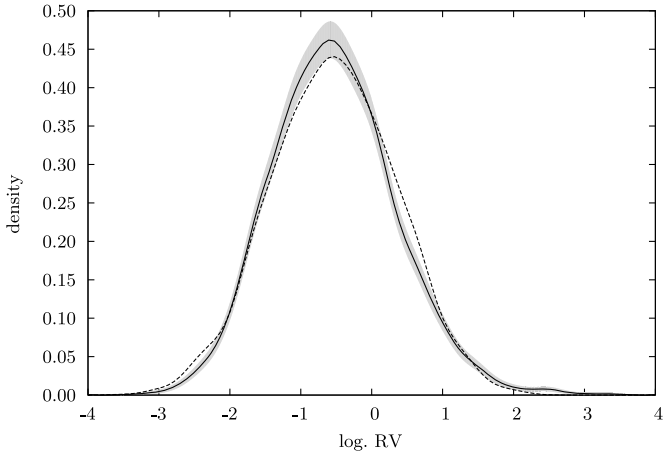

Figure 3. Kernel density estimate of logarithmic realized volatility of the S&P500 index futures (solid line). The shaded area corresponds to the pointwise 95% confidence intervals and the dashed line represents the kernel density estimate of iid random variables simulated from the fitted normal distribution.

of (logarithmic) realized volatility, Figure 1, exhibits the aforementioned hyperbolic decay. We evaluate this long-memory diagnosis in more detail in the empirical application. In the following, however, we first introduce our localized approach to realized volatility modeling.

3. THE LOCALIZED REALIZED VOLATILITY APPROACH

In this paper we adopt a local view on realized volatility modeling. The idea is simple. It is assumed that at each point in time there exists a past-time interval over which volatility can be well approximated by a local autoregressive (LAR) model. In contrast to fitting a global volatility model, we obtain at each point in time a potentially new set of parameters, which is estimated based on the so-called *interval of homogeneity*. For each point in time, the interval of homogeneity is selected in a sequential testing procedure, which starts from a small interval, where the local approximation holds and the AR parameters are approximately constant. The procedure then iteratively extends this interval and tests for time homogeneity until a structural break is found or data is exhausted. The local model is then fitted and can be used for volatility predictions.

The local (time-varying) autoregressive scheme is defined through a time-varying parameter set $\theta_t = (\theta_{0t}, \theta_{1t}, \ldots, \theta_{pt}, \sigma_t)^\top$:

$$\log RV_t = \theta_{0t} + \sum_{i=1}^{p} \theta_{it} \log RV_{t-i} + \varepsilon_t, \qquad \varepsilon_t \sim \mathrm{N}(0, \sigma_t^2), \tag{8}$$

where the Gaussian distributed innovations ε_t have mean zero and variance σ_t^2. Note that the specification also allows for time-varying volatility of volatility by letting σ_t rely on time.

Time-varying parameters at any point in time t are of course too flexible to really constitute a practical dynamic model. We therefore need to strike a balance between model flexibility and dimensionality. Traditional ways either estimate the time-varying parameters nonparametrically by assuming that the parameters are smooth functions of time (see, e.g., Cai,

Fan, and Li 2000) or assume that the time-varying parameters are piecewise constant functions provided that the number of changes are given (see, e.g., Bai and Perron 1998 and Mikosch and Stărică 2004a). Here we follow a different strategy by localizing (in time) a low-dimensional time series dynamics in the high-dimensional model (8). The basic idea is to approximate (8) at a fixed time point τ by a constant parameter vector $\theta_\tau = (\theta_{0\tau}, \theta_{1\tau}, \ldots, \theta_{p\tau}, \sigma_\tau)^\top$ over $I_\tau = [\tau - l_\tau, \tau)$ with $p + 2 \le l_\tau < \tau$. The interval I_τ is called the interval of homogeneity, whose length depends on time point τ. In the estimation of (8) at a particular time point τ, we only assume that an I_τ exists over which the local parametric model (approximately) holds for the process. This assumption nests the abovementioned "smooth transition" and "regime switching" assumptions as special cases: parameters can either smoothly vary over time or change abruptly. The question now is how to find I_τ or the value of l_τ over which the model parameters can be estimated.

The next section discusses the estimation and the test statistics employed to determine the interval of homogeneity. The sequential testing procedure is described in Section 3.2, while Section 3.3 discusses the choice of parameters involved in the procedure. The performance and sensitivity of the procedure are demonstrated in a set of simulations in Section 3.4.

3.1 Estimation and Test of Homogeneity

The estimation of the local parametric model is carried out via maximum likelihood. In particular, given an interval of homogeneity I_τ for time point τ, over which the process can be safely described by an AR model with constant parameters, the maximum likelihood (ML) estimator $\tilde{\theta}_\tau$ is defined as

$$\begin{aligned}\tilde{\theta}_\tau &= \operatorname*{argmax}_{\theta\in\Theta} L(\log RV; I_\tau, \theta)\\ &= \operatorname*{argmax}_{\theta\in\Theta}\left\{-\frac{l_\tau - p}{2}\log 2\pi - (l_\tau - p)\log\sigma\right.\\ &\quad\left. - \frac{1}{2\sigma^2}\sum_{t=\tau-l_\tau+p}^{\tau-1}\left(\log RV_t - \theta_0 - \sum_{i=1}^{p}\theta_i \log RV_{t-i}\right)^2\right\},\end{aligned}$$

where Θ denotes the parameter space and $L(\log RV; I_\tau, \theta)$ the local conditional log-likelihood function, for which we also use the short notation $L(I_\tau, \theta)$. We refer to the estimator $\tilde{\theta}_\tau$ as the *local ML estimator*.

The question now is how the interval of homogeneity I_τ can be determined. To this end likelihood ratio testing ideas are employed. Suppose that (log) *RV* is driven by an AR(p) process with a constant set of true parameters θ^*_τ at time point τ. The accuracy of estimation can be measured by the log-likelihood ratio (LR) (under homogeneity)

$$\mathrm{LR}(I_\tau, \tilde{\theta}_\tau, \theta^*_\tau) = L(I_\tau, \tilde{\theta}_\tau) - L(I_\tau, \theta^*_\tau). \tag{9}$$

Polzehl and Spokoiny (2006) derived a bound for LR and its power transformation $|\mathrm{LR}(I_\tau, \tilde{\theta}_\tau, \theta^*_\tau)|^r$ with $r > 0$ for an iid sequence of Gaussian innovations [in our case this refers to the innovations of the LAR(p) process]:

$$\mathsf{E}_{\theta^*_\tau}|\mathrm{LR}(I_\tau, \tilde{\theta}_\tau, \theta^*_\tau)|^r \le \xi_r. \tag{10}$$

This bound is nonasymptotic and can be applied to any interval I_τ. It allows to construct a confidence interval that can be used for testing homogeneity. The null hypothesis of time homogeneity means that the process follows the model (8) with a constant parameter, which implies that the ML estimator $\tilde{\theta}_\tau$ and the corresponding LR fulfill the risk bound (10). Therefore, the test of homogeneity can be performed, for example, by using the LR test statistic

$$|\mathrm{LR}(I_\tau, \tilde{\theta}_\tau, \theta^*_\tau)|^r.$$

In practice, the *hypothetical* AR(p) *parameters* θ^*_τ and also the risk bound ξ_r are unknown but can be computed empirically. Details on the feasible test procedure are given in the next section. In the estimation we are searching for an interval of *homogeneity* over which the process is well approximated by a parametric model. In other words, we mimic the unknown data-generating process by a local parametric model and simultaneously require that the modeling bias under this local parametric assumption is small. There exists a well-established theory addressing this local parametric assumption under a small modeling bias condition; see, e.g., Chen and Spokoiny (2010). Belomestny and Spokoiny (2007) shows that an optimal choice of an interval of local homogeneity can be obtained via an adaptive procedure. In the following, we concentrate on the construction details and its application to the dual view on the dependence structure of volatility. However, details of the results can be found in the cited literature and a comprehensive simulation study in Section 3.4 illustrates the performance of the adaptively selected estimators.

3.2 Adaptive Identification of the Interval of Homogeneity

This section presents a feasible adaptive selection algorithm of the interval of homogeneity for a particular point in time. Nevertheless, the procedure is general and is applied at every time point. The aim of the algorithm is to select the longest interval of homogeneity over which the parametric model is a good approximator for the process. The number of possible interval candidates is large, for example, the first interval may include just a few past observations and the intervals considered thereafter may be increased by just one observation in each step up to including all past observations. As this results in a large number of candidate intervals, it is practical to consider only a finite set of intervals $\mathbf{I}_\tau = \{I^1_\tau, \ldots, I^K_\tau\}$ with K candidates as suggested in Chen and Spokoiny (2010). For computational tractability, the intervals are increasingly ordered according to their length, that is, $I^1_\tau \subset \cdots \subset I^K_\tau$. To each interval there corresponds a local ML estimator, denoted by $\tilde{\theta}^k_\tau$ with $k = 1, \ldots, K$. In statistical learning theory those are called weak learners. Note that we are using the parametric assumption where the LAR model is only a good approximator of the process. Referring to the nonparametric smoothing literature, an increase in the length of intervals in (8) leads to an increase in modeling bias while the variance of the estimators is decreasing; see, for example, Härdle et al. (2004). In accordance with the chosen $\mathbf{I}_\tau$, the K weak learners therefore exhibit an increasing modeling bias and decreasing variance. Under the assumption that the interval of local homogeneity exists, the first interval I^1_τ is required to be short such that the modeling bias is small. Our interest here is to select an optimal estimator that has the smallest variance without violating the small modeling bias condition.

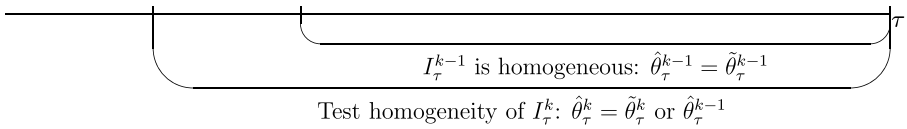

Figure 4. Sequential test of homogeneity: the longer interval I_τ^k is tested after the hypothesis of homogeneity over the shorter interval I_τ^{k-1} has been accepted.

The selection algorithm is based on a sequential testing procedure. The procedure starts from the shortest interval I_τ^1, over which local homogeneity holds by assumption. The weak learner $\tilde{\theta}_\tau^1$ is automatically accepted as an eligible local homogeneous estimator: $\hat{\theta}_\tau^1 = \tilde{\theta}_\tau^1$. Sequentially, at each step k with $2 \le k \le K$, we test the hypothesis of local homogeneity given that at the former step $k-1$ the null hypothesis has not been rejected; see Figure 4. The selected interval $\hat{I}_\tau$ corresponds to the largest accepted interval I_τ^k such that

$$|\,\mathrm{LR}(I_\tau^k, \tilde{\theta}_\tau^k, \hat{\theta}_\tau^{k-1})|^r \le \zeta_k, \tag{11}$$

where ζ_k is the critical value at step k and is described in more detail below. Note that this test (11) measures the difference of an estimator $\tilde{\theta}_\tau^k$ over a "possible" interval of local homogeneity I_τ^k to the most recently available optimal estimator $\hat{\theta}_\tau^{k-1}$. It differs from the LR test statistic implicitly linked to (10). Here the unknown hypothetical parameter θ_τ^* is replaced by the tentatively optimal estimator $\hat{\theta}_\tau^{k-1}$ since the latter is the possibly best estimator at the current step k. If there is no significant difference between the two estimators, it means that there is no significant change in the dynamics and the small modeling bias condition is not violated. We thus accept the null hypothesis of homogeneity and adopt the new estimator $\hat{\theta}_\tau^k = \tilde{\theta}_\tau^k$ as it has a smaller variance. On the other hand, if the test statistic is significant, it indicates that at least one structural change of the process exists and the LAR model is no longer a good approximator of the process. The sequential testing procedure terminates. This procedure then leads to the optimal estimator $\hat{\theta}_\tau$ that corresponds to the selected interval $\hat{I}_\tau$.

The formal definition of the procedure for a particular point in time τ is as follows:

1. Initialization: $\hat{\theta}_\tau^1 = \tilde{\theta}_\tau^1$.
2. $k = 2$: while $|\,\mathrm{LR}(I_\tau, \tilde{\theta}_\tau^k, \hat{\theta}_\tau^{k-1})|^r \le \zeta_k$ and $k \le K$,
$$k = k + 1,$$
$$\hat{\theta}_\tau^k = \tilde{\theta}_\tau^k.$$
3. Final estimate: $\hat{\theta}_\tau = \hat{\theta}_\tau^k$.

3.3 Choice of Parameters and Implementation Details

Clearly, the proposed procedure depends on a set of parameters, such as the lag order p in the LAR setup, the set of intervals, the power parameter r, and the critical values $\{\zeta_k\}_{k=1}^K$. In the following we address the choice of these parameters, and also discuss the computation of the critical values via Monte Carlo simulations.

3.3.1 Set of Intervals. We consider a finite set with $K = 13$ intervals in our study. This set is composed of the following interval lengths:

$$\{1\mathrm{w}, 1\mathrm{m}, 3\mathrm{m}, 6\mathrm{m}, 1\mathrm{y}, 1.5\mathrm{y}, 2\mathrm{y}, 2.5\mathrm{y}, 3\mathrm{y}, 3.5\mathrm{y}, 4\mathrm{y}, 4.5\mathrm{y}, 5\mathrm{y}\},$$

where w denotes a week (5 days), m refers to one month (21 days), and y to one year (252 days). In other words, $I_\tau^1 = [\tau - 1\mathrm{w}, \tau), I_\tau^2 = [\tau - 1\mathrm{m}, \tau), \ldots, I_\tau^{13} = [\tau - 5\mathrm{y}, \tau)$. This choice is motivated by the practical reason that investors are often concerned about special investment horizons. As the set $\mathbf{I}_\tau = \{I_\tau^k\}_{k=1}^{13}$ is used for each time point τ, we drop the subscript in the following for notational convenience. Other sets of intervals may be considered (see also Section 3.4). However, it is important to assure homogeneity over the shortest interval.

3.3.2 Selection of the Lag Order. While the lag selection in the (global) AR models is straightforward, it is more complicated in the LAR approach as the identification of the local intervals of homogeneity depends on the assumed lag order. The selection of the lag order p can be based, for example, on the minimum average value of the information criteria obtained from the log-likelihood values of the selected optimal estimators or on the minimum root mean square forecast errors. Depending on the number of lags, such a procedure may of course be computationally demanding (but still feasible).

Alternatively, we can exploit the flexibility of the LAR procedure, where the local parametric model, that is, the LAR(p) model, is only required to provide a good approximation of the true latent DGP over the interval of local homogeneity. The small modeling bias guarantees that the confidence set, built on the basis of the upper risk bound given in Equation (10), continues to hold with a slightly smaller coverage probability; see also Čížek, Härdle, and Spokoiny (2009). In other words, even if the assumed lag order of the LAR model is not the true one, but close to it, the procedure is appropriate. Section 3.4.3 addresses the issue of a wrong lag selection within a simulation study, which supports our expectation. To investigate the dual view on long memory, we therefore adopt in the empirical application the most extreme case of a short-memory model, that is, an AR(1) specification.

3.3.3 Parameter r. Belomestny and Spokoiny (2007) suggest to choose $r = 1/2$ in order to provide a stable performance and to minimize the computation error in the Monte Carlo simulation. We follow their recommendation in our empirical application. The sensitivity of the LAR procedure to different values of r is also assessed within a simulation study.

3.3.4 Critical Values. In the testing procedure, critical values measure the significance of ML estimators under the hypothesis of local homogeneity. The critical values are selected

using the general approach of testing theory: to provide a prescribed performance of the procedure under the null hypothesis. In particular, we generate global homogeneous processes, that is, AR(p) models with constant parameters in (8), ensuring homogeneity for every past interval. The critical values are then selected so that the ML estimators under homogeneity fulfill the risk bound (10) over each interval.

As an illustration, we calculate critical values for LAR(1) based on 100,000 generated AR(1) processes with $\theta_t = \theta^* = (\theta_0^*, \theta_1^*, \sigma^*)^\top$ for all t:

$$y_t = \theta_0^* + \theta_1^* y_{t-1} + \varepsilon_t, \qquad \varepsilon_t \sim \mathrm{N}(0, \sigma^{*2}).$$

The starting value is set to $y_0 = \theta_0^*/(1-\theta_1^*)$. The sample size of each process is 1261 in correspondence to the largest interval of $I^K = I^{13} = 5\mathrm{y}$. Under homogeneity, the ML estimator with respect to the largest interval is the optimal estimator (with the smallest variance among others), that is, $\hat\theta_t = \hat\theta_t^K = \tilde\theta_t^K$. Given a reasonable set of critical values, the risk bound (10) holds over the longest interval of homogeneity

$$\mathsf{E}_{\theta^*}\left|\mathrm{LR}\left(I^K, \tilde\theta_t^K, \hat\theta^K_{t(\zeta_1,\ldots,\zeta_K)}\right)\right|^r \le \xi_r. \tag{12}$$

We mimic here the environment of the sequential testing by replacing the unknown hypothetical AR(p) parameter θ^* with the most recently available optimal estimator $\hat\theta_t^k$. In addition, we use the notation $\hat\theta^k_{t(\zeta_1,\ldots,\zeta_k)}$ to emphasize that the adaptively selected estimator depends on the critical values $\{\zeta_1, \ldots, \zeta_k\}$. The bound $\xi_r = \mathsf{E}_{\theta^*}|\,\mathrm{LR}(I^K, \tilde\theta_t^K, \theta^*)|^r$ is empirically calculated. We also notice that the sequential testing procedure accumulates uncertainty in estimation due to the increase in the degrees of freedom. To take this into account, a condition similar to (12) is imposed at each step:

$$\mathsf{E}_{\theta^*}\left|\mathrm{LR}\left(I^k, \tilde\theta_t^k, \hat\theta^k_{t(\zeta_1,\ldots,\zeta_k)}\right)\right|^r \le \frac{k-1}{K-1}\xi_r, \qquad k = 1, \ldots, K. \tag{13}$$

The sequential testing procedure is adopted to compute the critical values. At step $k = 1$, we set $\zeta_1 = \infty$ in agreement with the local homogeneity in the shortest interval I^1 leading to $\hat\theta_t^1 = \tilde\theta_t^1$. In the computation of ζ_2 we set all the remaining $\zeta_k = \infty$ for $k \ge 3$ to specify the contribution of ζ_2 and choose the minimal value of ζ_2 that delivers the estimator satisfying the following risk function:

$$\mathsf{E}_{\theta^*}\left|\mathrm{LR}\left(I^k, \tilde\theta_t^k, \hat\theta^k_{t(\zeta_1,\zeta_2)}\right)\right|^r \le \frac{1}{K-1}\xi_r, \qquad k = 2, \ldots, K.$$

Consequently with $\zeta_1, \zeta_2, \ldots, \zeta_{k-1}$ fixed, we select the minimal value of ζ_k for $k = 3, \ldots, K$ which fulfills

$$\mathsf{E}_{\theta^*}\left|\mathrm{LR}\left(I^q, \tilde\theta_t^q, \hat\theta^q_{t(\zeta_1,\zeta_2,\ldots,\zeta_k)}\right)\right|^r \le \frac{k-1}{K-1}\xi_r, \qquad q = k, \ldots, K.$$

3.3.5 Hypothetical Parameters. Clearly, critical values also depend on the hypothetical parameters θ^* used for generating the homogeneous processes. In our study, we consider two ways for selecting θ^*: a global selection where θ^* is estimated over the full sample period or an adaptive selection where θ^* is reestimated at each time point using a rolling window with a fixed length. For the adaptive selection, a large rolling window size means that we put more attention to a time homogeneous situation. Such a choice leads to a rather conservative procedure with possibly low accuracy of estimation. On the contrary, a rolling window including fewer observations is more sensitive to structural shifts. Alternatively, the size of rolling window can be selected in a data driven way by minimizing some objective function, for example, by minimizing the forecast error, which is however computationally more intensive. In our empirical analysis we consider the predictive performance of the LAR procedure using both the global selection scheme as well as the adaptive selection based on rolling windows of 1 month, 6 months, 1 year, and 2.5 years. As expected, using the time dependent critical values (slightly) increases the accuracy of prediction.

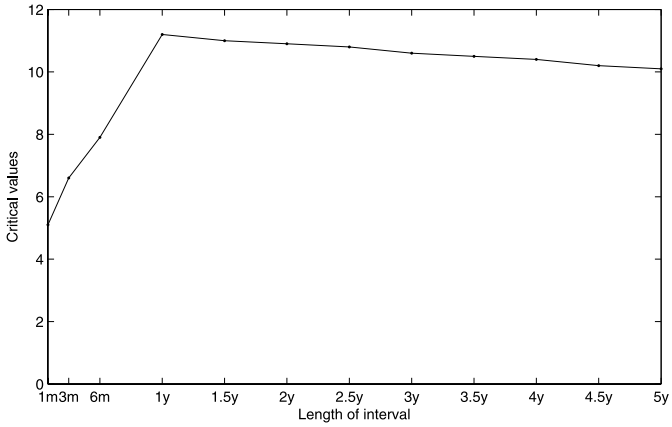

Figure 5. The set of critical values for LAR(1) model. They are based on $r = 1/2$ and on $\theta^* = (-0.1156, 0.7827, 0.5525)^\top$, which are calculated for the log realized volatility of the S&P500 index futures under the hypothesis of constant parameters in (8). The set of interval lengths is given on the x-axis.

Figure 5 depicts the global critical values calculated for a LAR(1) model with $r = 1/2$, the interval candidates given in Section 3.3.1 and the hypothetical AR(p) parameter $\theta^* = (-0.1156, 0.7827, 0.5525)^\top$, the estimates of an AR(1) model fitted to our real dataset—the logarithm of realized volatility of the S&P500 index data.

3.4 Simulation Experiments and Sensitivity Analysis

This section investigates the performance of the localized RV approach in a number of simulation studies focusing on the LAR(1) model. In particular, we assess its performance under different types of structural breaks, we analyze the impact of the parameters involved in the adaptive technique, and we assess the issue of model misspecification, such as a wrong lag selection.

3.4.1 Parameter Changes. In the following we consider the performance of the LAR(1) approach under various scenarios. Specifically, we simulate from an AR(1) with suddenly and gradually changing parameters in order to investigate the appropriateness of the LAR approach under different types of changes. The actual values of the parameters are again based on the estimates of an AR(1) model fitted to the full S&P500 realized volatility data. In each scenario, only one parameter varies over time while the other two remain constant. The processes of the changing parameters are displayed in Figures 6 to 8. The character S denotes a scenario with sudden changes of parameters, where big changes occur at time points $t = 1501$ and

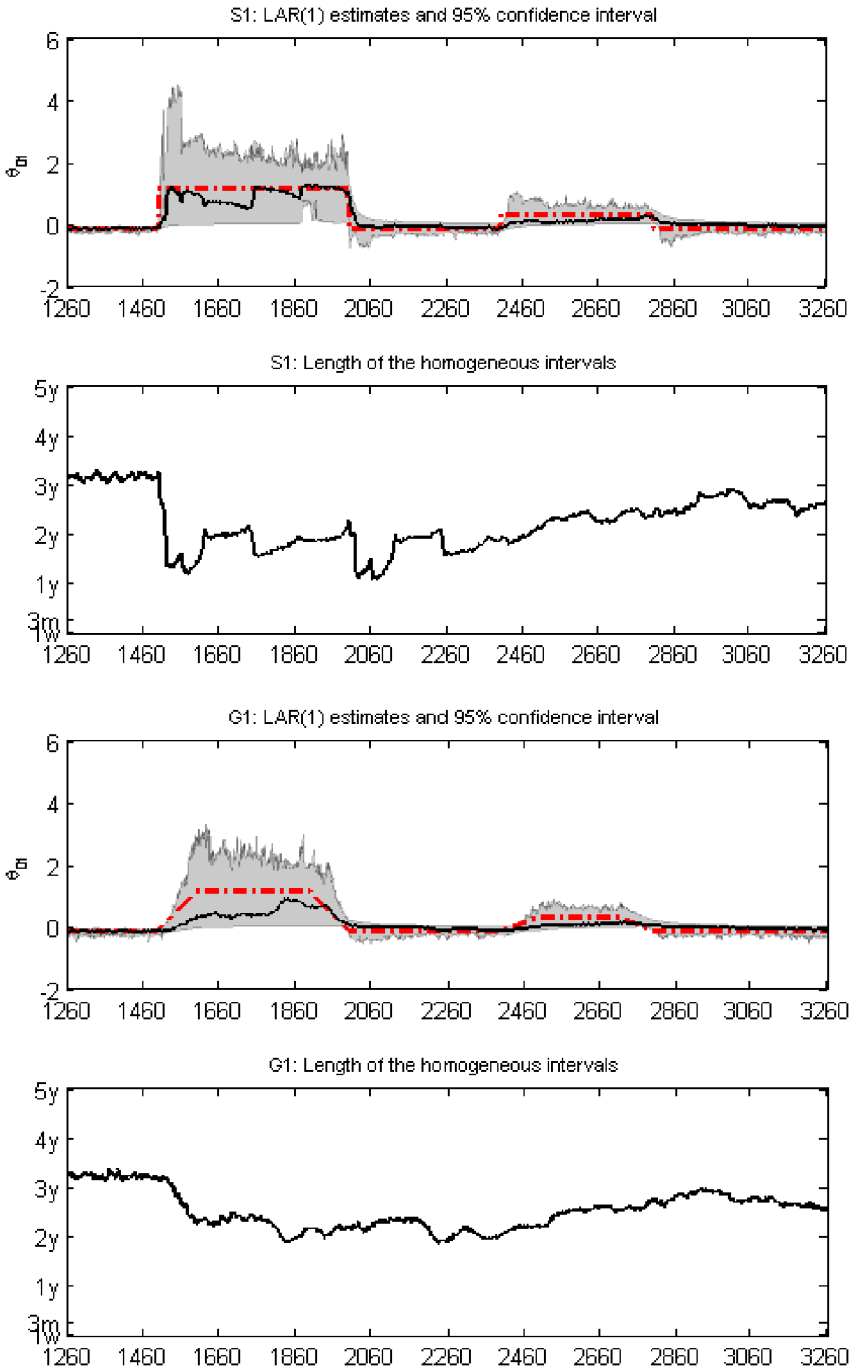

Figure 6. Simulation results for scenarios S1 and G1 (changing parameter: θ_{0t}). The red dashed line represents the process of the true time-varying parameter (S1: $\theta^*_{0t} = 1.1557$ for $t \in [1501, 2000]$, 0.3467 for $t \in [2401, 2800]$, -0.1156 otherwise) and the bold solid line gives the average value of the estimated parameter over 500 simulations. The shaded area corresponds to the pointwise 95% confidence intervals. The average values of the selected homogeneous intervals for each time point are presented in the bottom panel of each scenario. The online version of this figure is in color.

$t = 2000$ and small ones at $t = 2401$ and $t = 2800$, respectively. The G scenarios, on the contrary, denote gradual changes where the parameter gradually reaches to a new level within 100 steps after the change point. For example, in scenario G2 the autoregressive parameter θ_{1t} gradually changes from 0.7827 to -0.7827 over the period from 1501 to 1600, stays at the new level until it drops gradually back to 0.7827 over the period from 2001 to 2100. Similarly the small gradual changes occur over the periods $[2401, 2500]$ and $[2801, 2900]$. For each scenario, we generate 500 LAR(1) processes with 3261 observations. The first 1261 observations, corresponding to the largest interval $I^{13} = 5$ years, are used as training set.

The average value of the estimated parameters (solid line) and the pointwise 95% confidence intervals (shaded areas) are displayed in Figures 6 to 8 along with the true values of θ^* (dashed line). For each point in time the average value of the selected homogeneous intervals is also presented. Obviously, the selected homogeneous intervals are long when the parameters

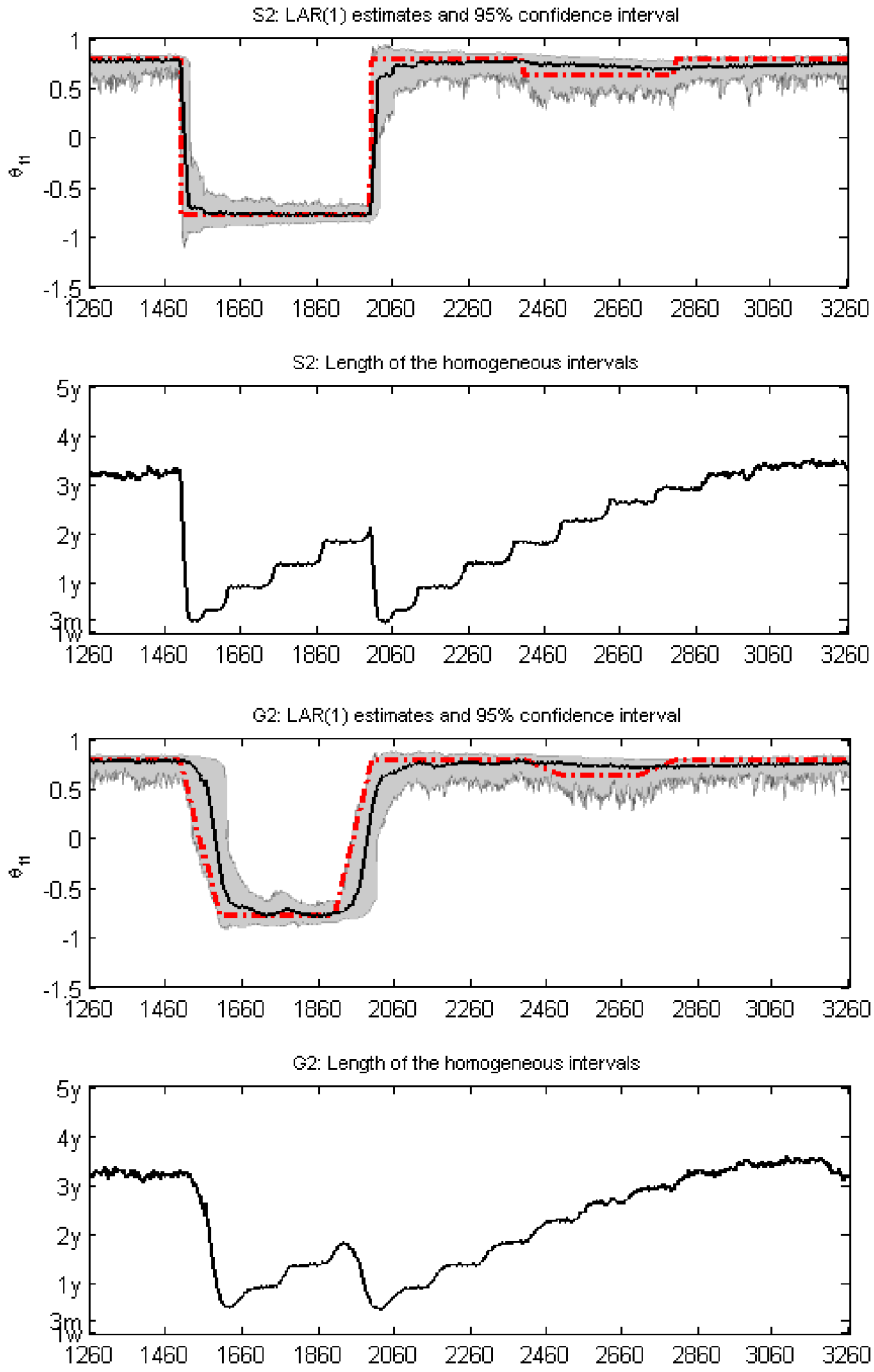

Figure 7. Simulation results for scenarios S2 and G2 (changing parameter: θ_{1t}). The red dashed line represents the process of the true time-varying parameter (S2: $\theta_{1t}^* = -0.7827$ for $t \in [1501, 2000]$, 0.6261 for $t \in [2401, 2800]$, 0.7827 otherwise) and the bold solid line gives the average values of the estimated parameter over 500 simulations. The shaded area corresponds to the pointwise 95% confidence intervals. The average value of the selected homogeneous intervals for each time point are presented in the bottom panel of each scenario. The online version of this figure is in color.

are constant over a long past time interval, but decline sharply when shifts occur. It indicates that the LAR procedure selects reasonable intervals of homogeneity.

In order to assess the performance of the local procedure in more detail, we additionally compute the detection speed, that is, the number of periods required for reaching 50% and 75% of the new level of the parameter. In the G scenarios the counting starts once the parameter has reached its new level, that is, after the gradual changes have finished. In the S scenarios the counting starts immediately from the change point. Table 2 reports the results. In general, the adaptive procedure works well. It shows that the procedure reacts quickly to a big shift, but slowly to a small shift. For example in the scenario S2, where the AR coefficient θ_{1t} jumps at $t = 1501$, the technique only needs 12 periods to catch up 50% of the big shift, while for the small shift at $t = 2401$ it takes 213 periods. This finding, however, is quite reasonable. After a small change of the parameters of the DGP the simulated observations may still

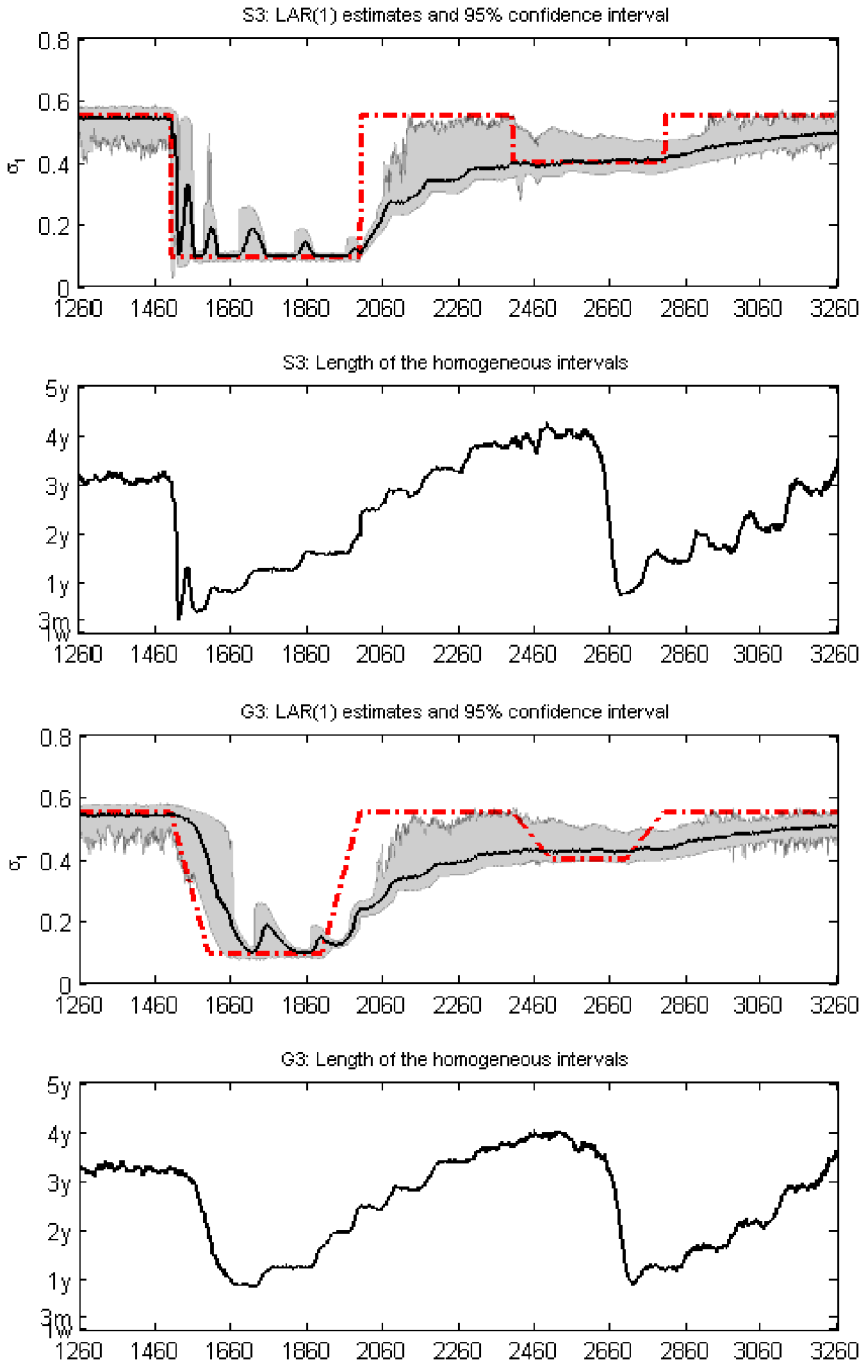

Figure 8. Simulation results for scenarios S3 and G3 (changing parameter: σ_t). The red dashed line represents the process of the true time-varying parameter (S3: $\sigma_t^* = 0.1000$ for $t \in [1501, 2000]$, 0.4000 for $t \in [2401, 2800]$, 0.5525 otherwise) and the bold solid line gives the average values of the estimated parameter over 500 simulations. The shaded area corresponds to the pointwise 95% confidence intervals. The average value of the selected homogeneous intervals for each time point are presented in the bottom panel of each scenario. The online version of this figure is in color.

be very close to those of the previous DGP and it is therefore hard for the procedure to differentiate between the two processes. In this case, more observations from the new DGP are needed for the identification of the parameter change. Nevertheless, the technique is able to detect the changes as more and more small shifts accumulate over time. Similar patterns are observed in the G scenarios which correspond to many small subsequent shifts. The results for the scenarios of σ_t further show that positive shifts, corresponding to an increase in the signal-to-noise ratio, can be only slowly detected; see also Figure 8.

3.4.2 Impact of Parameters. In the following we investigate the effect of the choice of parameters on the performance of the LAR procedure. Here we compute the detection speed of the LAR approach based on different sets of intervals $\{I_k\}_{k=1}^K$, different values of the power transformation parameter r and of the hypothetical AR(p) parameters θ^* used in the computation of the critical values. Moreover, we compute the root

Table 2. Detection speeds for the different scenarios

	S1		S2		S3			G1		G2		G3	
t	50%	75%	50%	75%	50%	75%	t	50%	75%	50%	75%	50%	75%
1501	21	23	12	18	18	20	1601	207	232	1	4	12	56
2001	9	19	13	19	169	>400	2101	1	1	1	6	88	>400
2401	66	374	213	>400	1	1	2501	169	>400	183	>400	1	1
2801	20	243	56	>400	293	>400	2901	1	166	1	346	176	>400

NOTE: Reported are the number of steps required for reaching 50% and 75% of the parameter change.

mean square forecast errors (RMSFEs) of LAR forecasts based on different choices of these parameters. The results are compared to our "default" case, where the parameters are set to the suggested values in Section 3.3, that is, the interval set is given by $\mathbf{I}_\tau = \{I_\tau^k\}_{k=1}^{13} = \{1\text{w}, 1\text{m}, 3\text{m}, 6\text{m}, 1\text{y}, 1.5\text{y}, 2\text{y}, 2.5\text{y}, 3\text{y}, 3.5\text{y}, 4\text{y}, 4.5\text{y}, 5\text{y}\}$, $r = 1/2$ and the vector of hypothetical AR(1) parameters $\theta^* = (-0.1156, 0.7827, 0.5525)^\top$. For the ease of exposition we only report the results for the scenarios with changes in the autoregressive parameter, that is, S2 and G2, as those are also particularly interesting in the model misspecification analysis discussed later. In particular, we consider two alternative sets of intervals. In order to assess the impact of the maximum length of the intervals, we truncate the default set of intervals at $K = 9$ (corresponding to 3 years), while the second scenario aims at investigating the sensitivity of the procedure towards a finer grid of intervals by including more intermediate subintervals, that is, introducing a three-months grid such that $K = 22$. We further evaluate the impact of a smaller value and a larger value of the power transformation parameter setting $r = 1/3$ and $r = 1$. As the critical values rely on the choice of the hypothetical parameters, we check the predictive performance using $80\%\theta^*$ and $120\%\theta^*$ to generate the homogeneous processes in the computation of critical values, which can be interpreted as an underestimation and overestimation of the actual parameter values, respectively.

Table 3 presents the results. In order to facilitate the comparison, we report here the relative average RSMFE of one-step ahead predictions, that is,

$$\text{R-RMSFE} = \sum_{j=1}^{500} \text{R-RMSFE}_j^{\text{nondefault}} \Big/ \sum_{j=1}^{500} \text{R-RMSFE}_j^{\text{default}},$$

where the average value of the RMSFEs with default choice is 0.5411 for S2 and 0.5374 for G2. We also define the relative detection speed as the difference of the average detection speed of the LAR procedure based on nondefault parameters to the average detection speed using the default choice. Thus, a positive/negative value indicates a slower/faster reaction of the technique with nondefault choices. The results illustrate well, that the LAR procedure is quite robust to the choice of the parameters. The "worst" cases appear when CVs are calculated based on imprecise hypothetical AR(p) parameters: a 2.74% improved predictability for $0.8\theta^*$ and a 3.95% worse performance for $1.2\theta^*$. It suggests that using alternative choices of the parameters delivers only small deviations from the default choices. Moreover, there are no crucial changes in the detection speed in the presence of large parameter changes, although

Table 3. Sensitivity analysis: impact of parameters

	Choice of parameters											
	$K = 9$		$K = 22$		$r = 1/3$		$r = 1$		$0.8\theta^*$		$1.2\theta^*$	
					Scenario S2							
R-RMSFE:	0.9956		1.0128		1.0102		0.9974		0.9726		1.0395	
R-DS:	50%	75%	50%	75%	50%	75%	50%	75%	50%	75%	50%	75%
$t = 1501$	−1	−1	2	2	0	0	2	2	−1	−4	6	5
$t = 2001$	0	0	3	0	1	0	2	0	−3	0	5	1
$t = 2401$	0	–	5	–	5	–	0	–	0	–	47	–
$t = 2801$	>344	–	>344	–	>344	–	>344	–	>344	–	>344	–
					Scenario G2							
R-RMSFE:	0.9946		1.0132		1.0143		0.9922		0.9728		1.0506	
R-DS:	50%	75%	50%	75%	50%	75%	50%	75%	50%	75%	50%	75%
$t = 1601$	0	0	0	0	0	0	0	0	0	0	0	4
$t = 2001$	0	0	0	0	0	0	0	0	0	0	0	6
$t = 2501$	0	–	0	–	0	–	0	–	0	–	184	–
$t = 2801$	>399	>54	>399	>54	>399	>54	>399	>54	>399	>54	>399	>54

NOTE: Reported are the relative one-step-ahead RMSFEs and the relative detections speeds (R-DS) in the scenarios S2 and G2 for different choices of the parameters. The default choice (i.e., the benchmark) is given by $K = 13$, $r = 1/2$, and $\theta^* = (-0.1156, 0.7827, 0.5525)^\top$. In the alternative choices only one parameter is changed fixing the other ones to the default values. $K = 9$ corresponds to a scenario with reduced maximum interval length while $K = 22$ is characterized by a finer grid of intervals. More details are given in the text. Scenarios S2 and G2 are displayed in Figure 7. "–" indicates that the detection speeds in both scenarios are greater than 400.

for small parameter changes the detection speed slows down. In general, the sensitivity analysis supports our default choice of parameters and the results suggest that for an adaptive, data-driven computation of the critical values the selection of the parameters may become even less important with respect to predictability.

3.4.3 Model Misspecification. In this section we investigate the robustness of the LAR procedure towards model misspecification, that is, if the true DGP has a different lag structure than the assumed one or, even worse, if the true DGP follows a different dynamic structure. The analysis is twofold: we first focus on short-memory models, which allow us to evaluate the impact of the lag order, and then consider the performance of the LAR procedure if the true DGP is a long-memory process.

For the short-memory scenarios we consider the local constant model, that is, $y_t = \theta_{0t} + \varepsilon_t, \varepsilon_t \sim \mathrm{N}(0, \sigma_t^2)$ and the LAR model with lag order $p = 2, 5$, and 10, $y_t = \theta_{0t} + \sum_{j=1}^{p} \theta_{jt} y_{t-j} + \varepsilon_t, \varepsilon_t \sim \mathrm{N}(0, \sigma_t^2)$, in order to account for the situations where the true DGPs either have less or more lags than is assumed in the LAR(1). The actual parameter values are again set to the ML estimates obtained by fitting the globally constant version of these models to the full S&P500 data sample. The design of the time variation in the parameters is similar to the scenarios in Figures 6 to 8, where the big and small changes are determined by a new level of the parameters (e.g., $-1\theta_p$ and $0.8\theta_p$ respectively). As the focus is on the impact of a misspecification in the lag order, we consider here only cases with changes in θ_{0t} in the local constant model and changes in θ_{pt}, the pth autoregressive part of the LAR(p) model. In the long-memory scenario we simulate from an ARFIMA$(2, 0.47, 0)$ with constant parameters. The specification of the ARFIMA model is also guided by the empirical results obtained for the full S&P500 sample; see Section 4. For each DGP, 500 series are simulated, each with a length of 3261 observations.

The sensitivity of the LAR procedure towards model misspecification is assessed in terms of predictability. In particular, we compute for each simulated series 2000 one-step-ahead forecasts based on: (i) the "wrong" but flexible LAR(1) approach, where the intervals of local homogeneity are selected by using the adaptive technique; (ii) the true data-generating model using optimally time-varying window size. More precisely, for a particular time point the optimal window is either identified using the LAR(p) procedure (for LAR scenarios) or assumed to be the interval used for generating the process (otherwise). In addition, the shortest/longest length of the intervals is set to include 15/1250 observations, which is in line with the assumption of homogeneity in the LAR procedure and assures the feasibility of estimation. The average value of the RMSFEs for different scenarios are reported in Table 4. In most cases, the forecasts based on the LAR(1) specification yield only slightly bigger RSMFEs than the true DGPs. It supports that the LAR procedure with the lag order $p = 1$ can provide a quite accurate approximation. In other words, the LAR procedure is quite robust to the selection of the lag order p. Moreover, the LAR(1) performs also well if the true source of the long-range dependence is a long-memory process, confirming that long memory can well be approximated by a short-memory model with breaks. In summary, the simulation shows that the local adaptive procedure with lag order $p = 1$ is a reasonable approximation, even if the underlying process deviates from the LAR(1) setup.

Table 4. Sensitivity analysis: model misspecification

DGP:	Local const. θ_{0t}	LAR(2) θ_{2t}	LAR(5) θ_{5t}	LAR(10) θ_{10t}	ARFIMA
$\widehat{\mathrm{DGP}}$	1.0225	0.6247	0.5664	0.5568	0.5105
LAR(1)	0.9339	0.6293	0.5848	0.5724	0.5619

NOTE: Reported are the average RMSFEs based on the LAR(1) procedure and the estimated data-generating processes, $\widehat{\mathrm{DGP}}$.

4. ALTERNATIVE MODELS

As we aim at a comparison of the LAR procedure to the long memory view of volatility, we primarily consider alternative models that emanate from this view. Nevertheless, we also compare our procedure to the smooth transition regression tree (STR-Tree) model, that is, a model with breaks.

The ARFIMA model is one of the standard models used in the realized volatility literature; see, for example, Andersen et al. (2003). Under the ARFIMA(p, d, q) model, the dynamics of logarithmic realized volatility is given by

$$\phi(L)(1-L)^d(\log RV_t - \mu) = \psi(L)u_t, \tag{14}$$

with $\phi(L) = 1 - \phi_1 L - \cdots - \phi_p L^p$, $\psi(L) = 1 + \psi_1 L + \cdots + \psi_q L^q$, L denoting the lag operator, and $d \in (0, 0.5)$ is the fractional difference parameter. Given the empirical distributional properties of logarithmic realized volatility, u_t is usually assumed to be a Gaussian white noise process, which facilitates the exact maximum-likelihood estimation of the model.

