Computing Science and Statistics, Interface Meeting Sydney 1996

Nonparametric Time Series Model Selection*

W. Härdle
Humboldt-Universität zu Berlin
Wirtschaftswissenschaftliche Fakultät
Institut für Statistik und Ökonometrie
Spandauer Strasse 1 D - 10178 Berlin
Germany

L. Yang
Humboldt-Universität zu Berlin
Wirtschaftswissenschaftliche Fakultät
Institut für Statistik und Ökonometrie
Spandauer Strasse 1 D - 10178 Berlin
Germany

July 3, 1996

Abstract

Nonparametric procedures are an interesting alternative to classical time series analysis. The nonparametric technique follows the principle of 'letting the data speak for themselves,' and provides guidance in choosing a parametric models. In this paper local polynomial estimators are given for vector conditional heteroskedastic autoregressive nonlinear (CHARN) model in which both the conditional mean and the conditional variance (volatility) matrix are unknown functions of the past. We examine the rates of convergence of these estimators and their asymptotic normality. These are applied to estimation of volatility matrices of foreign exchange rates. As the usual nonparametric models often have less than satisfactory performance when dealing with more than one lag, we also give the joint estimation of the additive mean and the multiplicative volatility, which fully exploits the additive/multiplicative structure. We then discuss the usefulness of this approach in selecting the correct lags without assumptions on the forms of the structures.

1 Nonparametric Vector Autoregression

There are two very important aspects of model selection: the selection of form of the function, and the selection of significant variables. We concentrate on the first aspect in Sections 1 to 3, which also provide the necessary tools for investigating the second aspect in Section 4.

Multivariate time series occur in the modelling of dynamics over time and help explaining interdependence among variables. A common model in this context is vector autoregression where the dynamics over time is modeled via a linear operation on the past values of the vector time series, see Lütkepohl (1991). One restrictive element is that the conditional covariance is assumed to be fixed or of specific form. Since the beginning of the eighties this drawback has been stressed by Engle (1982), Robinson (1983, 1984), Teräsvirta (1994) in the econometric literature and by Collomb (1984), Tjøstheim (1994), McKeague and Zhang (1994), and Vieu (1994) in the statistical literature. Nonlinear time series models in this context are threshold autoregressive (TAR) models of Tong (1978, 1983), the exponential autoregressive (EXPAR) models of Haggan and Ozaki (1981), the smooth-transition autoregressive (STAR) models of Chan and Tong (1986) and Granger and Teräsvirta (1992).

The nonparametric modelling of mean function and the volatility matrix offers a way out, since it does not depend on specific structures of these quantities. In the framework of ARCH models, recently non- and semiparametric approaches (Gregory, 1989; Engle and Gonzalez-Rivera, 1991) have been proposed. Engle and Ng (1993) measured the impact of news on volatility and found asymmetric volatility functions. Gouriéroux and Monfort (1992) models both the conditional mean and the conditional variance nonparametrically. Their model

$$Y_i = \sum_{j=1}^{J} \alpha_j \, I(X_i \in A_j) + \sum_{j=1}^{J} \beta_j \, I(X_i \in A_j)\xi_i, i = 1, 2, ...$$

$$X_i = (Y_{i-1}, Y_{i-2}, ..., Y_{i-m}) \in I\!R^{md}, Y_i \in I\!R^d \qquad (1.1)$$

is called a qualitative threshold ARCH model. Here $\{A_j\}_{j=1}^J$ with fixed J denotes a partition of the set of lagged values for $Y, (\alpha_j), (\beta_j)$ are unknown parameter vectors and matrices respectively, and ξ_i is white noise.

*Acknowledgements: We would like to thank Alexander Tsybakov, Rolf Tschernig, Helmut Lütkepohl and Christian Hafner for helpful discussions. The research was supported by Sonderforschungsbereich 373 "Quantifikation und Simulation Ökonomischer Prozesse" Deutsche Forschungsgemeinschaft.

(1997) Härdle, W. and Yang, L.J.
Nonparametric Time Series Analysis Model Selection.

It generalizes the threshold model of Tong (1983) but shares with it the drawback of a fixed number J of threshold points.

A generalization of model (1.1) to a wider class of conditional mean and variance functions can be seen as a limit of (1.1) for $J \to \infty$ thus allowing J to be unknown

$$Y_i = f(X_i) + \Sigma^{1/2}(X_i)\xi_i, \quad i = 1, 2, \ldots \tag{1.2}$$
$$X_i = (Y_{i-1}, Y_{i-2}, ..., Y_{i-m}) \in I\!R^{md}$$
$$Y_i = (Y_{i1}, Y_{i2}, ..., Y_{id}) \in I\!R^{d}$$

where $\xi_i = (\xi_{i1}, \xi_{i2}, ..., \xi_{id}) \in I\!R^d$, $i = m, m+1, ..., n$, are random vector variables; ξ_i are i.i.d.with $E(\xi_{1j}) = 0$, for any $1 \le j \le d$, $E(\xi_{1j}^2) = 1$. The mean vector function f: $I\!R^{md} \to I\!R^d$ and volatility matrix function Σ: $I\!R^{md} \to I\!R^d \times I\!R^d$ are unknown, $\Sigma(x)$ is symmetric and positive definite for any $x \in I\!R^{md}$, and the initial value $X_m = (Y_{m-1}, Y_{m-2}, ..., Y_0)$ is a random vector variable independent of $\{\xi_i\}$.

We call (1.2) a conditional heteroskedastic autoregressive nonlinear (CHARN) model, it neither makes structural assumptions on f and Σ nor distributional assumptions on ξ.

For the CHARN model, it is crucial not only to estimate the conditional mean function $f(\bullet)$, but also the conditional variance or volatility function $\Sigma(\bullet)$ at the same time. As a matter of fact, for the prediction of financial time series, the volatility function plays a more important role than the mean function. It is therefore beneficial to obtain the joint estimation of both $f(\bullet)$ and $\Sigma(\bullet)$ for model (1.2). Härdle, Tsybakov and Yang (1996) studied the estimation of $f(\bullet)$ and $\Sigma(\bullet)$ for the multivariate CHARN model which generalized the result of Härdle and Tsybakov (1996) on asymptotic normality of the local polynomial (LP) estimators. The idea of local polynomial estimation goes back to Stone (1977). The statistical properties of LP estimators in nonparametric regression were studied by Tsybakov (1986). For recent references, we refer to Ruppert and Wand (1994), Wand and Jones (1995), Fan and Gijbels (1996).

For $v, s \in I\!R^d$ that both have unit length and any $x \in I\!R^{md}$, the mean function of $v^T Y$ is $f(x; v) = v^T f(x)$ while the covariance function of $v^T Y$ and $s^T Y$ is $v^T \Sigma(x) s$. In particular, letting v, s be coordinate vectors $e_j = (0, ..., 1, ..., 0)^T$ gives $f_j(x)$ and $\sigma_{jk}(x)$, $j, k = 1, 2, ..., d$, the components of the vector function $f(x)$ and the matrix function $\Sigma(x)$ respectively. The LP method solves the following minimization problems

$$\begin{array}{l} c_n(x; v, s) = \arg\min_{c \in I\!R^{md+1}} \\ \sum_{i=m}^{n} (v^T Y_i Y_i^T s - c^T U_{in})^2 K_h (X_i - x), \\ c_n(x; v) = \arg\min_{c \in I\!R^{md+1}} \\ \sum_{i=m}^{n} (v^T Y_i - c^T U_{in})^2 K_h (X_i - x), \end{array} \tag{1.3}$$

where $K : I\!R^{md} \longrightarrow I\!R^1$ is a kernel $K_h(u) = \frac{1}{h^{md}} K(\frac{u}{h})$, $h = h_n$ is a positive number (bandwidth), $h_n \to 0$, as $n \to \infty$ and

$$U_{in} = F(u_{in}), u_{in} = \frac{X_i - x}{h}, \tag{1.4}$$

where $F(u) = \begin{pmatrix} 1 \\ u \end{pmatrix} \in I\!R^{md+1}$, for $u \in I\!R^{md}$. The estimator of $f(x; v)$ is defined as

$$\hat{f}(x; v) = c_n(x; v)^T F(0).$$

The estimator of $v^T \Sigma(x) s$ is defined as

$$\begin{array}{l} \hat{\sigma}(x; v, s) = c_n(x; v, s)^T F(0) \\ -\{c_n(x; v)^T F(0)\}\{c_n(x; s)^T F(0)\}. \end{array} \tag{1.5}$$

The following is assumed

(A1) The error variables $\xi_{1j}, 1 \le j \le d$, are independent. The density $p(\bullet)$ of ξ_1 exists and satisfies

$$\inf_{x \in \mathcal{K}} p(x) > 0$$

for any compact $\mathcal{K} \subset I\!R^d$. Also $E(\xi_{1j}) = E(\xi_{1j}^3) = 0$, $E(\xi_{1j}^2) = 1$, and $E(\xi_{1j}^4) = 1 + m_4 < \infty$.

(A2) There exist constants $C_1 > 0$, $C_2 > 0$ such that

$$|f(x)| \le C_1(1 + |x|), \tag{1.6}$$
$$\left|\Sigma^{1/2}(x)\right| \le C_2(1 + |x|). \tag{1.7}$$

(A3) The matrix function $\Sigma(x)$ is symmetric for any $x \in I\!R^{md}$, and satisfies

$$\inf_{x \in \mathcal{K}} \lambda_{\min}\{\Sigma(x)\} > \lambda_{\mathcal{K}} > 0,$$

for any compact $\mathcal{K} \subset I\!R^{md}$, where $\lambda_{\min}(\Sigma)$ denotes the minimal eigenvalue of a real symmetric matrix Σ.

(A4) $C_1 + C_2 E|\xi_1| < 1/m$.

(A5) The functions f and Σ are componentwise twice continuously differentiable at the point $x \in I\!R^{md}$.

(A6) The density $\mu(\bullet)$ of the stationary distribution $\pi(\bullet)$ exists, is bounded, continuous and strictly positive in a neighborhood of the point x.

(A7) The kernel K is a compactly supported bounded non-negative function on $I\!R^{md}$, such that

$$\int K(u)du = 1, \int uK(u)du = 0,$$

$$\int uu^T K(u)du = \sigma_K^2 I_{md},$$

where $\sigma_K^2 > 0$, and I_{md} denotes the identity matrix of dimension md.

(A8) $h_n = \beta n^{-1/(4+md)}$, where $\beta > 0$.

(A9) The initial value X_m is a fixed vector in $\mathbb{R}^{md}$.

Under the conditions (A1) to (A4), the Markov chain $\{X_i\}$ is geometrically ergodic, i.e. it is ergodic, with stationary probability measure $\pi(\bullet)$ such that, for almost every x,

$$||P^n(\bullet|x) - \pi(\bullet)||_{TV} = O(\rho^n),$$

for some $0 \leq \rho < 1$. Here

$$P^n(B|x) = P\{X_n \in B | X_m = x\},$$

for a Borel subset $B \subset \mathbb{R}^{md}$, and $||\bullet||_{TV}$ is the total variation distance, see Ango Nze (1992).

Denote $||K||_2^2 = \int K^2(u)du$. Härdle, Tsybakov and Yang (1996) proved the following theorem

Theorem 1 *Under the assumptions (A1) to (A9), as* $n \longrightarrow \infty$

$$n^{\frac{2}{4+md}} \left\{ \begin{array}{c} \widehat{f}_j(x) - f_j(x) \\ \widehat{f}_k(x) - f_k(x) \end{array} \right\} \xrightarrow{\mathcal{D}} \mathcal{N} \left[\left\{ \begin{array}{c} b_j(x) \\ b_k(x) \end{array} \right\}, \left\{ \begin{array}{cc} V_j(x) & c_{jk}(x) \\ c_{jk}(x) & V_k(x) \end{array} \right\} \right] \tag{1.8}$$

with

$$b_j(x) = \beta^2 \frac{\sigma_K^2}{2} \left[Tr\left\{ \nabla^2 f_j(x) \right\} \right]$$

and

$$V_j(x) = \beta^{-md} \frac{\sigma_{jj}(x)}{\mu(x)} ||K||_2^2, c_{jk}(x) = \beta^{-md} \frac{\sigma_{jk}(x)}{\mu(x)} ||K||_2^2.$$

Also, as $n \longrightarrow \infty$

$$n^{\frac{2}{4+md}} \left\{ \begin{array}{c} \widehat{\sigma}_{jk}(x) - \sigma_{jk}(x) \\ \widehat{\sigma}_{j'k'}(x) - \sigma_{j'k'}(x) \end{array} \right\} \xrightarrow{\mathcal{D}} \mathcal{N} \left[\left\{ \begin{array}{c} b_{jk}(x) \\ b_{j'k'}(x) \end{array} \right\}, \left\{ \begin{array}{cc} V_{jk}(x) & c_{jk,j'k'}(x) \\ c_{jk,j'k'}(x) & V_{j'k'}(x) \end{array} \right\} \right] \tag{1.9}$$

with

$$b_{jk}(x) = \beta^2 \frac{\sigma_K^2}{2} \left[Tr\left\{ \nabla^2 \sigma_{jk}(x) + 2\nabla f_j(x) \nabla^T f_k(x) \right\} \right],$$

$$V_{jk}(x) = c_{jk,jk}(x)$$

where

$$c_{jk,j'k'}(x) = \beta^{-md} \frac{||K||_2^2}{\mu(x)} (m_4 - 2) T^*_{jk,j'k'}(x)$$

$$+\beta^{-md} \frac{||K||_2^2}{\mu(x)} \sigma_{jj'}(x)\sigma_{kk'}(x) + \sigma_{jk'}(x)\sigma_{kj'}(x)$$

and

$$T^*_{jk,j'k'}(x) = \sum_{l=1}^{d} s_{jl}(x) s_{j'l}(x) s_{kl}(x) s_{k'l}(x)$$

in which $s_{jl}(x)$ *denotes the* (j,l)*-th entry of the matrix* $\Sigma^{1/2}(x)$. *Finally, as* $n \longrightarrow \infty$

$$n^{\frac{2}{4+md}} \left\{ \begin{array}{c} \widehat{\sigma}_{jk}(x) - \sigma_{jk}(x) \\ \widehat{f}_{j'}(x) - f_{j'}(x) \end{array} \right\} \xrightarrow{\mathcal{D}} \mathcal{N} \left[\left\{ \begin{array}{c} b_{jk}(x) \\ b_{j'}(x) \end{array} \right\}, \left\{ \begin{array}{cc} V_{jk}(x) & 0 \\ 0 & V_{j'}(x) \end{array} \right\} \right] \tag{1.10}$$

2 An Application

As an interesting example, estimates as described in Section 1 were obtained in Härdle, Tsybakov and Yang (1996) for the daily returns of Y_{i1} =DEM/USD (Deutsche Mark/US Dollar) and of Y_{i2} =DEM/GBP (Deutsche Mark/British Pound) for the period of January 2, 1980 to October 30, 1992, a total of 3212 observations.

The estimated conditional mean functions $\widehat{f}_1(x)$ and $\widehat{f}_2(x)$ of the lagged values $x_i = \begin{pmatrix} y_{1,i-1} \\ y_{2,i-1} \end{pmatrix}$ were found to be rather flat and around zero. Also found was a negative correlation when the two returns have opposite lagged values, while positive correlations were found elsewhere. This pattern in which the conditional covariance $\widehat{\sigma}_{12}(x)$ changes from negative to positive tends to resemble the multivariate GARCH Capital Asset Pricing Model as in Bollerslev, Engle and Wooldridge (1988), where the mean functions are also close to zero and therefore negligible. The implications of this to the foreign exchange market is not known to us at this time. We also should point out that effective construction of confidence bands for volatility estimation is yet to become available, and thus our result here is just a starting point. The computation was done in XploRe, using the WARPing technique (Härdle, Klinke, Turlach, 1995). More detailed numerical and graphical descriptions of the data and the various estimates are contained in the paper Härdle, Tsybakov and Yang (1996) cited above.

3 Integration Method

The practical performance of the estimators in Section 1 could suffer from the statistical imprecision introduced

by a large number of lags with smaller sample size, a phenomenon commonly referred to as the "curse of dimensionality". Stone (1982) showed for i.i.d. regression that if the mean function is a sum of univariate functions, then the one dimensional convergence rate can be achieved for its estimation, thus avoiding the "curse of dimensionality". Chen and Tsay (1993a,b) studied additive time series models using the BRUTO algorithm developed by Hastie and Tibshirani (1990). The "integration method" was introduced by Auestad and Tjøstheim (1991) and further explored by Tjøstheim and Auestad (1994a). It provides closed form bias and variance expressions of the one dimensional function estimator. It has been employed recently in the autoregression setting by Masry and Tjøstheim (1995a,b), and in the i.i.d regression setting by Linton and Härdle (1996), Linton and Nielsen (1995). The volatility function measures the scale and is always positive, thus it is more appropriate to model its changes multiplicatively rather than additively, as in the EGARCH model of Nelson (1991).

Consider, therefore, a CHARN model of the form

$$Y_i = f(Y_{i-1}, Y_{i-2}, ..., Y_{i-d}) + s(Y_{i-1}, Y_{i-2}, ..., Y_{i-d})\xi_i \tag{3.1}$$

where $\{\xi_i\}_{i>1}$ are as in Section 1, $Y_0, Y_1, ..., Y_{d-1}$ are random variables independent of the $\{\xi_i\}$'s and

$$f(Y_{i-1}, Y_{i-2}, ..., Y_{i-d}) = c_f + \sum_{\beta=1}^{d} f_\beta(Y_{i-\beta}), \tag{3.2}$$

$$\begin{aligned} &\sigma(Y_{i-1}, Y_{i-2}, ..., Y_{i-d}) = \\ &s(Y_{i-1}, Y_{i-2}, ..., Y_{i-d})^2 = c_\sigma \textstyle\prod_{\beta=1}^{d} \sigma_\beta(Y_{i-\beta}) \end{aligned} \tag{3.3}$$

where c_f and c_σ are constants, $\{f_\beta(\bullet)\}_{\beta=1}^{d}$ and $\{\sigma_\beta(\bullet)\}_{\beta=1}^{d}$ are sets of unknown functions satisfying certain identifiability conditions. Under assumptions similar to those in Section 1, Yang and Härdle (1996) proposed a set of estimators $\left\{\widehat{f}_\alpha(\bullet)\right\}$ and $\{\widehat{\sigma}_\alpha(\bullet)\}$ by integrating local polynomial estimators of degree $p > 0$, and proved the following

Theorem 2 *For any* $\mathbf{x} = (x_1, ..., x_d)$ *and any* $\alpha = 1, ..., d$

$$n^{\frac{p+1}{2p+3}} \left\{\widehat{f}_\alpha(x_\alpha) - f_\alpha(x_\alpha)\right\} \xrightarrow{D} N\left\{b_{f\alpha}(x_\alpha), v_{f\alpha}(x_\alpha)\right\} \tag{3.4}$$

$$n^{\frac{p+1}{2p+3}} \left\{\widehat{\sigma}_\alpha(x_\alpha) - \sigma_\alpha(x_\alpha)\right\} \xrightarrow{D} N\left\{b_{\sigma\alpha}(x_\alpha), v_{\sigma\alpha}(x_\alpha)\right\} \tag{3.5}$$

While for any $\alpha \neq \beta$ *one has*

$$cov\left[n^{\frac{p+1}{2p+3}} \left\{\widehat{f}_\alpha(x_\alpha) - f_\alpha(x_\alpha)\right\}, n^{\frac{p+1}{2p+3}} \left\{\widehat{f}_\beta(x_\beta) - f_\beta(x_\beta)\right\}\right] \to 0. \tag{3.6}$$

$$cov\left[n^{\frac{p+1}{2p+3}} \left\{\widehat{\sigma}_\alpha(x_\alpha) - \sigma_\alpha(x_\alpha)\right\}, n^{\frac{p+1}{2p+3}} \left\{\widehat{f}_\alpha(x_\alpha) - f_\alpha(x_\alpha)\right\}\right] \to c_{\sigma\alpha}(x_\alpha) \tag{3.7}$$

$$cov\left[n^{\frac{p+1}{2p+3}} \left\{\widehat{\sigma}_\alpha(x_\alpha) - \sigma_\alpha(x_\alpha)\right\}, n^{\frac{p+1}{2p+3}} \left\{\widehat{\sigma}_\beta(x_\beta) - \sigma_\beta(x_\beta))\right\}\right] \to 0 \tag{3.8}$$

Furthermore

$$n^{\frac{p+1}{2p+3}} \left\{\widehat{f}(\mathbf{x}) - f(\mathbf{x})\right\} \xrightarrow{D} N\left\{b_f(\mathbf{x}), v_f(\mathbf{x})\right\} \tag{3.9}$$

$$n^{\frac{p+1}{2p+3}} \left\{\widehat{\sigma}(\mathbf{x}) - \sigma(\mathbf{x})\right\} \xrightarrow{D} N\left\{b_\sigma(\mathbf{x}), v_\sigma(\mathbf{x})\right\} \tag{3.10}$$

The definitions of the functions $b_{f\alpha}(x_\alpha)$, $v_{f\alpha}(x_\alpha)$, $b_{\sigma\alpha}(x_\alpha)$, $v_{\sigma\alpha}(x_\alpha)$, $c_{\sigma\alpha}(x_\alpha)$, $b_f(\mathbf{x})$, $v_f(\mathbf{x})$, $b_\sigma(\mathbf{x})$, $v_\sigma(\mathbf{x})$, are given in Yang and Härdle (1996).

4 Lag Selection

Useful estimation of autoregressive (AR) time series requires the selection of correct lagged values. While many lag selection methods have been developed for stationary AR models, they all assume that the series itself is linear. Tjøstheim and Auestad (1994b) and Vieu (1994) had formulated nonparametric lag selection rules, using kernel based estimate of the Final Prediction Error (FPE) and the Cross Validation as selection criteria respectively. See also Chen and Tsay (1993a) for the best subset procedure.

While the application of nonparametric autoregression procedure needs a good selection of lagged variables, it also provides a tool for such selection that is easy to use. Applying the results on local polynomial autoregression as in Section 1, Tschernig and Yang (1996) have studied the nonparametric FPE estimate for a given finite lag AR process, using a local linear (LL) estimator with bandwidth h. It has the following expression

$$FPE = A + \frac{\|K\|_2^{2j}}{nh^j} B + \sigma_K^4 \frac{h^4}{4} C \tag{4.1}$$

where

$$A = \int \sigma(x) w^2(x) \mu(x) dx, B = \int \sigma(x) w^2(x) dx,$$

$$C = \int \left[\mathrm{Tr}\left\{\nabla^2 f(x)\right\}\right]^2 w^2(x) \mu(x) dx.$$

and where j is the total number of correct lags, $w(\bullet)$ a weight function, all other functions are the same as Section 1 except that now the series is one dimensional. The

bandwidth h is then set to the optimal h_{opt} by minimizing the FPE, thus

$$h_{opt} = \left(\frac{j \|K\|_2^{2j}}{nC\sigma_K^4} B \right)^{\frac{1}{j+4}} \tag{4.2}$$

The FPE of Tjøstheim and Auestad, FPETA, has essentially the same expression as that in (4.1) above, but by making $h = o(n^{-\frac{1}{j+4}})$, it throws out the term $\sigma_K^4 \frac{h^4}{4} C$, which measures some kind of curvature of the process. For smooth (here by smooth we mean second order continuously differentiable) processes that are highly nonlinear, this causes problem, as shown in the Monte Carlo study of Tschernig and Yang (1996), where the restriction on the overfitting tendency was insufficient. Another drawback of FPETA is that it used the Nadaraya-Watson (NW) instead of LL estimator, thus even if the term $\sigma_K^4 \frac{h^4}{4} C$ were included, it would involve not only $\mathrm{Tr}\nabla^2 f(x)$ and $\mu(x)$, but also $\nabla f(x)$ and $\nabla \mu(x)$, which would add extra difficulty.

Tschernig and Yang (1996) have shown that the lag selection rule based on the above FPE is consistent as $n \to \infty$, while the probability of overfitting (i.e., selecting all the correct lags plus some extra ones) goes to zero at a slower rate than that of underfitting (i.e., missing some correct lags). Therefore a new modified FPE (FPETY) is proposed

$$FPETY = \left(A + \frac{\|K\|_2^{2j}}{nh_{opt}^j} B + \sigma_K^4 \frac{h_{opt}^4}{4} C \right) \left(1 + \frac{j}{n^{\frac{4}{j+4}}} \right) \tag{4.3}$$

which possesses all the properties of the previous one in (4.1), but with some extra protection against overfitting, by multiplying the factor $1 + \frac{j}{n^{\frac{4}{j+4}}}$ which penalizes larger models. The Monte Carlo study of Tschernig and Yang (1996) has shown that this new FPETY outperforms the FPETA when the process is smooth, while for discontinous or nonsmooth processes, it performs somewhat worse than the FPETA. In most cases, the FPETY performs better than the AIC criteria of Tjøstheim and Auestad.

References

Ango Nze P. (1992) *Critères d'ergodicité de quelques modèles à représentation markovienne*, C.R. Acad. Sci. Paris, **315**, sér 1, 1301-1304.

Auestad, B. and Tjøstheim, D. (1991) Functional identification in nonlinear time series. In Nonparametric Functional Estimation and Related Topics, Ed. Roussas, G.G Amsterdam: Kluwer Academic Publishers, 493-507.

Bollerslev, T.; Engle, R.; Wooldridge, J. (1988) *A Capital Asset Pricing Model with Time-varying Covariances* Journal of Political Economy, vol.96, no.1.

Chan, K.S.; Tong, H. (1986) *On estimating thresholds in autoregressive models*, Journal of Time Series Analysis, **7**, 179-190

Chen, R.; Tsay, R.S. (1993a) *Nonlinear additive ARX models*, Journal of the American Statistical Association, **88**, 955-967.

Chen, R.; Tsay, R.S. (1993b) *Functional-coefficient autoregressive models*, Journal of the American Statistical Association, **88**, 298-308.

Collomb, G. (1984) *Propriétés de convergence presque complète du prédicteur à noyau.* Zeitschrift für Wahrscheinlichkeitstheorie und verwandte Gebiete, 66, 441-460.

Engle, R.F. (1982) *Autoregressive conditional heteroscedasticity with estimates of the variance of U.K. inflation*, Econometrica, **50**, 987-1008.

Engle, R.F.; Gonzalez-Rivera, G. (1991) *Semiparametric ARCH Models*, Journal of Business and Economic Statistics, **9**, 345-360.

Engle, R.F.; Ng, V. (1993) *Measuring and testing the impact of news on volatility*, Journal of Finance, **48**, 1749-1778.

Fan, J., Gijbels, I. (1996) *Local Polynomial Modelling and its Applications*, Chapman and Hall.

Granger, C.; Teräsvirta, T. (1993) *Modelling Nonlinear Dynamic Relationships*, Oxford University Press, Oxford

Haggan, V.; Ozaki, T. (1981) *Modelling nonlinear vibrations using an amplitude-dependent autoregressive time series model*, Biometrika, **68**, 189–196.

Gouriéroux, Ch.; Monfort, A. (1992) *Qualitative threshold ARCH models*, Journal of Econometrics 52, 159-199.

Granger, C.; Teräsvirta, T. (1993) *Modelling Nonlinear Dynamic Relationships*, Oxford University Press, Oxford

Gregory, A.W. (1989) *A Nonparametric Test for Autoregressive Conditional Heteroscedasticity: A Markov Chain Approach*, Journal of Business and Economic Statistics, 7, 107-115.

Haggan, V.; Ozaki, T. (1981) *Modelling nonlinear vibrations using an amplitude-dependent autoregressive time series model*, Biometrika **68**, 189–196.

Härdle, W.; Klinke, S.; Turlach, B. (1995) *XploRe - an interactive statistical computing environment*, Springer Verlag, Heidelberg.

Härdle, W.; Tsybakov, A.B. (1996) *Local polynomial estimators of the volatility function in nonparametric autoregression* , to appear in Journal of Econometrics.

Härdle, W.; Tsybakov, A.B.; Yang, L. (1996) *Nonparametric vector autoregression*, to appear in Journal of Statistical Planning and Inference.

Hastie, T. J.; Tibshirani, R. J. (1990) *Generalized Additive Models*, Monographs on Statistics and Applied Probability, 43, Chapman and Hall, London.

Linton, O. B.; Härdle, W. (1996) *Estimation of additive regression models with known links*, to appear in Biometrika.

Linton, O. and Nielsen, J.P. (1995) *A kernel method of estimating structured nonparametric regression based on marginal integration*, Biometrika, **82**, 93-100.

Lütkepohl, H. (1991) *Introduction to Multiple Time Series Analysis*, Springer-Verlag, Heidelberg.

Masry, E. and Tjøstheim, D. (1995a) *Nonparametric estimation and identification of ARCH nonlinear time series: Strong convergence and asymptotic normality*, Econometric Theory, **11**, 258-289.

Masry, E. and Tjøstheim, D. (1995b) *Additive nonlinear ARX time series and projection estimates*, to appear in Econometric Theory.

McKeague, I.W.; Zhang, M.J. (1994) *Identification of nonlinear time series from first order cumulative characteristics*, Annals of Statistics 22, 495-514.

Nelson, D.B. (1991) *Conditional heteroskedasticity in asset returns: A new approach*, Econometrica, **59**, 347-370.

Robinson, P.M. (1983) *Nonparametric Estimators for Time Series*, Journal of Time Series Analysis, **4**, 185-207.

Robinson, P.M. (1984) *Robust Nonparametric Autoregression*, In: *Robust and Nonlinear Time Series Analysis*, eds. Franke, Härdle and Martin, Lecture Notes in Statistics, 26, Springer-Verlag, Heidelberg.

Ruppert, D.; Wand, M.P. (1994) *Multivariate Locally Weighted Least Squares Regression*, Annals of Statistics, **22**, 1346-1370.

Stone, C.J. (1977) *Consistent Nonparametric Regression*, Annals of Statistics, **5**, 595-645.

Stone, C.J. (1982) *Optimal Global Rates of Convergence for Nonparametric Regression*, Annals of Statistics, **10**, 1040-1053.

Teräsvirta, T. (1994) *Specification, Estimation and Evaluation of Smooth Transition Autoregressive Models.*, Journal of the American Statistical Association, **89**, 208-218.

Tjøstheim, D. (1994) *Non-linear Time Series Analysis: A Selective Review*, Scandinavian Journal of Statistics, **21**, 97-130.

Tjøstheim, D. and Auestad, B. (1994a) *Nonparametric Identification of Nonlinear Time Series: Projections*, Journal of the American Statistical Association, **89**, 1398-1409.

Tjøstheim, D. and Auestad, B. (1994b) *Nonparametric Identification of Nonlinear Time Series: Selecting Significant Lags*, Journal of the American Statistical Association, **89**, 1410-1419.

Tong, H. (1978) *On a threshold model*, in C. H. Chen (ed.), *Pattern Recognition and Signal Processing*, Sijthoff and Noordholf, The Netherlands.

Tong, H. (1983) *Threshold Models in Nonlinear Time Series Analysis*, Vol. 21 of Lecture Notes in Statistics, Springer-Verlag, Heidelberg.

Tschernig, R.; Yang, L. (1996) *Nonparametric Lag Selection for Time Series*, in preparation.

Tsybakov, A.B. (1986) *Robust Reconstruction of Functions by the Local Approximation Method*, Problems of Information Transmission, **22**, 133-146.

Vieu, P. (1994) *Order Choice in Nonlinear Autoregressive Models*, Statistics **24**, 1-22.

Wand, M.P.; Jones, M.C. (1995) *Kernel Smoothing*, Chapman and Hall, London.

Yang, L.; Härdle, W. (1996)
Nonparametric Autoregression with Multiplicative Volatility and Additive Mean, submitted to the Journal of Time Series Analysis.

Finanzmarktanalyse und -prognose, 53-56, Physika Verlag

Financial Calculations on the Net

Wolfgang Härdle and Stefan Sperlich

Institut für Statistik und Ökonometrie, Wirtschaftswissenschaftliche Fakultät, Humboldt-Universität zu Berlin, D-10178 Berlin, Germany

1 Introduction to XploRe

XploRe is an interactive statistical computing environment. As the name indicates, support for exploratory data analysis, for simulation and intensive calculation scenarios is given by a variety of interactive tools. Like most of the statistic software packages it is based on vector- and matrix calculation, but unlikely many such packages, in XploRe matrices can be of up to nine dimensions. Further, XploRe offers extraordinary opportunities in dynamic and interactive graphics. For our purpose, that is calculating in finance, the interactivity of the user interface and the techniques of visualization are of special importance; particulary since this interactivity is net based and easy to implement for programmers who want to modify or extend existing methods.

A noteworthy quality in XploRe are the capabilities of network facilities, which in software development have become more and more important. From the very beginning, XploRe was geared towards inter- or intranet compatability. Thus, for example, the help system is available in HTML. Furthermore, the newly developed Java interface endows the user with the possibility to work with XploRe via inter- or intranet and thus without the necessity of installing it locally. Thus, a complete session can be opened just with a freely available WWW browser like Netscape (try the http address: *http://www.xplore_stat_de/*).

2 The `finance` library of XploRe

There is growing interest in quantifying and simulating economic processes, particulary in the statistical analysis of the behaviour of financial markets. The library `finance` is designed for this purpose. It offers macros to predict and estimate (time series) processes, for example stock returns, to determine option prices and to evaluate different scenarios (e. g. for portfolio strategies). To give a brief overview of this library we will present and illustrate some macros implemented in `finance`. It is called by `library("finance")` on the free Java interface of XploRe. We will pay special attention to the simulation of processes of stock returns.

(1998) Härdle, W. and Sperlich, S. Financial Calculations on the Net.

Finanzmarktanalyse und -prognose, 53-56, Physika Verlag

The basic method to determine option prices of the European kind is to use the analytic solution of Black and Scholes. Their formula allows to calculate the option price under strong assumptions on the model. The analogue for American options is the approximation by McMillan. Distributions of dividends for assets like stocks usually lead to changes in the value of the corresponding derivative. Unfortunately this often can not be taken into account in the analytic solutions and thus has to be done numerically for European as well as for American options. Therefore binomial trees are used. These thechniques can be performed interactively via the macro `optstart()`, as presented in figure 1 for illustration. It is also of interest to investigate and

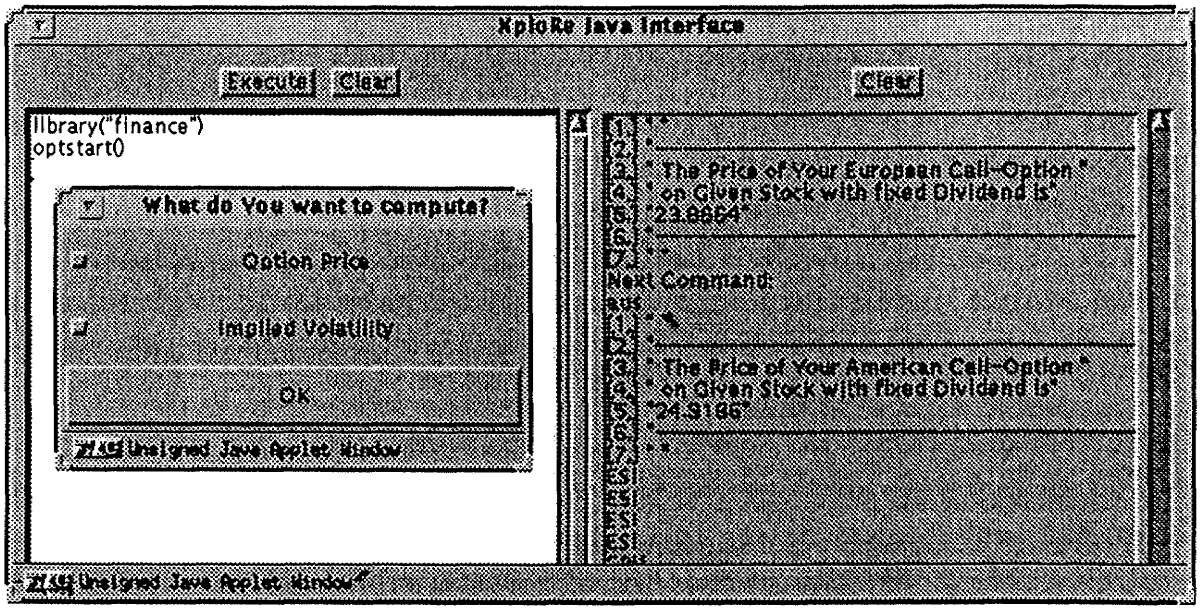


Fig. 1. The macro `optstart` running on Java

to visualize the influence of the various factors such as domestic interest rate, time to expiration on option prices. This can be done by the macro `influence()`. Figure 2 displays the surface of option prices versus exercise price and time to expiration while keeping all the other parameters fixed.

Obviously, scenarios for portfolios strategies like the construction of spreads or arbitrage possibilities can easily be obtained by table calculation. It has long be recognized that stock returns do not obey a simple Wiener process. Various alternatives have been introduced, including among others, jump processes or a mixture of Wiener and jump processes, see for example Streller (1995). Macros for simulating such processes and for estimating the parameters in those models are also available, e. g. `stocksim()`, `stockest()`.

A different approach to model stock prices is using typical time series models such as ARCH, GARCH, EGARCH, T-ARCH etc. , see Gourieroux (1997).

Finanzmarktanalyse und -prognose, 53-56, Physika Verlag

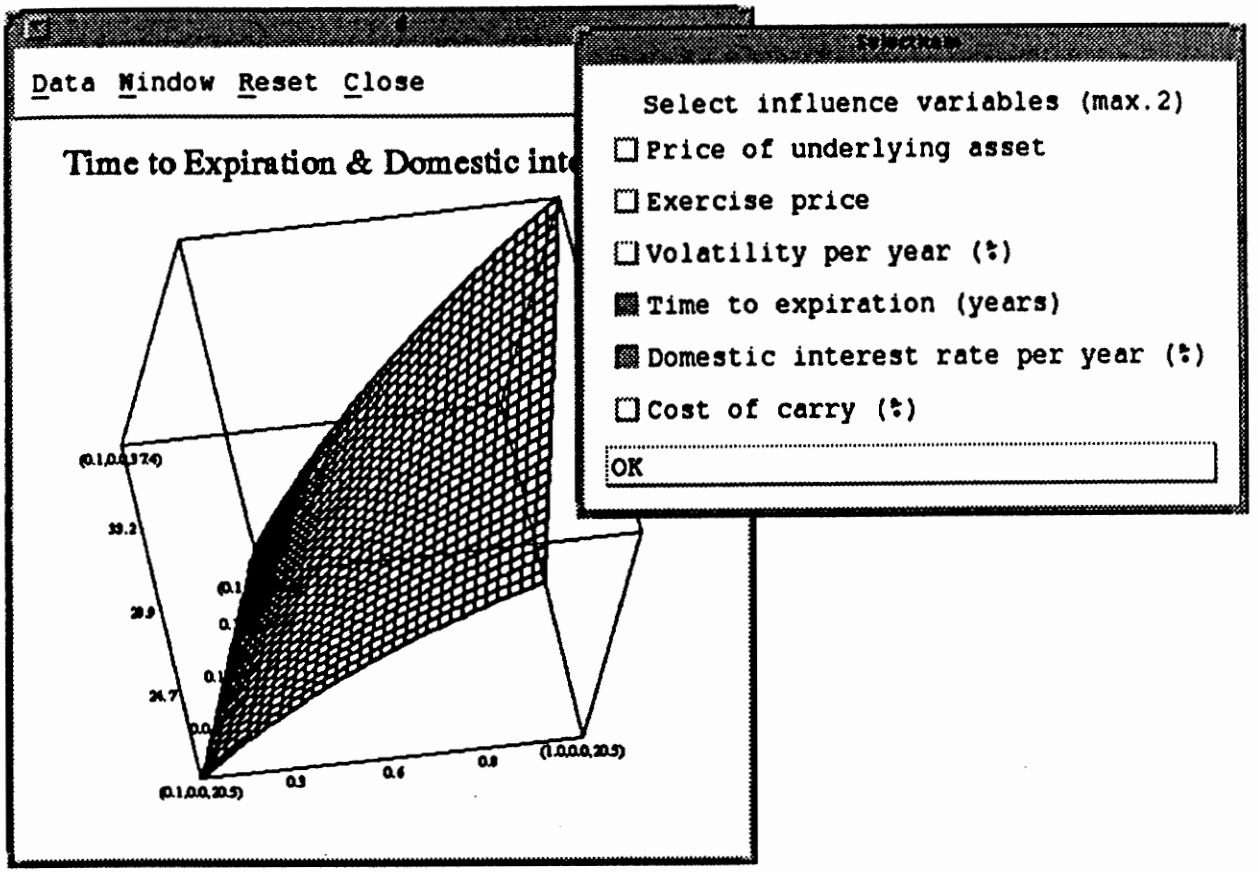


Fig. 2. The macro influence in action

The class of ARCH models, built upon the concept of stochastic volatility, is much more flexible than classic models in fitting financial data. In particular, the time discrete model of Duan (1995) using the GARCH(1,1) process is appealing for the theory of option pricing.

A crucial question is the correct specification of the stochastic volatility. Therefore it is necessary to develop programs to model the volatility more flexible, e. g. , for asymmetry of shocks. Examples of such a development are the Threshold ARCH model of Rabemananjara and Zakoian (1993) and the extension of it to T-GARCH by Härdle and Hafner (1997). The program to analyse and compare these models is implemented in `tgarsim()`.

In Figure 3 we present results of a comparison of GARCH and T-GARCH to the Black and Scholes option pricing model. Displayed are the generated processes, option prices, absolute and relative differences to Black and Scholes versus moneyness (S/K with S=stock price, K= strike price). The greyish curve indicates the T-GARCH model.

All programs are available on the net.

(1998) Härdle, W. and Sperlich, S. Financial Calculations on the Net.

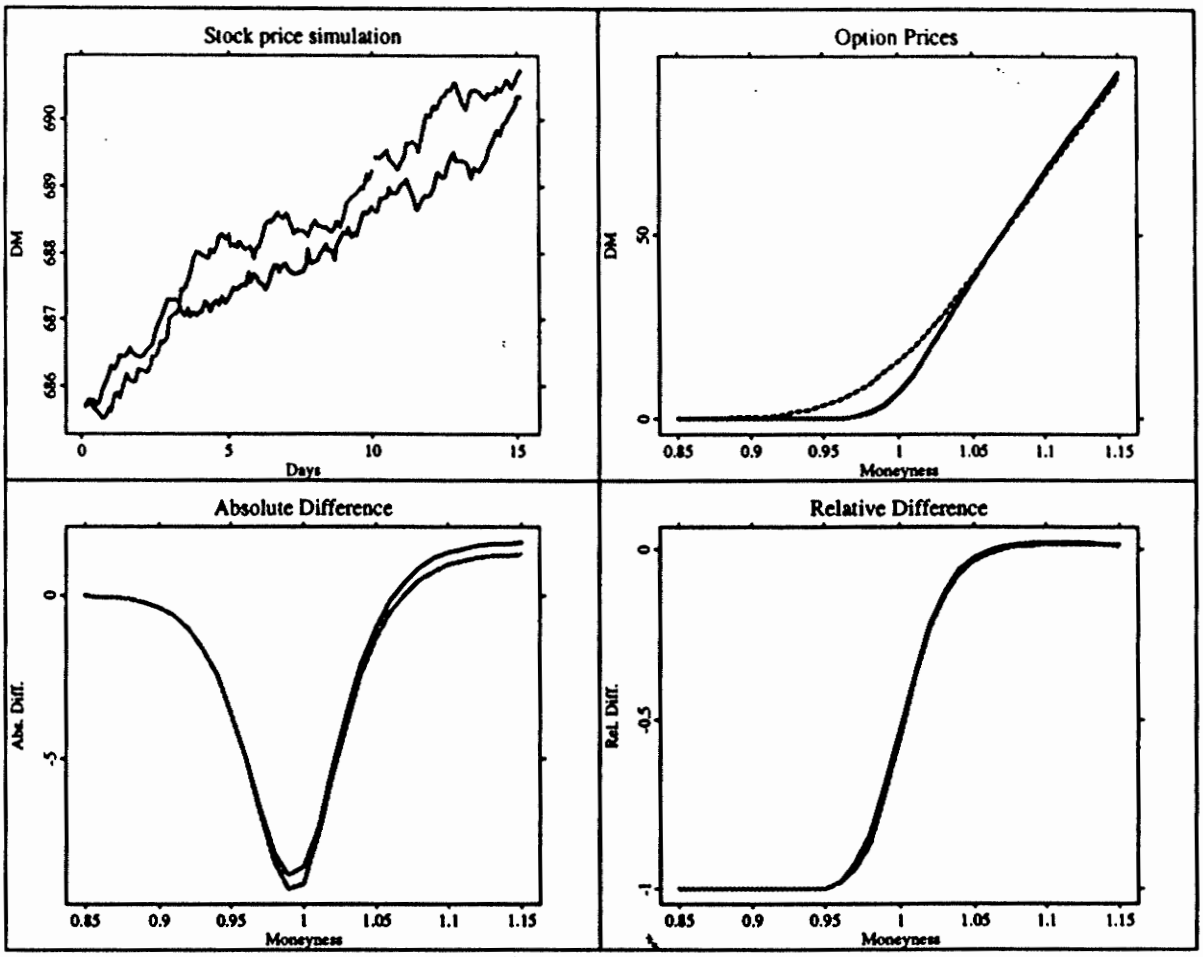


Fig. 3. The result of the macro `tgarsim`

References

Duan, J. C.: The GARCH option pricing model, Mathematical Finance. **5** (1995) 13–32

Gourieroux, C.: ARCH Models and Financial Applications, Springer Verlag, Heidelberg & Berlin. (1997)

Härdle, W. , Hafner, C.: Discrete Time Option Pricing with Flexible Volatility Estimation, forthcoming. (1997)

Rabemananjara, R. , Zakoian, J. M.: Threshold ARCH Models and Asymmetries in Volatility, Journal of Applied Econometrics. **8** (1993) 31–49

Streller, A.: Modellierung von Aktienkursen durch Diffusionen mit Sprüngen – Modelldiskussion und ein Weg zur Schätzung der Parameter, Discussion Paper 56., sfb 373. (1995)

TEACHING AND THE INTERNET

Connected Teaching of Statistics

By Wolfgang Härdle, Sigbert Klinke, and J. S. Marron

1. Introduction

The study of statistics is commonly considered difficult by students, since it requires a variety of skills including quantitative and graphical insights as well as mathematical ability. Yet an increasing number of people need facility with quantitative methods and students need to acquire statistical capabilities because they are confronted with more and more data sets to be understood. In addition these data sets grow rapidly in size and structural complexity. An example for such data are the files that are collected on mobile phone applications, transaction records, etc. Despite these changing needs the teaching methods used have been surprisingly constant in recent years. An attractive and potentially powerful new way of updating current teaching methodology is via tools based on an intra- or the Internet. In this article we suggest a set of criteria for effective web based teaching and propose the first net based approach to meet these criteria.

New technology is accepted more widely if its use is immediately understandable and easy for everybody. The same is true for teaching statistics in face of the new challenges in structure and size of data. It can only be effective if statistical methods are explained in a way that gives the student easy access to them. Two viewpoints are important for understanding this effective teaching, that of the student and that of the teacher. The new additional component of web based computing in teaching has an impact on both viewpoints. For example, large data sets, interactive graphics and on-line information were unusable for undergraduate statistics teaching a few years ago. Now easy availability of these features requires an update of the criteria behind "what is good teaching?" from both viewpoints.

The student will benefit from

- Quick and easy access to methodology and data via browsers
- Interactive examples since doing is one of the fastest methods of learning
- Smooth transition from classroom to homework to full scale statistical tools

The teacher will benefit from

- Quick and easy broadcast of methodology and data via browsers
- A user friendly environment
- A powerful and flexible environment for dissemination of research

Many current teachers of statistics have a lot of inertia and reservation against changing teaching practice. Hence a stepwise plan towards smooth integration of web based teaching elements will gain the largest following. A series of steps through levels which allows gradual involvement and time commitment is:

Level 1: Display off the shelf class examples via a web browser. This requires only standard display equipment, and minimal effort by the instructor assuming ready made examples are available.

Level 2: Do examples as in Level 1, live on the web and give interested students the link coordinates of exercises, data or further programs and suggest some enriching examples they try on their own. This requires web access for most students and in the classroom. Again the ready made examples can be used and student questions will be minimal because no requirement is made of less capable students.

Level 3: Do examples as in Level 2, but assign homework using web based examples and methodology. This requires web access for all students and much more instructional support to address the inevitable questions and problems.

Level 4: Become a developer of examples. This involves more time and energy on the part of the teacher (the amount depends on the friendliness of the environment and on the integrability of other web based documents), but also yields the most rewards in terms of the customizability that more creative teachers will want. Our goal for teachers who reach this stage is to provide tools which will maximize their individual creativity.

2. The solution

Our approach to meet the above criteria for teaching statistics in elementary courses is based on macros written in XploRe (`www.XploRe-stat.de`). Some examples are discussed in detail in Section 3. Here we develop the general framework and present the various outlets of XploRe for different platforms. XploRe is the interactive statistical computing environment which works equally well on single user machines, intranetworks and web connected clients. This is technically

Härdle, W., Klinke, S. and Marron, J. (1999) Connected Teaching Of Statistics

available via XploRe's server-client concept. The server (professor's machine) makes the course documents (data and statistical methods) available. The client (classroom machine or students PC) connects to the server via the web without additional software downloads. The Java technology (`www.javasoft.com`) and standard web browsers enable this universal access despite the well known heterogeneity across hardware platforms. The overhead of earlier methods of handing out software and data sets is thus dramatically reduced.

The quick and easy access to methodology and data via browsers comes from the good integrability of XploRe data, macros and tutorials into web documents. Since most students are familiar with using browsers on the Internet, there will be no overhead of learning the environment, which would otherwise distract from their learning the desired statistical lessons.

A set of interactive examples is discussed in Section 3 below. These are intended to illustrate the point that interactive learning is very effective and all based on a standard browser front end (with a Java Swing class from SUN). For example, when a student is involved in choosing numerical parameters for a particular case study, the level of thought needed, followed by anticipation of the answer, which is then immediately displayed, results in a deep type of learning. In particular, doing is the best method of learning.

Some of the earlier approaches (e.g. `www.stat.sc.edu/~west/webstat/`, `www.stat.berkeley.edu/users/stark/Java/` and `www.ruf.rice.edu/~lane/stat_sim/`) to web based teaching of statistics have included a smooth transition from classroom to homework by allowing the student to use the Java applet shown in class also for homework. A natural next step to producing truly quantitatively equipped students is to also provide a smooth transition to full scale statistical tools that will continue to be useful long after the class is over. Because our examples are based on the statistical computing environment XploRe it is simple to move from classroom examples to more elaborate data analysis.

Traditional methods of conveying data to students, such as writing on a chalkboard or piece of paper, have severe limitations, due to the effort involved at both ends of the process. Exchange of floppy disks allows software, and also larger data sets to be conveyed, but this involves a lot of overhead in terms of effort (e.g. control of homogeneity of hardware platforms) on the part of the teacher. The Internet clearly allows quick and easy broadcast of methodology and data via browsers.

Many teachers of statistics have not learned web development skills, and perhaps may not have even learned other types of computational skills. For such potential users a user friendly environment means class examples must be already completely developed and ready to use. We offer classroom ready examples on (`ise.wiwi.hu-berlin.de/statistik/lehrmaterial/statmat.html`).

Other teachers of statistics will be higher end users, who have their own ideas for class examples, or else would like to customize those that are provided. For them a user friendly environment means the existing examples are coded in a very high level language, which is easy to modify, and provides a convenient basis for other types of development. XploRe is matrix (array) based and thus development occurs at a higher level than is available from Java programming. An important advantage of XploRe over other high level statistical languages such as SPLUS (`www.mathsoft.com/Splus`), GAUSS (`www.aptech.com`) or STATA (`www.stata.com`) is that XploRe macros may be automatically converted to web transparent methodology via an HTML translator.

Teachers who wish to modify the given examples, or develop their own, will need more than just a user friendly environment. They will also need a powerful and flexible environment, which contains a wide range of quickly usable tools. XploRe has a wide range of statistical tools with the possibility of specialization for different fields like finance, econometrics, etc. Java based approaches to specializing software for teaching cannot provide this full scale since they are based on combinations of the limited set of fixed applets available in the toolbox provided by the applets' constructor.

Statistical Technology on the net

Three hardware platforms are in widespread use for statistical computing and graphical data interaction: Macintosh, UNIX, and Windows. The first has a simple graphically oriented user interface and allows highly interactive dialogues with data. UNIX is used for high-speed and distributed computing but is often less satisfactory in graphical interaction. Windows aims at facilitating both high-speed computing and graphics but is weaker at present than UNIX for Internet access. Distributed computing is simply not possible under Windows unless one uses certain add-ons. An overview of current Internet technology and statistics is given by Symanzik (1998) (`www.galaxy.gmu.edu/~symanzik`)

Many software platforms for statistical computing exist but are unfortunately not easily interchangeable. The

reasons for this include the history of software development, the targeted user groups, and the optimization of certain software for specific hardware configurations. The original version of GAUSS (`www.aptech.com`), for example, was optimized for INTEL chips and, therefore, could not be transferred to the Mac or UNIX platforms. Now GAUSS is available on UNIX, but the UNIX version does not have a graphical device that allows, for e.g. interactive changes in the layout of graphs. SPLUS (`www.mathsoft.com/Splus`) was developed for UNIX systems and was only later transferred to PCs. Consequently, the PC version is different from the UNIX version. EVIEWS was developed for DOS and is now available for Windows but not for UNIX or for the Mac. TSP is a DOS program and is not easily transferred to a Windows/NT platform. SPSS exists for Windows but has still a batch structure that makes many mouse clicks necessary in order to generate implicitly the batch commands. STATA (`www.stata.com`), SAS (`www.sas.com`) and SHAZAM (`shazam.econ.ubc.ca`) are unusual in that mutually compatible versions exist for all platforms. Besides the software that we mentioned here as examples, there are many other platforms which also share the property of heterogeneity.

Heterogeneity of software platforms creates relatively few problems if there is no need to exchange programs. Exchange of graphs, document files, and ASCII-based data sets can be carried out by FTP, provided that the user has the appropriate graphics plug-in and document reader (e.g., Ghostscript or Acrobat). However, there is also a need for exchangeable computer programs for implementing advanced statistical methods, as these are becoming increasingly complex mathematically, and writing the necessary programs can be a difficult and time-consuming task, which puts it effectively out of reach in many cases.

Graduate-level instruction in statistics provides one example of the usefulness of exchangeability. It is not unusual for a faculty member at one university to give a short course at another. In some cases, a faculty member at one university may use electronic communication to present a course at several geographically dispersed locations. Calculation of an estimator may require heavy computing that is available on the researcher's home machine. During the course, modifications of this estimator and different applications may be discussed, and these may require access to the software at multiple locations. Exchangeability of software is necessary to enable students at all locations to carry out computational and empirical exercises that the instructor has prepared at his own university.

Collaboration among researchers at different locations provides another example of the desirability of exchangeability. In this case, the goal is to enable each collaborator to carry out computations using the same software. Ideally such cooperation should be based on a pool of easily accessible software and computing power for all parties. For effective progress on a project that involves heterogeneous hardware and software, it is desirable for partners to have the ability to contribute methods despite being at different locations and working with different computing environments. In addition it may simply be a problem for a researcher who is a visitor in another establishment to be able to continue using his own programs.

On the other hand, heterogeneity of software does have the important advantage of enabling a developer of new methods to choose the software system that is best suited to the problem under consideration. Therefore the problem of exchangeability should not be solved by standardizing statistical software but by making software from different sources accessible to diversely equipped users.

3. Teachware Quantlet Examples

Example 1

This example illustrates that gathering a random sample is different from "just choosing some", i.e. proper random sampling requires some mechanism to ensure that "all samples are equally likely," which is quite different from "arbitrary human choice."

This is intended for a classroom setting, where students are asked to write down (to avoid changing during the course of the exercise) a "randomly chosen" number among 1, 2, 3, 4. Since most people choose 3, and most of the rest choose 2, the resulting distribution is quite far from the random uniform distribution.

Nonrandomness of the chosen numbers is demonstrated "on line" by entering numbers (the actual counts for each of 1,2,3,4) into the textbox on the left. The numbers currently shown come from an actual class. Clicking "OK" generates the result on the right, which is a bar graph showing the counts (for easy visual interpretation), together with a text window summarizing the results of some simple statistical analysis, including a confidence interval for the proportion of 2's and 3's. While confidence intervals have likely not been explained at this point in the course, it can simply be said "this range gives a feeling for the variability in the data, and it will contain the given proportion if these numbers are actually random." This provides motivation and

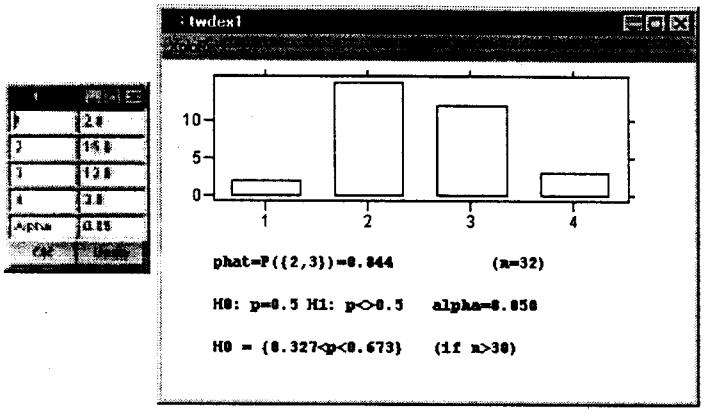


Figure 1: Demonstrates Example 1, "random" is different from "just some." Left: Menu to enter the number of people who have chosen 1, 2, 3 or 4 and the confidence level. Right: Resulting window with graphical and text output, which assesses the amount of "randomness" of the entered numbers.

interest for the time when confidence intervals and hypothesis tests are developed (when this example should be revisited).

For level 2 or level 3 teaching, students should be encouraged to experiment with changing the input values, and watching the change in the interval bracketing 0.5. For example, what happens for (25, 25, 25, 25)? For (0, 0, 100, 0)? What is the difference between (10, 70, 0, 20) and (10, 0, 70, 20)? Students could be challenged to "explore the boundary between random and not" by finding data vectors which are near each other, but give opposite test results. This example may be repeated as many times as desired and may be run directly from the XploRe web site, `www.XploRe-stat.de`. One opens the Java 1.1 interface (Swing classes have to be in the corresponding Java directory), enters `library("tware")` and then enters the quantlet name `twrandomsample()`.

Example 2

This example is intended to illustrate the concept of a p-value for hypothesis testing. For simplicity, it is done in the context of the Binomial(n,p) distribution. The hypothesis tested is: $H_0 : p < c$, for some choice of c.

The example starts with a menu of input boxes, which allows input of:

- The binomial parameter, n (number of Bernoulli trials),
- The binomial parameter, p (probability of success in the Bernoulli trials),
- The observed binomial value, x.

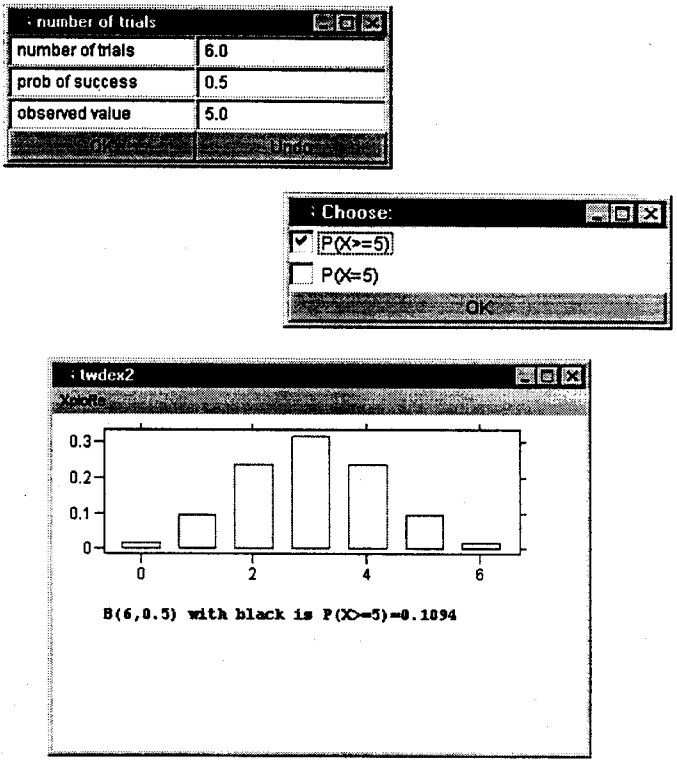


Figure 2: Demonstrates Example 2, how p-values work. Upper: Menu to choose the Binomial distribution parameters: the number of trials, n, the probability of success, p, and the observed value, x. Middle: menu allows choice between $P(X \geq x)$ or $P(X = x)$. Lower resulting plot under Java with area representing the p-value, $P(X \geq 5)$, shown as outline boxes.

The intention is to motivate the p-value, i.e. the "observed significance level" for the observed value x, through graphical display of the region represented by $P(X \geq x)$.

The main graphic is a bar chart, where bar heights show the Binomial(n,p) distribution. The bars corresponding to the event $\{X \geq x\}$ are shown with a black outline, which gives a visual impression for this probability. There is text added to the graph, which gives the numerical value of this probability.

There are two check boxes, which allow choice of the displayed probability as either $P(X \geq x)$ (the usual "p-value") or as $P(X = x)$ (another candidate for "observed significance"). See discussion below about this.

In class it is recommended to demonstrate:

i. When the observed value x becomes larger, the p-value decreases, i.e. the evidence against H_0 becomes stronger.

ii. When p becomes larger, the p-value increases, i.e. the evidence against H_0 becomes weaker. This makes sense, since then the null hypothesis has a better chance of explaining the observed value.

iii. To see why the p-value is $P(X \geq x)$, and not $P(X = x)$, use the checkboxes mentioned above. The parameters $p = .5$, and $x = n/2$ are recommended, and then take several values of n, such as $n = 10, 40, 160$. These show that $P(X = x)$ has the problem that it depends strongly on n, and worse gets small even when there is clearly no strong evidence against H_0. On the other hand, $P(X \geq x)$ is stable for increasing n, and stays large when there is no strong evidence.

This example may be run from the XploRe web site, `www.XploRe-stat.de` directly. One opens the Java 1.1 interface, enters `library("tware")` and then enters `twpvalue()`.

Example 3

This example illustrates two points. First how the normal distribution provides a good approximation to the binomial for large n. Second why it is both important and natural to subtract the mean, and divide by the standard deviation, when doing a normal approximation.

The example starts with an overlay of three theoretical probability histograms (bar graphs where heights are probabilities), representing the Binomial distribution, with a fixed value of p, say $p = 0.6$, and with three choices of n, say $n = 10, 20, 40$, as shown in the left in Figure 3a (bottom). The instructor points out that there is a common "mound shape" to the three graphs, but that they are not close to any fixed distribution, and will not get closer to anything as the sample size n grows, since the probability mass moves to the right. However the effect can be understood, and perhaps adjusted for, by the development of the concept of centerpoint of a probability distribution, e.g. the mean.

When the centerpoint is understood, its effect in the present example is illustrated in the right column of the main graphic. This shows the three theoretical probability histograms of the random variables minus their means. Now it is apparent that mean adjustment overcomes the problem of probability mass moving off to the right, but there is a second problem with the distribution becoming more spread as the sample size grows. Again the effect can be quantitatively understood, and adjusted for, by the development of a concept of spread of a distribution, i.e. the standard deviation.

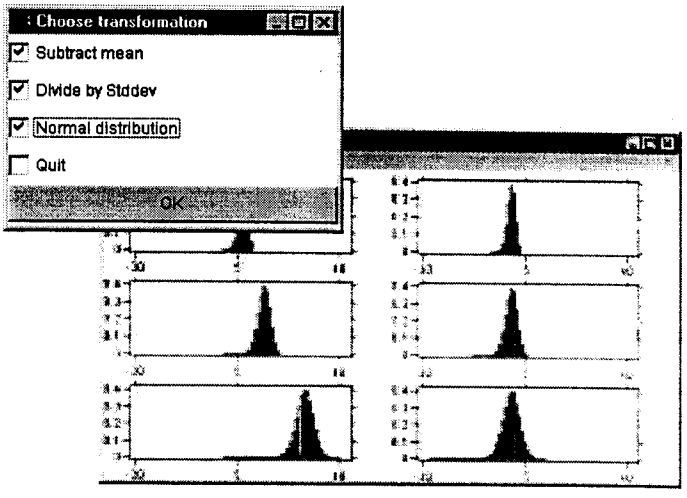


Figure 3a: Demonstrates Example 3, Standardization and Normal approximation of the Binomial. Upper: menus for controlling the transformation of the Binomial distributions. Lower: main graphic window, showing three Binomial distributions in the left column, and the corresponding transformed versions in the right.

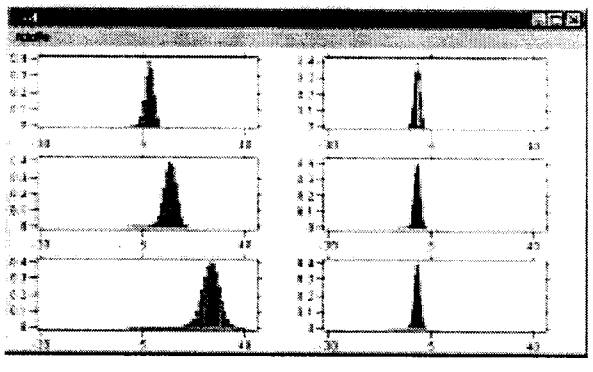


Figure 3b: Shows the effect of adjusting for the scale, on the histograms in the left part of the main output window in Figure 3a. Also shows the effect of overlaying the approximating Normal probability density.

When the spread is understood, its effect in this example is illustrated by checking the "divide by Stddev" box in the control menu. This changes the right column to plots of the probability histograms of the random variables minus their means, divided by their standard deviations as shown in Figure 3b. This shows that the distribution is clearly converging to a common shape. Then the instructor states that with more mathematics, it can be shown that this common distribution is the Gaussian, i.e. normal distribution, which is then overlaid using the "Normal Distribution" checkbox.

This example may be run from the XploRe web site `www.XploRe-stat.de` directly. One opens the Java 1.1 interface, enters `library("tware")` and then enters `twnormalize()`.

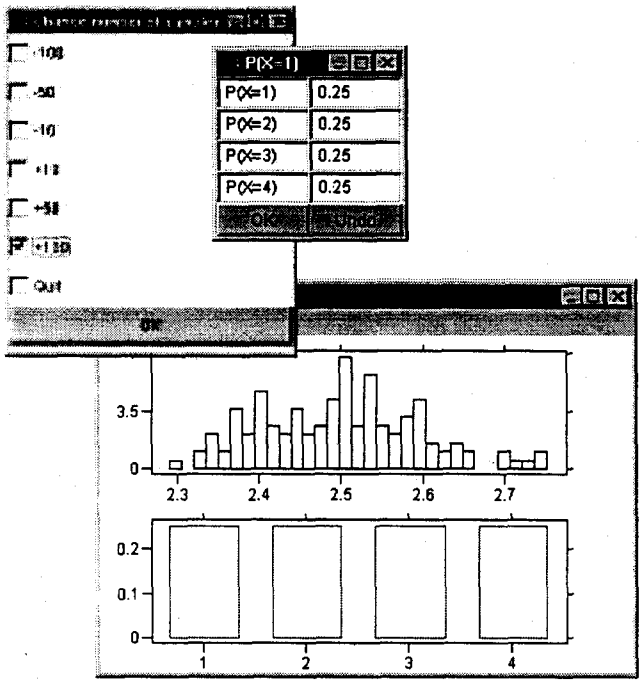


Figure 4: Demonstrates Example 4, Central Limit Theorem. Upper right: menu controlling probabilities of a 4 point distribution. Upper left: menu controlling number of realizations to average. Lower: main graphic window, showing result of repeated convolution, which demonstrates that the distribution of averages converges to the Gaussian.

Example 4

This example shows the main point of the Central Limit Theorem: that averages tend to have a Normal probability distribution, even when the individual underlying distribution is far from Normal. The example starts with a menu containing text boxes (shown in the upper right of Figure 4) for entering an initial discrete probability distribution. This distribution is supported on the integers 1,2,3,4, and after the probabilities are entered a bar graph is displayed (the lower main window shown in Figure 4), showing the probability histogram of the entered distribution. The upper left window controls the number of realizations to average, n.

Clicking OK in the upper left window shows the probability histogram (in the main window) of the average of $X_1, \ldots, X_n$ (computed by simple discrete convolution). This demonstrates how the shape tends towards that of the Normal distribution. Another push button will overlay the approximating normal distribution onto the current probability histogram. For level 2 and level 3 teaching, students could be encouraged to try this with other choices of the underlying probability distribution.

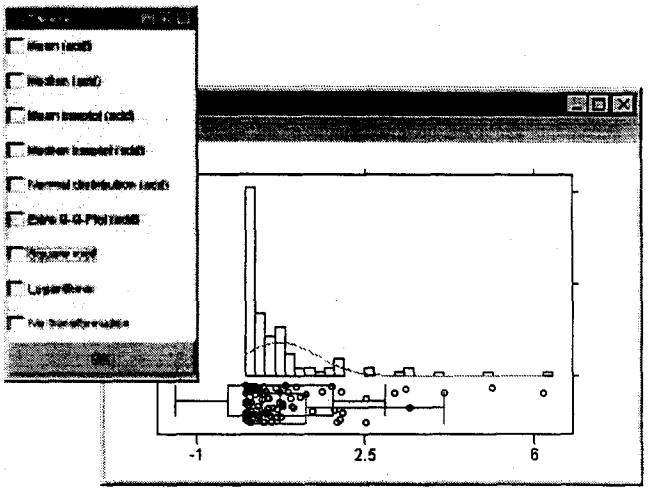


Figure 5: Demonstrates Example 5, Display of 1-d data. Left: Control menu, with checkboxes allowing different displays. Right: Main graphics window, currently showing histogram with overlaid Normal distribution, and jitter plot, with both types of boxplot (mean - standard deviation boxplot, median - quartile boxplot).

They could be challenged to find shapes which give rapid convergence to the Normal, and shapes which give very slow convergence. The student has the possibility to increase and decrease the number of the repetitions of the random drawing. This is designed for discovery of "how" and "when" the Normal limit distribution is a valid approximation as a function of sample size.

This example may be run from the web XploRe site `www.XploRe-stat.de` directly. One opens the Java 1.1 interface, enters `library("tware")` and then enters `twclt()`.

Example 5

This example illustrates the use of visual display devices for one dimensional data. It shows the relationship between the mean and median, shows how transformation can be used to make data have a distribution closer to Normal, and shows the resulting impact of transformations on the mean and the median. Display devices include histograms, jitter plots, and Q-Q (normal probability) plots. Currently considered transformations are the square root, and the logarithm.

The control menu, shown on the upper left of Figure 5, allows choice of which statstical graphics to include. Check boxes will allow the exploration of various notions of "center" and "spread" via overlaid boxplots. The "mean boxplot" is centered at the mean, with the box endpoints showing the mean plus and minus one standard deviation, and with the whiskers showing the mean plus and minus two standard deviations. The "median box plot" is centered at the median, with the box

endpoints showing the quartiles and the whiskers showing the 2.5 and 97.5 percentiles.

The chosen example has substantial skewness which shows that these two boxplots can be quite different, and furthermore that the percentile methods are giving a better notion of "center" and "spread". The square root and the logarithmic (base 10) transformations, show how this situation changes dramatically when the data are transformed. Closeness to normality, in each case, can also be studied via a Q-Q, i.e. normal probability, plot using that checkbox.

This example may be run from the XploRe web site. One opens the Java 1.1 interface, enters `library("tware")` and then enters `twld()`.

Example 6

This example gives a visual demonstration of the form of the Pearson correlation coefficient. In particular, it shows why the product moment gives a measure of "dependence," and why it is essential to "normalize," i.e. to subtract means, and divide by standard deviations, to preserve that property.

It uses simulated bivariate Gaussian data, with the number of data points, and the correlation entered through checkboxes as shown at the top of Figure 6. The data are shown with a scatterplot in the main graphics window appearing in the bottom of Figure 6. Text below shows the numerical value of the product moment, $\sum_i(x_i y_i)$, the recentered product moment, $\sum_i((x_i - \bar{x})(y_i - \bar{y}))$, and the rescaled, recentered product moment, $\sum_i((x_i - \bar{x})(y_i - \bar{y}))/(s_x s_y)$, which is the Pearson correlation coefficient.

Starting with $N(0,1)$ marginals shows how the ordinary product moment quantifies "dependence," since most values in the first and third quadrants make the product moment positive, and most values in the second and fourth quadrants make the product moment negative, while independence gives cancellation of these effects, so the product moment is essentially zero.

To understand the need for recentering and rescaling, the other menu (shown in the middle of Figure 6) allows changing the center point of the point cloud. When the centerpoint is changed, the point cloud moves accordingly (with the original position shown in gray) and the various moments also updated. The teacher can comment that the original product moment changes dramatically, while the recentered product moment stays the same. Another menu allows changing the scales, and again this change is apparent both visually, and in the product moments, which shows why normalizing by the product of the standard deviations is essential.

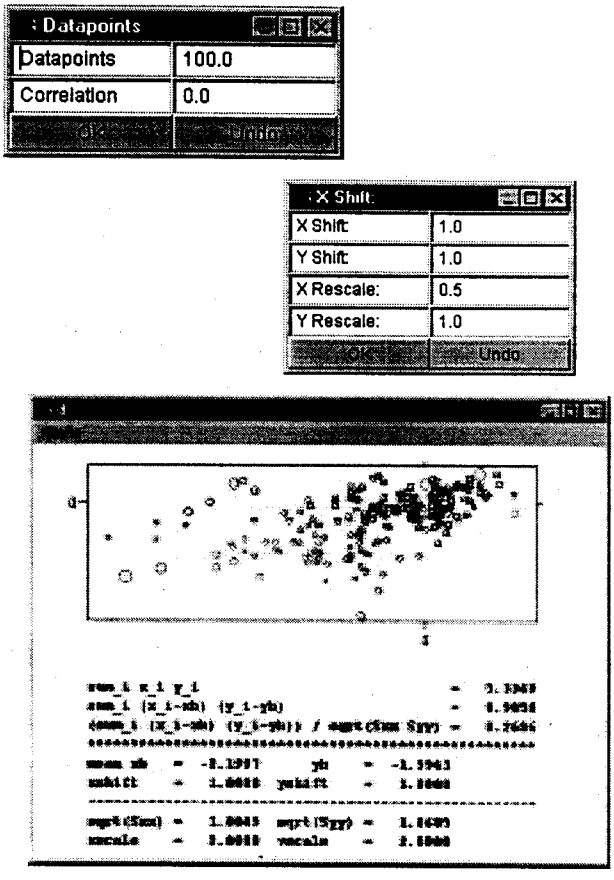


Figure 6: Demonstrates Example 6, Correlation Coefficient. Top: menu for controlling number and correlation of underlying normal data. Middle: menu for demonstrating how shifts and scales affect the product moment, but not the Pearson correlation coefficient.

This example may be run from the XploRe web site `www.XploRe-stat.de` directly. One opens the Java 1.1 interface, enters `library("tware")` and then enters `twpearson()`.

Example 7

This example gives visual insight into how least squares simple linear regression works, and the relationship between the regression of Y on X, X on Y, and total regression.

As for example 6 the data are bivariate Gaussian, and a menu (shown upper right in Figure 7) allows control of the number of data points, and the correlation. Intuitive understanding of least squares fitting is conveyed through interactive manipulation of a candidate fit line. The upper left menu in Figure 7 gives control over this process, through incremental adjustments that are selected by check boxes, followed by a push of the "OK" button. The main graphics window shows the data scatterplot, together with the least squares fit line.

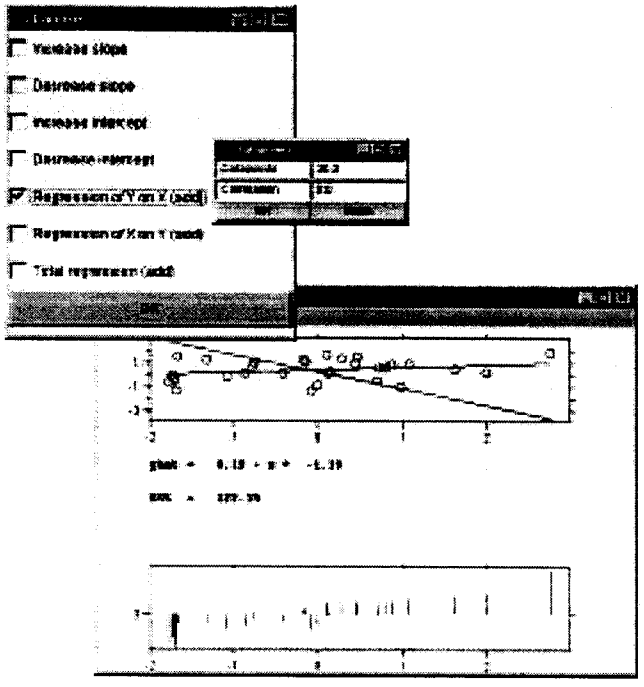


Figure 7: Demonstrates Example 7, Simple Linear Regression. Upper: menus for the changes of the regression line. Lower: main graphic window, showing result of repeated application of changing slope and intercept in comparison with LS line.

A text component shows the equation of the current line (which changes as the line is manipulated), together with the Residual Sum of Squares which gives a numerical summary of the goodness of fit. Very effective visual indication of what RSS means comes from the lower graphics part of this window, which represents the residuals as vertical lines. When the fit is poor (and hence the RSS is large), the residual plot shows why, and give a clear indication of how the line should be moved to improve the quality of the fit to the data.

Additional check boxes allow understanding the variations of regression of X on Y, and total variation, and result in appropriate shifts of the graphics. This example could be modified to allow other types of fitting, such as least L_1, or other types of robust fits.

This example may be run from the XploRe web site `www.XploRe-stat.de` directly. One opens the Java 1.1 interface, enters `library("tware")` and then enters `twlinreg()`.

Acknowledgements

We would like to thank Nathan Derby, Marlene Müller and Bernd Rönz for helpful suggestions and corrections. The paper was financially supported by the Sonderforschungsbereich 373 "Quantifikation und Simulation Ökonomischer Prozesse," Deutsche Forschungsgemeinschaft. The research of J. S. Marron was supported in part by NSF grant DMS-9504414.

References

Classroom ready examples in XploRe
`ise.wiwi.hu-berlin.de/statistik/lehrmaterial/statmat.html`

GLM tutorial
`www.xplore-stat.de/tutorials/glmstart.html`

GAUSS software
`www.aptech.com`

GAUSS programming for Econometricians
`eclab.econ.pdx.edu/gpe/`

Help system pages
`www.xplore-stat.de/help/_Xpl_Start.html`

Image processing with Java
`www.utdallas.edu/~degroat/javadip/JavaDIP.html`

MD*Tech - Method and Data Technologies
`www.mdtech.de`

Non- and Semiparametric Modeling course text
`www.quantlet.de/~scripts/scripts/spm/spm.html`

SAS software
`www.sas.com`

SHAZAM software
`shazam.econ.ubc.ca`

Splus software
`www.mathsoft.com/Splus`

Stata software
`www.stata.com`

STATLIB server of SPLUS
`lib.stat.cmu.edu/S/`

SticiGui(c) Java Tools
`www.stat.berkeley.edu/users/stark/Java`

SUN's Java Development Kit (JDK)
`www.javasoft.com`

Support Vector Machine
svm.dcs.rhbnc.ac.uk/pagesnew/1D-Reg.shtml

Symanzik (1998)
www.galaxy.gmu.edu/~symanzik

Virtual Stat Lab
www.ruf.rice.edu/~lane/stat_sim/index.html

wavelet book in PDF format
www.quantlet.de/~scripts/scripts/wav/wavpdf.pdf

Webstat Project
www.stat.sc.edu/~west/webstat/

XLISP-STAT
www.cern.ch/WebMaker/examples/xlisp/www/cldoc_1.html

XploRe
www.XploRe-stat.de

Wolfgang Härdle
Institut für Statistik und Ökonometrie
Wirtschaftswissenschaftliche Fakultät
Humboldt-Universität zu Berlin
stat@wiwi.hu-berlin.de

Sigbert Klinke
Institut für Statistik und Ökonometrie
Wirtschaftswissenschaftliche Fakultät
Humboldt-Universität zu Berlin
sigbert@wiwi.hu-berlin.de

J. S. Marron
Department of Statistics
University of North Carolina
marron@stat.unc.edu

TOPICS IN INFORMATION VISUALIZATION

Linked Data Views

By Graham Wills

Introduction

I think of a "data view" very generally as anything that gives the user a way of looking at data so as to gain insight and understanding. A data view is usually thought of as a bar chart, scatterplot, or other graphical tool, but I use the term to include a display of the results of a regression analysis, a neural net prediction or a set of descriptive statistics. In a simple case, a scroll bar is a view of a document, linked to a textual representation beside it. Selecting an area in the scroll bar using the thumb links to the associated text view to display new textual information. In general, a data view is a representation the user can look at and study to help understand relationships and determine features of interest in the data they are studying. Typically there are parameters or variations in the method of display so that some way of interacting with the view to modify its behavior is necessary.

Also typical is the desire to explain something of interest found in a view. Do data form two clusters under this particular projection of the grand tour? Is there a change in the relationship between salary and years playing baseball when the latter is greater than five years? When we see something interesting, we want to explain it, usually by considering other data views or by including additional variables. With some types of view, it is not hard to add in variables and see if they can explain the feature, or indeed if they have any effect whatsoever. In a regression analysis, you can just simply add a variable to the set of explanatory variables (taking due care with respect to multicollinearity and other confounding factors). If a histogram of X shows something of interest, you can "add" a variable Y to it by making a scatterplot of X against Y. If you want to explain something in a scatterplot, then it is possible to turn it into a rotating point cloud in 3D, and using projection pursuit or grand tour techniques, you can go to still higher dimensions.

Despite the undoubted utility of this approach, it does present some problems that prevent it from being a complete solution. The main ones are:

- As plots become increasingly complex, they become harder to interpret. Few people have problems with most one-dimensional plots. Scatterplots, tables and grouped boxplots or other displays involving two dimensions are easily learnable. But the necessity of spinning and navigating

Härdle, W., Klinke, S. and Marron, J. (1999) Connected Teaching Of Statistics

Econometric Theory, Problems and Solutions 15, 258-259

ASYMPTOTIC PROPERTIES OF THE NONPARAMETRIC PART IN PARTIALLY LINEAR HETEROSCEDASTIC REGRESSION MODELS

Wolfgang Härdle, Hua Liang and Axel Werwatz *

Abstract

This paper considers estimation of the unknown function $g(\bullet)$ in the partially linear regression model $Y_i = X_i^T\beta + g(T_i) + \varepsilon_i$ with heteroscedastic errors. We first construct a class of estimates g_n of g and prove that, under appropriate conditions, g_n is weak, mean square error consistent. Rates of convergence and asymptotic normality for the estimator g_n are also established.

Key Words and Phrases:Key words and phrases: Asymptotic normality, consistency, heteroscedasticity, kernel estimation, rates of convergence, partially linear model, semiparametric models.

1 INTRODUCTION

Semiparametric models combine the flexibility of nonparametric modeling with structural parametric components. One such model that has received a lot of attention in the literature is the semiparametric partially linear regression model

$$Y_i = X_i^T\beta + g(T_i) + \varepsilon_i, i = 1, \ldots, n, \tag{1.1}$$

where X and T are (possibly) multidimensional regressors, β a vector of unknown parameters, $g(\bullet)$ an unknown smooth function and ε an error term with mean zero conditional on X and T.

*Hua Liang is Associate Professor of Statistics, Institute of Systems Science, Chinese Academy of Sciences, Beijing 100080, China. Wolfgang Härdle is Professor of Econometrics, Axel Werwatz is Dr. of Economics. The first and third authors are at the Institut für Statistik und Ökonometrie, Humboldt-Universität zu Berlin, D-10178 Berlin, Germany. This research was supported by Sonderforschungsbereich 373 "Quantifikation und Simulation Ökonomischer Prozesse" . The second author was supported by Alexander von Humboldt Foundation.

Liang, H., Härdle, W. and Werwatz, A. (1999) Asymptotic Properties of the Nonparametric Part in Partially Linear Heteroskedastic Models.

Econometric Theory, Problems and Solutions 15, 258-259

Well-known applications in the econometrics literature that can be put in the form of (1.1) are the human capital earnings function (Willis (1986)) and the wage curve (Blanchflower and Oswald (1994)). In both cases, log-earnings of an individual are related to personal characteristics (sex, marital status) and measures of a person's human capital like schooling and labor market experience. Economic theory suggests a non-linear relationship between log-earnings and labor market experience, which therefore plays the role of the variable T in (1.1). The wage curve is obtained by including the local unemployment rate as an additional regressor, with a possibly non-linear influence. Rendtel and Schwarze (1995), for instance, estimate $g(\bullet)$ as a function of the local unemployment rate using smoothing-splines and find a U-shaped relationship.

Under various assumptions, several authors have considered estimation of β in (1.1) at a parametric rate. Chen (1988), Engle, et al. (1986), Heckman (1986), Robinson (1988), Schick (1996) and Speckman (1988) constructed $\sqrt{n}-$consistent estimators of β. Cuzick (1992a) studied efficient estimation of β when the error density is known. Efficient estimation when the error distribution is of an unknown form is treated in Cuzick (1992b) and Schick (1993).

In this paper, we will instead focus on deriving the asymptotic properties of an estimator of the unknown function $g(\bullet)$. We consider its consistency, weak convergence rate and asymptotic normality. We will derive these results for a specific version of (1.1) with nonstochastic regressors, heteroscedastic errors and T univariate.

The remainder of this paper is organized as follows. In the following section we will describe methods for estimating β and $g(\bullet)$. We prove consistency and asymptotic normality of the estimator of $g(\bullet)$ in sections 3 and 4. We illustrate the usefulness of the estimator and the relevance of the asymptotic distribution results for applied work by a small-scale Monte Carlo study and an empirical illustration in the final section of the paper.

2 THE ESTIMATOR

Specifically, we consider estimation of $g(\bullet)$ (and β) in the following partially linear, semiparametric regression model:

$$Y_i = X_i^T\beta + g(T_i) + \varepsilon_i, i = 1, \ldots, n \tag{2.1}$$

where β is an unknown p dimensional parameter vector, $g(\bullet)$ an unknown, smooth function from $[0,1]$ to $\mathbb{R}^1$, (X_1, T_1), (X_2, T_2) ... are known, nonrandom design points and $\varepsilon_1, \ldots, \varepsilon_n$ are independent mean zero random errors with nonconstant finite variance. We allow the variance of ε to depend on X and T in an arbitrary way.

Previous work in a heteroscedastic setting has focused on the nonparametric regression model (i.e. $\beta = 0$). Müller et al. (1987) proposed an estimate of the variance function by using kernel smoother, and then proved that the estimate is uniformly consistent. Hall and Carroll (1989) considered consistency of estimates of $g(\bullet)$. Eubank et al. (1990) proposed trigonometric series type estimators g_λ of g. They investigated asymptotic approximations of the integrated mean squared error and the partially integrated mean squared error of g_λ. The heteroscedastic version of (1.1) with $\beta \neq 0$ has been considered in Schick (1996) but he considers weighted least squares estimation of β. We focus on nonparametric estimation of $g(\bullet)$ as a function of T.

Liang, H., Härdle, W. and Werwatz, A. (1999) Asymptotic Properties of the Nonparametric Part in Partially Linear Heteroskedastic Models.

Suppose we knew β. Then we may estimate $g(\bullet)$ by nonparametric regression of $Y_i - X_i^T\beta$ (the variation in Y_i not accounted for by the linear component $X_i^T\beta$) on T_i.

In the literature one can find various methods for estimating $g(\bullet)$ nonparametrically, e.g., kernel, nearest neighbor, orthogonal series, piecewise polynomial and smoothing splines. See Härdle (1990) for an extensive discussion of their statistical properties. All these estimators may be written as weighted local averages of the observed values of the dependent variable with the weights depending on the values of the explanatory variables. In our case, we can write (still assuming that β is known):

$$\hat{g}(t) = \sum_{i=1}^{n} \omega_{ni}(t)(Y_i - X_i^T\beta), \tag{2.2}$$

where $\omega_{ni}(t) = \omega_{ni}(t; T_1, T_2, \ldots, T_n)$ are weight functions that depend on the observations $T_1, \ldots, T_n$.

For instance, a Gasser-Müller-type kernel estimator takes

$$\omega_{ni}(t) = \frac{1}{h}\int_{s_{i-1}}^{s_i} K\Big(\frac{t-s}{h_n}\Big)ds \quad 1 \le i \le n$$

for $s_0 = 0$, $s_n = 1$, $s_i = \frac{1}{2}(T_{(i)} + T_{(i+1)})$. Here $T_{(1)}, \ldots, T_{(n)}$ are sample order statistics, $K(\bullet)$ is the kernel function and h denotes the bandwidth. See Remark 4.1 below for details.

Given the estimator $\hat{g}(t)$ as defined in (2.2) we may estimate β by the least squares regression of

$$\begin{aligned} Y_i &= X_i^T\beta + \hat{g}(T_i) + \epsilon_i \\ Y_i - \sum_{j=1}^{n} \omega_{nj}(T_i)Y_j &= \Big\{X_i - \sum_{j=1}^{n} \omega_{nj}(T_i)X_j\Big\}^T\beta + \epsilon_i \\ \widetilde{Y}_i &= \widetilde{X}_i^T\beta + \epsilon_i \end{aligned} \tag{2.3}$$

That is, we estimate β by the least squares estimator

$$\beta_{LS} = (\widetilde{X}^T\widetilde{X})^{-1}\widetilde{X}^T\widetilde{Y} \tag{2.4}$$

where $\widetilde{X} = (\widetilde{X}_1, \ldots, \widetilde{X}_n)^T$ and $\widetilde{Y} = (\widetilde{Y}_1, \ldots, \widetilde{Y}_n)^T$ are the presmoothed design and response variables.

In the final step we obtain the feasible estimator of $g(\bullet)$ by substituting β_{LS} for the unknown β in (2.2):

$$g_n(t) = \sum_{i=1}^{n} \omega_{ni}(t)(Y_i - X_i^T\beta_{LS}). \tag{2.5}$$

Further motivation for the estimators defined in (2.4) and (2.5) is given in Speckman (1988) and Gao, et al. (1995). Note though that β_{LS} is not an efficient estimator in the sense of asymptotic normality.

In the following section we state and prove the weak, mean square error consistency and give the rates of convergence of $g_n(t)$ under various assumptions.

3 CONSISTENCY RESULTS

All technical preliminaries needed in the proofs of the following results are collected in Appendix A as lemmas. For convenience and simplicity, we always let C denote some positive constant not depending on n. We will use the following assumptions.

Assumption 1. *There exist continuous functions $h_j(\bullet)$ defined on $[0,1]$ such that each element of X_i satisfies*

$$x_{ij} = h_j(T_i) + u_{ij} \quad 1 \le i \le n, \quad 1 \le j \le p, \tag{3.1}$$

where u_{ij} is a sequence of real numbers which satisfy $\lim_{n\to\infty} \frac{1}{n}\sum_{i=1}^{n} u_i = 0$ and

$$\lim_{n\to\infty} \frac{1}{n}\sum_{i=1}^{n} u_i u_i^T = B \tag{3.2}$$

is a positive definite matrix, and

$$\limsup_{n\to\infty} \frac{1}{a_n} \max_{1\le k\le n} \left|\sum_{i=1}^{k} u_{j_i m}\right| < \infty \quad \text{for } m = 1, \ldots, p \tag{3.3}$$

holds for all permutations $(j_1, \ldots, j_n)$ of $(1, 2, \ldots, n)$ where $u_i = (u_{i1}, \ldots, u_{ip})^T$, $a_n = n^{1/2}\log n$.

Assumption 2.

(a) $\sum_{i=1}^{n} \omega_{ni}(t) \to 1$ as $n \to \infty$;

(b) $\sum_{i=1}^{n} |\omega_{ni}(t)| \le C$ for all t and some constant C;

(c) $\sum_{i=1}^{n} |\omega_{ni}(t)| I(|t - T_i| > a) \to 0$ as $n \to \infty$ for all $a > 0$;

(d) $\sup_{i\le n} |\omega_{ni}(t)| = O(\log^{-1} n)$.

Denote $n_t = \left\{\sum_{i=1}^{n} \omega_{ni}^2(t)\right\}^{-1}$.

Assumption 3.

(a) $\sup_n(n_t) \sup_{1\le i\le n} |\omega_{ni}(t)| < \infty$, and $n_t = o(n)$;

(b) $\sqrt{n_t} \sup_{1\le i\le n} |\omega_{ni}(t)| = O(n^{-\alpha/2})$ for some $1 > \alpha > 0$;

(c) $\sum_{i=1}^{n} \omega_{ni}^2(t) E\varepsilon_i^2 = \sigma_0^2/n_t + o(1/n_t)$ for some $\sigma_0^2 > 0$.

Remark 3.1 *Assumption 1 is a common requirement for proving consistency of β in the partially linear model (1.1). In fact, (3.1) of Assumption 1 is parallel to the case*

$$h_j(T_i) = E(x_{ij}|T_i) \quad \text{and } u_{ij} = x_{ij} - E(x_{ij}|T_i)$$

when (X_i, T_i) are random variables. (3.2) is similar to the result of the strong law of large numbers for random errors. (3.3) is similar to law of the iterated logarithm. More detailed discussions may be found in Speckman (1988) and Gao et al. (1995).

Theorem 3.1 *Under Assumptions 1 and 2,* $E\{g_n(t)\} \to g(t)$ *as* $n \to \infty$ *at every continuity point of the function* g.

Proof. Decompose the difference $g_n(t) - g(t)$ as follows by direct calculation.

$$g_n(t) - g(t) = \sum_{j=1}^{n} \omega_{nj}(t)\{g(T_i) + \varepsilon_i - X_i^T(\widetilde{X}^T\widetilde{X})^{-1}\widetilde{X}^T\tilde{g}(T) - X_j^T(\widetilde{X}^T\widetilde{X})^{-1}\widetilde{X}^T\tilde{\varepsilon}\} - g(t)$$

where $\tilde{g}(T) = \{\tilde{g}(T_1), \ldots, \tilde{g}(T_n)\}^T$ and $\tilde{g}(T_i) = g(T_i) - \sum_{j=1}^{n} \omega_{nj}(T_i)g(T_j)$ and $\tilde{\varepsilon}$ just like $\widetilde{X}$. It follows that

$$E\{g_n(t)\} - g(t) = \Big\{\sum_{i=1}^{n} \omega_{ni}(t)g(T_i) - g(t)\Big\} - \sum_{i=1}^{n} \omega_{ni}(t)X_i^T(\widetilde{X}^T\widetilde{X})^{-1}\widetilde{X}^T\tilde{g}(T) \tag{3.4}$$

The first term tends to zero by Lemma A.1 (i). By lemmas A.2 and A.1 (i) and Cauchy-Schwarz inequality, we know that every element of $(\widetilde{X}^T\widetilde{X})^{-1}\widetilde{X}^T\tilde{g}(T)$ is $o(n^{-1/2})$, i.e.,

$$\Big\{(\widetilde{X}^T\widetilde{X})^{-1}\widetilde{X}^T\tilde{g}(T)\Big\}_j = o(n^{-1/2}) \quad \text{for } j = 1, \ldots, p \tag{3.5}$$

It suffices to show that every element of $\sum_{i=1}^{n} \omega_{ni}(t)X_i$ is $O(n^{1/2})$. Observe that

$$\sum_{i=1}^{n} \omega_{ni}(t)x_{ij} = \sum_{i=1}^{n} \omega_{ni}(t)\{h_j(T_i) + u_{ij}\}$$

Since $h_j(\bullet)$ is continuous, $\sum_{i=1}^{n} \omega_{ni}(t)h_j(T_i)$ converges to $h(t)$ on the continuity point of $h(t)$ by the same proof as one for Lemma A.1(i). Moreover, by Abel's inequality and Assumption 2 (d),

$$\Big|\sum_{i=1}^{n} \omega_{ni}(t)u_{ij}\Big| \le \sup_{1\le i\le n} |\omega_{ni}(t)| \max_{1\le k\le n} \Big|\sum_{i=1}^{k} u_{j_i m}\Big| = O(n^{1/2})$$

Thus

$$\sum_{i=1}^{n} \omega_{ni}(t)x_{ij} = O(n^{1/2}) \tag{3.6}$$

and we complete the proof of Theorem 3.1.

Theorem 3.1 shows that $g_n(t)$ is an asymptotically unbiased estimator of $g(t)$ at every continuity point of $g(t)$. The next result, Theorem 3.2, will demonstrate that $g_n(t)$ is also mean square-error consistent.

Theorem 3.2 *Assume the conditions of Theorem 3.1 hold except Assumption 2 (d) which is replaced by* $\sup_{i\le n} |\omega_{ni}(t)| = o(\log^{-1} n)$. *Then* $E\{g_n(t) - g(t)\}^2 \to 0$ *as* $n \to \infty$.

Proof. Directly, $E\{g_n(t) - g(t)\}^2$ can be bounded by

$$CE\Big|\sum_{i=1}^{n} \omega_{ni}(t)g(T_i) - g(t)\Big|^2 + CE\Big|\sum_{i=1}^{n} \omega_{ni}(t)\varepsilon_i\Big|^2$$
$$+ CE\Big|\sum_{i=1}^{n} \omega_{ni}(t)(\widetilde{X}^T\widetilde{X})^{-1}\widetilde{X}^T\tilde{g}(T)\Big|^2 + CE\Big|\sum_{i=1}^{n} \omega_{ni}(t)X_i^T(\widetilde{X}^T\widetilde{X})^{-1}\widetilde{X}^T\tilde{\varepsilon}\Big|^2 \tag{3.7}$$

In the proof of Theorem 3.1 we obtained that the first and third terms of (3.7) converge to zero as n tends to infinity. The second can be shown to be order $o(1)$ by a direct calculation.

We shall now prove the fourth term also converges to zero. Denote $(\widetilde{X}^T\widetilde{X})^{-1}\widetilde{X}^T = (\eta_{ji})_{p\times n}$.

$$\begin{aligned} E\Big|\sum_{i=1}^{n}\omega_{ni}(t)X_i^T(\widetilde{X}^T\widetilde{X})^{-1}\widetilde{X}^T\varepsilon\Big|^2 &= E\Big\{\sum_{i=1}^{n}\omega_{ni}(t)\Big(\sum_{k=1}^{p}\sum_{l=1}^{n}x_{ik}\eta_{kl}\varepsilon_l\Big)\Big\}^2 \\ &= \sum_{l=1}^{n}\Big(\sum_{i=1}^{n}\sum_{k=1}^{p}\omega_{ni}(t)x_{ik}\eta_{kl}\Big)^2\sigma_l^2. \end{aligned}$$

It follows from the arguments for (3.6) that this equals to $o(n)\sum_{l=1}^{n}\eta_{kl}^2$. Since $\sum_{l=1}^{n}\eta_{kl}^2$ and the elements of the $k-$th row of $(\widetilde{X}^T\widetilde{X})^{-1}$ have the same order $O(n^{-1})$. It follows that

$$E\Big|\sum_{i=1}^{n}\omega_{ni}(t)X_i^T(\widetilde{X}^T\widetilde{X})^{-1}\widetilde{X}^T\varepsilon\Big|^2 = o(1). \tag{3.8}$$

Furthermore, we can easily show that

$$E\Big|\sum_{i=1}^{n}\omega_{ni}(t)X_i^T(\widetilde{X}^T\widetilde{X})^{-1}\sum_{k=1}^{n}\widetilde{X}_k\Big\{\sum_{l=1}^{n}\omega_{nl}(T_k)\varepsilon_k\Big\}\Big|^2 = o(1). \tag{3.9}$$

Combining (3.8) and (3.9) ensures that the fourth term of (3.7) is $o(1)$, and thus completes the proof of Theorem 3.2.

The following result gives the weak convergence rate of g_n under stronger assumptions on $\{\omega_{ni}(t)\}$ than those given in Assumption 2. Here we list these assumptions.

Assumption 2'. The weight functions $\omega_{ni}(t)$ satisfy:

(a) $\sup_t|\sum_{i=1}^{n}\omega_{ni}(t)-1| = O_P(n^{-1/3}\log n)$;

(b) $\sup_t\sum_{i=1}^{n}|\omega_{ni}(t)|I(|t-T_i|>c_n) = O(d_n)$, where d_n and c_n are $n^{-1/3}\log n$;

(c) $\sup_t\max_{1\le i\le n}|\omega_{ni}(t)| = O_P(n^{-2/3})$.

Theorem 3.3 *Assume $g(\bullet)$ and $h_j(\bullet)$ are Lipschitz continuous of order 1 and Assumptions 1 and 2' hold. Then*

$$g_n(t)-g(t) = O_P(n^{-1/3}\log n).$$

Proof. By Lemma A.1 (ii),

$$\sum_{i=1}^{n}\omega_{ni}(t)g(T_i)-g(t) = O_P(n^{-1/3}\log n).$$

Using Assumption 2' (c) and Chebyshev's inequality we have (See also Lemma 2.3 of Liang and Härdle, 1997)

$$\sum_{i=1}^{n}\omega_{ni}(t)\varepsilon_i = O_P(n^{-1/3}\log n).$$

The similar arguments as that for (3.5) and (3.6) yield

$$\sum_{i=1}^{n} \omega_{ni}(t) X_i^T (\widetilde{X}^T \widetilde{X})^{-1} \widetilde{X}^T \tilde{g}(T) = O_P(n^{-1/3} \log n).$$

Finally, observe that $\sum_{i=1}^{n} \omega_{ni}(t) u_{ij} = O(1)$ and then $\sum_{i=1}^{n} \omega_{ni}(t) x_{ij} = O(1)$ for $j = 1, \ldots, p$. Thus, by the arguments for (3.8) and (3.9),

$$E\Big|\sum_{i=1}^{n} \omega_{ni}(t) X_i^T (\widetilde{X}^T \widetilde{X})^{-1} \widetilde{X}^T \tilde{\varepsilon}\Big|^2 = O(n^{-1}) \tag{3.10}$$

which entails

$$\sum_{i=1}^{n} \omega_{ni}(t) X_i^T (\widetilde{X}^T \widetilde{X})^{-1} \widetilde{X}^T \tilde{\varepsilon} = O_P(n^{-1/3} \log n).$$

This completes the proof of Theorem 3.3.

Remark 3.2 *We can conclude from the above arguments that*

$$\limsup_{n \to \infty} (n^{2/3} \log^{-2} n) E\{g_n(t) - g(t)\}^2 < \infty.$$

Theorem 3.4 gives the asymptotic variance of $g_n(t)$.

Theorem 3.4 *Under Assumptions 1, 2' and 3,* $n_t Var\{g_n(t)\} \to \sigma_0^2$ *as* $n \to \infty$.

Proof.

$$\begin{aligned} n_t Var\{g_n(t)\} &= n_t E\Big\{\sum_{i=1}^{n} \omega_{ni}(t)\varepsilon_i\Big\}^2 + n_t E\Big\{\sum_{i=1}^{n} \omega_{ni}(t) X_i^T (\widetilde{X}^T \widetilde{X})^{-1} \widetilde{X}^T \tilde{\varepsilon}\Big\}^2 \\ &\quad -2n_t E\Big\{\sum_{i=1}^{n} \omega_{ni}(t)\varepsilon_i\Big\} \cdot \Big\{\sum_{i=1}^{n} \omega_{ni}(t) X_i^T (\widetilde{X}^T \widetilde{X})^{-1} \widetilde{X}^T \tilde{\varepsilon}\Big\} \end{aligned}$$

The first term converges to σ_0^2. The second term tends to zero by (3.10), and then the third term also tends to zero by the Cauchy-Schwarz inequality. #

4 ASYMPTOTIC NORMALITY

In the nonparametric regression model, Liang (1995) proved asymptotic normality for independent ε_i's under the mild conditions. In this section, we shall consider the asymptotic normality of g_n under the appropriate assumptions.

Assumption 2". *The weight functions* $\omega_{ni}(t)$ *satisfy:*

(a) $\sum_{i=1}^{n} \omega_{ni}(t) - 1 = o(n_t^{-1/2})$;

(b) $\sum_{i=1}^{n} |\omega_{ni}(t)| I(|t - T_i| > c_n') = o(d_n')$, *where* c_n' *and* d_n' *are* $o(n_t^{-1/2})$.