The HAR model aims at reproducing the observed volatility phenomenon. However, in contrast to the ARFIMA model, the HAR model is formally not a long-memory model. Instead, the correlation structure is approximated by the sum of a few multiperiod volatility components. The use of such components is motivated by the existence of heterogenous agents having different investment horizons; see Corsi (2009) and Müller et al. (1997). In particular, the HAR model put forward by Corsi (2009) builds on a daily, weekly, and monthly component, which are defined by

$$RV_{t+1-k:t} = \frac{1}{k}\sum_{j=1}^{k} RV_{t-j}$$

with $k = 1, 5, 21$, respectively. The HAR model is then given by

$$\log RV_t = \alpha_0 + \alpha_d \log RV_{t-1} + \alpha_w \log RV_{t-5:t-1} + \alpha_m \log RV_{t-21:t-1} + u_t \tag{15}$$

with u_t typically being also Gaussian white noise. Maximum-likelihood estimation is straightforward. Interestingly, the HAR and ARFIMA models have been found to obtain a similar forecasting performance with both models outperforming the traditional volatility models based on daily returns; see, for example, Andersen, Bollerslev, and Diebold (2007) and Koopman, Jungbacker, and Hol (2005).

It is sometimes argued that volatility exhibits both long memory and structural breaks. We therefore compare our procedure

also to the adaptive ARFIMA model, that has recently been developed in Baillie and Morana (2009a) for modeling inflation dynamics. The model is based on a time-dependent intercept that is given by a Flexible Fourier Form representation, and an innovation term that follows a stationary long-memory process. The flexible functional form of the intercept allows for smooth as well as sharp nonlinearities without the need to identify break points and the magnitude of the breaks. Baillie and Morana (2009b) have shown that a FIGARCH model with such a time dependent intercept provides superior volatility forecasts in comparison to alternative GARCH and adaptive GARCH specifications. We therefore adopt the adaptive ARFIMA model for modeling logarithmic realized volatility, which is given by

$$\log RV_t = \mu + \sum_{j=1}^{k} (\sin(2\pi jt/T) + \delta_j \cos(2\pi jt/T)) + u_t, \quad (16)$$

where

$$\phi(L)(1-L)^d u_t = \psi(L)\epsilon_t.$$

It is characterized as A-ARFIMA(p, d, q, k).

The STR-Tree model proposed in da Rosa, Veiga, and Medeiros (2008) provides an interesting alternative to the LAR procedure. It builds on the methodology of classification and regression tress, where it is assumed that the dependent variable is given by the sum of regression models, each of which is determined by recursive partitions of the covariate space. The structure of a regression tree model is usually represented in the format of a binary choice decision tree with a set of parent and terminal nodes, denoted here by $\mathbb{J}$ and $\mathbb{K}$, respectively. The splits at the parent nodes are sharp. The STR-Tree model instead smoothes the splits by replacing the indicator function by a logistic function

$$G(x; \gamma, c) = \frac{1}{1 + e^{-\gamma(x-c)}}.$$

Scharth and Medeiros (2009) advocate the use of the STR-Tree approach for modeling logarithmic realized volatility, that is,

$$\log RV_t = \boldsymbol{\alpha}^\top \mathbf{w}_t + \sum_{k \in \mathbb{K}} \boldsymbol{\theta}_k^\top \mathbf{z}_t B_{\mathbb{J}k}(\mathbf{x}_t, \boldsymbol{\beta}_k) + u_t, \quad (17)$$

where $\mathbf{x}_t = (x_{1,t}, \ldots, x_{m,t})^\top$ denotes the vector of explanatory variables, or in terms of the smooth transition literature the so-called transition variables, $\mathbf{z}_t = (\log RV_{t-1}, \ldots, \log RV_{t-p})^\top$, and $\mathbf{w}_t$ is a vector of linear regressors that are unaffected by the tree, $\mathbf{w}_t \nsubseteq \mathbf{x}_t$. Moreover,

$$B_{\mathbb{J}k}(\mathbf{x}_t, \boldsymbol{\beta}_k) = \prod_{j \in \mathbb{J}} G\big(x_{s_j,t}; \gamma_j, c_j\big)^{n_{k,j}(1+n_{k,j})/2} \times \big[1 - G\big(x_{s_j,t}; \gamma_j, c_j\big)\big]^{(1-n_{k,j})(1+n_{k,j})}, \quad (18)$$

where $s_j \in \{1, \ldots, m\}$ gives the transition variable being relevant at node j and

$$n_{k,j} = \begin{cases} -1 & \text{if parent node } j \text{ is not included} \\ & \text{in the path to terminal node } k \\ 0 & \text{if the right-child node of parent node } j \\ & \text{is included in the path to terminal node } k \\ 1 & \text{if the left-child node of parent node } j \\ & \text{is included in the path to terminal node } k. \end{cases}$$

The spirit of the STR-Tree model is similar to the LAR procedure in the sense that realized volatility is approximated by local AR(p) models. However, in the STR-Tree model the regimes are due to partitions of the transition variables, such as lagged returns (capturing the well-known leverage effect), which are determined globally, that is, over the full sample period. The LAR instead is more flexible, as the interval of homogeneity is determined locally. Moreover, it does not require the specification of a set of variables that may lead to parameter changes. In fact, any event or changes in variables that affect the parameters of the AR(p) model such that local homogeneity is rejected are automatically encountered in the procedure.

5. EMPIRICAL ANALYSIS

We now turn to the empirical investigation of the dual views on the dynamics of volatility. We focus our analysis on realized volatility of the S&P500 index futures from January 2, 1985 to February 4, 2005 (see Section 2). Like in the simulation exercise we use the first 5 years of our sample as a training set. For the local autoregressive procedure this means that January 2, 1990 is the first time point for which we estimate the LAR model and that we allow the longest interval of homogeneity ($K = 13$) to be 5 years with the remaining set of subintervals given as in the Section 3.3, that is, 1 week ($k = 1$), 1 month ($k = 2$), …, 4.5 years ($k = 12$).

The estimation of the LAR model is conducted for different sets of critical values, in order to assess also the empirical sensitivity of the approach with respect to the choice of the critical values. We therefore consider critical values obtained from a Monte Carlo simulation based on the parameter values of the AR model being estimated over the full sample period. We refer to this as the global LAR model. The other sets of critical values are obtained adaptively using a 1 month, 6 months, 1 year, and 2.5 years sample period. Figure 9 shows the distribution of the lengths of the selected homogenous intervals of the LAR(1) model over the evaluation period (January 2, 1990 to February 4, 2005) based on the global and the adaptive

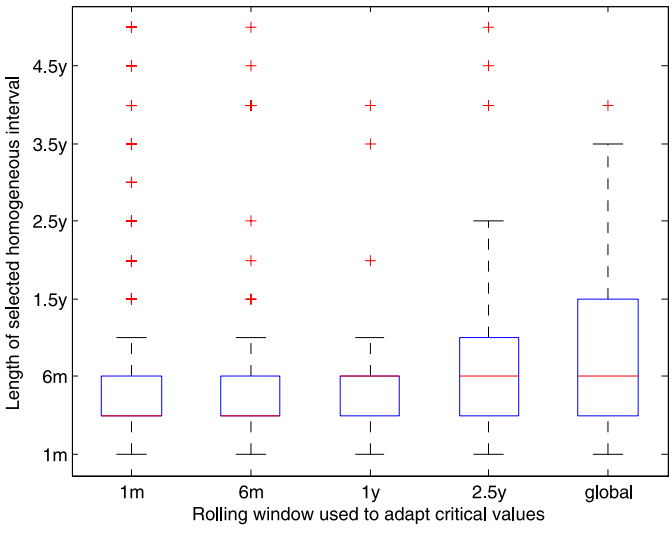

Figure 9. Boxplot of the homogenous intervals selected by the LAR(1) procedure with 1 month, 6 months, 1 year, 2.5 years adaptive critical values and the global LAR(1) procedure.

critical values. Obviously, the global LAR(1) model exhibits a slightly higher variation in the length of the selected intervals. Interestingly, with the exception of the adaptive 1 month and 6 months LAR(1) models for which the median interval length is at $k = 3$, we find that the median is $k = 4$, which corresponds to 6 months of homogeneity. Furthermore, note that the average interval length is for nearly all LAR(1) models about 6 months, which indicates only a weak sensitivity of the interval selection procedure to the sample size used in the computation of the critical values.

In our analysis we assess the forecasting performance for several periods into the future. Such multiperiod predictions may seem to be at odds with the idea of the LAR procedure, which builds on local homogeneity. Local homogeneity has the advantage that forecasts are based only on the most recent information being relevant at the particular forecast origins. But for iterative long-term predictions it also implies that the procedure may perform poor as for increasing forecast horizons it becomes more likely that the assumption of local homogeneity is violated. Nevertheless, the advantage of local homogeneity can also be transferred to the case of multiperiod predictions by incorporating the forecast horizon into the adaptive selection via a restricted LAR(h) specification:

$$\log RV_{t+h} = \theta_{0t} + \theta_{ht} \log RV_t + \varepsilon_t, \qquad \varepsilon_t \sim \mathrm{N}(0, \sigma_t^2), \quad (19)$$

which leads to a direct forecasting approach. We adopt this specification in the empirical analysis.

Table 5 presents the RMSFEs of the LAR model for the 1-day, 5-days, and 10-days ahead forecasts using the different sets of critical values. The empirical results reveal that an adaptive approach and a reduction of the sample period underlying the computation of the critical values introduces more flexibility into the procedure, which seems to result in an increase in forecast accuracy.

We investigate the dual views by comparing the forecasting performance of the LAR procedure to the alternative models. To this end, we recursively compute (logarithmic) realized volatility forecasts from all model types over the evaluation period. Moreover, as we have observed different degrees of persistence in log realized volatility for different lengths of the sample period (see Figure 1), we consider for each of the alternative models forecasts conditional on different information sets, that is, different sample sizes.

Table 5. Root mean square forecast errors of the LAR model based on different sets of critical values

Sample size used in the critical values	$h = 1$	$h = 5$	$h = 10$
1m	0.4823	**0.4619**	**0.4615**
6m	**0.4791**	0.4791	0.4873
1y	0.4842	0.4881	0.4945
2.5y	0.4898	0.5027	0.5056
Global	0.4986	0.5660	0.5884

NOTE: The table reports the root mean square forecast errors (RMSFE) of the h-day ahead logarithmic realized volatility forecasts of the S&P500 index futures based on the LAR(h) models. The first column refers to the information set that is used in the computation of the critical values. For example, the number reported in the first upper-left cell gives the RMSFE of forecasts based on the LAR(1) approach with critical values being computed adaptively over the previous month. Global indicates that the critical values have been computed based on the full sample. Bold numbers indicate the minimum RMSFE for each forecast horizon.

More precisely, forecasts of the ARFIMA, adaptive ARFIMA, and HAR models are based on a rolling window scheme, with rolling window sizes ranging from 3 months to 5 years, which is broadly consistent with our choice of subintervals in the LAR procedure. The conditioning on the different sample sizes is also an attempt to account for the possibility that both long memory and structural breaks are driving volatility. For the STR-Tree model we follow Scharth and Medeiros (2009), and form forecasts based on the recursive scheme. We additionally compute forecasts from constant AR models conditional on the set of rolling windows used also in the HAR and ARFIMA models, as this allows for a direct evaluation of the relevance of the local selection of the interval length employed in the LAR procedure. Such an evaluation requires that forecasts from AR models are also based on the direct forecasting approach.

The forecasts of the other models are computed iteratively, such that their specifications remain the same for all forecast horizons. In particular, the ARFIMA forecasts are based on an ARFIMA$(2, d, 0)$ specification, which was selected according to the Akaike as well as the Bayesian information criteria using the full sample period. For the adaptive ARFIMA model we obtain an A-ARFIMA$(1, d, 1, 2)$ specification with $\gamma_2 = 0$. Estimation and forecasting is carried out using the Ox ARFIMA 1.04 package; see Doornik and Ooms (2004, 2006). For the STR-Tree model we consider the daily lagged return as the transition variable in order to account for the most popular leverage specification. Moreover, for consistency with the short-memory models considered in this paper, we set $p = 1$, and let only the AR(1) coefficient be affected by the tree as indicated by statistical tests on the relevance of explanatory variables in the tree based on the full sample period. Over this period the model is characterized by two splits. In computing the forecasts we respecify the tree structure and reestimate the model every period. We are grateful to Marcel Scharth for providing us with his code. Multistep forecasts are based on conditional simulations as explained in the appendix of Scharth and Medeiros (2009).

For the ease of exposition we do not report all forecasting results but instead focus only on those models that yielded the minimal RMSFE within each model class. Table 6 thus reports the RMSFE of the "best" models along with the corresponding conditioning information set for which the forecasts have

Table 6. Root mean square forecast errors and information sets of the best models

	$h = 1$		$h = 5$		$h = 10$	
Model	RMSFE	Info set	RMSFE	Info set	RMSFE	Info set
LAR	0.4791	6m	0.4619	1m	0.4615	1m
AR	0.5047	3m	0.5712	3m	0.5873	3m
STR-Tree	0.5547	Rec.	0.7746	Rec.	0.8738	rec.
ARFIMA	0.4991	3y	0.5827	3y	0.6207	3y
A-ARFIMA	0.5020	4.5y	0.5904	4.5y	0.6312	4y
HAR	0.5014	3y	0.5848	2.5y	0.6232	2.5y

NOTE: The table reports the root mean square forecast errors (RMSFE) of the h-day ahead logarithmic realized volatility forecasts of the S&P500 index futures based on the various models. Reported are the results for the models yielding minimal RMSFE within each model class. "Info set" refers to the corresponding sample size used in the computation of the critical values (for the LAR procedure) or to the size of the rolling window used in model estimation and prediction (for the AR, ARFIMA, and HAR models). "Rec." refers to forecasts based on the STR-Tree model, for which the recursive forecasting scheme is employed.

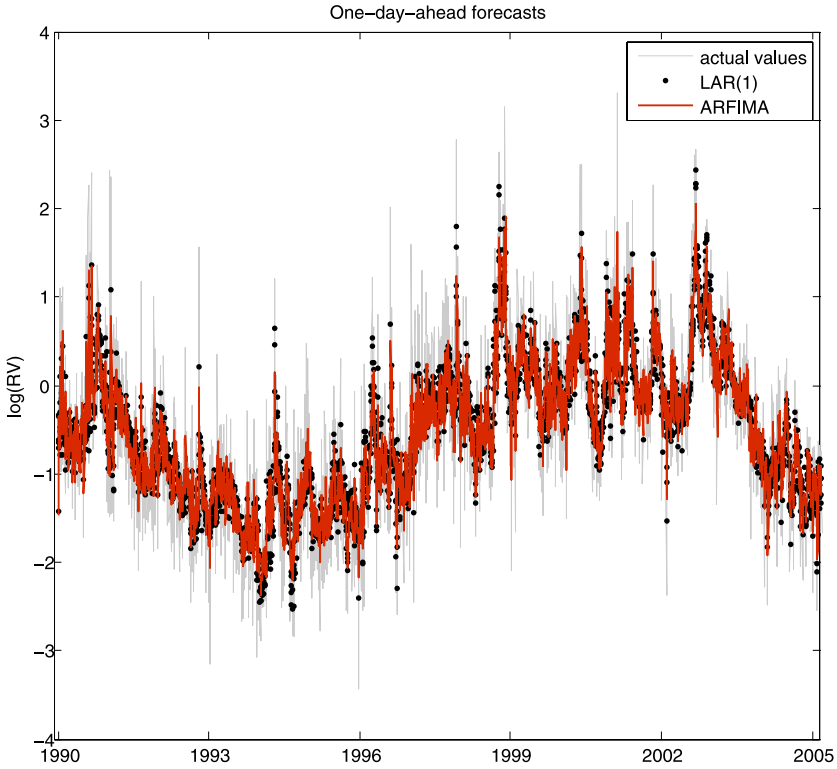

Figure 10. Time-evolvement of the actual log realized volatility (grey line in the background) and the one-step ahead forecasts of (i) the LAR(1) model with critical values being computed over 6 months and (ii) the ARFIMA model based on a 3-year rolling window. These model specifications yield the minimum RMSFE within each model class (see Table 6).

been found to minimize the RMSFE. That is the information set reports either the rolling window size or the sample size used in the computation of the critical values. An illustration of the time-evolvement of the forecasts is presented in Figure 10 which depicts the one-day ahead forecasts of the LAR(1) and ARFIMA models having minimal RMSFEs.

Interestingly, according to the RMSFEs our LAR procedure provides the most accurate forecasts at all forecast horizons. Note that this already holds for the forecasts based on the LAR model with globally computed critical values, which can be readily inferred by comparing the results reported in Tables 5 and 6.

The direct comparison of the LAR forecasts with those based on the constant AR models also reveals, that the selection of the locally homogenous intervals is indeed important. The adaptive procedure, which determines at each time point the adequate length of the time interval over which the AR model is appropriate, is superior. Note that for increasing window sizes, that is, larger information sets, the predictability of the constant AR model worsens (results are not reported gere, but are available from the authors upon request). This might be expected as for larger sample sizes, for example, more than 2 years, the autocorrelation function of realized volatility exhibits more persistence and, thus, an AR model tends to be misspecified. The STR-Tree model, instead, is better suited to generate long-range dependence as it picks local AR(1) specifications that depend on the state of the lagged daily return. It is therefore surprising that the model performs worse than those without leverage effect. However, this may be due to our model specification that makes only use of past daily returns. For a different dataset, Scharth and Medeiros (2009), for example, find a superior performance of the STR-Tree model where the splits are determined by returns accumulated over the past 90, 39, 5, and 2 days, indicating that long-term returns are important when modeling and forecasting realized volatility. A more thorough treatment of the leverage effect is the subject of future research.

In accordance to the empirical results reported in the realized volatility literature so far, the HAR and ARFIMA models exhibit similar forecast accuracy with a slight tendency of the ARFIMA model to outperform the HAR model. Interestingly, the results indicate that the inclusion of structural changes in the form of the adaptive ARFIMA model does not lead to improvements in the predictability of the S&P500 realized volatility. Moreover, all long-memory models are outperformed by the LAR method. This becomes even more pronounced for larger forecast horizons. In order to get a feeling of whether this is due to a comparison of direct with iterated forecasts we have additionally computed direct forecasts for the HAR model. We find that the iterated method provides better forecasts than the direct one (e.g., the RSMFE of the direct HAR forecasts based on a 2.5 year rolling window size is 0.5857 for $h = 5$ and 0.6240 for $h = 10$), which is consistent with the recent empirical findings reported in Ghysels, Rubia, and Valkanov (2009) and Marcellino, Stock, and Watson (2005).

We further evaluate the predictive performance of the different realized volatility models on the grounds of the so-called Mincer–Zarnowitz regressions, that is, by regressing the observed log realized volatility on the corresponding forecasts of model i:

$$\log RV_t = \alpha + \beta \log \widehat{RV}_{t,i} + \nu_t. \tag{20}$$

Table 7. Results of the Mincer–Zarnowitz regressions and Diebold–Mariano tests for the volatility models with minimal RMSFEs

Model	p-value	R^2	DM (best LAR) t-stat.
	$h = 1$		
LAR, 6m	0.7242	0.7180	
3m AR(1)	0.6203	0.6872	−9.5125
STR-Tree	0.0014	0.6242	−14.8673
3y ARFIMA	0.6375	0.6942	−6.2782
4.5y A-ARFIMA	0.0388	0.6909	−6.8551
3y HAR	0.8842	0.6910	−6.9275
	$h = 5$		
LAR, 1m	0.1265	0.7377	
3m AR(5)	0.2326	0.6004	−14.4786
STR-Tree	0.0000	0.4860	−15.3163
3y ARFIMA	0.5656	0.5835	−14.2157
4.5y A-ARFIMA	0.0233	0.5745	−14.1834
2.5y HAR	0.7427	0.5803	−14.7150
	$h = 10$		
LAR, 1m	0.0705	0.7392	
3m AR(10)	0.0897	0.5789	−13.8568
STR-Tree	0.0000	0.1911	−14.7564
3y ARFIMA	0.7593	0.5273	−12.8450
4y A-ARFIMA	0.2201	0.5123	−12.4366
2.5y HAR	0.6611	0.5236	−13.0511

NOTE: Reported are results of the Mincer–Zarnowitz regressions and of the modified Diebold–Mariano tests for the models yielding the minimum RMSFE within each model class (see Tables 5 to 6). The results are reported for different forecast horizons h (in days). The second column reports the p-value of a F-test for $H_0 : \alpha = 0$ and $\beta = 1$, and the third column reports the coefficient of determination (R^2) of the Mincer–Zarnowitz regression given in Equation (20). The last column gives the modified t-statistics of the Diebold–Mariano test on equal forecast performance, that is, $H_0 : \mu = 0$ in the regression $e^2_{t,\text{LAR}} - e^2_{t,i} = \mu + \nu_t$ with $e_{t,i}$ denoting the forecast error of model i. Results are based on heteroscedasticity and autocorrelation robust Newey–West (co)variances.

This allows to test for the unbiasedness of the different forecasts. Table 7 reports the coefficients of determination (R^2s) of this regression along with the p-value of the F-test on unbiased forecasts, i.e., $H_0 : \alpha = 0$ and $\beta = 1$. Note that for the ease of exposition we again solely present here the comparison of the models performing best in terms of the RMSFE.

The results indicate that, with the exception of the forecasts of the STR-Tree model, none of the forecasts is significantly biased at the 5% significance level. The coefficients of determination reported in Table 7 indicate a superior forecasting performance of the adaptive LAR models. We investigate this result further and test for the significance of the observed differences in the forecast accuracies. In particular, we conduct a pairwise test on the equality of the mean square forecast errors (MSFE) of the LAR procedure and the other models; see Diebold and Mariano (1995). To this end, we regress the difference between the squared forecast errors of the LAR model and those of the competing model i, that is, $e^2_{t,\text{LAR}} - e^2_{t,i}$, on a constant μ. The null hypothesis of equal MSFEs is equivalent to $H_0 : \mu = 0$. Table 7 reports the modified Diebold–Mariano test statistics proposed in Harvey, Leybourne, and Newbold (1997). Obviously, the null hypothesis is always strongly rejected in favor of a significant better forecasting performance of the adaptive LAR model, as indicated by the significant negative sign of the t-statistic. Overall, the LAR approach seems to be superior.

However, it should be noted that this conclusion is based on a pairwise comparison of the best models and there may be LAR models for which this is not the case. A simultaneous comparison of the predictive ability of all competing models would be desirable at this stage. However, the corresponding existing tests, like the test for superior predictive ability (SPA) of Hansen (2005) and the model confidence set approach of Hansen, Lunde, and Nason (2010) are not applicable here, as the forecasts are based on time varying window sizes (given by the locally selected interval of homogeneity and the recursive forecasting scheme employed in the STR-Tree model), which violates the assumption of strict stationarity of the loss differential. The Diebold–Mariano test, in contrast, can still be applied; see Giacomini and White (2006). To obtain a broader picture on the performance of the LAR procedure, we therefore extend the pairwise comparisons. In particular, we additionally conduct a pairwise comparison of forecasts of the alternative models conditional on a moderately small sample (1 year) and on a large sample (5 years) with forecasts from the LAR models based on 1 year adaptively and on globally computed critical values. Note that for the ease of exposition we do not report the corresponding results here, however, they are available from the authors upon request. Overall, the results are similar to the ones reported in Table 7. Only for the global LAR model, we fail to reject the null in the comparison with the one-step-ahead forecasts of the long-memory models. But also in those cases the t-statistics are negative.

6. CONCLUSION

This paper investigates a dual view on the long-range dependence of realized volatility. While the current realized volatility literature primarily advocates the use of long-memory models to explain this phenomenon, we argue that volatility can alternatively be described by short-memory models with structural breaks. To this end we propose localized realized volatility modeling where we consider the case of a dynamic short-memory model. In particular, at each point in time we determine an interval of homogeneity over which the volatility is approximated by an AR process. Our approach is based on local adaptive techniques developed in Belomestny and Spokoiny (2007), which make it flexible and allow for time-varying coefficients. It does neither require the specification of the type, magnitude or reasons of breaks. This contrasts to smooth transition or regime switching models.

Our procedure relies on parameters, that have to be predetermined. A simulation study, however, shows that the procedure is quite robust to the choice of parameters and to model misspecification. Interestingly, the method performs also well, even if the true source of the long-range dependence is a long-memory process. Moreover, we show, that an adaptive view on intervals of local homogeneity (and a decrease in the respective underlying sample size) is increasing the procedure's flexibility, yielding higher accuracy in estimation and a better forecasting performance. Furthermore, the choice of the underlying parameters can also be based upon criteria reflecting the user's objective, such as in sample fit or forecasting criteria. Although we have refrained from doing so in our empirical application, we find that our adaptive localized realized volatility procedure

provides accurate volatility forecasts and significantly outperforms the standard long-memory realized volatility models and two alternative models with breaks. It seems that our view on volatility is practical and realistic.

Extensions of the local parametric model to explicitly account for other important data characteristics, such as the leverage effect, are left for future research.

[Received January 2009. Revised June 2010.]

REFERENCES

Allen, D. E., McAleer, M., and Scharth, M. (2010), "Realized Volatility Risk," Discussion Paper CIRJE F-693, CIRJE, University of Tokyo, Faculty of Economics. [1377]

Andersen, T. G., and Bollerslev, T. (1998), "Answering the Skeptics: Yes, Standard Volatility Models Do Provide Accurate Forecasts," *International Economic Review*, 39, 885–905. [1378]

Andersen, T. G., Bollerslev, T., and Diebold, F. X. (2007), "Roughing It Up: Including Jump Components in the Measurement, Modeling and Forecasting of Return Volatility," *Review of Economics and Statistics*, 89, 701–720. [1387]

Andersen, T., Bollerslev, T., Diebold, F., and Ebens, H. (2001a), "The Distribution of Realized Stock Return Volatility," *Journal of Financial Economics*, 61, 43–76. [1378]

Andersen, T. G., Bollerslev, T., Diebold, F. X., and Labys, P. (2001b), "The Distribution of Realized Exchange Rate Volatility," *Journal of the American Statistical Association*, 96, 42–55. [1376]

——— (2003), "Modeling and Forecasting Realized Volatility," *Econometrica*, 71, 579–625. [1376,1387]

Bai, J., and Perron, P. (1998), "Estimating and Testing Linear Models With Multiple Structural Changes," *Econometrica*, 66, 47–78. [1380]

Baillie, R. T., and Morana, C. (2009a), "Investigating Inflation Dynamics and Structural Change With an Adaptive ARFIMA Approach," Working Paper 6/09, International Centre for Economic Research. [1377,1388]

——— (2009b), "Modelling Long Memory and Structural Breaks in Conditional Variances: An Adaptive FIGARCH Approach," *Journal of Economics Dynamics and Control*, 33, 1577–1592. [1377,1388]

Baillie, R. T., Bollerslev, T., and Mikkelsen, H. O. (1996), "Fractionally Integrated Generalized Autoregressive Conditional Heteroskedasticity," *Journal of Econometrics*, 74, 3–30. [1376]

Bandi, F. M., and Russell, J. R. (2005), "Microstructure Noise, Realized Volatility, and Optimal Sampling," *Review of Economic Studies*, 75, 339–369. [1378]

Barndorff-Nielsen, O. E., and Shephard, N. (2002a), "Econometric Analysis of Realized Volatility and Its Use in Estimating Stochastic Volatility Models," *Journal of the Royal Statistical Society, Ser. B*, 64, 253–280. [1378]

——— (2002b), "Estimating Quadratic Variation Using Realized Variance," *Journal of Applied Econometrics*, 17, 457–477. [1378]

Barndorff-Nielsen, O. E., and Veraart, A. E. D. (2009), "Stochastic Volatility of Volatility in Continuous Time," Research Paper 2009-25, CREATES. [1377]

Barndorff-Nielsen, O. E., Hansen, P. R., Lunde, A., and Shephard, N. (2008), "Designing Realised Kernels to Measure the Ex-Post Variation of Equity Prices in the Presence of Noise," *Econometrica*, 76, 1481–1536. [1378]

Belomestny, D., and Spokoiny, V. (2007), "Spatial Aggregation of Local Likelihood Estimates With Applications to Classification," *The Annals of Statistics*, 35, 2287–2311. [1380,1381,1391]

Cai, Z., Fan, J., and Li, R. (2000), "Efficient Estimation and Inferences for Varying Coefficient Models," *Journal of the American Statistical Association*, 95, 888–902. [1379]

Chen, J., and Gupta, A. (1997), "Testing and Locating Variance Changepoints With Application to Stock Prices," *Journal of the American Statistical Association*, 92, 739–747. [1377]

Chen, Y., and Spokoiny, V. (2010), "Modeling and Estimation for Nonstationary Time Series With Applications to Robust Risk Management," manuscript, C.A.S.E.—Center for Applied Statistics and Economics. [1380]

Čížek, P., Härdle, W., and Spokoiny, V. (2009), "Statistical Inference for Time-Inhomogeneous Volatility Models," *Econometrics Journal*, 12, 248–271. [1377,1381]

Corsi, F. (2009), "A Simple Approximate Long-Memory Model of Realized Volatility," *Journal of Financial Econometrics*, 7, 174–196. [1376,1387]

Corsi, F., Mittnik, S. M., Pigorsch, C., and Pigorsch, U. (2008), "The Volatility of Realized Volatility," *Econometric Reviews*, 27, 46–78. [1377]

da Rosa, J., Veiga, A., and Medeiros, M. C. (2008), "Tree-Structured Smooth Transition Regression Models," *Computational Statistics and Data Analysis*, 52, 2469–2488. [1388]

Diebold, F. X. (1986), Comment on "Modeling the Persistence of Conditional Variance," by R. F. Engle and T. Bollerslew, *Econometric Reviews*, 5, 51–56. [1376]

Diebold, F. X., and Inoue, A. (2001), "Long Memory and Regime Switching," *Journal of Econometrics*, 105, 131–159. [1376]

Diebold, F., and Mariano, R. (1995), "Comparing Predictive Accuracy," *Journal of Business & Economic Statistics*, 13, 253–263. [1391]

Doornik, J. A., and Ooms, M. (2004), "Inference and Forecasting for ARFIMA Models, With an Application to US and UK Inflation," *Studies in Nonlinear Dynamics and Econometrics*, 8, Article 14. [1389]

——— (2006), "A Package for Estimating, Forecasting and Simulating ARFIMA Models: ARFIMA," Ox package 1.04, available at *http://www.doornik.com/download.html*. [1389]

Ghysels, E., Rubia, A., and Valkanov, R. (2009), "Multi-Period Forecasts of Volatility: Direct, Iterated, and Mixed-Data Approaches," working paper, University of North Carolina, Chapel Hill. [1390]

Giacomini, R., and White, H. (2006), "Tests of Conditional Predictive Ability," *Econometrica*, 74, 1545–1578. [1391]

Granger, C. W. J. (1980), "Long Memory Relationships and the Aggregation of Dynamic Models," *Journal of Econometrics*, 14, 227–238. [1376]

Granger, C. W. J., and Hyung, N. (2004), "Occasional Structural Breaks and Long Memory With an Application to the S&P500 Absolute Stock Returns," *Journal of Empirical Finance*, 11, 399–421. [1376]

Granger, C. W., and Joyeux, R. (1980), "An Introduction to Long Memory Time Series Models and Fractional Differencing," *Journal of Time Series Analysis*, 1, 5–39. [1376]

Hamilton, J. D., and Susmel, R. (1994), "Autoregressive Conditional Heteroskedasticity and Changes in Regime," *Journal of Econometrics*, 64, 307–333. [1377]

Hansen, P. R. (2005), "A Test for Superior Predictive Ability," *Journal of Business & Economic Statistics*, 23, 365–380. [1391]

Hansen, P. R., Lunde, A., and Nason, J. M. (2010), "Model Confidence Sets for Forecasting Models," working paper, Federal Reserve Bank of Atlanta, available at *http://ssrn.com/abstract=522382*. [1391]

Härdle, W., Müller, M., Sperlich, S., and Werwatz, A. (2004), *Nonparametric and Semiparametric Models*, Berlin/Heidelberg/New York: Springer-Verlag. [1380]

Harvey, D., Leybourne, S., and Newbold, P. (1997), "Testing the Equality of Prediction Mean Squared Errors," *International Journal of Forecasting*, 13, 281–291. [1391]

Hasbrouck, J. (2007), *Empirical Market Microstructure*, New York: Oxford University Press. [1378]

Hillebrand, E., and Medeiros, M. C. (2008), "Asymmetries, Breaks, and Long-Range Dependence: An Estimation Framework for Time Series of Daily Realized Volatility," discussion paper, Pontifical Catholic University of Rio de Janeiro. [1377]

Hosking, J. R. M. (1981), "Fractional Differencing," *Biometrika*, 68, 165–176. [1376]

Koopman, S. J., Jungbacker, B., and Hol, E. (2005), "Forecasting Daily Variability of the S&P100 Stock Index Using Historical, Realised and Implied Volatility Measurements," *Journal of Empirical Finance*, 12, 445–475. [1387]

Lamoureux, C. G., and Lastrapes, W. D. (1990), "Persistence in Variance, Structural Change and the GARCH Model," *Journal of Business & Economic Statistics*, 8, 225–234. [1376]

Lanne, M. (2006), "A Mixture Multiplicative Error Model for Realized Volatility," *Journal of Financial Econometrics*, 4, 594–616. [1377]

Liebermann, O., and Phillips, P. C. B. (2008), "Refined Inference on Long Memory in Realized Volatility," *Econometric Reviews*, 27, 254–267. [1376]

Liu, C., and Maheu, J. M. (2008), "Are There Structural Breaks in Realized Volatility?" *Journal of Financial Econometrics*, 6, 326–360. [1377]

Marcellino, M., Stock, J. H., and Watson, M. (2005), "A Comparison of Direct and Iterated Multistep AR Methods for Forecasting Macroeconomic Time Series," *Journal of Econometrics*, 135, 499–526. [1390]

Martens, M., Dijk, D. v., and de Pooter, M. (2009), "Modeling and Forecasting S&P 500 Volatility: Long Memory, Level Shifts, Leverage Effects, Day-of-the-Week Seasonality, and Macroeconomic Announcements," *International Journal of Forecasting*, 25, 282–303. [1377]

McAleer, M., and Medeiros, M. (2008a), "Realized Volatility: A Review," *Econometric Reviews*, 27 (1), 10–45. [1378]

——— (2008b), "A Multiple Smooth Transition Heterogeneous Autoregressive Model for Long Memory and Asymmetries," *Journal of Econometrics*, 147, 104–119. [1377]

Mikosch, T., and Stărică, C. (2004a), "Changes of Structure in Financial Time Series and the GARCH Model," *REVSTAT Statistical Journal*, 2, 41–73. [1377,1380]

——— (2004b), "Non-Stationarities in Financial Time Series, the Long Range Dependence and the IGARCH Effects," *Review of Economics and Statistics*, 86, 378–390. [1377]

Morana, C., and Beltratti, A. (2004), "Structural Change and Long-Range Dependence in Volatility of Exchange Rates: Either, Neither or Both?" *Journal of Empirical Finance*, 11, 629–658. [1377]

Müller, U. A., Dacorogna, M. M., Dav, R. D., Olsen, R. B., Pictet, O. V., and von Weizsäcker, J. E. (1997), "Volatilities of Different Time Resolutions—Analyzing the Dynamics of Market Components," *Journal of Empirical Finance*, 4, 213–239. [1387]

Pigorsch, C., Pigorsch, U., and Popov, I. (2010), "Volatility Estimation Based on High-Frequency Data," in *Handbook of Computational Finance*, eds. J. C. Duan, J. E. Gentle, and W. K. Härdle, Heidelberg: Springer. [1378]

Polzehl, J., and Spokoiny, V. (2006), "Propagation-Separation Approach for Local Likelihood Estimation," *Probability Theory and Related Fields*, 135, 335–362. [1380]

Pong, S., Shackleton, M. B., Taylor, S. J., and Xu, X. (2004), "Forecasting Currency Volatility: A Comparison of Implied Volatilities and AR(FI)MA Models," *Journal of Banking & Finance*, 28, 2541–2563. [1376]

Scharth, M., and Medeiros, M. C. (2009), "Asymmetric Effects and Long Memory in the Volatility of Dow Jones Stocks," *International Journal of Forecasting*, 25, 304–327. [1377,1388-1390]

So, M. K. P., Lam, K., and Li, W. K. (1998), "A Stochastic Volatility Model With Markov Switching," *Journal of Business & Economic Statistics*, 16, 244–253. [1377]

Quantitative Finance
Publication details, including instructions for authors and subscription information:
http://www.tandfonline.com/loi/rquf20

Modeling default risk with support vector machines

Shiyi Chen [a], W. K. Härdle [b] & R. A. Moro [c d]

[a] China Center for Economic Studies (CCES), Fudan University, 220 Handan Road, 200433 Shanghai, PR China

[b] Center for Applied Statistics and Economics (CASE), Humboldt-Universität zu Berlin, Spandauer Str. 1, 10178 Berlin, Germany

[c] Department of Economics and Finance, Brunel University, Uxbridge, Middlesex UB8 3PH, UK

[d] DIW econ, Mohrenstr. 58, 10117 Berlin, Germany

Version of record first published: 20 Apr 2010

To cite this article: Shiyi Chen, W. K. Härdle & R. A. Moro (2011): Modeling default risk with support vector machines, Quantitative Finance, 11:1, 135-154

To link to this article: http://dx.doi.org/10.1080/14697680903410015

Quantitative Finance, Vol. 11, No. 1, January 2011, 135–154

Modeling default risk with support vector machines

SHIYI CHEN*†, W. K. HÄRDLE‡ and R. A. MORO§¶

†China Center for Economic Studies (CCES), Fudan University, 220 Handan Road, 200433 Shanghai, PR China
‡Center for Applied Statistics and Economics (CASE), Humboldt-Universität zu Berlin, Spandauer Str. 1, 10178 Berlin, Germany
§Department of Economics and Finance, Brunel University, Uxbridge, Middlesex UB8 3PH, UK
¶DIW econ, Mohrenstr. 58, 10117 Berlin, Germany

(*Received 10 January 2007; in final form 12 January 2009*)

Predicting default risk is important for firms and banks to operate successfully. There are many reasons to use nonlinear techniques for predicting bankruptcy from financial ratios. Here we propose the so-called Support Vector Machine (SVM) to predict the default risk of German firms. Our analysis is based on the Creditreform database. In all tests performed in this paper the nonlinear model classified by SVM exceeds the benchmark logit model, based on the same predictors, in terms of the performance metric, AR. The empirical evidence is in favor of the SVM for classification, especially in the linear non-separable case. The sensitivity investigation and a corresponding visualization tool reveal that the classifying ability of SVM appears to be superior over a wide range of SVM parameters. In terms of the empirical results obtained by SVM, the eight most important predictors related to bankruptcy for these German firms belong to the ratios of activity, profitability, liquidity, leverage and the percentage of incremental inventories. Some of the financial ratios selected by the SVM model are new because they have a strong nonlinear dependence on the default risk but a weak linear dependence that therefore cannot be captured by the usual linear models such as the DA and logit models.

Keywords: Statistical learning theory; Applications to default risk; Capital asset pricing; Economics of risk

1. Introduction

Predicting default probabilities and deducing the corresponding risk classification is becoming more and more important in order for firms to operate successfully and for banks to clearly grasp their clients' specific risk class. In particular, the implementation of the Basel II capital accord will further exert pressure on firms and banks. As both the risk premium and the credit costs are determined by the default risk, the firms' ratings will have a deeper economic impact on banks as well as on the firms themselves than ever before. Thus, from a risk management perspective, the choice of a correct rating model that can capture consistent predictive information concerning the probabilities of default over some successive time periods is of crucial importance.

There are strands of the literature that deal with the statistical and stochastic analysis of default risk (Burnham and Anderson 1998, Caouette *et al*. 1998, Shumway 1998, Sobehart *et al*. 2000, Saunders and Allen 2002, Gaeta 2003, Chakrabarti and Varadachari 2004, Giesecke 2004, Zagst and Hocht 2006). One models default events using accounting data, whereas other models recommend using market information. Market-based models can be further classified into structural models and reduced form models. There is also a hybrid approach that uses accounting data as well as market information to predict the probability of default. The market-based approach relies on the time series of company market data. Unfortunately, time series long enough to reliably estimate the risk is not available for most companies. Moreover, the majority of German firms are not listed and, therefore, their market price is unknown. This justifies the choice of a model for which only cross-sectional or pooled accounting data would be required. For this study, accounting data for bankrupt and operating German companies was provided by Creditreform.

Among the accounting-based models, the first attempts to identify the difference between the financial ratios of

*Corresponding author. Email: shiyichen@fudan.edu.cn

Quantitative Finance
ISSN 1469–7688 print/ISSN 1469–7696 online © 2011 Taylor & Francis
http://www.informaworld.com
DOI: 10.1080/14697680903410015

solvent and insolvent firms were the studies of Ramser and Foster (1931), Fitzpatrick (1932), Winakor and Smith (1935) and Merwin (1942). These studies settled the fundamentals for bankruptcy prediction research. It was not until the 1960s that the traditional research was changed. Beaver (1966) pioneeringly presented the univariate approach to discriminant analysis (DA) for bankruptcy prediction. Altman (1968) expanded this analysis to multivariate analysis. Up to the 1980s, DA was the dominant method in bankruptcy prediction. However, there are obvious modeling restrictions of this approach, some of which are the assumptions of normality, homoscedasticity of the disturbances, fulfillment of conditional expectation of the dependent variable between 0 and 1, and no adjustment for multicollinearity. During the 1980s the DA method was replaced by logistic analysis, which fits the logistic regression model for binary or ordinal response data by the method of maximum likelihood estimation (MLE). In fact, the logit model uses the logistic cumulative distribution function in modeling the default probability. Among the first users of logit analysis in the context of bankruptcy were Ohlson (1980), Collins and Green (1982), Lo (1986) and Platt *et al*. (1994). The advantage of the logit model is that it does not assume multivariate normality and equal variance disturbance, and its probability lies between 0 and 1 (Härdle and Simar 2003). However, the logit model is also sensitive to the collinearity among the variables. In addition, the key assumption behind the logit model is that the logarithm of odds is linear in the underlying random variable; therefore, common to DA and logit modeling is a linear classifying hyperplane that separates insolvent and solvent firms. This works well if the data are linearly separable. A linear separating hyperplane is, however, not suitable if there is doubt that the separation mechanism is of a nonlinear kind. There are good reasons to take the linear non-separability case seriously (Falkenstein *et al*. 2000).

Many nonlinear numerical methodologies have been developed to solve the linear non-separability problem: Maximum Expected Utility (MEU), Artificial Neural Networks (ANN) and Support Vector Machines (SVM). The MEU model was proposed at Standard & Poor's Risk Solutions Group, which allows models to incorporate the nonlinearity, non-monotonicity, and interactions present in the data, reducing the risk of overfitting. Friedman and Sandow (2003a, b) and Friedman and Huang (2003) demonstrated how the MEU method outperforms the Logit model. ANN was introduced to analyse bankrupt firms in the 1990s (see Hertz *et al*. (1991), Refenes (1995) and Härdle *et al*. (2004) for more details). This method also discards the assumption of linearity and mutual independence of explanatory variables for the default prediction function (Serrano *et al*. 1993, Back *et al*. 1994, 1996, Wilson and Sharda 1994). ANN models built using *K*-fold cross-validation techniques can be very robust and reduce over-fitting. Although the nonlinear ANN can classify a dataset much better than the linear models, it has often been criticized to be vulnerable to the multiple minima problem. Common to the OLS and MLE for linear models, ANN also makes use of the principle of minimizing empirical risk, which usually leads to a poor level of classification for out-of-sample data (Haykin 1999).

Based on statistical learning theory, an alternative nonlinear separation method, the Support Vector Machine (SVM), was recently introduced in default risk analysis. The SVM yields a single minimum without undesirable local fits as often produced by ANN. This property results from the minimized target function that is convex quadratic and linearly restricted. In addition, the SVM is also able to handle the interactions between the ratios and does not need any parameter restrictions and prior assumptions such as that concerning the distribution for latent errors. Furthermore, the biggest advantage of SVM among all the alternatives is its ability to minimize the risk associated with model misspecification, which endows SVM with an excellent separating ability. The current literature in statistical learning theory has produced strong evidence that SVM systematically outperforms standard pattern recognition/classification, function regression and data analysis techniques (Vapnik 1995, Haykin 1999). The application of SVM to company default analysis is less reported in the management science and finance literature. Härdle *et al*. (2005, 2007) report that, compared with the traditional DA and logit models in predicting the probabilities of default and rating firms, the SVM has a superior performance. Gestel *et al*. (2005) combined SVM and the logistic regression model to capture the multivariate nonlinear relations. This combination technique balances the interpretability and predictability required to rating banks.

In this study, we investigate the applicability of this new technique to predicting the risk scores and the probabilities of defaults (PDs) of German firms from the Creditreform database spanning from 1996 through 2002. The aim is to investigate (1) which of the accounting ratios are meaningful and have predictive character for bankruptcy, and (2) does a well-specified SVM-based nonlinear model consistently outperform the benchmark logit model in predicting PDs as predicted by theory?

The rest of the paper is organized as follows. In the next section we give a theoretical introduction to the Support Vector Machine (SVM) for classification. Section 3 describes the Creditreform database and the variables and ratios used in this study. In section 4, we present the validation procedures, re-sampling technique, performance measures and the ratios selection methods. Section 5 analyses the empirical results, including the predictors related to bankruptcy, the sensitivity analysis of SVM parameters, and a comparison of the predictive performance between SVM and the logit model. Section 7 offers conclusions.

2. The Support Vector Machine

The term Support Vector Machine (SVM) originates from Vapnik's statistical learning theory (Vapnik 1995, 1997), which formulates the classification problem as a quadratic

programming (QP) problem. The principles on which the SVM is based, especially the regularization principle for solving ill-posed problems, are also described by Tikhonov (1963), Tikhonov and Arsenin (1977) and Vapnik (1979). The SVM transforms by nonlinear mapping the input space (of covariates) into a high-dimensional feature space and then solves a linear separable classification problem in this feature space. Thus, linear separable classification in the feature space corresponds to linearly non-separable classification in the lower-dimensional input space. As the name implies, the design of the SVM hinges on the extraction of a subset of the training data that serves as support vectors and that represents a stable characteristic of the data.

Given a training data set $\{x_i, y_i\}_{i=1}^n \{(\mathbf{x}_i, y_i)\}_{i=1}^n$, with input vector $\mathbf{x}_i \in R^d$ (company financial ratios in this study) $x_i \in R^d$ and output scalar $y_i \in \{+1, -1\}$ $y_i = \{+1, -1\} \in R^1$ (-1 = 'successful', $+1$ = 'bankrupt'), we aim to find a classifying (score) function $f(\mathbf{x})$ to approximate the latent, unknown decision function $g(\mathbf{x})$. In the logistic and the DA case, this is simply a linear function. In the SVM case, the classifying function is

$$f(\mathbf{x}) = \sum_{l=1}^{l} w_l \phi_l(\mathbf{x}) + b = \mathbf{w}^T \phi(\mathbf{x}) + b, \qquad (1)$$

where $\phi(\mathbf{x}) = [\phi_1(\mathbf{x}), \ldots, \phi_l(\mathbf{x})]^T$ and $\mathbf{w} = [w_1, \ldots, w_l]^T$.