Theorem 4.1 *Suppose that $g(t)$ is Lipschitz continuous of order 1, and that the assumptions 1, 2"and 3 hold. Assume that $\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_n$ are independent random variables with $E\varepsilon_i = 0$ and $\inf_i \sigma_i^2 > c_\sigma > 0$ for some c_σ. There exists a function $G(u)$ satisfying*

$$\int_0^\infty uG(u)du < \infty \tag{4.1}$$

such that

$$P(|\varepsilon_i| > u) \leq G(u), \quad \text{for } i = 1, \ldots, n \quad \text{and large enough } u. \tag{4.2}$$

If

$$\frac{\max_{1\leq i\leq n} \omega_{ni}^2(t)}{\sum_{i=1}^n \omega_{ni}^2(t)} \to 0 \qquad \text{as} \quad n \to \infty. \tag{4.3}$$

Then

$$\frac{g_n(t) - g(t)}{\sqrt{Var\{g_n(t)\}}} \longrightarrow^{\mathcal{L}} N(0,1) \text{ as } n \to \infty.$$

Proof. At first, we can prove that, under the conditions of Theorem,

$$\frac{g_n(t) - Eg_n(t)}{\sqrt{Var\{g_n(t)\}}} \longrightarrow^{\mathcal{L}} N(0,1) \quad \text{as } n \to \infty.$$

It suffices to show that

$$\frac{Eg_n(t) - g(t)}{\sqrt{Var\{g_n(t)\}}} = \sqrt{n_t}\{Eg_n(t) - g(t)\} + o(1) = o(1).$$

For $c_n' = o(n_t^{-1/2})$. Note that

$$\begin{aligned} |Eg_n(t) - g(t)| &\leq \sum_{i=1}^n |\omega_{ni}(t)\{g(T_i) - g(t)\}|\{I(|T_i - t| > c_n') + I(|T_i - t| \leq c_n')\} \\ &\quad + |g(t)|\Big|\sum_{i=1}^n \omega_{ni}(t) - 1\Big| \\ &\leq \delta(g, c_n') \cdot B + 2C\sum_{i=1}^n |\omega_{ni}(t)| I(|T_i - t| > c_n') + C\Big|\sum_{i=1}^n \omega_{ni}(t) - 1\Big|, \end{aligned}$$

where $C = \sup_{t\in[0,1]} |g(t)|$ and $\delta(g, c_n') = \sup_{|t-t'|\leq c_n'} |g(t) - g(t')|$. Assumption 2" and the previous arguments yield the conclusion of Theorem 4.1. #

Remark 4.1 *In this remark, we shall give concrete weight functions $\{\omega_{ni}(t), i = 1, \ldots, n\}$ which satisfy the assumptions given in the former context, in order to explain the reasonability of the results established in previous sections carefully.*

Assume

$$\omega_{ni}(t) = \frac{1}{h_n}\int_{s_{i-1}}^{s_i} K\Big(\frac{t-s}{h_n}\Big)ds \quad 1 \le i \le n, \tag{4.4}$$

where $s_0 = 0$, $s_n = 1$ *and* $s_i = \frac{1}{2}(T_{(i)} + T_{(i+1)})$, $1 \le i \le n-1$. h_n *is a sequence of bandwidth parameters which tends to zero as* $n \to \infty$ *and* $K(\bullet)$ *is a kernel function, which is supposed to have compact support and to satisfy*

$$\text{supp}(K) = [-1,1], \sup|K(x)| \le C < \infty, \int K(u)du = 1 \text{ and } K(u) = K(-u).$$

Obviously Assumptions 2(a), (b) and (d) are satisfied for the weight functions given in (4.4). If

$$\int_{|u|\ge ah_n^{-1}} K(u)du = o(1).$$

Then Assumption 2 (c) hold also. In fact

$$\begin{aligned}\sum_{i=1}^{n}\omega_{ni}(t)I(|T_i - t| > a) &= \frac{1}{h_n}\sum_{i=1}^{n}\int_{s_{i-1}}^{s_i} K\Big(\frac{t-s}{h_n}\Big)dsI(|T_i - t| > a)\\ &\le \frac{1}{h_n}\int_{|T_i - s|\ge a - \max|T_i - T_{i-1}|} K\Big(\frac{t-s}{h_n}\Big)ds\\ &\le \frac{1}{h_n}\int_{|u|\ge h_n^{-1}(a-\max|T_i-T_{i-1}|)} K(u)du = o(1).\end{aligned}$$

Now let us take $h_n = Cn^{-1/3}$ *for some* $C > 0$ *and suppose*

$$\int_{|u|\ge ah_n^{-1}} K(u)du = O(n^{-1/3}\log n).$$

There exist constants $C_1, C_2 > 0$ *such that*

$$\frac{C_1}{n} \le \min_{1\le i\le n}|T_i - T_{i-1}| \le \max_{1\le i\le n}|T_i - T_{i-1}| \le \frac{C_2}{n}.$$

Then we can take $n_t = nh_n$, *and Assumptions 3 and 2" hold. Theorem 4.1 implies that*

$$\sqrt{nh_n}\{g_n(t) - g(t)\} \longrightarrow^{\mathcal{L}} N(0, \sigma_0^2) \text{ as } n \to \infty.$$

This is just the classical conclusion in nonparametric regression estimation.

5 NUMERICAL EXAMPLES

In this section we will illustrate the finite-sample behavior of the estimator by applying it to true data and by performing a small simulation study.

In the introduction we already mentioned the human-capital earnings function as a well-known econometric application that can be put into the form of a partially linear model.

It typically relates the logarithm of earnings to a set of explanatory variables describing an individual's skills, personal characteristics and labour market conditions. Specifically, we estimate β and $g(\bullet)$ in the model

$$\ln Y_i = X_i^T \beta + g(T_i) + \varepsilon_i, \tag{5.1}$$

where X contains two dummy variables indicating the level of secondary schooling a person has completed and T, is a measure of labour market experience (defined as the number of years spent in the labour market and approximated by subtracting (years of schooling + 6) from a person's age).

Under certain assumptions, the estimate of β can be interpreted as the rate of return from obtaining the respective level of secondary schooling. Regarding $g(T)$, human capital theory suggests a concave form: rapid human capital accumulation in the early stage of one's labor market career are associated with rising earnings that peak somewhere during midlife and decline thereafter as hours worked and the incentive to invest in human capital decrease. To allow for concavity, parametric specifications of the earnings-function typically include T and T^2 in the model and obtain a positive estimate for the coefficient of T and a negative estimate for the coefficient of T^2.

For nonparametric fitting, we use a Nadaraya-Watson weight function with quartic kernel

$$(15/16)(1-u^2)^2 I(|u| \leq 1)$$

and chose the bandwidth using cross-validation. The estimate of $g(T)$ is depicted in Figure 1. In a sample size that is lower than in most empirical investigations of the human capital earnings function we obtain an estimate that nicely agrees with the concave relationship envisioned by economic theory and often confirmed by parametric model fitting.

We also conducted a small simulation study to get further insights into the small-sample performance of the estimator of $g(\bullet)$. We consider the model

$$Y_i = X_i^T \beta + \sin(\pi T_i) + \sin(X_i^T \beta + T_i)\varepsilon_i, \quad i = 1, \ldots, n = 300$$

where ε_i is standard normally distributed and X_i and T_i are sampled from a uniform distribution on $[0,1]$. We set $\beta = (1, 0.75)^T$ and performed 100 replications of generating samples of size $n = 300$ and estimating $g(\bullet)$. Figure 2 depicts the "true" curve $g(T) = \sin(\pi T)$ (solid-line) and an average of the 100 estimates of $g(\bullet)$ (dashed-line). The average estimate nicely captures the shape of $g(\bullet)$.

A LEMMAS

In this appendix we state some useful lemmas.

Lemma A.1 *Suppose that Assumption 2 (a)-(c) hold and $g(\bullet)$ and $h_j(\bullet)$ are continuous. Then*

$$(i) \qquad \max_{1 \leq i \leq n} \left| G_j(T_i) - \sum_{k=1}^{n} \omega_{nk}(T_i) G_j(T_k) \right| = o(1)$$

Econometric Theory, Problems and Solutions 15, 258-259

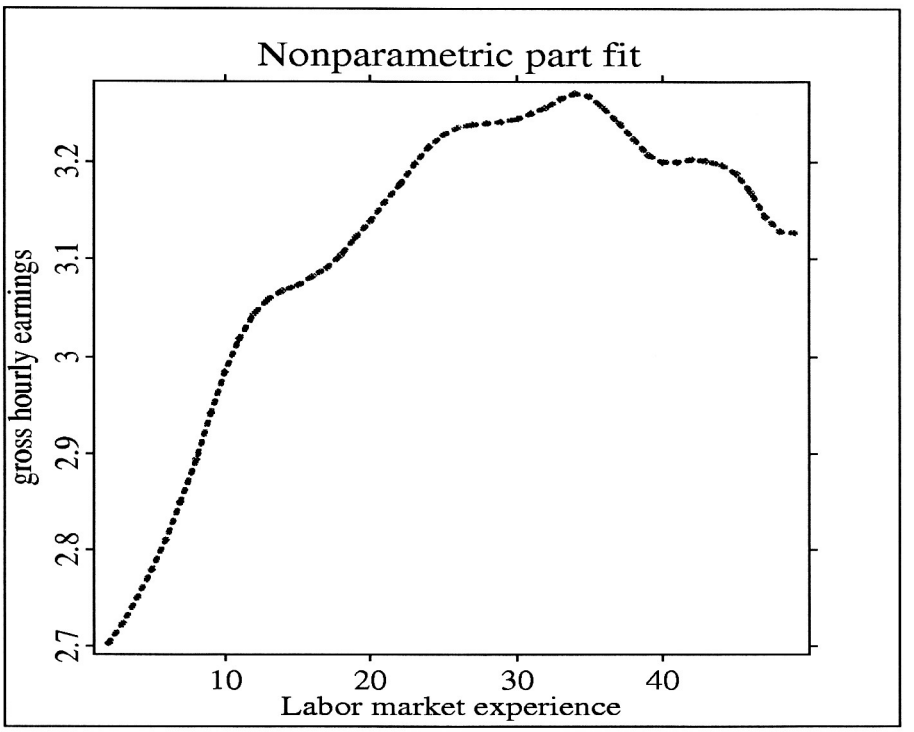


Figure 1: *Relationship of log-earnings and labour-market experience*

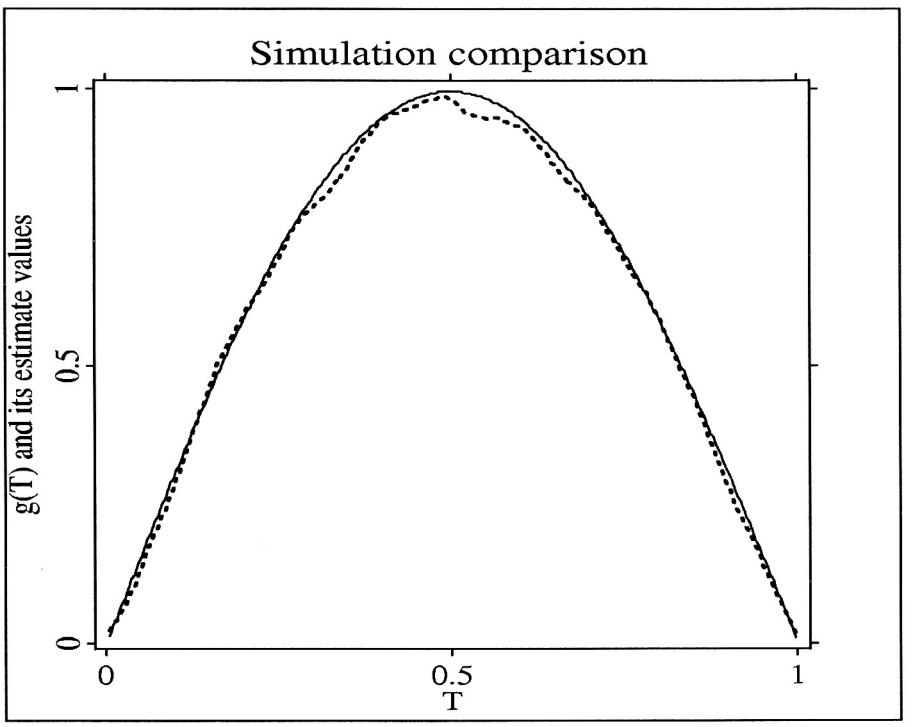


Figure 2: *Estimates of the function* $g(T)$

Liang, H., Härdle, W. and Werwatz, A. (1999) Asymptotic Properties of the Nonparametric Part in Partially Linear Heteroskedastic Models.

Furthermore, if $g(\bullet)$ and $h_j(\bullet)$ are Lipschitz continuous of order 1 and Assumption 2' (a)-(c) and 2 (b) hold. Then

$$(ii) \qquad \max_{1\le i\le n}\Big|G_j(T_i) - \sum_{k=1}^{n}\omega_{nk}(T_i)G_j(T_k)\Big| = O(c_n + d_n)$$

for $j = 0, \ldots, p$. Where $G_0(\bullet) = g(\bullet)$ and $G_l(\bullet) = h_l(\bullet)$ for $l = 1, \ldots, p$.

Proof. We only present the proof of (ii) for $g(\bullet)$. The proofs of other cases and (i) are similar. Observe that

$$\begin{aligned}\sum_{i=1}^{n}\omega_{ni}(t)\{g(T_i) - g(t)\} &= \sum_{i=1}^{n}\omega_{ni}(t)\{g(T_i) - g(t)\} + \Big\{\sum_{i=1}^{n}\omega_{ni}(t) - 1\Big\}g(t)\\ &= \sum_{i=1}^{n}\omega_{ni}(t)\{g(T_i) - g(t)\}I(|T_i - t| > c_n)\\ &\quad + \sum_{i=1}^{n}\omega_{ni}(t)\{g(T_i) - g(t)\}I(|T_i - t| \le c_n) + \Big\{\sum_{i=1}^{n}\omega_{ni}(t) - 1\Big\}g(t).\end{aligned}$$

By Assumption 2'(b) and Lipschitz continuity of $g(\bullet)$

$$\sum_{i=1}^{n}\omega_{ni}(t)\{g(T_i) - g(t)\}I(|T_i - t| > c_n) = O(d_n), \tag{A.1}$$

and

$$\sum_{i=1}^{n}\omega_{ni}(t)\{g(T_i) - g(t)\}I(|T_i - t| \le c_n) = O(c_n). \tag{A.2}$$

(A.1)-(A.2) and Assumption 2 (a) complete the proof of Lemma A.1.

Lemma A.2 *Under Assumptions 1 and 2.*

$$\lim_{n\to\infty}\frac{1}{n}\widetilde{X}^T\widetilde{X} = B$$

Proof. Denote $\bar{h}_{ns}(T_i) = h_s(T_i) - \sum_{k=1}^{n}\omega_{nk}(T_i)x_{ks}$. It follows from $x_{is} = h_s(T_i) + u_{is}$ that the (s, m) element of $\widetilde{X}^T\widetilde{X}$ $(s, m = 1, \ldots, p)$ is

$$\begin{aligned}\sum_{i=1}^{n}\tilde{x}_{is}\tilde{x}_{im} &= \sum_{i=1}^{n}u_{is}u_{im} + \sum_{i=1}^{n}\bar{h}_{ns}(T_i)u_{im}\\ &\quad + \sum_{i=1}^{n}\bar{h}_{nm}(T_i)u_{is} + \sum_{i=1}^{n}\bar{h}_{ns}(T_i)\bar{h}_{nm}(T_i)\\ &\stackrel{\text{def}}{=} \sum_{i=1}^{n}u_{is}u_{im} + \sum_{q=1}^{3}R_{nsm}^{(q)}\end{aligned}$$

Lemma A.1 means $R_{nsm}^{(3)} = o(n)$. This fact and Assumption 1 show that $R_{nsm}^{(1)} = o(n)$ and $R_{nsm}^{(2)} = o(n)$ using Cauchy-Schwarz inequality. We therefore complete the proof of the lemma.

The following Lemma is a slight version of Theorem 9.1.1 of Chow and Teicher (1988). We therefore do not give a proof.

Lemma A.3 *Let $\xi_{nk}, k = 1, \ldots, k_n$, be independent random variables with $E\xi_{nk} = 0$, and $E\xi_{nk}^2 = \sigma_{nk}^2 < \infty$. Assume that $\lim_{n\to\infty} \sum_{k=1}^{k_n} \sigma_{nk}^2 = 1$ and $\max_{1\le k\le k_n} \sigma_{nk}^2 \to 0$. Then $\sum_{k=1}^{k_n} \xi_{nk} \to^{\mathcal{L}} N(0,1)$ in distribution if and only if*

$$\sum_{k=1}^{k_n} E\xi_{nk}^2 I(|\xi_{nk}| > \delta) \to 0 \quad \text{for any } \delta > 0 \quad \text{as } n \to \infty.$$

Lemma A.4 *Let $V_1, \ldots, V_n$ be independent random variables with $EV_i = 0$ and $\inf_i EV_i^2 > C > 0$ for some constant number C. The function $H(v)$ satisfying $\int_0^\infty vH(v)dv < \infty$ such that*

$$P\{|V_k| > v\} \le H(v) \quad \text{for large enough } v > 0 \quad \text{and } k = 1, \ldots, n. \tag{A.3}$$

Also assume that $\{a_{ni}, i = 1, \ldots, n, n \ge 1\}$ is a sequence real numbers satisfying $\sum_{i=1}^n a_{ni}^2 = 1$. If $\max_{1\le i\le n} |a_{ni}| \to 0$, then for $a'_{ni} = a_{ni}/\sigma_i(V)$,

$$\sum_{i=1}^n a'_{ni} V_i \longrightarrow^{\mathcal{L}} N(0,1) \quad \text{as } n \to \infty.$$

Proof. Denote $\xi_{nk} = a'_{nk} V_k, k = 1, \ldots, n$. We have $\sum_{k=1}^n E\xi_{nk}^2 = 1$. Moreover, it follows that

$$\begin{aligned}
\sum_{k=1}^n E\xi_{nk}^2 I(|\xi_{nk}| > \delta) &= \sum_{k=1}^n {a'_{nk}}^2 EV_k^2 I(|a_{nk}V_k| > \delta) \\
&= \sum_{k=1}^n \frac{a_{nk}^2}{\sigma_k^2} EV_k^2 I(|a_{nk}V_k| > \delta) \\
&\le (\inf_k \sigma_k^2)^{-1} \sup_k E\{V_k^2 I(|a_{nk}V_k| > \delta)\}.
\end{aligned}$$

It follows from the condition (A.3) that

$$\sup_k E\{V_k^2 I(|a_{nk}V_k| > \delta)\} \to 0 \quad \text{as } n \to \infty.$$

Lemma A.4 is therefore derived from Lemma A.3.

REFERENCES

Blanchflower, D.G. and Oswald, A.J. (1994). *The Wage Curve*, MIT Press Cambridge, MA.

Chen, H. (1988). Convergence rates for parametric components in a partly linear model. *Annals of Statistics*, **16**, 136-146.

Chow, Y. S. and Teicher, H. (1988). *Probability Theory.* 2nd Edition, Springer-Verlag.

Cuzick, J. (1992a). Semiparametric additive regression. *Journal of the Royal Statistical Society, Series B*, **54**, 831-843.

Cuzick, J. (1992b). Efficient estimates in semiparametric additive regression models with unknown error distribution. *Annals of Statistics*, **20**, 1129-1136.

Engle, R. F., Granger, C.W.J., Rice,J. and Weiss, A. (1986). Semiparametric estimates of the relation between weather and electricity sales. *Journal of the American Statistical Association*, **81**, 310-320.

Eubank, R.L., Hart, J.D. and Speckman, P. (1990). Trigonometric series regression estimators with an application to partially linear model. *Journal of Multivariate Analysis*, **32**, 70-83.

Gao, J. T., Hong, S.Y. and Liang, H. (1995). Convergence rates of a class of estimates in partly linear models. *Acta Mathematica Sinica*, **38**, 658-669.

Hall, P. and Carroll, R.J. (1989) Variance function estimation in regression: the effect of estimating the mean. *Journal of the Royal Statistical Society, Series B*, **51**, 3-14.

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

Heckman, N.E. (1986). Spline smoothing in partly linear models. *Journal of the Royal Statistical Society, Series B*, **48**, 244-248.

Liang, H. (1995). A note on asymptotic normality for nonparametric multiple regression: the fixed design case. *Soo-Chow Journal of Mathematics,* 395-399.

Liang, H. and Härdle, W. (1997). *Asymptotic properties of parametric estimation in partially linear heteroscedastic models.* Discussion paper no. 33, Sonderforschungsbereich 373, Humboldt-Universität zu Berlin.

Müller, H. G. and Stadtmüller, U. (1987). Estimation of heteroscedasticity in regression analysis. *Annals of Statistics*, **15,** 610-625.

Rendtel, U. and Schwarze, J. (1995). *Zum Zusammenhang zwischen Lohnhoehe und Arbeitslosigkeit: Neue Befunde auf Basis semiparametrischer Schaetzungen und eines verallgemeinerten Varianz-Komponenten Modells.* German Institute for Economic Research (DIW) Discussion Paper 118, Berlin.

Robinson, P.M. (1988). Root-N-consistent semiparametric regression. *Econometrica*, **56**, 931-954.

Schick, A. (1993). On efficient estimation in regression models. *Annals of Statistics*, **21**, 1486-1521.

Schick, A. (1996). Weighted least squares estimates in partly linear regression models. *Statistics & Probability Letters*, **27**, 281-287.

Speckman, P. (1988). Kernel smoothing in partial linear models. *Journal of the Royal Statistical Society, Series B*, **50**, 413-436.

Willis, R. J. (1986). *Wage Determinants: A Survey and Reinterpretation of Human Capital Earnings Functions in: Ashenfelter, O. and Layard, R. The Handbook of Labor Economics*, **1**, North Holland-Elsevier Science Publishers Amsterdam, 525-602.

XploRe Learning Guide,p 285-305, Springer Verlag, Heidelberg

11 Finance

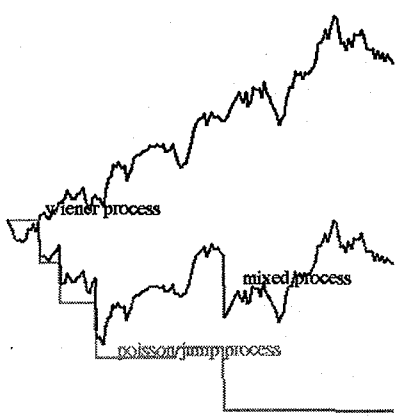
Stefan Sperlich and Wolfgang Härdle

There is growing interest in quantifying and simulating economic processes, particularly in the statistical analysis of the behavior of financial markets. The library `finance` is designed for this purpose. This chapter explains and illustrates the use of XploRe for theory and practice in this setting.

The `finance` library offers functions to predict, to simulate and to estimate time series processes as for example stock returns, to determine option prices and to evaluate different scenarios (e.g. for portfolio strategies). To give a survey of the library we will present the principal procedures implemented in the finance library and illustrate their use with several graphics.

Before starting to work with the finance library in XploRe you have to load all the functions contained in the library by typing the command

```
library("finance")
```



11.1 Outline of the Theory

In the last decades there has been a growing interest in the behavior of financial markets. Due to the increasing globalization of markets, they began to play a central role in international business and economic decision making. Thus, the meaning of "risk" became the central theme in this context.

Risk management is essential in a modern market economy. Financial markets enable firms and households to select an appropriate level of risk in their transactions, by redistributing risks towards other agents who are willing and able to assume them. Markets for options, futures and other so-called derivative securities — derivatives, for short — have a particular status. Futures allow agents to hedge against upcoming risks; such contracts promise the future delivery of a certain item at a certain price. As an example, a firm might decide to engage in copper mining after determining that the metal to be extracted can be sold in advance at the futures market for copper. The risk of future movements in the copper price is thereby transferred from the owner of the mine to the buyer of the contract. Due to their design, options allow agents to hedge against one-sided risks; options give the right, but not the obligation, to buy or sell something at a prespecified price in the future.

In avoiding the risk of long positions one could for example try to hedge the risk by going short in options on the corresponding asset and adapting the proportion held in assets and short-selled options according to the underlying price process of that asset. Therefore, formulas for the pricing of those derivative securities generated a lot of practical and theoretical interest.

Already in the year 1900, Bachelier introduced Brownian motion as a model for price fluctuations on a speculative market. In 1973, Black and Scholes founded their famous option pricing formula which calculates the "fair price" of an option (which means that there is no arbitrage). This has generated a lot of theoretical work relying on that basic model.

11.1.1 Some History

The valuation of derivatives has a long history. One of the earliest endeavors was undertaken by Louis Bachelier (thesis at Sorbonne, 1900). But his formula was based on such assumptions as zero interest rate, and a process that allowed for a negative share price.

This formula was improved by Case Sprenkle, James Boness and Paul Samuelson. They assumed that stock prices are log-normally distributed, guaranteeing that share prices are positive, and allowed for a nonzero interest rate. They also assumed that investors are risk averse and demand a risk premium additionally to the interest rate. In 1964, Boness suggested a formula that came close to the Black–Scholes formula, but still relied on an unknown interest rate, which included compensation for the risk associated with the stock.

Further attempts at valuation (before 1973) basically determined the expected value of a stock option at expiration and discounted its value back to the time of evaluation. Unfortunately, those approaches require taking a stance on which risk premium to use in the discounting. But assigning a risk premium is not straightforward, since it should reflect not only the risk for changes in the stock price, but also the investors attitude towards risk. The latter is hard or impossible to observe in reality.

11.1.2 The Black–Scholes Formula

A commonly used model for the description of fluctuations of asset prices is the following. $X(.)$ denotes the price process which is assumed to be the solution of the stochastic differential equation

$$dX(t) = s(t,x)\,dW(t) + m(t,x)\,dt.$$

Here $W(.)$ denotes Brownian motion, $s(.,X)$ is the volatility process and $m(.,X)$ is the trend or drift. Classical models suppose that $s(t,X) = \sigma X(t)$ and $m(t,X) = \mu X(t)$ which results in geometric Brownian motion.

Fischer Black, Robert Merton and Myron Scholes developed a new method of determining the value of derivatives. Their work (in the early 1970s) solved a longstanding problem in financial economics and has provided ways of dealing with financial risk, both in theory and in practice. Further, their methodology has proven general enough for a wide range of applications. It can thus be used to value not only the flexibility of physical investment projects but also insurance contracts and guarantees.

In the press release, when Scholes and Merton were awarded the Nobel Prize in 1997, was given the following example: Consider a European call option at a strike price of \$100 in three months. (A European option gives the right to buy or sell only at a certain date, whereas a so-called American option gives the same right at any point in time up to a certain date.) Clearly, the value of this

call option depends on the current share price; the higher the share price today the greater the probability that it will exceed \$100 in three months, in which case it will pay to exercise the option. A formula for option valuation should thus determine exactly how the value of the option depends on the current share price. How much the value of the option is altered by a change in the current share price is called the "delta" of the option — see also the `greeks`.

Assume that the value of the option increases by \$1 when the current share price goes up \$2 and decreases by \$1 when the stock goes down \$2. Assume also that an investor holds a portfolio of the underlying stock and wants to hedge against the risk of changes in the share price. He can then construct a risk-free portfolio by selling twice as many options as the number of shares he owns. For reasonably small increases in the share price, the profit the investor makes on the shares will be the same as the loss he incurs on the options, and vice versa for decreases in the share price. As the portfolio thus constructed is risk free, it must yield exactly the same return as a risk-free three-month treasury bill. If it did not, arbitrage trading would begin to eliminate the possibility of making risk-free profits. As the share price is altered over time and as the time to maturity draws nearer, the delta of the option changes. In order to maintain a risk-free stock-option portfolio, the investor has to change its composition.

Black, Merton and Scholes assumed that such trading can take place continuously without any transaction costs. The condition that the return on a risk-free stock-option portfolio yields the risk-free rate, at each point in time, implies a partial differential equation, the solution of which is the Black–Scholes formula for a call option:

$$C = SN(d) - Le^{-rt}N\left(d - \sigma\sqrt{t}\right)$$

where $N()$ is the standard normal distribution, S, t, r, L see below and d is defined by

$$d = \frac{\log(S/L) + (r + \sigma^2/2)t}{\sigma\sqrt{t}}.$$

According to this formula, the value of the call option C is given by the difference between the expected share price — the first term on the righthand side — and the expected cost — the second term — if the option is exercised. The higher the option value, the higher the current share price S, the higher the volatility of the share price σ, the higher the risk-free interest rate r, the longer the time to maturity t, the lower the strike price L, and the higher the probability that the option will be exercised — see also the quantlet influence.

All the parameters in the equation can be observed except sigma, which has to be estimated from market data. Alternatively, if the price of the call option is known, the formula can be used to solve for the market-implied volatility. Market equilibrium is not necessary for option valuation; it is sufficient that there are no arbitrage opportunities. The method described in the example above is based precisely on the absence of arbitrage. It generalizes to valuation of other types of derivatives. Mertons 1973 article included the Black–Scholes formula and some generalizations, for instance, he allowed the interest rate to be stochastic. The theory of Merton, Black and Scholes can also be used for many other or related fields such as:

- **Corporate liabilities**
  Black, Merton and Scholes realized already in 1973 that a share can be interpreted as an option on the whole firm. When loans mature and the value of the firm is lower than the nominal value of debt, the shareholders have the right, but not the obligation, to repay the loans. The method can thus be used for determining the value of shares, which can be important if the shares are not traded. Since other corporate liabilities are also derivative instruments (whose value, too, depends on the value of the firm), they can be valued using the same method.
- **Investment evaluation**
- **Guarantees and insurance contracts**
- **Complete markets**

11.2 Assets

There exist several quantlets for the simulation and estimation of asset prices. Implemented in the finance library are

`stocksim ()`
simulates random processes for stock prices

`stockest (data)`
estimates a diffusion model for stock price data

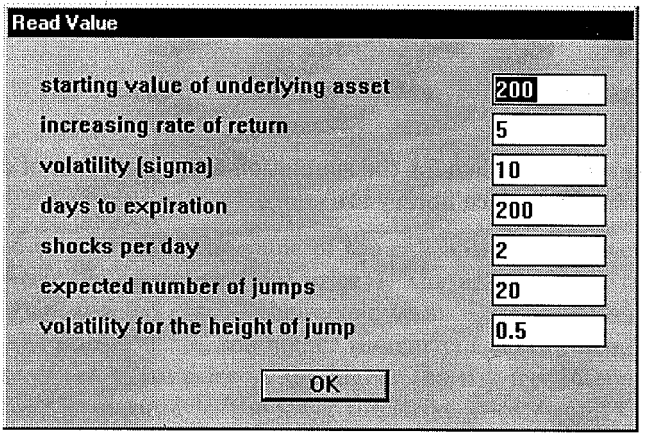
`stockestsim (data)`
simulates and estimates a Wiener process with Poisson jump

11.2.1 Stock Simulation

The quantlet `stocksim` simulates random processes for a stock price in three different ways:

- using a Wiener process,
- using a Poisson jump process,
- using a mixture of both.

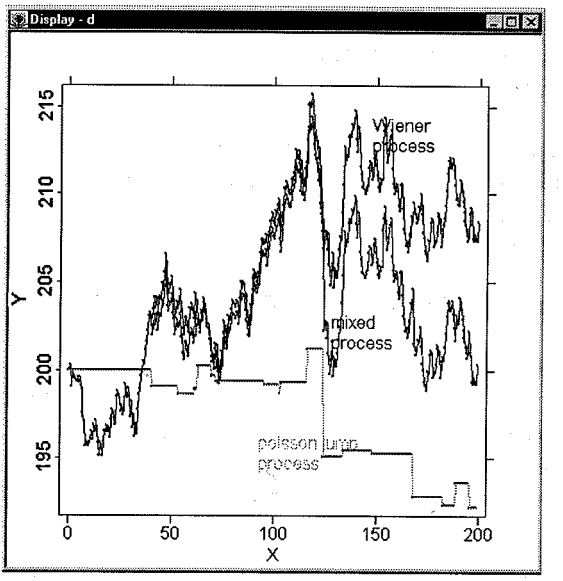
It is invoked by typing `stocksim()` An interactive window appears which asks for the values of the process to be simulated.



The function returns as output a display plotting the three processes and asks if one wants to repeat the simulation. In the interactive window one is asked for the starting values of the underlying asset, the increasing rate of return which corresponds to $m(t,x)$ in the underlying diffusion process. The volatility parameter σ corresponds to a constant $s(t,x)$. The expected number of jumps is the parameter for the underlying jump Poisson process. More precisely the geometric Brownian motion

$$\frac{dX(t)}{X(t)} = \mu dt + \sigma dW(t) \tag{11.1}$$

is simulated with an overlayed Poisson jump process. If we set the last two parameters (intensity and height of the jump) equal to zero we exactly simulate from (11.1) at discrete points. The first parameter equals $X(0)$ and the second is the drift μ. The volatility is given by b, the third parameter. If T^* denotes the days to expiration and nd is the observation frequency per day, the process is on the time interval $T^*/365 \subset [0,1]$ on exactly $T = ndT^*$ discretization points. This may be checked by variation of these parameters. The process is recursively calculated as $X(t+1) = X(t)\exp(\mu/nd + \sigma W(t)/\sqrt{nd})$ at the points $t = 1, 2, \ldots, T$.



11.2.2 Stock Estimation

The `stockest` quantlet assumes that the underlying diffusion processes models are the same as under 11.2.1, i.e. a mixture of a Poisson jump and a Wiener process with drift. For the estimation of such a process, we have to choose a dataset that we want to examine. Let's estimate the parameters for the price process of the Motorola stock. The data is loaded into XploRe by typing

```
data=read("motorola")
```

in the command line of the XploRe input window. The data consists of 591 observations. It has 6 columns. We choose the second column — which simply contains the price notations of the stock — with the command `data=data[,2]` Estimation now takes place by executing `stockest`. This quantlet is executed by typing the name of the variable representing the dataset in parentheses:

```
stockest(data)
```

Now the corresponding parameters of the model are displayed in the XploRe output window. As an example, take the estimation of the volatility: In the output window you find the following information:

```
Content of object _tmp.sigma2
[1,]    38.819
```

The other estimated parameters are `mue`, the increasing rate of return, `sigma` the volatility of returns, `lambda` the number of jumps in the Poisson model and `jump` the volatility of the height of the jump.

11.2.3 Stock Estimation and Simulation

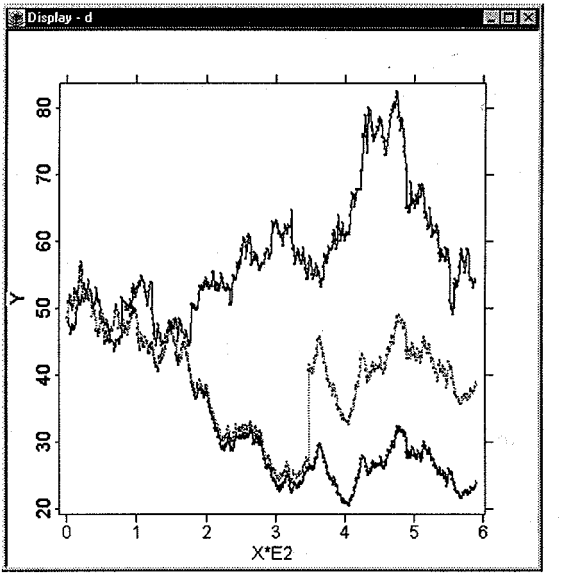
The quantlet `stockestsim` is a combination of the quantlets described in Subsections 11.2.1 and 11.2.2. At first it estimates with the first part of a given dataset the parameters of a random process. This is done for two kinds of models: a Wiener process and a combination of a Wiener and a Poisson jump process. Then both models are compared by a simulation with the rest of the real dataset.

As in the quantlet `stockest`, you need to choose the dataset first and then execute the function by putting the dataset as input parameter. This is done

in XploRe by typing the following sequence of commands in the command line of the input window:

```
data=read("motorola")
data=data[,2]
stockestsim(data)
```

The result is a graphical display showing the three processes:



XploRe Learning Guide,p 285-305, Springer Verlag, Heidelberg

11.3 Options

11.3.1 Calculation of Option Prices and Implied Volatilities

A calculation of option prices is possible by using one of the following functions:

`optstart ()`
starting program to calculate option prices or implied volatilities

`bitree (vers, task)`
calculates option prices using the Binomial tree

`{opvv,sel,ingred} = bs1 (task)`
calculates option prices using the Black–Scholes formula

`mcmillan (eopv, sel, task, ingred)`
calculates option prices using the McMillan formula

`american ()`
starting program to calculate option prices for american options

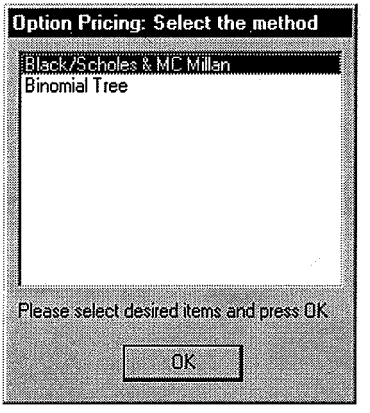
`european ()`
starting program to calculate option prices or implied volatilities for european options

`asset (vers)`
auxiliary quantlet to calculate option prices for american options

The interactive option pricing quantlet `optstart` is simply invoked by typing

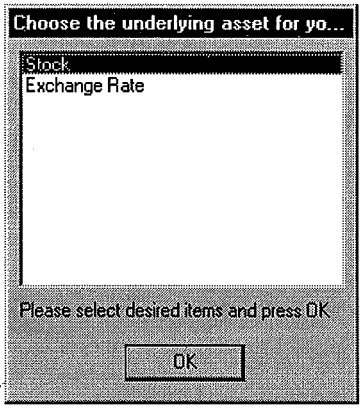
```
optstart()
```

in the XploRe command line. A selection box appears which starts the interactive option pricing procedure.

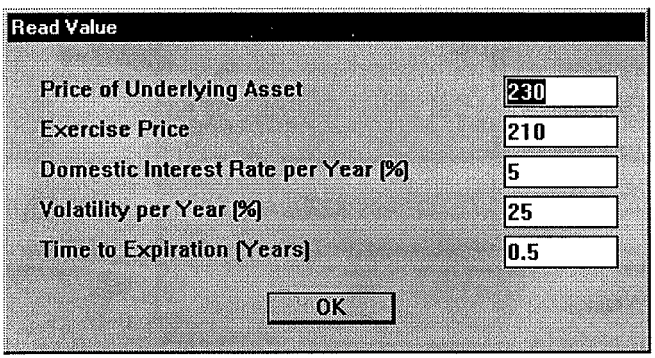


Simply select the method you want to use. If you wish to calculate the option price analytically, choose Black/Scholes & MC Millan; if you want XploRe to calculate it numerically, choose Binomial Tree. Let's choose Black/Scholes & MC Millan. In any case you will be asked whether you want to compute the price of an European (an option which can be executed only at a given date) or an American option (that can be executed anytime).

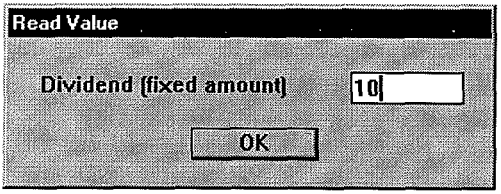
In this example we have chosen American. This is the kind of option that is usually traded e.g. in the USA or in Germany. The next decision is about the underlying asset (stock or exchange rate).



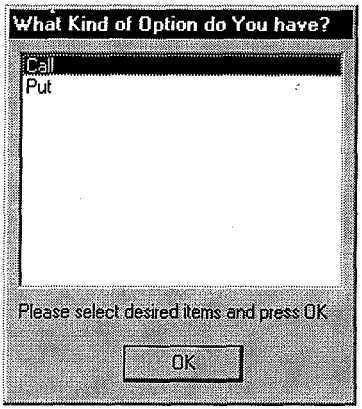
In our example we are regarding a stock as the underlying asset. In this setting you are questioned if you like to have dividends included in the stock and of what kind you want them to be. (If you choose Exchange Rate here, the next two menu items will be skipped.) Then you are asked whether you like to compute the price of an option or the implied volatility. Now we are ready to enter the parameters needed for the computation of the option prices. These are Price of the Underlying Asset, Exercise Price, Domestic Interest Rate per Year and Volatility per Year in percent as well as the Time to Expiration in years.



In case you have chosen a dividend payment, one more window will appear where you are asked to put the amount of the dividend.



Finally you can choose the kind of option you like to calculate. Let's say we wanted to know the price of a call option:



The price of our American call option on the given stock in the scenario (chosen through the corresponding parameters) with fixed dividend is now displayed in the XploRe output window. In case you have chosen a stock as underlying asset even the price of the European call option is displayed (in case you have not chosen a dividend, the price of a European call option equals that of an American call option):

```
[1,]
[2,] ------------------------------------
[3,]  The Price of Your European Call-Option
[4,]  on Given Stock with fixed Dividend is
[5,]  27.3669
[6,] ------------------------------------
[7,]
```

```
[1,]
[2,] ------------------------------------
[3,]  The Price of Your American Call-Option
[4,]  on Given Stock with fixed Dividend is
[5,]  27.4641
[6,] ------------------------------------
[7,]
```

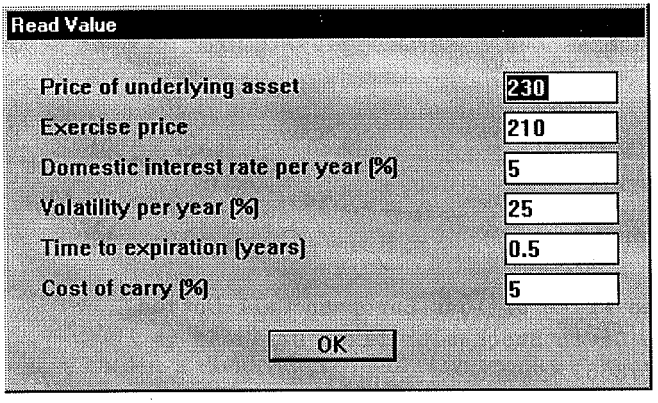
11.3.2 Option Price Determining Factors

`influence ()`
displays the influence of price determining parameters on options

The quantlet `influence` measures and visualizes the influence of different factors on the prices of options. It is simply started by typing

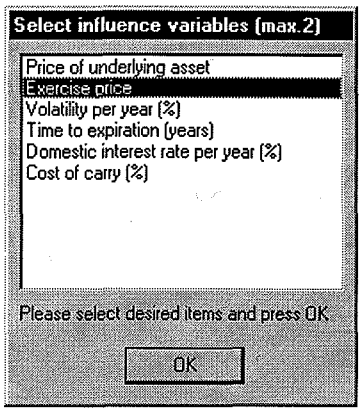
```
influence()
```

in the command line of XploRe. The option prices are calculated with the Black–Scholes formula. After starting the quantlet the following window appears:

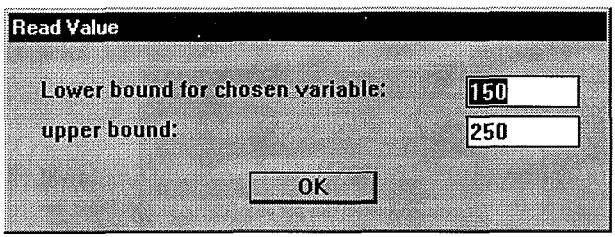


You may enter the different parameters needed to simulate the diffusion process. Next select the influence variables — you may select up to two variables. The

following example demonstrates the use of just one variable:

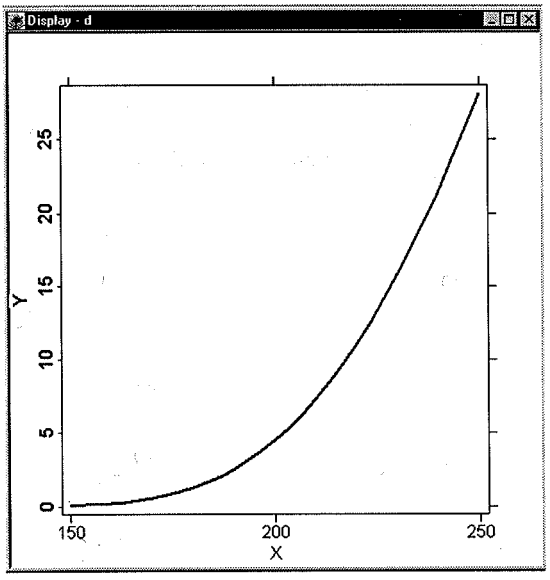


In this example we would like to calculate the influence of the exercise price on the option price. You must set the lower and upper bound for your chosen variable.

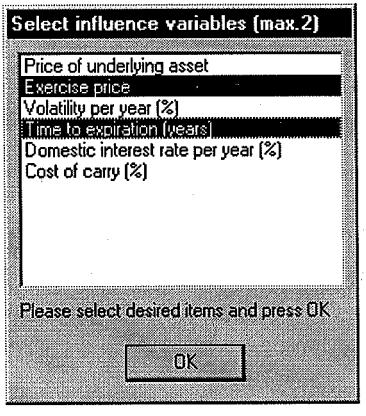


After pushing the OK button you will be asked for what kind of option the influence is to be calculated.

If you choose for example a Put option you will obtain the following graph which shows the influence of the factor (exercise price) on the price of the option:



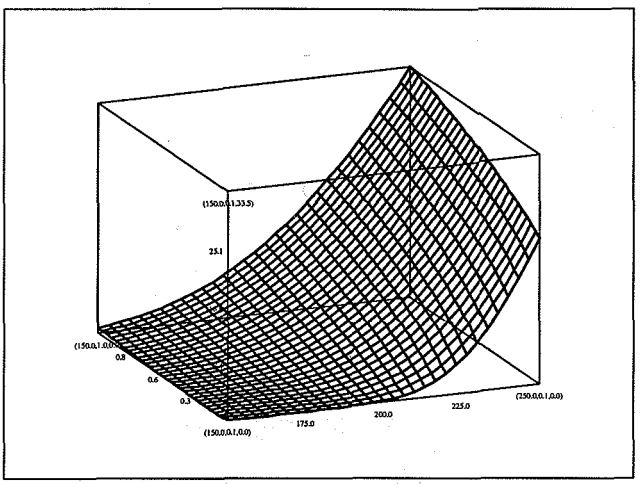
Using `influence` you can also select two variables as the following example demonstrates:



After selecting the two variables you wish to compute, XploRe asks you to set the lower and upper bound for both variables.

If you choose e.g. Put you will obtain a three-dimensional graphic with the two selected influence factors (exercise price and time to expiration) and the price

of the option. You may turn the graphic around by using the cursor buttons.



11.3.3 Greeks

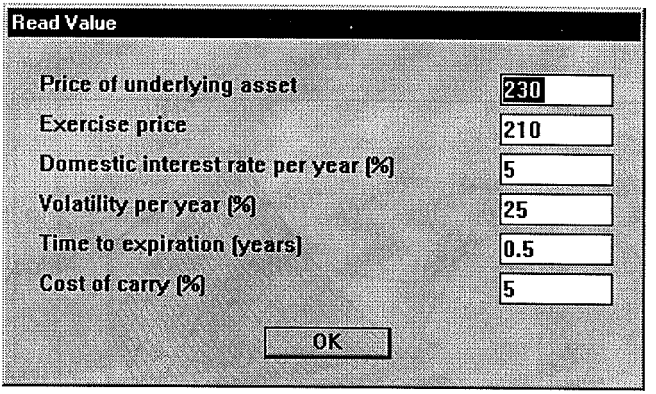
`greeks ()`
calculates and displays the different indices which are used for trading with options

The interactive function `greeks` calculates and displays the different indices used for analyzing and trading with options. You start it by

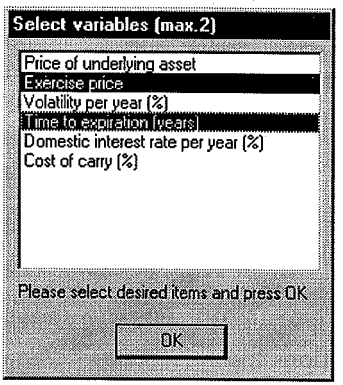
```
data=greeks()
```

The first step is to enter the asset's basic data:

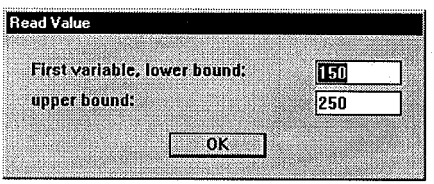
XploRe Learning Guide,p 285-305, Springer Verlag, Heidelberg

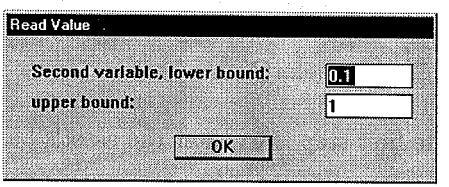


Next, you have to select the variables you want to analyze (at most two), e.g.

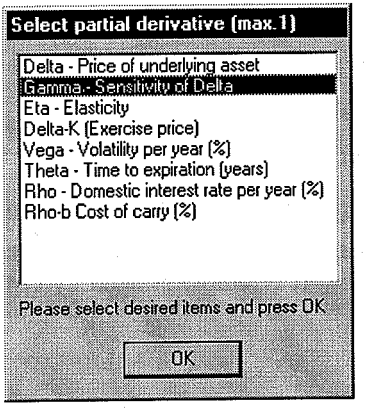


Select the ranges for the values of the chosen variables:

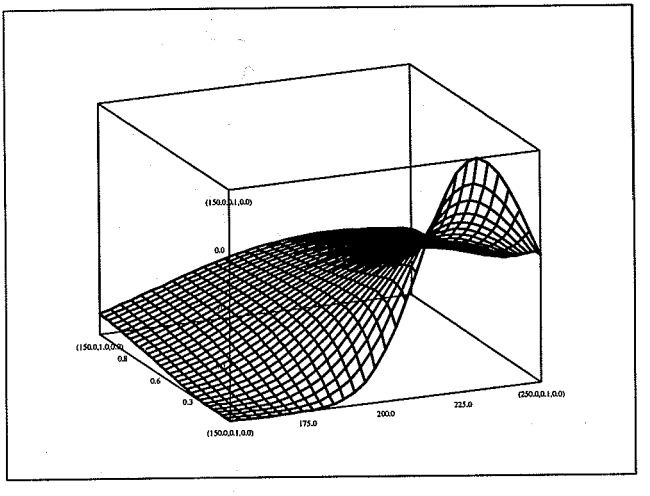




Now you can choose the index you are interested in:



After telling the program the kind of option you want, the quantlet `greeks` will produce a graphical output window for your result:



XploRe Learning Guide,p 285-305, Springer Verlag, Heidelberg

11.4 Portfolios and Hedging

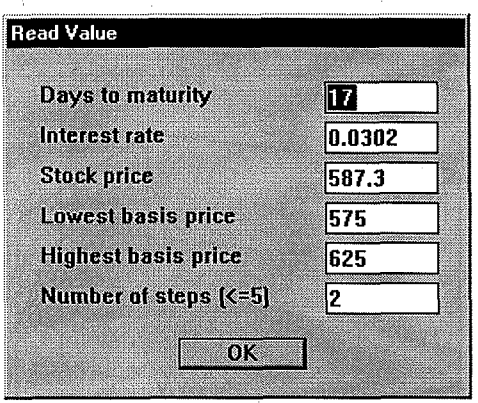
11.4.1 Calculation of Arbitrage

```
arbitrage ()
      calculates an arbitrage table
```

The function `arbitrage` calculates an arbitrage table considering puts and calls with the same strike price. It is simply started by typing

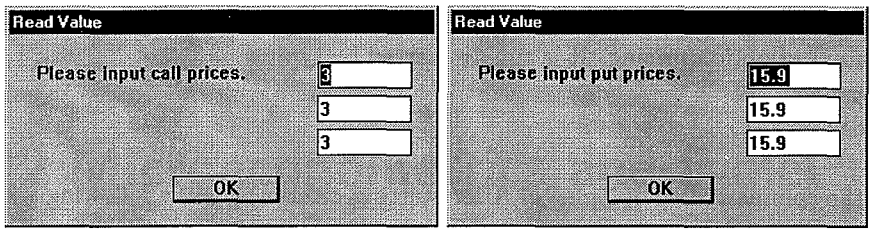
```
arbitrage()
```

in the command line of XploRe. After starting arbitrage the following window will appear:



Here you are asked to put in your given data — Days to maturity, Interest rate, Stock price, Lowest basis price, Highest basis price and the Number of steps.