The nonlinear functions $\phi(\mathbf{x})$ are the transformation functions from the input space to the feature space that represent the features of the input space. A simple example of features for a quadratic function in a two-dimensional space is $\phi_1 = x_1^2$, $\phi_2 = \sqrt{2}x_1x_2$ and $\phi_3 = x_2^2$. The dimension of the feature space is l, which is directly related to the capacity of the SVM to approximate a smooth input–output mapping; the higher the dimension of the feature space, the more accurate, at the cost of variability, the approximation will be. Parameter $\mathbf{w}$ denotes a set of linear weights connecting the feature space to the output space, and b is the bias or threshold. The optimal solution $\mathbf{w}^*$ and b^* can be used to construct the optimal hyperplane $\mathbf{w}^{*T}\phi(\mathbf{x}) + b^* = 0$ and the classification function $f(\mathbf{x}) = \mathbf{w}^{*T}\phi(\mathbf{x}) + b^*$. We can predict solvent and insolvent companies using the estimated function $f(\mathbf{x})$.

2.1. *Advantage of SVM for classification in theory*

The main superiority of nonlinear non-parametric SVM over the benchmarking methods in predicting company credit risk results from its special theoretical device in two ways: (1) it takes linearly non-separable situations into account, whereas the DA and logit models only work well if the data are linear separable; and (2) it adopts the principle of structural risk minimization rather than empirical risk minimization employed by the OLS, MLE, ANN (and other) models. We illustrate the principle in figure 1 using the simplest classifying function $f(\mathbf{x}) = -x_1 - 2x_2 + 2$, where $\mathbf{x} = (x_1, x_2)^T$, $\mathbf{w} = (-1, -2)$ and $b = 2$.

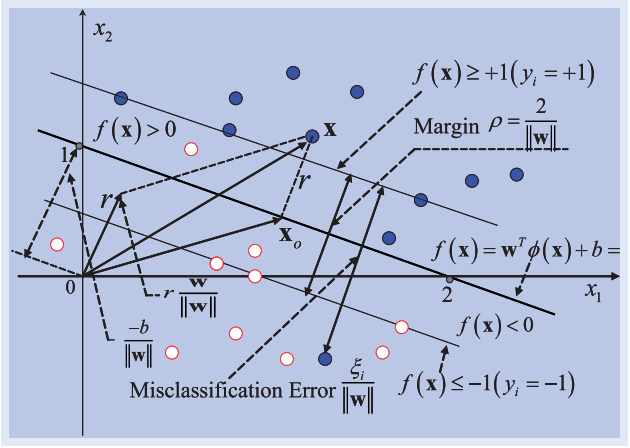

Figure 1. Separation margin, misclassification error and structural risk minimization for the SVM in two-dimensional input space.

The statistical problem is how to construct a classifying hyperplane (hypersurface) and obtain the classifying function $f(\mathbf{x})$. If the data set is linearly separable, the perfect classification hyperplane does exist. The function $f(\mathbf{x})$ gives an algebraic measure of the distance from $\mathbf{x}$ to the optimal hyperplane. Perhaps the easiest way to see this is to express $\mathbf{x}$ as $\mathbf{x} = \mathbf{x}_0 + r(\mathbf{w}/||\mathbf{w}||)$, where $\mathbf{x_0}$ is the normal projection of $\mathbf{x}$ onto the optimal hyperplane, r is the desired algebra distance from any point $\mathbf{x}$ to the optimal hyperplane (positive if $\mathbf{x}$ is on the positive side of the optimal hyperplane and negative otherwise), and $||\mathbf{w}||$ is the Euclidean norm of the weight vector $\mathbf{w}$. Since, by definition, $f(\mathbf{x}_0) = 0$, it follows that

$$f(\mathbf{x}) = \mathbf{w}^T\mathbf{x}_0 + \mathbf{w}^T r\frac{\mathbf{w}}{||\mathbf{w}||} + b = f(\mathbf{x}_0) + r||\mathbf{w}|| = r||\mathbf{w}||$$

or

$$r = \frac{f(\mathbf{x})}{||\mathbf{w}||}.$$

Because of the values of y_i being ± 1, the parameters $(\mathbf{w}, b)$ for the optimal hyperplane must satisfy the constraints $f(\mathbf{x}) \geq 1$ for $y_i = +1$ (insolvent) or $f(\mathbf{x}) \leq -1$ for $y_i = -1$ (solvent), that is $y_i \cdot f(\mathbf{x}) \geq 1$. The particular data points for which the constraint is satisfied with the equality sign are called *support vectors*, hence the name 'Support Vector Machine'. In conceptual terms, the support vectors are those data points that lie closest to the decision surface and are therefore the most difficult to classify. As such, they have a direct bearing on the optimum location of the classification hyperplane and play a prominent role in the operation of SVM. Now consider the support vectors; they are located on the upper and lower separation band for which $f(\mathbf{x}) = \pm 1$. Therefore, the algebraic distance from the support vectors to the optimal hyperplane is

$$r = \frac{f(\mathbf{x})}{||\mathbf{w}||} = \frac{\pm 1}{||\mathbf{w}||}.$$

Let ρ denote the optimum value of the margin of separation between solvent and insolvent companies.

Then it follows that $\rho = 2r = 2/||\mathbf{w}||$, which states that maximizing the margin of separation between classes is equivalent to minimizing the Euclidean norm of $\mathbf{w}$, $\|\mathbf{w}\|$. Thus, the classifying function in the linear separable case can be derived from maximizing the separation margin directly. Likewise, the distance from the origin to the optimal hyperplane is given by $-b/||\mathbf{w}||$, as shown in figure 1.

If the training set is linearly non-separable, the hyperplane that can correctly classify the training set no longer exists and, naturally, we need to find a hypersurface instead. For the hypersurface, however, we know less about the concept of the geometrical margin that is particular for the hyperplane; therefore, it is more difficult to find a hypersurface than a hyperplane. The transformation from the input space into higher-dimensional feature space, i.e. $\mathbf{x} \mapsto \phi(\mathbf{x})$, is then introduced in the SVM. It is possible that the new training set in the feature space $\{\phi(\mathbf{x}_i), y_i\}_{i=1}^n$ becomes linearly separable. Accordingly, the problem of finding a hypersurface in the input space is transformed into finding a hyperplane in the feature space and letting its margin or the 'safe' distance between classes, where in the perfectly separable case no observation can lie, be maximized.

It is not possible to construct a separating hyperplane without encountering classification errors. The margin of separation between classes is said to be soft if a data point violates the condition $y_i \cdot f(\mathbf{x}) \geq 1$. This violation can arise in one of two ways: (1) the data point falls inside the region of separation but on the right side of the decision surface; and (2) the data points falls on the wrong side of the decision surface. Note that we have correct classification in case (1), but misclassification in case (2). Therefore, a new set of non-negative slack variables $\{\xi_i\}_{i=1}^n$ are introduced and the condition is softened to $y_i \cdot f(\mathbf{x}) \geq 1 - \xi_i$. Note $0 < \xi_i \leq 1$ for case (1), $\xi_i \geq 1$ for case (2), and $\xi_i = 0$ for the linearly separable case. The support vectors are those particular data points that satisfy the soft condition precisely even if $\xi_i > 0$. The support vectors are thus defined in exactly the same way for both linearly separable and non-separable cases. In fact, using the soft constraints and the condition $\xi_i \geq 0$, the slack variables ξ_i can be represented as a hinge loss function which is the tightest convex upper bound of the misclassification loss and special and preferred to the loss function of the logit model because it allows a sparse solution, in the sense that some observations of the training set, if they are classified correctly, may not be necessary to construct the separating boundary. Sparseness of the solution also greatly simplifies the computation of SVM because then usually only few observations, so-called support vectors, are required to restore the solution, while for the logit regression, all observations are necessary.

The algebraic distance from the misclassification point to the optimal hyperplane is $r = [(1 - \xi_i)/\|\mathbf{w}\|]$, which can be derived making use of the same algebraic manipulation as in the linear separable case. Thus, the distance between the misclassification point and the upper band, the case in figure 1, is $\xi_i/\|\mathbf{w}\|$ and the tolerance to misclassification errors on the training set can be measured by $\sum_{i=1}^n \xi_i/||\mathbf{w}||$. Our goal is to find a separating hyperplane for which the misclassification error, averaged on the training set, is minimized, which is similar to minimize the sum of residual squares, the empirical risk in OLS and MLE estimation.

Thus, two targets exist for SVM in the linear non-separable case: still maximize the separation margin $2/||\mathbf{w}||$ and simultaneously minimize the misclassification distance $\sum_{i=1}^n \xi_i/||\mathbf{w}||$. The most intuitive form of the objective function to be minimized is

$$\min_{\mathbf{w}, b, \xi} \frac{1}{2}||\mathbf{w}|| + C \sum_{i=1}^{n} \frac{\xi_i}{||\mathbf{w}||}. \tag{2}$$

As shown above, the second term is the margin-based loss function, which is the sum of errors measured as the distance from a misclassified observation to the hyperplane boundary, its class weighted with the parameter C. Equation (2) exhibits the so-called structural risk minimizing principle held by the SVM method. The benchmark models such as the DA and logit estimated by OLS and MLE, and simple ANN-based nonlinear models with no constraints usually employ the principle of minimizing error functions calculated on the training sample. Therefore, SVM not only minimizes the traditional empirical risk, but also maximizes the separating margin, and finally obtains a trade-off between two targets. It is this kind of special design of minimizing the structural risk that endows SVM with stronger classifying ability than the benchmark methods.

2.2. *SVM algorithm*

To minimize the cost function (2), an equivalent quadratic cost function, $(1/2)||\mathbf{w}||^2 + C\sum_{i=1}^n \xi_i$, can be obtained from equation (2) multiplied by $||\mathbf{w}||$ ($||\mathbf{w}|| > 0$). Thus, the primary problem of the SVM for the non-separable case is expressed as

$$\min_{\mathbf{w}, b, \xi} \frac{1}{2}||\mathbf{w}||^2 + C \sum_{i=1}^{n} \xi_i, \tag{3}$$

s.t.

$$y_i \times \{\mathbf{w}^T \phi(\mathbf{x}_i) + b\} + \xi_i \geq 1, \tag{4}$$

$$\xi_i \geq 0, \quad i = 1, 2, \ldots, n. \tag{5}$$

As before, minimizing the first term of equation (3) is equivalent to maximizing the separation margin. The scaling factor 1/2 is included here for convenience of presentation. As for the second term, it is an upper bound on the number of misclassification errors. The formulation of the cost function in equation (3) is also therefore in perfect accord with the principle of structural risk minimization. The penalty parameter $C > 0$ is introduced to integrate the weights of two targets. It controls the trade-off between the complexity of the machine and the number of non-separable points; that is, the penalty parameter C controls the extent of penalization

(or the tolerance) to misclassification errors on the training set. Partially the optimization function is derived from the problem of separating the population of defaulters from non-defaulters. However, it contains a second part responsible for margin maximization that is introduced artificially. Although it introduces a bias to the original optimization problem, it reduces the complexity of the SVM and increases its accuracy on out-of-sample data. The value of parameter C has to be selected by the user (Haykin 1999). The optimization problem for non-separable patterns stated above includes the optimization problem for linearly separable patterns as a special case. Specifically, setting $\xi_i = 0$ for all i in both equations (3) and (4) reduces them to the corresponding forms for the linearly separable case.

The corresponding dual problem of SVM for non-separable patterns can be derived using the Karush–Kuhn–Tucker conditions (Fletcher 1987, Bertsekas 1995) as follows:

$$\min_{\alpha} \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n} y_i y_j \alpha_i \alpha_j K(\mathbf{x}_i, \mathbf{x}_j) - \sum_{i=1}^{n} \alpha_j, \tag{6}$$

s.t

$$\sum_{i=1}^{n} y_i \alpha_i = 0, \tag{7}$$

$$0 \le \alpha_i \le C, \quad i = 1, 2, \ldots, n, \tag{8}$$

where α_i and α_j are Lagrange multipliers. Note that neither the slack variables ξ_i nor their Lagrange multipliers appear in the dual problem. Thus, the objective function (6) to be minimized is the same in both the linear separable and non-separable cases. Deng and Tian (2004) demonstrate that the dual problem is easier to solve than the primal problem. We can then use the optimal solution α_i^* to obtain the solution of the primal problem:

$$\mathbf{w}^* = \sum_{i=1}^{n} y_i \alpha_i^* \phi(\mathbf{x}_i), \tag{9}$$

$$b^* = y_j - \sum_{i=1}^{n} y_i \alpha_i^* K(\mathbf{x}_i, \mathbf{x}_j), \quad \forall j \in \{j \,|\, 0 < \alpha_j^* < C\}. \tag{10}$$

By substitution, the nonlinear classifying (score) function can be obtained:

$$\begin{aligned} f(\mathbf{x}_j) &= \mathbf{w}^{*T}\phi(\mathbf{x}_j) + b^* = \sum_{i=1}^{n} y_i \alpha_i^* \phi^T(\mathbf{x}_i)\phi(\mathbf{x}_j) + b^* \\ &= \sum_{i=1}^{n} y_i \alpha_i^* K(\mathbf{x}_i, \mathbf{x}_j) + b^*, \end{aligned} \tag{11}$$

where $K(\mathbf{x}_i, \mathbf{x}_j) = \phi^T(\mathbf{x}_i)\phi(\mathbf{x}_j)$ is the inner product kernel function in which $\mathbf{x}_i$ belongs to the training set and $\mathbf{x}_j$ is the new company financial ratio, either in the training set or validating and forecasting set. For the classification problem, the decision function (11) is constructed to help us deduce in what kind of category, say $+1$ or -1, the new output $f(\mathbf{x}_j)$ corresponding to $\mathbf{x}_j$ is located. To the end, the intuitive way is to compare $\mathbf{x}_j$ with $\mathbf{x}_i$ pairwise; if $\mathbf{x}_j$ is closer to $\mathbf{x}_i$ on the positive side, then the new output $f(\mathbf{x}_j)$ nears $+1$, if $\mathbf{x}_j$ is closer to $\mathbf{x}_i$ on the negative side $f(\mathbf{x}_j)$ falls into the category -1. This is reasonable because a similar input should lead to the same output. Therefore, the decision function only depends on the proximity between two observations and the classification is in fact a proximity problem. In SVM, the inner product kernel function $K(\mathbf{x}_i, \mathbf{x}_j)$ is the key tool to measure this kind of proximity. In addition, the SVM theory considers the form of $K(\mathbf{x}_i, \mathbf{x}_j)$ in the Hilbert space without specifying $\phi(\cdot)$ explicitly and without computing all corresponding inner products, which provides the flexibility of the high-dimensional Hilbert space for low computational costs and greatly reduces the computational complexity. Thus, the kernel becomes the crucial part of SVM.

It is necessary to find an appropriate kernel in order to solve the optimization problem of SVM. The requirement on the kernel function is to satisfy Mercer's theorem (Mercer 1908, Courant and Hilbert 1970), such that the Kernel matrix, $\{K(\mathbf{x}_i, \mathbf{x}_j)\}_{i,j=1}^{n}$, is symmetric and semi-positive definite. Mercer's theorem tells us whether or not a candidate kernel is actually an inner-product kernel in some space and therefore admissible for use in a support vector machine. Within this requirement there is some freedom in how it is chosen. The usual chosen kernels are linear, polynomial and Gaussian kernel functions. A different kernel requires estimating the extent of proximity based on a different metric criterion. In this study, we choose an anisotropic Gaussian kernel for the SVM:

$$K(\mathbf{x}_i, \mathbf{x}_j) = \exp(-(\mathbf{x}_i - \mathbf{x}_j)^T r^{-2} \Sigma^{-1} (\mathbf{x}_i - \mathbf{x}_j)/2), \tag{12}$$

where Σ is the variance–covariance matrix of the data and r is the Gaussian, also known as the radial basis kernel coefficient which implicitly controls the complexity of the feature space and the solution—the larger r, the less the complexity. Therefore, based on expression (11), for any new company $\mathbf{x}_j$, those companies from the training sample $\mathbf{x}_i$ will have a greater impact on $f(\mathbf{x}_j)$ if $\mathbf{x}_j$ are closer to $\mathbf{x}_i$. The anisotropic Gaussian kernel offers a way of measuring the proximity between two companies; it is higher when the companies are close and smaller when they are far from each other.

3. Data and financial ratios

3.1. Data description

The data used in this study is the Creditreform database. It contains a random sample of 20,000 solvent and 1000 insolvent firms in Germany and spans the period from 1996 to 2002, although the data are concentrated in 2001 and 2002 with approximately 50% of the observations coming from this period. Most firms appear in the database several times in different years. Each firm is described by a set of financial statement variables such as those in balance sheets and

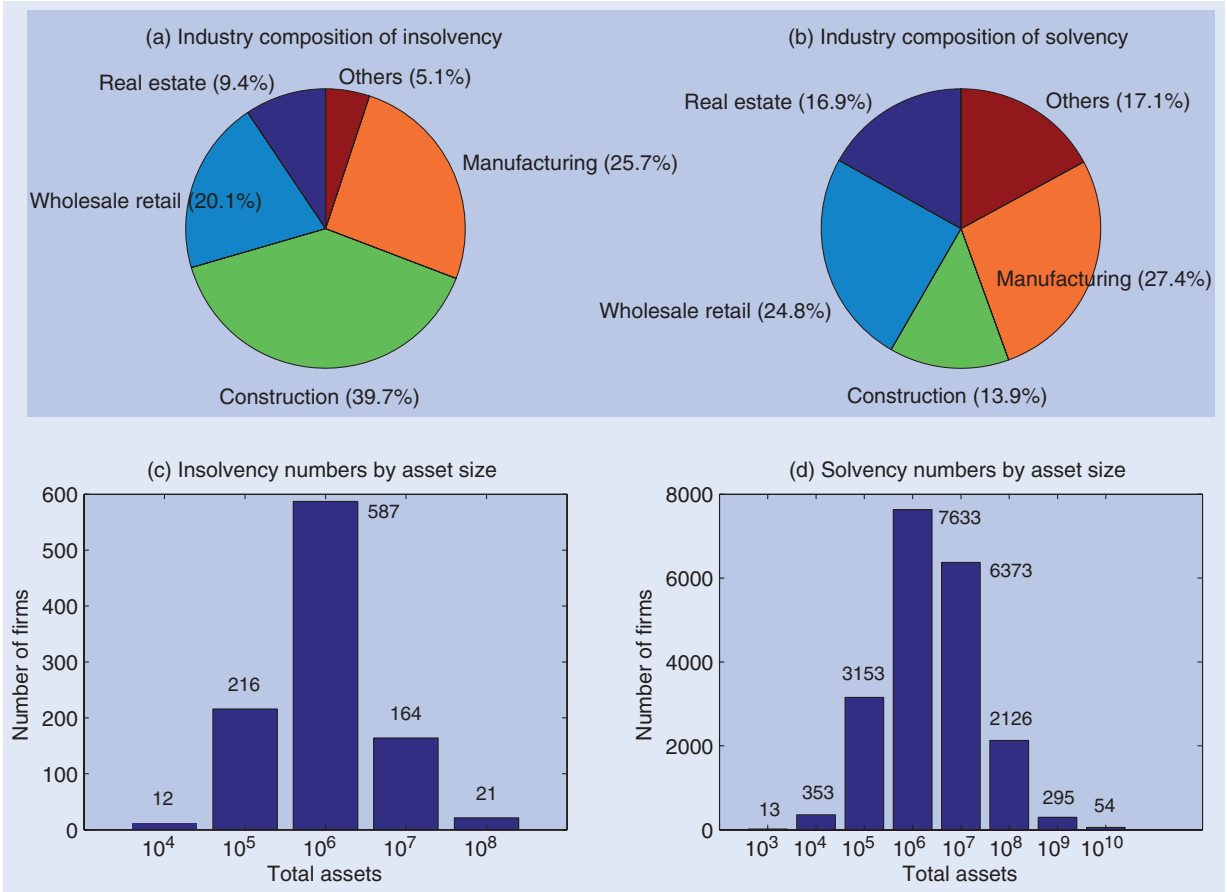

Figure 2. Industry composition and size distribution of the companies in the Creditreform database.

income statements. The data for the insolvent firms were collected two years prior to insolvency.

Figure 2 shows the industry composition and size distribution of the database. The industries to which each firm belongs can be systematically classified according to an internationally recognized system—Classification of Economic Activities, Edition 1993 (WZ 93)—published by the German Federal Statistical Office. WZ 93 uses a hierarchy of five different levels. The higher the level, the more precise the description of the main activity. In terms of the classification industry codes of WZ 93, as shown in figures 2(a) and (b), the 1000 insolvent firms consist of about 39.7% construction, 25.7% manufacturing, 20.1% the wholesale and retail trade, 9.4% real estate and 5.1% others. The others among the 1000 insolvent firms include agriculture, mining, electricity, gas and water supply, hotels and restaurants, transport and communication, financial intermediation and social service activities. The industries of the 20,000 solvent firms are manufacturing (27.4%), wholesale and retail trade (24.8%), real estate (16.9%), construction (13.9%) and others (17.1%). Different from the 'others' of insolvent firms, the others in solvency contain additional industries such as publishing, administration and defense, education and health.

The distribution of total assets can be regarded as being representative of the distribution of the firm size. In figures 2(c) and (d), the 1000 insolvent sample comprises 12 firms located in the size category 10^4 EUR, 216 in 10^5 EUR, 587 in 10^6 EUR, 164 in 10^7 EUR and 21 in 10^8 EUR. (Here, 10^4 EUR represents one category of asset size in which the firms have total assets of between 10,000 and 99,999 EUR. The definition of the other size categories is similar to that for 10^4 EUR.) The number of firms corresponding to each asset size category of the 20,000 solvent firms is 13 (10^3 EUR and below), 353 (10^4 EUR), 3153 (10^5 EUR), 7633 (10^6 EUR), 6373 (10^7 EUR), 2126 (10^8 EUR), 295 (10^9 EUR) and 54 (10^{10} EUR and above).

In an attempt to obtain a more homogeneous company sample, we cleaned the database of companies whose characteristics are very different from the others. That is to say, we do not attempt to cover all firms in the database for our study because of the very different nature of some firms. Thus, in focusing on predicting the PDs of German firms we eliminated the following types of firms from the whole sample.

- *Firms with a small percentage composition of industry*—that is, we eliminate the firms that belong to the 'other' industries in the insolvent and solvent databases, for example financial intermediation and public institutions. Thus only four main types of industry (Construction, Manufacturing, Wholesale & Retail Trade and Real Estate) remain in the study.
- *Smallest and largest firms*—that is, we exclude those firms that, because of their asset size,

Table 1. Variables used in the study.

Abbreviation	Variable	Abbreviation	Variable
CASH	Cash and cash equivalents	DEBT	Debt
INV	Inventories	AP	Accounts payable
CA	Current assets	SALE	Sales
ITGA	Intangible assets	AD	Amortization and depreciation
TA	Total assets	INTE	Interest expense
QA	Quick assets (=CA-INV)	EBIT	Earnings before interest and tax
AR	Accounts receivable	OI	Operating income
LB	Lands and buildings	NI	Net income
OF	Own funds	IDINV	Increase (decrease) inventories
CL	Current liabilities	IDL	Increase (decrease) liabilities
TL	Total liabilities	IDCASH	Increase (decrease) cash
WC	Working capital (=CA-CL)		

are not located in the categories 10^5, 10^6 and 10^7 EUR. As Khandani *et al*. (2001) noted, the credit quality of the smallest firms is often as dependent on the finances of a key individual as on the firm itself; the number of largest firms that go bankrupt is usually very small in Germany.

We further clean the database to ensure that the value of some variables, such as the denominator when calculating the ratios, should not be zero. We also exclude the firms solvent in 1996 because of missing insolvency values for this year.

Thus, 783 insolvent firms and 9583 solvent firms were chosen and analysed. The bankrupt firms are paired with non-bankrupt firms with a similar industry and total asset size. Correspondingly, the predicted default probabilities and rating results in this study are only suitable for German firms from four main industry sectors (Construction, Manufacturing, Wholesale & Retail Trade and Real Estate) and with medium asset size (lying within the categories 10^5, 10^6, and 10^7 EUR).

3.2. Ratio definitions

The Creditreform database provides many financial statement variables for each firm. In accordance with the existing literature, 28 ratios were selected for the bankruptcy analysis. In summary, there are 28 financial ratios (including one size variable) and a binary response, which records whether the firm went bankrupt within two years of the financial statements or not. There is also information on the industry distribution and on the year of the accounts. There are no missing values. These ratios can be grouped into the following six broad categories (factors): profitability, leverage, liquidity, activity, firm size and the percentage change for some variables. The variables applied to calculate these ratios are shown in table 1. Table 2 describes these ratios and how they were calculated. For simplicity, we provide short names for some ratios that capture the essence of what they measure. Table 3 summarizes the descriptive statistics of the 28 ratios for both the insolvency and solvency sample.

In previous studies, profitability ratios have appeared to be strong predictors related to bankruptcy. In addition, among all the potential risk factors, there are more profitability ratios than any other factor. The profitability ratios employed in our study are return on assets (ROA, NI/TA), net profit margin (NI/SALE), OI/TA, operating profit margin (OI/SALE), EBIT/TA, EBITDA and EBIT/SALE, denoted respectively as x1, x2, x3, x4, x5, x6 and x7.

The ROA figure gives investors an idea of how effectively the firm is deploying its assets to generate income. The higher the ROA number, the better, because the firm is earning more money on less investment. Net profit margin measures how much of every dollar of sales a firm actually keeps in earnings. A higher profit margin indicates a more profitable firm that has better control over its costs compared with its competitors. Some investors add extraordinary items back into net income when performing this calculation because they would like to use operating returns on assets, which represent a firm's true operating performance. Operating income is also required to calculate operating profit margin, which describes a firm's operating efficiency and pricing strategy. EBIT is all profits before taking into account interest payments and income taxes. An important factor contributing to the widespread use of EBIT is the way in which it nullifies the effects of different capital structures and tax rates used by different firms. By excluding both taxes and interest expenses the figure homes in on the firm's ability to profit and thus makes for easier cross-firm comparisons. EBIT is the precursor to EBITDA, which takes the process further by removing two non-cash items from the equation (depreciation and amortization). Thus, defaulting firms usually have lower profitability values; however, firms with extremely large and volatile profitability may also be likely to translate into higher default probabilities. We will try to capture this kind of complex nonlinear dependence in our database.

Leverage is also a key measure of firm risk. In this study, seven leverage ratios are analysed. They are simple and adjusted own funds ratio, CL/TA, net indebtedness, TL/TA, debt ratio (DEBT/TA) and interest coverage ratio (EBIT/INTE), represented by x8 through x14.

The own funds ratio measures the ratio of a firm's internal capital to its assets. The simple version is widely used in credit models, which is basically the mirror image

Table 2. Definitions of accounting ratios.

Ratio No.	Definition	Ratio	Category
x1	NI/TA	Return on assets (ROA)	Profitability
x2	NI/SALE	Net profit margin	Profitability
x3	OI/TA		Profitability
x4	OI/SALE	Operating profit margin	Profitability
x5	EBIT/TA		Profitability
x6	(EBIT + AD)/TA	EBITDA	Profitability
x7	EBIT/SALE		Profitability
x8	OF/TA	Own funds ratio (simple)	Leverage
x9	(OF-ITGA)/(TA-ITGA-CASH-LB)	Own funds ratio (adjusted)	Leverage
x10	CL/TA		Leverage
x11	(CL-CASH)/TA	Net indebtedness	Leverage
x12	TL/TA		Leverage
x13	DEBT/TA	Debt ratio	Leverage
x14	EBIT/INTE	Interest coverage ratio	Leverage
x15	CASH/TA		Liquidity
x16	CASH/CL	Cash ratio	Liquidity
x17	QA/CL	Quick ratio	Liquidity
x18	CA/CL	Current ratio	Liquidity
x19	WC/TA		Liquidity
x20	CL/TL		Liquidity
x21	TA/SALE	Asset turnover	Activity
x22	INV/SALE	Inventory turnover	Activity
x23	AR/SALE	Account receivable turnover	Activity
x24	AP/SALE	Account payable turnover	Activity
x25	Log(TA)		Size
x26	IDINV/INV	Percentage of incremental inventories	Percentage
x27	IDL/TL	Percentage of incremental Liabilities	Percentage
x28	IDCASH/CASH	Percentage of incremental cash flow	Percentage

Table 3. Descriptive statistics of the 28 accounting ratios. IQR is the interquartile range.

	Insolvent				Solvent			
Ratio	q0.05	Med.	q0.95	IQR	q0.05	Med.	q0.95	IQR
NI/TA	−0.19	0.00	0.09	0.04	−0.09	0.02	0.19	0.06
NI/SALE	−0.15	0.00	0.06	0.03	−0.07	0.01	0.10	0.03
OI/TA	−0.22	0.00	0.10	0.06	−0.11	0.03	0.27	0.09
OI/SALE	−0.16	0.00	0.07	0.04	−0.08	0.02	0.13	0.04
EBIT/TA	−0.19	0.02	0.13	0.07	−0.09	0.05	0.27	0.09
EBITDA	−0.13	0.07	0.21	0.08	−0.04	0.11	0.35	0.12
EBIT/SALE	−0.14	0.01	0.10	0.04	−0.07	0.02	0.14	0.05
OF/TA	0.00	0.05	0.40	0.13	0.00	0.14	0.60	0.23
(OF-ITGA) / (TA-ITGA-CASH-LB)	−0.01	0.05	0.56	0.17	0.00	0.16	0.95	0.32
CL/TA	0.18	0.52	0.91	0.36	0.09	0.42	0.88	0.39
(CL-CASH)/TA	0.12	0.49	0.89	0.36	−0.05	0.36	0.83	0.41
TL/TA	0.29	0.76	0.98	0.35	0.16	0.65	0.96	0.40
DEBT/TA	0.00	0.21	0.61	0.29	0.00	0.15	0.59	0.31
EBIT/INTE	−7.90	1.05	7.20	2.47	−6.78	2.16	73.95	5.69
CASH/TA	0.00	0.02	0.16	0.05	0.00	0.03	0.32	0.10
CASH/CL	0.00	0.03	0.43	0.11	0.00	0.08	1.40	0.29
QA/CL	0.18	0.68	1.90	0.54	0.25	0.94	4.55	1.00
CA/CL	0.56	1.26	3.73	0.84	0.64	1.58	7.15	1.56
WC/TA	−0.32	0.15	0.63	0.36	−0.22	0.25	0.73	0.41
CL/TL	0.34	0.84	1.00	0.37	0.22	0.85	1.00	0.44
SALE/TA	0.43	1.63	4.15	1.41	0.50	2.08	6.19	1.76
INV/SALE	0.02	0.16	0.89	0.26	0.01	0.11	0.56	0.16
AR/SALE	0.02	0.12	0.33	0.11	0.00	0.09	0.25	0.09
AP/SALE	0.03	0.14	0.36	0.10	0.01	0.07	0.24	0.08
Log(TA)	13.01	14.87	17.16	1.69	12.82	15.41	17.95	2.37
IDINV/INV	−1.20	0.00	0.75	0.34	−0.81	0.00	0.56	0.07
IDL/TL	−0.44	0.00	0.48	0.15	−0.53	0.00	0.94	0.14
IDCASH/CASH	−12.71	0.00	0.94	0.79	−7.13	0.00	0.91	0.52

of TL/TA, as expected: they are mathematical complements. We have made some adjustments to the simple own funds ratio to counter creative accounting practices, and to try to generate a better measure of firm credit strength. The adjustments are also used by Khandani *et al.* (2001). Net indebtedness measures the level of short-term liabilities not covered by the firm's most liquid assets as a proportion of its total assets. Thus, in addition to measuring the short-term leverage of a firm, it also provides a measure of the liquidity of a firm. While the debt ratio performs about as well as TL/TA for public firms, it does considerably worse for private firms, which makes TL/TA preferred. The difference between debt and liabilities is that liabilities is a more inclusive term that includes debt, deferred taxes, minority interest, accounts payable, and other liabilities. The interest coverage ratio is highly predictive. Falkenstein *et al.* (2000) argue that the interest coverage ratio turns out to be one of the most valuable explanatory variables in the public firm dataset in a multivariate context, although in the private firm database its relative power decreases significantly.

Six liquidity ratios, CASH/TA, cash ratio, quick ratio, current ratio, WC/TA and CL/TA (x15 through x20), are analysed in this paper. Liquidity is a common variable in most credit decisions and represents the ability to convert an asset into cash quickly. In the private dataset, CASH/TA is the most important single variable relative to default. Quick ratio is an indicator of a firm's short-term liquidity and measures a firm's ability to meet its short-term obligations with its most liquid assets. The larger the quick ratio, the better the position of the firm. The quick ratio is more conservative than the current ratio because it excludes inventory from current assets. Current ratio is mainly used to give an idea of the firm's ability to pay back its short-term liabilities (debt and payables) with its short-term assets (cash, inventory, receivables). If a firm is in default, its current ratio must be low. Yet, just as the cash in your wallet does not necessarily imply wealth, a high current ratio does not necessarily imply health. Working capital measures both a firm's efficiency and its short-term financial health. Altman (1968) reported that the WC/TA ratio is a measure of the net liquid assets of the firm relative to the total capitalization and proved to be more valuable than the current ratio and the quick ratio. Falkenstein *et al.* (2000) showed that, firstly, the CL/TL ratio appears of little use in forecasting, second that the quick ratio appears slightly more powerful than the WC/TA ratio, and third, the quick ratio and current ratio carry roughly similar information.

Activity ratios also capture important bankruptcy information and are frequently used when performing fundamental analysis for different firms. We analyse four different activity ratios: the asset turnover (TA/SALE, x21), the inventory turnover (INV/SALE, x22), the account receivable and payable turnover (AR/SALE, x23; AP/SALE, x24).

The asset turnover ratio is a standard financial ratio illustrating the sales-generating ability of the firm's assets. Usually, the asset turnover is non-monotonic and very flat. Note that some studies report that the asset turnover degrades model predictability, for example the *Z*-score that reduces the asset turnover performs better than the one that keeps it. The reciprocal of the inventory turnover shows how many times a firm's inventory is sold and replaced over a period. A high turnover implies poor sales and, therefore, excess inventory. High inventory levels are unhealthy because they represent an investment with a rate of return of zero. Accounts payable and receivable turnover ratios are more powerful predictors, the reciprocals of which also display how many times the firm's accounts are converted into sales over a period. The former is a short-term liquidity measure used to quantify the rate at which a firm pays off its suppliers. The latter is a measure used to quantify a firm's effectiveness in extending credit as well as collecting debts. By maintaining accounts receivable, firms are indirectly extending interest-free loans to their clients. The above description of the activity ratios is usually true in the manufacturing industry but is not the case for other industries. For instance, service firms may have no inventory to turn over.

Sales or total assets are almost indistinguishable as indicators of size risk, which makes the choice between the two measures arbitrary. In this study, we use the natural logarithm of total assets (log(TA), x25) to represent the firm size to investigate the default risk of small, medium (SMEs) and large firms. For example, access to capital for these firms is very different and may affect the prediction ability of some financial ratios and, consequently, the performance of the SVM model. Due to the available variables provided by the Creditreform database, we also compute three ratios of the percentage of incremental inventories, liabilities and cash flow (x26, x27, x28), respectively. For example, the increased (decreased) cash flow is the additional operating cash flow that an organization receives from taking on a new project. A positive incremental cash flow means that the firm's cash flow will increase with the acceptance of the project, the ratio of which is a good indication that an organization should spend some time and money investing in the project.

Previous empirical research has found that a firm is more likely to go bankrupt if it is unprofitable, highly leveraged, and suffers cashflow difficulties (Myers 1977, Aghion and Bolton 1992, Lennox 1999). Moreover, large firms are less likely to encounter credit constraints because of reputation effects. This is clearly demonstrated by the statistical description of financial ratios in table 3, which shows that insolvent firms are typically small, have poor profitability and liquidity, and are highly leveraged, compared with solvent firms, with only a few exceptions such as EBIT/SALE, OF/TA and EBIT/INTE. In addition, the firms that go on to default have higher values for the activity ratio. Except for the last three, all ratios for insolvent firms vary less than for solvent firms because of the smaller number of observations.

The statistics described in table 3 reveal that several of the ratios are highly skewed and there are many outliers; this may affect whether they can be of much help in

identifying insolvent and solvent firms. It is also possible that many of these outliers are errors of some kind. Therefore, the ratios used in the following analysis are processed as follows: if $x_i < q_{0.05}(x_i)$, then $x_i = q_{0.05}(x_i)$, and if $x_i > q_{0.95}(x_i)$, then $x_i = q_{0.95}(x_i)$, $i = 1, 2, \ldots, 28$. $q_\alpha(x_i)$ is an α quantile of x_i. Thus, the discriminating results obtained from both the SVM and the logit model are robust and not sensitive to outliers.

4. Prediction framework

4.1. The validation procedure

To compare the SVM and the logit models in a setting most close to the real situation in which these models are used in practice, the holdout method is chosen in this study for cross validation, namely training of the model on all available data up to the present period and the forecasting of default events for the next period. In this study, the training data are chosen from 1997 through 1999, and the validating set are selected from 2000 through 2002. Then the model is first estimated using the training data; once the model form and parameters are established, the model is used to identify insolvencies among all the firms available during the holdout period (2000–2002). Note that the predicted outputs for 2000 through 2002 are out of time for firms existing in the previous three years, and out of sample for all the firms whose data become available only after 2000. Such out-of-sample and out-of-time tests are the most appropriate way to compare model performance. The validation result set is the collection of all the out-of-sample and out-of-time model predictions that can then be used to analyse the performance of the model in more detail. For an introduction to the validation framework, see Sobehart *et al.* (2001).

Following the holdout validation procedure, we construct a training set containing 387 insolvent and 3534 solvent companies and a validation set containing 396 default events and 6049 non-defaulters. Note that the training and validation sets are themselves a subsample of the population and, therefore, may yield spurious model performance differences based only on data anomalies. A common approach to overcome this problem is to use the re-sampling techniques to leverage the available data and reduce the dependency on the particular sample at hand (Efron and Tibshirani 1993, Herrity *et al.* 1999, Horowitz 2001). Re-sampling approaches provide two related benefits (Sobehart *et al.* 2001). First, they give an estimate of the variability around the actual reported model performance. This variability can be used to determine whether differences in model performance are statistically significant, using familiar statistical tests. Second, because of the low numbers of defaults, re-sampling approaches decrease the likelihood that individual defaults (or non-defaults) will overly influence the chances of a particular model being ranked higher or lower than another model. Similar to previous bankruptcy studies, this paper also adopts a matched pairs approach for drawing subsamples for both the training and validation set. The advantage of the matching procedure is that it helps to cut the cost of data collection, as the proportion of insolvent firms in the population is very small. The problem that the use of relatively small samples could lead to over-fitting can be avoided by the re-sample techniques.

The re-sampling technique employed in this analysis is the bootstrap, which proceeds as follows. We use all insolvent firms, 387 in the training set and 396 in the validation set, and randomly select a subsample with the same number of solvencies from the 3534 solvencies in the training set and the 6049 solvencies in the validation set, respectively.

For the selected validation subset the performance measure is calculated and recorded. Then we perform a Monte Carlo experiment: another subsample is drawn, and the process is repeated. This continues for many repetitions until a distribution for each performance measure is established. In this paper the process will be repeated 30 times.

4.2. Performance measures

We now introduce two metrics for measuring and comparing the performance of credit risk models: the Accuracy Ratio (AR) and the misclassification error. These two measures aim to determine the power of discrimination that a model exhibits in warning of default risk. These techniques are quite general and can be used to compare different types of models even when the model outputs differ and are difficult to compare directly.

AR is a valuable and simple tool to determine the discriminative power of risk models. AR can be derived from the Cumulative Accuracy Profile (CAP) curve, which is particularly useful in that it simultaneously measures Type I and Type II errors (Herrity *et al.* 1999, Engelmann *et al.* 2003, Basle Committee on Banking Supervision 2005). In statistical terms, the CAP curve represents the cumulative probability distribution of default events for different percentiles of the risk score scale. To obtain CAP curves, firms are first ordered by their risk scores. For a given fraction $x\%$ of the total number of firms, a CAP curve is constructed by calculating the percentage $y(x)$ of the defaulters whose risk score is equal to or smaller than that for fraction x. In other words, for a given x, $y(x)$ measures the fraction of defaulters (of the total defaulters) whose risk scores are equal to or smaller than those of fraction x (of the total firms). One would expect a concentration of non-defaulters at the highest scores and defaulters at the lowest scores.

Figure 3 shows a CAP plot. The random CAP represents the case of zero information (which is equivalent to a random assignment of scores). The ideal CAP represents the case in which the model is able to discriminate perfectly, and all defaults are caught at the lowest model output. The actual CAP shows the performance of the model being evaluated. It depicts the percentage of defaults captured by the model.

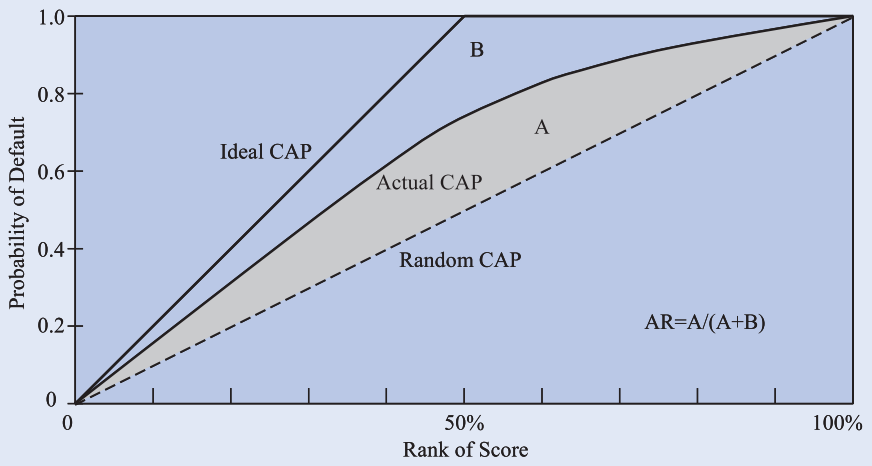

Figure 3. Cumulative accuracy profile (CAP) curve.

Therefore, AR is defined as the ratio of the area between a model's CAP curve and the random CAP curve to the area between the perfect CAP curve and the random CAP curve (see figure 3). The AR value is a fraction between zero and one. Risk measures with AR that approach zero have little advantage over a random assignment of risk scores, whereas those close to one display good predictive power. Mathematically, the AR value is defined as

$$AR = \frac{\int_0^1 y(x)\,\mathrm{d}x - (1/2)}{\int_0^1 y_{\text{ideal}}(x)\,\mathrm{d}x - (1/2)}. \tag{13}$$

If the number of bankruptcies equals the number of operating companies in the sample, then the AR becomes

$$AR \approx 2\int_0^1 y(x)\,\mathrm{d}x - 1. \tag{14}$$

In addition, when evaluating the explanatory power of the bankruptcy models, it is helpful to define two types of prediction error: a type I error, which indicates low default risk when in fact the risk is high, and a type II error, which conversely indicates a high default risk when in fact the risk is low. Usually, minimizing one type of error comes at the expense of increasing the other type of error. Clearly, the type I and type II error rates depend on the number of firms predicted to fail. The higher (lower) the number of firms predicted to go bankrupt, the smaller (larger) is the type I error rate and the larger (smaller) is the type II error rate. The number of predicted bankruptcies depends on the cut-off probability, which is equal to 0.5 in our study. From a supervisory viewpoint, type I errors are more problematic as they produce higher costs. Usually, the cost of a default is higher than the loss of prospective profits. Altman *et al.* (1977) estimated the relative costs of type I and type II errors for commercial bank loans as being 7:1. Sobehart *et al.* (2001) also described the cost scenarios schematically.

For more details on the performance measures, we refer to DeLong *et al.* (1988), Swets (1998), Keenan and Sobehart (1999), Swets *et al.* (2000), Sobehart *et al.* (2001) and Sobehart and Keenan (2004).

4.3. Predictor selection

In this study, the benchmark linear parametric probability model is the conditional logit model estimated by MLE, which is described as follows:

$$\Pr(y_i = 1|\mathbf{x}_{i1},\ldots,\mathbf{x}_{id}) = \frac{1}{1+\mathrm{e}^{-(\beta_0+\beta_1\mathbf{x}_{i1}+\cdots+\beta_d\mathbf{x}_{id}+\varepsilon_i)}}. \tag{15}$$

Based on equation (1) or (11), the target nonlinear non-parametric probability model estimated by the SVM can also be expressed in the following form:

$$\Pr(y_i = 1|\mathbf{x}_{i1},\ldots,\mathbf{x}_{id}) = f(\mathbf{x}_{i1},\ldots,\mathbf{x}_{id}) + \varepsilon_i, \tag{16}$$

where $y_i = 1$ indicates the bankrupt company, and $y_i = 0$ for the logit case and $y_i = -1$ for SVM represent the successful firm; the input vectors $\mathbf{x}_i$ are the relevant company financial ratios explaining the probability of bankrupcy. Before we begin to estimate the models, the process of predictor selection is illustrated.

For a parametric model we can estimate the distribution of the coefficients of the predictors and their confidence intervals. However, we cannot do so for non-parametric models. Instead, we can use the bootstrap technique, as described in the subsection on the validation procedure, to empirically estimate the distribution of the AR on many subsamples. In this study we randomly select 30 subsamples and compute the corresponding ARs 30 times. The median AR provides a robust measure to compare different ratios as predictors.

There are so many possible financial ratios that can be used as explanatory variables in credit scoring models that selection criteria are needed to obtain a parsimonious model. There are two main methods for selecting the appropriate ratios (Falkenstein *et al.* 2000). The first is forward stepwise selection. Start with the predictor that has the highest performance accuracy and then sequentially add the next predictor that also has the highest accuracy in the group and higher than the former until additional predictors have no additional improvement.

Table 4. Median of the AR measure for a univariate SVM model. Accounts payable turnover (AP/SALE, x24) produces the highest AR median.

No.	Ratio	AR median	No.	Ratio	AR median
x1	NI/TA	28.428	x15	CASH/TA	22.140
x2	NI/SALE	22.985	x16	CASH/CL	25.821
x3	OI/TA	36.358	x17	QA/CL	28.746
x4	OI/SALE	31.413	x18	CA/CL	16.983
x5	EBIT/TA	29.941	x19	WC/TA	14.264
x6	EBITDA	29.155	x20	CL/TL	−7.608
x7	EBIT/SALE	19.447	x21	SALE/TA	17.414
x8	OF/TA	32.941	x22	INV/SALE	24.764
x9	(OF-ITGA) / (TA-ITGA-CASH-LB)	31.938	x23	AR/SALE	17.468
x10	CL/TA	18.020	**x24**	**AP/SALE**	**49.174**
x11	(CL-CASH)/TA	23.319	x25	Log(TA)	23.816
x12	TL/TA	22.477	x26	IDINV/INV	15.493
x13	DEBT/TA	16.528	x27	IDL/TL	−9.528
x14	EBIT/INTE	28.270	x28	IDCASH/CASH	−6.562

The second is backward elimination in which one starts with all predictors, then reduces all of the poor variables. In this study, forward selection is preferred for the SVM method due to its relatively lower computational cost. The logit model, with forward selection, together with the investigation of the statistical significance and correct sign of the individual parameters of the predictors, is likely to choose different explanatory variables than the SVM. To compute and compare each method more conveniently, we will only report the results of the logit model with the same predictors as the SVM-based model. The discriminating power of each ratio is assessed using the median of the AR performance measures.