After pushing the OK button you can first put in the call prices and right afterwards the put prices:



As a result XploRe presents you the following table of arbitrage in the output window, where

- `Call_price` is the vector of call prices
- `Put_price` is the vector of put prices
- `Basis` is the vector of basis prices
- `Stock_flow` is the amount we pay/get for buying/selling stock
- `Call_flow` is the amount we pay/get for buying/selling call option
- `Put_flow` is the amount we pay/get for buying/selling put option
- `Bank_flow` is the investment to/loan from a bank
- `Arbitrage` is the vector of arbitrage gains/losses

```
[ 1,] Stock price:     587.30
[ 2,] Interest rate:     0.0302
[ 3,] Days to maturity: 17.00
[ 4,]
[ 5,] Call_price Put_price Basis  Stock_flow Call_flow Put_flow  Bank_flow  Arbitrage
[ 6,] ------------------------------------------------------------------------------
[ 7,]       3.00     15.90  575.00    587.30    -3.00    15.90    -574.18      26.02
[ 8,]       3.00     15.90  600.00    587.30    -3.00    15.90    -599.15       1.05
[ 9,]       3.00     15.90  625.00    587.30    -3.00    15.90    -624.11     -23.91
[10,]
```

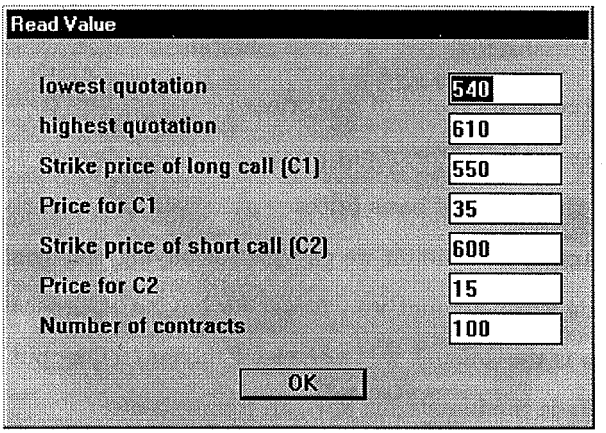
11.4.2 Bull-Call Spreads

`callbull ()`
calculates the results of a Bull-Call Spread for the context of option pricing

The function `callbull` calculates the results of a Bull-Call Spread for the context of option pricing. It is simply started by typing

```
callbull()
```

in the command line of XploRe. After starting `callbull()` the following window will appear:



After putting in all the basic data required just push the OK button. XploRe will calculate the results and present them in the XploRe output window in the following way:

```
[ 1,]
[ 2,]   Stock price    long Call   short Call    gain/loss
[ 3,] ----------------------------------------------------
[ 4,]        540.00     -3500.00      1500.00     -2000.00
[ 5,]        550.00     -3500.00      1500.00     -2000.00
[ 6,]        560.00     -2500.00      1500.00     -1000.00
[ 7,]        570.00     -1500.00      1500.00         0.00
[ 8,]        580.00      -500.00      1500.00      1000.00
[ 9,]        590.00       500.00      1500.00      2000.00
[10,]        600.00      1500.00      1500.00      3000.00
[11,]
```

XploRe Learning Guide,p 247-270, Springer Verlag

9 Time Series

Petr Franěk and Wolfgang Härdle

The purpose of this chapter is to show how XploRe may be used by practitioners for analyzing observed time series.

Some of the time series tools are standard in the literature. The more elaborated nonlinearity tests based on artificial neural networks are implemented for the nonadvanced use.

9.1 Time Domain and Frequency Domain Analysis

`x = acf (y {,k})`
computes the autocorrelation function of a time series

`acfplot(y {,k})`
plots the autocorrelation function of a time series

`x = pacf (y,k)`
computes the partial autocorrelation function of a time series

`pacfplot(y {,k})`
plots the partial autocorrelation function of a time series

`x = pgram (x { opt})`
computes and plots the raw (log) periodogram of a time series

`x = spec (y {,width {,opt}})`
estimates and plots the spectral density of a time series

(1999) Franek, P. and Härdle, W. Time Series.

```
timeplot(y {,len {,header}})
```
plots a time series in multiple windows with user-specified maximum length per window

A time series represents a path (realization) of a stochastic process $\{Y_t\}$. The subscript t ($t = 1, \ldots, T$) is usually understood as a time index (e.g. days or years). In this section we will consider that the underlying stochastic process is **weakly stationary**, i.e. we will assume that it satisfies the conditions

$$\begin{aligned} E|Y_t|^2 &< \infty \quad \forall t, \\ EY_t &= \mu \quad \forall t, \\ E(Y_t - \mu)(Y_s - \mu) &= E(Y_{t+\tau} - \mu)(Y_{s+\tau} - \mu) \quad \forall s, t, \tau. \end{aligned}$$

All functions for the time series analysis are part of the `times` library and become available after loading this library:

```
library("times")
```

To display a time series we use `timeplot`. The following example plots the first 250 observations of the time series `dmus58` (Deutschmark–Dollar FX rates in 1982):

```
y = read("dmus58")
y = y[1:250,]
timeplot(y)
```

`times01.xpl`

The resulting graph is shown in Figure 9.1.

9.1.1 Autocovariance and Autocorrelation Function

The sample **autocovariance** function at lag τ of a process Y_t is defined for $\tau \in \{0, \ldots, T-1\}$ as

$$\hat{\gamma}(\tau) = T^{-1} \sum_{t=\tau+1}^{T} (Y_t - \overline{Y})(Y_{t-\tau} - \overline{Y})$$

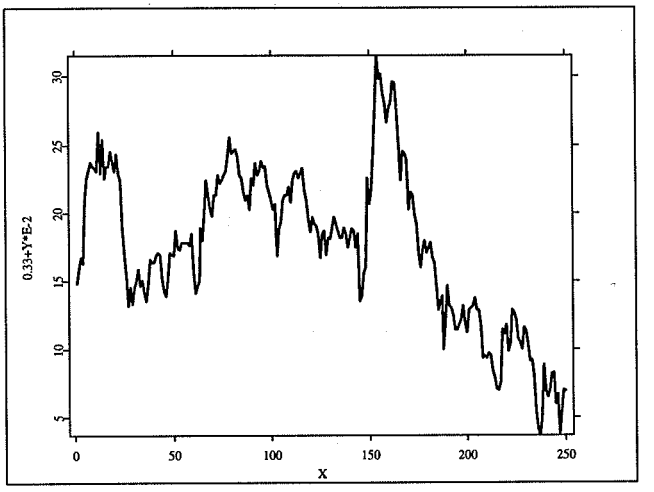


Figure 9.1. First 250 observations of dmus58.

where $\overline{Y} = T^{-1}\sum_{t=1}^{T} Y_t$ is the arithmetic mean of the time series Y_t. We define the sample **autocorrelations** $\widehat{\rho}(\tau)$ of the process by standardizing the sample autocovariance function $\widehat{\gamma}(\tau)$ by $\widehat{\gamma}(0)$ (the sample variance of the process), i.e.

$$\widehat{\rho}(\tau) = \frac{\widehat{\gamma}(\tau)}{\widehat{\gamma}(0)}.$$

Let's consider a sample of 500 independent realizations of a standard normal random variable:

```
randomize(0)
y = normal(500)
```

times02.xpl

Using acf we can evaluate the sample autocorrelations of the generated series. In the following example the result is stored in the vector x and the first five autocorrelations are displayed:

```
x = acf(y)
x[1:5]
```

The shape of the autocorrelation function may be easily analyzed using the function `acfplot` which displays the **correlogram**, i.e. the graph of the autocorrelation function $\widehat{\rho}(\tau)$. Type

```
acfplot(y)
```

to get the following graph in Figure 9.2. Confidence levels $\pm 2/\sqrt{T}$ are plotted to easily check the assumption that the series is a white noise.

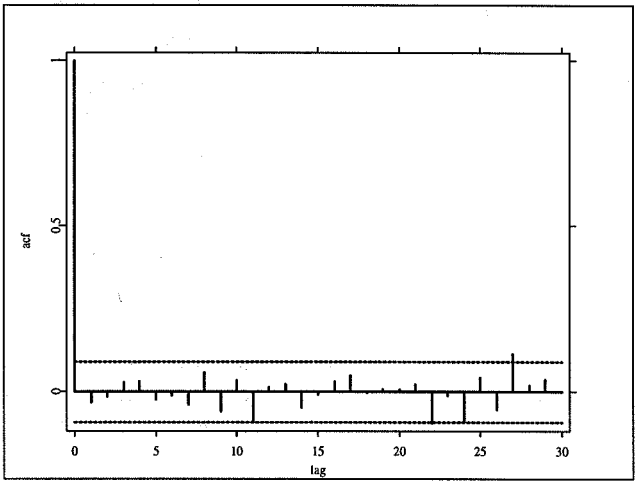


Figure 9.2. Autocorrelation function.

As another useful measure, the **partial autocorrelations** are implemented in `pacf`. The partial autocorrelation of order $\tau \geq 2$ is calculated as the correlation of the two residuals obtained after regressing $Y_{\tau+1}$ and Y_1 on the intermediate observations $Y_2, \ldots, Y_\tau$, the partial autocorrelation at lag 1 being defined as the correlation between Y_1 and Y_2.

XploRe also provides the function `pacfplot` as a partial autocorrelation equivalent to `acfplot`. The usage of these two quantlets is similar to the previously mentioned autocorrelation function. To evaluate the sample partial autocorrelations of the generated series and plot the evaluated values type the following instructions:

```
x = pacf(y,5)
```

```
x[1:5]
pacfplot(y)
```

The resulting plot is shown in Figure 9.3.

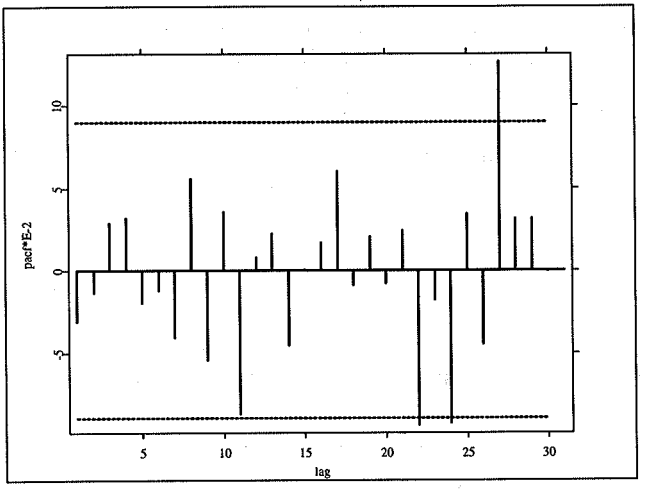


Figure 9.3. Partial autocorrelation function.

9.1.2 The Periodogram and the Spectrum of a Series

The frequency domain analysis is concerned with the decomposition of the observed series into periodic components. The main analytical tool for spectral analysis is the **spectrum** of a series defined as

$$f(\lambda) = (2\pi)^{-1} \left[\gamma(0) + 2 \sum_{\tau=1}^{\infty} \gamma(\tau) \cos(\lambda\tau) \right],$$

with λ the angular frequencies in $[-\pi, \pi]$ and $\gamma(\tau)$ the theoretical autocovariances. Since the spectrum is symmetric around zero, the analysis is restricted to the range of frequencies in $[0, \pi]$. For a sample of T observations, we consider the harmonic frequencies, or Fourier frequencies $\lambda_j = 2\pi j/T$, $j = 1, \ldots, [T/2]$.

The sample counterpart of the spectrum is the periodogram, defined as

$$I(\lambda) = (2\pi)^{-1} \left[\widehat{\gamma}(0) + 2 \sum_{\tau=1}^{T-1} \widehat{\gamma}(\tau) \cos(\lambda\tau) \right].$$

The periodogram is not a consistent estimator of the spectrum. When the sample size tends to infinity, more frequencies are considered without adding more precision on a particular one. The function `pgram` computes and plots the periodogram of the series. XploRe computes the periodogram for the frequencies $k_j \in [0, 0.5]$, which are linked to the angular frequencies by the relationship $k_j = \lambda_j/(2\pi)$. The periodogram of our generated series of independent normal random variables Y_t can be obtained as follows:

```
z = pgram(y)
```

`times03.xpl`

The periodogram for the series `y` is computed, the result is stored in the variable `z` and the periodogram is displayed on the screen (Figure 9.4.)

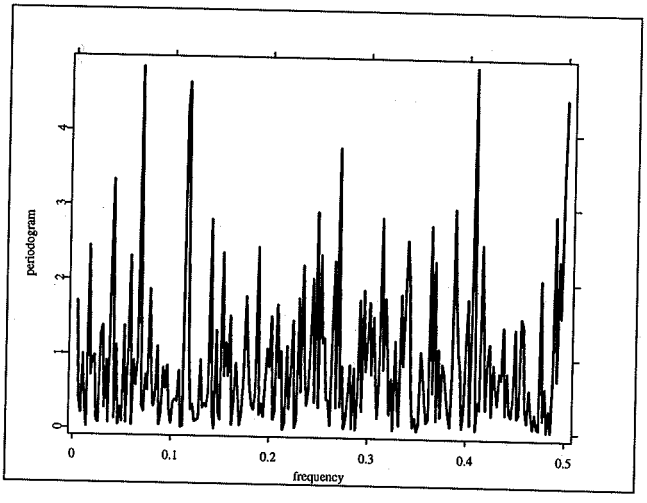


Figure 9.4. Periodogram.

The jagged shape of the periodogram illustrates the typical feature of a white noise series, i.e. a series of independent and identically distributed random

variables. The lack of smoothness of the periodogram makes it difficult to interpret. In order to estimate the spectrum, the periodogram can be smoothed using the function spec that will be introduced in more details in the following section.

9.2 Linear Models

```
est = armacls (y, p, q)
```
estimates parameters of an ARMA process using the conditional sum of squares

```
est = armalik (y)
```
estimates parameters of an ARMA(1,1) process using the maximum likelihood

```
y = genar (eps, startval, phi)
```
generates an AR process

```
y = genarma (a, b, eps)
```
generates an ARMA process with zero mean

In this section we focus our attention on the class of **linear models**, i.e. models driven by a general dynamic relationship of the form

$$Y_t = g(Y_{t-1}, Y_{t-2}, \ldots, Y_{t-p}),$$

where the function $g(.)$ is assumed to be linear.

9.2.1 Autoregressive Models

A process Y_t is called an autoregressive process of order p, AR(p), if it is driven by the relation

$$Y_t = \phi_1 Y_{t-1} + \phi_2 Y_{t-2} + \ldots + \phi_p Y_{t-p} + \varepsilon_t,$$

where the ε_t form a white noise process, they are also called **innovations**.

A sample of n observations of an AR(p) process can be generated using the command genar, which has the following syntax:

```
y = genar(eps, starval, phi)
```

where

- `eps` is the n-dimensional vector of white noises,
- `starval` is the p-dimensional vector of initial values,
- `phi` is the p-dimensional vector of autoregressive parameters $(\phi_1, \ldots, \phi_p)$.

In the following example a sample of 250 observations of an AR(1) process is generated. In this example, `x` is a vector of 250 independent realizations of standardized normal random variable.

```
randomize (0)
eps = normal(250)
starval = 0
phi = 0.5
y = genar(eps,starval,phi)
```

 `times04.xpl`

In the following example, using the function `spec`, the typical spectrum of an autoregressive process is displayed. Note that the lowest frequencies have the highest contribution to the variation of the process. `spec` also displays the periodogram of the series. Type the instruction

```
spec(y)
```

to obtain the plots in Figure 9.5.

9.2.2 Autoregressive Moving Average Models

A process Y_t is called a moving average process of order q, MA(q), if it is driven by the relation

$$Y_t = \varepsilon_t + \psi_1 \varepsilon_{t-1} + \psi_2 \varepsilon_{t-2} + \ldots + \psi_q \varepsilon_{t-q},$$

where ε_t is a white noise innovation process.

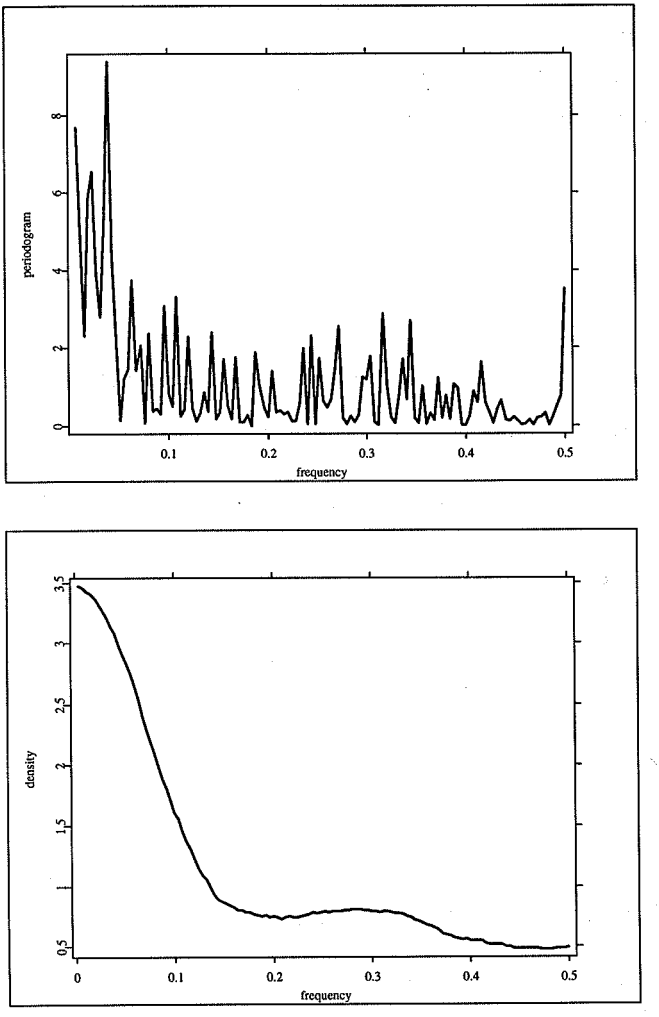


Figure 9.5. Periodogram and spectrum.

The structures of the autoregressive (AR) process and the moving average (MA) process may be combined into an autoregressive moving average process

$$Y_t = \phi_1 Y_{t-1} + \phi_2 Y_{t-2} + \ldots + \phi_p Y_{t-p} + \varepsilon_t + \psi_1 \varepsilon_{t-1} + \psi_2 \varepsilon_{t-2} + \ldots + \psi_q \varepsilon_{t-q}.$$

This process is denoted by ARMA(p, q), where p is the order of the autoregressive part and q is the order of the moving average part.

A sample of n observations of an ARMA(p, q) process can be generated using `genarma`, which has the following syntax:

```
y = genarma(a,b,eps)
```

where

- `a` is the p-dimensional vector of autoregressive parameters $(\phi_1, \ldots, \phi_p)$,
- `b` is the q-dimensional vector of moving average parameters $(\psi_1, \ldots, \psi_q)$,
- `eps` is the n-dimensional vector of white noises.

In the following example a sample of 250 observations of an ARMA$(1, 1)$ with Gaussian innovations process is generated:

```
randomize(0)
a = 0.5
b = 0.3
eps = normal(250)
y = genarma(a,b,eps)
```

`times05.xpl`

9.2.3 Estimating ARMA Processes

The parameters of an ARMA$(1, 1)$ process may be estimated using `armalik`, which has the following syntax:

```
est = armalik(y)
```

where

- `y` is the observed process,

- est is a list containing the estimated parameters, the corresponding asymptotic standard deviations, the asymptotic covariance and the estimate of the white noise variance.

To estimate the parameters of our generated ARMA(1,1) process, type:

```
est1 = armalik(y)
est1{1}
est1{2}
```

Q times06.xpl

As output, XploRe returns

```
Contents of a
[1,]  0.49957
[2,]  0.25991
Contents of stderr
[1,]  0.057633
[2,]  0.064244
```

The parameters of a general ARMA(p, q) process can be estimated by armacls. This quantlet minimizes the conditional sum of squares and has the following syntax:

```
est = armacls(y,p,q)
```

where

- y is the sample of observations,
- p is the order of the autoregressive part,
- q is the order of the moving average part,
- est is a list containing the estimated parameters and information about the convergence of the method.

Since the generated sample y is the realization of an ARMA(1,1) process, we may estimate the parameters of our sample with the following instructions:

```
est2 = armacls(y,1,1)
est2
```

times07.xpl

As a result XploRe shows

```
Contents of est2.y.minimum
[1,]  0.49504
[2,]  0.28265
Contents of est2.y.iter
[1,]       20
Contents of est2.y.converged
[1,]        1
Contents of est2.wnv
[1,]  0.90943
```

9.3 Nonlinear Models

```
z = archest(y, q, p)
```
estimates the parameters of a GARCH model

```
h = archtest (y {,lags {,testform}})
```
test for ARCH effects

```
h = annarchtest (y {,nlags {,nodes {,testform}}})
```
test for ARCH effects based on neural networks

```
y = genarch (a, b, n)
```
generates an GARCH process with Gaussian innovations

```
y = genbil (phi, psi, gamma, noise)
```
generates a bilinear process

```
y = genexpar (thrlag, gamma, phi0, phi1, noise)
```
generates an exponential AR process

```
y = gentar (nr, thrlag, thr, phi, noise)
```
generates a threshold AR process

```
gpplot(x, m, k)
```
returns the Grassberger–Procaccia plot for time series

Nonlinear time series models have been recently explored by many authors. These models became important especially in analyzing financial and economic time series with underlying theoretical models that contain nonlinear relations. Several nonlinear models are implemented in XploRe with special focus on estimating and testing in ARCH and GARCH models which are often used in financial applications.

9.3.1 Several Examples of Nonlinear Models

We speak about a **threshold model** if the parameters of the model depend on the state of the observed system (random process). In building these models, the real line (please note that we are concerned only with univariate time series here) is divided into k parts by the set of ordered values $r_1 < \cdots < r_{k-1}$

(these values are called the threshold parameters). Depending on which interval $(r_j, r_{j+1}]$ contains the value Y_t the jth set of parameters is used to generate Y_{t+d}. The parameter $d < k$ is then called a delay parameter. The zero mean threshold AR(p) process (TAR(p) process) then may be introduced by the equation

$$Y_t = \phi_1^j Y_{t-1} + \phi_2^j Y_{t-2} + \ldots + \phi_p^j Y_{t-p} + \varepsilon_t,$$

where ε_t is a white noise process and $j \in \{1, \ldots, k\}$ is an indicator of the set of parameters to be used, i.e. j is determined by the condition $Y_{t-d} \in (r_j, r_{j+1}]$.

The threshold AR(p) may be generated in XploRe using `gentar`, which has the following syntax:

```
y = gentar(nr,thrlag,thr,phi,noise)
```

where

- `nr` is the number of threshold regions,
- `thrlag` is the threshold lag (the delay parameter),
- `thr` is a `(nr-1)`-dimensional vector of the threshold parameters that separates the regions,
- `phi` is a `(nr*p)`-dimensional vector of the AR parameters for all regions sorted as follows
$$(\phi_1^1, \ldots, \phi_p^1, \ldots, \phi_1^{nr}, \ldots, \phi_p^{nr}),$$
- `noise` is an n-dimensional vector of the noise (n is the number of observations to be generated).

The following example generates and displays 250 observations of the process

$$\begin{aligned} Y_t &= 0.6\, Y_{t-1} + \varepsilon_t \quad \text{for} \quad Y_{t-1} \leq 0 \\ &= 0.4\, Y_{t-1} + \varepsilon_t \quad \text{for} \quad Y_{t-1} > 0 \end{aligned}$$

with $\varepsilon_t \sim N(0, 1)$.

```
y=gentar(2,1,0,#(0.6,0.4), normal(250))
timeplot(y)
```

times08.xpl

The class of **exponential autoregressive** (EAR(p)) models with the lag d is characterized by the relation

$$Y_t = \sum_{j=1}^{p} \left[\phi_j^0 Y_{t-j} + (\phi_j^1 - \phi_j^0) \exp(-\gamma Y_{t-d})\right] + \varepsilon_t,$$

where ε_t is a white noise process.

The EAR(p) process may be generated in XploRe using `genexpar`, which has the following syntax:

```
y = genexpar(thrlag,gamma,phi0,phi1,noise)
```

where

- `thrlag` is the threshold lag (the delay parameter),
- `gamma` is a positive parameter of the exponential function,
- `phi0` is a p-dimensional vector of AR parameters $(\phi_1^0, \ldots, \phi_p^0)$,
- `phi1` is a p-dimensional vector of AR parameters $(\phi_1^1, \ldots, \phi_p^1)$,
- `noise` is an n-dimensional vector of the noise (n is the number of observations to be generated).

The following example generates and displays 250 observations of an EAR(2) process:

```
y=genexpar(1,0.1,0.3|0.6, 2.2|-0.8,normal(250))
timeplot(y)
```

times09.xpl

The resulting time series is shown in Figure 9.6.

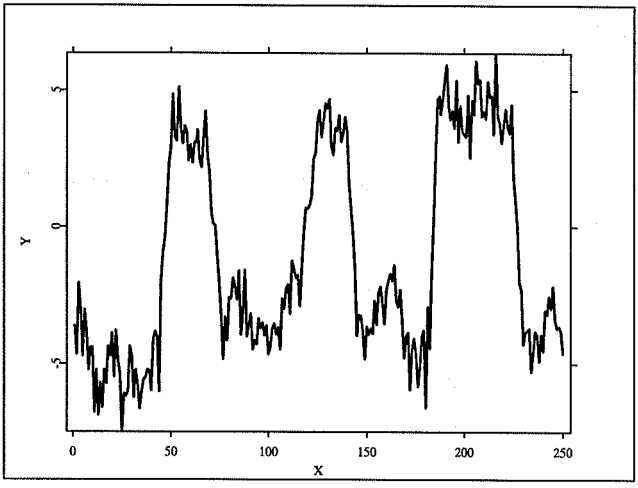


Figure 9.6. 250 observations of an EAR(2) process.

We speak about **bilinear model** if the process is driven by the relation

$$Y_t = \sum_{i=1}^{p} \phi_i Y_{t-i} + \sum_{j=1}^{q} \psi_j \varepsilon_{t-j} + \sum_{i=1}^{p} \sum_{j=1}^{q} \gamma_{i,j} Y_{t-i} \varepsilon_{t-j}$$

where ε_t is a white noise process.

The bilinear model may be generated in XploRe using `genbil`, which has the following syntax:

```
y = genbil(phi,psi,gamma,noise)
```

where

- `phi` is a p-dimensional vector of the AR parameters,
- `psi` is a q-dimensional vector of the MA parameters,
- `gamma` is a $p \cdot q$-dimensional vector of the bilinear parameters sorted as

$$(\gamma_{1,1}, \gamma_{1,2}, \ldots, \gamma_{1,q}, \gamma_{2,1}, \ldots, \gamma_{p,q}),$$

- `noise` is an n-dimensional vector of the noise (n is the number of observations to be generated).

The following example generates and displays 250 observations of a bilinear process, the resulting plot is shown in Figure 9.7.

```
y=genbil(0.5|0.2, 0.3|-0.3, 0.8|0|0|0.3,normal(250))
timeplot(y)
```

`times10.xpl`

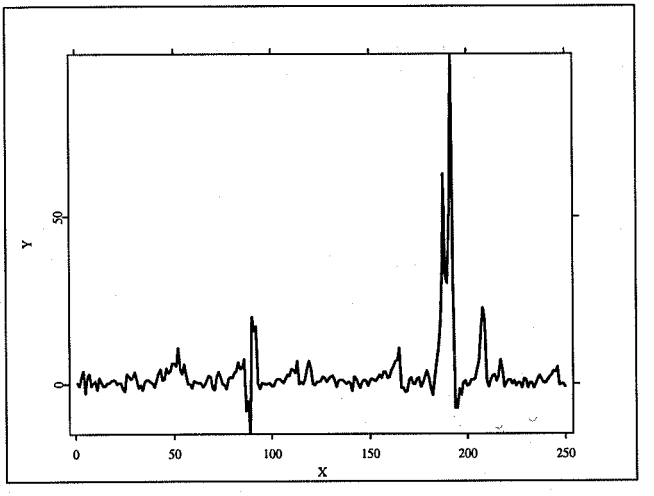


Figure 9.7. 250 observations of a bilinear process.

Based on the article Grassberger and Procaccia (1983), `gpplot` implements the Grassberger–Procaccia plot of time series. This function is applied as follows:

```
x=normal(100)
d=gpplot(x,2,10)
```

`times11.xpl`

Note that this example does not work in the Academic Edition of XploRe.

9.3.2 Nonlinearity in the Conditional Second Moments

A family of models with conditional heteroscedasticity is lately very popular especially among econometricians analyzing financial time series. This family may be introduced by a general formula

$$Y_t = f(Y_{t-1}, \ldots, Y_{t-\tau}) + \varepsilon_t, \quad t = 1, \ldots, T \quad \text{with} \quad \varepsilon_t | I_t \sim N(0, \sigma_t^2)$$

where I_t is the information set available at time t. The conditional variance σ_t^2 is a function of the information contained in the set I_t. We consider here the class of ARCH/GARCH models which restricts the information set I_t to the sequence of past squared disturbances ε_t^2.

ARCH models

The class of ARCH models represents the conditional variance σ_t^2 at time t as a function of the past squared disturbances

$$\sigma_t^2 = \omega + \sum_{i=1}^{q} \alpha_i \varepsilon_{t-i}^2.$$

This defines an ARCH(q) process where q is the order of the autoregressive lag polynomial $\alpha(L) = \alpha_1 L + \ldots + \alpha_q L^q$. This class of processes is generalized and nested in the class of generalized ARCH processes, denoted as GARCH and defined by the relation

$$\sigma_t^2 = \omega + \sum_{i=1}^{q} \alpha_i \varepsilon_{t-i}^2 + \sum_{j=1}^{p} \beta_j \sigma_{t-j}^2 = \omega + \alpha(L)\varepsilon_t^2 + \beta(L)\sigma_t^2.$$

This defines a GARCH(p, q) process, where p and q are the respective orders of the autoregressive and moving average lag polynomials. The moving average lag polynomial is defined as $\beta(L) = \beta_1 L + \ldots + \beta_p L^p$.

An GARCH process can be generated using `genarch`, which has the following syntax:

```
y = genarch(a,b,n)
```

where

- `a` is the vector which contains the constant and the coefficient of the autoregressive lag polynomial,

- b is the vector of parameters of the moving average lag polynomial,
- n is the size of the simulated series.

If the absolute value of the sum of coefficients of the autoregressive and moving average polynomials is greater than 1, i.e. if the generated process is not stationary, XploRe displays an error message.

In the following example, 250 observations of a GARCH(1,1) process are generated and displayed:

```
a = #(1,0.45)          ; AR part of GARCH errors
b = #(0.5)             ; MA part of GARCH errors
y = genarch(a,b,250)
```

times12.xpl

Figure 9.8 shows the process.

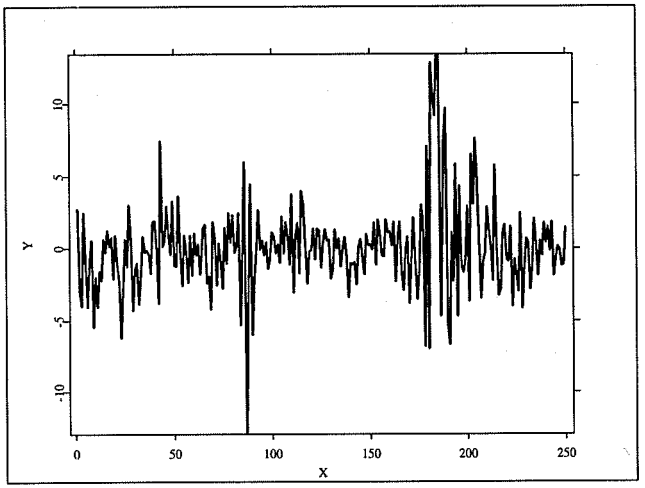


Figure 9.8. 250 observations of a GARCH(1,1) process.

The displayed time series shows that the GARCH process is characterized by the clustering of large deviating observations. This occurs typically in financial time series.

We can verify that the generated series is leptokurtic by evaluating its kurtosis with the command `kurtosis(y)` which yields the following output:

```
Contents of k
[1,]   7.1956
```

This value is greater than the kurtosis of the Gaussian distribution, i.e. 3. This feature is common to all ARCH and GARCH processes.

9.3.3 Estimating ARCH Models

ARCH/GARCH models are estimated in the time domain by the maximum likelihood method. This implies that we have to make an assumption about the distribution of the error terms. Usually the assumption of normality is sufficient and simplifies the estimation procedure. However, in some cases, this normality assumption is not sufficient for capturing the leptokurtosis of the sample under investigation. In that case we have to resort to other fat tailed distributions, such as the Student or the generalized exponential distribution.

Under the assumption of Gaussian innovations, we can estimate the parameters of a GARCH process using `archest`. Its syntax is as follows:

```
z = archest(y,q,p)
```

where

- `y` is the series of observations,
- `p` is the order of the $\beta(L)$ polynomial,
- `q` is the order of the $\alpha(L)$ polynomial.

To estimate the parameters of the previously generated process, type:

```
z = archest(y,1,1)
```

The quantlet `archest` estimates the parameters, their standard errors, and stores them in the form of a list. The first element of the list, `z{1}` contains the set of parameters, the second element of the list, `z{2}` contains the standard errors of estimates. Typing `z{1}` yields the parameter estimates

```
[1,]   1.9675
[2,]  0.40522
[3,]  0.42373
```

Typing `z{2}` returns the estimated standard errors

```
[1,]  0.57227
[2,]  0.07518
[3,]  0.077201
```

Note that `archest` returns additional values. For their description consult the APSS help file.

9.3.4 Testing for ARCH

We test for ARCH effects by investigating whether the components of the autoregressive lag polynomial are all equal to zero. Two formulations for a test for ARCH are usually considered:

- The Lagrange Multiplier test statistic, given by

$$\text{A-LM} = \frac{1}{2}\tilde{\varepsilon}^T\tilde{\varepsilon},$$

  where $\tilde{\varepsilon}$ is the T-dimensional vector of estimated endogenous variables in the auxiliary regression

$$\frac{\widehat{\varepsilon}_t^2}{\widehat{\sigma}^2} - 1 = \gamma_0 + \gamma_1\widehat{\varepsilon}_{t-1}^2 + \ldots + \gamma_n\widehat{\varepsilon}_{t-n}^2 + \zeta_t.$$

  Here $\widehat{\varepsilon}_t$ are the residuals from the regression model, and $\sigma_t^2 = T^{-1}\sum_t \widehat{\varepsilon}_t^2$.

- The R^2 form, with the test statistic equal to

$$TR^2,$$

  where R^2 is the squared multiple R^2 value of the regression of $\widehat{\varepsilon}_t^2$ on an intercept and n lagged values of $\widehat{\varepsilon}_t^2$.

Both tests are asymptotically equivalent, and asymptotically distributed as χ_n^2 under the null hypothesis.

The XploRe quantlet `archtest` performs these two tests. Its syntax is

```
h = archtest(y {,lags {,testform}})
```

where

- `y` is the vector of residuals.
- `lags` is the number of lags in the auxiliary regression. This argument may either be a scalar or a vector. In the latter case, the statistics is computed for all the order components of the vector. By default, the lag orders 2, 3, 4, and 5 are computed.
- `testform` is the form of the test. This argument is a string which can be either `"LM"` or `"TR2"`. In the former case the Lagrange multiplier form is evaluated, while in the latter case the R^2 form is computed.

This function returns a table: The first column contains the order of the test, the second column contains the value of the test, the third column contains the 95% critical value for the respective order, and the fourth column contains the p-value of the test.

In our generated sample of observations, we test for ARCH with the command

```
archtest(y,"TR2")
```

This calculates the R^2 form of the ARCH test for the default lags 2, 3, 4 and 5. The results are displayed in the form of a table:

```
[1,] Lag order  Statistic  95% Critical Value  P-Value
[2,] ------------------------------------------------
[3,]
[4,]       2     85.45238          5.97378      0.00000
[5,]       3    105.05328          7.80251      0.00000
[6,]       4    104.68014          9.47844      0.00000
[7,]       5    105.15906         11.06309      0.00000
```

We recommend to consult the APSS help file of `archest` and to play around with the numerous variants of this quantlet.

Kamstra (1993), Caulet and Péguin–Feissolle (1997) and Péguin–Feissolle (1999) consider a general nonparametric test for ARCH based on neural networks, in the spirit of the Lee, White and Granger (1993) nonlinearity test presented

above. Caulet and Péguin–Feissolle (1997) consider the following parameterization of the conditional variance

$$\sigma_t^2 = \beta_0 + \sum_{j=1}^{q} \frac{\beta_j}{1 + \exp\{-(\gamma_{j0} + \gamma_{j1}\varepsilon_{t-1}^2 + \ldots + \gamma_{jn}\varepsilon_{t-n}^2)\}}. \qquad (9.1)$$

Péguin–Feissolle (1999) considers the more general case

$$\sigma_t^2 = \beta_0 + \sum_{j=1}^{q} \frac{\beta_j}{1 + \exp\{-(\gamma_{j0} + \gamma_{j1}\varepsilon_{t-1} + \ldots + \gamma_{jn}\varepsilon_{t-n})\}},$$

i.e. extends the σ-field of the information set from the set of squared residuals to the set of residuals. This extended test appears to be more powerful than other tests for ARCH when the data generating process is a nonstandard one. The parameters $\gamma_{i,j}$ are randomly generated for solving the problem of parameter identification under the null hypothesis. All these neural network based test statistics are asymptotically χ_n^2 distributed under the null hypothesis.

The quantlet `annarchtest` evaluates the Lagrange multiplier form and the R^2 form for the specification (9.1). Its syntax is

```
h = annarchtest(y, {,nlags {,nodes {,testform}}})
```

where

- `y` is the vector of residuals.
- `nlags` is the number of lags in the auxiliary regression. This second argument may either be a vector or a scalar. If this argument is a vector, the test will be calculated for all order components of this vector. By default, the number of lags is set to 2, 3, 4 and 5.
- `nodes` is the number of hidden nodes in the neural network architecture. This argument may either be a vector or a scalar. By default, `nodes` is set to 3.
- If both the second and third arguments are vectors, the statistic will be calculated for all combinations of the second and third arguments.
- `testform` is the form of the test, which — as in the previous case — is either `"LM"` or `"TR2"` depending on the form of the test you wish to compute. By default, the `"LM"` form is calculated.

This function returns the results in the form of a table: The first column contains the number of lags in the auxiliary regression, the second column contains the number of hidden nodes, the third column contains the calculated statistic, the fourth column contains the 95% critical value for that order, and the last column contains the p-value of the test.

We calculate the test for ARCH based on neural networks with the command

```
annarchtest(y)
```

This returns

```
[1,] Lag order  Nb of hidden Statistic  95% Critical Value  P-Value
[2,]            units
[3,] ------------------------------------------------------------
[4,]
[5,]     2          3        26.56036        5.97378        0.00000
[6,]     3          3        60.80811        7.80251        0.00000
[7,]     4          3        14.44156        9.47844        0.00601
[8,]     5          3        27.32019       11.06309        0.00005
```

Bibliography

Andrews, D. W. K. (1991). Heteroskedasticity and Autocorrelation Consistent Covariance Matrix Estimation, *Econometrica* **59**: 817–858.

Brockwell, P. J. and Davis, R. A. (1991). *Time Series: Theory and Methods*, Springer-Verlag, New York.

Caulet, R. and Péguin–Feissolle, A. (1997). A Test for Conditional heteroskedasticity Based on Artificial Neural Networks, *GREQAM DT* **97A09**.

Grassberger and Procaccia (1983). Measuring the Strangeness of Strange Attractors, *Physica* **9D**: 189–208.

Harvey, A.C. (1993). *Time Series Models*, Harvester Wheatsheaf.

Kamstra, M. (1993). A Neural Network Test for Heteroskedasticity, *Simon Fraser Working Paper.*

Kwiatkowski, D., Phillips, P. C. B., Schmidt, P. and Shin, Y. (1992). Testing the Null Hypothesis of Stationarity Against the Alternative of a Unit Root: How Sure Are We that Economic Series Have a Unit Root, *Journal of Econometrics* **54**: 159–178.

Lee, T. H., White, H. and Granger, C. W. J. (1993). Testing for Neglected Nonlinearity in Time Series Models – A Comparison of Neural Network Methods and Alternative Tests, *Journal of Econometrics* **56**: 269–290.

Lee, D. and Schmidt, P. (1996). On the Power of the KPSS Test of Stationarity Against Fractionally-Integrated Alternatives, *Journal of Econometrics* **73**: 285–302.

Newey, W. K. and West, K. D. (1987). A Simple Positive Definite, Heteroskedasticity and Autocorrelation Consistent Covariance Matrix, *Econometrica* **55**: 703–705.

Péguin–Feissolle, A (1999). A Comparison of the Power of Some Tests for Conditional Heteroskedasticity, *Economics Letters*, forthcoming.

Siddiqui, M. (1976). The Asymptotic Distribution of the Range and Other Functions of Partial Sums of Stationary Processes, *Water Resources Research* **12**: 1271–1276.

Teräsvirta, T. (1998). Modeling Economic Relationships with Smooth Transition Regression Function, *Handbook of Applied Economic Statistics.*

Velasco, C. (1998). Gaussian Semiparametric Estimation of Non–Stationary Time Series, *Journal of Time Series Analysis*, forthcoming.

XploRe Learning Guide,p 229-246, Springer Verlag

8 Neural Networks

Wolfgang Härdle and Heiko Lehmann

A neural network consists of many simple processing units that are connected by communication channels. Much of the inspiration for the field of neural networks came from the desire to perform artificial systems capable of sophisticated, perhaps intelligent computations similar to those of the human brain.

Neural networks usually learn from examples and exhibit some capability for generalization beyond the data used for training. They are able to approximate highly nonlinear functional relationships in data sets.

The smallest part of a neural network is one single neuron as shown in Figure 8.1. It takes a set of individual **inputs** $x = (x_1, \ldots, x_I)$ and determines (through the learning algorithm) the optimal **connection weights** $w = (w_1, \ldots, w_I)$ that are appropriate to each input. Next, the neuron aggregates these weighted values to a single value

$$u = w_0 + \sum_{i=1}^{I} w_i x_i \,.$$

An **activation function** $F(\bullet)$ is then applied to the aggregated weighted value to produce an individual output

$$F(u)$$

for the specific neuron. A typical activation function is the logistic distribution function

$$F(u) = \frac{1}{1 + \exp(-u)}.$$

The aim of a neural network is to explain the **outputs** $y = (y_1, \ldots, y_Q)$ by the input variables $x = (x_1, \ldots, x_I)$. More exactly, we want to find functions $f_k(\bullet)$ such that $f_k(x)$ explains the output variable y_k.

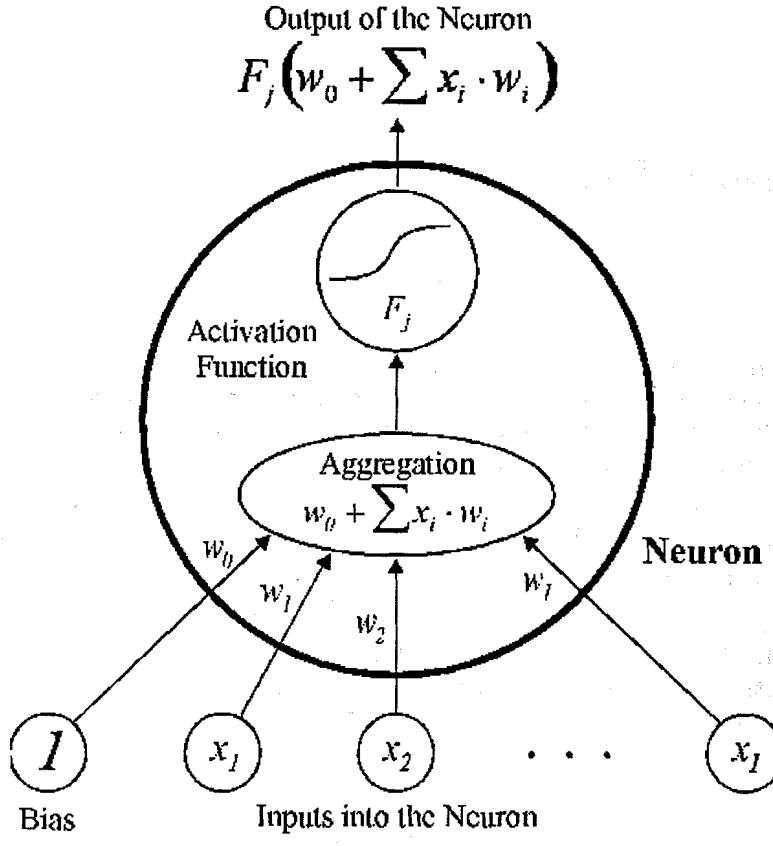


Figure 8.1. A neuron within a neural network.

A neural network with one **hidden layer** (single hidden layer) consists of neurons of three basic types:

- The **input neurons** collect the external information and send it to the layer of hidden units.
- The **hidden neurons** aggregate the information and send it to the output neuron(s).
- The **output neurons** contain the aggregated information passed through the activation function.

8.1 Feed-Forward Networks

Figure 8.2 shows a **feed-forward** network with one hidden layer. This network attempts to fit the model

$$f_k(x) = F\left\{w_{0k}^{(2)} + \sum_{j=1}^{J} w_{jk}^{(2)} F\left(w_{0j}^{(1)} + \sum_{i=1}^{I} w_{ij}^{(1)} x_i\right)\right\}$$

for the output unit y_k. Feed-forward means that information can only flow forward from the input units to the first hidden layer, from the first hidden layer to the second hidden layer, and so on. Information cannot flow between the units of one layer.

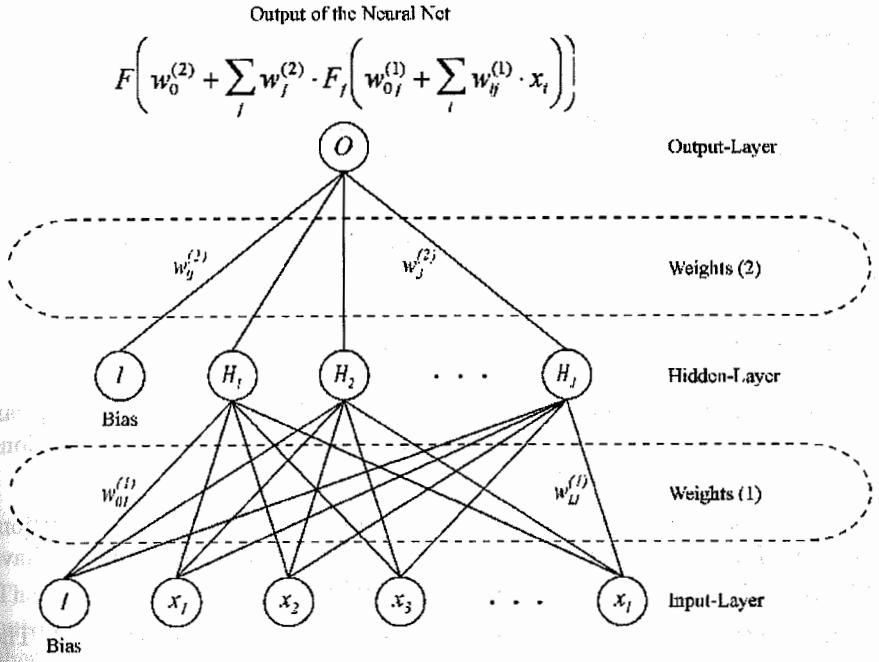


Figure 8.2. Feed-forward network with one hidden layer.

8.2 Computing a Neural Network

`net = nnrnet (x, y, w, size, {, param, wts})`
trains a single layer feed-forward network with input `x`, output `y`, prior weights `w`, and number of hidden units `size`; optionally the type of the network can be determined by `param` and initial weights `wts` can be given

`net = nnrpredict (x, net)`
predicts the responses for given variables `x` and network `net`

`net = nnrinfo (net)`
shows information about network `net`

`nnrsave(net, "nnfile")`
saves network `net` to files `nnfile.*`

`net = nnrload ("nnfile")`
loads network `net` from files `nnfile.*`

The function `nnrnet` allows for constructing and training a single hidden layer network with maximal 100 units. The call looks like

```
net = nnrnet (x, y, w, size, param, wts)
```

where `x` and `y` are the input and output variables. Note that `x` as well as `y` can consist of several variables (columns). We assume that `x` and `y` have dimensions $n \times I$ and $n \times Q$, respectively.

With the `w` parameter, we can associate a prior weight to each observation. This is useful, e.g. for ties in the data. Note that the prior weights `w` have nothing in common with the weights calculated in the net.

The parameter `size` determines the number of units in the hidden layer. The total number of units must not exceed 100, i.e.

$$\text{columns of } x + \text{columns of } y + \text{units in hidden layer} \leq 100.$$

The default network is a classification network: logistic output units, no softmax, no "skip-layer" connections, no weight decay and the training stops after

100 iterations. The default model for the output units y_k, $k = 1, \ldots, Q$, is hence

$$f_k(x) = F\left\{ w_{0k}^{(2)} + \sum_{j=1}^{\text{size}} w_{jk}^{(2)} F\left(w_{0j}^{(1)} + \sum_{i=1}^{I} w_{ij}^{(1)} x_i \right) \right\}$$

with $F(\bullet)$ the logistic function. If a model different to the default is fitted, the parameter `param` needs to be modified. We explain this in more detail in Subsection 8.2.1.

The result of `nnrnet` is a composed object `net`. More information on the components of `net` can be found in Subsection 8.2.2. The function `nnrinfo` shows a short information about the fitted network. The result of

```
nnrinfo(net)
```

could for example print the following information in the output window:

```
[ 1,] "A 2 - 1 - 1 network:"
[ 2,] "# weights     : 5"
[ 3,] "linear output : no"
[ 4,] "error function: least squares"
[ 5,] "log prob model: no"
[ 6,] "skip links    : no"
[ 7,] "decay         : 0"
[ 8,] ""
[ 9,] " From    To Weights"
[10,] "    0     3  -0.751"
[11,] "    1     3    0.81"
[12,] "    2     3   0.575"
[13,] "    0     4   -4.95"
[14,] "    3     4    14.8"
```

The abbreviation 2 – 1 – 1 means two input units, one hidden layer and one output unit. Altogether five weights w_{st} have been calculated, the values of these weights are given in the last lines. The other items show which parameters have been specified for the network.

Typically, a neural network is applied to a subsample of the data which is used as a training data set. The remaining observations are then used to validate the network. To compute predicted values for the validation set, `nnrpredict` is used:

```
ypred = nnrpredict (xval, net)
```

Since the result of a neural network fitting is a composed object, two convenient functions for saving and loading neural networks are provided. The network net can be stored into a set of files by

```
nnrsave (net, "mynet")
```

All created files start with the prefix mynet. The network can be reloaded by

```
net = nnrload ("mynet")
```

8.2.1 Controlling the Parameters of the Neural Network

The type of a network and the control parameters for the iteration are determined by the parameter param of nnrnet. If, for instance, a model different to the default is fitted, this parameter needs to be modified. param is a vector of eight elements:

param[1]
: determines if the activation function for the output is the logistic function (default value 0). Setting param[1] to the value 1 changes the activation function of the output unit to the identity function.

param[2]
: determines the error function (the optimization criterion). The default value 0 indicates the quadratic least squares error function

$$\sum_{k=1}^{\text{size}} \sum_{i=1}^{n} \{f_k(x_i) - y_{i,k}\}^2 .$$

Setting param[2] to the value 1 changes the error function to the entropy for the classification case

$$\sum_{k=1}^{\text{size}} \sum_{i=1}^{n} \left\{ f_k(x_i) \log \left(\frac{f_k(x_i)}{y_{i,k}} \right) + \{1 - f_k(x_i)\} \log \left(\frac{1 - f_k(x_i)}{1 - y_{i,k}} \right) \right\} .$$

`param[3]`
If `param[3]` is set to the value `1`, then the softmax activation function is used for the outputs. This means the output is

$$f_k(x_i) = \frac{\exp\{f_k^*(x_i)\}}{\sum_{\ell=1}^{Q} \exp\{f_\ell^*(x_i)\}}.$$

The default value is `0`, which means no softmax.

`param[4]`
includes "skip-layer" connections. Setting `param[4]` to the value `1` generates "skip-layer" connections, i.e.

$$f_k(x) = w_{0k}^{(2)} + \sum_{i=1}^{p} w_{ij}^{(2)} x_i + \sum_{j=1}^{\text{size}} w_{kj}^{(2)} F\left(w_{0j}^{(1)} + \sum_{i=1}^{p} w_{ij}^{(1)} x_i \right).$$

The default value is `0`, which means no "skip-layer" connections.

`param[5]`
sets the maximal value δ for the initial weights. If the optional input parameter `wts` is not given, uniform random numbers from $[-\delta, \delta]$ are used. The default value is $\delta = 0.7$.

`param[6]`
sets the weight decay, the default is `0`.

`param[7]`
sets the maximal number of iterations, the default is `100`.

`param[8]`
shows information about the iteration. Setting `param[8]` to the value `1` produces control output in the output window during the optimization. The default is `0`, i.e. not to show control output.

8.2.2 The Resulting Neural Network

The result of `nnrnet` is a composed object, the list `net`, which contains the resulting fit and information about the network. The components are the following:

`net.n`
three-dimensional vector that contains the number of input, hidden and output units, respectively

`net.nunits`, `net.nconn`, `net.conn`
internal information about the network topology

`net.decay`
scalar, the weight decay parameter (=`param[6]`)

`net.entropy`
scalar, the value of the entropy

`net.softmax`
scalar, softmax indicator (=`param[3]`)

`net.value`
scalar, the value of the error function

`net.wts`
vector of final weights

`net.yh.result`
$n \times Q$ matrix, the estimated outputs

`net.yh.hess`
the Hessian matrix

8.3 Running a Neural Network

In the following two sections we run simple neural nets on clustered data. Before proceeding to the examples, the following libraries need to be loaded:

```
library ("plot")
library ("nn")
```

The `nn` library contains the functions for running the networks. The `plot` library is used to produce scatter plots of the clusters.

8.3.1 Implementing a Simple Discriminant Analysis

In the following, we will use a single hidden layer network with one hidden unit to perform a discriminant analysis on an artificially generated data set with two clusters.

All XploRe codes for this subsection can be found in nn1.xpl. The first step is to generate the training data set:

```
randomize(0)
n  = 200
xt = normal(n,2)+#(-1,-1)' | normal(n,2)+#(+1,+1)'
```

Here, a mixture of two two-dimensional normal distributions is generated. Each cluster consists of n = 200 observations. The variances are identical (equal to 1 in both directions) whereas the means are shifted by (+1,+1) and (−1,−1), respectively. The following code lines can be used to display the data set graphically:

```
color  = string("red",1:n) | string("blue",1:n)
symbol = string("circle",1:n) | string("triangle",1:n)
xt     = setmask(xt, color, symbol)
plot(xt)
xl="x1"
yl="x2"
tl="Training Data Set"
setgopt(plotdisplay,1,1,"title",tl,"xlabel",xl,"ylabel",yl)
```

The generated two-dimensional data are shown in Figure 8.3. We have labeled the observations from the first cluster by red circles, whereas the observation from the second cluster are labeled as blue triangles.

To apply the neural network, we need to create now the output variable y and the prior weights w. For y, we use a value of 0 for the first and a value of 1 for the second cluster. The prior weights are all set to 1. The last statement of the following code computes the neural network using one hidden unit and assigns the result to net.

```
yt = (matrix(n)-1)|matrix(n)
w  = matrix(2*n)
```

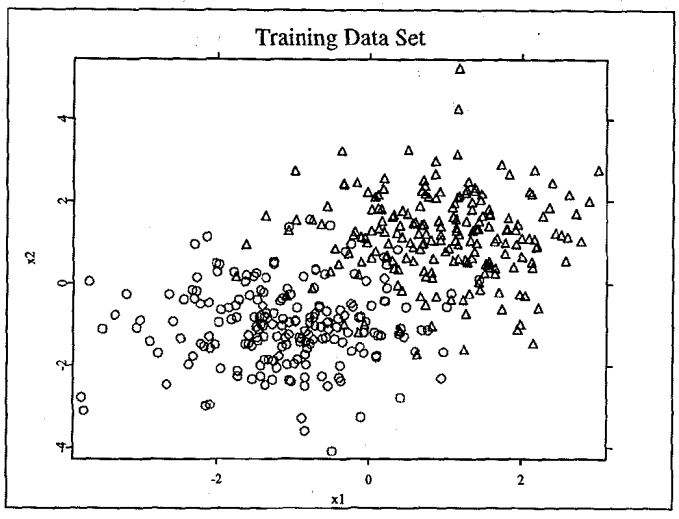


Figure 8.3. A generated training data set with two clusters.

```
param = 1
net = nnrnet(xt,yt,w,1)
```

We can obtain a summary of the fitted network from

```
nnrinfo(net)
```

which prints into the output window:

```
Contents of ts
[ 1,] "A 2 - 1 - 1 network:"
[ 2,] "# weights      : 5"
[ 3,] "linear output  : no"
[ 4,] "error function: least squares"
[ 5,] "log prob model: no"
[ 6,] "skip links     : no"
[ 7,] "decay          : 0"
[ 8,] ""
[ 9,] " From    To Weights"
```

```
[10,] "    0     3   -1.18"
[11,] "    1     3  -0.285"
[12,] "    2     3  -0.198"
[13,] "    0     4    99.8"
[14,] "    3     4    44.8"
```

To validate the obtained network, we generate new random data from the same mixture of two-dimensional normal distributions. The classification of these data using the network `net` is done by `nnrpredict`.

```
x = normal(n,2)+#(-1,-1)' | normal(n,2)+#(+1,+1)'
pred  = nnrpredict(x, net)
prob  = pred.result
```

The macro `nnrpredict` calculates the predicted values and the Hessian matrix. `pred.result` extracts the predicted values.

Now we compute the misclassified observations and show them in comparison with the original data `x`.

```
y  = (matrix(n)-1)|matrix(n) ; true
yp = prob > 0.5              ; predicted
misc = paf(1:2*n,y!=yp)      ; misclassified
good = paf(1:2*n,y==yp)      ; correctly classified
nm = rows(misc)
sm = string("fill",1:nm)+symbol[misc]
xm = setmask(x[misc],color[misc],sm,"huge")
xg = setmask(x[good],color[good],symbol[good])

pm = 100*nm/(2*n)            ; percentage of misclassified
spm = string("%1.2f",pm)+"%"
Network = createdisplay(1,1)
show(Network,1,1,xg,xm)
tl="Network: misclassified = "+spm
setgopt(Network,1,1,"title",tl,"xlabel",xl,"ylabel",yl)
```

Figure 8.4 shows the two-dimensional data that we used for validation. As before, observations from the first cluster are labeled by red circles, whereas the observation from the second cluster are labeled as blue triangles. All misclassified data are labeled by large filled symbols.

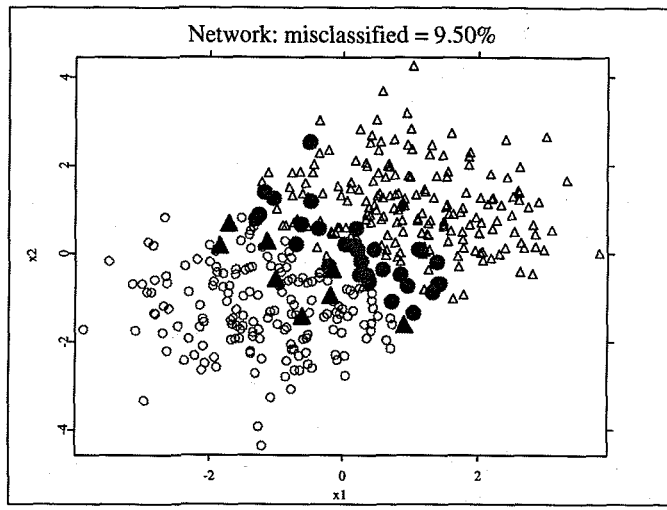


Figure 8.4. Neural network classification.

Let's compare the classification obtained from the neural network with that from a classical linear discriminant analysis. Apart from the discrimination rule that is used for the prediction here, the code is almost identical to the above.

```
mu0 = mean(xt[1:n])
mu1 = mean(xt[n+1:2*n])
mu  = (mu0+mu1)/2
lin = inv(cov(xt))*(mu0-mu1)'

y   = (matrix(n)-1)|matrix(n) ; true
yp  = (x-mu)*lin<=0           ; predicted
misc = paf(1:2*n,y!=yp)       ; misclassified
good = paf(1:2*n,y==yp)       ; correctly classified
nm = rows(misc)
sm = string("fill",1:nm)+symbol[misc]
xm = setmask(x[misc],color[misc],sm,"huge")
xg = setmask(x[good],color[good],symbol[good])
x  = setmask(x, color, symbol)
```

```
pm = 100*nm/(2*n)                ; percentage of misclassified
spm = string("%1.2f",pm)+"%"
Discrim = createdisplay(1,1)
show(Discrim,1,1,xg,xm)
tl="Linear misclassified = "+spm
setgopt(Discrim,1,1,"title",tl,"xlabel",xl,"ylabel",yl)
```

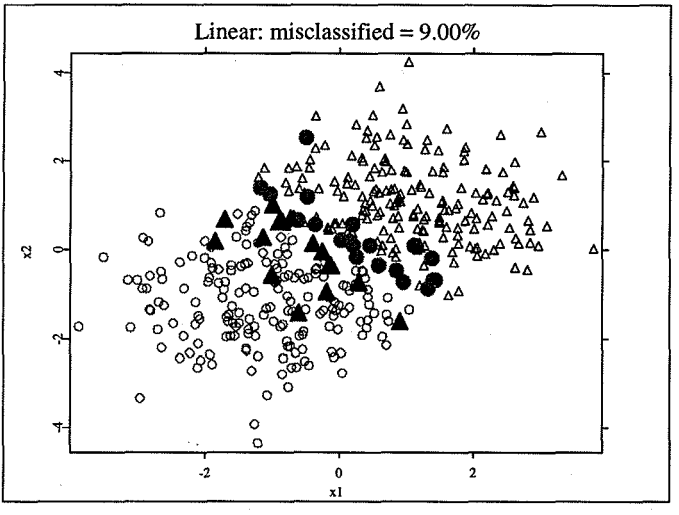


Figure 8.5. Linear discriminant analysis.

Figure 8.5 shows the resulting classification. Again, all misclassified data are labeled by large filled symbols. Comparing Figures 8.4 and 8.5 shows that the percentage of misclassification is nearly equal for both methods. The linear discriminant analysis performs slightly better. This is not astonishing, since the linear discriminant analysis is designed to handle the data that we generated.

8.3.2 Implementing a More Complex Discriminant Analysis

In contrast to the previous subsection, we will now consider a generated data set where the linear discriminant analysis performs worse than the neural network.

The XploRe codes are largely identical to the previous examples and can b found in nn2.xpl.

As before we generate a training data set, which features two clusters.

```
randomize(0)
n  = 100
xt = normal(n,2)+#(-1,-1)' | normal(n,2)+#(+1,-2)'
xt = xt | normal(n,2)+#(+4, 0)' | normal(n,2)+#(+1,+1)'

color  = string("red",1:3*n) | string("blue",1:n)
symbol = string("circle",1:3*n) | string("triangle",1:n)
xt     = setmask(xt, color, symbol)
plot(xt)
xl="x1"
yl="x2"
tl="Training Data Set"
setgopt(plotdisplay,1,1,"title",tl,"xlabel",xl,"ylabel",yl)
```

The generated two-dimensional data are shown in Figure 8.6. It is obvious that here the points from the second group (labeled by blue triangles) overlap the points from the first group (red circles) in a more complicated way.

We proceed in the same way as before, i.e. we create the output variable y and set all prior weights w to 1. Then the neural network is fitted. In contrast to the previous section, we now use 3 hidden layers to take the more complex structure of the data into account.

```
yt = (matrix(3*n)-1)|matrix(n)
w  = matrix(4*n)
param = 1
net = nnrnet(xt,yt,w,3)
nnrinfo(net)
```

The resulting fit is summarized as follows:

```
Contents of ts
[ 1,] "A 2 - 3 - 1 network:"
[ 2,] "# weights      : 13"
[ 3,] "linear output : no"
```

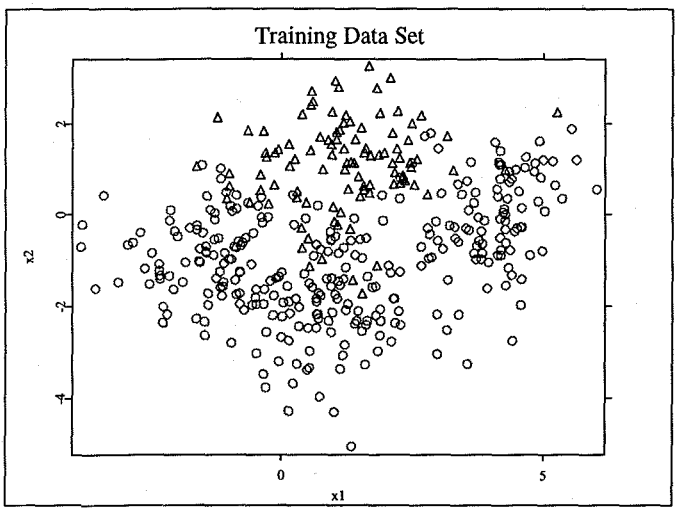


Figure 8.6. A generated training data set with two clusters.

```
[ 4,] "error function: least squares"
[ 5,] "log prob model: no"
[ 6,] "skip links    : no"
[ 7,] "decay         : 0"
[ 8,] ""
[ 9,] " From    To Weights"
[10,] "    0     3    1.26"
[11,] "    1     3  -0.106"
[12,] "    2     3   -5.15"
[13,] "    0     4    2.92"
[14,] "    1     4   -1.32"
[15,] "    2     4    0.37"
[16,] "    0     5   -7.61"
[17,] "    1     5   -56.1"
[18,] "    2     5   -28.3"
[19,] "    0     6   -2.76"
[20,] "    3     6   -4.38"
[21,] "    4     6    7.64"
[22,] "    5     6   -4.71"
```

Again, we assess the quality of the obtained network by counting the misclassified observations for a validation data set.

```
x = normal(n,2)+#(-1,-1)' | normal(n,2)+#(+1,-2)'
x = x | normal(n,2)+#(+4, 0)' | normal(n,2)+#(+1,+1)'
pred  = nnrpredict(x, net)
prob  = pred.result

y  = (matrix(3*n)-1)|matrix(n) ; true
yp = prob > 0.5                ; predicted
misc = paf(1:4*n,y!=yp)        ; misclassified
good = paf(1:4*n,y==yp)        ; correctly classified
nm = rows(misc)
sm = string("fill",1:nm)+symbol[misc]
xm = setmask(x[misc],color[misc],sm,"huge")
xg = setmask(x[good],color[good],symbol[good])

pm = 100*nm/(4*n)                  ; percentage of misclassified
spm = string("%1.2f",pm)+"%"
Network = createdisplay(1,1)
show(Network,1,1,xg,xm)
tl="Network: misclassified = "+spm
setgopt(Network,1,1,"title",tl,"xlabel",xl,"ylabel",yl)
```

Figure 8.7 shows the resulting plot of the two-dimensional data that we used for prediction, with misclassified data labeled by large filled symbols.

The comparison with the classical linear discriminant analysis is implemented in the following lines:

```
mu0 = mean(xt[1:3*n])
mu1 = mean(xt[3*n+1:4*n])
mu  = (mu0+mu1)/2
lin = inv(cov(xt))*(mu0-mu1)'

y  = (matrix(3*n)-1)|matrix(n) ; true
yp = (x-mu)*lin<=0             ; predicted
misc = paf(1:4*n,y!=yp)        ; misclassified
good = paf(1:4*n,y==yp)        ; correctly classified
nm = rows(misc)
```

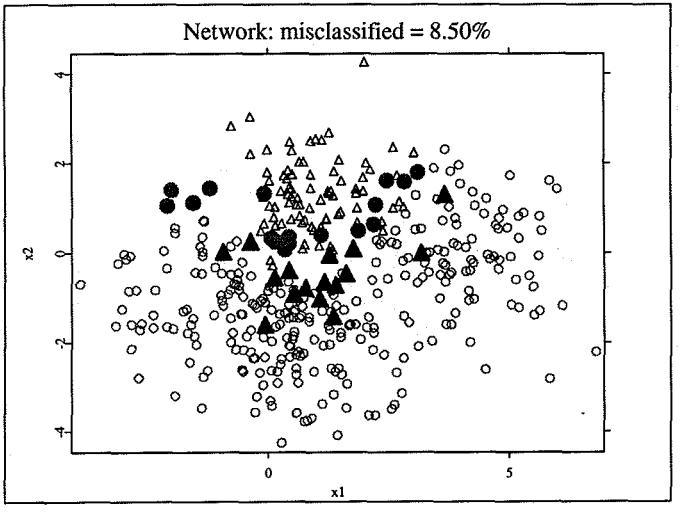


Figure 8.7. Neural network classification.

```
sm = string("fill",1:nm)+symbol[misc]
xm = setmask(x[misc],color[misc],sm,"huge")
xg = setmask(x[good],color[good],symbol[good])
x  = setmask(x, color, symbol)

pm = 100*nm/(4*n)              ; percentage of misclassified
spm = string("%1.2f",pm)+"%"
Discrim = createdisplay(1,1)
show(Discrim,1,1,xg,xm)
tl="Linear misclassified = "+spm
setgopt(Discrim,1,1,"title",tl,"xlabel",xl,"ylabel",yl)
```

Figure 8.8 shows the resulting classification. The comparison of Figures 8.7 and 8.8 reveals now that the neural network separates the clusters more accurately. This is due to the fact that the neural network with three hidden units can better adapt to a nonlinear discrimination rule.

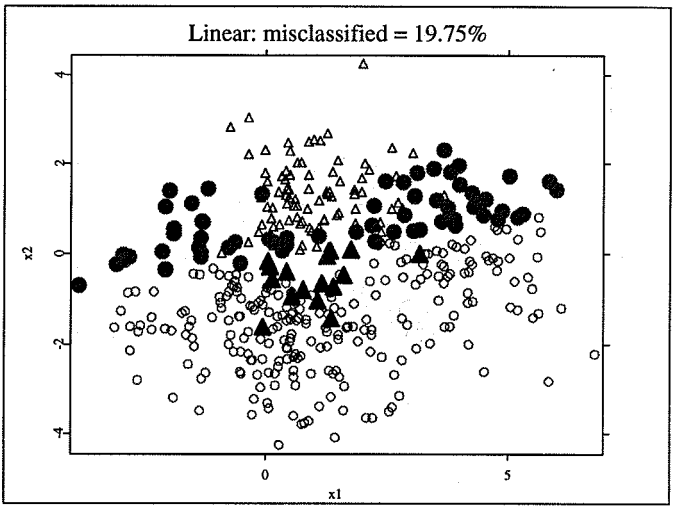


Figure 8.8. Linear discriminant analysis.

Bibliography

Bishop, C. (1995). *Neural Networks for Pattern Recognition*, Clarendon Press, Oxford.

Ripley, B. (1996). *Pattern Recognition and Neural Networks*, Cambridge University Press.

Smoothing and Regression, Approaches, Computation and Application
Chapter 12 M.Schimek (ed), Wiley Publishers,357-391

1 Multivariate and Semiparametric Kernel Regression

Wolfgang HÄRDLE and Marlene MÜLLER

Institut für Statistik und Ökonometrie, Wirtschaftswissenschaftliche Fakultät, Humboldt–Universität zu Berlin, Germany

1.1 INTRODUCTION

Nonparametric smoothing methods serve three essential needs in statistical data analysis. First they provide a flexible analysis tool, often based on interactive graphical data representation (Scott, 1992). Second they help in constructing a model from observations, for example by comparison with concurrent models (Müller, 1988). Third they provide pilot estimators in adaptation problems, see Newey and Stoker (1993). Here we present the multivariate kernel smoother, examine the asymptotic properties of both density and regression estimators, and review applications of this technique in semiparametric statistics.

Multivariate nonparametric density estimation is an often used pilot tool for examining the structure of data. Regression smoothing helps in investigating the association between covariates and responses. We concentrate on kernel smoothing using local polynomial fitting which includes the Nadaraya–Watson estimator. Some theory on the asymptotic behavior and bandwidth selection is provided. In the applications of the kernel technique, we focus on the semiparametric paradigm. In more detail we describe the single index model (SIM) and the generalized partial linear model (GPLM).

1.2 MULTIDIMENSIONAL SMOOTHING WITH KERNELS

In this section we review kernel smoothing methods for density and regression function estimation. Many ideas, in particular for asymptotics, bandwidth

Smoothing and Regression, Approaches, Computation and Application
Chapter 12 M.Schimek (ed), Wiley Publishers,357-391

choice and graphical representation, are similar for both purposes.

We can however only introduce a small part on the available material. In particular, for the regression case we restrict the presentation on the random design case. For a more detailed presentation of the subject we refer to the monographs by Härdle (1990; 1991), Scott (1992), Wand and Jones (1995) and Fan and Gijbels (1995). Kernel regression for univariate data is discussed in detail by Sarda and Vieu (1998) in this volume. For more aspects of multivariate kernel density smoothing see also Scott (1998) in this book.

1.2.1 Multivariate kernel density estimation

The goal of multivariate nonparametric density estimation is to approximate the probability density function (pdf) $f(t) = f(t_1, \ldots, t_q)$ of the random variables $T = (T_1, \ldots, T_q)^T$. The multivariate kernel density estimator in the q–dimensional case is defined as

$$\widehat{f}_h(t) = \frac{1}{n} \sum_{i=1}^{n} \frac{1}{h_1 \ldots h_q} \mathcal{K}\left(\frac{T_{i1} - t_1}{h_1}, \ldots, \frac{T_{iq} - t_q}{h_q} \right), \tag{1.1}$$

$\mathcal{K}$ denoting a multivariate kernel function $\mathcal{K} : I\!R^q \to I\!R$. Note, that (1.1) assumes that the bandwidth h is a vector of bandwidths $h = (h_1, \ldots, h_q)^T$.

What form shall the multidimensional kernel function $\mathcal{K}(u) = \mathcal{K}(u_1, \ldots, u_q)$ take on? The easiest solution is to use a *multiplicative* kernel

$$\mathcal{K}(u) = K(u_1) \cdot \ldots \cdot K(u_q)$$

with K denoting an univariate kernel function. For univariate kernels with support $[-1, 1]$ (as the Epanechnikov kernel $K(u) = 0.75(1 - u^2)\,\mathrm{I}(|u| \leq 1)$) observations in a cube around t are used to estimate the density at the point t. An alternative is to use a genuine multivariate kernel function $\mathcal{K}(u)$, as e.g. the multivariate Epanechnikov

$$\mathcal{K}(u) \propto (1 - u^T u)\ \mathrm{I}(u^T u \leq 1).$$

This type of multivariate kernels can be obtained from univariate by defining

$$\mathcal{K}(u) \propto K(\|u\|), \tag{1.2}$$

where $\|u\| = \sqrt{u^T u}$ denotes the Euclidean norm of the vector u. Note that we use $\propto$ to indicate that the appropriate constant has to be multiplied. Kernels of the form (1.2) use observations from a ball around t to estimate the pdf at t. This type of kernels is usually called *spherical* or *radially symmetric* since $\mathcal{K}(u)$ has the same value for all u on a sphere around zero. Figure 1.1 shows the contour lines from a bivariate product and a bivariate radially symmetric kernel on the left and right hand side, respectively.

Smoothing and Regression, Approaches, Computation and Application
Chapter 12 M.Schimek (ed), Wiley Publishers,357-391

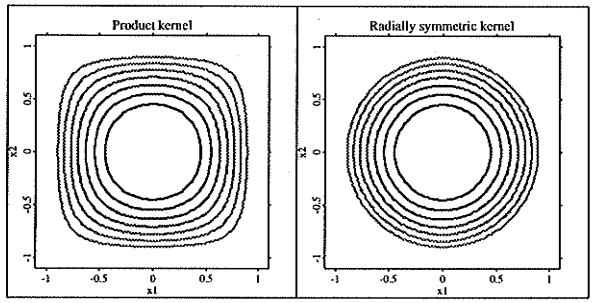


Fig. 1.1 Contours from bivariate product (left) and bivariate radially symmetric (right) Epanechnikov kernel.

Note that the kernel weights in Figure 1.1 correspond to equal bandwidth in each direction, i.e. $h = (h_1, h_2)^T = (1, 1)^T$. When we use different bandwidths, the observations around t in the density estimate $\widehat{f}_h(x)$ will be used with different weights in both dimensions.

Another approach is to use a nonsingular, symmetric *bandwidth matrix* $\mathbf{H}$. The general form for the multivariate density estimator is then

$$\widehat{f}_{\mathbf{H}}(t) = \frac{1}{n}\sum_{i=1}^{n} \frac{1}{\det(\mathbf{H})} \mathcal{K}\left\{\mathbf{H}^{-1}(T_i - t)\right\} = \frac{1}{n}\sum_{i=1}^{n} \mathcal{K}_{\mathbf{H}}\left(T_i - t\right), \qquad (1.3)$$

see Silverman (1986) and Scott (1992). Here we introduce the short notation

$$\mathcal{K}_{\mathbf{H}}(\bullet) = \frac{1}{\det(\mathbf{H})} \mathcal{K}(\mathbf{H}^{-1}\bullet)$$

analogously to $K_h = K(\bullet/h)/h$ in the one–dimensional case. A bandwidth matrix includes all simpler cases as special cases. An equal bandwidth h in all dimensions as in (1.1) corresponds to $\mathbf{H} = h\mathbf{I}_q$ where $\mathbf{I}_q$ denotes the $q \times q$ identity matrix. Different bandwidths as in (1.1) are equivalent to $\mathbf{H} = \mathrm{diag}(h_1, \ldots, h_q)$, the diagonal matrix with elements $h_1, \ldots, h_q$.

What effect does the inclusion of off–diagonal elements have? We will see that a good rule of thumb is to use a bandwidth matrix proportional to $\widehat{\mathbf{\Sigma}}^{-1/2}$ where $\widehat{\mathbf{\Sigma}}$ is the covariance matrix of the data. Hence, using such a bandwidth corresponds to a transformation of the data to identity covariance matrix. As a consequence we can use bandwidth matrices to correct for correlation between the components of T. We have plotted the contour curves of product

and radially symmetric Epanechnikov weights with bandwidth matrix

$$\mathbf{H} = \begin{pmatrix} 1 & 0.5 \\ 0.5 & 1 \end{pmatrix}^{1/2},$$

i.e. $\mathcal{K}_{\mathbf{H}}(u) = \mathcal{K}(\mathbf{H}^{-1}u)/\det(\mathbf{H})$, in Figure 1.2.

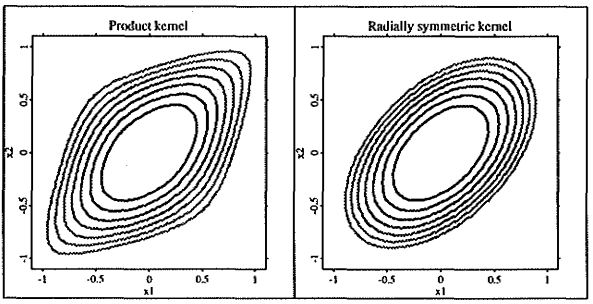


Fig. 1.2 Contours from bivariate product (left) and bivariate radially symmetric (right) Epanechnikov kernel. Bandwidth matrix $\mathbf{H}$.

In the following we will consider statistical properties such as bias, variance, the issue of bandwidth selection and applications for this estimator. We formulate all results for estimators with bandwidth matrices and multivariate kernel function $\mathcal{K}$.

1.2.1.1 Bias, variance and asymptotics A consequence of the standard assumption on the non–negative kernel $\mathcal{K}$

$$\int \mathcal{K}(u)\, du = 1 \tag{1.4}$$

is that the estimate $\widehat{f}_{\mathbf{H}}$ is a density function, i.e. $\int \widehat{f}_{\mathbf{H}}(t)\, dt = 1$. The estimate is consistent in any point t of continuity of f:

$$\widehat{f}_{\mathbf{H}}(t) = \frac{1}{n}\sum_{i=1}^{n} \mathcal{K}_{\mathbf{H}}\left(T_i - t\right) = f(t) + o_p(1), \tag{1.5}$$

if $n \to \infty$, $\det(\mathbf{H}) \to 0$ and $n\det(\mathbf{H}) \to \infty$, see e.g. Ruppert and Wand (1994). The derivation of the mean squared error *MSE* and the mean integrated squared error *MISE* is analogous to the one–dimensional case. We will sketch the asymptotic expansions and concentrate on the asymptotic mean integrated squared error *AMISE*.

As usual, *AMISE* has a bias part *AIB* and a variance part *AIV*. The bias

Smoothing and Regression, Approaches, Computation and Application
Chapter 12 M.Schimek (ed), Wiley Publishers,357-391

of $\widehat{f}_{\mathbf{H}}(t)$ is $E\,\widehat{f}_{\mathbf{H}}(t) - f(t)$ and the integrated squared bias is

$$IB(\mathbf{H}) = \int \{E\,\widehat{f}_{\mathbf{H}}(t) - f(t)\}^2\,dt.$$

The asymptotic integrated squared bias $AIB(\mathbf{H})$ is the first order term of $IB(\mathbf{H})$, i.e.

$$\frac{IB(\mathbf{H}) - AIB(\mathbf{H})}{AIB(\mathbf{H})} = o(1)$$

as $\det(\mathbf{H}) \to 0$, $n \to \infty$ and $n\det(\mathbf{H}) \to \infty$. Define now the integrated variance

$$IV(\mathbf{H}) = \int E\{\widehat{f}_{\mathbf{H}}(t) - E\,\widehat{f}_{\mathbf{H}}(t)\}^2\,dt$$

and the asymptotic integrated variance AIV analogous to AIB. Then the asymptotic mean integrated squared error $AMISE$ can be calculated as

$$AMISE(\mathbf{H}) = AIB(\mathbf{H}) + AIV(\mathbf{H}).$$

A detailed derivation of the components of $AMISE$ can be found in Scott (1992) or Wand and Jones (1995) and the references therein. As in the univariate case we use a second order Taylor expansion. Here and in the following we denote with ∇_f the gradient and with $\mathcal{H}_f$ the Hessian matrix of second order partial derivatives of a function (here f). Then the Taylor expansion of $f(\bullet)$ around t is

$$f(t+u) = f(t) + u^T\nabla_f(t) + \frac{1}{2}u^T\mathcal{H}_f(t)t + o(u^Tu),$$

see Wand and Jones (1995), p. 94. This leads to the expression

$$\begin{aligned} E\,\widehat{f}_{\mathbf{H}}(t) &= \int \mathcal{K}_{\mathbf{H}}(u-t)\,f(u)\,du = \int \mathcal{K}(s)\,f(t+\mathbf{H}s)\,ds \\ &\approx \int \mathcal{K}(s)\left\{f(t) + s^T\mathbf{H}^T\nabla_f(t) + \frac{1}{2}s^T\mathbf{H}^T\mathcal{H}_f(t)\mathbf{H}s\right\}ds. \quad (1.6) \end{aligned}$$

If we assume additionally to (1.4)

$$\int u\mathcal{K}(u)\,du = 0_q,$$
$$\int uu^T\mathcal{K}(u)\,du = \mu_2(\mathcal{K})\mathbf{I}_q,$$

then (1.6) yields $E\,\widehat{f}_{\mathbf{H}}(t) - f(t) \approx \frac{1}{2}\mu_2(\mathcal{K})\,\mathrm{tr}\{\mathbf{H}^T\mathcal{H}_f(t)\mathbf{H}\}$, hence

$$AIB(\mathbf{H}) = \frac{1}{4}\mu_2^2(\mathcal{K})\int \left[\mathrm{tr}\{\mathbf{H}^T\mathcal{H}_f(t)\mathbf{H}\}\right]^2\,dt.$$

Smoothing and Regression, Approaches, Computation and Application
Chapter 12 M.Schimek (ed), Wiley Publishers,357-391

As in univariate density estimation, the leading term of the variance part is the second moment of the estimate, i.e.

$$\begin{aligned} Var\left\{\widehat{f}_{\mathbf{H}}(t)\right\} &= \frac{1}{n}\int \{\mathcal{K}_{\mathbf{H}}(u-t)\}^2\, du - \frac{1}{n}\left\{E\,\widehat{f}_{\mathbf{H}}(t)\right\}^2 \\ &\approx \int \frac{1}{n\det(\mathbf{H})}\mathcal{K}^2(s)\, f(t+\mathbf{H}s)\, ds \\ &\approx \int \frac{1}{n\det(\mathbf{H})}\mathcal{K}^2(s)\, \{f(t)+s^T\mathbf{H}^T\nabla_f(t)\}\, ds \\ &\approx \frac{1}{n\det(\mathbf{H})}\|\mathcal{K}\|_2^2\, f(t), \end{aligned}$$

with $\|\mathcal{K}\|_2$ denoting the q–dimensional L_2–norm of $\mathcal{K}$. Hence

$$AIV(\mathbf{H}) = \frac{1}{n\det(\mathbf{H})}\|\mathcal{K}\|_2^2$$

and in summary we get the following *AMISE* formula for the multivariate kernel density estimator

$$AMISE(\mathbf{H}) = \frac{1}{4}\mu_2^2(\mathcal{K})\int \left[\mathrm{tr}\{\mathbf{H}^T\mathcal{H}_f(t)\mathbf{H}\}\right]^2\, dt + \frac{1}{n\det(\mathbf{H})}\|\mathcal{K}\|_2^2. \qquad (1.7)$$

Let us now turn to the problem how to choose the *AMISE* optimal bandwidth. Again this is the bandwidth which balances bias–variance tradeoff in *AMISE*. Denote h a scalar, such that $\mathbf{H} = h\mathbf{H}_0$ and $\det(\mathbf{H}_0) = 1$. Then *AMISE* can be written as

$$AMISE(\mathbf{H}) = \frac{1}{4}h^4\,\mu_2^2(\mathcal{K})\int \left[\mathrm{tr}\{\mathbf{H}_0^T\mathcal{H}_f(t)\mathbf{H}_0\}\right]^2\, dt + \frac{1}{nh^q}\|\mathcal{K}\|_2^2.$$

If we only allow changes in h the optimal orders for the smoothing parameter h and *AMISE* are

$$h_0 = O(n^{-1/(4+q)}), \quad AMISE(h_0\mathbf{H}_0) = O(n^{-4/(4+q)}).$$

Hence, this density estimator has a rather slow rate of convergence, especially if q is large. If we consider $\mathbf{H} = h\mathbf{I}_q$ (the same bandwidth in all q dimensions) and we fix the sample size n, then the *AMISE* optimal bandwidth has to be considerably larger than in the one–dimensional case to make sure that the estimate has reasonably small variability. Some ideas of comparable sample sizes to reach the same quality of the density estimates over different dimensions can be found in Silverman (1986), p. 94, and Scott and Wand (1991). Moreover, the computational effort of this technique increases with the number of dimensions q. Therefore, multidimensional density estimation is usually not practically applied if $q \geq 5$.

Smoothing and Regression, Approaches, Computation and Application
Chapter 12 M.Schimek (ed), Wiley Publishers,357-391

1.2.1.2 Bandwidth selection and graphical representation The problem of an automatic, data-driven choice of the bandwidth $\mathbf{H}$ is of great importance in the multivariate case. In one or two dimensions we may choose an "appropriate" bandwidth interactively by looking at the sequence of density estimates for different bandwidths. But how can this be done in three, four or more dimensions? The problem of graphical representation arises, which we address next.

Theoretically the bandwidth selection problem can be handled as in the one–dimensional case. Typically, one searches for a global bandwidth $\mathbf{H}$ or a local bandwidth $\mathbf{H}(t)$. Two approaches are frequently used in both cases

- plug–in bandwidths, in particular "rule–of–thumb" bandwidths,
- resampling methods, in particular cross–validation and bootstrap.

We will introduce generalizations for Silverman's rule–of–thumb and least squares cross–validation to stress the analogy with the one–dimensional bandwidth selectors.

Rule–of–thumb bandwidth Rule–of–thumb bandwidth selection provides a formula arising from a reference distribution. Obviously, the pdf of a multivariate normal distribution $N_q(\mu, \mathbf{\Sigma})$ is a good candidate for a reference distribution in the multivariate case. Suppose that the kernel $\mathcal{K}$ is Gaussian, i.e. the pdf of $N_q(0_q, \mathbf{I}_q)$. Note that $\mu_2(\mathcal{K}) = 1$ and $\|\mathcal{K}\|_2^2 = 2^{-q}\pi^{-q/2}$ in this case. Hence, from (1.7) and the fact that

$$\begin{aligned}&\int [\mathrm{tr}\{\mathbf{H}^T \mathcal{H}_f(t) \mathbf{H}\}]^2 \, dt \\ &= \frac{1}{2^{q+2}\pi^{q/2} \det(\mathbf{\Sigma})^{1/2}} \left[2\,\mathrm{tr}(\mathbf{H}^T \mathbf{\Sigma}^{-1} \mathbf{H})^2 + \{\mathrm{tr}(\mathbf{H}^T \mathbf{\Sigma}^{-1} \mathbf{H})\}^2\right]\end{aligned}$$

we can easily derive rule–of–thumb formulae for different assumptions on $\mathbf{H}$ and $\mathbf{\Sigma}$.

In the simplest case, i.e. that we consider $\mathbf{H}$ and $\mathbf{\Sigma}$ to be diagonal matrices $\mathbf{H} = \mathrm{diag}(h_1, \ldots, h_q)$ and $\mathbf{\Sigma} = \mathrm{diag}(\sigma_1^2, \ldots, \sigma_q^2)$, this leads to

$$\widetilde{h}_j = \left(\frac{4}{q+2}\right)^{1/(q+4)} n^{-1/(q+4)} \sigma_j. \tag{1.8}$$

Note that this formula coincides with Silverman's rule–of–thumb in the case $q = 1$, see Silverman (1986), p. 45. Replacing the σ_j's by estimates and noting the first factor is always between 0.924 and 1.059, we arrive at Scott's rule

$$\widehat{h}_j = n^{-1/(q+4)} \widehat{\sigma}_j,$$

see Scott (1992), p. 152.

Smoothing and Regression, Approaches, Computation and Application
Chapter 12 M.Schimek (ed), Wiley Publishers,357-391

It is difficult to derive the rule–of–thumb for general $\mathbf{H}$ and $\mathbf{\Sigma}$. However, (1.8) shows that it might be a good idea to choose the bandwidth matrix $\mathbf{H}$ proportional to $\mathbf{\Sigma}^{1/2}$. In this case we get as generalization of Scott's rule

$$\widehat{\mathbf{H}} = n^{-1/(q+4)} \widehat{\mathbf{\Sigma}}^{1/2}. \tag{1.9}$$

We remark that this rule is equivalent to applying a Mahalanobis transformation on the data (to transform the estimated covariance matrix to identity), then to compute the kernel estimate with equal bandwidths $h = n^{1/(q+4)}$ and finally to retransform the estimated pdf back to the original scale.

But before we go on with applications, let us consider what we can do, if we want to use a kernel different from the Gaussian. The idea of canonical kernels by Marron and Nolan (1988) can be easily extended to the multivariate case. Consider a kernel $\mathcal{K}$ and all equivalent kernel functions $\mathcal{K}_\delta = \delta^{-1}\mathcal{K}(\bullet/\delta)$ with $\delta \geq 0$. Although δ is a scalar, it is working on the q–variate argument of $\mathcal{K}$. Now we have $\|\mathcal{K}_\delta\|_2^2 = \delta^{-q}\|\mathcal{K}\|_2^2$ and $\mu_2(\mathcal{K}_\delta) = \delta^2\mu_2(\mathcal{K})$. As in the one–dimensional case we choose δ such that the bias-variance tradeoff in $AMISE(\mathbf{H}, \mathcal{K}_\delta)$ is independent of $\mathcal{K}_\delta$. This yields

$$\mu_2^2(\mathcal{K}_{\delta_0}) = \|\mathcal{K}_{\delta_0}\|_2^2 \iff \delta_0 = \left\{ \frac{\|\mathcal{K}\|_2^2}{\mu_2^2(\mathcal{K})} \right\}^{1/(q+4)}.$$

δ_0 again is called *canonical bandwidth* of the kernel $\mathcal{K}$. Denote now $\mathcal{K}^A$ a kernel function with canonical bandwidth δ_0^A and $\mathcal{K}^B$ a kernel function with canonical bandwidth δ_0^B. Suppose we have used $\mathbf{H}_A$ with kernel $\mathcal{K}^A$ and we want to recompute the kernel density estimate with kernel $\mathcal{K}^B$. Then it holds

$$AMISE(\mathbf{H}_A, \mathcal{K}^A) \approx AMISE(\mathbf{H}_B, \mathcal{K}^B)$$

if

$$\mathbf{H}_B = \frac{\delta_0^B}{\delta_0^A} \mathbf{H}_A,$$

which allows to adjust bandwidths for different kernel as in the one–dimensional case.

Let us consider an example. Suppose we want to use the product Quartic kernel $\mathcal{K}^Q$ instead of the q-dimensional Gaussian $\mathcal{K}^G$ which is faster in direct computation because of its compact support on $[-1, 1]$. Which is the equivalent rule–of–thumb to (1.9) in this case? Here we have $\delta_0^G = \{1/(2\sqrt{\pi})\}^{q/(q+4)}$ and $\delta_0^Q = (49 \cdot 5^q/7^q)^{1/(q+4)}$ which gives the canonical bandwidths in Table 1.1 for dimensions $q = 1, \ldots, 5$.

The fourth column of Table 1.1 gives the factor which the rule–of–thumb bandwidth matrix in (1.9) needs to be multiplied with to obtain the rule–of–thumb bandwidth for the multiplicative Quartic kernel. Of course all rule–of–thumb bandwidths for other kernel functions can be calculated in a similar way.

Smoothing and Regression, Approaches, Computation and Application
Chapter 12 M.Schimek (ed), Wiley Publishers,357-391

q	δ_0^G	δ_0^Q	δ_0^Q/δ_0^G
1	0.7764	2.0362	2.6226
2	0.6558	1.7100	2.6073
3	0.5814	1.5095	2.5964
4	0.5311	1.3747	2.5883
5	0.4951	1.2783	2.5820

Table 1.1 Bandwidth adjusting factors for Gaussian and multiplicative Quartic Kernel for different dimensions q.

For a product kernel $\mathcal{K}$, constructed from an univariate kernel K, $\mu_2(\mathcal{K}) = \mu_2(K)$ and $\|\mathcal{K}\|_2 = \|K\|_2^q$. A table of values $\mu_2(K)$, $\|K\|_2^2$ can be found in Härdle (1991), p. 239, for example.

Principally, all plug–in methods for the one–dimensional kernel density estimation can be extended to the multivariate case. See Wand and Jones (1994) for details on multivariate plug–in bandwidth selection.

Cross–validation As we mentioned before, the cross–validation method is fairly independent of the special structure of the parameter or function estimate. Considering the bandwidth choice problem, cross–validation techniques allow to adapt to a wider class of density functions f than the rule–of–thumb approach. (Remember that the rule–of–thumb bandwidth is optimal for the reference pdf, hence it may fail for multimodal densities for instance.)

Recall, that in contrast to the rule–of–thumb approach, least squares cross–validation for density estimation aims to estimate the *ISE* optimal bandwidth. Here we approximate the integrated squared error

$$\begin{aligned} ISE(\mathbf{H}) &= \int \{\hat{f}_{\mathbf{H}}(t) - f(t)\}^2 \, dt \\ &= \int \hat{f}_{\mathbf{H}}^2(t) \, dt - 2 \int \hat{f}_{\mathbf{H}}(t) f(t) \, dt + \int f^2(t) \, dt. \end{aligned} \tag{1.10}$$

Apparently, this is the same formula as in the the one–dimensional case and, since the last term of(1.10) does not involve $\mathbf{H}$, it can be ignored. The first term can be easily calculated from the data. Only the second term of (1.10) is unknown and has to be estimated. However, observe that $\int \hat{f}_{\mathbf{H}}(t) f(t) \, dt = E\, \hat{f}_{\mathbf{H}}(T)$, where the only new aspect now is that T is q–dimensional. As in the one–dimensional case we estimate this term by a leave–one–out estimator

$$\widehat{E\, \hat{f}_{\mathbf{H}}(T)} = \frac{1}{n} \sum_{i=1}^{n} \hat{f}_{\mathbf{H},-i}(T_i)$$

Smoothing and Regression, Approaches, Computation and Application
Chapter 12 M.Schimek (ed), Wiley Publishers,357-391

where

$$\widehat{f}_{\mathbf{H},-i}(t) = \frac{1}{n-1} \sum_{i \neq j, j=1}^{n} \mathcal{K}_{\mathbf{H}}(T_j - t).$$

This yields the multivariate cross–validation criterion as a straightforward generalization of CV in the one–dimensional case:

$$\begin{aligned} CV(\mathbf{H}) &= \frac{1}{n^2 \det(\mathbf{H})} \sum_{i=1}^{n} \sum_{j=1}^{n} \mathcal{K} \star \mathcal{K} \left\{ \mathbf{H}^{-1}(T_j - t_i) \right\} \\ &\quad - \frac{2}{n(n-1)} \sum_{i=1}^{n} \sum_{\substack{j=1 \\ j \neq i}}^{n} \mathcal{K}_{\mathbf{H}}(T_j - T_i). \end{aligned}$$

$\mathcal{K} \star \mathcal{K}$ denotes the convolution of $\mathcal{K}$ with itself. The difficulty comes in by the fact that the bandwidth is now a $q \times q$ matrix $\mathbf{H}$. In the most general case, this means, we have to minimize over $q(q+1)/2$ parameters. Still, if we assume $\mathbf{H}$ to be a diagonal matrix, this remains a q–dimensional optimization problem. This holds as well for other cross–validation approaches. Multivariate resampling methods for bandwidth selection are discussed in more detail in Sain, Baggerly and Scott (1994).

Graphical representation Consider now the problem to graphically display a multivariate density estimate. Assume first $q = 2$. Here we are still able to show the density estimate in a 3-dimensional plot. This is in particular useful if the estimated function can be rotated on the computer screen interactively. For a two-dimensional presentation a contour plot often gives more insight to the structure of the data.

In the following, we will use the credit data from Fahrmeir and Hamerle (1984), Fahrmeir and Tutz (1994) for illustration. This data set consists of $n = 1000$ clients, 700 paid a credit back without problems, 300 did not. Among a number of categorical variables (running account, previous credits, purpose, personal attributes etc.) three continuous variables are available: duration, amount of credit, and age.

Figures 1.3, 1.4 (upper panels) display a two-dimensional density estimate

$$\widehat{f}_h(t) = \widehat{f}_h(t_1, t_2)$$

for log(duration), log(amount) and log(amount), log(age, respectively. We use the subscript h to indicate that we used a diagonal bandwidth matrix $\mathbf{H} = \mathrm{diag}(h_1, h_2)$.

Additionally, Figures 1.3, 1.4 (lower panels) gives contour plots of these density estimates. It is easily observed, that both distributions are rather symmetric. This is due to the logarithmic transformation. In the duration direction a typical bimodal structure can be recognized. This slightly repro-

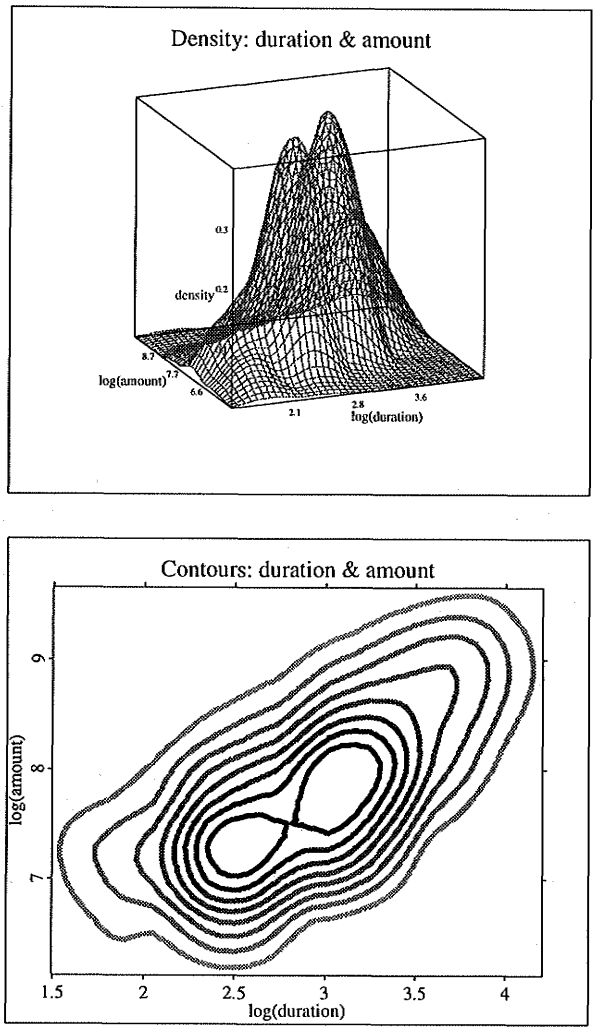


Fig. 1.3 Two-dimensional density estimate (upper panel) and density contours (lower panel) for duration and amount. Rule-of-thumb bandwidths $h_1 = 0.48$, $h_2 = 0.64$. Credit data, Fahrmeir and Hamerle (1984).

duces in the amount direction. Obviously, both variables are related with positive correlation.

Here, the bandwidth was chosen accordingly to the general "rule-of-thumb" (1.9), which tends to oversmooth multimodal structures of the data. In fact, the durations of credits are multiples of 6 months in most case. The two clear

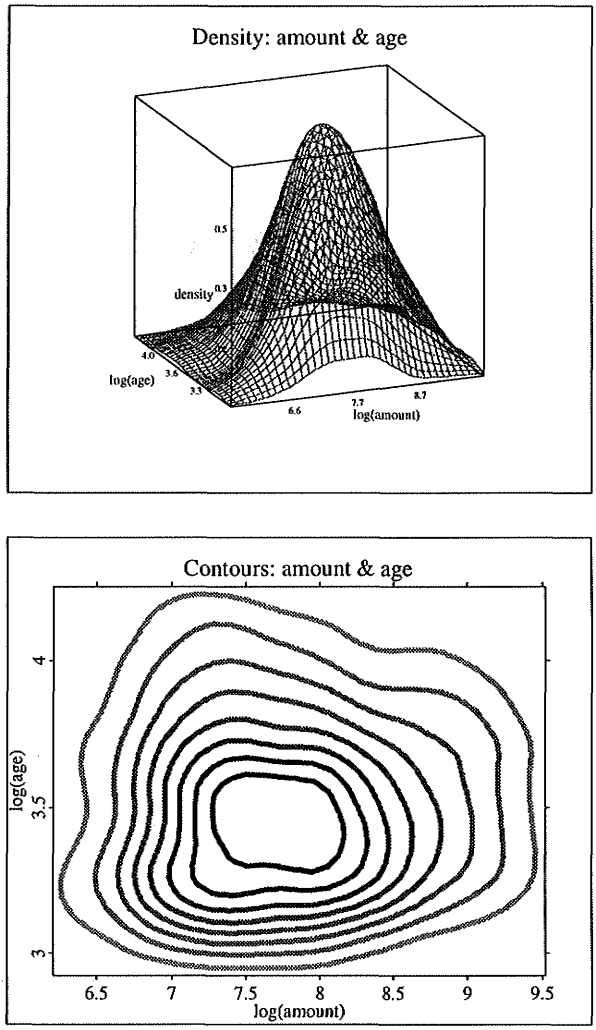


Fig. 1.4 Two-dimensional density estimate (upper panel) and density contours (lower panel) for amount and age. Rule-of-thumb bandwidths $h_1 = 0.64$, $h_2 = 0.25$. Credit data, Fahrmeir and Hamerle (1984).

modes that we observe are those for durations 12 and 24 months. In all applications of this paper we use the Quartic (Biweight) product kernel. Recall that the the univariate Quartic kernel is $K(u) = 0.9375(1-u^2)^2 \cdot \mathrm{I}(|u| \leq 1)$.

For three-dimensional density estimates, it is always possible to hold one variable fixed and to plot the density function only in dependence of the other

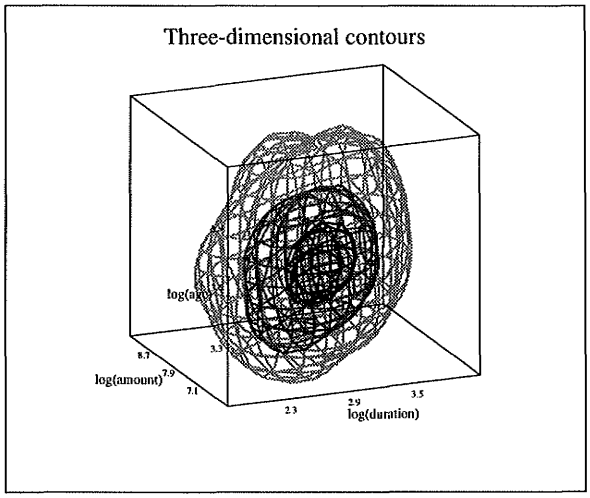


Fig. 1.5 Three-dimensional density contours for duration, amount and age. Rule-of-thumb bandwidths $h_1 = 0.56$, $h_2 = 0.75$, $h_3 = 0.29$. Credit data, Fahrmeir and Hamerle (1984).

variables. Alternatively, we can again plot contours of the density estimate, which are now three-dimensional surfaces. Figure 1.5 shows this for the credit scoring variables. In the original version of this plot, red, green and blue surfaces show the values of the density estimate at the levels (in percent) indicated on the right. Colors and the possibility to rotate the contours on the computer screen eases the exploration of the data structures. Of course, we are restricted to two-dimensional plots here. However, one can clearly recognize the ellipsoidal structure of the contour which indicates a relatively symmetric distribution.