5. Empirical results

This section discusses the empirical results for each stage of the analysis of the German bankruptcy data using an SVM model. The prediction horizon in each case is two years, i.e. the data were recorded two years prior to bankruptcy for the companies that would become bankrupt. The balance sheet and income statement data for 20,000 solvent and 1000 insolvent firms in Germany were selected randomly by Creditreform. These data are represented as the financial ratios listed in table 4. They cover the period from 1996 to 2002. Each company may appear several times in different years.

5.1. Selection of the first predictor and the sensitivity of the SVM parameters

The first stage of analysing default risk is the selection of the first best predictor related to bankruptcy among the 28 ratios using the median of the AR metric in which the SVM model has one input. It is often argued that the SVM lacks interpretability of the results as is the case for the logit model. Most importantly, since there are no distributional assumptions underlying the SVM modeling, it is impossible to test the significance of variables within the SVM framework. Therefore, we will identify the most significant variable in an additional procedure before analysing the SVM model.

Based on table 4 we can see that Accounts Payable Turnover (AP/SALE, x24) provides the highest median AR of 49.17%. We can also see that CL/TL (x20), IDL/TL (x27) and IDCASH/CASH (x28) have a very low accuracy: their median AR values are below zero. For the next step we will select Accounts Payable Turnover (x24) as the first best single predictor related to German default firms, which is somewhat different from previous studies in which it was usually argued that the most significant predictors were profitability or leverage ratios. In fact, the SVM-based nonlinear model is able to search the nonlinear dependence of the data automatically as opposed to the logit model and it is Accounts Payable Turnover selected by SVM as the first predictor that greatly improves the classifying performance of SVM by more than 10%. Using most of the other ratios as the first predictor, the SVM-based model does not exceed the logit model by much in modeling the default risk.

The accounts payable turnover ratio is calculated by taking the average accounts payable and dividing it by the total sales during the same period. Its reciprocal shows investors how many times per period the firm pays its average payable amount. If the turnover ratio increases from one period to another, this is a sign that it takes the firm longer to pay off its suppliers than before. The opposite is true when the turnover ratio is falling, which means that the firm is paying off suppliers at a faster rate. Therefore, the firms with higher accounts payable turnover values will have less ability to convert their accounts into sales, have lower revenues, and go bankrupt more readily.

The SVM model has two control parameters, the influence of which was investigated in this study: the penalty parameter C and the Gaussian kernel coefficient r. C controls the tolerance to misclassification errors on the training set, while r represents the complexity of classifying functions. The possibility of fine-tuning SVM using these parameters, besides the flexibility of its classification function, further contributed to the higher performance of the SVM compared with the logit model,

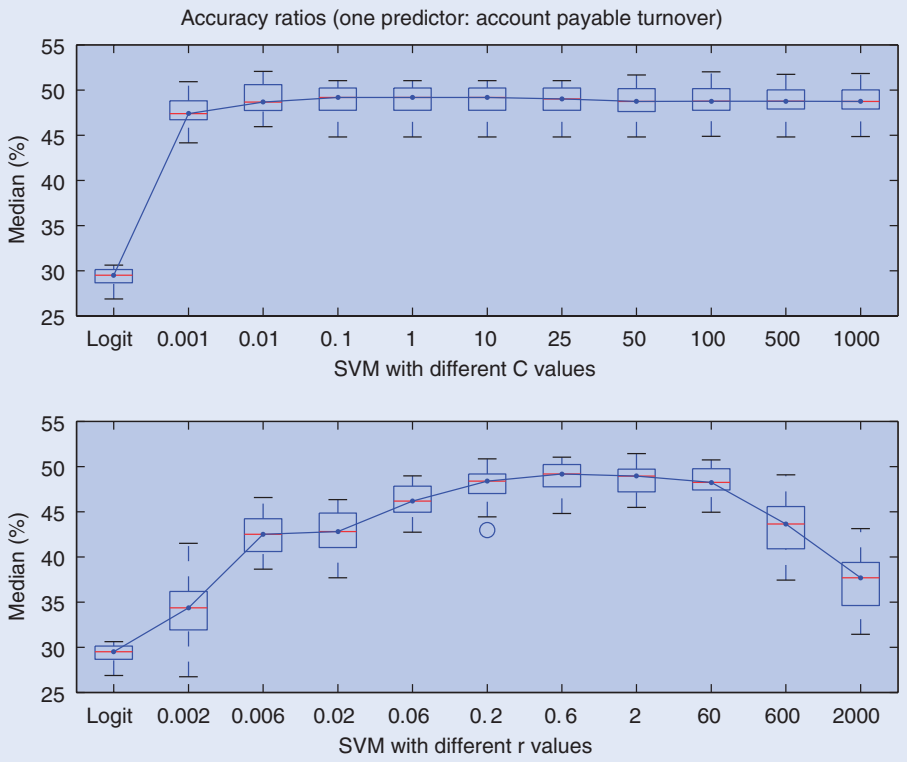

Figure 4. Sensitivity of the SVM to different parameters.

Table 5. Misclassification error (30 randomly selected samples; one predictor AP/SALE, x24).

	Parameter		Type I error		Type II error		Total error	
Model	*C*	*r*	Mean	Std	Mean	Std	Mean	Std
SVM	0.001	0.6	40.57	0.1167	23.43	0.9812	32.01	0.5723
	0.1	0.6	38.42	0.5125	24.45	1.1938	31.44	0.7014
	10	0.6	34.43	1.2126	27.86	1.637	31.15	0.9433
	100	0.6	25.22	0.6176	34.66	1.3541	29.94	0.8086
	1000	0.6	25.76	0.7705	34.26	1.3805	30.01	0.8712
	10	0.002	37.2	2.4512	32.79	2.5753	34.99	1.7611
	10	0.06	31.86	3.1527	29.25	2.2887	30.56	1.1405
	10	0.6	34.43	1.2126	27.86	1.637	31.15	0.9433
	10	60	37.27	0.5112	25.87	1.2134	31.57	0.7798
	10	2000	41.09	0.0791	24.85	0.3265	32.97	0.1123
Logit			38.15	0.5625	32.77	1.1888	35.46	0.7151

which has no similar adjustment parameters. Moreover, a grater SVM performance is a consequence of the SVM loss function, which is a tighter upper bound on the {0,1} step loss function. For univariate models, as figure 4 illustrates, the gain in performance of the SVM over the logit model is substantial and greater than for multivariate models since the former intrinsically has a larger number of degrees of freedom than the latter, which is limited by the number of variables.

The results in table 4 were obtained from the SVM with parameters $C=10$ and $r=0.6$, which were chosen according to the following sensitivity investigation of the SVM parameters (see box plot in figure 4 and table 5). That is to say, the values of parameters C and r could be determined experimentally via the standard use of a re-sampling training data set. Obviously, the SVM differs in different values of the penalty parameter C and the Gaussian kernel coefficient r. The ratio AP/SALE (x24) is exemplified here and the result for the benchmark logit model is also reported.

Here the median ARs are also estimated on 30 bootstrapped subsamples. On the whole, the discriminating ability of the SVM seems to be more sensitive to the value of r rather than to that of C. In figure 4(top), with fixed $r=0.6$, the median of the AR starts from 47.4% for $C=0.001$ and reaches the highest value 49.2% for $C=10$ and slightly decreases to 48.7% when $C=1000$. The varying range of AR is very small. Figure 4(bottom) illustrates the AR of the SVM versus r with fixed $C=10$. Within the interval, r is found to have a strong impact on the AR value, which starts at 34.4% when $r=0.002$ and drastically increases to the highest value 49.2%

when $r = 0.6$ and then decreases to 37.7% when $r = 2000$. In both parts of the figure the discriminating performance of the logit model is inferior to that of the SVM-based model with different parameter values.

As we have seen, $C = 10$ and $r = 0.6$ seem to be the best choice of parameter combination for the study in this paper. Thus, if we do not mention it particularly, the results of the SVM in the remaining part of this paper are all obtained using these parameter values. Note that this is not the case for the other data sample. The appropriate values of the C and r parameters will vary from sample to sample, therefore the sensitivity investigation of the SVM parameters should be carried out before classifying different data samples.

Table 5 shows the percentage of misclassified out-of-sample observations for the logit model and the SVM-based model with different parameters using a single predictor, the Account Payable Turnover. These errors are also obtained by bootstrap, and are all significant according to the standard deviations listed in table 5. Smaller values indicate better model accuracy. As shown in the table, the logit model has higher type I, type II and total error rates than the SVM-based model with only a few exceptions, suggesting that a well-specified SVM-based nonlinear model is superior to a logit model. For the SVM, with an increase of C from 0.001 to 1000, type II errors also increase, but type I errors decrease, and the total errors first decrease and then increase slightly. With increasing r values, type I and total errors also follow a U-shaped trend and type II errors have a monotonic negative relation with the r value. Therefore, $C = 10$ and $r = 0.6$ also appear to be the appropriate trade-off choice for our study in the following part of this paper. They produce only 34.43% type I errors, 27.86% type II errors and 31.15% total errors, whereas logit analysis produces 38.15% type I errors, 32.77% type II errors and 35.46% total errors.

As is evident from figure 5, which shows a univariate dependence of PD on AP/SALE, this dependence is not monotonously increasing or following any distinctive pattern, e.g. a logistic function. The SVM, being a more flexible non-parametric approach, is better suited for describing a broader class of dependence, such as this one, than the logit model. Another advantage of the SVM is its smaller bias in the estimation of the boundary between the solvent and insolvent companies in a situation when the number of the former is much larger than the number of the latter, as is almost always the case. The score of the logit model, which is interpreted as a PD, can be significantly biased for score values much lower or higher than 0.5. Subsequently, the threshold score for the boundary between solvent and insolvent companies is also biased. This is one reason for the substantial improvement in accuracy of the SVM compared with the logit model, as illustrated in figure 4. Because of this feature the SVM gains an additional improvement over the logit model if instead of subsamples with a 50/50 ratio of insolvent versus solvent companies we use subsamples where solvent companies prevail.

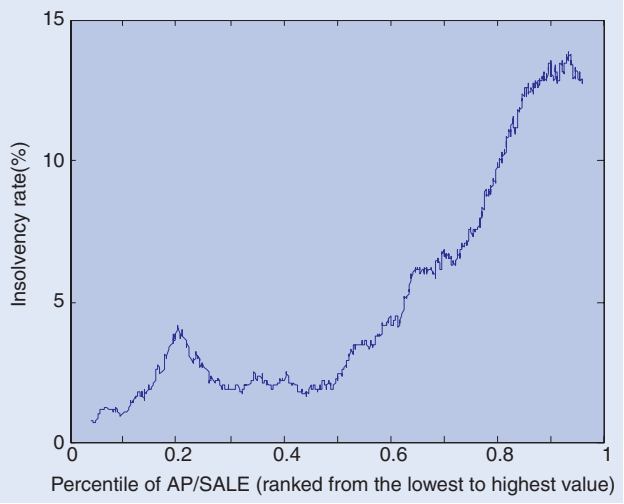

Figure 5. Insolvency rate evaluated for the financial ratio AP/SALE (x24) from the German Creditreform database. The k-nearest-neighbors procedure was used with the size of the window around 1/12 of 18,800 observations (the observations with zero values of sales used as the denominator to calculate the ratios were deleted from all 21,000 observations).

5.2. *Comparison of models with two predictors and PD visualization*

Table 6 shows the identifying performance of bivariate SVM-based models using the best predictor from the univariate model (AP/SALE) and one other. The values of the median of the AR direct us to the profitability ratio OI/TA (x3), the value of which increases to the highest of 56.46%, which indicates that OI/TA (x3) is the best choice for the second predictor.

Therefore, different from the usual result that NI/TA dominates other profitability ratios related to default risk, our study reveals that OI/TA performs better than the others in identifying bankrupt German firms. As the operating income does not include items such as investments in other firms, taxes, interest expenses and depreciation, the ratio represents a firm's true operating performance.

For two dimensions (i.e. two predictors), graphs are obviously an extremely useful tool for studying the data and assessing the quality of different default risk models. In addition, because of its nonlinearity it is more necessary for the SVM-based model to use visual tools than for the logit model to represent classification results. We demonstrate an application of visualization techniques for default analysis and parameter sensitivity investigation based on the SVM in figure 6. In the case of the logit model, the scores can be directly explained as the default probabilities, whereas for the SVM-based model the probabilities of default need to be calculated using the risk scores predicted by the estimated classifying function. Making use of the monotonic logistic cumulative distribution function, the default probabilities of German companies by SVM are calculated from the scores and then plotted as the background contour in figure 6 (corresponding to the right-hand bar in each sub-figure). The two predictors are the ratios AP/SALE (x24) and OI/TA (x3). These graphs are a subset of those used in

Table 6. Median of AR measure for a bivariate SVM model. AP/SALE (x24) and OI/TA (x3) produce the highest AR median.

No.	Ratio	AR median	No.	Ratio	AR median
x1	NI/TA	54.362	x15	CASH/TA	53.011
x2	NI/SALE	53.809	x16	CASH/CL	52.233
x3	**OI/TA**	**56.460**	x17	QA/CL	50.553
x4	OI/SALE	55.652	x18	CA/CL	44.678
x5	EBIT/TA	54.409	x19	WC/TA	48.676
x6	EBITDA	53.847	x20	CL/TL	49.725
x7	EBIT/SALE	52.948	x21	SALE/TA	49.624
x8	OF/TA	51.907	x22	INV/SALE	51.305
x9	(OF-ITGA) / (TA-ITGA-CASH-LB)	51.316	x23	AR/SALE	49.604
x10	CL/TA	48.197	**x24**	**AP/SALE**	
x11	(CL-CASH)/TA	49.680	x25	Log(TA)	51.545
x12	TL/TA	51.080	x26	IDINV/INV	49.904
x13	DEBT/TA	52.231	x27	IDL/TL	49.013
x14	EBIT/INTE	46.517	x28	IDCASH/CASH	46.617

the study. White and black points represent the 396 insolvent and 396 solvent firms from one random subsample of the validation set. The outliers were capped at the 5% and 95% quantiles as described in section 3.2 and kept in the subsample. In most panels of figure 6 they appear at the border. The classifying decision function (optimal hyperplane) is represented by the line denoted 0.5, along which the default probability is 0.5 and the risk scores are zero for SVM. The lines denoted 0.3 and 0.7 (or, more accurately, 0.27 and 0.73) are the lower and upper boundaries of the separation margin corresponding to scores of -1 and $+1$ in SVM. As shown in figure 6, clearly most successful firms lying in the blue area have positive profitability (OI/TA) and relatively lower account payable turnover (AP/SALE), while a majority of bankrupt firms is located in the opposite area. As known, low profitability usually indicates a high default risk, but extremely high profitability may also indicate a high cash flow volatility that is likely to translate into a higher default probability. Although the SVM-based model is sufficiently flexible to reveal a nonlinear dependence between profitability and PD, different from the logit model, for the Creditreform data in this study, the dependence could be too weak to be captured by SVM. Also, the sensitivity investigation results of the free parameters, C and r, of SVM could easily be determined from the figure.

Figure 6(a) shows the classification results for the logit model. Because the disadvantage of the logit model is the linearity of its solution, we see a straight classification line that is the linear combination of two predictors. Figure 6(b) shows the discriminating results obtained with the SVM-based model using a classifying function of moderate complexity ($r = 0.6$) and $C = 10$. This nonlinear classifying line (score 0 and PD 0.5) seems to identify the two types of firms very well with the areas in which solvent and insolvent firms are localized.

Fix $r = 0.6$. If the penalty is too low (C decreases to 0.01 and 0.1 as in figures 6(c) and (d)), the discriminating curve becomes flatter than that in figure 6(b). The calculated default probabilities are too small to display the two boundaries. That is, most of the firms fall inside the separation region but the insolvent and solvent firms are still clustered in their own areas. If the penalty increases, for example $C = 500$ as in figure 6(e), the identifying ability of SVM cannot be increased further than shown in figure 6(b).

Fix $C = 10$. If the complexity of the classifying functions increases (the r value decreases to 0.06 as illustrated in figure 6(f)), the SVM will try to capture each observation, although the majority of the insolvent firms still lie inside the band (0.5, 0.7) and above, with the solvent firms inside (0.5, 03) and below. The complexity in this case is too high for the given sample. If the r value increases to 60 (figure 6(g)), the classifying curve becomes flatter than that with $r = 0.6$; if r increases further to 2000 (figure 6(h)), the discriminating curve can be approximated as a linear combination of two predictors and is similar to the benchmark logit model, although the coefficients of the predictors may be different. The calculated default probabilities are also very small. The complexity here is too low to obtain a more detailed picture.

Although two cases of high complexity clearly demonstrate overfitting, (f) when $C = 10$ and $r = 0.06$, and (e) when $C = 500$ and $r = 0.6$, in all other cases the separating line is moderately nonlinear and for the case of a virtually linear SVM (h) with $C = 10$ and $r = 2000$ the separating line resembles that for the logit regression (a), with a different slope. Perfect separation for out-of-sample observations is not possible in any case. Nevertheless, comparing panel (a) for the logit with panel (f) for the SVM that achieved the maximum separation power, we observe that the most important difference between the two is in the area where the density of observations is the highest and even a small change in shape can lead to a substantial change in the classification ability.

The sensitivity analysis information obtained from this graphical analysis is similar to Härdle *et al.* (2005) and also confirms the choice combination of parameters as described in the sensitivity investigation of section 5.1. A set of alternative random subsamples as extracted from the validation set also display similar findings using the same visualization technique.

While the analysis here has been restricted to only two classes, namely bankruptcy and solvency, it can easily be

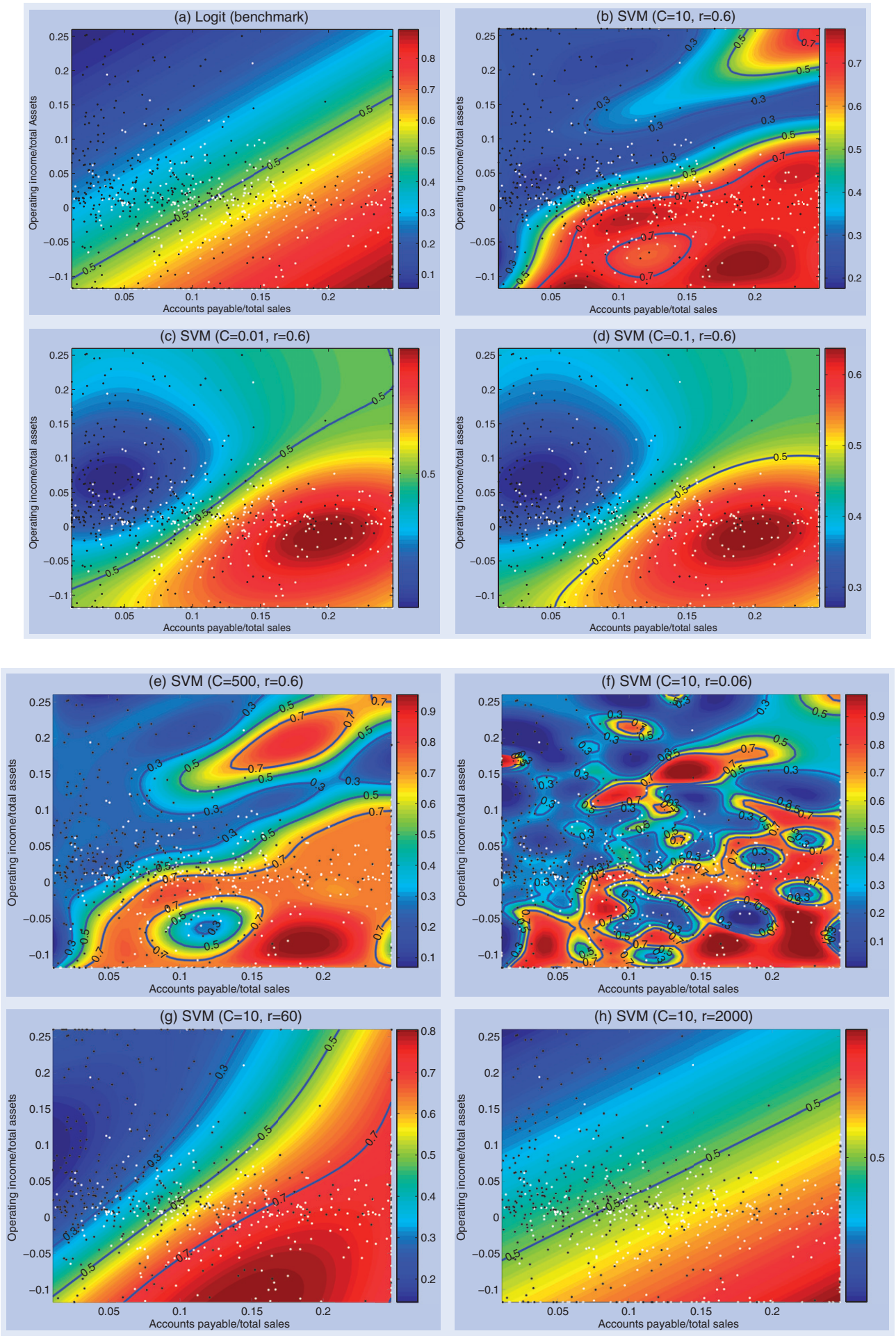

Figure 6. Default probabilities predicted for one random subsample and sensitivity analysis for the SVM.

Table 7. Median of AR measure for the best SVM model with eight important financial ratios calculated on 50/50 subsamples.

		AR median					AR median		
No.	Ratio	Logit	SVM	Predictors	No.	Ratio	Logit	SVM	Predictors
x1	NI/TA	35.12	59.93		**x15**	**CASH/TA**			**3**
x2	**NI/SALE**	**35.15**	**60.51**	**8**	x16	CASH/CL	34.87	59.42	
x3	**OI/TA**			**2**	x17	QA/CL	34.66	55.62	
x4	OI/SALE	35.06	60.44		x18	CA/CL	34.41	54.93	
x5	**EBIT/TA**			**7**	x19	WC/TA	34.72	59.48	
x6	EBITDA	34.93	59.85		x20	CL/TL	33.91	57.45	
x7	EBIT/SALE	35.14	60.4		x21	SALE/TA	35.05	56.61	
x8	OF/TA	35.04	59.64		**x22**	**INV/SALE**			**6**
x9	(OF-ITGA)/(TA-ITGA-CASH-LB)	34.94	59.42		x23	AR/SALE	35.15	59.81	
x10	CL/TA	33.94	58.19		**x24**	**AP/SALE**			**1**
x11	(CL-CASH)/TA	34.01	57.76		x25	Log(TA)	36.14	55.77	
x12	**TL/TA**			**4**	**x26**	**IDINV/INV**			**5**
x13	DEBT/TA	34.97	59.07		x27	IDL/TL	35.22	58.88	
x14	EBIT/INTE	35.03	54.37		x28	IDCASH/CASH	35.06	55.08	

generalized to multiple classes. In a multiple class case, financial analysts usually pre-specify rating classes (i.e. AAA, A, BB, C, etc.). A certain range of scores and default probabilities is associated with each rating class. The ranges are computed on the basis of historical data. According to the similarity of the scores, a new firm is assigned to one particular class. Therefore, we can draw more than one classifying function in the figure above to separate different rating classes.

5.3. Powerful predictors related to insolvent German firms

The selection procedure will be repeated for each new ratio added. The values of the AR increase until the model includes eight ratios, then they slowly decline. The medians of the AR for the models with eight ratios are shown in table 7. Most of the models tested here had AR values in the range 43.50–60.51% for out-of-sample and out-of-time tests. The results reported here are the product of the bootstrap approach described in the previous section. Obviously, the SVM-based model including ratios AP/SALE (x24), OI/TA (x3), CASH/TA (x15), TL/TA (x12), IDINV/INV (x26), INV/SALE (x22), EBIT/TA (x5) and NI/SALE (x2) attains the highest median AR, 60.51%. For comparison, we also report the median AR for the benchmark logit model with the same ratios. We can see that, for models containing the former seven ratios and one of the remaining, the medians of the AR are always higher for the SVM. This clearly reveals that the SVM-based model is always consistently superior to the benchmark logit model in identifying bankrupt firms and confirms the theoretical advantage of SVM for classification in the linear non-separable case. With respect to the percentage of correctly classified out-of-sample observations, a similar result is achieved (71.85% for the SVM-based model vs. 67.24% for the logit model).

It is noteworthy that, because the insolvency data was collected two years prior to insolvency, the predicted risk scores and calculated performance metrics in this study measure the model's ability to identify the firms that are going to default within the next two years. For example, the predicted default probability for 2002 denotes the probability that a firm defaults in 2003 or 2004.

We could not significantly improve upon our results by adding more ratios, and no model with fewer ratios performed as well. The eight selected predictors related to bankrupt German firms are AP/SALE (account payable turnover, x24), OI/TA (x3), CASH/TA (x15), TL/TA (x12), IDINV/INV (percentage of changing inventories, x26), INV/SALE (inventory turnover, x22), EBIT/TA (x5) and NI/SALE (net profit margin, x2). The size of the company was controlled in the analysis by the logarithm of the total assets (log(TA), x25). This can serve as a proxy for the cost of capital. In contrast to other studies, firm size has been shown to have no important effects on the probability of bankruptcy, which could be the result of pre-selecting only medium-sized companies.

Among the powerful predictors in identifying bankrupt German firms, there are two activity ratios (Account Payable Turnover and Inventory Turnover), three profitability ratios (OI/TA, EBIT/TA and Net Profit Margin), one liquidity ratio (CASH/TA), one leverage ratio (TL/TA) and one percentage of change ratio (Percentage of Incremental Inventories). It seems that activity ratios play the most important role in predicting the default probabilities of German firms. The activity ratio measures a firm's ability to convert different positions of their balance sheets into cash or sales. German firms will typically try to turn their accounts payable and inventories into sales as fast as possible because these will actually lead to higher revenues. Instead of ROA, EBIT/TA has a more powerful impact on insolvent German firms. In essence, it measures the operating performance and true productivity of firm assets on whose earning power the existence of the firm is based. Of course, the earnings of a firm only cannot tell the entire story. High earnings are good, but an increase in earnings does not mean that the net profit margin of a firm is improving.

For instance, if a firm has costs that have increased at a greater rate than sales, it leads to a lower profit margin. This is an indication that costs need to be under better control. Therefore, net profit margin is also very useful when analysing German bankruptcy data. In our study the liquidity ratio CASH/TA is only inferior to activity and profitability ratios when explaining German bankruptcies. Its strong explanatory power may result because the sample used in this study is mainly composed of private firms and this might not be true for public firms used in previous studies. The leverage ratio TL/TA also has a powerful influence on the identification of German bankruptcies. This metric is used to measure a firm's financial risk by determining how much of its assets have been financed by debt. This is a very broad ratio as it includes short- and long-term liabilities (debt) as well as all types of both tangible and intangible assets. The higher a firm's degree of leverage, the more the firm is considered risky. A firm with high leverage is more vulnerable to downturns in the business cycle because the firm must continue to service its debt regardless of how bad sales are. The incremental inventories provided by the Creditreform database also contain useful information for studying insolvent German firms.

To summarize our results, a German firm is most likely to go bankrupt when it has high turnover, low profits, low cash flows, is highly leveraged and has a high percentage of changing inventories. Although these results are similar to those of previous studies, the discovery of significant effects of the activity ratio and incremental inventories for predicting defaults in Germany is new.

6. Conclusions

We use a discrimination technique, the Support Vector Machine for classification, to analyse the German bankrupt company database spanning from 1996 through 2002. The identifying ability of an SVM-based nonlinear and non-parametric model is compared with that of the benchmark logit model with regard to two performance metrics (AR and misclassification error) on the basis of bootstrapped subsamples. The evidence from empirical results consistently shows that a credit risk model based on SVM significantly outperforms the benchmark linear parametric model in modeling the default risk of German firms out of sample and out of time. The sensitivity of the SVM to the penalty parameter C and Gaussian kernel coefficient r is examined according to the median of the AR using box plots (see figure 4), classification errors (see table 5) and two-dimensional visualization tools (figure 6). It is found that the discriminating ability of the SVM seems to be more sensitive to the values of r than C. Thus, appropriate trade-off values of parameters C and r should be chosen for bankruptcy analysis; for example, $C = 10$ and $r = 0.6$ in this study for the formal empirical analysis.

In addition to the unique minimum, no prior assumptions and it not being necessary to adjust the collinearity between the ratios, in particular the principle of structural risk minimization, endows the SVM approach with the most excellent classifying ability among all alternatives. Also, the SVM-based model is good at searching the linear non-separable hypersurface, which the logit model cannot do. As shown in table 4, the ratio Account Payable Turnover was selected by SVM among 28 candidates as the first best predictor to model the risk, which drastically upgrades the classifying accuracy, AR, of SVM by more than 10% as opposed to most of the other ratios selected. Otherwise, the performance gap between the SVM-based and logit model would not be so great, as shown in table 7. If the data are nonlinear, e.g. the Creditreform database, no linear model is able to separate the populations optimally, regardless of the DA, and the logit and probit models. The SVM method (as well as other pattern-recognition techniques) provides a more consistent way of finding the nonlinearities in the data, as opposed to performing an *ad-hoc* search of all possible combinations of the logit model. The holdout validation method, the most appropriate for modeling the real risk in practice, and the bootstrap re-sampling technique, guarantee the robustness and stability of the SVM approach. Due to the application of a kernel function and the sparseness of the algorithm, the achievement of such an improvement by SVM is not at a cost of much computational time, just a few seconds. Therefore, the empirical evidence confirms the theoretical advantage of SVM for classification and justifies it as applicable in practice. Of course, the non-parametric nature behind the SVM will come at the expense of understanding and insight; that is, the impact (the magnitude and direction and its significance) of the predictors on the default probabilities cannot be interpreted explicitly, in contrast to the parametric logit model. What the SVM is good at is capturing the nonlinearities better and forecasting the default probabilities more accurately than the benchmark.

As described in section 5.3, there are eight accounting ratios that are powerful predictors related to the bankruptcy of German companies. It turns out that activity ratios such as Account Payable and Inventory Turnover play the most important role in predicting the default probabilities. The percentage of incremental inventories provided by the Creditreform database also contains useful information for German bankruptcy analysis. These findings are new and somewhat different from the other default risk studies. The ability to automatically find the nonlinear dependence of the SVM model and the application of a widely accepted forward stepwise selection procedure in our case provides adequate selection that cannot be done by the usual linear classifying techniques such as the DA, logit model. That is to say, for German companies, Account Payable and Inventory Turnover, the percentage of incremental inventories selected have a strong nonlinear dependence on PDs, but a weak linear dependence that may lead to their unpopularity. Consistent with previous research, the profitability ratios, e.g. OI/TA, EBIT/TA and NI/SALE (net profit margin), are also powerful predictors related to German insolvency. Other results are similar to published research, e.g. that liquidity and leverage ratios also have

important effects on the probability of default for German companies. But, in contrast to the others, firm size (log(TA), x25) was not chosen by the forward selection procedure as a predictor, which could be the result of pre-selecting only medium-sized companies.

Acknowledgements

The authors thank the Editors, Philip Angell, David Burgoyne, Beth Cawte, Collette Teasdale, and two referees for their constructive suggestions that significantly improved the paper. This work was supported by Deutsche Forschungsgemeinschaft through SFB 649 'Economic Risk'. Shiyi Chen was also sponsored by the Shanghai Pujiang Program, the Shanghai Leading Academic Discipline Project (No. B101) and the State Innovative Institute of Project 985 at Fudan University. W.K. Härdle was also partially supported by the National Center for Theoretical Sciences (South), Taiwan. R.A. Moro was supported by the German Academic Exchange Service (DAAD).

References

Aghion, P. and Bolton, P., An 'incomplete contracts' approach to financial contracting. *Rev. Econ. Stud.*, 1992, **59**(3), 473–494.

Altman, E.I., Financial ratios, discriminant analysis and the prediction of corporate bankruptcy. *J. Finance*, 1968, **23**(4), 589–609.

Altman, E.I., Haldeman, R. and Narayanan, P., Zeta analysis: a new model to identify bankruptcy risk of corporations. *J. Bank. Finance*, 1977, **1**(1), 29–54.

Back, B., Laitinen, T. and Sere, K., Neural networks and bankruptcy prediction, in *17th Annual Congress of the European Accounting Association*, Venice, Italy, 1994. Abstract in *Collected Abstracts of the 17th Annual Congress of the European Accounting Association* 116.

Back, B., Laitinen, T., Sere, K. and Wezel, M., Choosing bankruptcy predictors using discriminant analysis, logit analysis, and genetic algorithms. Technical Report 40, TUCS Research Group, 1996.

Basle Committee on Banking Supervision, Studies on the validation of internal rating systems. AIG/RTF BIS Working Paper No. 14, 2005.

Beaver, W., Financial ratios as predictors of failures. Empirical research in accounting: Selected studies. *J. Account. Res.*, 1966, **5**(suppl.), 71–111.

Bertsekas, D.P., *Nonlinear Programming*, 1995 (Athenas Science: Belmont, MA).

Burnham, K.P. and Anderson, D.R., *Model Selection and Inference*, 1998 (Springer: New York).

Caouette, J.B., Altman, E.I. and Narayanan, P., *Managing Credit Risk: The Next Great Financial Challenge*, 1998 (Wiley: New York).

Chakrabarti, B. and Varadachari, R., Quantitative methods for default probability estimation – a first step towards Basel II. i-flex solutions, 2004.

Collins, R. and Green, R., Statistical methods for bankruptcy prediction. *J. Econ. Business*, 1982, **34**(4), 349–354.

Courant, R. and Hilbert, D., *Methods of Mathematical Physics*, Vol. I and II, 1970 (Wiley Interscience: New York).

DeLong, E.R., DeLong, D.M. and Clarke-Pearson, D.L., Comparing the areas under two or more correlated receiver operating characteristic curves: a non-parametric approach. *Biometrics*, 1988, **44**(3), 837–845.

Deng, N.Y. and Tian, Y.J., *New Methods in Data Mining: Support Vector Machine*, 2004 (Science Press: Beijing).

Efron, B. and Tibshirani, R.J., *An Introduction to the Bootstrap*, 1993 (Chapman & Hall: New York).

Engelmann, B., Hayden, E. and Tasche, D., Testing rating accuracy. *Risk*, 2003, **January**, 82–86.

Falkenstein, E., Boral, A. and Carty, L., Riskcalc for private companies: Moody's default model, Report Number: 56402, Moody's Investors Service, Inc., New York, 2000.

Fitzpatrick, P., *A Comparison of the Ratios of Successful Industrial Enterprises With Those of Failed Companies,* 1932 (The Accountants Publishing Company: Washington, DC).

Fletcher, R., *Practical Methods of Optimization*, 2nd ed., 1987 (Wiley: New York).

Friedman, C. and Sandow, S., Model performance measures for expected utility maximizing investors. *Int. J. Theor. Appl. Finance*, 2003a, **6**(4), 355–401.

Friedman, C. and Sandow, S., Learning probabilistic models: an expected utility maximization approach. *J. Mach. Learn. Res.*, 2003b, **4**, 257–291.

Friedman, C. and Huang, J., Default probability modeling: a maximum expected utility approach. Standard & Poor's Risk Solutions Group, New York, 2003.

Gaeta, G., editor, *The Certainty of Credit Risk: Its Measurement and Management*, 2003 (Wiley Finance (Asia): Singapore).

Gestel, T.V., Baesens, B., Dijcke, P.V., Suykens, J., Garcia, J. and Alderweireld, T., Linear and nonlinear credit scoring by combining logistic regression and support vector machines. *J. Credit Risk*, 2005, **1**(4), 31–60.

Giesecke, K., Credit risk modeling and valuation: An introduction. In *Credit Risk: Modeling and Management*, 2nd ed., edited by D. Shimko, pp. 487–526, 2004 (Risk Books: London).

Hanley, A. and McNeil, B., The meaning and use of the area under a receiver operating characteristics (ROC) curve. *Diagn. Radiol.*, 1982, **143**(1), 29–36.

Härdle, W., Moro, R.A. and Schäfer, D., Predicting bankruptcy with support vector machines. In *Statistical Tools for Finance and Insurance*, edited by P. Cizek, W. Härdle, and R. Weron, pp. 225–248, 2005 (Springer: Berlin).

Härdle, W., Moro, R.A. and Schäfer, D., Graphical data representation in bankruptcy analysis. In *Handbook for Data Visualization*, edited by Ch.-H. Chen, W. Härdle, and A. Unwin, pp. 853–872, 2007 (Springer: Berlin).

Härdle, W., Müller, M., Sperlich, S. and Werwatz, A., *Nonparametric and Semiparametric Models*, 2004 (Springer: Heidelberg).

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

Haykin, S., *Neural Networks: A Comprehensive Foundation*, 1999 (Prentice-Hall: Engelwood Cliffs, NJ).

Herrity, J.V., Keenan, S.C., Sobehart, J.R., Carty, L.V. and Falkenstein, E.G., Measuring private firm default risk. Moody's Investors Service Special Comment, 1999.

Hertz, J., Krogh, A. and Palmer, R.G., *The Theory of Neural Network Computation*, 1991 (Addison Welsey: Redwood, CA).

Horowitz, J.L., *The Bootstrap*, Vol. 5, 2001 (Elsevier: Amsterdam).

Keenan, S.C. and Sobehart, J.R., Performance measures for credit risk models. Research report #1-10-10-99, Moody's Risk Management Services, 1999.

Khandani, B., Lozano, M. and Carty, L., Moody's riskcalc for private companies: The German model. Rating Methodology, Moody's Investors Service, 2001.

Lennox, C., Identifying failing companies: A re-evaluation of the logit, probit and DA approaches. *J. Econ. Business*, 1999, **51**, 347–364.

Lo, A.W., Logit versus discriminant analysis: A specification test and application to corporate bankruptcies. *J. Econometr.*, 1986, **31**(2), 151–178.

Mercer, J., Functions of positive and negative type, and their connection with the theory of integral equations. *Trans. London Philos. Soc. A*, 1908, **209**, 415–446.
Merwin, C., Financing small corporations in five manufacturing industries, 1926–36. National Bureau of Economic Research, 1942.
Myers, S., Determinants of corporate borrowing. *J. Financial Econ.*, 1977, **5**(2), 147–175.
Ohlson, J., Financial ratios and the probabilistic prediction of bankruptcy. *J. Account. Res.*, 1980, **18**, 109–131.
Platt, H., Platt, M. and Pedersen, J., Bankruptcy discrimination with real variables. *J. Business, Finance Account.*, 1994, **21**(4), 491–510.
Ramser, J. and Foster, L., A demonstration of ratio analysis. Bulletin No. 40, Bureau of Business Research, University of Illinois, 1931.
Refenes, A.P., *Neural Networks in the Capital Markets*, 1995 (Wiley: Chichester).
Saunders, A. and Allen, L., *Credit Risk Measurement*, 2nd ed., 2002 (Wiley: New York).
Serrano, C., Martin, B. and Gallizo, J.L., Artificial neural networks in financial statement analysis: Ratios versus accounting data. Technical report, paper presented at the 16th Annual Congress of the European Accounting Association, Turku, Finland, April 28–30, 1993.
Shumway, T., Forecasting bankruptcy more accurately: a simple hazard model. Working Paper, University of Michigan Business School, 1998.
Sobehart, J.R., Stein, R.M., Mikityanskaya, V. and Li, L., Moody's public firm risk model: a hybrid approach to modeling default risk. Moody's Investors Service Rating Methodology, 2000.
Sobehart, J., Keenan, S. and Stein, R., Benchmarking quantitative default risk models: A validation methodology. *Algo Res. Q.*, 2001, **4**(1/2), 57–72.
Sobehart, J.R. and Keenan, S.C., Performance evaluation for credit spread and default risk models. In *Credit Risk: Models and Management*, 2nd ed., edited by D. Shimko, pp. 275–305, 2004 (Risk Books: London).
Swets, J.A., Measuring the accuracy of diagnostic systems. *Science*, 1998, **240**(4857), 1285–1293.
Swets, J.A., Dawes, R.M. and Monahan, J., Better decisions through science. *Sci. Am.*, 2000, **October**, 82–87.
Tikhonov, A.N., On solving ill-posed problem and method regularization. *Dokl. Akad. Nauk USSR*, 1963, **153**, 501–504.
Tikhonov, A.N. and Arsenin, V.Y., *Solution of Ill-posed Problems*, 1977 (W.H. Winston: Washington, DC).
Vapnik, V., *Estimation of Dependencies Based on Empirical Data*, 1979 (Nauka: Moscow).
Vapnik, V., *The Nature of Statistical Learning Theory*, 1995 (Springer: New York).
Vapnik, V., *Statistical Learning Theory*, 1997 (Wiley: New York).
Wilson, R.L. and Sharda, R., Bankruptcy prediction using neural networks. *Decis. Supp. Syst.*, 1994, **11**, 545–557.
Winakor, A. and Smith, R., Changes in the financial structure of unsuccessful industrial corporations. Bulletin No. 51, Bureau of Business Research, University of Illinois, 1935.
Zagst, R. and Hocht, S., Comparing default probability models. Working Paper, Munich University of Technology, 2006.

AStA Adv Stat Anal
DOI 10.1007/s10182-011-0182-1

ORIGINAL PAPER

Simultaneous confidence bands for expectile functions

Mengmeng Guo · Wolfgang Karl Härdle

Received: 25 February 2011 / Accepted: 9 November 2011

Abstract Expectile regression, as a general M smoother, is used to capture the tail behaviour of a distribution. Let $(X_1, Y_1), \ldots, (X_n, Y_n)$ be i.i.d. rvs. Denote by $v(x)$ the unknown τ-expectile regression curve of Y conditional on X, and by $v_n(x)$ its kernel smoothing estimator. In this paper, we prove the strong uniform consistency rate of $v_n(x)$ under general conditions. Moreover, using strong approximations of the empirical process and extreme value theory, we consider the asymptotic maximal deviation $\sup_{0\le x\le 1} |v_n(x) - v(x)|$. According to the asymptotic theory, we construct simultaneous confidence bands around the estimated expectile function. Furthermore, we apply this confidence band to temperature analysis. Taking Berlin and Taipei as an example, we investigate the temperature risk drivers to these two cities.

Keywords Expectile regression · Consistency rate · Simultaneous confidence bands · Asymmetric least squares · Kernel smoothing

1 Introduction

In regression function estimation, most investigations are concerned with the conditional mean. Geometrically, the observations $\{(X_i, Y_i),\ i = 1, \ldots, n\}$ form a cloud of points in a Euclidean space. The mean regression function focuses on the center of the point-cloud, given the covariate X, see Efron (1991). However, more insights

M. Guo (✉)
Institute for Statistics and Econometrics, Humboldt-Universität zu Berlin, Unter den Linden 6, 10099, Berlin, Germany
e-mail: guomengm@cms.hu-berlin.de

W.K. Härdle
C.A.S.E.—Center for Applied Statistics and Economics, Humboldt-Universität zu Berlin, Unter den Linden 6, 10099, Berlin, Germany
e-mail: haerdle@wiwi.hu-berlin.de

Published online: 03 December 2011

about the relation between Y and X can be gained by considering the tails of the conditional distribution.

Asymmetric least squares estimation provides a convenient and relatively efficient method of summarizing the conditional distribution of a dependent variable given the regressors. It turns out that similar to conditional percentiles, the conditional expectiles also characterize the distribution. Breckling and Chambers (1988) proposed M-quantiles, which extend this idea by a "quantile-like" generalization of regression based on asymmetric loss functions. Expectile regression, and more general M-quantile regression, can be used to characterize the relationship between a response variable and explanatory variables when the behaviour of "non-average" individuals is of interest. Jones (1994) described that expectiles and M-quantiles are related to means and quantiles are related to the median, and moreover expectiles are indeed quantiles of a transformed distribution. However, Koenker (2005) pointed out that expectiles have a more global dependence on the form of the distribution.

The expectile curves can be key aspects of inference in various economic problems and are of great interest in practice. Kuan et al. (2009) considered the conditional autoregressive expectile (CARE) model to calculate the VaR. Expectiles are also applied to calculate the expected shortfall in Taylor (2008). Moreover, Schnabel and Eilers (2009a) analyzed the relationship between gross domestic product per capita (GDP) and average life expectancy using expectile curves. Several well-developed methods already existed to estimate expectile curves. Schnabel and Eilers (2009b) combined asymmetric least square and P-splines to calculate a smooth expectile curve. In this paper, we apply the kernel smoothing techniques for the expectile curve, and construct the simultaneous confidence bands for the expectile curve, which describes a picture about the global variability of the estimator.

Let $(X_1, Y_1), \ldots, (X_n, Y_n)$ be i.i.d. rvs. We denote the joint probability density function (pdf) of the rvs is $f(x, y)$, $F(x, y)$ is the joint cumulative distribution function (cdf), conditional pdf is $f(y|x)$, $f(x|y)$ and conditional cdf $F(y|x)$, $F(x|y)$. Further, $x \in J$ with J a possibly infinite interval in $\mathbb{R}^d$ and $y \in \mathbb{R}$. In general, X may be a multivariate covariate.