1.2.2 Multivariate kernel regression

Multivariate nonparametric regression aims to estimate the functional relation between a response variable Y and a multivariate explanatory variable T, i.e. the conditional expectation

$$E(Y|T) = E\left(Y|T_1, \ldots, T_q\right) = m(T),$$

where as before $T = (T_1, \ldots, T_d)^T$. The relation

$$E(Y|T) = \int y f(y|t)\, dy = \frac{\int y f(y,t)\, dy}{f(x)}$$

Smoothing and Regression, Approaches, Computation and Application
Chapter 12 M.Schimek (ed), Wiley Publishers,357-391

leads by replacing the multivariate densities $f(y,t)$ by the kernel density estimate

$$\widehat{f}_{h,\mathbf{H}}(y,t) = \frac{1}{n}\sum_{i=1}^{n} K_h(Y_i - y)\,\mathcal{K}_{\mathbf{H}}(t_i - t)$$

and $f(t) = f_T(t)$ by (1.3) to the multivariate generalization of the Nadaraya–Watson estimator:

$$\widehat{m}_{\mathbf{H}}(t) = \frac{\sum_{i=1}^{n} \mathcal{K}_{\mathbf{H}}(T_i - t)\,Y_i}{\sum_{i=1}^{n} \mathcal{K}_{\mathbf{H}}(T_i - t)}..$$

Hence, the multivariate kernel regression estimator is just a weighted sum of the observed responses Y_i. The denominator ensures that the weights sum up to 1. Depending on the choice of the kernel, $\widehat{m}_{\mathbf{H}}(t)$ is a weighted average of those Y_i where T_i lies in a ball or cube around t.

Note that the multivariate Nadaraya–Watson estimator is a local constant estimator, i.e. the solution of

$$\min_{\beta_0} \sum_{i=1}^{n} \{Y_i - \beta_0\}^2 \mathcal{K}_{\mathbf{H}}(T_i - t).$$

Replacing β_0 by a polynomial in $T_i - t$ yields a local polynomial kernel regression estimator. This definition of a local polynomial kernel regression is a straightforward generalization of the univariate case. For details see Ruppert and Wand (1994). Let us illustrate this with the example of a local linear regression estimate. The minimization problem is

$$\min_{\beta_0,\beta_1} \sum_{i=1}^{n} \left\{Y_i - \beta_0 - (T_i - t)^T\beta_1\right\}^2 \mathcal{K}_{\mathbf{H}}(T_i - t).$$

The solution of the problem can be written as

$$\widehat{\beta} = (\widehat{\beta}_0, \widehat{\beta}_1^T)^T = \left(\mathbf{T}^T\mathbf{W}\mathbf{T}\right)^{-1}\mathbf{T}^T\mathbf{W}\mathbf{Y} \tag{1.11}$$

using the notations

$$\mathbf{T} = \begin{pmatrix} 1 & (T_1 - t)^T \\ \vdots & \vdots \\ 1 & (T_n - t)^T \end{pmatrix}, \quad \mathbf{Y} = \begin{pmatrix} Y_1 \\ \vdots \\ Y_n \end{pmatrix},$$

and $\mathbf{W} = \mathrm{diag}\left(\mathcal{K}_{\mathbf{H}}(T_1 - t), \ldots, \mathcal{K}_{\mathbf{H}}(T_n - t)\right)$. In (1.11) $\widehat{\beta}_0$ estimates the regression function itself, whereas $\widehat{\beta}_1$ estimates the partial derivatives w.r.t. the components T. In the following we denote the multivariate local linear

Smoothing and Regression, Approaches, Computation and Application
Chapter 12 M.Schimek (ed), Wiley Publishers,357-391

estimator as

$$\widehat{m}_{1,\mathbf{H}}(t) = \widehat{\beta}_0(t).$$

1.2.2.1 Bias, variance and asymptotics The asymptotic conditional variance of the Nadaraya–Watson estimator $\widehat{m}_{\mathbf{H}}$ and the local linear $\widehat{m}_{1,\mathbf{H}}$ is identical and its derivation can be found in detail in Ruppert and Wand (1994):

$$Var\left\{\widehat{m}_{\mathbf{H}}(t)|T_1,\ldots,T_n\right\} = \frac{1}{n\det(\mathbf{H})}\|\mathcal{K}\|_2^2\frac{\sigma^2(t)}{f(t)}\{1+o_p(1)\},$$

with $\sigma^2(t)$ denoting the variance function in $Var(Y|t)$.

We sketch the derivation of the asymptotic conditional bias since we find remarkable differences between both estimators. Denote M the second order Taylor expansion of $(m(T_1),\ldots,m(T_n))^T$, i.e.

$$M \approx m(t)\mathbf{1}_n + L(t) + \frac{1}{2}Q(t) = \mathbf{T}\begin{pmatrix} m(t) \\ \nabla_m(t)\end{pmatrix} + \frac{1}{2}Q(t), \tag{1.12}$$

with

$$L(t) = \begin{pmatrix} (T_1-t)^T\nabla_m(t) \\ \vdots \\ (T_n-t)^T\nabla_m(t)\end{pmatrix}, \quad Q(t) = \begin{pmatrix} (T_1-t)^T\mathcal{H}_m(t)(T_1-t) \\ \vdots \\ (T_n-t)^T\mathcal{H}_m(t)(T_n-t)\end{pmatrix}.$$

Additionally to (1.5) it can be shown that

$$\begin{aligned}\frac{1}{n}\sum_{i=1}^n \mathcal{K}_{\mathbf{H}}(T_i-t)\,(T_i-t) &= \mu_2(\mathcal{K})\mathbf{H}\mathbf{H}^T\nabla_f(t) + o_p(\mathbf{H}\mathbf{H}^T\mathbf{1}_d),\\ \frac{1}{n}\sum_{i=1}^n \mathcal{K}_{\mathbf{H}}(T_i-t)\,(T_i-t)(T_i-t)^T &= \mu_2(\mathcal{K})f(t)\mathbf{H}\mathbf{H}^T\nabla_f(t) + o_p(\mathbf{H}\mathbf{H}^T),\end{aligned}$$

see Ruppert and Wand (1994). Therefore the denominator of the conditional asymptotic expectation of the Nadaraya–Watson estimator $\widehat{m}_{\mathbf{H}}$ is approximately $f(t)$. Using $E(\mathbf{Y}|T_1,\ldots,T_n) = M$ and the Taylor expansion for M we have

$$\begin{aligned}&E\left\{\widehat{m}_{\mathbf{H}}|T_1,\ldots,T_n\right\}\\ &\approx \{f(t)+o_p(1)\}^{-1}\Big\{\frac{1}{n}\sum_{i=1}^n \mathcal{K}_{\mathbf{H}}(T_i-t)m(t) + \sum_{i=1}^n \mathcal{K}_{\mathbf{H}}(T_i-t)(T_i-t)^T\nabla_m(t)\\ &\qquad + \sum_{i=1}^n \mathcal{K}_{\mathbf{H}}(T_i-t)(T_i-t)^T\mathcal{H}_m(t)(T_i-t)\Big\}\end{aligned}$$

Smoothing and Regression, Approaches, Computation and Application
Chapter 12 M.Schimek (ed), Wiley Publishers,357-391

$$\approx \{f(t)\}^{-1}\left[f(t)m(t) + \mu_2(\mathcal{K})\nabla_m \mathbf{H}\mathbf{H}^T\nabla_f + \frac{1}{2}\mu_2(\mathcal{K})f(t)\,\mathrm{tr}\{\mathbf{H}^T\mathcal{H}_m(t)\mathbf{H}\}\right].$$

We use $\approx$ to indicate asymptotic equality. The results for the Nadaraya–Watson estimator are summarized in the following theorem.

THEOREM 1
The conditional asymptotic bias and variance of the multivariate Nadaraya–Watson kernel regression estimator are

$$\begin{aligned} E\{\widehat{m}_{\mathbf{H}}|T_1,\ldots,T_n\} - m(t) &\approx \mu_2(\mathcal{K})\frac{\nabla_m(t)^T\mathbf{H}\mathbf{H}^T\nabla_f(t)}{f(t)} \\ &\quad + \frac{1}{2}\mu_2(\mathcal{K})\, tr\{\mathbf{H}^T\mathcal{H}_m(t)\mathbf{H}\} \\ Var\{\widehat{m}_{\mathbf{H}}|T_1,\ldots,T_n\} &\approx \frac{1}{n\det(\mathbf{H})}\|\mathcal{K}\|_2^2\frac{\sigma^2(t)}{f(t)} \end{aligned}$$

in the interior of the support of f_T.

Recall the notation $e_1 = (1,0,\ldots,0)^T$ for the first unit vector in $\mathbb{R}^d$. Then we can write the local linear estimator as

$$\widehat{m}_{1,\mathbf{H}}(t) = e_1^T\left(\mathbf{T}^T\mathbf{W}\mathbf{T}\right)^{-1}\mathbf{T}^T\mathbf{W}\mathbf{Y}.$$

Now we have using (1.11) and (1.12)

$$\begin{aligned} & E\{\widehat{m}_{1,\mathbf{H}}|T_1,\ldots,T_n\} - m(t) \\ &= e_1^T\left(\mathbf{T}^T\mathbf{W}\mathbf{T}\right)^{-1}\mathbf{T}^T\mathbf{W}\mathbf{T}\left\{\begin{pmatrix} m(t) \\ \nabla_m(t)\end{pmatrix} + \frac{1}{2}Q(t)\right\} - m(t) \\ &= \frac{1}{2}e_1^T\left(\mathbf{T}^T\mathbf{W}\mathbf{T}\right)^{-1}\mathbf{T}^T\mathbf{W}Q(t) \end{aligned}$$

since $e_1^T[m(t),\nabla_m(t)^T] = m(t)$. Hence, the numerator of the asymptotic conditional bias only depends on the quadratic term. This is one of the key points in asymptotics for local polynomial estimators. If we would use local polynomials of order d and expand M up to order $d+1$, then only the term of order $d+1$ would appear in the numerator of the asymptotic conditional bias. Of course, this leads to a more complicated structure for the denominator.

THEOREM 2
The conditional asymptotic bias and variance of the multivariate local linear regression estimator are

$$\begin{aligned} E\{\widehat{m}_{1,\mathbf{H}}|T_1,\ldots,T_n\} - m(t) &\approx \frac{1}{2}\mu_2(\mathcal{K})\, tr\{\mathbf{H}^T\mathcal{H}_m(t)\mathbf{H}\} \\ Var\{\widehat{m}_{1,\mathbf{H}}|T_1,\ldots,T_n\} &\approx \frac{1}{n\det(\mathbf{H})}\|\mathcal{K}\|_2^2\frac{\sigma^2(t)}{f(t)} \end{aligned}$$

Smoothing and Regression, Approaches, Computation and Application
Chapter 12 M.Schimek (ed), Wiley Publishers,357-391

in the interior of the support of f_T.

For the proof of Theorem 2 we refer again to Ruppert and Wand (1994). They also show that the local linear estimate has same order conditional bias in the interior as well as in the boundary of the support of f_T. Fan, Gasser, Gijbels, Brockmann and Engel (1993) point out that the multivariate local linear fit with Epanechnikov kernel is a best linear estimator and has a minimax efficiency of at least 89.4% among all estimators.

1.2.2.2 Bandwidth selection and practical aspects Principally, the methods to choose a smoothing parameter in nonparametric regression are the same as in density estimation. Again, plug–in and resampling ideas are employed for finding a global bandwidth $\mathbf{H}$ or a local bandwidth $\mathbf{H}(t)$.

For our presentation, we concentrate on the classical cross–validation bandwidth selector. As a motivation, we introduce the *residual sum of squares* (*RSS*) as a (naive) way to asses the goodness of fit

$$RSS(\mathbf{H}) = n^{-1} \sum_{i=1}^{n} \{Y_i - \widehat{m}_{\mathbf{H}}(X_i)\}^2 , \tag{1.13}$$

which is also called the resubstitution estimate for the *averaged squared error* (*ASE*). Note, that we concentrate on the Nadaraya–Watson estimator at the moment.

There is a problem with the *RSS*: Y_i is used in $\widehat{m}_{\mathbf{H}}(X_i)$ to predict itself. As a consequence, $ASE(\mathbf{H})$ can be made arbitrarily small by letting $\mathbf{H} \to 0$ (in which case $\widehat{m}_{\mathbf{H}}$ is an interpolation of the Y_i's). This leads to the *cross–validation* function

$$CV(\mathbf{H}) = n^{-1} \sum_{i=1}^{n} \{Y_i - \widehat{m}_{\mathbf{H},-i}(X_i)\}^2$$

This function replaces $\widehat{m}_{\mathbf{H}}(X_i)$ in (1.13) with the *leave–one–out*-estimator

$$\widehat{m}_{\mathbf{H},-i}(X_i) = \frac{\sum_{j \neq i} \mathcal{K}_{\mathbf{H}}(X_i - X_j) Y_j}{\sum_{j \neq i} \mathcal{K}_{\mathbf{H}}(X_i - X_j)}$$

and is equivalent to multiplying each term in $RSS(\mathbf{H})$ by a *penalizing function* that is correcting for the downward bias of the resubstitution estimate. For the Nadaraya–Watson estimator

$$\begin{aligned} CV(\mathbf{H}) &= \frac{1}{n} \sum_{i=1}^{n} \{Y_i - \widehat{m}_{\mathbf{H},-i}(X_i)\}^2 \\ &= \frac{1}{n} \sum_{i=1}^{n} \{Y_i - \widehat{m}_{\mathbf{H}}(X_i)\}^2 \left\{ \frac{Y_i - \widehat{m}_{\mathbf{H},-i}(X_i)}{Y_i - \widehat{m}_{\mathbf{H}}(X_i)} \right\}^2 \end{aligned} \tag{1.14}$$

Smoothing and Regression, Approaches, Computation and Application
Chapter 12 M.Schimek (ed), Wiley Publishers,357-391

and

$$\begin{aligned}\frac{Y_i - \widehat{m}_{\mathbf{H}}(X_i)}{Y_i - \widehat{m}_{\mathbf{H},-i}(X_i)} &= \frac{\sum_j \mathcal{K}_{\mathbf{H}}(X_i - X_j)Y_j - Y_i \sum_j \mathcal{K}_{\mathbf{H}}(X_i - X_j)}{\sum_{j\neq i} \mathcal{K}_{\mathbf{H}}(X_i - X_j)Y_j - Y_i \sum_{j\neq i} \mathcal{K}_{\mathbf{H}}(X_i - X_j)} \\ &\quad \cdot \frac{\sum_{j\neq i} \mathcal{K}_{\mathbf{H}}(X_i - X_j)}{\sum_j \mathcal{K}_{\mathbf{H}}(X_i - X_j)} \\ &= 1 - \frac{\mathcal{K}_{\mathbf{H}}(0)}{\sum_j \mathcal{K}_{\mathbf{H}}(X_i - X_j)}. \end{aligned} \tag{1.15}$$

Note that (1.15) is a function of the i–th diagonal element of the smoother matrix. In this case, cross–validation is equivalent with *generalized cross–validation* (Craven and Wahba, 1979). Härdle, Hall and Marron (1988) show asymptotic optimality of the selected bandwidth, although the rate of convergence is rather slow. An improved bandwidth selection is discussed in Härdle, Hall and Marron (1992).

We want to remark that (1.14) and (1.15) also imply that the computation of $CV(\mathbf{H})$ does not require more computational effort than the computation of $m_{\mathbf{H}}(X_1), \ldots, m_{\mathbf{H}}(X_n)$. However, the optimization over a matrix $\mathbf{H}$ might be cumbersome, hence diagonal bandwidth matrices (or even $\mathbf{H} = h\mathbf{I}_q$ with appropriate standardization of the data) are still preferred in practice.

Before we consider cross–validation bandwidth selection in the local linear case, we want to comment on the practical computation of the estimator. Principally, since multivariate kernel regression estimators can be expressed as local polynomial estimators, their computation can be done by any statistical package that is able to run weighted least squares regression. However, since we estimate a function, this weighted least squares regression has to be performed in all observation points or on a grid of points in $\mathbb{R}^q$. Therefore, explicit formulae are useful.

We will give a formula for the multivariate local linear estimator in the following. For a fixed point t consider the sums

$$\begin{aligned}\mathcal{S}_0 = \mathcal{S}_0(t) &= \sum_{i=1}^n \mathcal{K}_{\mathbf{H}}(T_i - t) \\ \mathcal{S}_1 = \mathcal{S}_1(t) &= \sum_{i=1}^n \mathcal{K}_{\mathbf{H}}(T_i - t)(T_i - t) \\ \mathcal{S}_2 = \mathcal{S}_2(t) &= \sum_{i=1}^n \mathcal{K}_{\mathbf{H}}(T_i - t)(T_i - t)(T_i - t)^T\end{aligned}$$

Smoothing and Regression, Approaches, Computation and Application
Chapter 12 M.Schimek (ed), Wiley Publishers,357-391

$$\mathcal{T}_0 = \mathcal{T}_0(t) = \sum_{i=1}^{n} \mathcal{K}_{\mathbf{H}}(T_i - t) Y_i$$
$$\mathcal{T}_1 = \mathcal{T}_1(t) = \sum_{i=1}^{n} \mathcal{K}_{\mathbf{H}}(T_i - t)(T_i - t) Y_i.$$

Note that $\mathcal{S}_1$ and $\mathcal{T}_1$ are q–variate vectors and that $\mathcal{S}_2$ is a $q \times q$ matrix. Then for the local linear estimate we can write

$$\widehat{\beta} = \begin{pmatrix} \mathcal{S}_0 & \mathcal{S}_1^T \\ \mathcal{S}_1 & \mathcal{S}_2 \end{pmatrix}^{-1} \begin{pmatrix} \mathcal{T}_0 \\ \mathcal{T}_1 \end{pmatrix}.$$

For the regression function we need only the first component $e_1^T \widehat{\beta}$. Applying block–wise matrix inversion we obtain

$$e_1^T \begin{pmatrix} \mathcal{S}_0 & \mathcal{S}_1^T \\ \mathcal{S}_1 & \mathcal{S}_2 \end{pmatrix}^{-1} = \left(\mathcal{S}_0 - \mathcal{S}_1^T \mathcal{S}_2^{-1} \mathcal{S}_1\right)^{-1} \cdot \begin{pmatrix} 1 & -\mathcal{S}_1^T \mathcal{S}_2^{-1} \end{pmatrix}$$

and hence

$$\widehat{m}_{1,\mathbf{H}}(t) = \frac{\mathcal{T}_0 - \mathcal{S}_1^T \mathcal{S}_2^{-1} \mathcal{T}_1}{\mathcal{S}_0 - \mathcal{S}_1^T \mathcal{S}_2^{-1} \mathcal{S}_1}. \tag{1.16}$$

The cross–validation criterion here is a weighted RSS as in (1.14). If we denote the leave–one–out estimator $\widehat{m}_{1,\mathbf{H},-i}(t)$ and define its components accordingly, we observe

$$\begin{aligned} \mathcal{S}_{0,-i} &= \mathcal{S}_0 - \mathcal{K}_{\mathbf{H}}(0), \quad \mathcal{S}_{1,-i} = \mathcal{S}_1, \quad \mathcal{S}_{2,-i} = \mathcal{S}_2 \\ \mathcal{T}_{0,-i} &= \mathcal{T}_0 - Y_i \mathcal{K}_{\mathbf{H}}(0), \quad \mathcal{T}_{1,-i} = \mathcal{T}_1. \end{aligned}$$

This means

$$\widehat{m}_{1,\mathbf{H},-i}(t) = \frac{\mathcal{T}_0 - Y_i \mathcal{K}_{\mathbf{H}}(0) - \mathcal{S}_1^T \mathcal{S}_2^{-1} \mathcal{T}_1}{\mathcal{S}_0 - \mathcal{K}_{\mathbf{H}}(0) - \mathcal{S}_1^T \mathcal{S}_2^{-1} \mathcal{S}_1}$$

which yields in analogy to (1.15)

$$\frac{Y_i - \widehat{m}_{\mathbf{H}}(X_i)}{Y_i - \widehat{m}_{\mathbf{H},-i}(X_i)} = 1 - \frac{\mathcal{K}_{\mathbf{H}}(0)}{\mathcal{S}_0 - \mathcal{S}_1^T \mathcal{S}_2^{-1} \mathcal{S}_1}. \tag{1.17}$$

As in the Nadaraya–Watson case, (1.17) is a function of the i–th diagonal element of the smoother matrix. A summary of bandwidth selection methods other than cross–validation can be found in particular in Fan and Gijbels (1995). They also cover rule–of–thumb approaches.

Recall that (1.16) estimates the regression function only in one point t. To estimate the regression plane we have to apply (1.16) on a two–dimensional grid of points. The WARPing technique (binning) described in Härdle and Scott (1992) and applied to local polynomial kernel regression by Fan and

Marron (1994), Fan and Müller (1995), can be used to speed up calculations. See also Wand (1994) for an analysis of fast computation methods for multivariate kernel estimation.

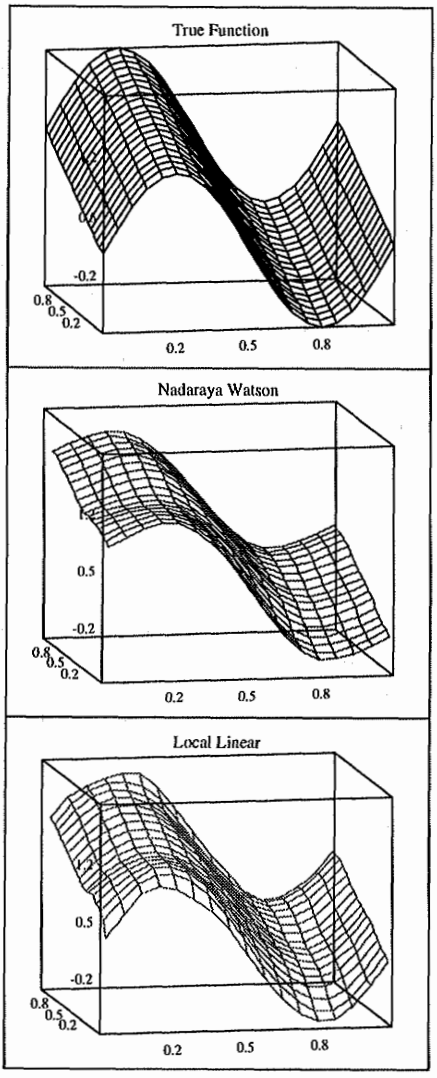


Fig. 1.6 Two-dimensional Nadaraya–Watson and local linear estimate. Simulated data.

Figure 1.6 shows the bivariate Nadaraya–Watson and local linear estimate

Smoothing and Regression, Approaches, Computation and Application
Chapter 12 M.Schimek (ed), Wiley Publishers,357-391

for simulated data. The underlying curve is in fact an additive combination of a sine function in the first and a linear function in the second argument. Note, that we have chosen the same bandwidth in both estimates. The multivariate Nadaraya–Watson regression estimator can be improved in the boundary regions by a boundary correction method. See Staniswalis, Messer and Finston (1994) and Staniswalis and Messer (1996) for more details.

Of course, nonparametric kernel regression estimation is not limited to bivariate distributions. A practical issue is the graphical display for higher dimensional multivariate functions. This was already considered when we discussed the graphical representation of multivariate density estimates. The corresponding remarks apply here again. The general problem in multivariate nonparametric estimation is the *curse of dimensionality*. Recall that the nonparametric regression estimators are based on the idea of local (weighted) averaging. In higher dimensions the observations are usually sparsely distributed for reasonable sample sizes, and consequently estimators based on local averaging perform unsatisfactorily in this situation.

Technically, one can explain this effect by looking at the *AMISE* again. Consider a multivariate regression estimator with the same bandwidth h for all components, e.g. a Nadaraya–Watson or local linear estimator with bandwidth matrix $\mathbf{H} = h\mathbf{I}_q$. Here the asymptotic *MISE* also depends on q:

$$AMISE(n,h) = \frac{1}{nh^q}C_1 + h^4 C_2.$$

where C_1 and C_2 are constants that neither depend on n nor h. If we derive the optimal bandwidth we find that $h_{opt} \sim n^{-1/(4+q)}$ and hence the rate of convergence for *AMISE* is $n^{-4/(4+q)}$. One can clearly see that the speed of convergence decreases dramatically for higher dimensions q.

1.3 SEMIPARAMETRIC GENERALIZED REGRESSION MODELS

As the name suggests, semiparametric models combine two elements, one of them to be estimated nonparametrically, the other one requiring the estimation of a set of finite dimensional parameters. In this section we concentrate on single index and generalized partial linear models.

Often a canonical partitioning of the explanatory variables exists. In particular, if there are binary or discrete explanatory variables we keep them separate from the other design variables. In the following we denote by $T = (T_1, \ldots, T_q)^T$ a vector of continuous explanatory variables and refer to $X = (X_1, \ldots, X_p)^T$ as the discrete part of the variables.

Semiparametric generalized linear models are widely used in modeling binary choice, i.e. in situations where the response variable has two alternatives.

Smoothing and Regression, Approaches, Computation and Application
Chapter 12 M.Schimek (ed), Wiley Publishers,357-391

Recall the example on credit scoring which was introduced previously. In the analysis of discrete response variables one typically models the expected value of the response as a nonlinear monotone function of a linear combination of the explanatory variables. Examples are probit or logit models where the nonlinear (link) function is the cumulative distribution function of a normal respectively logistic distribution, see McCullagh and Nelder (1989). Then the so-called *generalized linear model* has the form

$$E(Y|X,T) = G(X^T\beta + T^T\gamma), \tag{1.18}$$

with a known monotone function G and unknown parameters β and γ. The model (1.18) combines computational feasibility (especially for discrete covariates) with good interpretability of the "index" $X^T\beta + T^T\gamma$ and therefore has found wide application in all fields of applied statistics, see e.g. Fahrmeir and Tutz (1994), Maddala (1983). However, for some applications it may be argued that the assumption of (1.18) is too restrictive (Horowitz, 1993). Indeed it may not even be clear if the relationship between the influential variables and the response is monotone.

Several approaches have been proposed to generalize parametric regression models in order to allow nonmonotone relationships between explanatory variables and the dependent variable Y. We will focus on two classes of semiparametric models that have received a lot of attention.

- A generalization of the known (parametric) link function G to an unknown (nonparametric) link function $g(\bullet)$ yields the *single index model* (SIM)

  $$E(Y|X,T) = g(X^T\beta + T^T\gamma),$$

  also called a *one term projection pursuit model* in statistics. Obviously, due to the nonparametric character of the link function conventional parametric estimation procedures can no longer be applied in this case. Instead, nonparametric estimators will now be necessary. In this chapter we give an overview on how this model can be estimated using kernel methods.

- A generalization of the linear form $X^T\beta + T^T\gamma$ to a partial linear form $X^T\beta + m(T)$ yields the *generalized partial linear model* (GPLM)

  $$E(Y|X,T) = G\left\{X^T\beta + m(T)\right\},$$

  G denoting a known link function as in the GLM model. Here, the $m(\bullet)$ will be a multivariate nonparametric function of the variable T.

  In high dimensions of T the estimate of the nonparametric function $m(\bullet)$ faces the same problems as the fully nonparametric multidimensional regression function estimates: the curse of dimensionality and the practical problem of interpretability.

Härdle, W. and Müller, M. (2000) Multivariate and Semiparametric Kernel regression.

Smoothing and Regression, Approaches, Computation and Application
Chapter 12 M.Schimek (ed), Wiley Publishers,357-391

Hence it might be reasonable to think about a lower dimensional nonparametric model for the nonparametric part. A possible alternative is the GPLM with an additive structure in the nonparametric component, i.e. the *generalized additive model* (GAM).

$$E(Y|X,T) = G\left\{X^T\beta + m_1(T_1) + \ldots + m_d(T_d)\right\}.$$

Here, the $m_j(\bullet)$ will be univariate nonparametric functions of the variables T_j.

1.3.1 Generalizing the link function: single index models

Single index models derive their name from the economic term "index", a summary of different variables into one number. Meanwhile, there have been a number of methods proposed do deal with these models. A straightforward semiparametric GLM extension is provided by Weisberg and Welsh (1994). They estimated the unknown link function and its derivative (for the Fisher scoring algorithm) with an kernel smoother. Ichimura (1993) uses a similar idea within a least squares criterion. Klein and Spady (1993) show an asymptotic efficiency result for a pseudo–likelihood binary choice estimator.

All these three methods require optimization of a pseudo–likelihood of possibly complicated structure. We present here a direct approach which avoids numerical iterations. The estimation of the single index model

$$E(Y|X,T) = g(X^T\beta + T^T\gamma)$$

is carried out in two steps. First the coefficients vectors β, γ are estimated, then using the obtained index values $X_i^T\widehat{\beta} + T_i^T\widehat{\gamma}$ one can estimate g by usual univariate nonparametric regression.

1.3.1.1 Average derivative estimation Consider for a moment only the continuous part of the variables, $T = (T_1, \ldots, T_q)^T$. Denote the regression function to be estimated by $m(\bullet)$, i.e. $E(Y|T) = m(T)$. The vector of average derivatives is given by

$$\delta = E\left\{\nabla_m(T)\right\} = E\left\{g'(T^T\beta)\right\}\ \beta, \tag{1.19}$$

where $\nabla_m(t)$ is the vector of partial derivatives of $m(\bullet)$ and g' the derivative of $g(\bullet)$.

Looking at (1.19) shows that δ equals β up to scale. Hence, any estimate of δ determines β up to scale. The estimation of δ can be carried out by means of several *average derivative estimation* (ADE) methods. We will concentrate on estimators based on the density function of T, however a variety of other methods exist. For an overview see Stoker (1991).

The key idea on ADE based on the density $f(\bullet)$ of T lies in "transferring"

Smoothing and Regression, Approaches, Computation and Application
Chapter 12 M.Schimek (ed), Wiley Publishers,357-391

the derivative of the regression function m on to the derivative of the density f. Consider

$$\delta = E\{\nabla_m(T)\} = \int \nabla_m(t) f(t)\, dt.$$

If $f(t)\, m(t) \to 0$ is assumed for $\|t\| \to \infty$, then partial integration yields $E\{\nabla_m(T)\} = -\int m(t) \nabla_f(t)\, dt$. Hence by introducing the *score vector*

$$\ell(t) = \nabla_{\log f}(t) = \frac{\nabla_f(t)}{f(t)}$$

one arrives at

$$\delta = -\int \frac{\nabla_f(t)}{f(t)} f(t)\, m(t)\, dt = \int \ell(t)\, m(t)\, f(t)\, dt = E\{\ell(T)\, m(T)\}.$$

Employing $E\{\ell(T)\, m(T)\} = E\{\ell(T)\, Y\}$ immediately allows for the estimation of δ by the sample analog

$$\widehat{\delta} = n^{-1} \sum_{i=1}^{n} \widehat{\ell}_{\mathbf{H}}(T_i)\, Y_i,$$

where $\ell(t)$ is approximated by $\widehat{\ell}_{\mathbf{H}}(t) = -\{\widehat{f}_{\mathbf{H}}(t)\}^{-1} \left(-\partial_1 \widehat{f}_{\mathbf{H}}(t), \ldots, -\partial_q \widehat{f}_{\mathbf{H}}(t)\right)$. Here, $\widehat{f}_{\mathbf{H}}(t)$ is the multivariate kernel density estimator and ∂_j are the partial derivatives w.r.t. the j–th dimension (i.e. $\partial_j \widehat{f}_{\mathbf{H}}(t)$ are used for estimating the partial derivatives of the density)

$$\partial_j \widehat{f}_{\mathbf{H}}(t) = \frac{1}{n \det(\mathbf{H})} \sum_{j=1}^{n} \partial_j K_{\mathbf{H}}\,(t - T_j)\,.$$

Due to the sparseness of data in high dimensions, the use of $\widehat{f}_{\mathbf{H}}$ can also be problematic since $\widehat{\ell}_{\mathbf{H}}$ might behave bad in regions of small density. Hence Härdle and Stoker (1989) propose to use the ADE estimator

$$\widehat{\delta} = n^{-1} \sum_{i=1}^{n} \widehat{\ell}_{\mathbf{H}}(T_i)\, Y_i\, \mathrm{I}\{\widehat{f}_{\mathbf{H}}(T_i) > b_n\},$$

where $\mathrm{I}\{\widehat{f}_{\mathbf{H}}(T_i) > b_n\}$ is an indicator that excludes density values which are too small. The trimming bounds b_n are chosen such that $b_n \to 0$ for $n \to \infty$.

Härdle and Stoker (1989) have shown that

$$\sqrt{n}(\widehat{\delta} - \delta) \xrightarrow[n\to\infty]{\mathcal{L}} N(\mathbf{0}, \mathbf{\Sigma}_\delta),$$

where $\mathbf{\Sigma}_\delta$ is the covariance matrix of $\ell(T)Y + \{\nabla_m(T) - \ell(T)m(T)\}$. Note that

Smoothing and Regression, Approaches, Computation and Application
Chapter 12 M.Schimek (ed), Wiley Publishers,357-391

$\widehat{\delta}$ achieves $\sqrt{n}$-convergence, a rate that is typically achieved by parametric estimators.

The need for "trimming" the ADE is one of the problems associated with a random denominator. Random denominators also complicate the derivation of the distributional properties. These difficulties are overcome by *density weighted average derivative estimation* (WADE) of Powell, Stock and Stoker (1989). Observe that the density weighted average derivative shares the property of the (unweighted) average derivative of being proportional to the coefficient vector β in index models:

$$\delta = E\left\{\nabla_m(T)\, w(T)\right\} = E\left\{g'(T^T\beta)\, w(T)\right\}\ \beta,$$

A "natural" weight function is given by the density f itself. Calculations similar to those for the unweighted ADE with $w(t) = f(t)$ yield

$$\begin{aligned}\delta &= \int \nabla_m(t)\, f^2(t)\, dt = -2\int m(t)\nabla_f(t)\, f(t)\, dt \\ &= -2\, E\{Y\, \nabla_f(T)\}.\end{aligned}$$

Thus one may estimate β up to scale by

$$\widehat{\delta} = -\frac{2}{n}\sum_{i=1}^{n} Y_i\, (\partial_1 \widehat{f}_{\mathrm{H}}(t), \ldots, \partial_q \widehat{f}_{\mathrm{H}}(t))^T. \tag{1.20}$$

The WADE estimator defined in (1.20) shares the desirable distributional features of the ADE estimator ($\sqrt{n}$-consistency, asymptotic normality) while not requiring any trimming in practice.

Finally, an estimate for $g(\bullet)$ can be found by applying an univariate estimation method to $\widehat{\delta}^T T_i$ and Y_i. For the Nadaraya–Watson estimator, when $h \sim n^{-1/5}$, Härdle and Stoker (1989) showed the usual rate of convergence $\sqrt{nh}$ for the pointwise convergence of the regression function.

1.3.1.2 Including discrete explanatory variables By definition, derivatives can only be calculated if the variable under study is continuous. Thus, the method of weighted or unweighted ADE fails when discrete variables $X = (X_1, \ldots, X_d)^T$ needs to be included into the model. Before giving a more general solution, let us explain how the coefficient of one dichotomous variable is entered in the model. Recall the SIM

$$E(Y|T, X) = g(X^T\beta + T^T\gamma)$$

with T the continuous and X the discrete part of the covariates. In the simplest case, we suppose that is X is binary, i.e. either $X = 1$ or $X = 0$.

Smoothing and Regression, Approaches, Computation and Application
Chapter 12 M.Schimek (ed), Wiley Publishers,357-391

Then, this model can be "split" into two submodels

$$\begin{aligned} E(Y|T,X) &= g(T^T\gamma) \qquad \text{if } X=0 \\ E(Y|T,X) &= g(T^T\gamma+\beta) \quad \text{if } X=1. \end{aligned}$$

These are in fact two models to be estimated, one for $X = 0$ and one for $X = 1$. Note that γ alone could be estimated from the first equation only.

Theoretically, the same T_i can be associated with either $X_i = 0$ yielding an index value of $\gamma^T T_i$ or with $X_i = 1$ leading to an index value of $\gamma^T T_i + \beta$. Thus the difference between the two indices is exactly β. In practice finding these *horizontal* differences will be rather difficult. A very simple estimator is proposed in Korostelev and Müller (1995), using the observation, that the *integral* difference between the two link functions also equals β. Essentially, the coefficient of the binary explanatory variable can be estimated by

$$\widehat{\beta} = \widehat{J}^{(1)} - \widehat{J}^{(0)}$$

with

$$\widehat{J}^{(0)} = \sum_{i=0}^{n_0} \gamma^T (T_{i+1}^{(0)} - T_i^{(0)}) Y_i^{(0)}, \quad \widehat{J}^{(1)} = \sum_{i=0}^{n_1} \gamma^T (T_{i+1}^{(1)} - T_i^{(1)}) Y_i^{(1)},$$

where the superscripts $^{(0)}$ and $^{(1)}$ denote the observations coming from the subsamples according to $X_i = 0$ and $X_i = 1$. In the simplest case of a binary Y variable the estimator is $\sqrt{n}$-consistent and can be improved for efficiency by a one–step estimator, see Korostelev and Müller (1995).

Horowitz and Härdle (1996) extend this approach to multivariate multi–categorical X and arbitrary Y. Again, this approach is based on a split of the whole sample into subsamples according to the categories of X. Consider the thresholded link function

$$\widetilde{g} = c_o \, \mathrm{I}(g < c_o) + g \, \mathrm{I}(c_o \le g \le c_1) + c_1 \, \mathrm{I}(g > c_1).$$

Denote $x^{(k)}$ a possible realization of X, then the integrated link function conditional on $x^{(k)}$ is

$$J^{(k)} = \int_{v_o}^{v_1} \widetilde{g}(v + \beta^T x^{(k)}) \, dv,$$

where $v_o = g^{c_o}$ and $v_1 = g^{c_1}$. Now compare the integrated link functions for all X–categories $x^{(k)}$ $(k = 1, \ldots, M)$ to the first X–category $x^{(0)}$. It holds

$$J^{(k)} - J^{(0)} = (c_1 - c_0) \left\{ x^{(k)} - x^{(0)} \right\} \beta,$$

Smoothing and Regression, Approaches, Computation and Application
Chapter 12 M.Schimek (ed), Wiley Publishers,357-391

hence with

$$\Delta J = \begin{pmatrix} J^{(1)} - J^{(0)} \\ \cdots \\ J^{(M)} - J^{(0)} \end{pmatrix}, \quad \Delta x = \begin{pmatrix} x^{(1)} - x^{(0)} \\ \cdots \\ x^{(M)} - x^{(0)} \end{pmatrix}$$

one gets $\Delta J = (c_1 - c_0)\, \Delta x\, \beta$. This yields finally

$$\beta = (c_1 - c_o)^{-1} (\Delta x^T \Delta x)^{-1} \Delta x^T \Delta J \tag{1.21}$$

to determine β. The estimation of β is based on replacing $J^{(k)}$ in (1.21) by

$$\widehat{J}^{(k)} = \int_{v_o}^{v_1} \widehat{\widetilde{g}}(v + \beta^T x^{(k)})\, dv$$

with $\widehat{\widetilde{g}}$ a nonparametric estimate of the thresholded link function $\widetilde{g}$. This estimator is obtained by a univariate regression of the estimated "continuous" indices $\widehat{\gamma}^T T_i^{(k)}$ on $Y_i^{(k)}$. Horowitz and Härdle (1996) show that using a $\sqrt{n}$–consistent estimate $\widehat{\gamma}$ and a Nadaraya–Watson estimator $\widehat{\widetilde{g}}$ for $\widetilde{g}$ the estimated coefficient $\widehat{\beta}$ is itself $\sqrt{n}$–consistent and has an asymptotic normal distribution.

1.3.2 Generalizing the index: generalized partial linear models

An alternative way to incorporate an nonmonotone dependence of the response on the continuous variables is given by a *generalized partial linear model* (GPLM)

$$E(Y|X,T) = G\{X^T\beta + m(T)\}, \tag{1.22}$$

where $\beta = (\beta_1, \ldots, \beta_p)^T$ is a finite dimensional parameter and $m(\bullet)$ is a smooth function. These models allow a nonparametric inclusion of a part of the explanatory variables. In practice this might be only those continuous variables which have most influence on the dependent variable Y. In this section we will deal with the GPLM in general and shortly with *generalized partial linear partial additive models* (GAM).

Estimators for β and $m(\bullet)$ have been proposed by Hastie and Tibshirani (1990), Severini and Wong (1992), Severini and Staniswalis (1994) and Hunsberger (1994). Carroll, Fan, Gijbels and Wand (1997) proposed an extension to *generalized partial linear single index model* (GPLSIM) which uses a single index model instead of the fully nonparametric function $m(\bullet)$.

1.3.2.1 Semiparametric maximum likelihood The estimation of model (1.22) can be motivated by the fact that an estimate $\widehat{\beta}$ can be found for known m, and an estimate $\widehat{m}$ can be found for known β. An overview on different algorithms for the GPLM can be found in Müller (1997). We will concentrate here on the profile likelihood algorithm proposed by Severini and Wong

Smoothing and Regression, Approaches, Computation and Application
Chapter 12 M.Schimek (ed), Wiley Publishers,357-391

(1992) and Severini and Staniswalis (1994). An extented presentation of the backgrounds of this approach appears in Staniswalis and Thael (1997).

Define

$$\begin{aligned} \mu &= E(Y|X,T) = G\{\beta^T X + m(T)\} \\ \sigma^2 V(\mu) &= Var(Y|X,T) \end{aligned}$$

and denote by $\ell(\mu, y)$ the individual log–likelihood or quasi–likelihood function (if the distribution of Y does not belong to an exponential family). The "parametric" likelihood function

$$\mathcal{L}(\beta) = \sum_{i=1}^{n} \ell\left[G\{X_i^T\beta + m_\beta(T_i), Y_i\}\right]$$

is used to obtain $\widehat{\beta}$. A "smoothed" or "local" likelihood

$$\mathcal{L}^S(\eta) = \sum_{i=1}^{n} \mathcal{K}_{\mathbf{H}}(t - T_i)\, \ell\left\{G(X_i^T\beta + \eta, Y_i)\right\} \tag{1.23}$$

is optimized to estimate the smooth function $m_\beta(t) = \eta$ at point t. Note that the use of this smoothed likelihood function leads to the equivalent of the Nadaraya–Watson estimator $\widehat{m}_{\mathbf{H}}$ in ordinary regression. To obtain a local polynomial estimator of the nonparametric part $m(\bullet)$ we need to incorporate polynomial terms into the smoothed likelihood. In the local linear case we would use

$$\mathcal{L}^S(\eta_0, \eta_1) = \sum_{i=1}^{n} \mathcal{K}_{\mathbf{H}}(t - T_i)\, \ell\left[G\{X_i^T\beta + \eta_0 + (T_i - t)^T\eta_1, Y_i\}\right]$$

and get $m_\beta(t) = \eta_0$ at point t. Analogous to local linear regression η_1 points to the gradient of $m(\bullet)$ in t.

The computational algorithm consists in searching maxima of both likelihoods simultaneously. We stay in the framework of a Nadaraya–Watson type estimation of m. Severini and Staniswalis (1994) show that the resulting estimator $\widehat{\beta}$ is $\sqrt{n}$–consistent and asymptotically normal, and that estimators $\widehat{m} = \widehat{m}_{\widehat{\beta}}$ are consistent in supremum norm. Note that m is estimated as a function of the parametric component β which yields an asymptotically efficient estimate $\widehat{\beta}$ (Severini and Wong, 1992). The possible scale parameter σ can be estimated by

$$\widehat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}(Y_i - \widehat{\mu}_i)^2 / V(\widehat{\mu}_i),$$

where $\widehat{\mu}_i = G\{\widehat{\beta}^T X_i + \widehat{m}(T_i)\}$.

Smoothing and Regression, Approaches, Computation and Application
Chapter 12 M.Schimek (ed), Wiley Publishers,357-391

The algorithm which we will present here corresponds to that proposed in Severini and Staniswalis (1994) for some special cases of link function and distributions of Y. In order to avoid boundary effects, one can use a weight function in the convergence criterion or trimming in the estimation of β as in Severini and Staniswalis (1994).

Define $\widehat{\eta}_j(\beta) = \widehat{m}_\beta(t_j)$ and $\ell_i(u) = \ell\{(G(u), Y_i\}$. For example, in a binary response model we have $\ell_i(u) = Y_i \log G(u) + (1 - Y_i) \log\{1 - G(u)\}$. In the following, ℓ_i' and ℓ_i'' denote the derivatives of $\ell_i(u)$ with respect to u. The maximization of the smoothed quasi–likelihood (1.23) requires to solve

$$0 = \sum_{i=1}^{n} \ell_i'\{X_i^T\beta + \widehat{\eta}_j(\beta)\}\mathcal{K}_{\mathbf{H}}(T_i - T_j) \tag{1.24}$$

w.r.t. $\widehat{\eta}_j(\beta)$. In some models (in particular for identity and exponential link functions G) equation (1.24) can be solved explicitly for $\widehat{\eta}_j(\beta)$. Differentiation of (1.24) leads to an estimate for $\widehat{\eta}_j'$ as a function of β

$$\widehat{\eta}_j'(\beta) = \frac{-\sum_{i=1}^{n} \ell_i''\{X_i^T\beta + \widehat{\eta}_j(\beta)\}\mathcal{K}_{\mathbf{H}}(T_i - T_j)X_i}{\sum_{i=1}^{n} \ell_i''\{X_i^T\beta + \widehat{\eta}_j(\beta)\}\mathcal{K}_{\mathbf{H}}(T_i - T_j)}. \tag{1.25}$$

For β we have to solve

$$0 = \sum_{i=1}^{n} \ell_i'\{X_i^T\beta + \widehat{\eta}_i(\beta)\}\ \{X_i + \widehat{\eta}_i'(\beta)\}. \tag{1.26}$$

Equations (1.24)–(1.26) imply the following iterative Newton–Raphson type algorithm to find $\widehat{\beta}$ and $\widehat{m}(t_j) = \widehat{\eta}_j(\widehat{\beta})$, $j = 1, \ldots, n$.

- *initialization*
  Different strategies to obtain starting values are possible:
  - Start with $\widehat{\beta}^{(0)}$, $\widehat{\eta}_j^{(0)}$ from the parametric (GLM) fit. Higher order polynomial terms in T may be included to allow for a nonlinear function $\widehat{\eta}_j^{(0)}$.
  - Alternatively, it is possible to use $\widehat{\beta}^{(0)} = 0$ and as in GLM $\widehat{\eta}_j^{(0)} = G^{-1}\{(Y_j + \overline{Y})/2\}$ (but $\widehat{\eta}_j^{(0)} = G^{-1}\{(Y_j + 0.5)/(m+1)\}$ for binomial responses).
  - Severini and Staniswalis (1994) propose to start with $\widehat{\beta}^{(0)} = 0$ and $\widehat{\eta}_j^{(0)} = G^{-1}(Y_j)$ (with an adjustment for binomial responses).
- *updating step for* $\widehat{\eta}_j(\beta) = \widehat{m}_\beta(T_j)$

The function $\widehat{\eta}_j(\beta)$ is updated by

$$\widehat{\eta}_j^{(k+1)} = \widehat{\eta}_j^{(k)} - \frac{\sum_{i=1}^{n} \ell_i'(X_i^T\widehat{\beta}^{(k)} + \widehat{\eta}_j^{(k)})\, K_{\mathbf{H}}(T_i - T_j)}{\sum_{i=1}^{n} \ell_i''(X_i^T\widehat{\beta}^{(k)} + \widehat{\eta}_j^{(k)})\, K_{\mathbf{H}}(T_i - T_j)}.$$

- *updating step for β*
  The parameter β is updated by

$$\widehat{\beta}^{(k+1)} = \widehat{\beta}^{(k)} - \mathcal{B}^{-1} \sum_{i=1}^{n} \ell_i'(X_i^T\widehat{\beta}^{(k)} + \widehat{\eta}_i^{(k+1)})\widetilde{X}_i^{(k)}$$

with a Hessian type matrix

$$\mathcal{B} = \sum_{i=1}^{n} \ell_i''(X_i^T\widehat{\beta}^{(k)} + \widehat{\eta}_i^{(k+1)})\, \widetilde{X}_i^{(k)} \widetilde{X}_i^{(k)T}$$

and

$$\widetilde{X}_j^{(k)} = X_j - \frac{\sum_{i=1}^{n} \ell_i''(X_i^T\widehat{\beta}^{(k)} + \widehat{\eta}_j^{(k+1)})\, K_{\mathbf{H}}(T_i - T_j)\, X_i}{\sum_{i=1}^{n} \ell_i''(X_i^T\widehat{\beta}^{(k)} + \widehat{\eta}_j^{(k+1)})\, K_{\mathbf{H}}(T_i - T_j)}.$$

As an alternative, the functions $\ell_i''(u)$ can be replaced by their expectations (w.r.t. to Y) to obtain a Fisher scoring type procedure.

1.3.2.2 Practical application Let us illustrate the semiparametric estimation with the previously introduced credit scoring example, (Fahrmeir and Tutz, 1994; Fahrmeir and Hamerle, 1984). Recall that the data set consists of $n = 1000$ clients, among which 700 paid a credit back without problems and 300 did not. We define the binary variable Y with value 1 for those who paid back and 0 if not. The data set contains observations from three continuous variables (duration and amount of credit, age of client) and 17 discrete variables. It is of interest how the explanatory variables can be used to predict credit worthiness.

A parametric logit model leads to the parameter estimates listed in Table 1.2. We omit the parameter estimates for the discrete explanatory variables. The linear influence of duration is highly significant. Amount and age have no significant coefficients if we include them linearly. We will see that the insignificant coefficients are a sign for a more complex structured influence.

In a next step we fitted a generalized partially linear model according to the algorithm presented above. Here, the influence of amount and age has been fitted nonparametrically. Figure 1.7 shows the two–variate estimate $\widehat{m}$ in the upper panel. The bandwidths were chosen as 40% of the range in both

Smoothing and Regression, Approaches, Computation and Application
Chapter 12 M.Schimek (ed), Wiley Publishers,357-391

	Coeff.	(t–val.)	Coeff.	(t–val.)	Coeff.	(t–val.)
const.	-17.605	(-1.91)	-34.909	(-2.51)	–	–
duration	-0.036	(-3.85)	-0.033	(-3.48)	-0.037	(-4.23)
log(**amount**)	1.654	(1.41)	4.847	(2.57)	–	–
log(**amount**) square	–	–	-0.229	(-2.26)	–	–
log(**age**)	4.119	(1.59)	6.949	(1.11)	–	–
log(**age**) square	–	–	-0.501	(-0.60)	–	–
log(**age**)∗ log(**amount**)	-0.484	(-1.47)	-0.384	(-1.17)	–	–
...	...	...	...	...	...	...
	Linear		Quadratic		Part. Linear	

Table 1.2 Parametric logit coefficients and GPLM coefficients (t–values in parentheses). Bandwidths are 40% of range for GPLM. Credit data, Fahrmeir and Hamerle (1984).

dimensions which gives $h_1 = 1.72$ and $h_2 = 0.55$. A scatterplot of amount versus age is given in the lower panel of Figure 1.7. The good clients ($Y = 1$) are marked by +, the bad clients ($Y = 0$) are marked by o.

We have carried out the analysis for different bandwidths. For all these different bandwidths, the nonparametric estimates $\widehat{m}$ are obviously nonlinear. However, it is difficult to judge whether a nonparametric estimate gives a significant improvement. The high values of $\widehat{m}$ are caused by only a few observations (as can be seen from the scatterplot). For a closer inspection of $\widehat{m}$ Figure 1.8 shows also a contour plot of $\widehat{m}$. In general, it cannot be excluded that the visual difference between the nonparametric and the linear fit may be caused by boundary and bias problems of $\widehat{m}$. Additionally, some of the other covariables have a quite dominant influence on credit worthiness.

Härdle, Mammen and Müller (1996) proposed a procedure for testing GLM versus GPLM. We applied this test using and computed critical values using the bootstrap procedure proposed in Härdle, Mammen and Müller (1996). Table 1.3 shows the observed significance levels for rejection. As before, we report the results for the bandwidths expressed in percent of the ranges in both dimensions. The decision of the test depends obviously on the bandwidth.

We see from Table 1.3 that linearity is clearly rejected for bandwidths up to 40%. The significance levels for rejection increase when we include interaction and quadratic terms. But altogether, we conclude that the correct model should be of more complicated structure than the quadratic.