From an optimization point of view, both quantile and expectile can be expressed as minimum contrast parameter estimators. Define $\rho_\tau(u) = |\mathbf{I}(u \le 0) - \tau||u|$ for $0 < \tau < 1$, then the τth quantile is expressed as $\arg\min_\theta \mathsf{E}\,\rho_\tau(y - \theta)$, where

$$\mathsf{E}\,\rho_\tau(y-\theta) = (1-\tau)\int_{-\infty}^{\theta} |y-\theta|\,dF(y|x) + \tau\int_{\theta}^{\infty} |y-\theta|\,dF(y|x)$$

where θ is the estimator of the τ expectile, and define $\theta \in I$, where the compact set $I \subset \mathbb{R}$. With the interpretation of the contrast function $\rho_\tau(u)$ as the negative log likelihood of asymmetric Laplace distribution, we can see the τth quantile as a quasi maximum estimator in the location model. Changing the loss (contrast) function to

$$\rho_\tau(u) = \big|\mathbf{I}(u \le 0) - \tau\big|u^2, \quad \tau \in (0,1) \tag{1}$$

leads to expectile. Note that for $\tau = \frac{1}{2}$, we obtain the mean respective to the sample average. Putting this into a regression framework, we define the conditional expectile

function (to level τ) as

$$v(x) = \arg\min_{\theta} \mathsf{E}\big\{\rho_\tau(y-\theta)|X=x\big\} \tag{2}$$

Inserting (1) into (2), we obtain the expected loss function:

$$\mathsf{E}\big\{\rho_\tau(y-\theta)|X=x\big\} = (1-\tau)\int_{-\infty}^{\theta}(y-\theta)^2\,dF(y|x) + \tau\int_{\theta}^{\infty}(y-\theta)^2\,dF(y|x) \tag{3}$$

From now on, we silently assume τ is fixed therefore we suppress the explicit notion. Recall that the conditional quantile $l(x)$ at level τ can be considered as

$$l(x) = \inf\big\{y \in \mathbb{R}|F(y|x) \geq \tau\big\}$$

Therefore, the proposed estimate $l_n(x)$ can be expressed:

$$l_n(x) = \inf\big\{y \in \mathbb{R}|\widehat{F}(y|x) \geq \tau\big\}$$

where $\widehat{F}(y|x)$ is the kernel estimator of $F(y|x)$:

$$\widehat{F}(y|x) = \frac{\sum_{i=1}^{n} K_h(x-X_i)\mathbf{I}(Y_i \leq y)}{\sum_{i=1}^{n} K_h(x-X_i)}$$

In the same spirit, define $G_{Y|x}(\theta)$ as

$$G_{Y|x}(\theta) = \frac{\int_{-\infty}^{\theta}|y-\theta|\,dF(y|x)}{\int_{-\infty}^{\infty}|y-\theta|\,dF(y|x)}$$

Replacing θ by $v(x)$, we get

$$G_{Y|x}(v) = \frac{\int_{-\infty}^{v(x)}|y-v(x)|\,dF(y|x)}{\int_{-\infty}^{\infty}|y-v(x)|\,dF(y|x)} = \tau$$

so $v(x)$ can be equivalently seen as solving: $G_{Y|x}(\theta) - \tau = 0$ (w.r.t. θ). Therefore,

$$v(x) = G_{Y|x}^{-1}(\tau)$$

with the τth expectile curve kernel smoothing estimator:

$$v_n(x) = \hat{G}_{Y|x}^{-1}(\tau)$$

where the nonparametric estimate of $G_{Y|x}(v)$ is

$$\hat{G}_{Y|x}(\theta) = \frac{\sum_{i=1}^{n} K_h(x-X_i)\,\mathbf{I}(Y_i < y)|y-\theta|}{\sum_{i=1}^{n} K_h(x-X_i)|y-\theta|}$$

Quantiles and expectiles both characterize a distribution function although they are different in nature. As an illustration, Fig. 1 plots curves of quantiles and expectiles

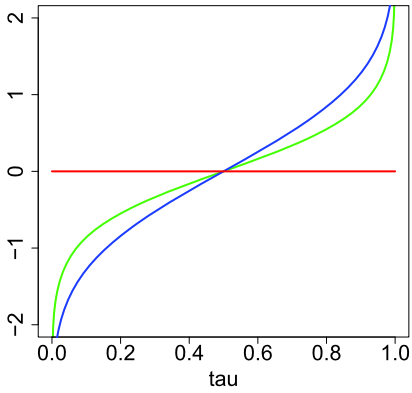

Fig. 1 (Color online) Quantile curve (*blue*) and expectile curve (*green*) for standard normal distribution

of the standard normal $N(0,1)$. Obviously, there is a one-to-one mapping between quantile and expectile, see Yao and Tong (1996). For fixed x, define $w(\tau)$ such that $v_{w(\tau)}(x) = l(x)$, then $w(\tau)$ is related to the τth quantile curve $l(x)$ via

$$w(\tau) = \frac{\tau l(x) - \int_{-\infty}^{l(x)} y\, dF(y|x)}{2\,\mathsf{E}(Y|x) - 2\int_{-\infty}^{l(x)} y\, dF(y|x) - (1-2\tau)l(x)} \tag{4}$$

$l(x)$ is an increasing function of τ, therefore, $w(\tau)$ is also a monotonically increasing function. Expectiles correspond to quantiles with this transformation w. However, it is not straightforward to apply (4), since it depends on the conditional distribution of the regressors. For very simple distributions, it is not hard to calculate the transformation $w(\tau)$, for example, $Y \sim U(-1,1)$, then $w(\tau) = \tau^2/(2\tau^2 - 2\tau + 1)$. However, if the distribution is more complicated, even worse, the conditional distribution is unknown, it is hard to apply this transformation, see Jones (1994). Therefore, it is not feasible to calculate expectiles from the corresponding quantiles.

In the current paper, we apply the methodology to weather studies. Weather risk is an uncertainty caused by weather volatility. Energy companies take positions in weather risk if it is a source of financial uncertainty. However, weather is also a local phenomenon, since the location, the atmosphere, human activities and some other factors influence the temperature. We investigate whether such local factors exist. Taking two cities, Berlin and Taipei, as an example, we check whether the performance of high expectiles and low expectiles of temperature varies over time. To this end, we calculate the expectiles of trend and seasonality corrected temperature.

The structure of this paper is as follows. In Sect. 2, the stochastic fluctuation of the process $\{v_n(x) - v(x)\}$ is studied and the simultaneous confidence bands are presented through the equivalence of several stochastic processes. We calculate the asymptotic distribution of $v_n(x)$, and the strong uniform consistency rate of $\{v_n(x) - v(x)\}$ is discussed in this section. In Sect. 3, a Monte Carlo study is to investigate the behaviour of $v_n(x)$ when the data are generated with the error terms standard normally distributed. Section 4 considers an application in the temperature of Berlin and Taipei. All proofs are attached in Appendix.

2 Results

In light of the concepts of M-estimation as in Huber (1981), if we define $\psi(u)$ as

$$\begin{aligned}\psi(u) &= \frac{\partial \rho(u)}{\partial u} \\ &= \big|\mathbf{I}(u \le 0) - \tau\big| u \\ &= \big\{\tau - \mathbf{I}(u \le 0)\big\}|u|\end{aligned}$$

$v_n(x)$ and $v(x)$ can be treated as a zero (w.r.t. θ) of the function:

$$H_n(\theta, x) \stackrel{\text{def}}{=} n^{-1} \sum_{i=1}^{n} K_h(x - X_i)\psi(Y_i - \theta) \tag{5}$$

$$H(\theta, x) \stackrel{\text{def}}{=} \int_{\mathbb{R}} f(x, y)\psi(y - \theta)\, dy \tag{6}$$

respectively.

Härdle (1989) has constructed the uniform confidence bands for general M-smoothers. Härdle and Song (2009) studied the uniform confidence bands for quantile curves. In our paper, we investigate expectile curves, one kind of M-smoother. The loss function for quantile regression is not differentiable, however it is differentiable for expectile when it is in the asymmetric quadratic form. Therefore, by employing similar methods as those developed in Härdle (1989), it is shown in this paper that

$$\begin{aligned}&\mathrm{P}\Big[(2\delta \log n)^{1/2}\Big\{\sup_{x \in J} r(x)\big|v_n(x) - v(x)\big|/\lambda(K)^{1/2} - d_n\Big\} < z\Big] \\ &\quad \longrightarrow \quad \exp\big\{-2\exp(-z)\big\}, \quad \text{as } n \to \infty\end{aligned} \tag{7}$$

with some adjustment of $v_n(x)$, we can see that the supreme of $v_n(x) - v(x)$ follows the asymptotic Gumbel distribution, where $r(x)$, δ, $\lambda(K)$, d_n are suitable scaling parameters. The asymptotic result (7) therefore allows the construction of simultaneous confidence bands for $v(x)$ based on specifications of the stochastic fluctuation of $v_n(x)$. The strong approximation with Brownian bridge techniques is applied in this paper to prove the asymptotic distribution of $v_n(x)$.

To construct the confidence bands, we make the following necessary assumptions about the distribution of (X, Y) and the score function $\psi(u)$ in addition to the existence of an initial estimator whose error is a.s. uniformly bounded.

(A1) The kernel $K(\cdot)$ is positive, symmetric, has compact support $[-A, A]$ and is Lipschitz continuously differentiable with bounded derivatives.

(A2) $(nh)^{-1/2}(\log n)^{3/2} \to 0$, $(n \log n)^{1/2}h^{5/2} \to 0$, $(nh^3)^{-1}(\log n)^2 \le M$, M is a constant.

(A3) $h^{-3}(\log n) \int_{|y|>a_n} f_Y(y)\, dy = \mathcal{O}(1)$, $f_Y(y)$ the marginal density of Y, $\{a_n\}_{n=1}^{\infty}$ a sequence of constants tending to infinity as $n \to \infty$.

(A4) $\inf_{x \in J} |p(x)| \ge p_0 > 0$, where $p(x) = \partial\, \mathsf{E}\{\psi(Y - \theta)|x\}/\partial\theta|_{\theta = v(x)} \cdot f_X(x)$, where $f_X(x)$ is the marginal density of X.

(A5) The expectile function $v(x)$ is Lipschitz twice continuously differentiable, for all $x \in J$.

(A6) $0 < m_1 \leq f_X(x) \leq M_1 < \infty$, $x \in J$, and the conditional density $f(\cdot|y)$, $y \in \mathbb{R}$, is uniform locally Lipschitz continuous of order $\tilde{\alpha}$ (ulL-$\tilde{\alpha}$) on J, uniformly in $y \in \mathbb{R}$, with $0 < \tilde{\alpha} \leq 1$, and $\psi(x)$ is piecewise twice continuously differentiable.

Define also

$$\sigma^2(x) = \mathsf{E}\big[\psi^2\big\{Y - v(x)\big\}|x\big]$$

$$H_n(x) = (nh)^{-1} \sum_{i=1}^{n} K\big\{(x - X_i)/h\big\}\psi\big\{Y_i - v(x)\big\}$$

$$D_n(x) = (nh)^{-1} \frac{\partial \sum_{i=1}^{n} K\{(x - X_i)/h\}\psi\{Y_i - \theta\}}{\partial\theta}\bigg|_{\theta=v(x)}$$

and assume that $\sigma^2(x)$ and $f_X(x)$ are differentiable.

Assumption (A1) on the compact support of the kernel could possibly be relaxed by introducing a cutoff technique as in Csörgö and Hall (1982) for density estimators. Assumption (A2) has purely technical reasons: to keep the bias at a lower rate than the variance and to ensure the vanishing of some non-linear remainder terms. Assumption (A3) appears in a somewhat modified form also in Johnston (1982). Assumption (A4) guarantees that the first derivative of the loss function, i.e. $\psi(u)$ is differentiable. Assumptions (A5) and (A6) are common assumptions in robust estimation as in Huber (1981), Härdle et al. (1988) that are satisfied by exponential, and generalized hyperbolic distributions.

Zhang (1994) has proved the asymptotic normality of the nonparametric expectile. Under the Assumptions (A1) to (A4), we have

$$\sqrt{nh}\big\{v_n(x) - v(x)\big\} \xrightarrow{\mathcal{L}} \mathrm{N}\big\{0, V(x)\big\} \tag{8}$$

with

$$V(x) = \lambda(K) f_X(x)\sigma^2(x)/p(x)^2$$

where we can denote

$$\begin{aligned}
\lambda(K) &= \int_{-A}^{A} K^2(u)\,du \\
\sigma^2(x) &= \mathsf{E}\big[\psi^2\big\{Y - v(x)\big\}|x\big] \\
&= \int \psi^2\big\{y - v(x)\big\}\,dF(y|x) \\
&= \tau^2 \int_{v(x)}^{\infty} \big\{y - v(x)\big\}^2\,dF(y|x) + (1-\tau)^2 \int_{-\infty}^{v(x)} \big\{y - v(x)\big\}^2\,dF(y|x) \qquad (9)
\end{aligned}$$

$$p(x) = \mathsf{E}\big[\psi'\big\{Y - v(x)\big\}|x\big] \cdot f_X(x)$$
$$= \left\{\tau \int_{v(x)}^{\infty} dF(y|x) + (1-\tau)\int_{-\infty}^{v(x)} dF(y|x)\right\} \cdot f_X(x) \tag{10}$$

For the uniform strong consistency rate of $v_n(x) - v(x)$, we apply the result of Härdle et al. (1988) by taking $\beta(y) = \psi(y-\theta)$, $y \in \mathbb{R}$, for $\theta \in I$, $q_1 = q_2 = -1$, $\gamma_1(y) = \max\{0, -\psi(y-\theta)\}$, $\gamma_2(y) = \min\{0, -\psi(y-\theta)\}$ and $\lambda = \infty$ to satisfy the representations for the parameters there. We have the following lemma under some specified assumptions:

Lemma 1 *Let $H_n(\theta, x)$ and $H(\theta, x)$ be given by (5) and (6). Under Assumption* (A6) *and $(nh/\log n)^{1/2} \to \infty$ through Assumption* (A2), *for some constant A^* not depending on n, we have a.s. as $n \to \infty$*

$$\sup_{\theta\in I}\sup_{x\in J}\big|H_n(\theta, x) - H(\theta, x)\big| \le A^* \max\big\{(nh/\log n)^{-1/2}, h^{\tilde{\alpha}}\big\} \tag{11}$$

For our result on $v_n(\cdot)$, we shall also require

$$\inf_{x\in J}\left|\int \psi\big\{y - v(x) + \varepsilon\big\}\, dF(y|x)\right| \ge \tilde{q}|\varepsilon|, \quad \text{for } |\varepsilon| \le \delta_1 \tag{12}$$

where δ_1 and $\tilde{q}$ are some positive constants, see also Härdle and Luckhaus (1984). This assumption is satisfied if there exists a constant $\tilde{q}$ such that $f\{v(x)|x\} > \tilde{q}/p$, $x \in J$.

Theorem 1 *Under the conditions of Lemma 1 and also assuming (12) holds, we have a.s. as $n \to \infty$*

$$\sup_{x\in J}\big|v_n(x) - v(x)\big| \le B^* \max\big\{(nh/\log n)^{-1/2}, h^{\tilde{\alpha}}\big\} \tag{13}$$

with $B^* = A^*/m_1\tilde{q}$ not depending on n and m_1 a lower bound of $f_X(x)$. If additionally $\tilde{\alpha} \ge \{\log(\sqrt{\log n}) - \log(\sqrt{nh})\}/\log h$, it can be further simplified to

$$\sup_{x\in J}\big|v_n(x) - v(x)\big| \le B^*\big\{(nh/\log n)^{-1/2}\big\}$$

Theorem 2 *Let $h = n^{-\delta}$, $\frac{1}{5} < \delta < \frac{1}{3}$ with $\lambda(K)$ as defined before, and*

$$d_n = (2\delta\log n)^{1/2} + (2\delta\log n)^{-1/2}\left[\log\big\{c_1(K)/\pi^{1/2}\big\} + \frac{1}{2}(\log\delta + \log\log n)\right]$$
$$\text{if } c_1(K) = \big\{K^2(A) + K^2(-A)\big\}/\big\{2\lambda(K)\big\} > 0$$
$$d_n = (2\delta\log n)^{1/2} + (2\delta\log n)^{-1/2}\log\big\{c_2(K)/2\pi\big\}$$
$$\text{otherwise with } c_2(K) = \int_{-A}^{A}\big\{K'(u)\big\}^2\, du/\big\{2\lambda(K)\big\}$$

Then (7) *holds with*

$$r(x) = (nh)^{-\frac{1}{2}} p(x) \left\{ \frac{f_X(x)}{\sigma^2(x)} \right\}^{\frac{1}{2}}$$

This theorem can be used to construct uniform confidence intervals for the regression function as stated in the following corollary.

Corollary 1 *Under the assumptions of the theorem above, an approximate* $(1-\alpha) \times 100\%$ *confidence band over* $[0, 1]$ *is*

$$v_n(x) \pm (nh)^{-1/2} \left\{ \hat{\sigma}^2(x) \lambda(K) / \hat{f}_X(x) \right\}^{1/2} \hat{p}^{-1}(x) \left\{ d_n + c(\alpha)(2\delta \log n)^{-1/2} \right\}$$

where $c(\alpha) = \log 2 - \log |\log(1-\alpha)|$ *and* $\hat{f}_X(x)$, $\hat{\sigma}^2(x)$ *and* $\hat{p}(x)$ *are consistent estimates for* $f_X(x)$, $\sigma^2(x)$ *and* $p(x)$.

With $\sqrt{V(x)}$ introduced, we can further write Corollary 1 as

$$v_n(x) \pm (nh)^{-1/2} \left\{ d_n + c(\alpha)(2\delta \log n)^{-1/2} \right\} \sqrt{\hat{V}(x)}$$

where $\hat{V}(x)$ is the nonparametric estimator of $V(x)$. Bandwidth selection is quite crucial in kernel smoothing. In this paper, we use the optimal bandwidth discussed in Zhang (1994), which has the following form

$$h_n^{\text{opt}} = \left(\frac{\sigma^2(x) \lambda(K)}{n[\Lambda\{v(x)|x\}]^2 [\int \{y - v(x)\}^2 K^2\{y - v(x)\} \, dF(y|x)]^2} \right)^{1/5} \tag{14}$$

where

$$\Lambda(\theta|x) = \frac{\partial^2 \psi(\theta|x-u)}{\partial u^2}|_{u=0}$$

The proof is essentially based on a linearization argument after a Taylor series expansion. The leading linear term will then be approximated in a similar way as in Johnston (1982), Bickel and Rosenblatt (1973). The main idea behind the proof is a strong approximation of the empirical process of $\{(X_i, Y_i)_{i=1}^n\}$ by a sequence of Brownian bridges as proved by Tusnady (1977).

As $v_n(x)$ is the zero (w.r.t. θ) of $H_n(\theta, x)$, it follows by applying second-order Taylor expansions to $H_n(\theta, x)$ around $v(x)$ that

$$v_n(x) - v(x) = \left\{ H_n(x) - \mathsf{E}\, H_n(x) \right\} \big/ p(x) + R_n(x) \tag{15}$$

where $\{H_n(x) - \mathsf{E}\, H_n(x)\}/p(x)$ is the leading linear term and the remainder term is written as

$$\begin{aligned} R_n(x) = {} & H_n(x) \left\{ p(x) - D_n(x) \right\} \big/ \left\{ D_n(x) \cdot p(x) \right\} + \mathsf{E}\, H_n(x)/p(x) \\ & + \frac{1}{2} \left\{ v_n(x) - v(x) \right\}^2 \cdot \left\{ D_n(x) \right\}^{-1} \end{aligned} \tag{16}$$

$$\cdot (nh)^{-1} \sum_{i=1}^{n} K\{(x - X_i)/h\} \psi''\{Y_i - v(x) + r_n(x)\}, \tag{17}$$

$$|r_n(x)| < |v_n(x) - v(x)|.$$

We show in Appendix that (Lemma 4) that $\|R_n\| = \sup_{x \in J} |R_n(x)| = \mathcal{O}_p\{(nh \log n)^{-1/2}\}$.

Furthermore, the rescaled linear part

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

is approximated by a sequence of Gaussian processes, leading finally to the Gaussian process

$$Y_{5,n}(x) = h^{-1/2} \int K\{(x - t)/h\}\, dW(x) \tag{18}$$

Drawing upon the result of Bickel and Rosenblatt (1973), we finally obtain asymptotically the Gumbel distribution.

We also need the Rosenblatt (1952) transformation,

$$T(x, y) = \{F_{X|y}(x|y), F_Y(y)\}$$

which transforms (X_i, Y_i) into $T(X_i, Y_i) = (X_i', Y_i')$ mutually independent uniform rv's. In the event that x is a d-dimension covariate, the transformation becomes

$$T(x_1, x_2, \ldots, x_d, y) = \{F_{X_1|y}(x_1|y), F_{X_2|y}(x_2|x_1, y), \ldots, F_{X_k|x_{d-1},\ldots,x_1,y}(x_k|x_{d-1}, \ldots, x_1, y), F_Y(y)\} \tag{19}$$

With the aid of this transformation, Theorem 1 of Tusnady (1977) may be applied to obtain the following lemma.

Lemma 2 *On a suitable probability space a sequence of Brownian bridges B_n exists that*

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

where $Z_n(x, y) = n^{1/2}\{F_n(x, y) - F(x, y)\}$ denotes the empirical process of $\{(X_i, Y_i)\}_{i=1}^n$.

For $d > 2$, it is still an open problem which deserves further research.

Before we define the different approximating processes, let us first rewrite (18) as a stochastic integral w.r.t. the empirical process $Z_n(x, y)$,

$$Y_n(x) = \{hg'(x)\}^{-1/2} \iint K\{(x - t)/h\} \psi\{y - v(x)\}\, dZ_n(t, y)$$

$$g'(x) = \sigma^2(x) f_X(x)$$

The approximating processes are now

$$Y_{0,n}(x) = \{hg(x)\}^{-1/2} \iint_{\Gamma_n} K\{(x-t)/h\}\psi\{y - v(x)\}\, dZ_n(t, y)$$

where $\Gamma_n = \{|y| \le a_n\}$,

$$g(t) = \mathsf{E}\big[\psi^2\{y - v(x)\} \cdot \mathbf{I}(|y| \le a_n)\big|X = x\big] \cdot f_X(x) \tag{20}$$

$$Y_{1,n}(x) = \{hg(x)\}^{-1/2} \iint_{\Gamma_n} K\{(x-t)/h\}\psi\{y - v(x)\}\, dB_n\{T(t, y)\}$$

$\{B_n\}$ being the sequence of Brownian bridges from Lemma 2 (21)

$$Y_{2,n}(x) = \{hg(x)\}^{-1/2} \iint_{\Gamma_n} K\{(x-t)/h\}\psi\{y - v(x)\}\, dW_n\{T(t, y)\}$$

$\{W_n\}$ being the sequence of Wiener processes satisfying

$$B_n(t', y') = W_n(t', y') - t'y'W_n(1, 1) \tag{22}$$

$$Y_{3,n}(x) = \{hg(x)\}^{-1/2} \iint_{\Gamma_n} K\{(x-t)/h\}\psi\{y - v(t)\}\, dW_n\{T(t, y)\} \tag{23}$$

$$Y_{4,n}(x) = \{hg(x)\}^{-1/2} \int g(t)^{1/2} K\{(x-t)/h\}\, dW(t) \tag{24}$$

$$Y_{5,n}(x) = h^{-1/2} \int K\{(x-t)/h\}\, dW(t)$$

$\{W(\cdot)\}$ being the Wiener process (25)

Lemmas 5 to 10 ensure that all these processes have the same limit distributions. The result then follows from

Lemma 3 (Theorem 3.1 in Bickel and Rosenblatt 1973) *Let d_n, $\lambda(K)$, δ as in Theorem 2. Let*

$$Y_{5,n}(x) = h^{-1/2} \int K\{(x-t)/h\}\, dW(t)$$

Then, as $n \to \infty$, the supremum of $Y_{5,n}(x)$ has a Gumbel distribution.

$$\mathrm{P}\Big\{(2\delta \log n)^{1/2}\Big[\sup_{x\in J}|Y_{5,n}(x)|/\{\lambda(K)\}^{1/2} - d_n\Big] < z\Big\} \to \exp\{-2\exp(-z)\}$$

Same as quantile, the supremum of a nonparametric expectile converge to its limit at a rate $(\log n)^{-1}$. We do not check the bootstrap confidence bands in this paper, which can be future work. Instead, we point out several well documented literature about this issue. For example, Claeskens and Keilegom (2003) discussed the bootstrap confidence bands for regression curves and their derivatives. Partial linear quantile regression and bootstrap confidence bands are well studied in Härdle et al. (2010). They proved that the convergence rate by bootstrap approximation to the dis-

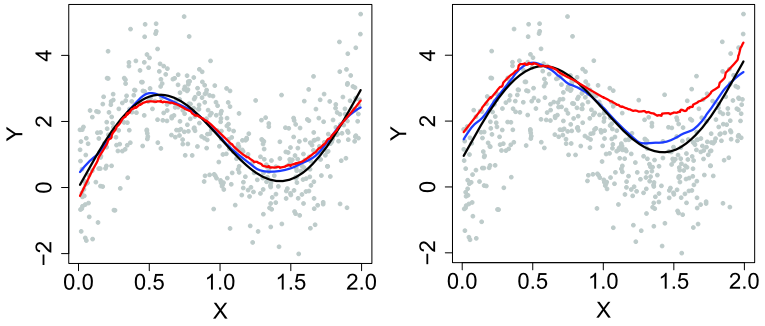

Fig. 2 (Color online) $\tau = 0.5$ (*left*) and $\tau = 0.9$ (*right*) estimated quantile and expectile plot. Quantile curve, theoretical expectile curve, estimated expectile curve

tribution of the supremum of a quantile estimate has been improved from $(\log n)^{-1}$ to $n^{-2/5}$.

3 A Monte Carlo study

In the design of the simulation, we generate bivariate random variables $\{(X_i, Y_i)\}_{i=1}^{n}$ with sample size $n = 50, n = 100, n = 200, n = 500$. The covariate X is uniformly distributed on $[0, 2]$

$$Y = 1.5X + 2\sin(\pi X) + \varepsilon \tag{26}$$

where $\varepsilon \sim \mathrm{N}(0, 1)$.

Obviously, the theoretical expectiles (fixed τ) are determined by

$$v(x) = 1.5x + 2\sin(\pi x) + v_N(\tau) \tag{27}$$

where $v_N(\tau)$ is the τth expectile of the standard Normal distribution.

Figure 2 (in the left part) describes the simulated data (the grey points), together with the 0.5 estimated quantile and estimated expectile and theoretical expectile curves, which represents, respectively, the conditional median and conditional mean. The conditional mean and conditional median coincide with each other, since the error term is symmetrically distributed, which is obvious in Fig. 2. In the right part of the figure, we consider the conditional 0.9 quantile and expectile curves. Via a transformation (4), there is a gap between the quantile curve and the expectile curve. By calculating $w(\tau)$ for the standard normal distribution, the 0.9 quantile can be expressed by the around 0.96 expectile. The estimated expectile curve is close to the theoretical one.

Figure 3 shows the 95% uniform confidence bands for expectile curve, which are represented by the two red dashed lines. We calculate both 0.1 (left) and 0.9 (right) expectile curves. The black lines stand for the corresponding 0.1 and 0.9 theoretical expectile curves, and the blue lines are the estimated expectile curves. Obviously, the theoretical expectile curves locate in the confidence bands.

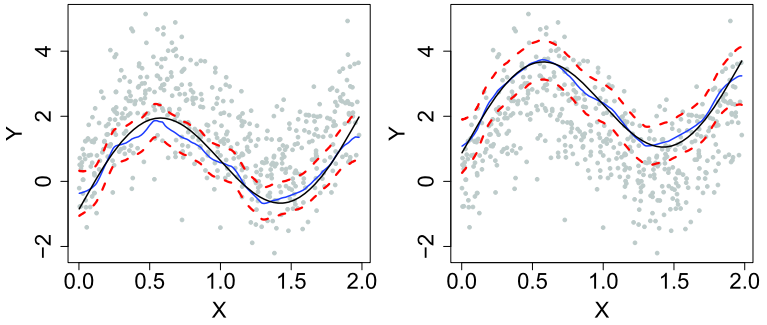

Fig. 3 Uniform confidence bands for expectile curve for $\tau = 0.1$ (*left*) and $\tau = 0.9$ (right). Theoretical expectile curve, estimated expectile curve and 95% uniform confidence bands

Table 1 Simulated coverage probabilities of 95% confidence bands for 0.9 expectile with 500 runs of simulation. *cp* stands for the coverage probability, and *h* is the width of the band

n	*cp*	*h*
50	0.526	1.279
100	0.684	1.093
200	0.742	0.897
500	0.920	0.747

Table 2 Simulated coverage probabilities of 95% confidence bands for 0.1 expectile with 500 runs of simulation. *cp* stands for the coverage probability, and *h* is the width of the band

n	*cp*	*h*
50	0.386	0.859
100	0.548	0.768
200	0.741	0.691
500	0.866	0.599

To check the performance of the calculated confidence bands, we compare the simulated coverage probability with the nominal values for coverage probability 95% for different sample sizes. We apply this method to both 0.9 and 0.1 expectile. Table 1 and Table 2 present the corresponding results. We run the simulation 500 times for each scenario. Obviously, the coverage probabilities improve with the increased the sample size, and the width of the bands h becomes smaller for both 0.9 and 0.1 expectile. It is noteworthy that when the number of observation is large enough, for example $n = 500$, the coverage probability is very close to the nominal probability, especially for the 0.9 expectile.

4 Application

In this part, we apply the expectile into the temperature study. We consider the daily temperature both of Berlin and Taipei, ranging from 19480101 to 20071231, together 21900 observations for each city. The statistical properties of the temperature are

Table 3 Statistical summary of the temperature in Berlin and Taipei

	Mean	SD	Skewness	Kurtosis	Max	Min
Berlin	9.66	7.89	−0.315	2.38	30.4	−18.5
Taipei	22.61	5.43	−0.349	2.13	33.0	6.5

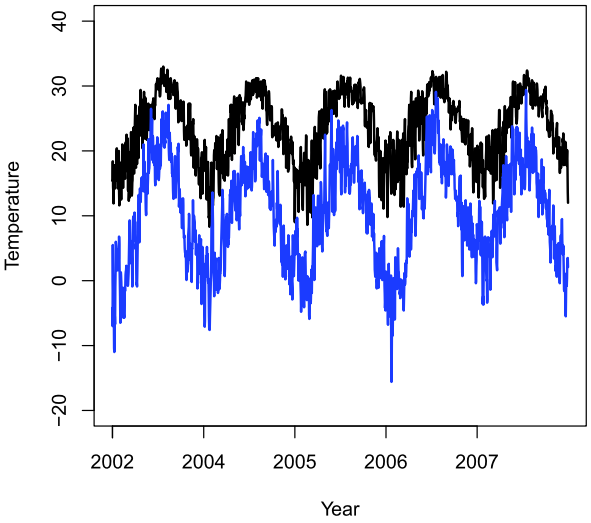

Fig. 4 (Color online) The time series plot of the temperature in Berlin and Taipei from 2002–2007. *The black line* stands for the temperature in Taipei, and *the blue line* is in Berlin

summarized in Table 3. The Berlin temperature data were obtained from Deutscher Wetterdienst, and the Taipei temperature data were obtained from the center for adaptive data analysis in National Central University.

Before proceeding to detailed modeling and forecasting results, it is useful to get an overall view of the daily average temperature data. Figure 4 displays the average temperature series of the sample from 2002 to 2007. The black line stands for the temperature in Taipei, and the blue line describes for the temperature in Berlin. The time series plots reveal strong and unsurprising seasonality in average temperature: in each city, the daily average temperature moves repeatedly and regularly through periods of high temperature (summer) and low temperature (winter). It is well documented that seasonal volatility in the regression residuals appears highest during the winter months where the temperature shows high volatility. Importantly, however, the seasonal fluctuations differ noticeably across cities both in terms of amplitude and detail of pattern.

Based on the observed pattern, we apply a stochastic model with seasonality and inter temporal autocorrelation, as in Benth et al. (2007). To understand the model clearly, let us introduce the time series decomposition of the temperature, with $t =$

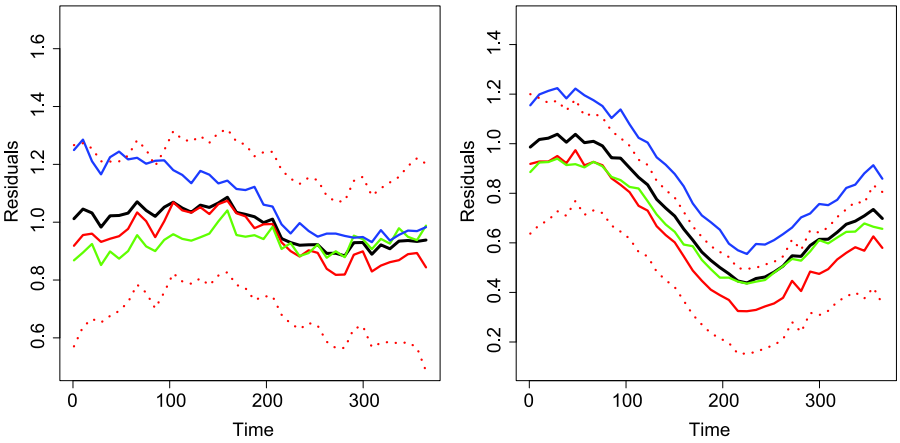

Fig. 5 0.9 expectile curves for Berlin (*left*) and Taipei (*right*) daily temperature residuals from 1948–2007 with the 95% uniform confidence bands for the first 20 years expectile

$1, \ldots, 365$ days, and $j = 0, \ldots, J$ years:

$$\begin{aligned} X_{365j+t} &= T_{t,j} - \Lambda_t \\ X_{365j+t} &= \sum_{l=1}^{L} \beta_{lj} X_{365j+t-l} + \varepsilon_{t,j} \\ \Lambda_t &= a + bt + \sum_{m=1}^{M} c_l \cos\left\{\frac{2\pi(t-d_m)}{l \cdot 365}\right\} \end{aligned} \tag{28}$$

where $T_{t,j}$ is the temperature at day t in year j, and Λ_t denotes the seasonality effect. Motivation of this modeling approach can be found in Diebold and Inoue (2001). Further studies as Campbell and Diebold (2005) has provided evidence that the parameters β_{lj} are likely to be j independent and hence estimated consistently from a global autoregressive process $AR(L_j)$ model with $L_j = L$. The analysis of the partial autocorrelations and Akaike's Information Criterion (AIC) suggests that a simple $AR(3)$ model fits well the temperature evolution both in Berlin and Taipei.

In this paper, the risk factor of temperature, which is the residual $\hat{\varepsilon}_{t,j}$ from (28), is studied in the expectile regression. We intend to construct the confidence bands for the 0.01 and 0.9 expectile curves for the volatility of temperature. It is interesting to check whether the extreme values perform differently in different cities.

The left part of the figures describes the expectile curves for Berlin, and the right part is for Taipei. In each figure, the thick black line depicts the average expectile curve with the data from 1948 to 2007. The red line is the expectile for the residuals from (28) with the data of the first 20 years temperature, i.e. in the period from 1948 to 1967. The 0.9 expectile for the second 20 years (1968–1987) residuals is described by the green line, and the blue line stands for the expectile curve in the latest 20 years (1988–2007). The dotted lines are the 95% confidence bands corresponding to the expectile curve with the same color. Figures 5, 6 and 7 describe the 0.9 expectile curves

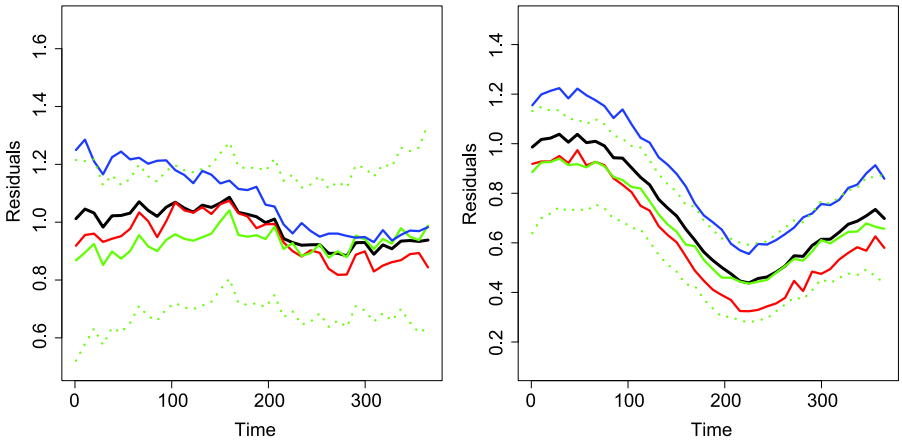

Fig. 6 0.9 expectile curves for Berlin (*left*) and Taipei (*right*) daily temperature residuals from 1948–2007 with the 95% uniform confidence bands for the second 20 years expectile

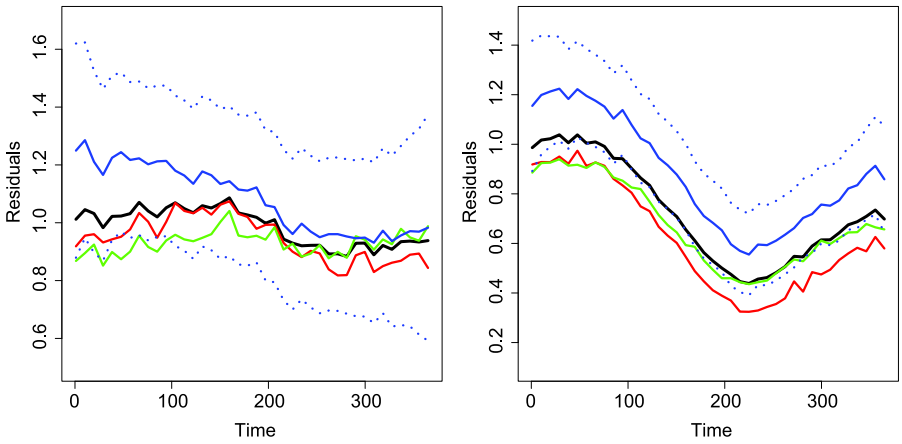

Fig. 7 0.9 expectile curves for Berlin (*left*) and Taipei (*right*) daily temperature residuals from 1948–2007 with the 95% uniform confidence bands for the latest 20 years expectile

for Berlin and Taipei, as well as their corresponding confidence bands. Obviously, the variance is higher in winter–earlier summer both in Berlin and Taipei.

Note that the behaviour of expectile curves in Berlin and Taipei is quite different. Firstly, the variation of the expectiles in Berlin is smaller than that of Taipei. All the expectile curves cross with each other in the last 100 observations of the year for Berlin, and the variance in this period is smaller. Moreover, all of these curves nearly locate in the corresponding three confidence bands. However, the performance of the expectile in Taipei is quite different from that of Berlin. The expectile curves for Taipei have similar trends for each 20 years. They have highest volatilities in January, and lowest volatility in July. More interestingly, the expectile curve for the latest 20 years does not locate in the confidence bands constructed using the data from the

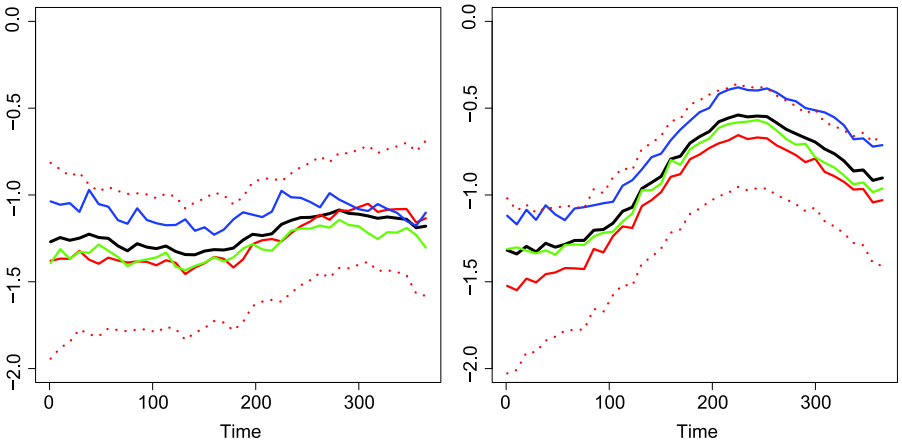

Fig. 8 0.01 expectile curves for Berlin (*left*) and Taipei (*right*) daily temperature residuals from 1948–2007 with the 95% uniform confidence bands for the first 20 years expectile

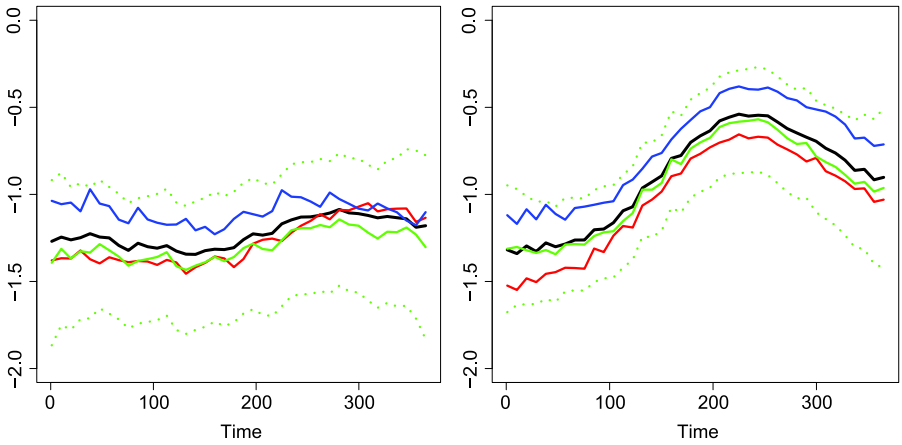

Fig. 9 0.01 expectile curves for Berlin (*left*) and Taipei (*right*) daily temperature residuals from 1948–2007 with the 95% uniform confidence bands for the second 20 years expectile

first 20 years and second 20 years, see Figs. 5 and 7. Similarly, the expectile curve for the first 20 years does not locate in the confidence bands constructed using the information from the latest 20 years.

Further, let us study low expectile for the residuals of the temperature in Berlin and Taipei. It is hard to calculate very small percentage of quantile curves, due to the sparsity of the data, expectiles though can overcome this drawback. One can calculate very low or very high expectiles, such as 0.01 and 0.99 expectile curves, even when there are not so many observations. Display of the 0.01 expectiles for the residuals and their corresponding confidence bands is given in Figs. 8, 9 and 10. One can detect that the shapes of the 0.01 expectile for Berlin and Taipei are different. It does not fluctuate a lot during the whole year in Berlin, while the variation in Taipei

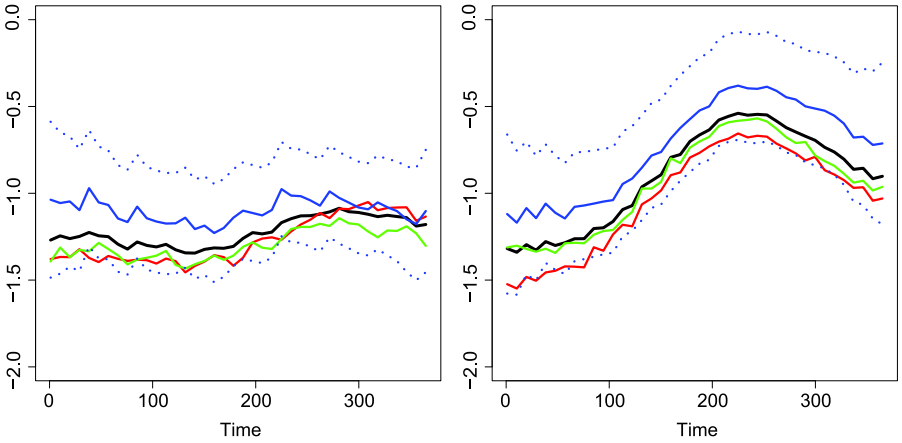

Fig. 10 0.01 expectile curves for Berlin (*left*) and Taipei (*right*) daily temperature residuals from 1948–2007 with the 95% uniform confidence bands for the latest 20 years expectile

is much bigger. However, all the curves both for Berlin and Taipei locate in their corresponding confidence bands.

As depicted in the figures, the performance of the residuals are quite different from Berlin and Taipei, especially for high expectiles. The variation of the temperature in Taipei is more volatile. One interpretation is that in the last 60 years, Taiwan has been experiencing a fast developing period. Industrial expansion, burning of fossil fuel and deforestation and other sectors, could be an important factor for the bigger volatility in the temperature of Taipei. However, Germany is well-developed in this period, especially in Berlin, where there are no intensive industries. Therefore, one may say the residuals reveals the influence of the human activities, which induce the different performance of the residuals of temperature.

Acknowledgements The financial support from the Deutsche Forschungsgemeinschaft via SFB 649 "Ökonomisches Risiko", Humboldt-Universität zu Berlin is gratefully acknowledged. China Scholarship Council (CSC) is gratefully acknowledged.

Appendix

Proof of Theorem 1 By the definition of $v_n(x)$ as a zero of (5), we have, for $\varepsilon > 0$,

$$\text{if } v_n(x) > v(x) + \varepsilon, \quad \text{and} \quad \text{then } H_n\big\{v(x) + \varepsilon, x\big\} > 0 \tag{29}$$

Now

$$H_n\big\{v(x) + \varepsilon, x\big\} \le H\big\{v(x) + \varepsilon, x\big\} + \sup_{\theta \in I} \big|H_n(\theta, x) - H(\theta, x)\big| \tag{30}$$

Also, by the identity $H\{v(x), x\} = 0$, the function $H\{v(x) + \varepsilon, x\}$ is not positive and has a magnitude $\ge m_1 \tilde{q} \varepsilon$ by assumption (A6) and (12), for $0 < \varepsilon < \delta_1$. That is, for

$0 < \varepsilon < \delta_1$,

$$H\{v(x) + \varepsilon, x\} \leq -m_1 \tilde{q} \varepsilon \tag{31}$$

Combining (29), (30) and (31), we have, for $0 < \varepsilon < \delta_1$:

$$\text{if } v_n(x) > v(x) + \varepsilon, \quad \text{and} \quad \text{then } \sup_{\theta \in I} \sup_{x \in J} \left| H_n(\theta, x) - H(\theta, x) \right| > m_1 \tilde{q} \varepsilon$$

With a similar inequality proved for the case $v_n(x) < v(x) + \varepsilon$, we obtain, for $0 < \varepsilon < \delta_1$:

$$\text{if } \sup_{x \in J} \left| v_n(x) - v(x) \right| > \varepsilon, \quad \text{and} \quad \text{then } \sup_{\theta \in I} \sup_{x \in J} \left| H_n(\theta, x) - H(\theta, x) \right| > m_1 \tilde{q} \varepsilon \tag{32}$$

It readily follows that (32) and (11) imply (13). □

Below we first show that $\|R_n\|_\infty = \sup_{x \in J} |R_n(x)|$ vanishes asymptotically faster than the rate $(nh \log n)^{-1/2}$; for simplicity we will just use $\|\cdot\|$ to indicate the sup-norm.

Lemma 4 *For the remainder term $R_n(t)$ defined in (16) we have*

$$\|R_n\| = \mathcal{O}_p\left\{(nh \log n)^{-1/2}\right\} \tag{33}$$

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

$$\|R_n\| \leq \left[\inf_{0 \leq x \leq 1} \left\{ \left| D_n(x) \right| \cdot p(x) \right\} \right]^{-1} \left\{ \|H_n\| \cdot \|p - D_n\| + \|D_n\| \cdot \|\mathsf{E} H_n\| \right\}$$
$$+ C_1 \cdot \|v_n - l\|^2 \cdot \left\{ \inf_{0 \leq t \leq 1} \left| D_n(x) \right| \right\}^{-1} \cdot \|f_n\|$$

where $f_n(x) = (nh)^{-1} \sum_{i=1}^n K\{(x - X_i)/h\}$.

The desired result (4) will then follow if we prove

$$\|H_n\| = \mathcal{O}_p\left\{(nh)^{-1/2} (\log n)^{1/2}\right\} \tag{34}$$

$$\|p - D_n\| = \mathcal{O}_p\left\{(nh)^{-1/4} (\log n)^{-1/2}\right\} \tag{35}$$

$$\|\mathsf{E} H_n\| = \mathcal{O}\left(h^2\right) \tag{36}$$

$$\|v_n - v\|^2 = \mathcal{O}_p\left\{(nh)^{-1/2} (\log n)^{-1/2}\right\} \tag{37}$$

Since (36) follows from the well-known bias calculation

$$\mathsf{E} H_n(x) = h^{-1} \int K\left\{(x - u)/h\right\} \mathsf{E}\left[\psi\left\{y - v(x)\right\} | X = u\right] f_X(u)\, du = \mathcal{O}\left(h^2\right)$$

where $\mathcal{O}(h^2)$ is independent of x in Parzen (1962), we have from assumption (A2) that $\|\mathsf{E} H_n\| = \mathcal{O}_p\{(nh)^{-1/2} (\log n)^{-1/2}\}$.

According to Lemma A.3 in Franke and Mwita (2003),

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

and the following inequality:

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

Statement (34) thus is obtained.

Statement (35) follows in the same way as (34) using assumption (A2) and the Lipschitz continuity properties of K, ψ', l.

According to the uniform consistency of $v_n(x)-v(x)$ shown before, we have

$$\|v_n-v\|=\mathcal{O}_p\left\{(nh)^{-1/2}(\log n)^{1/2}\right\}$$

which implies (37).

Now the assertion of the lemma follows, since by tightness of $D_n(x)$, $\inf_{0\le t\le 1}|D_n(x)|\ge q_0$ a.s. and thus

$$\|R_n\|=\mathcal{O}_p\left\{(nh\log n)^{-1/2}\right\}\left(1+\|f_n\|\right)$$

Finally, by Theorem 3.1 of Bickel and Rosenblatt (1973), $\|f_n\|=\mathcal{O}_p(1)$; thus the desired result $\|R_n\|=\mathcal{O}_p\{(nh\log n)^{-1/2}\}$ follows. □

We now begin with the subsequent approximations of the processes $Y_{0,n}$ to $Y_{5,n}$.

Lemma 5

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

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

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

$$
\begin{aligned}
&+ K(A)\left\{\int_{-a_n}^{a_n} Z_n(x - h\cdot A, y)\, dL(y)\right.\\
&\left.+ L(a_n)(a_n)Z_{n_a}(x - h\cdot A, a_n) - L(-a_n)(-a_n)Z_n(x - h\cdot A, -a_n)\right\}\\
&- K(-A)\left\{\int_{-a_n}^{a_n} Z_n(x + h\cdot A, y)\, dL(y) + L(a_n)(a_n)Z_n(x + h\cdot A, a_n)\right.\\
&\left.- L(-a_n)(-a_n)Z_n(x + h\cdot A, -a_n)\right\}
\end{aligned}
$$

If we apply the same operation to $Y_{1,n}$ with $B_n\{T(x, y)\}$ instead of $Z_n(x, y)$ and use Lemma 2, we finally obtain

$$\sup_{0\le x\le 1} h^{1/2} g(x)^{1/2} \left|Y_{0,n}(x) - Y_{1,n}(x)\right| = \mathcal{O}\left\{n^{-1/2}(\log n)^2\right\} \quad \text{a.s.}$$

□

Lemma 6 $\|Y_{1,n} - Y_{2,n}\| = \mathcal{O}_p(h^{1/2})$.

Proof Note that the Jacobian of $T(x, y)$ is $f(x, y)$. Hence

$$
\begin{aligned}
&Y_{1,n}(x) - Y_{2,n}(x)\\
&\quad = \left|\left\{g(x)h\right\}^{-1/2} \iint_{\Gamma_n} \psi\left\{y - v(x)\right\} K\left\{(x - t)/h\right\} f(t, y)\, dt\, dy\right| \cdot \left|W_n(1, 1)\right|
\end{aligned}
$$

It follows that

$$
\begin{aligned}
h^{-1/2}\|Y_{1,n} - Y_{2,n}\| \le{}& \left|W_n(1, 1)\right| \cdot \left\|g^{-1/2}\right\|\\
&\cdot \sup_{0\le t\le 1} h^{-1} \iint_{\Gamma_n} \left|\psi\left\{y - v(x)\right\} K\left\{(x - t)/h\right\}\right| f(t, y)\, dt\, dy
\end{aligned}
$$

Since $\|g^{-1/2}\|$ is bounded by assumption, we have

$$h^{-1/2}\|Y_{1,n} - Y_{2,n}\| \le \left|W_n(1, 1)\right| \cdot C_4 \cdot h^{-1} \int K\left\{(x - t)/h\right\} dx = \mathcal{O}_p(1)$$

□

Lemma 7 $\|Y_{2,n} - Y_{3,n}\| = \mathcal{O}_p(h^{1/2})$.

Proof The difference $|Y_{2,n}(x) - Y_{3,n}(x)|$ may be written as

$$\left|\left\{g(x)h\right\}^{-1/2} \iint_{\Gamma_n} \left[\psi\left\{y - v(x)\right\} - \psi\left\{y - v(t)\right\}\right] K\left\{(x - t)/h\right\} dW_n\left\{T(t, y)\right\}\right|$$

If we use the fact that l is uniformly continuous, this is smaller than

$$h^{-1/2}\left|g(x)\right|^{-1/2} \cdot \mathcal{O}_p(h)$$

and the lemma thus follows. □

Lemma 8 $\|Y_{4,n} - Y_{5,n}\| = \mathcal{O}_p(h^{1/2})$.

Proof

$$\begin{aligned}
\big|Y_{4,n}(x) - Y_{5,n}(x)\big| &= h^{-1/2}\left|\int\left[\left\{\frac{g(t)}{g(x)}\right\}^{1/2} - 1\right]K\big\{(x-t)/h\big\}dW(x)\right| \\
&\le h^{-1/2}\left|\int_{-A}^{A} W(x-hu)\frac{\partial}{\partial u}\left[\left\{\frac{g(x-hu)}{g(x)}\right\}^{1/2} - 1\right]K(u)\,du\right| \\
&\quad + h^{-1/2}\left|K(A)W(t-hA)\left[\left\{\frac{g(x-Ah)}{g(x)}\right\}^{1/2} - 1\right]\right| \\
&\quad + h^{-1/2}\left|K(-A)W(x+hA)\left[\left\{\frac{g(x+Ah)}{g(x)}\right\}^{1/2} - 1\right]\right| \\
&\quad S_{1,n}(x) + S_{2,n}(x) + S_{3,n}(x), \quad \text{say}
\end{aligned}$$

The second term can be estimated by

$$h^{-1/2}\|S_{2,n}\| \le K(A)\cdot \sup_{0\le x\le 1}\big|W(x-Ah)\big|\cdot \sup_{0\le x\le 1} h^{-1}\left|\left[\left\{\frac{g(x-Ah)}{g(x)}\right\}^{1/2} - 1\right]\right|$$

by the mean value theorem it follows that

$$h^{-1/2}\|S_{2,n}\| = \mathcal{O}_p(1)$$

The first term $S_{1,n}$ is estimated as

$$\begin{aligned}
h^{-1/2}S_{1,n}(x) &= \left|h^{-1}\int_{-A}^{A} W(x-uh)K'(u)\left[\left\{\frac{g(x-uh)}{g(x)}\right\}^{1/2} - 1\right]du\right. \\
&\quad \left.\times\frac{1}{2}\int_{-A}^{A} W(x-uh)K(u)\left\{\frac{g(x-uh)}{g(x)}\right\}^{1/2}\left\{\frac{g'(x-uh)}{g(x)}\right\}du\right| \\
&= \big|T_{1,n}(x) - T_{2,n}(x)\big|, \quad \text{say}
\end{aligned}$$

$\|T_{2,n}\| \le C_5 \cdot \int_{-A}^{A}|W(t-hu)|\,du = \mathcal{O}_p(1)$ by assumption on $g(x) = \sigma^2(x)\cdot f_X(x)$. To estimate $T_{1,n}$ we again use the mean value theorem to conclude that

$$\sup_{0\le x\le 1} h^{-1}\left|\left\{\frac{g(x-uh)}{g(x)}\right\}^{1/2} - 1\right| < C_6\cdot|u|$$

hence

$$\|T_{1,n}\| \le C_6\cdot \sup_{0\le x\le 1}\int_{-A}^{A}\big|W(x-hu)\big|K'(u)u/\,du = \mathcal{O}_p(1)$$

Since $S_{3,n}(x)$ is estimated as $S_{2,n}(x)$, we finally obtain the desired result. □

The next lemma shows that the truncation introduced through $\{a_n\}$ does not affect the limiting distribution.

Lemma 9 $\|Y_n - Y_{0,n}\| = \mathcal{O}_p\{(\log n)^{-1/2}\}$.

Proof We shall only show that $g'(x)^{-1/2}h^{-1/2}\iint_{\mathbb{R}-\Gamma_n}\psi\{y - v(x)\} \times K\{(x-t)/h\}\,dZ_n(t,y)$ fulfills the lemma. The replacement of $g'(x)$ by $g(x)$ may be proved as in Lemma A.4 of Johnston (1982). The quantity above is less than $h^{-1/2}\|g^{-1/2}\| \cdot \|\iint_{\{|y|>a_n\}}\psi\{y - v(x)\}K\{(x-t)/h\}\,dZ(t,y)\|$. It remains to be shown that the last factor tends to zero at a rate $\mathcal{O}_p\{(\log n)^{-1/2}\}$. We show first that

$$V_n(x) = (\log n)^{1/2}h^{-1/2}\iint_{\{|y|>a_n\}}\psi\big\{y - v(x)\big\}K\big\{(x-t)/h\big\}\,dZ_n(t,y)$$

$$\xrightarrow{p} 0 \quad \text{for all } x$$

and then we show tightness of $V_n(x)$, the result then follows:

$$V_n(x) = (\log n)^{1/2}(nh)^{-1/2}\sum_{i=1}^{n}\big[\psi\big\{Y_i - v(x)\big\}\mathbf{I}\big(|Y_i| > a_n\big)K\big\{(x-X_i)/h\big\}$$

$$- \mathsf{E}\,\psi\big\{Y_i - v(x)\big\}\mathbf{I}\big(|Y_i| > a_n\big)K\big\{(x-X_i)/h\big\}\big]$$

$$= \sum_{i=1}^{n} X_{n,x}(x)$$

where $\{X_{n,x}(x)\}_{i=1}^{n}$ are i.i.d. for each n with $\mathsf{E}\,X_{n,x}(x) = 0$ for all $x \in [0,1]$. We then have

$$\mathsf{E}\,X_{n,x}^2(x) \le (\log n)(nh)^{-1}\,\mathsf{E}\,\psi^2\big\{Y_i - v(x)\big\}\mathbf{I}\big(|Y_i| > a_n\big)K^2\big\{(x-X_i)/h\big\}$$

$$\le \sup_{-A\le u\le A} K^2(u)\cdot(\log n)(nh)^{-1}\,\mathsf{E}\,\psi^2\big\{Y_i - v(x)\big\}\mathbf{I}\big(|Y_i| > a_n\big)$$

hence

$$\mathsf{Var}\big\{V_n(x)\big\} = \mathsf{E}\left\{\sum_{i=1}^{n} X_{n,x}(x)\right\}^2 = n\cdot\mathsf{E}\,X_{n,x}^2(x)$$

$$\le \sup_{-A\le u\le A} K^2(u)h^{-1}(\log n)\int_{\{|y|>a_n\}} f_y(y)\,dy\cdot M_\psi$$

where M_ψ denotes an upper bound for ψ^2. This term tends to zero by assumption (A3). Thus by Markov's inequality we conclude that

$$V_n(x) \xrightarrow{p} 0 \quad \text{for all } x \in [0,1]$$

To prove tightness of $\{V_n(x)\}$ we refer again to the following moment condition as stated in Lemma 4:

$$\mathsf{E}\big\{\big|V_n(x) - V_n(x_1)\big| \cdot \big|V_n(x_2) - V_n(x)\big|\big\} \le C' \cdot (x_2 - x_1)^2$$

C' denoting a constant, $x \in [x_1, x_2]$

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

$$\mathsf{E}\big\{V_n(x) - V_n(x_1)\big\}^2 = (\log n)(nh)^{-1}\,\mathsf{E}\Bigg[\sum_{i=1}^{n} \Psi_n(x, x_1, X_i, Y_i) \cdot \mathbf{I}\big(|Y_i| > a_n\big) - \mathsf{E}\big\{\Psi_n(x, x_1, X_i, Y_i) \cdot \mathbf{I}\big(|Y_i| > a_n\big)\big\}\Bigg]^2$$

where $\Psi_n(x, x_1, X_i, Y_i) = \psi\{Y_i - v(x)\}K\{(x - X_i)/h\} - \psi\{Y_i - v(x_1)\}K\{(x_1 - X_1)/h\}$. Since ψ, K are Lipschitz continuous except at one point and the expectation is taken afterwards, it follows that

$$\begin{aligned}&\big[\mathsf{E}\big\{V_n(x) - V_n(x_1)\big\}^2\big]^{1/2}\\ &\quad\le C_7 \cdot (\log n)^{1/2} h^{-3/2} |x - x_1| \cdot \left\{\int_{\{|y| > a_n\}} f_y(y)\,dy\right\}^{1/2}\end{aligned}$$

If we apply the same estimation to $V_n(x_2) - V_n(x_1)$ we finally have

$$\begin{aligned}&\mathsf{E}\big\{\big|V_n(x) - V_n(x_1)\big| \cdot \big|V_n(x_2) - V_n(x)\big|\big\}\\ &\quad\le C_7^2 (\log n) h^{-3} |x - x_1||x_2 - x| \times \int_{\{|y| > a_n\}} f_y(y)\,dy\\ &\quad\le C' \cdot |x_2 - x_1|^2 \quad \text{since } x \in [x_1, x_2] \text{ by (A3)}\end{aligned}$$

□

Lemma 10 *Let $\lambda(K) = \int K^2(u)\,du$ and let $\{d_n\}$ be as in the theorem. Then*

$$(2\delta \log n)^{1/2}\big[\|Y_{3,n}\|/\big\{\lambda(K)\big\}^{1/2} - d_n\big]$$

has the same asymptotic distribution as

$$(2\delta \log n)^{1/2}\big[\|Y_{4,n}\|/\big\{\lambda(K)\big\}^{1/2} - d_n\big]$$

Proof $Y_{3,n}(x)$ is a Gaussian process with

$$\mathsf{E}\, Y_{3,n}(x) = 0$$

and covariance function

$$
\begin{aligned}
r_3(x_1, x_2) &= \mathsf{E}\, Y_{3,n}(x_1) Y_{3,n}(x_2) \\
&= \big\{g(x_1)g(x_2)\big\}^{-1/2} h^{-1} \iint_{\Gamma_n} \psi^2\big\{y - v(x)\big\} K\big\{(x_1 - x)/h\big\} \\
&\quad \times K\big\{(x_2 - x)/h\big\} f(t, y)\, dt\, dy \\
&= \big\{g(x_1)g(x_2)\big\}^{-1/2} h^{-1} \iint_{\Gamma_n} \psi^2\big\{y - v(x)\big\} f(y|x) dy K\big\{(x_1 - x)/h\big\} \\
&\quad \times K\big\{(x_2 - x)/h\big\} f_X(x)\, dx \\
&= \big\{g(x_1)g(x_2)\big\}^{-1/2} h^{-1} \int g(x) K\big\{(x_1 - x)/h\big\} K\big\{(x_2 - x)/h\big\}\, dx \\
&= r_4(x_1, x_2)
\end{aligned}
$$

where $r_4(x_1, x_2)$ is the covariance function of the Gaussian process $Y_{4,n}(x)$, which proves the lemma. □

References

Benth, F., Benth, J., Koekebakker, S.: Putting a price on temperature. Scand. J. Stat. **34**(4), 746–767 (2007)

Bickel, P., Rosenblatt, M.: On some global measures of the deviation of density function estimates. Ann. Stat. **1**, 1071–1095 (1973)

Breckling, J., Chambers, R.: M-quantiles. Biometrika **74**(4), 761–772 (1988)

Campbell, S., Diebold, F.: Weather forecasting for weather derivatives. J. Am. Stat. Assoc. **100**, 6–16 (2005)

Claeskens, G., Keilegom, I.V.: Bootstrap confidence bands for regression curves and their derivatives. Ann. Stat. **31**(6), 1852–1884 (2003)

Csörgö, S., Hall, P.: Upper and lower classes for triangular arrays. Z. Wahrscheinlichkeitstheor. Verw. Geb. **61**, 207–222 (1982)

Diebold, F., Inoue, A.: Long memory and regime switching. J. Econom. **105**, 131–159 (2001)

Efron, B.: Regression percentiles using asymmetric squared loss. Stat. Sin. **1**, 93–125 (1991)

Franke, J., Mwita, P.: Nonparametric estimates for conditional quantiles of time series. Report in Wirtschaftsmathematik, 87, University of Kaiserslautern (2003)

Härdle, W.: Asymptotic maximal deviation of M-smoothers. J. Multivar. Anal. **29**, 163–179 (1989)

Härdle, W., Luckhaus, S.: Uniform consistency of a class of regression function estimators. Ann. Stat. **12**, 612–623 (1984)

Härdle, W., Song, S.: Confidence bands in quantile regression. Econom. Theory **3**, 1–21 (2009)

Härdle, W., Janssen, P., Serfling, R.: Strong uniform consistency rates for estimators of conditional functionals. Ann. Stat. **16**, 1428–1429 (1988)

Härdle, W., Ritov, Y., Song, S.: Partial linear regression and bootstrap confidence bands. SFB 649 Discussion Paper 2010-002. J. Multivar. Anal. (2010, submitted)

Huber, P.: Robust Statistics. Wiley, New York (1981)

Johnston, G.: Probabilities of maximal deviations of nonparametric regression function estimates. J. Multivar. Anal. **12**, 402–414 (1982)

Jones, M.: Expectiles and M-quantiles are quantiles. Stat. Probab. Lett. **20**, 149–153 (1994)

Koenker, R.: Quantile Regression. Cambridge University Press, Cambridge (2005)

Kuan, C.M., Yeh, Y.H., Hsu, Y.C.: Assessing value at risk with care, the conditional autoregressive expectile models. J. Econom. **150**, 261–270 (2009)

Parzen, M.: On estimation of a probability density function and mode. Ann. Math. Stat. **32**, 1065–1076 (1962)

Rosenblatt, M.: Remarks on a multivariate transformation. Ann. Math. Stat. **23**, 470–472 (1952)

Schnabel, S., Eilers, P.: An analysis of life expectancy and economic production using expectile frontier zones. Demogr. Res. **21**, 109–134 (2009a)

Schnabel, S., Eilers, P.: Optimal expectile smoothing. Comput. Stat. Data Anal. **53**, 4168–4177 (2009b)
Taylor, J.: Estimating value at risk and expected shortfall using expectiles. J. Financ. Econom. **6**, 231–252 (2008)
Tusnady, G.: A remark on the approximation of the sample distribution function in the multidimensional case. Period. Math. Hung. **8**, 53–55 (1977)
Yao, Q., Tong, H.: Asymmetric least squares regression estimation: a nonparametric approach. J. Nonparametr. Stat. **6**(2–3), 273–292 (1996)
Zhang, B.: Nonparametric regression expectiles. J. Nonparametr. Stat. **3**, 255–275 (1994)

The Annals of Statistics
2011, Vol. 39, No. 3, 1658–1688
DOI: 10.1214/10-AOS871

THE EFM APPROACH FOR SINGLE-INDEX MODELS

BY XIA CUI[1], WOLFGANG KARL HÄRDLE[2] AND LIXING ZHU[3]

Sun Yat-sen University, Humboldt-Universität zu Berlin and National Central University, and Hong Kong Baptist University and Yunnan University of Finance and Economics

Single-index models are natural extensions of linear models and circumvent the so-called curse of dimensionality. They are becoming increasingly popular in many scientific fields including biostatistics, medicine, economics and financial econometrics. Estimating and testing the model index coefficients $\boldsymbol{\beta}$ is one of the most important objectives in the statistical analysis. However, the commonly used assumption on the index coefficients, $\|\boldsymbol{\beta}\| = 1$, represents a nonregular problem: the true index is on the boundary of the unit ball. In this paper we introduce the EFM approach, a method of estimating functions, to study the single-index model. The procedure is to first relax the equality constraint to one with $(d-1)$ components of $\boldsymbol{\beta}$ lying in an open unit ball, and then to construct the associated $(d-1)$ estimating functions by projecting the score function to the linear space spanned by the residuals with the unknown link being estimated by kernel estimating functions. The root-n consistency and asymptotic normality for the estimator obtained from solving the resulting estimating equations are achieved, and a Wilks type theorem for testing the index is demonstrated. A noticeable result we obtain is that our estimator for $\boldsymbol{\beta}$ has smaller or equal limiting variance than the estimator of Carroll et al. [*J. Amer. Statist. Assoc.* **92** (1997) 447–489]. A fixed-point iterative scheme for computing this estimator is proposed. This algorithm only involves one-dimensional nonparametric smoothers, thereby avoiding the data sparsity problem caused by high model dimensionality. Numerical studies based on simulation and on applications suggest that this new estimating system is quite powerful and easy to implement.

1. Introduction. Single-index models combine flexibility of modeling with interpretability of (linear) coefficients. They circumvent the curse of dimensionality and are becoming increasingly popular in many scientific fields. The reduction of dimension is achieved by assuming the link function to be a univariate function applied to the projection of explanatory covariate vector on to some direction.

Received April 2010; revised December 2010.

[1]Supported by NNSF project (11026194) of China, RFDP (20100171120042) of China and "the Fundamental Research Funds for the Central Universities" (11lgpy26) of China.

[2]Supported by Deutsche Forschungsgemeinschaft SFB 649 "Ökonomisches Risiko."

[3]Supported by a Grant (HKBU2030/07P) from Research Grants Council of Hong Kong, Hong Kong, China.

MSC2010 subject classifications. 62G08, 62G08, 62G20.

Key words and phrases. Single-index models, index coefficients, estimating equations, asymptotic properties, iteration.

In this paper we consider an extension of single-index models where, instead of a distributional assumption, assumptions of only the mean function and variance function of the response are made. Let $(Y_i, \mathbf{X}_i)$, $i = 1, \ldots, n$, denote the observed values with Y_i being the response variable and $\mathbf{X}_i$ as the vector of d explanatory variables. The relationship of the mean and variance of Y_i is specified as follows:

$$E(Y_i|\mathbf{X}_i) = \mu\{g(\boldsymbol{\beta}^\top \mathbf{X}_i)\}, \qquad \mathrm{Var}(Y_i|\mathbf{X}_i) = \sigma^2 V\{g(\boldsymbol{\beta}^\top \mathbf{X}_i)\}, \tag{1.1}$$

where μ is a known monotonic function, V is a known covariance function, g is an unknown univariate link function and $\boldsymbol{\beta}$ is an unknown index vector which belongs to the parameter space $\Theta = \{\boldsymbol{\beta} = (\beta_1, \ldots, \beta_d)^\top : \|\boldsymbol{\beta}\| = 1, \beta_1 > 0, \boldsymbol{\beta} \in \mathbb{R}^d\}$. Here we assume the parameter space is Θ rather than the entire $\mathbb{R}^d$ in order to ensure that $\boldsymbol{\beta}$ in the representation (1.1) can be uniquely defined. This is a commonly used assumption on the index parameter [see Carroll et al. (1997), Zhu and Xue (2006), Lin and Kulasekera (2007)]. Another reparameterization is to let $\beta_1 = 1$ for the sign identifiability and to transform $\boldsymbol{\beta}$ to $(1, \beta_2, \ldots, \beta_d)/(1 + \sum_{r=2}^d \beta_r^2)^{1/2}$ for the scale identifiability. Clearly $(1, \beta_2, \ldots, \beta_d)/(1+\sum_{r=2}^d \beta_r^2)^{1/2}$ can also span the parameter space Θ by simply checking that $\|(1, \beta_2, \ldots, \beta_d)/(1 + \sum_{r=2}^d \beta_r^2)^{1/2}\| = 1$ and the first component $1/(1 + \sum_{r=2}^d \beta_r^2)^{1/2} > 0$. However, the fixed-point algorithm recommended in this paper for normalized vectors may not be suitable for such a reparameterization. Model (1.1) is flexible enough to cover a variety of situations. If μ is the identity function and V is equal to constant 1, (1.1) reduces to a single-index model Härdle, Hall and Ichimura (1993). Model (1.1) is an extension of the generalized linear model McCullagh and Nelder (1989) and the single-index model. When the conditional distribution of Y is logistic, then $\mu\{g(\boldsymbol{\beta}^\top \mathbf{X})\} = \exp\{g(\boldsymbol{\beta}^\top \mathbf{X})\}/[1 + \exp\{g(\boldsymbol{\beta}^\top \mathbf{X})\}]$ and $V\{g(\boldsymbol{\beta}^\top \mathbf{X})\} = \exp\{g(\boldsymbol{\beta}^\top \mathbf{X})\}/[1 + \exp\{g(\boldsymbol{\beta}^\top \mathbf{X})\}]^2$.

For single-index models: $\mu\{g(\boldsymbol{\beta}^\top \mathbf{X})\} = g(\boldsymbol{\beta}^\top \mathbf{X})$ and $V\{g(\boldsymbol{\beta}^\top \mathbf{X})\} = 1$, various strategies for estimating $\boldsymbol{\beta}$ have been proposed in the last decades. Two most popular methods are the average derivative method (ADE) introduced in Powell, Stock and Stoker (1989) and Härdle and Stoker (1989), and the simultaneous minimization method of Härdle, Hall and Ichimura (1993). Next we will review these two methods in short. The ADE method is based on that $\partial E(Y|\mathbf{X} = \mathbf{x})/\partial \mathbf{x} = g'(\boldsymbol{\beta}^\top \mathbf{x})\boldsymbol{\beta}$ which implies that the gradient of the regression function is proportional to the index parameter $\boldsymbol{\beta}$. Then a natural estimator for $\boldsymbol{\beta}$ is $\hat{\boldsymbol{\beta}} = n^{-1} \sum_{i=1}^n \widehat{\nabla G}(\mathbf{X}_i)/\|n^{-1} \sum_{i=1}^n \widehat{\nabla G}(\mathbf{X}_i)\|$ with $\nabla G(\mathbf{x})$ denoting $\partial E(Y|\mathbf{X} = \mathbf{x})/\partial \mathbf{x}$ and $\|\cdot\|$ being the Euclidean norm. An advantage of the ADE approach is that it allows estimating $\boldsymbol{\beta}$ directly. However, the high-dimensional kernel smoothing used for computing $\widehat{\nabla G}(\mathbf{x})$ suffers from the "curse of dimensionality" if the model dimension d is large. Hristache, Juditski and Spokoiny (2001) improved the ADE approach by lowering the dimension of the kernel gradually. The method of Härdle, Hall and Ichimura (1993) is carried out by minimizing a least squares criterion based on nonparametric estimation of the link g with respect

to $\boldsymbol{\beta}$ and bandwidth h. However, the minimization is difficult to implement since it depends on an optimization problem in a high-dimensional space. Xia et al. (2002) proposed to minimize average conditional variance (MAVE). Because the kernel used for computing $\boldsymbol{\beta}$ is a function of $\|\mathbf{X}_i - \mathbf{X}_j\|$, MAVE meets the problem of data sparseness. All the above estimators are consistent under some regular conditions. Asymptotic efficiency comparisons of the above methods have been discussed in Xia (2006) resulting in the MAVE estimator of $\boldsymbol{\beta}$ having the same limiting variance as the estimators of Härdle, Hall and Ichimura (1993), and claiming alternative versions of the ADE method having larger variance. In addition, Yu and Ruppert (2002) fitted the partially linear single-index models using a penalized spline method. Huh and Park (2002) used the local polynomial method to fit the unknown function in single-index models. Other dimension reduction methods that were recently developed in the literature are sliced inverse regression, partial least squares and canonical correlation method. These methods handle high-dimensional predictors; see Zhu and Zhu (2009a, 2009b) and Zhou and He (2008).

The main challenges of estimation in the semiparametric model (1.1) are that the support of the infinite-dimensional nuisance parameter $g(\cdot)$ depends on the finite-dimensional parameter $\boldsymbol{\beta}$, and the parameter $\boldsymbol{\beta}$ is on the boundary of a unit ball. For estimating $\boldsymbol{\beta}$ the former challenge forces us to deal with the infinite-dimensional nuisance parameter g. The latter one represents a nonregular problem. The classic assumptions about asymptotic properties of the estimates for $\boldsymbol{\beta}$ are not valid. In addition, as a model proposed for dimension reduction, the dimension d may be very high and one often meets the problem of computation. To attack the above problems, in this paper we will develop an estimating function method (EFM) and then introduce a computational algorithm to solve the equations based on a fixed-point iterative scheme. We first choose an identifiable parameterization which transforms the boundary of a unit ball in $\mathbb{R}^d$ to the interior of a unit ball in $\mathbb{R}^{d-1}$. By eliminating β_1, the parameter space Θ can be rearranged to a form $\{((1-\sum_{r=2}^{d}\beta_r^2)^{1/2}, \beta_2, \ldots, \beta_d)^\top : \sum_{r=2}^{d}\beta_r^2 < 1\}$. Then the derivatives of a function with respect to $(\beta_2, \ldots, \beta_d)^\top$ are readily obtained by the chain rule and the classical assumptions on the asymptotic normality hold after transformation. The estimating functions (equations) for $\boldsymbol{\beta}$ can be constructed by replacing $g(\boldsymbol{\beta}^\top\mathbf{X})$ with $\hat{g}(\boldsymbol{\beta}^\top\mathbf{X})$. The estimate $\hat{g}$ for the nuisance parameter g is obtained using kernel estimating functions and the smoothing parameter h is selected using K-fold cross-validation. For the problem of testing the index, we establish a quasi-likelihood ratio based on the proposed estimating functions and show that the test statistics asymptotically follow a χ^2-distribution whose degree of freedom does not depend on nuisance parameters, under the null hypothesis. Then a Wilks type theorem for testing the index is demonstrated.

The proposed EFM technique is essentially a unified method of handling different types of data situations including categorical response variable and discrete explanatory covariate vector. The main results of this research are as follows:

(a) *Efficiency*. A surprising result we obtain is that our EFM estimator for $\boldsymbol{\beta}$ has smaller or equal limiting variance than the estimator of Carroll et al. (1997).
(b) *Computation*. The estimating function system only involves one-dimensional nonparametric smoothers, thereby avoiding the data sparsity problem caused by high model dimensionality. Unlike the quasi-likelihood inference [Carroll et al. (1997)] where the maximization is difficult to implement when d is large, the reparameterization and the explicit formulation of the estimating functions facilitate an efficient computation algorithm. Here we use a fixed-point iterative scheme to compute the resultant estimator. The simulation results show that the algorithm adapts to higher model dimension and richer data situations than the MAVE method of Xia et al. (2002).

It is noteworthy that the EFM approach proposed in this paper cannot be obtained from the SLS method proposed in Ichimura (1993) and investigated in Härdle, Hall and Ichimura (1993). SLS minimizes the weighted least squares criterion $\sum_{j=1}^{n}[Y_j - \mu\{\hat{g}(\boldsymbol{\beta}^\top \mathbf{X}_j)\}]^2 V^{-1}\{\hat{g}(\boldsymbol{\beta}^\top \mathbf{X}_j)\}$, which leads to a biased estimating equation when we use its derivative if $V(\cdot)$ does not contain the parameter of interest. It will not in general provide a consistent estimator [see Heyde (1997), page 4]. Chang, Xue and Zhu (2010) and Wang et al. (2010) discussed the efficient estimation of single-index model for the case of additive noise. However, their methods are based on the estimating equations induced from the least squares rather than the quasi-likelihood. Thus, their estimation does not have optimal property. Also their comparison is with the one from Härdle, Hall and Ichimura (1993) and its later development. It cannot be applied to the setting under study. In this paper, we investigate the efficiency and computation of the estimates for the single-index models, and systematically develop and prove the asymptotic properties of EFM.

The paper is organized as follows. In Section 2, we state the single-index model, discuss estimation of g using kernel estimating functions and of $\boldsymbol{\beta}$ using profile estimating functions, and investigate the problem of testing the index using quasi-likelihood ratio. In Section 3 we provide a computation algorithm for solving the estimating functions and illustrate the method with simulation and practical studies. The proofs are deferred to the Appendix.

2. Estimating function method (EFM) and its large sample properties. In this section, which is concerned with inference based on the estimating function method, the model of interest is determined through specification of mean and variance functions, up to an unknown vector $\boldsymbol{\beta}$ and an unknown function g. Except for Gaussian data, model (1.1) need not be a full semiparametric likelihood specification. Note that the parameter space $\Theta = \{\boldsymbol{\beta} = (\beta_1, \ldots, \beta_d)^\top : \|\boldsymbol{\beta}\| = 1, \beta_1 > 0, \boldsymbol{\beta} \in \mathbb{R}^d\}$ means that $\boldsymbol{\beta}$ is on the boundary of a unit ball and it represents therefore a nonregular problem. So we first choose an identifiable parameterization which transforms the boundary of a unit ball in $\mathbb{R}^d$ to the interior of a unit ball in $\mathbb{R}^{d-1}$. By eliminating β_1, the parameter space Θ can be rearranged to a form

$\{((1-\sum_{r=2}^{d}\beta_r^2)^{1/2},\beta_2,\ldots,\beta_d)^\top:\sum_{r=2}^{d}\beta_r^2<1\}$. Then the derivatives of a function with respect to $\boldsymbol{\beta}^{(1)}=(\beta_2,\ldots,\beta_d)^\top$ are readily obtained by chain rule and the classic assumptions on the asymptotic normality hold after transformation. This reparameterization is the key to analyzing the asymptotic properties of the estimates for $\boldsymbol{\beta}$ and to facilitating an efficient computation algorithm. We will investigate the estimation for g and $\boldsymbol{\beta}$ and propose a quasi-likelihood method to test the statistical significance of certain variables in the parametric component.

2.1. *The kernel estimating functions for the nonparametric part* g. If $\boldsymbol{\beta}$ is known, then we estimate $g(\cdot)$ and $g'(\cdot)$ using the local linear estimating functions. Let h denote the bandwidth parameter, and let $K(\cdot)$ denote the symmetric kernel density function satisfying $K_h(\cdot)=h^{-1}K(\cdot/h)$. The estimation method involves local linear approximation. Denote by α_0 and α_1 the values of g and g' evaluating at t, respectively. The local linear approximation for $g(\boldsymbol{\beta}^\top\mathbf{x})$ in a neighborhood of t is $\tilde{g}(\boldsymbol{\beta}^\top\mathbf{x})=\alpha_0+\alpha_1(\boldsymbol{\beta}^\top\mathbf{x}-t)$. The estimators $\hat{g}(t)$ and $\hat{g}'(t)$ are obtained by solving the kernel estimating functions with respect to α_0,α_1:

$$
\begin{cases}
\displaystyle\sum_{j=1}^{n} K_h(\boldsymbol{\beta}^\top\mathbf{X}_j-t)\mu'\{\tilde{g}(\boldsymbol{\beta}^\top\mathbf{X}_j)\}V^{-1}\{\tilde{g}(\boldsymbol{\beta}^\top\mathbf{X}_j)\} \\
\qquad\qquad \times[Y_j-\mu\{\tilde{g}(\boldsymbol{\beta}^\top\mathbf{X}_j)\}]=0, \\
\displaystyle\sum_{j=1}^{n}(\boldsymbol{\beta}^\top\mathbf{X}_j-t)K_h(\boldsymbol{\beta}^\top\mathbf{X}_j-t)\mu'\{\tilde{g}(\boldsymbol{\beta}^\top\mathbf{X}_j)\}V^{-1}\{\tilde{g}(\boldsymbol{\beta}^\top\mathbf{X}_j)\} \\
\qquad\qquad \times[Y_j-\mu\{\tilde{g}(\boldsymbol{\beta}^\top\mathbf{X}_j)\}]=0.
\end{cases}
\tag{2.1}
$$

Having estimated α_0,α_1 at t as $\hat{\alpha}_0,\hat{\alpha}_1$, the local linear estimators of $g(t)$ and $g'(t)$ are $\hat{g}(t)=\hat{\alpha}_0$ and $\hat{g}'(t)=\hat{\alpha}_1$, respectively.

The key to obtain the asymptotic normality of the estimates for $\boldsymbol{\beta}$ lies in the asymptotic properties of the estimated nonparametric part. The following theorem will provide some useful results. The following notation will be used. Let $\mathcal{X}=\{\mathbf{X}_1,\ldots,\mathbf{X}_n\}$, $\rho_l(z)=\{\mu'(z)\}^lV^{-1}(z)$ and $\mathbf{J}=\frac{\partial\boldsymbol{\beta}}{\partial\boldsymbol{\beta}^{(1)}}$ the Jacobian matrix of size $d\times(d-1)$ with

$$
\mathbf{J}=\begin{pmatrix}-\boldsymbol{\beta}^{(1)\top}/\sqrt{1-\|\boldsymbol{\beta}^{(1)}\|^2}\\ \mathbf{I}_{d-1}\end{pmatrix},\qquad \boldsymbol{\beta}^{(1)}=(\beta_2,\ldots,\beta_d)^\top.
$$

The moments of K and K^2 are denoted, respectively, by, $j=0,1,\ldots,$

$$
\gamma_j=\int t^jK(t)\,dt\quad\text{and}\quad \nu_j=\int t^jK^2(t)\,dt.
$$

PROPOSITION 1. *Under regularity conditions* (a), (b), (d) *and* (e) *given in the Appendix*, *we have*:

(i) *With $h \to 0$, $n \to \infty$ such that $h \to 0$ and $nh \to \infty$, $\forall \boldsymbol{\beta} \in \Theta$, the asymptotic conditional bias and variance of $\hat{g}$ are given by*

$$
\begin{aligned}
&E\{\{\hat{g}(\boldsymbol{\beta}^\top \mathbf{x}) - g(\boldsymbol{\beta}^\top \mathbf{x})\}^2 | \mathcal{X}\} \\
&\quad = \{\tfrac{1}{2}\gamma_2 h^2 g''(\boldsymbol{\beta}^\top \mathbf{x})\}^2 \\
&\qquad + \nu_0 \sigma^2 / [nh f_{\boldsymbol{\beta}^\top \mathbf{x}}(\boldsymbol{\beta}^\top \mathbf{x}) \rho_2\{g(\boldsymbol{\beta}^\top \mathbf{x})\}] \\
&\qquad + \mathcal{O}_P(h^4 + n^{-1}h^{-1}).
\end{aligned} \tag{2.2}
$$

(ii) *With $h \to 0$, $n \to \infty$ such that $h \to 0$ and $nh^3 \to \infty$, for the estimates of the derivative g', it holds that*

$$
\begin{aligned}
&E\{\{\hat{g}'(\boldsymbol{\beta}^\top \mathbf{x}) - g'(\boldsymbol{\beta}^\top \mathbf{x})\}^2 | \mathcal{X}\} \\
&\quad = \{\tfrac{1}{6}\gamma_4 \gamma_2^{-1} h^2 g'''(\boldsymbol{\beta}^\top \mathbf{x}) \\
&\qquad + \tfrac{1}{2}(\gamma_4 \gamma_2^{-1} - \gamma_2) h^2 g''(\boldsymbol{\beta}^\top \mathbf{x}) \\
&\qquad\quad \times [\rho_2'\{g(\boldsymbol{\beta}^\top \mathbf{x})\}/\rho_2\{g(\boldsymbol{\beta}^\top \mathbf{x})\} + f'_{\boldsymbol{\beta}^\top \mathbf{x}}(\boldsymbol{\beta}^\top \mathbf{x})/f_{\boldsymbol{\beta}^\top \mathbf{x}}(\boldsymbol{\beta}^\top \mathbf{x})]\}^2 \\
&\qquad + \nu_2 \gamma_2^{-2} \sigma^2 / [nh^3 f_{\boldsymbol{\beta}^\top \mathbf{x}}(\boldsymbol{\beta}^\top \mathbf{x}) \rho_2\{g(\boldsymbol{\beta}^\top \mathbf{x})\}] \\
&\qquad + \mathcal{O}_P(h^4 + n^{-1}h^{-3}).
\end{aligned} \tag{2.3}
$$

(iii) *With $h \to 0$, $n \to \infty$ such that $h \to 0$ and $nh^3 \to \infty$, we have that*

$$
E\left\{\left\| \frac{\partial \hat{g}(\boldsymbol{\beta}^\top \mathbf{x})}{\partial \boldsymbol{\beta}^{(1)}} - g'(\boldsymbol{\beta}^\top \mathbf{x}) \mathbf{J}^\top \{\mathbf{x} - E(\mathbf{x}|\boldsymbol{\beta}^\top \mathbf{x})\} \right\|^2 \Big| \mathcal{X} \right\} = \mathcal{O}_P(h^4 + n^{-1}h^{-3}). \tag{2.4}
$$

The proof of this proposition appears in the Appendix. Results (i) and (ii) in Proposition 1 are routine and similar to Carroll, Ruppert and Welsh (1998). In the situation where $\sigma^2 V = \sigma^2$ and the function μ is identity, results (i) and (ii) coincide with those given by Fan and Gijbels (1996). From result (iii), it is seen that $\partial \hat{g}(\boldsymbol{\beta}^\top \mathbf{x})/\partial \boldsymbol{\beta}^{(1)}$ converges in probability to $g'(\boldsymbol{\beta}^\top \mathbf{x}) \mathbf{J}^\top \{\mathbf{x} - E(\mathbf{x}|\boldsymbol{\beta}^\top \mathbf{x})\}$, rather than $g'(\boldsymbol{\beta}^\top \mathbf{x}) \mathbf{J}^\top \mathbf{x}$ as if g were known. That is, $\lim_{n\to\infty}\{\partial \hat{g}(\boldsymbol{\beta}^\top \mathbf{x})/\partial \boldsymbol{\beta}^{(1)}\} \neq \partial\{\lim_{n\to\infty} \hat{g}(\boldsymbol{\beta}^\top \mathbf{x})\}/\partial \boldsymbol{\beta}^{(1)}$, which means that the convergence in probability and the derivation of the sequence $\hat{g}_n(\boldsymbol{\beta}^\top \mathbf{x})$ (as a function of n) cannot commute. This is primarily caused by the fact that the support of the infinite-dimensional nuisance parameter $g(\cdot)$ depends on the finite-dimensional projection parameter $\boldsymbol{\beta}$. In contrast, a semiparametric model where the support of the nuisance parameter is independent of the finite-dimensional parameter is a partially linear regression model having form $Y = \mathbf{X}^\top \boldsymbol{\theta} + \eta(T) + \varepsilon$. It is easy to check that the limit of $\partial \hat{\eta}(T)/\partial \boldsymbol{\theta}$ is equal to $E(\mathbf{X}|T)$, which is the derivative of $\lim_{n\to\infty} \hat{\eta}(T) = E(Y|T) - E(\mathbf{X}^\top|T)\boldsymbol{\theta}$ with respect to $\boldsymbol{\theta}$. Result (iii) ensures that the proposed estimator does not require undersmoothing of $g(\cdot)$ to obtain a root-n consistent estimator for $\boldsymbol{\beta}$ and it is also of its own interest in inference theory for semiparametric models.

2.2. *The asymptotic distribution for the estimates of the parametric part* $\boldsymbol{\beta}$. We will now proceed to the estimation of $\boldsymbol{\beta} \in \Theta$. We need to estimate the $(d-1)$-dimensional vector $\boldsymbol{\beta}^{(1)}$, the estimator of which will be defined via

$$\sum_{i=1}^{n} [\partial \mu\{\hat{g}(\boldsymbol{\beta}^\top \mathbf{X}_i)\}/\partial \boldsymbol{\beta}^{(1)}] V^{-1}\{\hat{g}(\boldsymbol{\beta}^\top \mathbf{X}_i)\}[Y_i - \mu\{\hat{g}(\boldsymbol{\beta}^\top \mathbf{X}_i)\}] = 0. \tag{2.5}$$

This is the direct analogue of the "ideal" estimating equation for known g, in that it is calculated by replacing $g(t)$ with $\hat{g}(t)$. An asymptotically equivalent and easily computed version of this equation is

$$\begin{aligned} \hat{\mathbf{G}}(\boldsymbol{\beta}) &\stackrel{\text{def}}{=} \sum_{i=1}^{n} \mathbf{J}^\top \hat{g}'(\boldsymbol{\beta}^\top \mathbf{X}_i)\{\mathbf{X}_i - \hat{\mathbf{h}}(\boldsymbol{\beta}^\top \mathbf{X}_i)\}\rho_1\{\hat{g}(\boldsymbol{\beta}^\top \mathbf{X}_i)\}[Y_i - \mu\{\hat{g}(\boldsymbol{\beta}^\top \mathbf{X}_i)\}] \\ &= 0 \end{aligned} \tag{2.6}$$

with $\mathbf{J} = \frac{\partial \boldsymbol{\beta}}{\partial \boldsymbol{\beta}^{(1)}}$ the Jacobian mentioned above, $\hat{g}$ and $\hat{g}'$ are defined by (2.1), and $\hat{\mathbf{h}}(t)$ the local linear estimate for $\mathbf{h}(t) = E(\mathbf{X}|\boldsymbol{\beta}^\top \mathbf{X} = t) = (h_1(t), \ldots, h_d(t))^\top$,

$$\hat{\mathbf{h}}(t) = \sum_{i=1}^{n} b_i(t)\mathbf{X}_i \Big/ \sum_{i=1}^{n} b_i(t),$$

where $b_i(t) = K_h(\boldsymbol{\beta}^\top \mathbf{X}_i - t)\{S_{n,2}(t) - (\boldsymbol{\beta}^\top \mathbf{X}_i - t)S_{n,1}(t)\}$, $S_{n,k} = \sum_{i=1}^{n} K_h(\boldsymbol{\beta}^\top \times \mathbf{X}_i - t)(\boldsymbol{\beta}^\top \mathbf{X}_i - t)^k$, $k = 1, 2$. We use (2.6) to estimate $\boldsymbol{\beta}^{(1)}$ in the single-index model, and then use the fact that $\beta_1 = \sqrt{1 - \|\boldsymbol{\beta}^{(1)}\|^2}$ to obtain $\hat{\beta}_1$. The use of (2.6) constitutes in our view a new approach to estimating single-index models; since (2.6) involves smooth pilot estimation of g, g' and $\mathbf{h}$ we call it the Estimation Function Method (EFM) for $\boldsymbol{\beta}$.