For higher dimensions in T the possible nonlinearities in (1.22) cannot be graphically displayed and face the above mentioned problems of interpretability. An additive structured partial linear index may be considered. This is considered in Hastie and Tibshirani (1990) on basis of the backfitting algo-

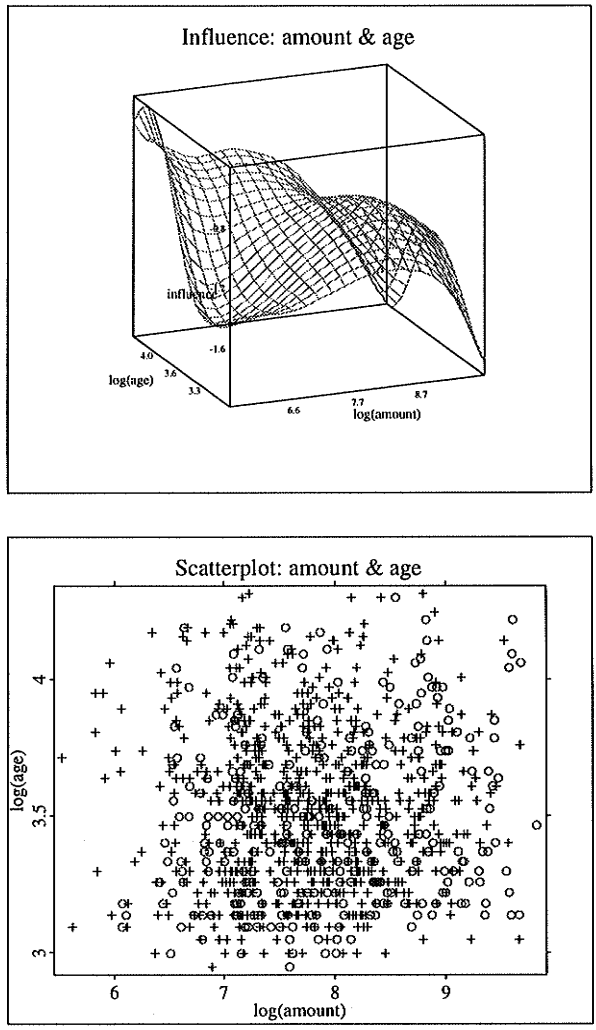


Fig. 1.7 Two-dimensional nonparametric function of amount and age in GPLM (upper panel). Bandwidths are 40% of range. Scatterplot of of amount and age (lower panel). Credit data, Fahrmeir and Hamerle (1984).

rithm. A variant based on the integration method introduced by Linton and Nielsen (1995) is currently under development, see Härdle, Huet, Mammen and Sperlich (1996).

Smoothing and Regression, Approaches, Computation and Application
Chapter 12 M.Schimek (ed), Wiley Publishers,357-391

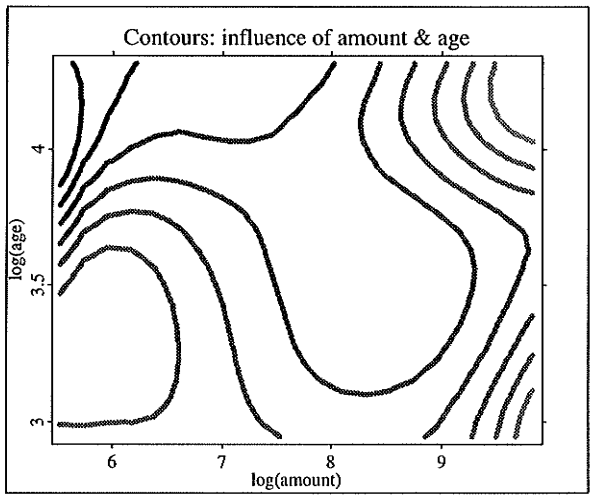


Fig. 1.8 Contours for nonparametric function of amount and age in GPLM. Bandwidths are 40% of range. Credit data, Fahrmeir and Hamerle (1984).

h	20%	30%	40%	50%	60%
linear	<0.01	<0.01	<0.01	0.01	0.29
linear & interaction	<0.01	<0.01	<0.01	0.07	0.40
quadratic	<0.01	<0.01	<0.01	0.35	0.55

Table 1.3 Observed significance levels for test of GLM against GPLM. Bootstrap sample size $n^* = 100$. Credit data, Fahrmeir and Hamerle (1984).

Software

Routines for kernel density and kernel regression estimation are included in virtually any modern software package. Semiparametric procedures are typically add–ons in programming environments which allow a user side integration of kernel estimation procedures. We want to mention XploRe and Splus as examples here.

Smoothing and Regression, Approaches, Computation and Application
Chapter 12 M.Schimek (ed), Wiley Publishers,357-391

Acknowledgments

The authors wish to thank several participants of classes in Non– and Semiparametric Modelling and two referees for numerous typo corrections and valuable proposals for improvement of the paper.

Härdle, W. and Müller, M. (2000) Multivariate and Semiparametric Kernel regression.

References

Carroll, R. J., Fan, J., Gijbels, I. and Wand, M. P. (1997). Generalized partially linear single–index models, *Journal of the American Statistical Association* **92**(438): 477–489.

Craven, P. and Wahba, G. (1979). Smoothing noisy data with spline functions, *Numer. Math.* **31**: 377–403.

Fahrmeir, L. and Hamerle, A. (1984). *Multivariate Statistische Verfahren*, De Gruyter, Berlin.

Fahrmeir, L. and Tutz, G. (1994). *Multivariate Statistical Modelling Based on Generalized Linear Models*, Springer.

Fan, J., Gasser, T., Gijbels, I., Brockmann, M. and Engel, J. (1993). Local polynomial fitting: A standard for nonparametric regression, *Discussion Paper 9315*, Institut de Statistique, Université Catholique, Louvain–La–Neuve.

Fan, J. and Gijbels, I. (1995). *Local Polynomial Modeling and Its Application — Theory and Methodologies*, Chapman and Hall, New York.

Fan, J. and Marron, J. S. (1994). Fast implementations of nonparametric curve estimators, *Journal of Computational and Graphical Statistics* **3**(1): 35–56.

Fan, J. and Müller, M. (1995). Density and regression smoothing, *in* W. Härdle, S. Klinke and B. A. Turlach (eds), *XploRe – an interactive statistical computing environment*, Springer, pp. 77–99.

Härdle, W. (1990). *Applied Nonparametric Regression*, Econometric Society Monographs No. 19, Cambridge University Press.

Härdle, W. (1991). *Smoothing Techniques, With Implementations in S*, Springer, New York.

Härdle, W., Hall, P. and Marron, J. S. (1988). How far are automatically chosen regression estimators from their optimum?, *Journal of the American Statistical Association* **83**: 86–97.

Härdle, W., Hall, P. and Marron, J. S. (1992). Regression smoothing estimators that are not far from their optimum, *Journal of the American Statistical Association* **87**: 227–233.

Härdle, W., Huet, S., Mammen, E. and Sperlich, S. (1996). Semiparametric additive indices for binary response, *Technical report*, Sonderforschungsbereich 373, Humboldt-Universität zu Berlin.

Härdle, W., Mammen, E. and Müller, M. (1996). Testing parametric versus semiparametric modelling in generalized linear models, *SFB 373 Discussion Paper 28*, Sonderforschungsbereich 373, Humboldt-Universität zu Berlin.

Härdle, W. and Scott, D. (1992). Smoothing in by weighted averaging using rounded points, *Computational Statistics* **7**: 97–128.

Härdle, W. and Stoker, T. (1989). Investigating smooth multiple regression by the method of average derivatives, *Journal of the American Statistical Association* **84**: 986–995.

Hastie, T. J. and Tibshirani, R. J. (1990). *Generalized Additive Models*, Vol. 43 of *Monographs on Statistics and Applied Probability*, Chapman and Hall, London.

Horowitz, J. L. (1993). Semiparametric and nonparametric estimation of quantal response models, *in* G. S. Madala, C. R. Rao and H. D. Vinod (eds), *Handbook of Statistics*, Elsevier Science Publishers, pp. 45–72.

Horowitz, J. L. and Härdle, W. (1996). Direct semiparametric estimation of single index models with discrete covariates, *Journal of the American Statistical Association* . to appear.

Hunsberger, S. (1994). Semiparametric regression in likelihood–based models, *Journal of the American Statistical Association* **89**: 1354–1365.

Ichimura, H. (1993). Semiparametric least squares (SLS) and weighted SLS estimation of single–index models, *Journal of Econometrics* **58**: 71–120.

Klein, R. and Spady, R. (1993). An efficient semiparametric estimator for binary response models, *Econometrica* **61**: 387–421.

Korostelev, A. and Müller, M. (1995). Single index models with mixed discrete-continuous explanatory variables, *Discussion Paper 26*, Sonderforschungsbereich 373, Humboldt-Universität zu Berlin.

Linton, O. and Nielsen, J. P. (1995). A kernel method of estimating structured nonparametric regression based on marginal integration, *Biometrika* **82**: 93–100.

Maddala, G. S. (1983). *Limited-dependent and qualitative variables in econometrics*, Econometric Society Monographs No. 4, Cambridge University Press.

Marron, J. S. and Nolan, D. (1988). Canonical kernels for density estimation, *Statistics & Probability Letters* **7**(3): 195–199.

McCullagh, P. and Nelder, J. A. (1989). *Generalized Linear Models*, Vol. 37 of *Monographs on Statistics and Applied Probability*, 2 edn, Chapman and Hall, London.

Müller, H.-G. (ed.) (1988). *Nonparametric Regression Analysis of Longitudinal Data*, Springer, Berlin.

Müller, M. (1997). Computer–assisted generalized partial linear models, in *Computing Science and Statistics*, 29th Symposium of the Interface – Proceedings, Houston, Texas.

Newey, W. and Stoker, T. (1993). Efficiency of weighted average derivative estimators and index models, *Econometrica* **5**: 1199–1223.

Powell, J. L., Stock, J. H. and Stoker, T. M. (1989). Semiparametric estimation of index coefficients, *Econometrica* **57**(6): 1403–1430.

Ruppert, D. and Wand, M. P. (1994). Multivariate locally weighted least squares regression, *Annals of Statistics* **22**(3): 1346–1370.

Sain, S. R., Baggerly, K. A. and Scott, D. W. (1994). Cross–validation of multivariate densities, *Journal of the American Statistical Association* **89**(427): 807–817.

Scott, D. W. (1992). *Multivariate Density Estimation: Theory, Practice, and Visualization*, John Wiley & Sons, New York, Chichester.

Scott, D. and Wand, M. (1991). Feasibility of multivariate density estimates, *Biometrika* **78**: 197–205.

Severini, T. A. and Staniswalis, J. G. (1994). Quasi-likelihood estimation in semiparametric models, *Journal of the American Statistical Association* **89**: 501–511.

Severini, T. A. and Wong, W. H. (1992). Generalized profile likelihood and conditionally parametric models, *Annals of Statistics* **20**: 1768–1802.

Silverman, B. W. (1986). *Density Estimation for Statistics and Data Analysis*, Vol. 26 of *Monographs on Statistics and Applied Probability*, Chapman and Hall, London.

Splus, see http://www.mathsoft.com/Splus.

Staniswalis, J. G. and Messer, K. (1996). Addendum to "Kernel estimators for multiple regression", *Journal of Nonparametric Statistics* **7**: 67–68.

Staniswalis, J. G., Messer, K. and Finston, D. R. (1994). Kernel estimators for multiple regression, *Journal of Nonparametric Statistics* **3**: 103–121.

Staniswalis, J. G. and Thael, P. F. (1997). An explanation of generalized profile likelihood, in *Computing Science and Statistics*, 29th Symposium of the Interface – Proceedings, Houston, Texas.

Stoker, T. M. (1991). Equivalence of direct, indirect and slope estimators of average derivatives, *in* W. A. Barnett, J. Powell and G. Tauchen (eds), *Nonparametric and Semiparametric Methods in Econometrics and Statistics*, Proceedings of the fifth international symposium in Economic Theory and Econometrics, Cambridge University Press.

Wand, M. P. (1994). Fast computation of multivariate kernel estimators, *Journal of Computational and Graphical Statistics* **3**(4): 433–445.

Wand, M. P. and Jones, M. C. (1994). Multivariate plug–in bandwidth selection, *Computational Statistics* **9**: 97–911.

Wand, M. P. and Jones, M. C. (1995). *Kernel Smoothing*, Vol. 60 of *Monographs on Statistics and Applied Probability*, Chapman and Hall, London.

Weisberg, S. and Welsh, A. H. (1994). Adapting for the missing link, *Annals of Statistics* **22**: 1674–1700.

XploRe, see http://www.xplore-stat.de.

Statistics 147, 119-130, Springer Verlag, Heidelberg

Backtesting Beyond VaR

W. Härdle and G. Stahl
Humboldt-Universität zu Berlin and
Bundesaufsichtsamt für das Kreditwesen, Berlin

Abstract

VaR models are related to statistical forecast systems. Within that framework different forecast tasks including Value-at-Risk and shortfall are discussed and motivated. A backtesting method based on the shortfall is developed and applied to VaR forecasts of a real portfolio. The analysis shows that backtesting based on shortfall is very sensitive with respect to the underlying assumptions.

1 Forecast tasks and VaR Models

With the implementation of Value-at-Risk (VaR) models a new chapter of risk management was opened. Their ultimate goal is to quantify the uncertainty about the amount that may be lost or gained on a portfolio over a given period of time. Most generally, the uncertainty is expressed by a forecast distribution P_{t+1} for period $t+1$ associated with the random variable L_{t+1}, denoting the portfolio's profits and losses (P&L).
In practice, the prediction P_{t+1} is conditioned on an information set at time t and, typically calculated through a plug-in approach, see Dawid (1984). In this case, P_{t+1} is output of a statistical forecast system, here the VaR model, consisting of a (parametric) family of distributions, denoted by $\mathcal{P} = \{P_\theta \mid \theta \in \Theta\}$ together with a prediction rule. Assumed that P_{t+1} belongs to this parametrized family $\mathcal{P}$ the estimates $\hat{\theta}_t$ are calculated by the prediction rule on the basis of a forward rolling data history $\mathcal{H}_t$ of fixed length n (typically $n = 250$ trading days) for all t, i.e.

$$P_{t+1}(\cdot) = P_{\hat{\theta}_t}(\cdot \mid \mathcal{H}_t).$$

Härdle, W. and Stahl, G. (2000) Backtesting beyond VaR.

Statistics 147, 119-130, Springer Verlag, Heidelberg

One example for $\mathcal{P}$ also pursued in this paper is the RiskMetrics (1996) delta normal framework, i.e., the portfolios considered are assumed to consist of linear (or linearised) instruments and the common distribution of the underlyings' returns $Y \in I\!R^d$, i.e., the log price changes $Y_{t+1} = logX_{t+1} - logX_t$, is a (conditional) multinormal distribution, $N_d(0, \Sigma_t)$, where Σ_t (resp. and σ_t^2) denotes a conditional variance, i.e., $\mathcal{H}_t$ measurable function.
Consider for simplicity a position of λ_t shares in a single asset (i.e., d = 1) whose market value is x_t. The conditional distribution of L_{t+1} for this position with exposure $w_t = \lambda_t\, x_t$ is (approximately)

$$\begin{aligned} \mathcal{L}(L_{t+1} \mid \mathcal{H}_t) &= \mathcal{L}(\lambda_t(X_{t+1} - x_t) \mid \mathcal{H}_t) = \mathcal{L}\Big(w_t \frac{X_{t+1} - x_t}{x_t} \mid \mathcal{H}_t\Big) \\ &\approx \mathcal{L}(w_t Y_{t+1} \mid \mathcal{H}_t) = N(0, w_t^2\, \sigma_t^2), \end{aligned}$$

where the approximation refers to

$$lnX_{t+1} - lnx_t = \frac{X_{t+1} - x_t}{x_t} + o(X_{t+1} - x_t).$$

The generalization to a portfolio of (linear) assets is straightforward. Let w_t denote a $d-$dimensional exposure vector, i.e., $w_t = (\lambda_t^1 x_t^1, \cdots, \lambda_t^d x_t^d)$. Hence, the distribution of the random variable $w_t^T Y_{t+1}$ belongs to the family

$$\mathcal{P}_{t+1} = \{N(0, \sigma_t^2) : \sigma_t^2 \in [0, \infty)\}, \tag{1}$$

where $\sigma_t^2 = w_t^T \Sigma_t w_t$.
The aim of the VaR analysis is to estimate $\theta = \sigma_t$ and thereby to establish a prediction rule. For L_{t+1} we adopt therefore the following framework:

$$\begin{aligned} L_{t+1} &= \sigma_t\, Z_{t+1} && (2) \\ Z_{t+1} &\overset{iid}{\sim} N(0,1) && (3) \\ \sigma_t^2 &= w_t^T \Sigma_t w_t. && (4) \end{aligned}$$

For a given $(n \times d)$ data matrix $\mathcal{X}_t = \{y_i\}_{i=t-n+1,\cdots,t.}$ of realisations of the underlying vector of returns with dimension d, two estimators for Σ_t will be considered. The first is a naive estimator, based on a rectangular moving average (RMA)

$$\hat{\Sigma}_t = \frac{1}{n} \mathcal{X}_t^T \mathcal{X}_t. \tag{5}$$

This definition of $\hat{\Sigma}_t$ makes sense since the expectation of Y_t is assumed zero. The second, also recommended by Taylor (1986) to forecast volatility, is

built by an exponential weighting scheme (EMA) applied to the data matrix $\tilde{\mathcal{X}}_t = \{ diag(\lambda^d, \lambda^{d-1}, \cdots, \lambda, 1)^{1/2} y_i \}_{i=t-n+1,\cdots,t}$:

$$\hat{\Sigma}_t = (1-\lambda)\tilde{\mathcal{X}}_t^T \tilde{\mathcal{X}}_t \tag{6}$$

These estimates are plugged-into (4) and (2), yielding two prediction rules for

$$P_{t+1} \in \mathcal{P} = \{N(0, \sigma_t^2) \mid \sigma_t^2 \in [0, \infty)\}.$$

By their very nature VaR models contribute to several aspects of risk management. Hence, a series of parameters of interest - all derived from P_{t+1} - arise in natural ways. The particular choice is motivated by specific forecast tasks, e.g., driven by external (e.g., regulatory issues) or internal requirements or needs (e.g., VaR-limits, optimisation issues).
A very important part of risk management is the implementation of a systematic process for limiting risk. In the light of that task, it is at hand that forecast intervals defined by the $\widehat{VaR}_t$,

$$\widehat{VaR}_t = F_{t+1}^{-1}(\alpha) := inf\{x \mid F_{t+1}(x) \geq \alpha\},$$

where F_{t+1} denotes the cdf of P_{t+1}, are substantial.
If the main focus is to evaluate the forecast quality of the prediction rule associated to a VaR model, transformations of F_t should be considered, see Dawid (1984), Sellier-Moiseiwitsch (1993) and Crnkovic and Drachman (1996). For a given sequence of prediction-realisation pairs (P_t, l_t) - where l_t denotes a realisation of L_t - the prediction rules works fine if the sample $u = \{u_t\}_{t=1}^k = \{F_t(l_t)\}_{t=1}^k$ looks like an iid random sample from $U[0,1]$. A satisfactory forecast quality is often interpreted as an adequate VaR model. The focus of this paper is to consider the expected shortfall of L_{t+1}, as the parameter of interest and to derive backtesting methods related to this parameter - this will be done in the next section. The expected shortfall - also called tail VaR - is defined by

$$\begin{aligned} E(L_{t+1} \mid L_{t+1} > VaR_t) &= E(L_{t+1} \mid L_{t+1} > z_\alpha \, \sigma_t) && (7) \\ &= \hat{\sigma}_t \, E(L_{t+1}/\sigma_t \mid L_{t+1}/\sigma_t > z_\alpha) && (8) \end{aligned}$$

where z_α is a α-quantile of a standard normal distribution. The motivation to consider this parameter is threefold. Firstly, McAllister and Mingo (1996) worked out the advantage of (7) compared to VaR if these parameters are plugged-into the denominator of a risk performance measures, e.g. a Sharpe-ratio or a RAROC (risk-adjusted return - that constitutes the numerator - on capital) numbers which are used to benchmark divisional performance, see Matten (1996) and CorporateMetrics (1999), - the economic motivation.

Statistics 147, 119-130, Springer Verlag, Heidelberg

Secondly, Artzner et al. (1997) pointed out that (7) defines a coherent risk measure, a conceptual consideration. Thirdly, Leadbetter (1995) emphasized in the context of environmental regulation the need for incorporating the height of exceedances violating regulatory thresholds and critized those methods solely based on counts, neglecting the heights - statistical arguments. The paper is organised as follows. In the next section we present our approach on backtesting using the expected shortfall risk. In section 3 we apply this methodology to real data and visualise the difference betweeen RMA and EMA based VaRs. Section 4 presents the conclusions of this work.

2 Backtesting based on the expected shortfall

As pointed out by Baille and Bollerslev (1992), the accuracy of predictive distributions is critically dependent upon the knowledge of the correct (conditional) distribution of the innovations Z_t in (2). For given past returns $\mathcal{H}_t = \{y_t, y_{t-1}, \cdots, y_{t-n}\}$, σ_t in (4) can be estimated either by (5) or (6) and then $\mathcal{L}(L_{t+1} \mid \mathcal{H}_t) = N(0, \hat{\sigma}_t)$. Hence,

$$\mathcal{L}(L_{t+1}/\hat{\sigma}_t \mid \mathcal{H}_t) = N(0,1).$$

This motivates to standardize the observations l_t by the predicted STD

$$\frac{l_{t+1}}{\sigma_t}$$

and to interprete thes as realisations of (2). see also RiskMetrics (1996):

$$Z_{t+1} = \frac{L_{t+1}}{\sigma_t} \sim N(0,1) \tag{9}$$

For a fixed u we get for Z_{t+1} in (2)

$$\vartheta = E(Z_{t+1} \mid Z_{t+1} > u) = \frac{\varphi(u)}{1-\Phi(u)} \tag{10}$$

$$\varsigma^2 = Var(Z_{t+1} \mid Z_{t+1} > u) = 1 + u \cdot \vartheta - \vartheta^2 \tag{11}$$

where φ, Φ denotes the density, resp. the cdf of a standard normal distributed random variable.
For a given series of standardized prediction-realisation pairs $(F_{t+1}(\cdot/\hat{\sigma}_t), l_{t+1}/\hat{\sigma}_t)$

$$(E(L_{t+1} \mid L_{t+1} > \widehat{VaR}_t), l_{t+1})_{t=0}^{n}$$

and a fixed u, ϑ is estimated by

$$\hat{\vartheta} = \frac{\sum_{t=0}^{n} z_{t+1}\, I(z_{t+1} > u)}{\sum_{t=0}^{n} I(z_{t+1} > u)} \tag{12}$$

where z_{t+1} denotes the realisations of the variable (2). Inference about the statistical significance of $\hat{\vartheta} - \vartheta$ will be based on the following asymptotic relationship:

$$\sqrt{N(u)}\,\left(\frac{\hat{\vartheta} - \vartheta}{\hat{\varsigma}}\right) \xrightarrow{\mathcal{L}} N(0,1) \tag{13}$$

where $N(u)$ is the (random) number of exceedances over u and $\hat{\vartheta}$ is plugged-into (11) yielding an estimate $\hat{\varsigma}$ for ς. The convergence in (13) follows from an appropriate version of the CLT for a random number of summands in conjunction with Slutzky's Lemma, see Leadbetter (1995) for details. Under sufficient conditions and properly specified null hypothesis it is straight forward to prove the complete consistency and an asymptotic $\alpha-$level for a test based on (13), see Witting and Müller-Funk (1995), pp. 236.
Though these asymptotic results are straight forward they should be applied with care. Firstly, because the truncated variables involved have a shape close to an exponential distribution, hence, $\hat{\vartheta}$ will be also skewed for moderate sample sizes, implying that the convergence in (13) will be rather slow. Secondly, in the light of the skewness, outliers might occur. In such a case, they will have a strong impact on an inference based on (13) because the means in the nominator and in the denominator as well are not robust. The circumstance that the truncated variables' shape is close to an exponential distribution motivates classical tests for an exponential distribution as an alternative to (13).

3 Backtesting in Action

The Data The prediction-realisation (P_t, l_t) pairs to be analysed are stemming from a real bond portfolio of a German bank that was hold fixed over the two years 94 and 95,i.e., $w_t \equiv w$. For that particular (quasi) linear portfolio the assumptions met by (2) - (4) are reasonable and common practice in the line of RiskMetrics.
The VaR forecasts are based on a history $\mathcal{H}_t$ of 250 trading days and were calculated by two prediction rules for a 99%-level of significance. The first rule applies a RMA, the second is based an EMA with decay factor $\lambda = 0.94$ as proposed by RiskMetrics to calculate an estimate of $\hat{\Sigma}_t$ different from (5). Remembering the bond crisis in 1994, it is of particular interest to see how

these different forecast rules perform under that kind of stress. Their comparison will also highlight those difficulties to be faced with the expected shortfall if it would be applied e.g. in a RAROC framework.

Exploratory Statistics The following analysis is based on two distinctivce features in order to judge the difference of the quality of prediction rules by elementary exploratory means: calibration and resolution, see Murphy and Winkler (1987), Dawid (1984) and Sellier-Moiseiwitsch (1993). The exploratory tools are timeplots of prediction- realisation pairs (Fig. 1) and indicator variables (Fig. 4) for the exceedances to analyse the resolution and Q-Q-plots of the variable

$$\frac{L_{t+1}}{\widehat{VaR}_t} = \frac{L_{t+1}}{2.33\hat{\sigma}_t} \tag{14}$$

to analyse the calibration (Fig 2, 3). A further motivation to consider variable (14) instead of (2) is that their realisations greater than one are just the exceedances of the VaR forcasts. Of course these realisations are of particular interest. If the predictions are perfect, the Q-Q-plot is a straight line and the range of the Y-coordinate of the observations should be containded in the interval $[-1, 1]$. Hence, the Q-Q-plot for (14) visualises not only the calibration but also the height of exceedances. A comparison of Figure 2 with Figure 3 shows clearly that EMA predictions are better calibrated than RMA ones. The second feature, resolution, refers to the *iid* assumption, see Murphy and Winkler (1987). Clusters in the timeplots of exceedances, Figure 4,

$$(t, I(l_{t+1} > \widehat{VaR}_t)_{t=1}^{260}$$

indicate a serial correlation of exceedances. Again EMA outperforms RMA. From Figure 1, we conclude that in 94 (95) 9 (4) exceedances were recorded for the EMA and 13 (3) for the RMA. Evidently, the window-length of 250 days causes an underestimation of risk for RMA if the market moves from a tranquile regime to a volatile one, and overestimates vice versa. On the other hand the exponential weighting scheme adapts changes of that kind much quickier.

Statistics 147, 119-130, Springer Verlag, Heidelberg

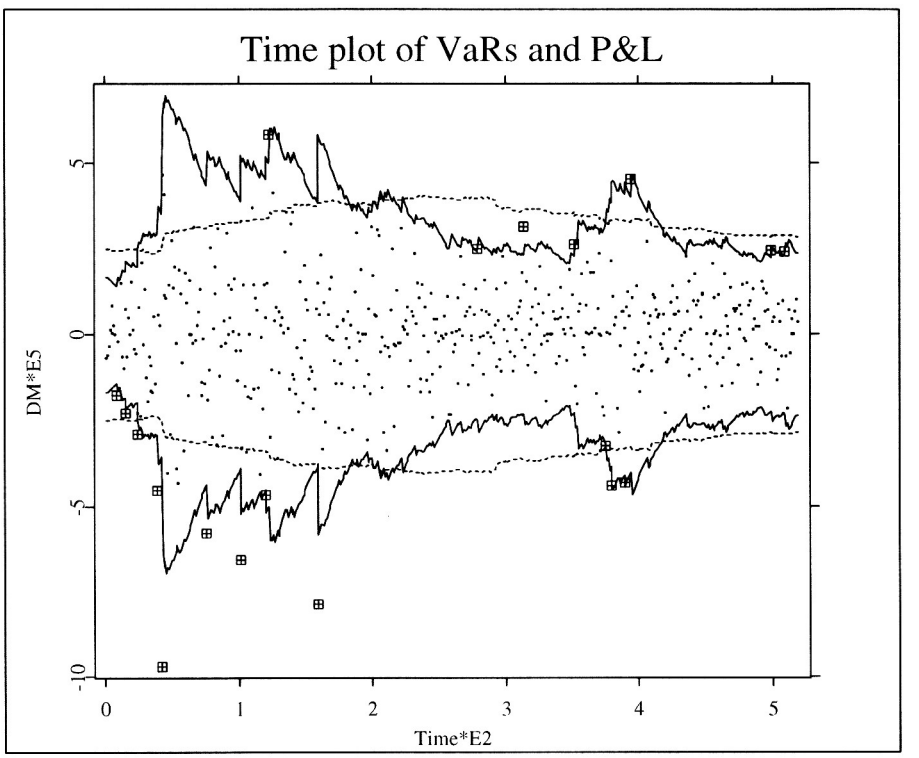


Figure 1: The dots show the observed change of the portfolio values, l_t. The dashed lines show the predicted VaRs based on RMA (99% and 1%). The solid lines show the same for EMA.

Statistics 147, 119-130, Springer Verlag, Heidelberg

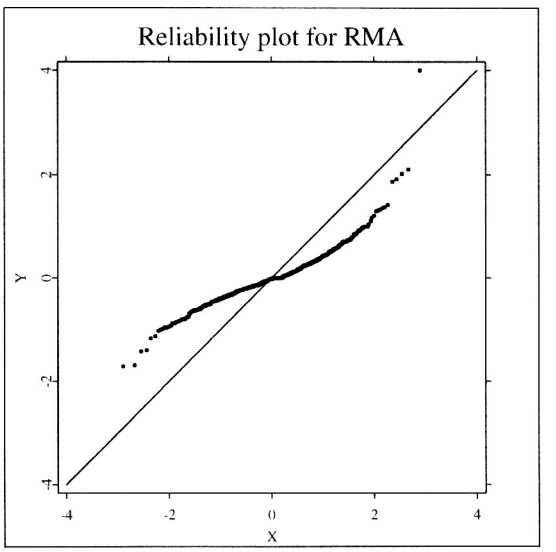


Figure 2: Q-Q plot of $l_{t+1}/\widehat{VaR}_t$ for RMA in 94.

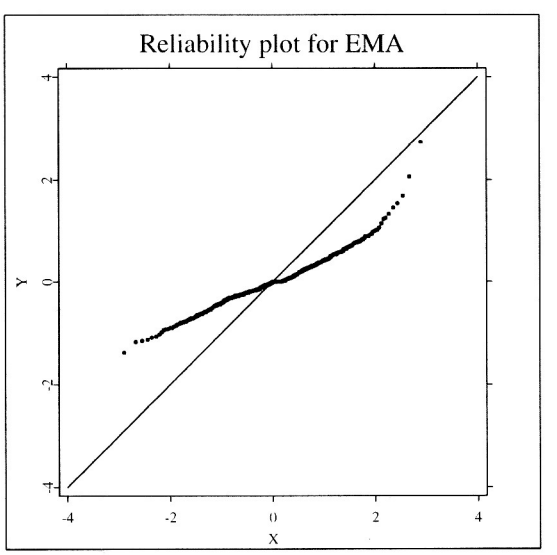


Figure 3: Q-Q plot of $l_{t+1}/\widehat{VaR}_t$ for EMA in 94.

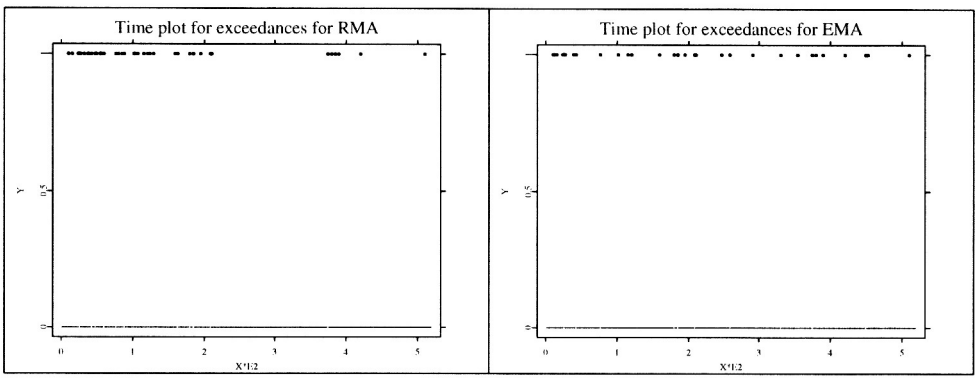


Figure 4: Timeplots of the exceedances over VaR of 80% level for RMA (left) and EMA. The better resolution of EMA is evident.

The poor forecast performance, especially for the upper tail is evident. The asymmetry and outliers are caused by the market trend. For a particular day the VaR forecast is exceeded by almost 400 %. If the model (2) - (4) would be correct, the variable (14) has a STD of 0.41. The STD calculated from the data is 0.62. Hence, in terms of volatility the RMA underestimates risk on the average of about 50%.

The plot for shows the same characteristics as that in Figure 2 but the EMA yields a better calibration. The STD from the data yields 0.5. Hence, an underestimation on the average of 25%. This indicates clearly that EMA gives a better calibration then RMA. Q-Q-plots for 95 are omitted. The two models give similar results, though even in that case the EMA is slightly better.

Inference The exploratory analysis has shown notable differences between the acurracy of RMA and EMA for the year 94. In this paragraph their statistical significance will be investigated. The inference will be based on the observations

$$\frac{l_{t+1}}{\hat{\sigma}_t}$$

and the underlying model (2) - (4). The threshold u is set to the 80%-quantile of L_{t+1}/σ_t yielding $\vartheta = 1.4$, by (10). Now, based on (13) an asymptotic significance test for the hypothesis

$$H_0 \ : \vartheta \overset{(<)}{=} 1.4 \tag{15}$$

will be used. This setting - especially (2) - seems reasonable for RMA and the given sample of size $n = 250$.
As mentioned by Skouras and Dawid (1996) plug-in forecasting systems have the disadvantage that the uncertainty of the estimator for σ_t is not incorporated in the predictive distribution P_{t+1}. This applies especially to Z_{t+1} if the EMA is used. In that case a $t(n)$-distribution is indicated. A reasonable choice - motivated by generalized degrees of freedom - is

$$Z_{t+1} = \frac{L_{t+1}}{\hat{\sigma}_t} \sim t(20). \tag{16}$$

Though the particular thresholds $u_N = 0.854$ - for the normal distribution - and $u_t = 0.86$ - for the $t(20)$ distribution differ only slightly (0.5 %), the associated means ϑ change about 5 % and the STD ς even about 18%. Parallel to (15) the hypothesis

$$H_0 \ : \vartheta \overset{(<)}{=} 1.47 \tag{17}$$

will be tested.
Tables 1 to 4 summarise the empirical results.

Method	$\vartheta = 1.4$	$\varsigma = 0.46$	$\frac{\sqrt{N(u)}(\hat{\vartheta}-\vartheta)}{\hat{\varsigma}}$	significance	nobs
EMA	$\hat{\vartheta} = 1.72$	$\hat{\varsigma} = 1.01$	2.44	0.75%	61
RMA	$\hat{\vartheta} = 1.94$	$\hat{\varsigma} = 1.3$	3.42	0.03%	68

Table 1: $H_0 \ : \vartheta \overset{(<)}{=} 1.4$

Method	$\vartheta = 1.47$	$\varsigma = 0.546$	$\frac{\sqrt{N(u)}(\hat{\vartheta}-\vartheta)}{\hat{\varsigma}}$	significance	nobs
EMA	$\hat{\vartheta} = 1.72$	$\hat{\varsigma} = 1.01$	2.01	2.3%	61
RMA	$\hat{\vartheta} = 1.94$	$\hat{\varsigma} = 1.3$	3.04	0.14%	68

Table 2: $H_0 \ : \vartheta \overset{(<)}{=} 1.47$

Firstly from tables 1 and 2, the observed exceedances over threshold u indicate again that the EMA is superior than the RMA. For a sample of 260 prediction-realisation pairs 52 exceedances are to be expected (STD 6.45). For the EMA 61 (61 - 52 $\approx$ 1.5 STD) exceedances were observed and 68 (68 - 52 $\approx$ 2.5 STD) for the RMA.
A comparison of table 1 with 2 shows that random errors strongly influence

the significance of the test. Recalling the impressive outliers in the Q-Q-plots it is worthwile to exclude these from the data and re-run the test. The results are given in tables 3 and 4. Again, a serious change in the level

Method	$\vartheta = 1.4$	$\varsigma = 0.46$	$\frac{\sqrt{N(u)}(\hat{\vartheta}-\vartheta)}{\hat{\varsigma}}$	significance	nobs
EMA	$\hat{\vartheta} = 1.645$	$\hat{\varsigma} = 0.82$	2.31	1%	60
RMA	$\hat{\vartheta} = 1.83$	$\hat{\varsigma} = 0.93$	3.78	0.00%	67

Table 3: $H_0 : \vartheta \overset{(<)}{=} 1.4$ - largest outlier excluded

Method	$\vartheta = 1.47$	$\varsigma = 0.546$	$\frac{\sqrt{N(u)}(\hat{\vartheta}-\vartheta)}{\hat{\varsigma}}$	significance	nobs
EMA	$\hat{\vartheta} = 1.645$	$\hat{\varsigma} = 0.82$	1.65	5%	60
RMA	$\hat{\vartheta} = 1.83$	$\hat{\varsigma} = 0.93$	3.1	0.15%	67

Table 4: $H_0 : \vartheta \overset{(<)}{=} 1.47$ - largest outlier excluded

of significance for the RMA is observed indicating the non robustness of the test. These results show furthermore that inference about the tails of a distribution is subtle. In addition the *iid* assumption - cluster of exceedances - might also be violated. One possible source for that is the overlap of the $\mathcal{H}_t$. Hence, the estimates may correlate. Techniques like moving blocks and resampling methods see Diebold and Mariano (1995) and Carlstein (1993) are good remedies.

To overcome the problems related to the slow convergence of (13) an exponential distribution may be fitted to the data and then, again a classical test will be applied. The following table reports the significance levels based on a one-sided Kolmogoroff-Smirnov test. Again, the results emphasize the

Method	$\sigma = 0.46$	$\sigma = 0.546$
EMA	0.25%	10% (14%)
RMA	$< 0.1\%$	$< 0.1\%$

Table 5: Kolmogoroff-Smirnov Test

impact of random errors. The number in brackets refers to that case, where the largest outlier is deleted.

Statistics 147, 119-130, Springer Verlag, Heidelberg

4 Conclusions

VaR models were introduced as specific statistical forecast systems. The backtesting procedure was formulated in terms of measuring forecast quality. The empirical results highlight the better calibration and resolution of VaR forecasts based on (exponentially weights) EMA compared to (uniformly weights) RMA. However, more interesting is the impressive difference in amount (50%). A surprising result is the strong dependence of inferences based on expected shortfall from the underlying distribution. Hence, if expected shortfall will be used in practice in order to calculate performance measures like RAROC the inferences resp. the estimates should be robustified.

Acknowledgements: The authors would like to express their warmest thanks to Zdeněk Hlávka for his help by providing the graphics in XploRe. They also wish to thank for the support by the Sonderforschungsbereich 373. Last but not least the second author disclaims that the views expressed herein should not be construed as being endorsed by the Bundesaufsichtsamt.

References

Artzner, P., Dealban, F., Eber, F.-J. & Heath, D. (1997)Thinking Coherently, *RISK MAGAZINE.*

Baille, R. T. & T. Bollerslev (1992) Prediction in Dynamic Models with Time-Dependent Conditional Variances. *Econometrica*, **50**: 91–114.

Carlstein, A. (1993) Resampling Techniques For Stationary Time Series: Some Recent Developments. *In New Directions in Time Series*, Ed. Brillinger, D. et al. Springer, New York. 75–82

Crnkovic, C. & J. Drachman (1996)A Universal Tool to Discriminate Among Risk Measurement Techniques, *RISK MAGAZINE.*

Dawid, A. P. (1984) The prequential approach. *J. R. Statist. Soc.*, A, **147**: 278–292.

Diebold, F. X. & R. S. Mariano (1995)Comparing Predictive Accuray. *Journal of Business and Statistics*, **13**: 253–263.

Härdle, W. & Klinke, S. & Müller, M. (1999) XloRe Learning Guide. www.xplore-stat.de

Leadbetter, M. R. (1995) On high level exceedance modeling and tail inference. *Journal of Planning and Inference*, **45**: 247–260.

Matten, C. (1996) *Managing Bank Capital.* John Wiley & Sons: Chicheseter.

McAllister, P. H. & J.J. Mingo (1996) Bank Capital requirements for securitzed loan portfolios. *Journal of Banking and Finance*, **20**: 1381–1405.

Murphy, A. H. & R. L. Winkler (1987) A General Framework for Forecast Verification. *Monthly Weather Review*, **115**: 1330–1338.

RiskMetrics (1996) Technical Dokument, 4th Ed.

CorporateMetrics (1999) Technical Dokument, 1st. Ed.

Sellier-Moiseiwitsch, F. (1993) Sequential Probability Forecasts and the Probability Integral Transform. *Int. Stat. Rev.*, **61**: 395–408.

Skouras, K. and A. P. Dawid (1996) On efficient Probability Forecasting Systems. *Research Report* **No. 159**, Dep. of Statistical Science, University College London.

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

Witting H. and U. Müller-Funk (1995) *Mathematische Statistik II.* Teubner, Stuttgart.

XploRe Application Guide, 397-458, Springer Verlag, Heidelberg

Flexible Time Series Analysis

Wolfgang Härdle and Rolf Tschernig

In this chapter we present nonparametric methods and available quantlets for **nonlinear modelling of univariate time series**. A general nonlinear time series model for an univariate stochastic process $\{Y_t\}_{t=1}^T$ is given by the **heteroskedastic nonlinear autoregressive (NAR) process**

$$Y_t = f(Y_{t-i_1}, Y_{t-i_2}, \ldots, Y_{t-i_m}) + \sigma(Y_{t-i_1}, Y_{t-i_2}, \ldots, Y_{t-i_m})\xi_t, \qquad (1)$$

where $\{\xi_t\}$ denotes an i.i.d. noise with zero mean and unit variance and $f(\cdot)$ and $\sigma(\cdot)$ denote the conditional mean function and conditional standard deviation with lags $i_1, \ldots, i_m$, respectively. In practice, the conditional functions $f(\cdot)$ and $\sigma(\cdot)$ as well as the number of lags m and the lags itself $i_1, \ldots, i_m$ are unknown and have to be estimated.

In Section 1 we discuss nonparametric estimators for the conditional mean function of nonlinear autoregressive processes of order one. While this case has been most intensively studied in theory, in practice models with several lags are often more appropriate. Section 2 covers the estimation of the latter, including the selection of appropriate lags. For all models we discuss methods of bandwidth selection which aim at an optimal trade-off between variance and bias of the presented estimators.

Both sections contain practical examples. The corresponding quantlets for fitting nonlinear autoregressive processes of order one are contained in the quantlib `smoother`. A number of quantlets for fitting higher order models are found in the third party quantlib `tp/cafpe/cafpe`.

Although obvious we would like to mention that in the following we only discuss methods for which quantlets are available. For an overview of alternative methods and models we would like to refer the reader to the surveys of Tjøstheim (1994) or Härdle, Lütkepohl, and Chen (1997).

Financial support was received by the Deutsche Forschungsgemeinschaft, SFB 373 ("Quantifikation und Simulation Ökonomischer Prozesse"), Humboldt Universität zu Berlin.

Härdle, W. and Tschernig, R. (2000) Flexible Time Series Analysis.

1 Nonlinear Autoregressive Models of Order One

1.1 Estimation of the Conditional Mean

`mh = regxest(x{, h, K, v})`
computes the univariate conditional mean function using the Nadaraya-Watson estimator

`mh = regest(x{, h, K, v})`
computes the univariate conditional mean function using the Nadaraya-Watson estimator and WARPing

`mh = lpregxest(x{, h, p, v})`
computes the univariate conditional mean function using local polynomial estimation

`mh = lpregest(x{, h, p, K, d})`
computes the univariate conditional mean function using local polynomial estimation and WARPing

Let us turn to estimating the conditional mean function $f(\cdot)$ of a nonlinear autoregressive processes of order one (NAR(1) process)

$$Y_t = f(Y_{t-1}) + \sigma(Y_{t-1})\xi_t \tag{2}$$

using nonparametric techniques. The basic idea is to estimate a Taylor approximation of order p of the unknown function $f(\cdot)$ around a given point y. The simplest Taylor approximation is obtained if its order p is chosen to be zero. One then approximates the unknown function by a constant. Of course, this approximation may turn out to be very bad if one includes observations Y_{t-1} that are distant to y since this might introduce a large approximation bias. One therefore weights those observations less in the estimation. Using the least squares principle, the estimated function value $\widehat{f}(y,h)$ is provided by the estimated constant $\widehat{c}_0$ of a local constant estimate around y

$$\widehat{c}_0 = \arg\min_{\{c_0\}} \sum_{t=2}^{T} \{Y_t - c_0\}^2 K_h(Y_{t-1} - y), \tag{3}$$

where K denotes the weighting function, which is commonly called a kernel function, and $K_h(Y_{t-1} - y) = h^{-1}K\{(Y_{t-1} - y)/h\}$. A number of kernel functions are used in practice, e.g. the Gaussian density function or the quartic kernel $K(u) = 15/16(1-u^2)^2$ on the range $[-1, 1]$ and $K(u) = 0$ elsewhere. $\widehat{f}(y, h) = \widehat{c}_0$ is known as the Nadaraya-Watson or local constant function estimator and can be written as

$$\widehat{f}(y, h) = \frac{\sum_{t=2}^{T} K_h(Y_{t-1} - y)Y_t}{\sum_{t=2}^{T} K_h(Y_{t-1} - y)}. \tag{4}$$

The parameter h is called bandwidth parameter and controls the weighting of the lagged variables Y_{t-1} with respect to their distance to y. While choosing h too small and therefore including only few observations in the estimation procedure leads to a too large estimation variance, taking h too large implies a too large approximation bias. Methods for bandwidth selection are presented in Subsection 1.2.

Before one applies Nadaraya-Watson estimation one should be aware of the conditions that the underlying data generating mechanism has to fulfil such that the estimator has nice asymptotic properties: most importantly, the function $f(\cdot)$ has to be continuous, the stochastic process has to be stationary and the dependence among the observations must decline fast enough if the distance among the observations increases. For measuring dependence in nonlinear time series one commonly uses various mixing concepts. For example, a sequence is said to be α-mixing (strong mixing) (Robinson 1983) if

$$\sup_{A \in \mathcal{F}_1^n, B \in \mathcal{F}_{n+k}^{\infty}} |P(A \cap B) - P(A)P(B)| \leq \alpha_k,$$

where $\alpha_k \to 0$ and $\mathcal{F}_i^j$ is the σ-field generated by $X_i, \ldots, X_j$. An alternative and stronger condition is given by the β-mixing condition (absolute regularity)

$$E \sup \{|P(B|A) - P(B)|\} \leq \beta(k)$$

for any $A \in \mathcal{F}_1^n$ and $B \in \mathcal{F}_{n+k}^{\infty}$. An even stronger condition is the ϕ-mixing (uniformly mixing) condition (Billingsley 1968) where

$$|P(A \cap B) - P(A)P(B)| \leq \phi_k P(A)$$

for any $A \in \mathcal{F}_1^n$ and $B \in \mathcal{F}_{n+k}^{\infty}$ and ϕ_k tends to zero for $k \to \infty$. The rate at which α_k, β_k or ϕ_k go to zero plays an important role in showing asymptotic properties of the nonparametric smoothing procedures. We note that these

conditions are in general difficult to check. However, if the process follows a stationary Markov chain, then geometric ergodicity implies absolute regularity, which in turn implies strong mixing conditions. Techniques exist for checking geometric ergodicity, see e.g. Doukhan (1994) or Lu (1998). Further and more detailed conditions will be discussed in Subsection 2.2.

The quantlet `regxest` allows to compute Nadarya-Watson estimates of $f(\cdot)$ for an array of different y's. Its syntax is

```
mh = regxest(x{, h, K, v})
```

with the input variables

`x`
: $(T-1) \times 2$ matrix, in the first column the independent, in the second column the dependent variable,

`h`
: scalar, bandwidth for which if not given, 20% of the range of the values in the first column of `x` is used,

`K`
: string, kernel function on [-1,1] or Gaussian kernel `"gau"` for which if not given, the Quartic kernel `"qua"` is used,

`v`
: $m \times 1$ vector of values of the independent variable on which to compute the regression for which if not given, `x` is used.

This quantlet returns a $(T-1) \times 2$ or $m \times 2$ matrix `mh`, where the first column is the sorted first column of `x` or the sorted `v`, the second column contains the regression estimate on the values of the first column.

In order to illustrate the methods presented in this chapter, we model the dynamics underlying the famous annual Canadian lynx trappings in 1821–1934, see e.g. Brockwell and Davis (1991, Appendix, Series G). Figures 1 and 2 of their original and logged time series are obtained with the quantlet

```
library("plot")
setsize(640,480)
lynx        = read("lynx.dat")  ; read data
```

```
d1          = createdisplay(1,1)
x1          = #(1821:1934)~lynx
setmaskl (x1, (1:rows(x1))', 0, 1)
show(d1,1,1,x1)                  ; plot data
setgopt(d1,1,1,"title","Annual Canadian Lynx
                                 Trappings, 1821-1934")
setgopt(d1,1,1,"xlabel","Years","ylabel","Lynx")
d2          = createdisplay(1,1)
x2          = #(1821:1934)~log(lynx)
setmaskl (x2, (1:rows(x2))', 0, 1)
show(d2,1,1,x2)                  ; plot data
setgopt(d2,1,1,"title","Logs of Annual Canadian
                                 Lynx Trappings, 1821-1934")
setgopt(d2,1,1,"xlabel","Years","ylabel","Lynx")
```

flts01.xpl

Their inspection indicates that taking logarithms is required to make the time series look stationary. The following quantlet reads the `lynx` data set, constructs the vectors of the dependent and lagged variables, computes the Nadaraya-Watson estimator and plots the resulting function including the scatter plot which is displayed in Figure 3. For selecting the bandwidth we use here the primitive rule to take one fifth of the data range.

```
library("smoother")
library("plot")
setsize(640,480)
;                        data preparation
lynx       = read("lynx.dat")
lynxrows   = rows(lynx)
lag1       = lynx[1:lynxrows-1]     ; vector of first lag
y          = lynx[2:lynxrows]       ; vector of dep. var.
data       = lag1~y
data       = log(data)
;                        estimation
h          = 0.2*(max(data[,1])-min(data[,1])); crude bandwidth
"Bandwidth used" h
mh         = regxest(data,h)        ; N-W estimation
;                        graphics
mh         = setmask(mh,"line","blue")
xy         = setmask(data,"cross","small")
```

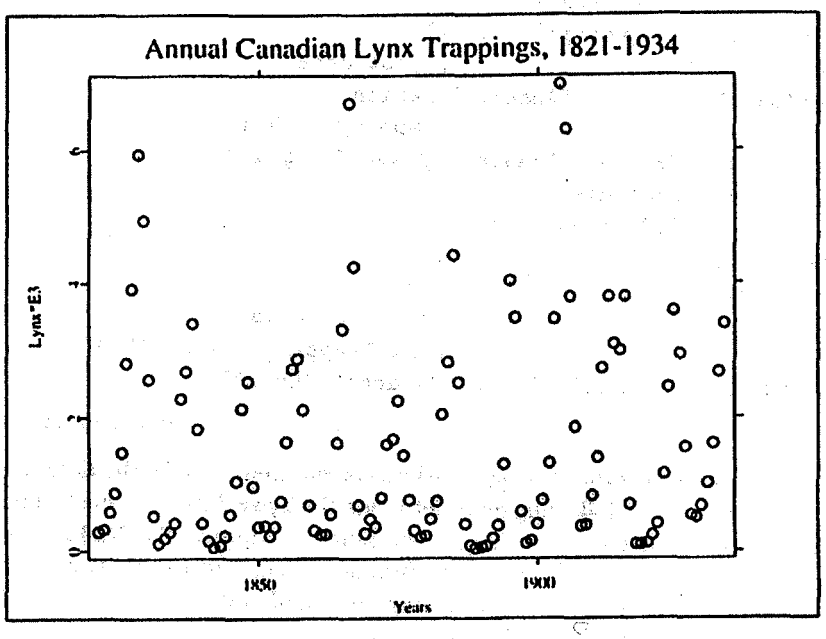


Figure 1: Time series of annual Canadian Lynx Trappings, 1821–1934

```
plot(xy,mh)
setgopt(plotdisplay,1,1,"title","Estimated NAR(1)
                                              mean function")
setgopt(plotdisplay,1,1,"xlabel","First Lag","ylabel","Lynx")
```

flts02.xpl

For long time series the computation of the Nadaraya-Watson estimates may become quite slow since there are more points at which to estimate the function and each estimation involves more data. In this case one may use the WARPing, **weighted average of rounded points**, technique. The basic idea is the "binning" of the data in bins of length d. Each observation is then replaced by the bincenter of the corresponding bin which means that each point is rounded to the precision given by d. A typical choice for d is $h/5$ or $(\max Y_{t-1} - \min Y_{t-1})/100$. In the latter case, the effective sample size r, i.e. the number

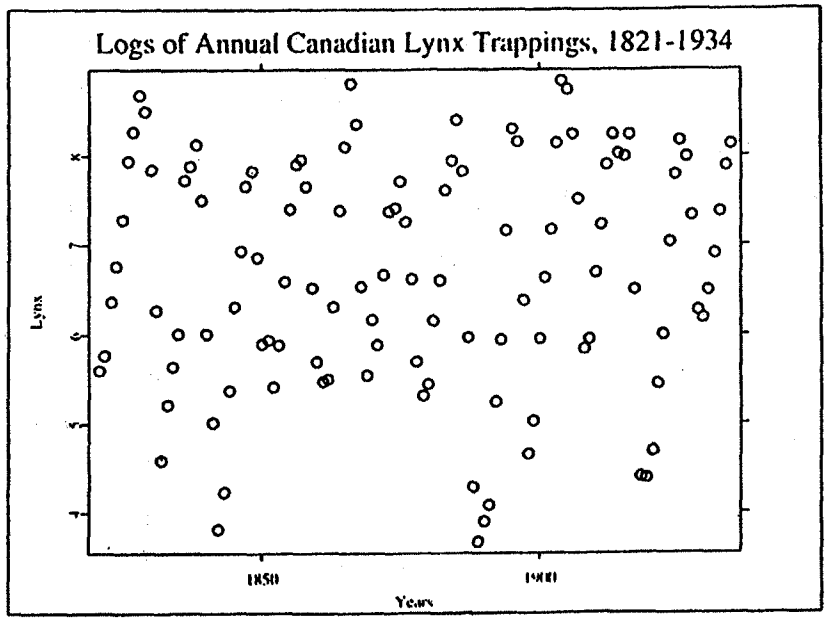


Figure 2: Time series of logarithm of annual Canadian Lynx Trappings, 1821–1934

of nonempty bins, for computation is at most 101. If WARPing is necessary, just call the quantlet **regest** which has the same parameters as the quantlet **regxest**.