REMARK 1. The estimating equations $\hat{\mathbf{G}}(\boldsymbol{\beta})$ can be represented as the gradient vector of the following objective function:

$$\hat{Q}(\boldsymbol{\beta}) = \sum_{i=1}^{n} Q[\mu\{\hat{g}(\boldsymbol{\beta}^\top \mathbf{X}_i)\}, Y_i]$$

with $Q[\mu, y] = \int_{\mu}^{y} \frac{s-y}{V\{\mu^{-1}(s)\}}\, ds$ and $\mu^{-1}(\cdot)$ the inverse function of $\mu(\cdot)$. The existence of such a potential function makes $\hat{\mathbf{G}}(\boldsymbol{\beta})$ to inherit properties of the ideal likelihood score function. Note that $\{\|\boldsymbol{\beta}^{(1)}\| < 1\}$ is an open, connected subset of $\mathbb{R}^{d-1}$. By the regularity conditions assumed on $\mu(\cdot), g(\cdot), V(\cdot)$ (for details see the Appendix), we know that the quasi-likelihood function $\hat{Q}(\boldsymbol{\beta})$ is twice continuously differentiable on $\{\|\boldsymbol{\beta}^{(1)}\| < 1\}$ such that the global maximum of $\hat{Q}(\boldsymbol{\beta})$ can be achieved at some point. One may ask whether the so-

lution is unique and also consistent. Some elementary calculations lead to the Hessian matrix $\partial^2\hat{Q}(\boldsymbol{\beta})/\partial\boldsymbol{\beta}^{(1)}\partial\boldsymbol{\beta}^{(1)\top}$, because the partial derivative $\frac{\partial\mu\{\hat{g}(\boldsymbol{\beta}^\top\mathbf{X}_i)\}}{\partial\boldsymbol{\beta}^{(1)}} = \mu'\{\hat{g}(\boldsymbol{\beta}^\top\mathbf{X}_i)\}\hat{g}'(\boldsymbol{\beta}^\top\mathbf{X}_i)\{\mathbf{X}_i - \hat{\mathbf{h}}(\boldsymbol{\beta}^\top\mathbf{X}_i)\}$, then

$$
\begin{aligned}
&\frac{1}{n}\frac{\partial^2\hat{Q}(\boldsymbol{\beta})}{\partial\boldsymbol{\beta}^{(1)}\partial\boldsymbol{\beta}^{(1)\top}} \\
&\quad= \frac{1}{n}\frac{\partial\hat{\mathbf{G}}(\boldsymbol{\beta})}{\partial\boldsymbol{\beta}^{(1)}} \\
&\quad= \frac{1}{n}\sum_{i=1}^{n}\frac{\partial[\mathbf{J}^\top\hat{g}'(\boldsymbol{\beta}^\top\mathbf{X}_i)\{\mathbf{X}_i - \hat{\mathbf{h}}(\boldsymbol{\beta}^\top\mathbf{X}_i)\}\rho_1\{\hat{g}(\boldsymbol{\beta}^\top\mathbf{X}_i)\}]}{\partial\boldsymbol{\beta}^{(1)}}[Y_i - \mu\{\hat{g}(\boldsymbol{\beta}^\top\mathbf{X}_i)\}] \\
&\qquad- \frac{1}{n}\sum_{i=1}^{n}\mathbf{J}^\top\hat{g}'(\boldsymbol{\beta}^\top\mathbf{X}_i)\{\mathbf{X}_i - \hat{\mathbf{h}}(\boldsymbol{\beta}^\top\mathbf{X}_i)\}\rho_1\{\hat{g}(\boldsymbol{\beta}^\top\mathbf{X}_i)\}\frac{\partial\mu\{\hat{g}(\boldsymbol{\beta}^\top\mathbf{X}_i)\}}{\partial\boldsymbol{\beta}^{(1)}} \\
&\quad= \frac{1}{n}\sum_{i=1}^{n}\Bigg[-\frac{\partial\{\boldsymbol{\beta}^{(1)}/\sqrt{1-\|\boldsymbol{\beta}^{(1)}\|^2}\}}{\partial\boldsymbol{\beta}^{(1)}}\hat{g}'(\boldsymbol{\beta}^\top\mathbf{X}_i)\{\mathbf{X}_{1i} - \hat{h}_1(\boldsymbol{\beta}^\top\mathbf{X}_i)\}\rho_1\{\hat{g}(\boldsymbol{\beta}^\top\mathbf{X}_i)\} \\
&\qquad\qquad+\mathbf{J}^\top\{\mathbf{X}_i - \hat{\mathbf{h}}(\boldsymbol{\beta}^\top\mathbf{X}_i)\}\frac{\partial\hat{g}'(\boldsymbol{\beta}^\top\mathbf{X}_i)}{\partial\boldsymbol{\beta}^{(1)\top}}\rho_1\{\hat{g}(\boldsymbol{\beta}^\top\mathbf{X}_i)\} \\
&\qquad\qquad+\mathbf{J}^\top\hat{g}'(\boldsymbol{\beta}^\top\mathbf{X}_i)\{\mathbf{X}_i - \hat{\mathbf{h}}(\boldsymbol{\beta}^\top\mathbf{X}_i)\}\frac{\partial\rho_1\{\hat{g}(\boldsymbol{\beta}^\top\mathbf{X}_i)\}}{\partial\boldsymbol{\beta}^{(1)\top}} \\
&\qquad\qquad\qquad-\mathbf{J}^\top\hat{g}'(\boldsymbol{\beta}^\top\mathbf{X}_i)\frac{\partial\hat{\mathbf{h}}(\boldsymbol{\beta}^\top\mathbf{X}_i)}{\partial\boldsymbol{\beta}^{(1)}}\rho_1\{\hat{g}(\boldsymbol{\beta}^\top\mathbf{X}_i)\}\Bigg] \\
&\qquad\times[Y_i - \mu\{\hat{g}(\boldsymbol{\beta}^\top\mathbf{X}_i)\}] \\
&\qquad- \frac{1}{n}\sum_{i=1}^{n}\mathbf{J}^\top\hat{g}'^2(\boldsymbol{\beta}^\top\mathbf{X}_i)\{\mathbf{X}_i - \hat{\mathbf{h}}(\boldsymbol{\beta}^\top\mathbf{X}_i)\}\{\mathbf{X}_i - \hat{\mathbf{h}}(\boldsymbol{\beta}^\top\mathbf{X}_i)\}^\top\rho_2\{\hat{g}(\boldsymbol{\beta}^\top\mathbf{X}_i)\}\mathbf{J}.
\end{aligned}
$$

By the regularity conditions in the Appendix, the multipliers of the residuals $[Y_i - \mu\{\hat{g}(\boldsymbol{\beta}^\top\mathbf{X}_i)\}]$ in the first sum of (2.7) are bounded. Mimicking the proof of Proposition 1, the first sum can be shown to converge to 0 in probability as n goes to infinity. The second sum converges to a negative semidefinite matrix. If the Hessian matrix $\frac{1}{n}\frac{\partial^2\hat{Q}(\boldsymbol{\beta})}{\partial\boldsymbol{\beta}^{(1)}\partial\boldsymbol{\beta}^{(1)\top}}$ is negative definite for all values of $\boldsymbol{\beta}^{(1)}$, $\hat{\mathbf{G}}(\boldsymbol{\beta})$ has a unique root. At sample level, however, estimating functions may have more than one root. For the EFM method, the quasi-likelihood $\hat{Q}(\boldsymbol{\beta})$ exists, which can be used to distinguish local maxima from minima. Thus, we suppose (2.6) has a unique solution in the following context.

REMARK 2. It can be seen from the proof in the Appendix that the population version of $\hat{\mathbf{G}}(\boldsymbol{\beta})$ is

$$\mathbf{G}(\boldsymbol{\beta}) = \sum_{i=1}^{n} \mathbf{J}^\top g'(\boldsymbol{\beta}^\top \mathbf{X}_i)\{\mathbf{X}_i - \mathbf{h}(\boldsymbol{\beta}^\top \mathbf{X}_i)\}\rho_1\{g(\boldsymbol{\beta}^\top \mathbf{X}_i)\}[Y_i - \mu\{g(\boldsymbol{\beta}^\top \mathbf{X}_i)\}], \tag{2.7}$$

which is obtained by replacing $\hat{g}, \hat{g}', \hat{\mathbf{h}}$ with $g, g', \mathbf{h}$ in (2.6). One important property of (2.7) is that the second Bartlett identity holds, for any $\boldsymbol{\beta}$:

$$E\{\mathbf{G}(\boldsymbol{\beta})\mathbf{G}^\top(\boldsymbol{\beta})\} = -E\left\{\frac{\partial \mathbf{G}(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}^{(1)}}\right\}.$$

This property makes the semiparametric efficiency of the EFM (2.6) possible.

Let $\boldsymbol{\beta}^0 = (\beta_1^0, \boldsymbol{\beta}^{(1)0^\top})^\top$ denote the true parameter and $\mathbf{B}^+$ denote the Moore–Penrose inverse of any given matrix $\mathbf{B}$. We have the following asymptotic result for the estimator $\hat{\boldsymbol{\beta}}^{(1)}$.

THEOREM 2.1. *Assume the estimating function* (2.6) *has a unique solution and denote it by* $\hat{\boldsymbol{\beta}}^{(1)}$. *If the regularity conditions* (a)–(e) *in the Appendix are satisfied, the following results hold*:

(i) *With* $h \to 0$, $n \to \infty$ *such that* $(nh)^{-1}\log(1/h) \to 0$, $\hat{\boldsymbol{\beta}}^{(1)}$ *converges in probability to the true parameter* $\boldsymbol{\beta}^{(1)0}$.

(ii) *If* $nh^6 \to 0$ *and* $nh^4 \to \infty$,

$$\sqrt{n}\big(\hat{\boldsymbol{\beta}}^{(1)} - \boldsymbol{\beta}^{(1)0}\big) \xrightarrow{\mathcal{L}} N_{d-1}(\mathbf{0}, \boldsymbol{\Sigma}_{\boldsymbol{\beta}^{(1)0}}), \tag{2.8}$$

where $\boldsymbol{\Sigma}_{\boldsymbol{\beta}^{(1)0}} = \{\mathbf{J}^\top \boldsymbol{\Omega}\mathbf{J}\}^+|_{\boldsymbol{\beta}^{(1)}=\boldsymbol{\beta}^{(1)0}}$, $\mathbf{J} = \frac{\partial \boldsymbol{\beta}}{\partial \boldsymbol{\beta}^{(1)}}$ *and*

$$\boldsymbol{\Omega} = E[\{\mathbf{X}\mathbf{X}^\top - E(\mathbf{X}|\boldsymbol{\beta}^\top \mathbf{X})E(\mathbf{X}^\top|\boldsymbol{\beta}^\top \mathbf{X})\}\rho_2\{g(\boldsymbol{\beta}^\top \mathbf{X})\}\{g'(\boldsymbol{\beta}^\top \mathbf{X})\}^2/\sigma^2].$$

REMARK 3. Note that $\boldsymbol{\beta}^\top \boldsymbol{\Omega}\boldsymbol{\beta} = 0$, so the nonnegative matrix $\boldsymbol{\Omega}$ degenerates in the direction of $\boldsymbol{\beta}$. If the mean function μ is the identity function and the variance function is equal to a scale constant, that is, $\mu\{g(\boldsymbol{\beta}^\top \mathbf{X})\} = g(\boldsymbol{\beta}^\top \mathbf{X})$, $\sigma^2 V\{g(\boldsymbol{\beta}^\top \mathbf{X})\} = \sigma^2$, the matrix $\boldsymbol{\Omega}$ in Theorem 2.1 reduces to be

$$\boldsymbol{\Omega} = E[\{\mathbf{X}\mathbf{X}^\top - E(\mathbf{X}|\boldsymbol{\beta}^\top \mathbf{X})E(\mathbf{X}^\top|\boldsymbol{\beta}^\top \mathbf{X})\}\{g'(\boldsymbol{\beta}^\top \mathbf{X})\}^2/\sigma^2].$$

Technically speaking, Theorem 2.1 shows that an undersmoothing approach is unnecessary and that root-n consistency can be achieved. The asymptotic covariance $\boldsymbol{\Sigma}_{\boldsymbol{\beta}^{(1)0}}$ in general can be estimated by replacing terms in its expression by estimates of those terms. The asymptotic normality of $\hat{\boldsymbol{\beta}} = (\hat{\beta}_1, \hat{\boldsymbol{\beta}}^{(1)^\top})^\top$ will follow from Theorem 2.1 with a simple application of the multivariate delta-method,

since $\hat{\beta}_1 = \sqrt{1 - \|\hat{\boldsymbol{\beta}}^{(1)}\|^2}$. According to the results of Carroll et al. (1997), the asymptotic variance of their estimator is $\boldsymbol{\Omega}^+$. Define the block partition of matrix $\boldsymbol{\Omega}$ as follows:

$$\boldsymbol{\Omega} = \begin{pmatrix} \boldsymbol{\Omega}_{11} & \boldsymbol{\Omega}_{12} \\ \boldsymbol{\Omega}_{21} & \boldsymbol{\Omega}_{22} \end{pmatrix}, \tag{2.9}$$

where $\boldsymbol{\Omega}_{11}$ is a positive constant, $\boldsymbol{\Omega}_{12}$ is a $(d-1)$-dimensional row vector, $\boldsymbol{\Omega}_{21}$ is a $(d-1)$-dimensional column vector and $\boldsymbol{\Omega}_{22}$ is a $(d-1) \times (d-1)$ nonnegative definite matrix.

COROLLARY 1. *Under the conditions of Theorem 2.1, we have*

$$\sqrt{n}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^0) \xrightarrow{\mathcal{L}} N_p(\mathbf{0}, \boldsymbol{\Sigma}_{\boldsymbol{\beta}^0}) \tag{2.10}$$

with $\boldsymbol{\Sigma}_{\boldsymbol{\beta}^0} = \mathbf{J}\{\mathbf{J}^\top \boldsymbol{\Omega} \mathbf{J}\}^+ \mathbf{J}^\top|_{\boldsymbol{\beta}=\boldsymbol{\beta}^0}$. *Further,*

$$\boldsymbol{\Sigma}_{\boldsymbol{\beta}^0} \le \boldsymbol{\Omega}^+|_{\boldsymbol{\beta}=\boldsymbol{\beta}^0}$$

and a strict less-than sign holds when $\det(\boldsymbol{\Omega}_{22}) = 0$. *That is, in this case EFM is more efficient than that of Carroll et al.* (1997).

The possible smaller limiting variance derived from the EFM approach partly benefits from the reparameterization so that the quasi-likelihood can be adopted. As we know, the quasi-likelihood is often of optimal property. In contrast, most existing methods treat the estimation of $\boldsymbol{\beta}$ as if it were done in the framework of linear dimension reduction. The target of linear dimension reduction is to find the directions that can linearly transform the original variables vector into a vector of one less dimension. For example, ADE and SIR are two relevant methods. However, when the link function $\mu(\cdot)$ is identity, the limiting variance derived here may not be smaller or equal to the ones of Wang et al. (2010) and Chang, Xue and Zhu (2010) when the quasi-likelihood of (2.5) is applied.

2.3. *Profile quasi-likelihood ratio test.* In applications, it is important to test the statistical significance of added predictors in a regression model. Here we establish a quasi-likelihood ratio statistic to test the significance of certain variables in the linear index. The null hypothesis that the model is correct is tested against a full model alternative. Fan and Jiang (2007) gave a recent review about generalized likelihood ratio tests. Bootstrap tests for nonparametric regression, generalized partially linear models and single-index models have been systematically investigated [see Härdle and Mammen (1993), Härdle, Mammen and Müller (1998),

Härdle, Mammen and Proenca (2001)]. Consider the testing problem:

$$
\begin{aligned}
&H_0 : g(\cdot) = g\left(\sum_{k=1}^{r} \beta_k X_k\right) \\
&\longleftrightarrow \quad H_1 : g(\cdot) = g\left(\sum_{k=1}^{r} \beta_k X_k + \sum_{k=r+1}^{d} \beta_k X_k\right).
\end{aligned} \tag{2.11}
$$

We mainly focus on testing $\beta_k = 0, k = r+1, \ldots, d$, though the following test procedure can be easily extended to a general linear testing $\mathbf{B}\tilde{\boldsymbol{\beta}} = 0$ where $\mathbf{B}$ is a known matrix with full row rank and $\tilde{\boldsymbol{\beta}} = (\beta_{r+1}, \ldots, \beta_d)^\top$. The profile quasi-likelihood ratio test is defined by

$$
T_n = 2\Big\{\sup_{\boldsymbol{\beta}\in\Theta} \hat{Q}(\boldsymbol{\beta}) - \sup_{\boldsymbol{\beta}\in\Theta, \tilde{\boldsymbol{\beta}}=0} \hat{Q}(\boldsymbol{\beta})\Big\}, \tag{2.12}
$$

where $\hat{Q}(\boldsymbol{\beta}) = \sum_{i=1}^{n} Q[\mu\{\hat{g}(\boldsymbol{\beta}^\top \mathbf{X}_i)\}, Y_i]$, $Q[\mu, y] = \int_{\mu}^{y} \frac{s-y}{V\{\mu^{-1}(s)\}}\, ds$ and $\mu^{-1}(\cdot)$ is the inverse function of $\mu(\cdot)$. The following Wilks type theorem shows that the distribution of T_n is asymptotically chi-squared and independent of nuisance parameters.

THEOREM 2.2. *Under the assumptions of Theorem 2.1, if* $\beta_k = 0, k = r+1, \ldots, d$, *then*

$$
T_n \xrightarrow{\mathcal{L}} \chi^2(d-r). \tag{2.13}
$$

3. Numerical studies.

3.1. *Computation of the estimates.* Solving the joint estimating equations (2.1) and (2.6) poses some interesting challenges, since the functions $\hat{g}(\boldsymbol{\beta}^\top \mathbf{X})$ and $\hat{g}'(\boldsymbol{\beta}^\top \mathbf{X})$ depend on $\boldsymbol{\beta}$ implicitly. Treating $\boldsymbol{\beta}^\top X$ as a new predictor (with given $\boldsymbol{\beta}$), (2.1) gives us $\hat{g}, \hat{g}'$ as in Fan, Heckman and Wand (1995). We therefore focus on (2.6), as estimating equations. It cannot be solved explicitly, and hence one needs to find solutions using numerical methods. The Newton–Raphson algorithm is one of the popular and successful methods for finding roots. However, the computational speed of this algorithm crucially depends on the initial value. We propose therefore a fixed-point iterative algorithm that is not very sensitive to starting values and is adaptive to larger dimension. It is worth noting that this algorithm can be implemented in the case that d is slightly larger than n, because the resultant procedure only involves one-dimensional nonparametric smoothers, thereby avoiding the data sparsity problem caused by high dimensionality.

Rewrite the estimating functions as $\hat{\mathbf{G}}(\boldsymbol{\beta}) = \mathbf{J}^\top \hat{\mathbf{F}}(\boldsymbol{\beta})$ with

$$
\hat{\mathbf{F}}(\boldsymbol{\beta}) = (\hat{F}_1(\boldsymbol{\beta}), \ldots, \hat{F}_d(\boldsymbol{\beta}))^\top
$$

and

$$\hat{F}_s(\boldsymbol{\beta}) = \sum_{i=1}^{n} \{X_{si} - \hat{h}_s(\boldsymbol{\beta}^\top \mathbf{X}_i)\} \mu'\{\hat{g}(\boldsymbol{\beta}^\top \mathbf{X}_i)\} \hat{g}'(\boldsymbol{\beta}^\top \mathbf{X}_i) V^{-1} \{\hat{g}(\boldsymbol{\beta}^\top \mathbf{X}_i)\} \times [Y_i - \mu\{\hat{g}(\boldsymbol{\beta}^\top \mathbf{X}_i)\}].$$

Setting $\hat{\mathbf{G}}(\boldsymbol{\beta}) = 0$, we have that

$$\begin{cases} -\beta_2 \hat{F}_1(\boldsymbol{\beta}) / \sqrt{1 - \|\boldsymbol{\beta}^{(1)}\|^2} + \hat{F}_2(\boldsymbol{\beta}) = 0, \\ -\beta_3 \hat{F}_1(\boldsymbol{\beta}) / \sqrt{1 - \|\boldsymbol{\beta}^{(1)}\|^2} + \hat{F}_3(\boldsymbol{\beta}) = 0, \\ \cdots \\ -\beta_d \hat{F}_1(\boldsymbol{\beta}) / \sqrt{1 - \|\boldsymbol{\beta}^{(1)}\|^2} + \hat{F}_d(\boldsymbol{\beta}) = 0. \end{cases} \tag{3.1}$$

Note that $\|\boldsymbol{\beta}^{(1)}\|^2 = \sum_{r=2}^{d} \beta_r^2$, $\beta_1 = \sqrt{1 - \|\boldsymbol{\beta}^{(1)}\|^2}$ and after some simple calculations, we can get that

$$\begin{cases} \beta_1 = |\hat{F}_1(\boldsymbol{\beta})| / \|\hat{\mathbf{F}}(\boldsymbol{\beta})\|, & s = 1, \\ \beta_s^2 = \hat{F}_s^2(\boldsymbol{\beta}) / \|\hat{\mathbf{F}}(\boldsymbol{\beta})\|^2, & s \geq 2, \end{cases}$$

and $\operatorname{sign}\{\beta_s \hat{F}_1(\boldsymbol{\beta})\} = \operatorname{sign}\{\hat{F}_s(\boldsymbol{\beta})\}$, $s \geq 2$. The above equation can also be rewritten as

$$\boldsymbol{\beta} \frac{\hat{F}_1(\boldsymbol{\beta})}{\|\hat{\mathbf{F}}(\boldsymbol{\beta})\|} = \frac{|\hat{F}_1(\boldsymbol{\beta})|}{\|\hat{\mathbf{F}}(\boldsymbol{\beta})\|} \times \frac{\hat{\mathbf{F}}(\boldsymbol{\beta})}{\|\hat{\mathbf{F}}(\boldsymbol{\beta})\|}. \tag{3.2}$$

Then solving the equation (2.6) is equivalent to finding a fixed point for (3.2). Though $\|\boldsymbol{\beta}^{(1)}\| < 1$ holds almost surely in (3.2) and always $\|\boldsymbol{\beta}\| = 1$, there will be some trouble if (3.2) is directly used as iterative equations. Note that the value of $\|\hat{\mathbf{F}}(\boldsymbol{\beta})\|$ is used as denominator that may sometimes be small, which potentially makes the algorithm unstable. On the other hand, the convergence rate of the fixed-point iterative algorithm derived from (3.2) depends on L, where $\|\frac{\partial\{\hat{\mathbf{F}}(\boldsymbol{\beta})|/\|\hat{\mathbf{F}}(\boldsymbol{\beta})\|\}}{\partial \boldsymbol{\beta}}\| \leq L$. For a fast convergence rate, it technically needs a shrinkage value L. An ad hoc fix introduces a constant M, adding $M\boldsymbol{\beta}$ on both sides of (3.2) and dividing by $\hat{F}_1(\boldsymbol{\beta})/\|\hat{\mathbf{F}}(\boldsymbol{\beta})\| + M$:

$$\boldsymbol{\beta} = \frac{M}{\hat{F}_1(\boldsymbol{\beta})/\|\hat{\mathbf{F}}(\boldsymbol{\beta})\| + M} \boldsymbol{\beta} + \frac{|\hat{F}_1(\boldsymbol{\beta})|/\|\hat{\mathbf{F}}(\boldsymbol{\beta})\|^2}{\hat{F}_1(\boldsymbol{\beta})/\|\hat{\mathbf{F}}(\boldsymbol{\beta})\| + M} \hat{\mathbf{F}}(\boldsymbol{\beta}),$$

where M is chosen such that $\hat{F}_1(\boldsymbol{\beta})/\|\hat{\mathbf{F}}(\boldsymbol{\beta})\| + M \neq 0$. In addition, to accelerate the rate of convergence, we reduce the derivative of the term on the right-hand side of the above equality, which can be achieved by choosing some appropriate M. This is the iteration formulation in Step 2. Here the norm of $\boldsymbol{\beta}_{new}$ is not equal to 1 and we have to normalize it again. Since the iteration in Step 2 makes $\boldsymbol{\beta}_{new}$

to violate the identifiability constraint with norm 1, we design (3.2) to include the whole $\boldsymbol{\beta}$ vector. The possibility of renormalization for $\boldsymbol{\beta}_{new}$ avoids the difficulty of controlling $\|\boldsymbol{\beta}_{new}^{(1)}\| < 1$ in each iteration in Step 2.

Based on these observations, the fixed-point iterative algorithm is summarized as:

Step 0. Choose initial values for $\boldsymbol{\beta}$, denoted by $\boldsymbol{\beta}_{old}$.

Step 1. Solve the estimating equation (2.1) with respect to $\boldsymbol{\alpha}$, which yields $\hat{g}(\boldsymbol{\beta}_{old}^{\top}\mathbf{x}_i)$ and $\hat{g}'(\boldsymbol{\beta}_{old}^{\top}\mathbf{x}_i)$, $1 \le i \le n$.

Step 2. Update $\boldsymbol{\beta}_{old}$ with $\boldsymbol{\beta}_{old} = \boldsymbol{\beta}_{new}/\|\boldsymbol{\beta}_{new}\|$ by solving the equation (2.6) in the fixed-point iteration

$$\boldsymbol{\beta}_{new} = \frac{M}{\hat{F}_1(\boldsymbol{\beta}_{old})/\|\hat{F}(\boldsymbol{\beta}_{old})\| + M}\boldsymbol{\beta}_{old} + \frac{|\hat{F}_1(\boldsymbol{\beta}_{old})|/\|\hat{F}(\boldsymbol{\beta}_{old})\|^2}{\hat{F}_1(\boldsymbol{\beta}_{old})/\|\hat{F}(\boldsymbol{\beta}_{old})\| + M}\hat{\mathbf{F}}(\boldsymbol{\beta}_{old}),$$

where M is a constant satisfying $\hat{F}_1(\boldsymbol{\beta})/\|\hat{F}(\boldsymbol{\beta})\| + M \neq 0$ for any $\boldsymbol{\beta}$.

Step 3. Repeat Steps 1 and 2 until $\max_{1\le s\le d} |\beta_{new,s} - \beta_{old,s}| \le tol$ is met with *tol* being a prescribed tolerance.

The final vector $\boldsymbol{\beta}_{new}/\|\boldsymbol{\beta}_{new}\|$ is the estimator of $\boldsymbol{\beta}^0$. Similarly to other direct estimation methods [Horowitz and Härdle (1996)], the preceding calculation is easy to implement. Empirically the initial value for $\boldsymbol{\beta}$, $(1, 1, \ldots, 1)^{\top}/\sqrt{d}$ can be used in the calculations. The Epanechnikov kernel function $K(t) = 3/4(1 - t^2)I(|t| \le 1)$ is used. The bandwidth involved in Step 1 can be chosen to be optimal for estimation of $\hat{g}(t)$ and $\hat{g}'(t)$ based on the observations $\{\boldsymbol{\beta}_{old}^{\top}\mathbf{X}_i, Y_i\}$. So the standard bandwidth selection methods, such as K-fold cross-validation, generalized cross-validation (GCV) and the rule of thumb, can be adopted. In this step, we recommend K-fold cross-validation to determine the optimal bandwidth using the quasi-likelihood as a criterion function. The K-fold cross-validation is not too computationally intensive while making K not take too large values (e.g., $K = 5$). Here we recommend trying a number of smoothing parameters that smooth the data and picking the one that seems most reasonable. As an adjustment factor, M will increase the stability of iteration. Ideally, in each iteration an optimum value for M should be chosen guaranteeing that the derivative on the right-hand side of the iteration formulation in Step 2 is close to zero. Following this idea, M will be depending the changes of $\boldsymbol{\beta}$ and $\hat{\mathbf{F}}(\boldsymbol{\beta})/\|\hat{\mathbf{F}}(\boldsymbol{\beta})\|$. This will be an expensive task due to the computation for the derivative on the right-hand side of the iteration formulation in Step 2. We therefore consider M as constant nonvarying in each iteration, and select M by the K-fold cross-validation method, according to minimizing the model prediction error. When the dimension d gets larger, M will get smaller. In our simulation runs, we empirically search M in the interval $[2/\sqrt{d}, d/2]$. This choice gives pretty good practical performance.

3.2. *Simulation results.*

EXAMPLE 1 (Continuous response). We report a simulation study to investigate the finite-sample performance of the proposed estimator and compare it with the rMAVE [refined MAVE; for details see Xia et al. (2002)] estimator and the EDR estimator [see Hristache et al. (2001), Polzehl and Sperlich (2009)]. We consider the following model similar to that used in Xia (2006):

$$\begin{aligned} &E(Y|\boldsymbol{\beta}^\top\mathbf{X}) = g(\boldsymbol{\beta}^\top\mathbf{X}), \qquad g(\boldsymbol{\beta}^\top\mathbf{X}) = (\boldsymbol{\beta}^\top\mathbf{X})^2\exp(\boldsymbol{\beta}^\top\mathbf{X}); \\ &\mathrm{Var}(Y|\boldsymbol{\beta}^\top\mathbf{X}) = \sigma^2, \qquad \sigma = 0.1. \end{aligned} \tag{3.3}$$

Let the true parameter $\boldsymbol{\beta} = (2, 1, 0, \ldots, 0)^\top/\sqrt{5}$. Two sets of designs for $\mathbf{X}$ are considered: Design (A) and Design (B). In Design (A), $(X_s + 1)/2 \sim \mathrm{Beta}(\tau, 1)$, $1 \le s \le d$ and, in Design (B), $(X_1 + 1)/2 \sim \mathrm{Beta}(\tau, 1)$ and $P(X_s = \pm 0.5) = 0.5$, $s = 2, 3, 4, \ldots, d$. The data generated in Design (A) are not elliptically symmetric. All the components of Design (B) are discrete except for the first component X_1. Y is generated from a normal distribution. This simulation data set consists of 400 observations with 250 replications. The results are shown in Table 1. All rMAVE, EDR and EFM estimates are close to the true parameter vector for $d = 10$. However, the average estimation errors from rMAVE and EDR estimates for $d = 50$ are about 2 and 1.5 times as large as those of the EFM estimates, respectively. This indicates that the fixed-point algorithm is more adaptive to high dimension.

EXAMPLE 2 (Binary response). This simulation design assumes an underlying single-index model for binary responses with

$$\begin{aligned} P(Y = 1|\mathbf{X}) &= \mu\{g(\boldsymbol{\beta}^\top\mathbf{X})\} = \exp\{g(\boldsymbol{\beta}^\top\mathbf{X})\}/[1 + \exp\{g(\boldsymbol{\beta}^\top\mathbf{X})\}], \\ g(\boldsymbol{\beta}^\top\mathbf{X}) &= \exp(5\boldsymbol{\beta}^\top\mathbf{X} - 2)/\{1 + \exp(5\boldsymbol{\beta}^\top\mathbf{X} - 3)\} - 1.5. \end{aligned} \tag{3.4}$$

The underlying coefficients are assumed to be $\boldsymbol{\beta} = (2, 1, 0, \ldots, 0)^\top/\sqrt{5}$. We consider two sets of designs: Design (C) and Design (D). In Design (C), X_1 and X_2

TABLE 1
Average estimation errors $\sum_{s=1}^{d} |\hat{\beta}_s - \beta_s|$ *for model* (3.3)

		Design (A)			Design (B)		
d	τ	rMAVE	EDR	EFM	rMAVE	EDR	EFM
10	0.75	0.0559*	0.0520	0.0792	0.0522*	0.0662	0.0690
10	1.5	0.0323*	0.0316	0.0298	0.0417*	0.0593	0.0457
50	0.75	0.9900	0.7271	0.5425	0.9780	0.7712	0.4515
50	1.5	0.3776	0.3062	0.1796	0.4693	0.4103	0.2211

*The values are adopted from Xia (2006).

TABLE 2
Average estimation errors $\sum_{s=1}^{d} |\hat{\beta}_s - \beta_s|$ *for model* (3.4)

	Design (C)			Design (D)		
d	rMAVE	EDR	EFM	rMAVE	EDR	EFM
10	0.5017	0.5281	0.4564	0.9614	0.9574	0.7415
50	2.0991	1.2695	1.1744	2.5040	2.4846	1.9908

follow the uniform distribution $U(-2, 2)$. In Design (D), X_1 is also assumed to be uniformly distributed in interval $(-2, 2)$ and $(X_2 + 1)/2 \sim \text{Beta}(1, 1)$. Similar designs for generalized partially linear single-index models are assumed in Kane, Holt and Allen (2004). Here a sample size of 700 is used for the case $d = 10$ and 3,000 is used for $d = 50$. Different sample sizes from Example 1 are used due to varying complexity of the two examples. For this example, 250 replications are simulated and the results are displayed in Table 2. In this set of simulations, the average estimation errors from rMAVE estimates and EDR estimates are about 1.5 and 1.2 times as large as EFM estimates, under both Design (C) and Design (D) for $d = 10$ or $d = 50$. The values in the row marked by $d = 50$ look a little bigger. However, it is reasonable because the number of summands in the average estimate error for $d = 50$ is five times as large as that for $d = 10$. Again it appears that the EFM procedure achieves more precise estimators.

EXAMPLE 3 (A simple model). To illustrate the adaptivity of our algorithm to high dimension, we consider the following simple single-index model:

$$Y = (\boldsymbol{\beta}^\top \mathbf{X})^2 + \varepsilon. \tag{3.5}$$

The true parameter is $\boldsymbol{\beta} = (2, 1, 0, \ldots, 0)^\top/\sqrt{5}$; $\mathbf{X}$ is generated from $N_d(2, \mathbf{I})$. Both homogeneous errors and heterogeneous ones are considered. In the former case, $\varepsilon \sim N(0, 0.2^2)$ and in the latter case, $\varepsilon = \exp(\sqrt{5}\boldsymbol{\beta}^\top \mathbf{X}/14)\widetilde{\varepsilon}$ with $\widetilde{\varepsilon} \sim N(0, 1)$. The latter case is designed to show whether our method can handle heteroscedasticity. A similar modeling setup was also used in Wang and Xia (2008), Example 5. The simulated results given in Table 3 are based on 250 replicates with a sample of $n = 100$ observations. An important observation from this simulation is that the proposed EFM approach still works even when the dimension of the parameter is equal to or slightly larger than the number of observations. It can be seen from Table 3 that our approach also performs well under the heteroscedasticity setup.

EXAMPLE 4 (An oscillating function model). A single-index model is designed as

$$Y = \sin(a\boldsymbol{\beta}^\top \mathbf{X}) + \varepsilon, \tag{3.6}$$

TABLE 3
Average estimation errors $\sum_{s=1}^{d} |\hat{\beta}_s - \beta_s|$ *for model* (3.5)

ε		$d = 10$	$d = 50$	$d = 100$	$d = 120$
$\varepsilon \sim N(0, 0.2^2)$	rMAVE	0.0318	0.3484	—	—
	EDR	0.0363	0.5020	—	—
	EFM	0.0272	0.2302	2.9409	5.0010
$\varepsilon \sim N(0, \exp(\frac{2X_1+X_2}{7}))$	rMAVE	0.3427	4.6190	—	—
	EDR	0.2542	2.1112	—	—
	EFM	0.2201	1.7937	4.1435	6.4973

— means that the values cannot be calculated by rMAVE and EDR because of high dimension.

where $\boldsymbol{\beta} = (2, 1, 0, \ldots, 0)^\top/\sqrt{5}$, $\mathbf{X}$ is generated from $N_d(2, \mathbf{I})$ and $\varepsilon \sim N(0, 0.2^2)$. The number of replications is 250 and the sample size $n = 400$. The simulation results are shown in Table 4. In these chosen values for a, we see that EFM performs better than rMAVE and EDR. But as is understood, more oscillating functions are more difficult to handle than those less oscillating functions.

EXAMPLE 5 (Comparison of variance). To make our simulation results comparable with those of Carroll et al. (1997), we mimic their simulation setup. Data of size 200 are generated according to the following model:

$$Y_i = \sin\{\pi(\boldsymbol{\beta}^\top \mathbf{X}_i - A)/(B - A)\} + \alpha Z_i + \varepsilon_i, \tag{3.7}$$

where $\mathbf{X}_i$ are trivariate with independent $U(0, 1)$ components, Z_i are independent of $\mathbf{X}_i$ and $Z_i = 0$ are for i odd and $Z_i = 1$ for i even, and ε_i follow a normal distribution $N(0, 0.01)$ independent of both $\mathbf{X}_i$ and Z_i. The parameters are taken to be $\boldsymbol{\beta} = (1, 1, 1)^\top/\sqrt{3}$, $\alpha = 0.3$, $A = \sqrt{3}/2 - 1.645/\sqrt{12}$ and $B = \sqrt{3}/2 + 1.645/\sqrt{12}$. Note that the EFM approach can still be applicable for this model as the conditionally centered response Y given Z has the model as, because of the independence between $\mathbf{X}$ and Z,

$$Y_i - E(Y_i|Z_i) = a + \sin\{\pi(\boldsymbol{\beta}^\top \mathbf{X}_i - A)/(B - A)\} + \varepsilon_i.$$

TABLE 4
Average estimation errors $\sum_{s=1}^{d} |\hat{\beta}_s - \beta_s|$ *for model* (3.6)

	$a = \pi/2$			$a = 3\pi/4$		
d	rMAVE	EDR	EFM	rMAVE	EDR	EFM
10	0.0981	0.0918	0.0737	0.0970	0.0745	0.0725
50	0.5247	0.6934	0.4355	0.6350	1.8484	0.5407

TABLE 5
Estimation for $\boldsymbol{\beta}$ of model (3.7) based on two randomly chosen samples

	One group of sample			Another group of sample		
	X_1	X_2	X_3	X_1	X_2	X_3
GPLSIM est.	0.595*	0.568*	0.569*	0.563*	0.574*	0.595*
GPLSIM s.e.	0.013*	0.013*	0.013*	0.010*	0.010*	0.010*
EFM est.	0.579	0.575	0.577	0.573	0.577	0.580
EFM s.e.	0.011	0.011	0.011	0.010	0.010	0.010

*The values are adopted from Carroll et al. (1997). We abbreviate "estimator" to "est." and "standard error" to "s.e.," which are computed from the sample version of $\boldsymbol{\Sigma}_{\hat{\boldsymbol{\beta}}}$ defined in (2.10).

As Z_i are dummy variables, estimating $E(Y_i|Z_i)$ is simple. Thus, when we regard $Y_i - E(Y_i|Z_i)$ as response, the model is still a single-index model. Here the number of replications is 100. The method derived from Carroll et al. (1997) is referred to be the GLPSIM approach. The numerical results are reported in Table 5. It shows that compared with the GPLSIM estimates, the EFM estimates have smaller bias and smaller (or equal) variance. Also in this example both EFM and GPLSIM can provide reasonably accurate estimates.

Performance of profile quasi-likelihood ratio test. To illustrate how the profile quasi-likelihood ratio performs for linear hypothesis problems, we simulate the same data as above, except that we allow some components of the index to follow the null hypothesis:

$$H_0 : \beta_4 = \beta_5 = \cdots = \beta_d = 0.$$

We examine the power of the test under a sequence of the alternative hypotheses indexed by parameter δ as follows:

$$H_1 : \beta_4 = \delta, \qquad \beta_s = 0 \qquad \text{for } s \geq 5.$$

When $\delta = 0$, the alternative hypothesis becomes the null hypothesis.

We examine the profile quasi-likelihood ratio test under a sequence of alternative models, progressively deviating from the null hypothesis, namely, as δ increases. The power functions are calculated at the significance level: 0.05, using the asymptotic distribution. We calculate test statistics from 250 simulations by employing the fixed-point algorithm and find the percentage of test statistics greater than or equal to the associated quantile of the asymptotic distribution. The pictures in Figures 1, 2 and 3 illustrate the power function curves for two models under the given significance levels. The power curves increase rapidly with δ, which shows the profile quasi-likelihood ratio test is powerful. When δ is close to 0, the test sizes are all approximately the significance levels.

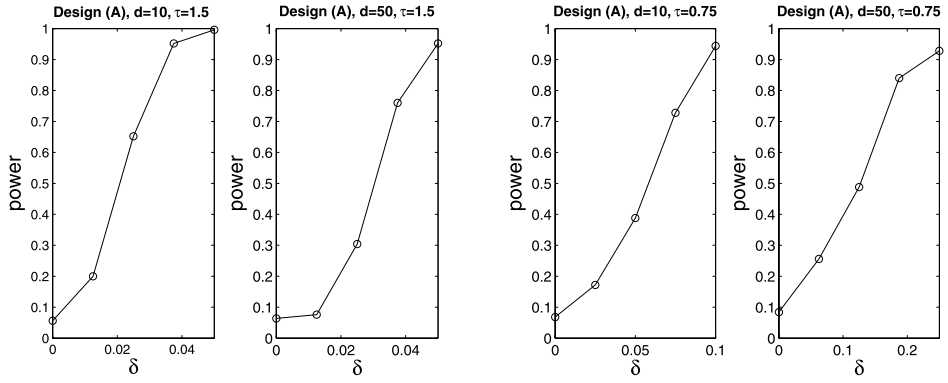

FIG. 1. *Simulation results for Design* (A) *in Example* 1. *The left graphs depict the case* $\tau = 1.5$ *with* τ *the first parameter in* Beta$(\tau, 1)$. *The right graphs are for* $\tau = 0.75$.

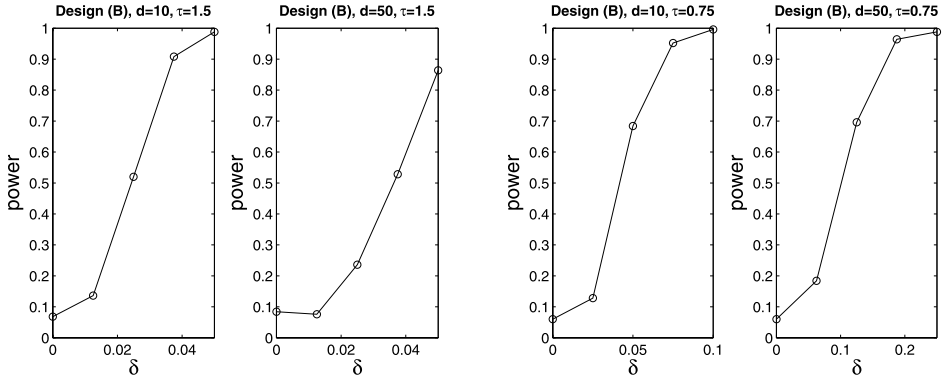

FIG. 2. *Simulation results for Design* (B) *in Example* 1. *The left graphs depict the case* $\tau = 1.5$ *with* τ *the first parameter in* Beta$(\tau, 1)$. *The right graphs are for* $\tau = 0.75$.

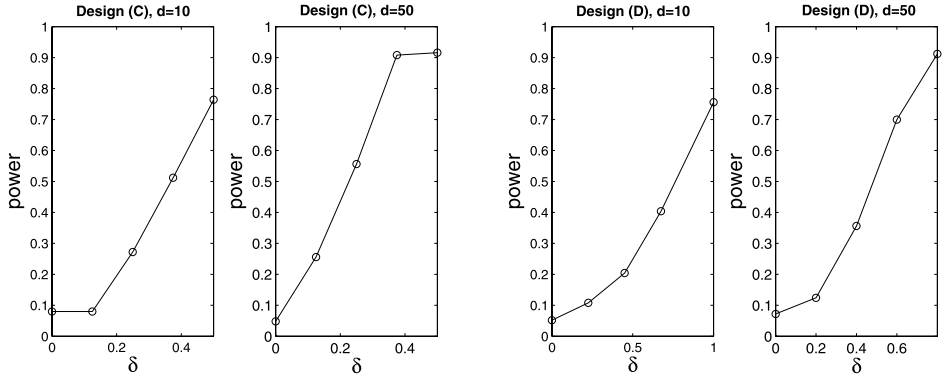

FIG. 3. *Simulation results for Example* 2. *The left graphs depict the case of Design* (C) *with parameter dimension being* 10 *and* 50. *The right graphs are for Design* (D).

3.3. *A real data example*. Income, to some extent, is considered as an index of a successful life. It is generally believed that demographic information, such as education level, relationship in the household, marital status, the fertility rate and gender, among others, has effects on amounts of income. For example, Murray (1997) illustrated that adults with higher intelligence have higher income. Kohavi (1996) predicted income using a Bayesian classifier offered by a machine learning algorithm. Madalozzo (2008) examined income differentials between married women and those who remain single or cohabit by using multivariate linear regression. Here we will use the single-index model to explore the relationship between income and some of its possible determinants.

We use the "Adult" database, which was extracted from the Census Bureau database and is available on website: http://archive.ics.uci.edu/ml/datasets/Adult. It was originally used to model income exceeds over USD 50,000/year based on census data. The purpose of using this example is to understand the personal income patterns and demonstrate the performance of the EFM method in real data analysis. After excluding a few missing data, the data set in our study includes 30,162 subjects. The selected explanatory variables are:

- *sex* (categorical): 1 = Male, 0 = Female.
- *native-country* (categorical): 1 = United-States, 0 = others.
- *work-class* (categorical): 1 = Federal-gov, 2 = Local-gov, 3 = Private, 4 = Self-emp-inc (self-employed, incorporated), 5 = Self-emp-not-inc (self-employed, not incorporated), 6 = State-gov.
- *marital-status* (categorical): 1 = Divorced, 2 = Married-AF-spouse (married, armed forces spouse present), 3 = Married-civ-spouse (married, civilian spouse present), 4 = Married-spouse-absent [married, spouse absent (exc. separated)], 5 = Never-married, 6 = Separated, 7 = Widowed.
- *occupation* (categorical): 1 = Adm-clerical (administrative support and clerical), 2 = Armed-Forces, 3 = Craft-repair, 4 = Exec-managerial (executive-managerial), 5 = Farming-fishing, 6 = Handlers-cleaners, 7 = Machine-op-inspct (machine operator inspection), 8 = Other-service, 9 = Priv-house-serv (private household services), 10 = Prof-specialty (professional specialty), 11 = Protective-serv, 12 = Sales, 13 = Tech-support, 14 = Transport-moving.
- *relationship* (categorical): 1 = Husband, 2 = Not-in-family, 3 = Other-relative, 4 = Own-child, 5 = Unmarried, 6 = Wife.
- *race* (categorical): 1 = Amer-Indian-Eskimo, 2 = Asian-Pac-Islander, 3 = Black, 4 = Other, 5 = White.
- *age* (integer): number of years of age and greater than or equal to 17.
- *fnlwgt* (continuous): The final sampling weights on the CPS files are controlled to independent estimates of the civilian noninstitutional population of the United States.
- *education* (ordinal): 1 = Preschool (less than 1st Grade), 2 = 1st–4th, 3 = 5th–6th, 4 = 7th–8th, 5 = 9th, 6 = 10th, 7 = 11th, 8 = 12th (12th Grade no

Diploma), 9 = HS-grad (high school Grad-Diploma or Equiv), 10 = Some-college (some college but no degree), 11 = Assoc-voc (associate degree-occupational/vocational), 12 = Assoc-acdm (associate degree-academic program), 13 = Bachelors, 14 = Masters, 15 = Prof-school (professional school), 16 = Doctorate.

- *education-num* (continuous): Number of years of education.
- *capital-gain* (continuous): A profit that results from investments into a capital asset.
- *capital-loss* (continuous): A loss that results from investments into a capital asset.
- *hours-per-week* (continuous): Usual number of hours worked per week.

Note that all the explanatory variables up to "age" are categorical with more than two categories. As such, we use dummy variables to link up the corresponding categories. Specifically, for every original explanatory variable up to "age," we use dummy variables to indicate it in which the number of dummy variables is equal to the number of categories minus one. By doing so, we then have 41 explanatory variables, where the first 35 ones are dummy and the remaining ones are continuous. After a preliminary data check, we find that the explanatory variables X_{37} = "fnlwgt," X_{39} = "capital-gain" and X_{40} = "capital-loss" are very skewed to the left and the latter two often take zero value. Before fitting (3.8) we first make a logarithm transformation for these three variables to have log("fnlwgt"), log(1 + "capital-gain") and log(1 + "capital-loss"). To make the explanatory variables comparable in scale, we standardize each of them individually to obtain mean 0 and variance 1. Since "education" and "education-num" are correlated, "education" is dropped from the model and it results in a significantly smaller mean residual deviance.

The single-index model will be used to model the relationship between income and the relevant 43 predictors $\mathbf{X} = (X_1, \ldots, X_{43})^\top$:

$$P(\text{"income"} > 50{,}000|\mathbf{X}) = \exp\{g(\boldsymbol{\beta}^\top \mathbf{X})\}/[1 + \exp\{g(\boldsymbol{\beta}^\top \mathbf{X})\}], \tag{3.8}$$

where $Y = I(\text{"income"} > 50{,}000)$ and $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_{43})^\top$ and β_s represents the effect of the sth predictor. Formally, we are testing the effect of gender, that is,

$$H_0 : \beta_1 = 0 \quad \longleftrightarrow \quad H_1 : \beta_1 \neq 0. \tag{3.9}$$

The fixed-point iterative algorithm is employed to compute the estimate for $\boldsymbol{\beta}$. To illustrate further the practical implications of this approach, we compare our results to those obtained by using an ordinary logistic regression (LR). The coefficients of the two models are given in Table 6. To make the analyses presented in the table comparable, we consider two standardizations. First, we standardize every explanatory variable with mean 0 and variance 1 so that the coefficients can be used to compare the relative influence from different explanatory variables. However, such a standardization does not allow us to compare between the single-index

TABLE 6
Fitted coefficients for model (3.8) (*estimated standard errors in parentheses*)

Variables	$\hat{\boldsymbol{\beta}}$ **of SIM**	$\hat{\boldsymbol{\beta}}$ **of LR**
Sex	0.1102 (0.0028)	0.1975 (0.0181)
Native-country	0.0412 (0.0027)	0.0354 (0.0116)
Work-class		
Federal-gov	0.1237 (0.0059)	0.0739 (0.0108)
Local-gov	0.2044 (0.0065)	0.0155 (0.0135)
Private	−0.2603 (0.0075)	0.0775 (0.0200)
Self-em-inc	0.1252 (0.0068)	0.0520 (0.0112)
Self-emp-not-inc	0.1449 (0.0066)	−0.0157 (0.0147)
Marital-Status		
Divorced	−0.0353 (0.0061)	−0.0304 (0.0264)
Married-AF-spouse	0.0195 (0.0036)	0.0333 (0.0079)
Married-civ-spouse	0.3257 (0.0150)	0.4545 (0.0754)
Married-spouse-absent	−0.0115 (0.0029)	−0.0095 (0.0146)
Never-married	−0.1876 (0.0085)	−0.1452 (0.0370)
Separated	−0.0412 (0.0050)	−0.0221 (0.0179)
Occupation		
Adm-clerical	−0.0302 (0.0050)	0.0131 (0.0164)
Armed-Forces	−0.0086 (0.0031)	−0.0091 (0.0131)
Craft-repair	−0.0913 (0.0050)	0.0263 (0.0146)
Exec-managerial	0.1813 (0.0061)	0.1554 (0.0148)
Farming-fishing	−0.0370 (0.0036)	−0.0772 (0.0125)
Handlers-cleaners	−0.0947 (0.0033)	−0.0662 (0.0153)
Machine-op-inspct	−0.1067 (0.0038)	−0.0290 (0.0133)
Other-service	−0.1227 (0.0045)	−0.1192 (0.0195)
Priv-house-serv	−0.0501 (0.0020)	−0.0833 (0.0379)
Prof-specialty	0.2502 (0.0065)	0.1153 (0.0160)
Protective-serv	0.1954 (0.0061)	0.0508 (0.0095)
Sales	0.0316 (0.0050)	0.0615 (0.0147)
Tech-support	0.0181 (0.0037)	0.0619 (0.0102)
Relationship		
Husband	−0.1249 (0.0093)	−0.3264 (0.0254)
Not-in-family	−0.0932 (0.0093)	−0.2074 (0.0612)
Other-relative	−0.0958 (0.0038)	−0.1498 (0.0219)
Own-child	−0.2218 (0.0076)	−0.3769 (0.0498)
Unmarried	−0.1124 (0.0067)	−0.1739 (0.0446)
Race		
Amer-Indian-Eskimo	−0.0252 (0.0024)	−0.0226 (0.0109)
Asian-Pac-Islander	0.0114 (0.0030)	0.0062 (0.0101)
Black	−0.0300 (0.0024)	−0.0182 (0.0111)
Other	−0.0335 (0.0021)	−0.0286 (0.0129)

TABLE 6
(*Continued*)

Variables	$\hat{\boldsymbol{\beta}}$ **of SIM**	$\hat{\boldsymbol{\beta}}$ **of LR**
Age	0.2272 (0.0042)	0.1798 (0.0111)
Fnlwgt	0.0099 (0.0028)	0.0414 (0.0092)
Education-num	0.4485 (0.0045)	0.3732 (0.0122)
Capital-gain	0.2859 (0.0055)	0.2582 (0.0084)
Capital-loss	0.1401 (0.0042)	0.1210 (0.0078)
Hours-per-week	0.2097 (0.0035)	0.1823 (0.0101)

model and the ordinary logistic regression model. We then further normalize the coefficients to be with Euclidean norm 1, and then the estimates of their standard errors are also adjusted accordingly. The single-index model provides more reasonable results: X_{38} = "education-num" has its strongest positive effect on income; those who got a bachelor's degree or higher seem to have much higher income than those with lower education level. In contrast, results derived from a logistic regression show that "married-civ-spouse" is the largest positive contributor.

Some other interesting conclusions could be obtained by looking at the output. Both "sex" and "native-country" have a positive effect. Persons who worked without pay in a family business, unpaid childcare and others earn a lower income than persons who worked for wages or for themselves. The "fnlwgt" attribute has a positive relation to income. Males are likely to make much more money than females. The expected sign for marital status except the *married* (married-AF-spouse, married-civ-spouse) is negative, given that the household production theory affirms that division of work is efficient when each member of a family dedicates his or her time to the more productive job. Men usually receive relatively better compensation for their time in the labor market than in home production. Thus, the expectation is that married women dedicate more time to home tasks and less to the labor market, and this would imply a different probability of working given the marital status choice.

Also "race" influences the income and Asian or Pacific Islanders seem to make more money than other races. And also, one's income significantly increases as working hours increase. Both "capital-gain" and "capital-loss" have positive effects, so we think that people make more money who can use more money to invest. The presence of young children has a negative influence on the income. "age" accounts for the experience effect and has a positive effect. Hence the conclusion based on the single-index model is consistent with what we expect.

To help with interpretation of the model, plots of $\boldsymbol{\beta}^\top \mathbf{X}$ versus predicted response probability and $\hat{g}(\boldsymbol{\beta}^\top \mathbf{X})$ are generated, respectively, and can be found on the right column in Figure 4. When the estimated single-index is greater than 0, $\hat{g}(\hat{\boldsymbol{\beta}}\mathbf{X})$ shows some degree of curvature. An alternative choice is to fit the data

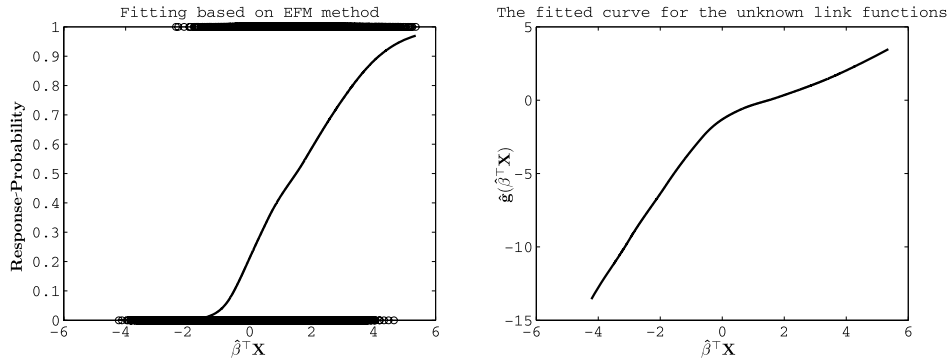

FIG. 4. *Adult data*: *The left graph is a plot of predicted response probability based on the single-index model. The right graph is the fitted curve for the unknown link function* $g(\cdot)$.

using generalized partially linear additive models (GPLAM) with nonparametric components of continuous explanatory variables. The relationships among "age," "fnlwgt," "capital-gain," "capital-loss" and "hours-per-week" all show nonlinearity. The mean residual deviances of SIM, LR and GPLAM are 0.7811, 0.6747 and 0.6240, respectively. SIM under study provides a slightly worse fit than the others. However, we note that LR is, up to a link function, linear about $\mathbf{X}$, and, according to the results of GPLAM, which is a more general model than LR, the actual relationship cannot have such a structure. SIM can reveal nonlinear structure. On the other hand, although the minimum mean residual deviance can be not surprisingly attained by GPLAM, this model has, respectively, ≈ 34 and 41 more degrees of freedom than SIM and LR have.

We now employ the quasi-likelihood ratio test to the test problem (3.9). The QLR test statistic is 166.52 with one degree of freedom, resulting in a P-value of $< 10^{-5}$. Hence this result provides strong evidence that gender has a significant influence on high income.

The Adult data set used in this paper is a rich data set. Existing work mainly focused on the prediction accuracy based on machine learning methods. We make an attempt to explore the semiparametric regression pattern suitable for the data. Model specification and variable selection merit further study.

APPENDIX: OUTLINE OF PROOFS

We first introduce some regularity conditions.

Regularity Conditions:

(a) $\mu(\cdot), V(\cdot), g(\cdot), \mathbf{h}(\cdot) = E(\mathbf{X}|\boldsymbol{\beta}^\top \mathbf{X} = \cdot)$ have two bounded and continuous derivatives. $V(\cdot)$ is uniformly bounded and bounded away from 0.

(b) Let $q(z, y) = \mu'(z)V^{-1}(z)\{y - \mu(z)\}$. Assume that $\partial q(z, y)/\partial z < 0$ for $z \in \mathbb{R}$ and y in the range of the response variable.

(c) The largest eigenvalue of $\mathbf{\Omega}_{22}$ is bounded away from infinity.

(d) The density function $f_{\boldsymbol{\beta}^\top \mathbf{x}}(\boldsymbol{\beta}^\top \mathbf{x})$ of random variable $\boldsymbol{\beta}^\top \mathbf{X}$ is bounded away from 0 on $T_{\boldsymbol{\beta}}$ and satisfies the Lipschitz condition of order 1 on $T_{\boldsymbol{\beta}}$, where $T_{\boldsymbol{\beta}} = \{\boldsymbol{\beta}^\top \mathbf{x} : \mathbf{x} \in T\}$ and T is a compact support set of $\mathbf{X}$.

(e) Let $Q^*[\boldsymbol{\beta}] = \int Q[\mu\{g(\boldsymbol{\beta}^\top \mathbf{x})\}, y] f(y|\boldsymbol{\beta}^{0\top}\mathbf{x}) f(\boldsymbol{\beta}^{0\top}\mathbf{x})\, dy\, d(\boldsymbol{\beta}^{0\top}\mathbf{x})$ with $\boldsymbol{\beta}^0$ denoting the true parameter value and $Q[\mu, y] = \int_\mu^y \frac{s-y}{V\{\mu^{-1}(s)\}}\, ds$. Assume that $Q^*[\boldsymbol{\beta}]$ has a unique maximum at $\boldsymbol{\beta} = \boldsymbol{\beta}^0$, and

$$E\Big[\sup_{\boldsymbol{\beta}^{(1)}} \sup_{\boldsymbol{\beta}^\top \mathbf{X}} |\mu'\{g(\boldsymbol{\beta}^\top \mathbf{X})\} V^{-1}\{g(\boldsymbol{\beta}^\top \mathbf{X})\}[Y - \mu\{g(\boldsymbol{\beta}^\top \mathbf{X})\}]|^2\Big] < \infty$$

and $E\|\mathbf{X}\|^2 < \infty$.

(f) The kernel K is a bounded and symmetric density function with a bounded derivative, and satisfies

$$\int_{-\infty}^{\infty} t^2 K(t)\, dt \neq 0 \quad \text{and} \quad \int_{-\infty}^{\infty} |t|^j K(t)\, dt < \infty, \qquad j = 1, 2, \ldots.$$

Condition (a) is some mild smoothness conditions on the involved functions of the model. We impose condition (b) to guarantee that the solutions of (2.1), $\hat{g}(t)$ and $\hat{g}'(t)$, lie in a compact set. Condition (c) implies that the second moment of estimating equation (2.7), $\operatorname{tr}(\mathbf{J}^\top \mathbf{\Omega} \mathbf{J})$, is bounded. Then the CLT can be applied to $G(\boldsymbol{\beta})$. Condition (d) means that $\mathbf{X}$ may have discrete components and the density function of $\boldsymbol{\beta}^\top \mathbf{X}$ is positive, which ensures that the denominators involved in the nonparametric estimators, with high probability, are bounded away from 0. The uniqueness condition in condition (e) can be checked in the following case for example. Assume that Y is a Poisson variable with mean $\mu\{g(\boldsymbol{\beta}^\top \mathbf{x})\} = \exp\{g(\boldsymbol{\beta}^\top \mathbf{x})\}$. The maximizer β_0 of $Q^*[\boldsymbol{\beta}]$ is equal to the solution of the equation $E[E\{[\exp\{g(\boldsymbol{\beta}^{0\top}\mathbf{X})\} - \exp\{g(\boldsymbol{\beta}^\top \mathbf{X})\}] g'(\boldsymbol{\beta}^\top \mathbf{X}) \mathbf{J}^\top \mathbf{X} | \boldsymbol{\beta}^{0\top}\mathbf{X}\}] = 0$. $\boldsymbol{\beta}_0$ is unique when $g'(\cdot)$ is not a zero-valued constant function and the matrix $\mathbf{J}^\top E(\mathbf{X}\mathbf{X}^\top)\mathbf{J}$ is not singular. Under the second part of condition (e), it is permissible to interchange differentiation and integration when differentiating $E[Q[\mu\{g(\boldsymbol{\beta}^\top \mathbf{X})\}, Y]]$. Condition (f) is a commonly used smoothness condition, including the Gaussian kernel and the quadratic kernel. All of the conditions can be relaxed at the expense of longer proofs.

Throughout the Appendix, $Z_n = \mathcal{O}_P(a_n)$ denotes that $a_n^{-1} Z_n$ is bounded in probability and the derivation for the order of Z_n is based on the fact that $Z_n = \mathcal{O}_P\{\sqrt{E(Z_n^2)}\}$. Therefore, it allows to apply the Cauchy–Schwarz inequality to the quantity having stochastic order a_n.

A.1. Proof of Proposition 1. We outline the proof here, while the details are given in the supplementary materials [Cui, Härdle and Zhu (2010)].

(i) Conditions (a), (b), (d) and (f) are essentially equivalent conditions given by Carroll, Ruppert and Welsh (1998), and as a consequence the derivation of bias and variance for $\hat{g}(\boldsymbol{\beta}^\top \mathbf{x})$ and $\hat{g}'(\boldsymbol{\beta}^\top \mathbf{x})$ is similar to that of Carroll, Ruppert and Welsh (1998).

(ii) The first equation of (2.1) is

$$0 = \sum_{j=1}^{n} K_h(\boldsymbol{\beta}^\top \mathbf{X}_j - \boldsymbol{\beta}^\top \mathbf{x}) \mu'\{\hat{\alpha}_0 + \hat{\alpha}_1(\boldsymbol{\beta}^\top \mathbf{X}_j - \boldsymbol{\beta}^\top \mathbf{x})\}$$
$$\times V^{-1}\{\hat{\alpha}_0 + \hat{\alpha}_1(\boldsymbol{\beta}^\top \mathbf{X}_j - \boldsymbol{\beta}^\top \mathbf{x})\}[Y_j - \mu\{\hat{\alpha}_0 + \hat{\alpha}_1(\boldsymbol{\beta}^\top \mathbf{X}_j - \boldsymbol{\beta}^\top \mathbf{x})\}].$$

Taking derivatives with respect to $\boldsymbol{\beta}^{(1)}$ on both sides, direct observations lead to

$$\frac{\partial \hat{\alpha}_0}{\partial \boldsymbol{\beta}^{(1)}} = \{B(\boldsymbol{\beta}^\top \mathbf{x})\}^{-1}\{A_1(\boldsymbol{\beta}^\top \mathbf{x}) + A_2(\boldsymbol{\beta}^\top \mathbf{x}) + A_3(\boldsymbol{\beta}^\top \mathbf{x})\},$$

where

$$B(\boldsymbol{\beta}^\top \mathbf{x}) = -\sum_{j=1}^{n} K_h(\boldsymbol{\beta}^\top \mathbf{X}_j - \boldsymbol{\beta}^\top \mathbf{x}) q_z'\{\hat{\alpha}_0 + \hat{\alpha}_1(\boldsymbol{\beta}^\top \mathbf{X}_j - \boldsymbol{\beta}^\top \mathbf{x}), Y_j\},$$

$$A_1(\boldsymbol{\beta}^\top \mathbf{x}) = \sum_{j=1}^{n} K_h(\boldsymbol{\beta}^\top \mathbf{X}_j - \boldsymbol{\beta}^\top \mathbf{x}) \mathbf{J}^\top (\mathbf{X}_j - \mathbf{x}) q_z'\{\hat{\alpha}_0 + \hat{\alpha}_1(\boldsymbol{\beta}^\top \mathbf{X}_j - \boldsymbol{\beta}^\top \mathbf{x}), Y_j\}\hat{\alpha}_1,$$

$$A_2(\boldsymbol{\beta}^\top \mathbf{x}) = \sum_{j=1}^{n} K_h(\boldsymbol{\beta}^\top \mathbf{X}_j - \boldsymbol{\beta}^\top \mathbf{x}) q_z'\{\hat{\alpha}_0 + \hat{\alpha}_1(\boldsymbol{\beta}^\top \mathbf{X}_j - \boldsymbol{\beta}^\top \mathbf{x}), Y_j\}$$
$$\times (\boldsymbol{\beta}^\top \mathbf{X}_j - \boldsymbol{\beta}^\top \mathbf{x}) \frac{\partial \hat{\alpha}_1}{\partial \boldsymbol{\beta}^{(1)}},$$

$$A_3(\boldsymbol{\beta}^\top \mathbf{x}) = \sum_{j=1}^{n} h^{-1} K_h'(\boldsymbol{\beta}^\top \mathbf{X}_j - \boldsymbol{\beta}^\top \mathbf{x}) \mathbf{J}^\top (\mathbf{X}_j - \mathbf{x}) q\{\hat{\alpha}_0 + \hat{\alpha}_1(\boldsymbol{\beta}^\top \mathbf{X}_j - \boldsymbol{\beta}^\top \mathbf{x}), Y_j\}$$

with $K_h'(\cdot) = h^{-1}K'(\cdot/h)$. Note that $\partial \hat{\alpha}_0 / \partial \boldsymbol{\beta}^{(1)} = \partial \hat{g}(\boldsymbol{\beta}^\top \mathbf{x}) / \partial \boldsymbol{\beta}^{(1)}$; then we have

$$\frac{\partial \hat{g}(\boldsymbol{\beta}^\top \mathbf{x})}{\partial \boldsymbol{\beta}^{(1)}} = \{B(\boldsymbol{\beta}^\top \mathbf{x})\}^{-1} A_1(\boldsymbol{\beta}^\top \mathbf{x})$$
$$+ \{B(\boldsymbol{\beta}^\top \mathbf{x})\}^{-1} A_2(\boldsymbol{\beta}^\top \mathbf{x}) + \{B(\boldsymbol{\beta}^\top \mathbf{x})\}^{-1} A_3(\boldsymbol{\beta}^\top \mathbf{x}). \tag{A.1}$$

We will prove that

$$E\|\{B(\boldsymbol{\beta}^\top \mathbf{x})\}^{-1} A_1(\boldsymbol{\beta}^\top \mathbf{x}) - g'(\boldsymbol{\beta}^\top \mathbf{x}) \mathbf{J}^\top \{\mathbf{x} - \mathbf{h}(\boldsymbol{\beta}^\top \mathbf{x})\}\|^2$$
$$= \mathcal{O}_P(h^4 + n^{-1}h^{-3}), \tag{A.2}$$

the second term in (A.1) is of order $\mathcal{O}_P(h^4 + n^{-1}h)$, and the third term is of order $\mathcal{O}_P(h^4 + n^{-1}h^{-3})$. The combination of (A.1) and these three results can directly

lead to result (ii) of Proposition 1. The detailed proof is summarized in three steps and is given in the supplementary materials [Cui, Härdle and Zhu (2010)].

(iii) By mimicking the proof of (ii), we can show that (iii) holds. See supplementary materials for details.

A.2. Proofs of (2.6) and (2.7). It is proved in the supplementary materials [Cui, Härdle and Zhu (2010)].

A.3. Proof of Theorem 2.1. (i) Note that the estimating equation defined in (2.6) is just the gradient of the following quasi-likelihood:

$$\hat{Q}(\boldsymbol{\beta}) = \sum_{i=1}^{n} Q[\mu\{\hat{g}(\boldsymbol{\beta}^\top \mathbf{X}_i)\}, Y_i]$$

with $Q[\mu, y] = \int^{\mu} \frac{y-s}{V\{\mu^{-1}(s)\}}\, ds$ and $\mu^{-1}(\cdot)$ is the inverse function of $\mu(\cdot)$. Then for $\boldsymbol{\beta}^{(1)}$ satisfying $(\sqrt{1 - \|\boldsymbol{\beta}^{(1)}\|^2}, \boldsymbol{\beta}^{(1)\top})^\top \in \Theta$, we have

$$\hat{\boldsymbol{\beta}}^{(1)} = \arg\max_{\boldsymbol{\beta}^{(1)}} \hat{Q}(\boldsymbol{\beta}).$$

The proof is based on Theorem 5.1 in Ichimura (1993). In that theorem the consistency of $\boldsymbol{\beta}^{(1)}$ is proved by means of proving that

$$\sup_{\boldsymbol{\beta}^{(1)}} \left| \frac{1}{n} \sum_{i=1}^{n} Q[\mu\{\hat{g}(\boldsymbol{\beta}^\top \mathbf{X}_i)\}, Y_i] - \frac{1}{n} \sum_{i=1}^{n} Q[\mu\{g(\boldsymbol{\beta}^\top \mathbf{X}_i)\}, Y_i] \right| = o_P(1), \tag{A.3}$$

$$\sup_{\boldsymbol{\beta}^{(1)}} \left| \frac{1}{n} \sum_{i=1}^{n} Q[\mu\{g(\boldsymbol{\beta}^\top \mathbf{X}_i)\}, Y_i] - \frac{1}{n} \sum_{i=1}^{n} E[Q[\mu\{g(\boldsymbol{\beta}^\top \mathbf{X}_i)\}, Y_i]] \right| = o_P(1) \tag{A.4}$$

and

$$\left| \frac{1}{n} \sum_{i=1}^{n} Q[\mu\{\hat{g}(\boldsymbol{\beta}_0^\top \mathbf{X}_i)\}, Y_i] - \frac{1}{n} \sum_{i=1}^{n} E[Q[\mu\{g(\boldsymbol{\beta}_0^\top \mathbf{X}_i)\}, Y_i]] \right| = o_P(1). \tag{A.5}$$

Regarding the validity of (A.5), this directly follows from (A.3) and (A.4). The type of uniform convergence result such as (A.4) has been well established in the literature; see, for example, Andrews (1987). We now verify the validity of (A.3), which reduces to showing the uniform convergence of the estimator $\hat{g}(t)$ under condition (e) [see Ichimura (1993)]. This can be obtained in a similar way as in Kong, Linton and Xia (2010), taking into account that the regularity conditions imposed in Theorem 2.1 are stronger than the corresponding ones in that paper.

(ii) Recall the notation $\mathbf{J}$, $\boldsymbol{\Omega}$ and $\mathbf{G}(\boldsymbol{\beta})$ introduced in Section 2. By (2.7), we have shown that

$$\sqrt{n}(\hat{\boldsymbol{\beta}}^{(1)} - \boldsymbol{\beta}^{(1)0}) = \frac{1}{\sqrt{n}} \{\mathbf{J}^\top \boldsymbol{\Omega} \mathbf{J}\}^{+} \mathbf{G}(\boldsymbol{\beta}) + o_P(1). \tag{A.6}$$

Theorem 2.1 follows directly from the above asymptotic expansion and the fact that $E\{\mathbf{G}(\boldsymbol{\beta})\mathbf{G}^\top(\boldsymbol{\beta})\} = n\mathbf{J}^\top \boldsymbol{\Omega} \mathbf{J}$. □

A.4. Proof of Corollary 1. The asymptotic covariance of $\hat{\boldsymbol{\beta}}$ can be obtained by adjusting the asymptotic covariance of $\hat{\boldsymbol{\beta}}^{(1)}$ via the multivariate delta method, and is of form $\mathbf{J}(\mathbf{J}^\top\boldsymbol{\Omega}\mathbf{J})^+\mathbf{J}^\top$. Next we will compare this asymptotic covariance with that (denoted by $\boldsymbol{\Omega}^+$) given in Carroll et al. (1997). Write $\boldsymbol{\Omega}$ as

$$\boldsymbol{\Omega} = \begin{pmatrix} \boldsymbol{\Omega}_{11} & \boldsymbol{\Omega}_{12} \\ \boldsymbol{\Omega}_{21} & \boldsymbol{\Omega}_{22} \end{pmatrix},$$

where $\boldsymbol{\Omega}_{22}$ is a $(d-1)\times(d-1)$ matrix. We will next investigate two cases, respectively: $\det(\boldsymbol{\Omega}_{22}) \neq 0$ and $\det(\boldsymbol{\Omega}_{22}) = 0$. Let $\boldsymbol{\alpha} = -\boldsymbol{\beta}^{(1)}/\sqrt{1-\|\boldsymbol{\beta}^{(1)}\|^2} = -\boldsymbol{\beta}^{(1)}/\beta_1$.

Consider the case that $\det(\boldsymbol{\Omega}_{22}) \neq 0$. Because $\operatorname{rank}(\boldsymbol{\Omega}) = d-1$, $\det(\boldsymbol{\Omega}_{11}\boldsymbol{\Omega}_{22} - \boldsymbol{\Omega}_{21}\boldsymbol{\Omega}_{12}) = 0$. Note that $\boldsymbol{\Omega}_{22}$ is nondegenerate; it can be easily shown that $\boldsymbol{\Omega}_{11} = \boldsymbol{\Omega}_{12}\boldsymbol{\Omega}_{22}^{-1}\boldsymbol{\Omega}_{21}$. Combining this with the following fact:

$$\begin{aligned}\mathbf{J}^\top\boldsymbol{\Omega}\mathbf{J} &= (\boldsymbol{\alpha} \quad \mathbf{I}_{d-1}) \begin{pmatrix} \boldsymbol{\Omega}_{11} & \boldsymbol{\Omega}_{12} \\ \boldsymbol{\Omega}_{21} & \boldsymbol{\Omega}_{22} \end{pmatrix} \begin{pmatrix} \boldsymbol{\alpha}^\tau \\ \mathbf{I}_{d-1} \end{pmatrix} \\ &= \boldsymbol{\Omega}_{22} + (\boldsymbol{\Omega}_{21}/\sqrt{\boldsymbol{\Omega}_{11}} + \sqrt{\boldsymbol{\Omega}_{11}}\boldsymbol{\alpha})(\boldsymbol{\Omega}_{12}/\sqrt{\boldsymbol{\Omega}_{11}} + \sqrt{\boldsymbol{\Omega}_{11}}\boldsymbol{\alpha}^\top) - \boldsymbol{\Omega}_{21}\boldsymbol{\Omega}_{12}/\boldsymbol{\Omega}_{11},\end{aligned}$$

we can get that $\mathbf{J}^\top\boldsymbol{\Omega}\mathbf{J}$ is nondegenerate. In this situation, its inverse $(\mathbf{J}^\top\boldsymbol{\Omega}\mathbf{J})^+$ is just the ordinary inverse $(\mathbf{J}^\top\boldsymbol{\Omega}\mathbf{J})^{-1}$. Then $\mathbf{J}(\mathbf{J}^\top\boldsymbol{\Omega}\mathbf{J})^+\mathbf{J}^\top = \{\mathbf{J}(\mathbf{J}^\top\boldsymbol{\Omega}\mathbf{J})^{-1/2}\}\{(\mathbf{J}^\top \times \boldsymbol{\Omega}\mathbf{J})^{-1/2}\mathbf{J}^\top\}$, a full-rank decomposition. Then

$$\begin{aligned}\{\mathbf{J}(\mathbf{J}^\top\boldsymbol{\Omega}\mathbf{J})^+\mathbf{J}^\top\}^+ &= \{\mathbf{J}(\mathbf{J}^\top\boldsymbol{\Omega}\mathbf{J})^{-1/2}\} \\ &\quad \times \{(\mathbf{J}^\top\boldsymbol{\Omega}\mathbf{J})^{-1/2}\mathbf{J}^\top\mathbf{J}(\mathbf{J}^\top\boldsymbol{\Omega}\mathbf{J})^{-1}\mathbf{J}^\top\mathbf{J}(\mathbf{J}^\top\boldsymbol{\Omega}\mathbf{J})^{-1/2}\}^{-1} \\ &\quad \times \{(\mathbf{J}^\top\boldsymbol{\Omega}\mathbf{J})^{-1/2}\mathbf{J}^\top\} \\ &= \mathbf{J}(\mathbf{J}^\top\mathbf{J})^{-1}\mathbf{J}^\top\boldsymbol{\Omega}\mathbf{J}(\mathbf{J}^\top\mathbf{J})^{-1}\mathbf{J}^\top \\ &= \boldsymbol{\Omega}.\end{aligned}$$

This means that $\mathbf{J}(\mathbf{J}^\top\boldsymbol{\Omega}\mathbf{J})^+\mathbf{J}^\top = \boldsymbol{\Omega}^+$.

When $\det(\boldsymbol{\Omega}_{22}) = 0$, we can obtain that

$$\boldsymbol{\Omega}^+ = \begin{pmatrix} 1/\boldsymbol{\Omega}_{11} + \boldsymbol{\Omega}_{12}\boldsymbol{\Omega}_{22.1}^+\boldsymbol{\Omega}_{21}/\boldsymbol{\Omega}_{11}^2 & -\boldsymbol{\Omega}_{12}\boldsymbol{\Omega}_{22.1}^+/\boldsymbol{\Omega}_{11} \\ -\boldsymbol{\Omega}_{22.1}^+\boldsymbol{\Omega}_{21}/\boldsymbol{\Omega}_{11} & \boldsymbol{\Omega}_{22.1}^+ \end{pmatrix}$$

with $\boldsymbol{\Omega}_{22.1} = \boldsymbol{\Omega}_{22} - \boldsymbol{\Omega}_{21}\boldsymbol{\Omega}_{12}/\boldsymbol{\Omega}_{11}$. Write $\mathbf{J}(\mathbf{J}^\top\boldsymbol{\Omega}\mathbf{J})^+\mathbf{J}^\top$ as

$$\begin{pmatrix} \boldsymbol{\alpha}^\top(\mathbf{J}^\top\boldsymbol{\Omega}\mathbf{J})^+\boldsymbol{\alpha} & \boldsymbol{\alpha}^\top(\mathbf{J}^\top\boldsymbol{\Omega}\mathbf{J})^+ \\ (\mathbf{J}^\top\boldsymbol{\Omega}\mathbf{J})^+\boldsymbol{\alpha} & (\mathbf{J}^\top\boldsymbol{\Omega}\mathbf{J})^+ \end{pmatrix}.$$

Note that $\mathbf{J}^\top\boldsymbol{\Omega}\mathbf{J} = \boldsymbol{\Omega}_{22.1} + (\boldsymbol{\Omega}_{21}/\sqrt{\boldsymbol{\Omega}_{11}} + \sqrt{\boldsymbol{\Omega}_{11}}\boldsymbol{\alpha})(\boldsymbol{\Omega}_{12}/\sqrt{\boldsymbol{\Omega}_{11}} + \sqrt{\boldsymbol{\Omega}_{11}}\boldsymbol{\alpha}^\top)$, so $\mathbf{J}^\top\boldsymbol{\Omega}\mathbf{J} \geq \boldsymbol{\Omega}_{22.1}$. Combining this with $\operatorname{rank}(\boldsymbol{\Omega}_{22}) = d-2$, we have that $(\mathbf{J}^\top\boldsymbol{\Omega}\mathbf{J})^+ \leq \boldsymbol{\Omega}_{22.1}^+$. It is easy to check that $\boldsymbol{\alpha}^\top\boldsymbol{\Omega}_{22.1} = 0$, so $\boldsymbol{\alpha} \perp \operatorname{span}(\boldsymbol{\Omega}_{22.1})$ and $\boldsymbol{\alpha}^\top\boldsymbol{\Omega}_{22.1}^+\boldsymbol{\alpha} = 0$, and then $\boldsymbol{\alpha}^\top(\mathbf{J}^\top\boldsymbol{\Omega}\mathbf{J})^+ = 0$. In this situation, $\mathbf{J}(\mathbf{J}^\top\boldsymbol{\Omega}\mathbf{J})^+\mathbf{J}^\top \leq \boldsymbol{\Omega}^+$ and the stick less-than sign holds since $\mathbf{J}^\top\boldsymbol{\Omega}\mathbf{J} \neq \boldsymbol{\Omega}_{22.1}$ and $1/\boldsymbol{\Omega}_{11} > 0$. □

A.5. Proof of Theorem 2.2. Under H_0, we can rewrite the index vector as $\boldsymbol{\beta} = [\mathbf{e}\ \mathbf{B}]^\top(\sqrt{1-\|\boldsymbol{\omega}^{(1)}\|^2}, \boldsymbol{\omega}^{(1)\tau})^\top$ where $\mathbf{e} = (1, 0, \ldots, 0)^\top$ is an r-dimensional vector,

$$\mathbf{B} = \begin{pmatrix} \mathbf{0}^\top & 0 \\ \mathbf{I}_{r-1} & \mathbf{0} \end{pmatrix}$$

is an $r \times (d-1)$ matrix and $\boldsymbol{\omega}^{(1)} = (\beta_2, \ldots, \beta_r)^\top$ is an $(r-1) \times 1$ vector. Let $\boldsymbol{\omega} = (\sqrt{1-\|\boldsymbol{\omega}^{(1)}\|^2}, \boldsymbol{\omega}^{(1)\top})^\top$. So under H_0 the estimator is also the local maximizer $\hat{\boldsymbol{\omega}}$ of the problem

$$\hat{Q}([\mathbf{e}\quad \mathbf{B}]^\top\hat{\boldsymbol{\omega}}) = \sup_{\|\boldsymbol{\omega}^{(1)}\|<1} \hat{Q}([\mathbf{e}\quad \mathbf{B}]^\top\boldsymbol{\omega}).$$

Expanding $\hat{Q}(\mathbf{B}^\top\hat{\boldsymbol{\omega}})$ at $\hat{\boldsymbol{\beta}}^{(1)}$ by a Taylor's expansion and noting that $\partial\hat{Q}(\boldsymbol{\beta})/\partial\boldsymbol{\beta}^{(1)}|_{\boldsymbol{\beta}^{(1)}=\hat{\boldsymbol{\beta}}^{(1)}} = 0$, then $\hat{Q}(\hat{\boldsymbol{\beta}}) - \hat{Q}(\mathbf{B}^\top\hat{\boldsymbol{\omega}}) = T_1 + T_2 + \mathcal{O}_P(1)$, where

$$T_1 = -\frac{1}{2}(\hat{\boldsymbol{\beta}}^{(1)} - \mathbf{B}^\top\hat{\boldsymbol{\omega}})^\top \frac{\partial^2\hat{Q}(\boldsymbol{\beta})}{\partial\boldsymbol{\beta}^{(1)}\partial\boldsymbol{\beta}^{(1)\tau}}\bigg|_{\boldsymbol{\beta}^{(1)}=\hat{\boldsymbol{\beta}}^{(1)}} (\hat{\boldsymbol{\beta}}^{(1)} - \mathbf{B}^\top\hat{\boldsymbol{\omega}}),$$

$$T_2 = \frac{1}{6}(\hat{\boldsymbol{\beta}}^{(1)} - \mathbf{B}^\top\hat{\boldsymbol{\omega}})^\top \times \frac{\partial\{(\hat{\boldsymbol{\beta}}^{(1)} - \mathbf{B}^\top\hat{\boldsymbol{\omega}})^\top\partial^2\hat{Q}(\boldsymbol{\beta})/(\partial\boldsymbol{\beta}^{(1)}\,\partial\boldsymbol{\beta}^{(1)\tau})|_{\boldsymbol{\beta}^{(1)}=\hat{\boldsymbol{\beta}}^{(1)}}(\hat{\boldsymbol{\beta}}^{(1)} - \mathbf{B}^\top\hat{\boldsymbol{\omega}})\}}{\partial\boldsymbol{\beta}^{(1)}}.$$

Assuming the conditions in Theorem 2.1 and under the null hypothesis H_0, it is easy to show that

$$\sqrt{n}(\mathbf{B}^\top\hat{\boldsymbol{\omega}} - \mathbf{B}^\top\boldsymbol{\omega}) = \frac{1}{\sqrt{n}}\mathbf{B}^\top\mathbf{B}(\mathbf{J}^\top\boldsymbol{\Omega}\mathbf{J})^+\mathbf{G}(\boldsymbol{\beta}) + \mathcal{O}_P(1).$$

Combining this with (A.6), under the null hypothesis H_0,

$$\begin{aligned}\sqrt{n}(\hat{\boldsymbol{\beta}}^{(1)} - \mathbf{B}^\top\hat{\boldsymbol{\omega}}^{(1)}) &= \frac{1}{\sqrt{n}}(\mathbf{J}^\top\boldsymbol{\Omega}\mathbf{J})^{1/2+}\{\mathbf{I}_{d-1} - (\mathbf{J}^\top\boldsymbol{\Omega}\mathbf{J})^{1/2}\mathbf{B}^\top\mathbf{B}(\mathbf{J}^\top\boldsymbol{\Omega}\mathbf{J})^{1/2+}\} \\ &\quad \times (\mathbf{J}^\top\boldsymbol{\Omega}\mathbf{J})^{1/2+}\mathbf{G}(\boldsymbol{\beta}) + o_P(1).\end{aligned} \tag{A.7}$$

Since $\frac{1}{\sqrt{n}}\mathbf{G}(\boldsymbol{\beta}) = \mathcal{O}_P(1)$, $\frac{\partial^2\hat{Q}(\boldsymbol{\beta})}{\partial\boldsymbol{\beta}^{(1)}\partial\boldsymbol{\beta}^{(1)\tau}}|_{\boldsymbol{\beta}^{(1)}} = -n\mathbf{J}^\top\boldsymbol{\Omega}\mathbf{J} + \mathcal{O}_P(n)$ and matrix $\mathbf{J}^\top\boldsymbol{\Omega}\mathbf{J}$ has eigenvalues uniformly bounded away from 0 and infinity, we have $\|\hat{\boldsymbol{\beta}}^{(1)} - \mathbf{B}^\top\hat{\boldsymbol{\omega}}^{(1)}\| = \mathcal{O}_P(n^{-1/2})$ and then $|T_2| = \mathcal{O}_P(1)$. Combining this and (A.7), we have

$$\begin{aligned}\hat{Q}(\hat{\boldsymbol{\beta}}) - \hat{Q}(\mathbf{B}^\top\hat{\boldsymbol{\omega}}) &= \frac{n}{2}(\hat{\boldsymbol{\beta}}^{(1)} - \mathbf{B}^\top\hat{\boldsymbol{\omega}}^{(1)})^\top\mathbf{J}^\top\boldsymbol{\Omega}\mathbf{J}(\hat{\boldsymbol{\beta}}^{(1)} - \mathbf{B}^\top\hat{\boldsymbol{\omega}}^{(1)}) \\ &= \frac{n}{2}\mathbf{G}^\top(\boldsymbol{\beta})(\mathbf{J}^\top\boldsymbol{\Omega}\mathbf{J})^{1/2+}\mathbf{P}(\mathbf{J}^\top\boldsymbol{\Omega}\mathbf{J})^{1/2+}\mathbf{G}(\boldsymbol{\beta})\end{aligned}$$

with $\mathbf{P} = \mathbf{I}_{d-1} - (\mathbf{J}^\top \boldsymbol{\Omega} \mathbf{J})^{1/2} \mathbf{B}^\top \mathbf{B} (\mathbf{J}^\top \boldsymbol{\Omega} \mathbf{J})^{1/2+}$. Here $\mathbf{P}$ is idempotent having rank $d - r$, so it can be written as $\mathbf{P} = \mathbf{S}^\top \mathbf{S}$ where $\mathbf{S}$ ia a $(d - r) \times (d - 1)$ matrix satisfying $\mathbf{S}\mathbf{S}^\top = \mathbf{I}_{d-r}$. Consequently,

$$\begin{aligned} 2\{\hat{Q}(\hat{\boldsymbol{\beta}}) - \hat{Q}(\mathbf{B}^\top \hat{\boldsymbol{\omega}})\} &= \big(\sqrt{n}\mathbf{S}(\mathbf{J}^\top \boldsymbol{\Omega} \mathbf{J})^{1/2+} \mathbf{G}(\boldsymbol{\beta})\big)^\top \big(\sqrt{n}\mathbf{S}(\mathbf{J}^\top \boldsymbol{\Omega} \mathbf{J})^{1/2+} \mathbf{G}(\boldsymbol{\beta})\big) \\ &\xrightarrow{\mathcal{L}} \chi^2(d - r). \end{aligned}$$

Acknowledgments. The authors thank the Associate Editor and two referees for their constructive comments and suggestions which led to a great improvement over an early manuscript.

SUPPLEMENTARY MATERIAL

Supplementary materials (DOI: 10.1214/10-AOS871SUPP; .pdf). Complete proofs of Proposition 1, (2.6) and (2.7).

REFERENCES

ANDREWS, D. W. K. (1987). Consistency in nonlinear econometric models: A genetic uniform law of large numbers. *Econometrica* **55** 1465–1471. MR0923471

CARROLL, R. J., RUPPERT, D. and WELSH, A. H. (1998). Local estimating equations. *J. Amer. Statist. Assoc.* **93** 214–227. MR1614624

CARROLL, R. J., FAN, J., GIJBELS, I. and WAND, M. P. (1997). Generalized partially linear single-index models. *J. Amer. Statist. Assoc.* **92** 447–489. MR1467842

CHANG, Z. Q., XUE, L. G. and ZHU, L. X. (2010). On an asymptotically more efficient estimation of the single-index model. *J. Multivariate Anal.* **101** 1898–1901. MR2651964

CUI, X., HÄRDLE, W. and ZHU, L. (2010). Supplementary materials for "The EFM approach for single-index models." DOI:10.1214/10-AOS871SUPP.

FAN, J. and GIJBELS, I. (1996). *Local Polynomial Modeling and Its Applications*. Chapman & Hall, London. MR1383587

FAN, J., HECKMAN, N. E. and WAND, M. P. (1995). Local polynomial kernel regression for generalized linear models and quasi-likelihood functions. *J. Amer. Statist. Assoc.* **90** 141–150. MR1325121

FAN, J. and JIANG, J. (2007). Nonparametric inference with generalized likelihood ratio test. *Test* **16** 409–478. MR2365172

HÄRDLE, W., HALL, P. and ICHIMURA, H. (1993). Optimal smoothing in single-index models. *Ann. Statist.* **21** 157–178. MR1212171

HÄRDLE, W. and MAMMEN, E. (1993). Testing parametric versus nonparametric regression. *Ann. Statist.* **21** 1926–1947. MR1245774

HÄRDLE, W., MAMMEN, E. and MÜLLER, M. (1998). Testing parametric versus semiparametric modelling in generalized linear models. *J. Amer. Statist. Assoc.* **93** 1461–1474. MR1666641

HÄRDLE, W., MAMMEN, E. and PROENCA, I. (2001). A bootstrap test for single index models. *Statistics* **35** 427–452. MR1880174

HÄRDLE, W. and STOKER, T. M. (1989). Investigating smooth multiple regression by method of average derivatives. *J. Amer. Statist. Assoc.* **84** 986–995. MR1134488

HEYDE, C. C. (1997). *Quasi-likelihood and Its Application*: *A General Approach to Optimal Parameter Estimation*. Springer, New York. MR1461808

HOROWITZ, J. L. and HÄRDLE, W. (1996). Direct semiparametric estimation of a single-index model with discrete covariates. *J. Amer. Statist. Assoc.* **91** 1632–1640. MR1439104

HRISTACHE, M., JUDITSKI, A. and SPOKOINY, V. (2001). Direct estimation of the index coefficients in a single-index model. *Ann. Statist.* **29** 595–623. MR1865333

HRISTACHE, M., JUDITSKY, A., POLZEHL, J. and SPOKOINY, V. (2001). Structure adaptive approach for dimension reduction. *Ann. Statist.* **29** 1537–1566. MR1891738

HUH, J. and PARK, B. U. (2002). Likelihood-based local polynomial fitting for single-index models. *J. Multivariate Anal.* **80** 302–321. MR1889778

ICHIMURA, H. (1993). Semiparametric least squares (SLS) and weighted SLS estimation of single-index models. *J. Econometrics* **58** 71–120. MR1230981

KANE, M., HOLT, J. and ALLEN, B. (2004). Results concerning the generalized partially linear single-index model. *J. Stat. Comput. Simul.* **72** 897–912. MR2100843

KOHAVI, R. (1996). Scaling up the accuracy of naive-Bayes classifiers: A decision-tree hybrid. In *Proceedings of the Second International Conference on Knowledge Discovery and Data Mining* 202–207. AAAI Press, Menlo Park, CA.

KONG, E., LINTON, O. and XIA, Y. (2010). Uniform Bahadur representation for local polynomial estimates of M-regression and its application to the additive model. *Econometric Theory* **26** 1529–1564. MR2684794

LIN, W. and KULASEKERA, K. B. (2007). Identifiability of single-index models and additive-index models. *Biometrika* **94** 496–501. MR2380574

MADALOZZO, R. C. (2008). An analysis of income differentials by marital status. *Estudos Econômicos* **38** 267–292.

MCCULLAGH, P. and NELDER, J. A. (1989). *Generalized Linear Models*, 2nd ed. Champman & Hall, London.

MURRAY, C. (1997). IQ and economic success. *The Public Interest* **128** 21–35.

POLZEHL, J. and SPERLICH, S. (2009). A note on structural adaptive dimension reduction. *J. Stat. Comput. Simul.* **79** 805–818. MR2751594

POWELL, J. L., STOCK, J. H. and STOKER, T. M. (1989). Semiparametric estimation of index coefficients. *Econometrica* **57** 1403–1430. MR1035117

WANG, H. and XIA, Y. (2008). Sliced regression for dimension reduction. *J. Amer. Statist. Assoc.* **103** 811–821. MR2524332

WANG, J. L., XUE, L. G., ZHU, L. X. and CHONG, Y. S. (2010). Estimation for a partial-linear single-index model. *Ann. Statist.* **38** 246–274. MR2589322

XIA, Y. (2006). Asymptotic distributions for two estimators of the single-index model. *Econometric Theory* **22** 1112–1137. MR2328530

XIA, Y., TONG, H., LI, W. K. and ZHU, L. (2002). An adaptive estimation of dimension reduction space (with discussions). *J. R. Stat. Soc. Ser. B Stat. Methodol.* **64** 363–410. MR1924297

YU, Y. and RUPPERT, D. (2002). Penalized spline estimation for partially linear single index models. *J. Amer. Statist. Assoc.* **97** 1042–1054. MR1951258

ZHOU, J. and HE, X. (2008). Dimension reduction based on constrained canonical correlation and variable filtering. *Ann. Statist.* **36** 1649–1668. MR2435451

ZHU, L. X. and XUE, L. G. (2006). Empirical likelihood confidence regions in a partially linear single-index model. *J. R. Stat. Soc. Ser. B Stat. Methodol.* **68** 549–570. MR2278341

ZHU, L. P. and ZHU, L. X. (2009a). Nonconcave penalized inverse regression in single-index models with high dimensional predictors. *J. Multivariate Anal.* **100** 862–875. MR2498719

ZHU, L. P. and ZHU, L. X. (2009b). On distribution weighted partial least squares with diverging number of highly correlated predictors. *J. R. Stat. Soc. Ser. B Stat. Methodol.* **71** 525–548. MR2649607

X. CUI
SCHOOL OF MATHEMATICS
AND COMPUTATIONAL SCIENCE
SUN YAT-SEN UNIVERSITY
GUANGZHOU
GUANGDONG PROVINCE, 510275
P.R. CHINA
E-MAIL: cuixia@mail.sysu.edu.cn

W. K. HÄRDLE
CASE-CENTER FOR APPLIED STATISTICS
AND ECONOMICS
HUMBOLDT-UNIVERSITÄT ZU BERLIN
WIRTSCHAFTSWISSENSCHAFTLICHE FAKULTÄT
SPANDAUER STR. 1
10178 BERLIN
GERMANY
E-MAIL: haerdle@wiwi.hu-berlin.de

L. ZHU
FSC1207, FONG SHU CHUEN BUILDING
DEPARTMENT OF MATHEMATICS
HONG KONG BAPTIST UNIVERSITY
KOWLOON TONG
HONG KONG
P.R. CHINA
E-MAIL: lzhu@hkbu.edu.hk