While the Nadaraya-Watson function estimate is simple to compute it may suffer from a substantial estimation bias due to the zero order Taylor expansion. Therefore, it seems natural to increase the order p of the expansion. For example, by selecting $p = 1$ one obtains the local linear estimator which corresponds to the following weighted minimiziation problem

$$\{\widehat{c}_0, \widehat{c}_1\} = \arg\min_{\{c_0,c_1\}} \sum_{t=2}^{T} \{Y_t - c_0 - c_1(Y_{t-1} - y)\}^2 K_h(Y_{t-1} - y), \qquad (5)$$

where the estimated function value $\widehat{f}_2(y, h)$ is provided as before by the esti-

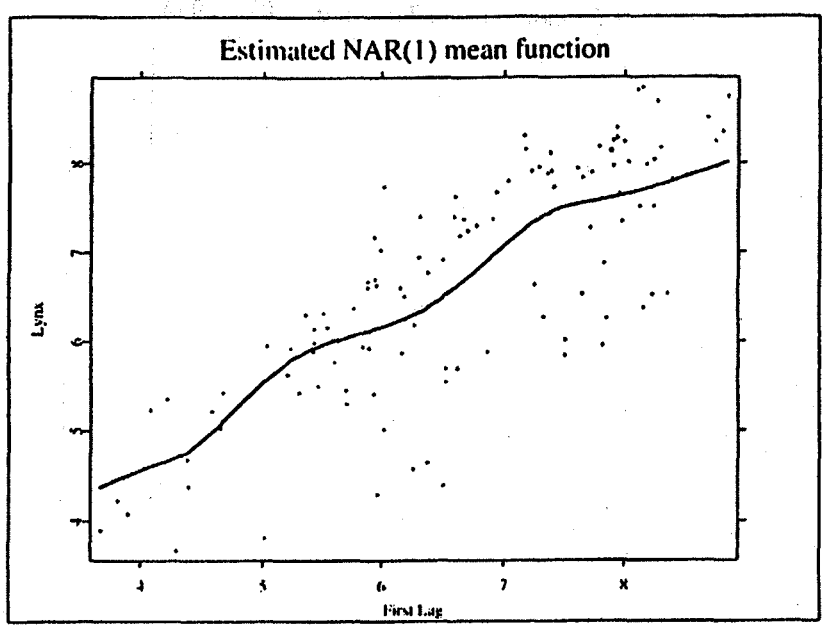


Figure 3: Nadaraya-Watson estimates of NAR(1) mean function for lynx data and scatter plot

mated constant $\widehat{c}_0$. In a similar way one obtains the local quadratic estimator if one chooses $p = 2$. The quantlet `lpregxest` allows to compute local linear or local quadratic function estimates using the quartic kernel. Its syntax is

```
y = lpregxest (x,h {,p {,v}})
```

where the inputs are:

x

$(T - 1) \times 2$ matrix, in the first column the independent, in the second column the dependent variable,

h

scalar, bandwidth for which if not given, the rule-of-thumb bandwidth

computed by the quantlet `lpregrot` is used,

`p`

integer, order of polynomial: p=0 yields the Nadaraya-Watson estimator, p=1 yields local linear estimation (which is default), p=2 (local quadratic) is the highest possible order,

`v`

$m \times 1$, values of the independent variable on which to compute the regression for which if not given, x is used.

The output is given by the

`mh`

$(T-1) \times 2$ or $m \times 2$ matrix, the first column is the sorted first column of x or the sorted v, the second column contains the regression estimate on the values of the first column.

The following quantlet allows to visualize the difference between local constant and local linear estimation of the first order nonlinear autoregressive mean function for the `lynx` data. It produces Figure 4 where the solid and dotted lines display the local linear and local constant estimates, respectively. One notices that the local linear function estimate shows less variation.

```
library("smoother")
library("plot")
setsize(640,480)
;                       data preparation
lynx       = read("lynx.dat")
lynxrows   = rows(lynx)
lag1       = lynx[1:lynxrows-1]     ; vector of first lag
y          = lynx[2:lynxrows]       ; vector of dep. var.
data       = lag1~y
data       = log(data)
;                       estimation
h          = 0.2*(max(data[,1])-min(data[,1])); crude bandwidth
mh         = regxest(data,h)        ; N-W estimation
mhlp       = lpregxest(data,h)      ; local linear estimation
;                       graphics
mh         = setmask(mh,"line","blue","dashed")
```

```
mhlp      = setmask(mhlp,"line","red")
xy        = setmask(data,"cross","small")
plot(xy,mh,mhlp)
setgopt(plotdisplay,1,1,"title","Estimated NAR(1)
                                                mean function")
setgopt(plotdisplay,1,1,"xlabel","First Lag","ylabel","Lynx")
                                                Q flts03.xpl
```

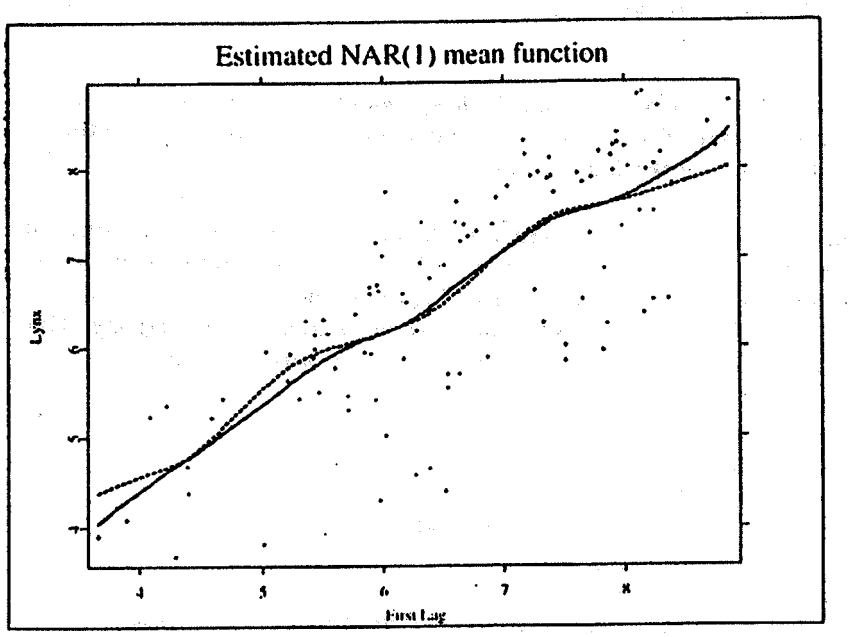


Figure 4: Local linear estimates (solid line) and Nadaraya-Watson estimates (dotted line) of NAR(1) mean function for lynx data and scatter plot

Like Nadaraya-Watson estimation local linear estimation may become slow for long time series. In this case, one may use the quantlet `lpregest` which uses the WARPing technique.

1.2 Bandwidth Selection

`{hcrit, crit} = regxbwsel(x{, h, K})`
interactive tool for bandwidth selection in univariate kernel regression estimation.

`{hcrit, crit} = regbwsel(x{, h, K, d})`
interactive tool for bandwidth selection in univariate kernel regression estimation using the WARPing method.

So far we have used a primitive way of selecting the bandwidth parameter h. Of course, there are better methods for bandwidth choice. They are all based on minimizing some estimated distance measures. Since we are interested in one bandwidth for various y, we look at "global" distances like, for instance, the **integrated squared error** (ISE)

$$d_I(h) = \int \left\{ f(y) - \hat{f}(y,h) \right\}^2 w(y)\mu(y)dy. \tag{6}$$

Here $\mu(\cdot)$ denotes the density of the stationary distribution and $w(\cdot)$ is a weight function with compact support. Note that the bandwidth which minimizes the ISE $d_I(h)$ in generally varies from sample to sample. In practice, one may want to avoid the integration and consider an approximation of the ISE, namely the **average squared error** (ASE)

$$d_A(h) = \frac{1}{T-1} \sum_{t=2}^{T} \left\{ f(Y_{t-1}) - \hat{f}(Y_{t-1},h) \right\}^2 w(Y_{t-1}). \tag{7}$$

Since the measure of accuracy $d_A(h)$ involves the unknown autoregression function $f(\cdot)$, it cannot be used directly. Instead, one may estimate $f(Y_{t-1})$ by Y_t. One then obtains the **average squared error of prediction** (ASEP)

$$d_{AP}(h) = \frac{1}{T-1} \sum_{t=2}^{T} \left\{ Y_t - \hat{f}(Y_{t-1},h) \right\}^2 w(Y_{t-1}). \tag{8}$$

This, however, implies the new problem that $d_{AP}(h)$ can be driven to zero by choosing h small enough. To see this consider the Nadaraya-Watson estimator (4) and imagine that the bandwidth h is chosen so small that (4) becomes $\hat{f}(Y_{t-1},h) = Y_t$. This implies $d_{AP}(h) = 0$. This estimation problem can easily

be solved by always leaving out Y_t in computing (4) which leads to

$$\hat{f}_{-t}(y) = \frac{\sum_{i=2, i\neq t}^{T} K_h(Y_{i-1} - y)Y_i}{\sum_{i=2, i\neq t}^{T} K_h(Y_{i-1} - y)} \tag{9}$$

and is called the **leave-one-out cross-validation estimate** of the autoregression function. One therefore estimates $d_{AP}(h)$ with the cross-validation function

$$CV(h) = \frac{1}{T-1} \sum_{t=2}^{T} \left\{ Y_t - \hat{f}_{-t}(Y_{t-1}, h) \right\}^2 w(Y_{t-1}). \tag{10}$$

Let $\hat{h}$ be the bandwidth that minimizes $CV(h)$. Härdle (1990) and Härdle and Vieu (1992) proved that under an α-mixing condition,

$$\frac{d_A(\hat{h})}{\inf_h d_A(h)} \to 1 \quad \text{in probability.}$$

The interactive quantlet `regxbwsel` offers cross-validation and other bandwidth selection methods. The latter may be used in case of independent data. It is called by

```
{hcrit, crit} = regxbwsel(x{, h, K})
```

with the input variables:

`x`
: $(T-1) \times 2$ vector of the data,

`h`
: $m \times 1$ vector of bandwidths,

`K`
: string, kernel function on $[-1, 1]$ e.g. quartic kernel `"qua"` (default) or Gaussian kernel `"gau"`.

The output variables are:

`hcrit`
: $p \times 1$ vector, selected bandwidths by the different criteria,

`crit`
$p \times 1$ string vector, criteria considered for bandwidth selection.

If one wants to use WARPing one has to use the quantlet `regbwsel`. Using the following quantlet one may estimate the cross-validation bandwidth for the `lynx` data set and obtains $\widehat{h} = 1.12085$.

```
library("smoother")
library("plot")
setsize(640,480)
;                        data preparation
lynx       = read("lynx.dat")
lynxrows   = rows(lynx)
lag1       = lynx[1:lynxrows-1]          ; vector of first lag
y          = lynx[2:lynxrows]            ; vector of dep. var.
data       = lag1~y
data       = log(data)
;
tmp        = regxbwsel(data)
```


`flts04.xpl`

It was already noted that the optimal bandwidth with respect to ISE (6) or ASE (7) may vary across samples. In order to obtain a sample independent optimal bandwidth one may consider the **mean integrated squared error** (MISE)

$$d_M(h) = E\left[\int \left\{f(y) - \hat{f}(y,h)\right\}^2 w(y)\mu(y)dy\right]. \qquad (11)$$

Like $d_I(h)$ or $d_A(h)$, it also cannot be used directly. It is, however, possible to derive the asymptotic expansion of $d_M(h)$. This allows to obtain an explicit formula for the asymptotically optimal bandwidth h_{opt} which, however, contains unknown constants. In Subsection 2.2 we show how one can estimate these unknown quantities in order to obtain a plug-in bandwidth $\widehat{h}_{opt}$.

1.3 Diagnostics

`acfplot(x)`
generates plot of autocorrelation function of time series contained in vector `x`.

`{jb, probjb, sk, k} = jarber(x, 1)`
checks for normality of the data contained in vector `x` using the Jarque-Bera test.

It is well known that if a fitted model is misspecified, then resulting inference can be misleading like, for example, for confidence intervals or significance tests. One way to check whether a chosen model is correctly specified is to investigate the resulting residuals. Most importantly, one checks for autocorrelation remaining in the residuals. This can easily be done by inspecting the graph of the autocorrelation function using the quantlet `acfplot`. It only requires the $(T-1) \times 1$ vector `x` with the estimated residuals as input variable. The quantlet also draws 95% confidence intervals for the case of no autocorrelation.

Another issue is to check the normality of the residuals. This is commonly done by using the Bera-Jarque test suggested by Bera and Jarque (1982). It is commonly called JB-test and can be computed with the quantlet `jarber` which is called by

```
{jb, probjb, sk, k} = jarber(resid, printout)
```

with input variables

`resid`
$(T-1) \times 1$ matrix of residuals,

`printout`
scalar, 0 no printout, 1 printout,

and output variables

`jb`
scalar, test statistic of Jarque-Bera test,

`probjb`
scalar, probability value of test statistics,

sk

scalar, skewness,

k

scalar, kurtosis.

In the following quantlet these diagnostics are applied to the residuals of the NAR(1) model fitted to the lynx data using the Nadaraya-Watson estimator (4) with the cross-validation bandwidth $\widehat{h} = 1.12085$

```
;               load required quantlets
  library("smoother")
  library("plot")
  func("acfplot")
  func("jarber")
  setsize(640,480)
;               data preparation
  lynx      = read("lynx.dat")
  lynxrows  = rows(lynx)
  lag1      = lynx[1:lynxrows-1]          ; vector of first lag
  y         = lynx[2:lynxrows]            ; vector of dep. var.
  data      = lag1~y
  data      = log(data)
  datain    = data~#(1:lynxrows-1)        ; add index to data
  dataso    = sort(datain,1)              ; sorted data
;               estimation
  h         = 1.12085               ; Cross-validation bandwidth
  mhlp      = regxest(dataso[,1|2],h)
                                    ; local constant estimation
;               graphics
  mhlp      = setmask(mhlp,"line","red")
  xy        = setmask(data,"cross","small")
  plot(xy,mhlp)
  setgopt(plotdisplay,1,1,"title",
                              "Estimated NAR(1) mean function")
  setgopt(plotdisplay,1,1,"xlabel","First Lag","ylabel","Lynx")
;               diagnostics
  yhatso    = mhlp.data[,2]~dataso[,3]  ; sorted est. fct. values
  yhat      = sort(yhatso,2)            ; undo sorting
  eps       = data[,2] - yhat[,1]       ; compute residuals
```

```
  acfplot(eps)              ; plot autocorrelation function of res.
  setgopt(dacf,1,1,"title","Autocorrelation function of NAR(1)
                                                    residuals")
;
  {jb,probjb,sk,k} = jarber(eps,1)
        ; compute Jarque-Bera test for normality of residuals
```

Q flts05.xpl

The plot of the resulting autocorrelation function of the residuals is shown in Figure 5. It clearly shows that the residuals are not white noise. This indicates that one should use a higher order nonlinear autoregressive process for modelling the dynamics of the lynx data. This will be discussed in Section 2. Moreover, normality is rejected even at the 1% significance level since the JB-test statistic is 11.779 which implies a p-value of 0.003.

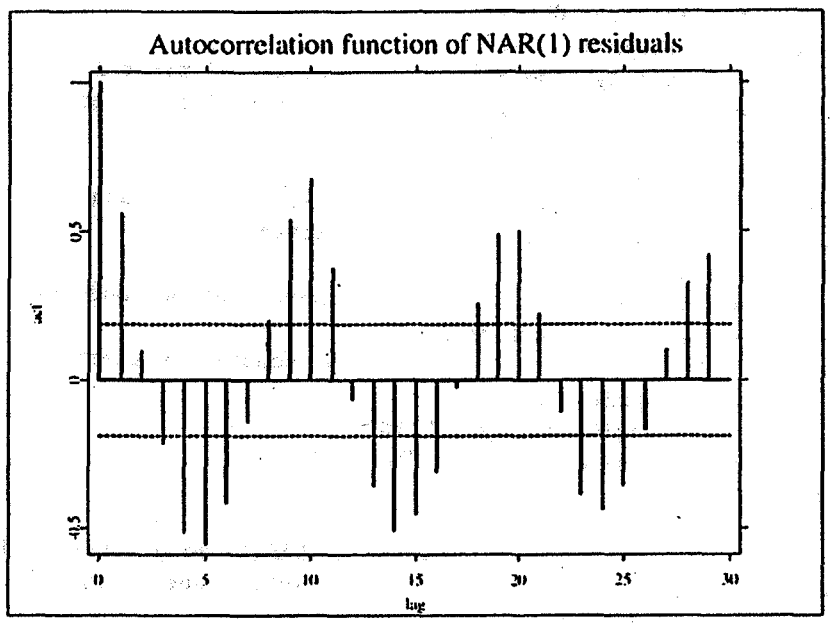


Figure 5: Autocorrelation function of estimated residuals based on a NAR(1) model for the lynx data

1.4 Confidence Intervals

`{mh, clo, cup} = regxci(x{, h, alpha, K, xv})`
computes pointwise confidence intervals with prespecified confidence level for univariate regression using the Nadaraya-Watson estimator.

`{mh, clo, cup} = regci(x{, h, alpha, K, d})`
computes pointwise confidence intervals with prespecified confidence level for univariate regression using the Nadaraya-Watson estimator. The computation uses WARPing.

Once one selected the bandwidth and checked the residuals one often wants to investigate the variance of estimating the autoregression function. Under appropriate conditions, the variance of both the Nadaraya-Watson and the local linear estimator can be approximated by

$$\mathrm{Var}(\widehat{f}(y,h)) \approx \frac{1}{Th}\frac{\sigma^2(y)}{\mu(y)}\|K\|_2^2 \tag{12}$$

as will be seen in Subsection 2.1. (12) can be used for constructing confidence intervals for $\widehat{f}(\cdot)$ since one can estimate the conditional variance $\sigma^2(y)$ by the kernel estimate

$$\widehat{\sigma}^2(y,h) = \frac{\sum_{t=2}^{T} K_h(Y_{t-1}-y)Y_t^2}{\sum_{t=2}^{T} K_h(Y_{t-1}-y)} - \widehat{f}(y,h) \tag{13}$$

and the density $\mu(y)$ by the kernel estimate

$$\widehat{\mu}(y,h) = \sum_{t=1}^{T} K_h(Y_t - y). \tag{14}$$

Based on these estimates the quantlet `regxci` computes pointwise confidence intervals using the Nadaraya-Watson estimator. It is called with

```
{mh, clo, cup} = regxci(x{, h, alpha, K, xv})
```

with input variables:

`x`
: $(T-1) \times 2$ matrix of the data with the independent and the dependent variable in the first and second column, respectively,

`h`
: scalar, bandwidth for which if not given 20% of the range of the values in the first column `x` is used,

`alpha`
: confidence level with 0.05 as default value,

`K`
: string, kernel function on $[-1, 1]$ and the quartic kernel `"qua"` as default,

`xv`
: $m \times 1$ matrix of the values of the independent variable on which to compute the regression and `x` as default.

The output variables are:

`mh`
: $(T-1) \times 2$ or $m \times 2$ matrix, the first column is the sorted first column of `x` or the sorted `xv`, the second column contains the regression estimate on the values of the first column,

`clo`
: $(T-1) \times 2$ or $m \times 2$ matrix, the first column is the sorted first column of `x` or the sorted `xv`, the second column contains the lower confidence bounds on the values of the first column,

`cup`
: $(T-1) \times 2$ or $m \times 2$ matrix, the first column is the sorted first column of `x` or the sorted `xv`, the second column contains the upper confidence bounds on the values of the first column.

If the WARPing technique is required, one uses the quantlet `regci`.

In Subsection 1.3 we found that the NAR(1) model for the `lynx` data is misspecified. Therefore, it is not appropriate for illustrating the computation of pointwise confidence intervals. Instead we will use a simulated time series. The quantlet below generates 150 observations of a stationary exponential AR(1)

process

$$Y_t = 0.3Y_{t-1} + 2.2Y_{t-1}\exp\left(-0.1Y_{t-1}^2\right) + \xi_t, \quad \xi \sim N(0,1), \qquad (15)$$

calls the interactive quantlet **regxbwsel** for bandwidth selection where one has to choose for the first time cross-validation and for the second time stop, computes the confidence intervals and plots the true and estimated function (solid and dashed line) as well as the pointwise confidence intervals (dotted line) as shown in Figure 6.

```
library("smoother")
library("plot")
library("times")
setsize(640,480)

;                    generate exponential AR(1) process
phi1     = 0.3
phi2     = 2.2
g        = 0.1
randomize(0)
x        = genexpar(1,g,phi1,phi1+phi2,normal(150))

;                    data preparation
xrows    = rows(x)
lag1     = x[1:xrows-1]              ; vector of first lag
y        = x[2:xrows]                ; vector of dep. var.
data     = lag1~y

;                    true function
f        = sort(lag1~(phi1*lag1 + phi2*lag1.*exp(-g*lag1~2)),1)

;                    estimation
{hcrit,crit}     = regxbwsel(data)
{mh, clo, cup}   = regxci(data,hcrit)

f        = setmask(f,"line","solid","red")
data     = setmask(data,"cross")
mh       = setmask(mh,"line","dashed","blue")
clo      = setmask(clo,"line","blue","thin","dotted")
cup      = setmask(cup,"line","blue","thin","dotted")
```

```
plot(data,f,mh,clo,cup)
setgopt(plotdisplay,1,1,"title","Confidence intervals of
                              estimated NAR(1) mean function")
setgopt(plotdisplay,1,1,"xlabel","First Lag","ylabel","Y")
```

Q flts06.xpl

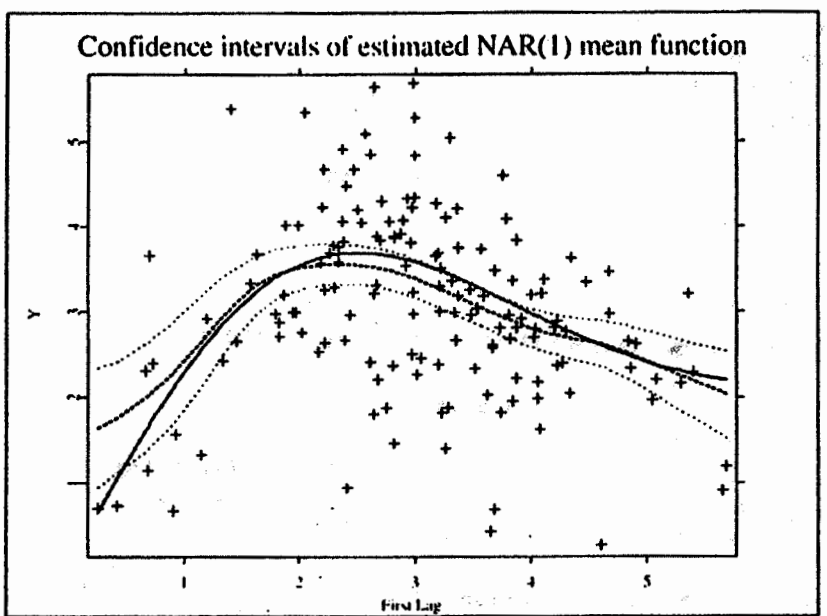


Figure 6: True and estimated mean function plus pointwise confidence intervals for a generated exponential AR(1) process

1.5 Derivative Estimation

`mh = lpderxest(x, h{, q, p, K, v})`
estimates the q-th derivative of a regression function using local polynomial kernel regression with quartic kernel.

`mh = lpderest(x, h{, q, p, K, d})`
estimates the q-th derivative of a autoregression function using local polynomial kernel regression. The computation uses WARPing.

When investigating the properties of a conditional mean function, one is often interested in its derivatives. The estimation of derivatives can be accomplished by using local polynomial estimation as long as the order p of the polynomial is at least as large as the order q of the derivative to be estimated. Using a local quadratic estimator

$$\{\hat{c}_0, \hat{c}_1, \hat{c}_2\}$$
$$= \underset{\{c_0,c_1,c_2\}}{\operatorname{argmin}} \sum_{t=1}^{T} \left\{Y_t - c_0 - c_1(Y_{t-1} - y) - c_2(Y_{t-1} - y)^2\right\}^2 K_h(Y_{t-1} - y)$$

one estimates the first and second derivative of $f(y)$ at y with

$$\hat{f}'(y,h) = \hat{c}_1, \quad \hat{f}''(y,h) = 2\hat{c}_2.$$

In general, one uses a $q+1$ instead of a q-th order polynomial for the estimation of the q-th derivative since this reduces the complexity of the estimation bias, see e.g. Fan and Gijbels (1995). The estimated derivative is then obtained as $\hat{f}^{(q)} = q!\hat{c}_q$. The quantlet `lpderxest` allows to estimate first and second order derivatives where maximally a second order polynomial is used. It is called by

```
mh = lpderxest (x, h{, q, p, K, v})
```

with input variables

x
$(T-1)\times 2$ matrix of the data with the independent and dependent variable in the first and second column, respectively.

h

scalar, bandwidth for which if not given the rule-of-thumb bandwidth is computed with `lpderrot`,

q

integer ≤ 2, order of derivative for which if not given, q=1 (first derivative) is chosen,

p

integer, order of polynomial for which if not given, p=q + 1 is used for q< 2 and p=q is used for q=2.

v

$m \times 1$, values of the independent variable on which to compute the regression for which if not given, x is used.

The output variable is

mh

$(T-1) \times 2$ or $m \times 2$ matrix where the first column is the sorted first column of x or the sorted v and the second column contains the derivative estimate on the values of the first column.

The quantlet `lpderest` which applies the WARPing technique (Fan and Marron 1994) allows for $p \leq 5$ and $q \leq 4$. We note, however, that WARPing may waste a lot of information. Bandwidth selection remains an important issue and can be done using the quantlet `lpderrot`.

In the following quantlet we estimate the first and second derivatives of the conditional mean function of the exponential AR(1) process (15) based on 150 observations. The true derviatives (solid lines) and their estimates (dashed lines) are shown in Figures 7 and 8.

```
library("smoother")
library("plot")
library("times")
setsize(640,480)
;                    generate exponential AR(1) process
phi1     = 0.3
phi2     = 2.2
g        = 0.1
```

```
  randomize(0)
  x       = genexpar(1,g,phi1,phi1+phi2,normal(150))

;                       data preparation
  xrows   = rows(x)
  lag1    = x[1:xrows-1]              ; vector of first lag
  y       = x[2:xrows]                ; vector of dep. var.
  data    = lag1~y
  ffder   = sort(lag1~(phi1 + exp(-g*lag1^2).*
                                  phi2.*(1-2.*g.*lag1^2)),1)
  fsder   = sort(lag1~(exp(-g*lag1^2).*(-2*g.*lag1)*
                                  phi2.*(3-2.*g.*lag1^2)),1)

;                       estimate first derivative
  ffder   = setmask(ffder,"line","solid","red")
  mhfder  = lpderxest(data)
  mhfder  = setmask(mhfder, "line","dashed","blue")
  plotder = createdisplay(1,1)
  show(plotder,1,1,ffder,mhfder)
  setgopt(plotder,1,1,"title","Estimated first derivative
                                  of mean function")
  setgopt(plotder,1,1,"xlabel","First lag","ylabel",
                                  "First derivative")
;                       estimate second derivative
  fsder   = setmask(fsder,"line","solid","red")
  hrot    = 2*lpderrot(data,2)
  mhsder  = lpderxest(data,hrot,2)
  mhsder  = setmask(mhsder, "line","dashed","blue")
  plot(fsder,mhsder)
  setgopt(plotdisplay,1,1,"title","Estimated second
                                  derivative of mean function")
  setgopt(plotdisplay,1,1,"xlabel","First lag","ylabel",
                                  "Second derivative")
```


flts07.xpl

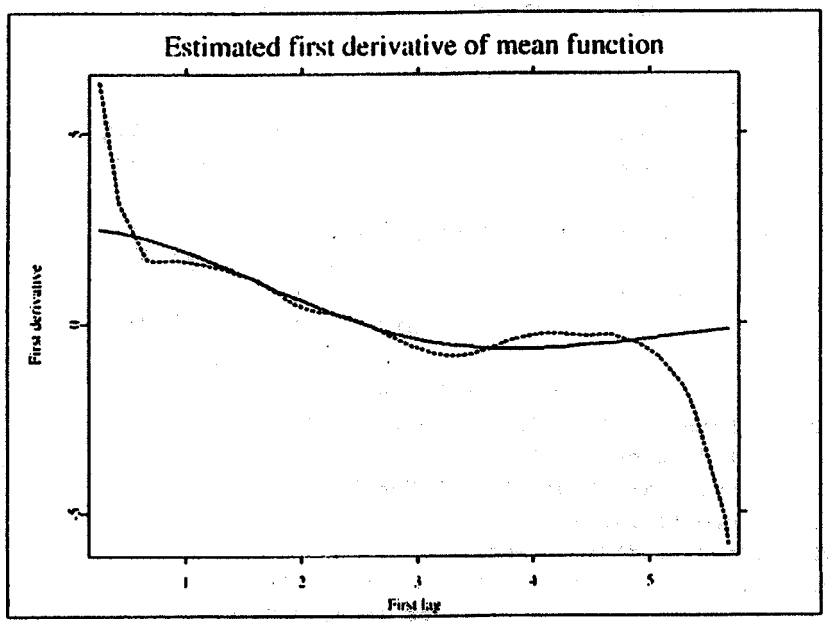


Figure 7: True and estimated first derivative for a generated exponential AR(1) process

2 Nonlinear Autoregressive Models of Higher Order

In Subsection 1.3 we briefly discussed diagnostics to check for the correct specification of a time series model. There we found for the lynx data set that the nonlinear autoregressive model of order one (2) is of too low order to capture the linear correlation in the data. For practical flexible time series modelling it is therefore necessary to allow for higher order nonlinear autoregressive models (1). Their estimation and the selection of relevant lags will be discussed in this section. To simplify notation, we introduce the vector of lagged variables $X_l = (Y_{i_1}, Y_{i_2}, \ldots, Y_{i_m})^T$ such that (1) can be written as

$$Y_l = f(X_l) + \sigma(X_l)\xi_l \tag{16}$$

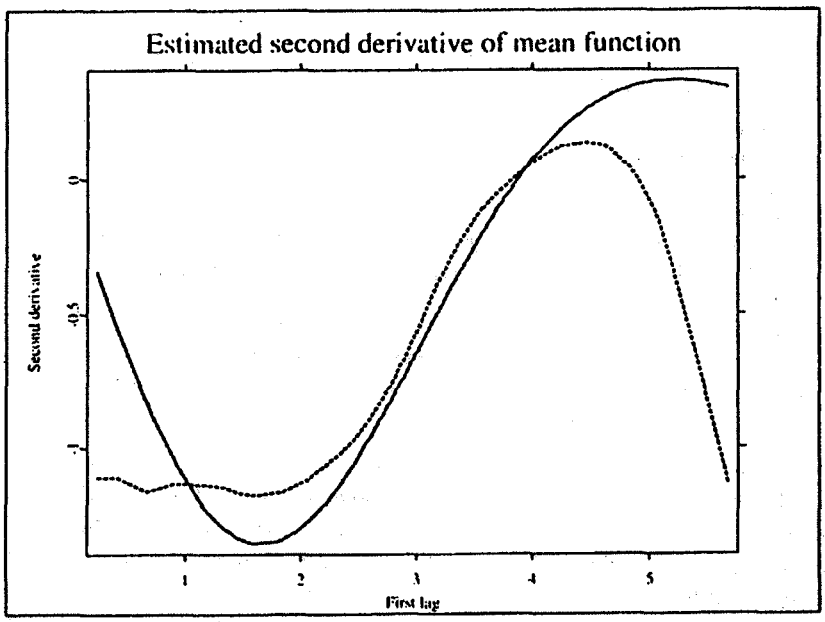


Figure 8: True and estimated second derivative a generated exponential AR(1) process

2.1 Estimation of the Conditional Mean

`mh = regxestp(x{, h, K, v})`
computes the Nadaraya-Watson estimator for multivariate autoregression.

`mh = regestp(x{, h, K, d})`
Nadaraya-Watson estimator for multivariate regression. The computation uses WARPing.

```
mh = lpregxestp(x{, h, K, v})
```
estimates a multivariate regression function using local polynomial kernel regression with quartic kernel.

```
mh = lpregestp(x{, h, K, d})
```
estimates a multivariate regression function using local polynomial kernel regression. The computation uses WARPing.

```
{mA, gsqA, denA, err} = fvllc(Xsj, Yorig, h, Xtj,
                              kernreg, lorq, fandg, loo)
```
estimates a multivariate regression function using local linear regression with Gaussian kernel.

It is not difficult to extend the Nadaraya-Watson estimator (4) and local linear estimator (5) to several lags in the conditional mean function $f(\cdot)$. One then simply uses Taylor expansions of order p for several variables. In the weighted minimization problem of the local constant estimator (3) one has to extend the kernel function $K_h(\cdot)$ for several lagged variables. The simplest way of doing this is to use a product kernel

$$K_h(X_t - x) = \prod_{j=1}^{m} h_j^{-1} K\left(\frac{X_{t,j} - x_j}{h_j}\right) \tag{17}$$

where one $h = (h_1, h_2, \ldots, h_m)^T$ is a vector of bandwidths for each lag or variable. Of course, one may also use the same bandwidth $h = (h, h, \ldots, h)^T$ for all lags in which case we write $K_h(X_t - x)$. Using a scalar bandwidth, (3) becomes

$$\widehat{c}_0 = \arg\min_{\{c_0\}} \sum_{t=i_m+1}^{T} \{Y_t - c_0\}^2 K_h(X_t - x) \tag{18}$$

and the Nadaraya-Watson estimator is given by

$$\widehat{f}_1(x, h) = \widehat{c}_0 = \frac{\sum_{t=i_m+1}^{T} K_h(X_t - x) Y_t}{\sum_{t=i_m+1}^{T} K_h(X_t - x)}. \tag{19}$$

Note that from now on we indicate the Nadaraya-Watson estimator and local linear estimator by the indices 1 and 2, respectively.

The local linear estimator with $p = 1$ is derived from the weighted minimization

$$\{\widehat{c}_0, \widehat{c}_1\} = \arg\min_{(c_0,c_1)} \sum_{t=i_m+1}^{T} \{Y_t - c_0 - c_1(X_t - x)\}^2 K_h(X_t - x). \quad (20)$$

Using the notation

$$Z_2 = \begin{pmatrix} 1 & \cdots & 1 \\ X_{i_m+1} - x & \cdots & X_T - x \end{pmatrix}^T, \quad Y = (Y_{i_m+1}, \ldots, Y_T)^T$$

$$e = (1, 0_{1\times m})^T, \quad W = \operatorname{diag}\left\{\frac{1}{T - i_m} K_h(X_t - x)\right\}_{t=i_m+1}^{T},$$

the estimate $\widehat{f}_2(x, h) = \widehat{c}_0$ can be written for any $x \in \mathbb{R}^m$ as

$$\widehat{f}_2(x) = e^T \left(Z_2^T W Z_2\right)^{-1} Z_2^T W Y. \quad (21)$$

Under suitable conditions which are listed in Subsection 2.2 the Nadaraya-Watson estimator (19) and local linear estimator (21) have an asymptotic normal distribution

$$T^{2/(4+m)} \left(\widehat{f}_a(x, h) - f(x)\right) \to N\left(\beta^2 \frac{\sigma_K^2}{2} r_a, \beta^{-m} \frac{\sigma(x)}{\mu(x)} ||K||_2^2\right), \quad a = 1, 2 \quad (22)$$

where

$$r_1(x) = \operatorname{Tr}\left\{\nabla^2 f(x)\right\} + 2\nabla^T \mu(x) \nabla f(x)/\mu(x), \quad r_2(x) = \operatorname{Tr}\left\{\nabla^2 f(x)\right\}. \quad (23)$$

Thus, the rate of convergence deteriorates with the number of lags. This feature is commonly called the 'curse of dimensionality' and often viewed as a substantial drawback of nonparametric methods. One should keep in mind, however, that the $\sqrt{T}$-rate of parametric models only holds if one estimates a model with an a priori chosen finite number of parameters which may imply a large estimation bias in case of misspecified models. If, however, one allows the number of parameters of parametric models to grow with sample size, $\sqrt{T}$-convergence may no longer hold.

The quantlets `regxestp` and `lregxestp` compute the Nadaraya-Watson estimator (19) and local linear estimator (21) for higher order autoregressions. They are called by

```
mh = regxestp(x{, h, K, v})
```

or

```
mh = lregxestp(x{, h, K, v})
```

with input variables

x

$(T - i_m) \times (m + 1)$ matrix of the data with the m lagged variables in the first m columns and the dependent variable in the last column,

h

scalar or $m \times 1$ or $1 \times m$ vector of bandwidth for which if not given 20% of the range of the values in the first column of **x** is used,

K

string, kernel function on [-1,1] or Gaussian kernel "gau" for which if not given, the quartic kernel "qua" is used,

v

$n \times m$ matrix of values of the independent variable on which to compute the regression for which if not given, a grid of length 100 ($m = 1$), length 30 ($m = 2$) and length 8 ($m = 3$) is used in case of $m < 4$. When $m \geq 4$ then **v** is set to **x**.

The output variable is a

mh

$(T - i_m) \times (m + 1)$ or $n \times (m + 1)$ matrix where the first m columns contain the grid or the sorted first m columns of **x**, the $m + 1$ column contains the regression estimate on the values of the first m columns.

As before, there are also quantlets which apply WARPing. They are called `regestp` and `lregestp`, respectively.

Since we found in Subsection 1.3 that a NAR(1) model is not sufficient to capture the dynamics of the lynx trappings, we compute and plot in the following quantlet the autoregression function for lag 1 and 2 for both estimators using the crude bandwidth of 20% of the data range. Note that you have to click on the graph and rotate it in order to see the regression surface.

```
  library("smoother")
  library("plot")
  setsize(640,480)

;                       data preparation
  lynx      = read("lynx.dat")
  lynxrows  = rows(lynx)
  lag1      = lynx[1:lynxrows-2]        ; vector of first lag
  lag2      = lynx[2:lynxrows-1]        ; vector of second lag
  y         = lynx[3:lynxrows]          ; vector of dep. var.
  data      = lag1~lag2~y
  data      = log(data)

;                       estimation
  h         = 0.2*(max(data[,1])-min(data[,1])) ; crude bandwidth
  mh        = regxestp(data,h)          ; local constant estimation
  mhlp      = lregxestp(data,h)         ; local constant estimation

;                       graphics
  mhplot    = createdisplay(1,1)
  mh        = setmask(mh,"surface","blue")
  show(mhplot,1,1,data,mh)              ; surface plot
  setgopt(mhplot,1,1,"title",
                          "Nadaraya-Watson estimate -- ROTATE!")
  mhlpplot  = createdisplay(1,1)
  mhlp      = setmask(mhlp,"surface","red")
  show(mhlpplot,1,1,data,mhlp)          ; surface plot
  setgopt(mhlpplot,1,1,"title",
                          "Local linear estimate -- ROTATE!")
                                                  flts08.xpl
```

Figures 9 and 10 show three-dimensional plots of the observations and the estimated regression function. In Figure 9 one can clearly see the problem of boundary effects, i.e. in regions where are no or only few data points the estimated function values may easily become erratic if the bandwidth is too small. Therefore, a selected bandwidth may be appropriate for regions with plenty of observations while inappropriate elsewhere. As can be seen from Figure 10, this boundary problem turns out to be worse for the local linear estimator where

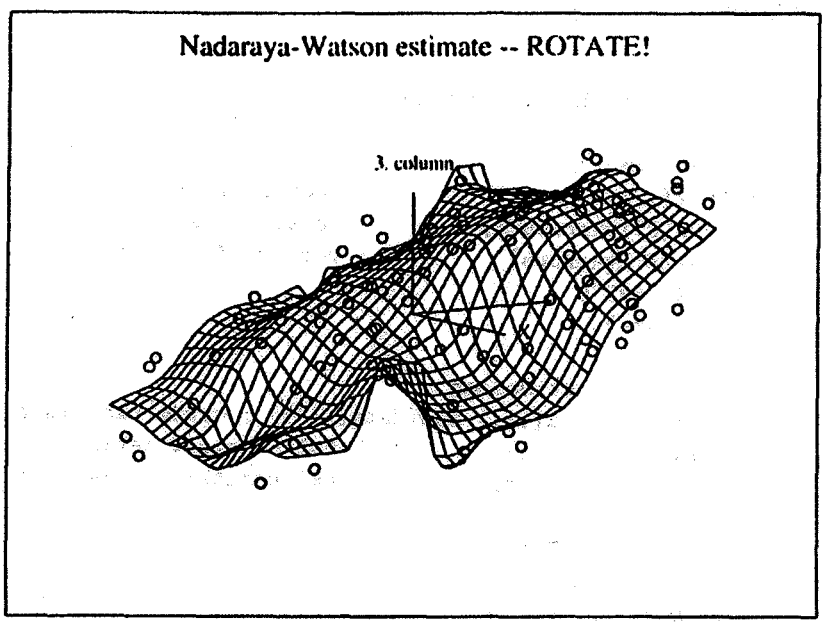


Figure 9: Observations and Nadaraya-Watson estimate of NAR(2) regression function for the lynx data

one observes a large outlier for one grid point. Such terrible estimates happen if the inversion in (20) is imprecise due to a too small bandwidth. One then has to increase the bandwidth. Try the quantlet `flts08.xpl` with replacing in the crude bandwidth choice the factor 0.2 by 2. Note that increasing the bandwidth makes the estimated regression surfaces of the two estimators look flat and closer to linearity, respectively. This, however, can increase the estimation bias. Therefore, an appropriate bandwidth choice is important. It will be discussed in the next section.

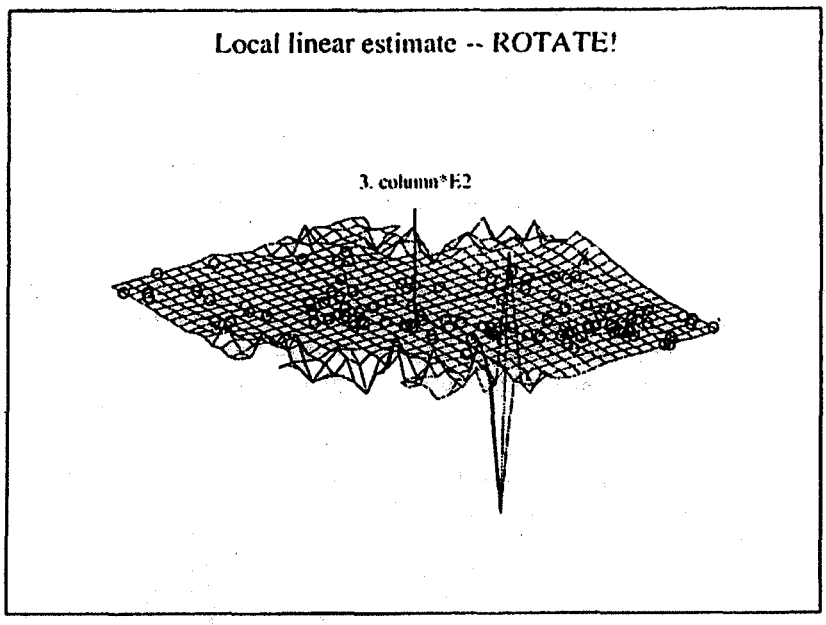


Figure 10: Observations and local linear estimate of NAR(2) regression function for the lynx data

2.2 Bandwidth and Lag Selection

```
{Bhat, Bhatr, hB, Chat, sumwc, hC, hA} = hoptest (xsj,
     yorig, xtj, estimator, kernel, ntotal, sigy2, perB,
                                          lagmax, robden)
```

quantlet to compute plug-in bandwidth for multivariate regression or nonlinear autoregressive processes of higher order.

```
{crmin, crpro} = cafpe(y, truedat, xdataln, xdatadif,
                       xdatastand, lagmax, searchmethod, dmax)
```
quantlet for local linear lag selection for the conditional mean function based on the Asymptotic Final Prediction Error ($AFPE_2$) or its corrected versions ($CAFPE_2$) using default settings.

```
{crmin, crpro, crstore, crstoreadd, hstore, hstoretest}
    = cafpefull(y, truedat, xresid, trueres, xdataln,
       xdatadif, xdatastand, lagmax, volat, searchmethod,
            dmax, selcrit, robden, perA, perB, startval,
                                        noutputf, outpath)
```
quantlet for local linear lag selection for the conditional mean or volatility function based on the asymptotic final prediction error ($AFPE_2$) or its corrected version ($CAFPE_2$).

```
{mA, gsqA, denA, err} = fvllc(Xsj, Yorig, h, Xtj,
                                    kernreg, lorq, fandg, loo)
```
can estimate the multivariate regression function, first or second direct derivatives using local linear or partial local quadratic regression with Gaussian kernel.

The example of the previous section showed that the bandwidth choice is very important for higher order autoregressive models. Equally important is the selection of the relevant lags. Both will be discussed in this section. The presented procedures are based on Tschernig and Yang (2000).We start with the problem of selecting the relevant lags. For this step it is necessary to a priori specify a set of possible lag vectors by choosing the maximal lag M. Denote the full lag vector containing all lags up to M by $X_{t,M} = (Y_{t-1}, Y_{t-2}, \ldots, Y_{t-M})^T$. The lag selection task is now to eliminate from the full lag vector $X_{t,M}$ all lags that are redundant. Let us first state the assumptions that Tschernig and Yang (2000) require:

(A1) For some $M \geq i_m$ the vector process $X_{t,M}$ is strictly stationary and β-mixing with $\beta(T) \leq k_0 T^{-(2+\delta)/\delta}$ for some $\delta > 0$, $k_0 > 0$.

(A2) The stationary distribution of the process $X_{t,M}$ has a continuous density $\mu_M(x_M)$, $x_M \in \mathbb{R}^M$. Note that $\mu(\cdot)$ is used for denoting $\mu_M(\cdot)$ and all of its marginal densities.

(A3) The function $f(\cdot)$ is twice continuously differentiable while $\sigma(\cdot)$ is continuous and positive on the support of $\mu(\cdot)$.

(A4) The errors $\{\xi_t\}_{t \geq i_m}$ have a finite fourth moment m_4.

(A5) The support of the weight function $w(\cdot)$ is compact with nonempty interior. The function $w(\cdot)$ is continuous, nonnegative and $\mu(x_M) > 0$ for x_M in the support of $w(\cdot)$.

(A6) The kernel function $K : \mathbb{R} \to \mathbb{R}$ is a symmetric probability density and the bandwidth h is a positive number with $h \to 0$, $nh^m \to \infty$ as $n \to \infty$.

For the definition of β-mixing see Section 1.1 or Doukhan (1994). Conditions (A1) and (A2) can be checked using e.g. Doukhan (1994, Theorem 7 and Remark 7, pp. 102, 103). Further conditions can be found in Lu (1998).

For comparing the quality of competing lag specifications, one needs an appropriate measure of fit, as for example the **final prediction error** (FPE)

$$FPE_a(h, i_1, \ldots, i_m) = E\left[\left(\check{Y}_t - \hat{f}_a(\check{X}_t, h)\right)^2 w(\check{X}_{t,M})\right], \quad a = 1, 2. \tag{24}$$

In the definition of the $FPE(\cdot)$ the process $\{\check{Y}_t\}$ is assumed to be independent of the process $\{Y_t\}$ but to have the same stochastic properties. If we now indicate the vector of lagged values of the data generating process by the superscript * and assume its largest lag is smaller than the chosen M, we can easily relate the definition of the FPE (24) to the MISE

$$d_{a,M}(h, i_1, \ldots, i_m) = E\left[\int \left\{f(x^*) - \hat{f}_a(x)\right\}^2 w(x_M)\mu(x_M)dx_M\right], \tag{25}$$

which here extends (11) to functions with several lags. First note that

$$\begin{aligned} FPE_a(h, i_1, \ldots, i_m) &= E\left\{E\left[\left(\check{Y}_t - \hat{f}_a(\check{X}_t, h)\right)^2 w(\check{X}_{t,M}) | Y_1, \ldots, Y_T\right]\right\} \\ &= E\left\{\int \left(y - \hat{f}_a(x)\right)^2 w(x_M)\mu(y, x_M)dydx_M\right\}. \end{aligned}$$

Using $\left\{y - \hat{f}(x)\right\}^2 = \left\{y - f(x)^* + f(x)^* - \hat{f}(x)\right\}^2$ one obtains the decomposition

$$FPE_a(h, i_1, \ldots, i_m) = A + d_{a,M}(h, i_1, \ldots, i_m), \tag{26}$$

where

$$A = \int \sigma^2(x^*)w(x_M)\mu(x_M)dx_M \tag{27}$$

denotes the mean variance or final prediction error for the true function $f(x^*)$. Therefore, it follows from (26) that the FPE measures the sum of the mean variance and the MISE.

In the literature mainly two approaches were suggested for estimating the unknown $FPE_a(\cdot)$ or variants thereof, namely cross-validation (Vieu 1994), (Yao and Tong 1994) or estimation of an asymptotic expression of the $FPE_a(\cdot)$ (Auestad and Tjøstheim 1990), (Tjøstheim and Auestad 1994), (Tschernig and Yang 2000). Given Assumptions (A1) to (A6), Tschernig and Yang (2000, Theorem 2.1) showed that for the local constant estimator, $a = 1$, and the local linear estimator, $a = 2$, one has $FPE_a(h, i_1, \ldots, i_m) = AFPE_a(h, i_1, \ldots, i_m) + o\{h^4 + (T - i_m)^{-1}h^{-m}\}$ where

$$AFPE_a(h, i_1, \ldots, i_m) = A + b(h)B + c(h)C_a \tag{28}$$

denotes the **asymptotic final prediction error**. The terms $b(h)B$ and $c(h)C$ denote the expected variance and squared bias of the estimator, respectively, with the constants

$$B = \int \sigma^2(x^*)w(x_M)\mu(x_M)/\mu(x)dx_M, \tag{29}$$

$$C_a = \int r_a(x)^2 w(x_M)\mu(x_M)dx_M \tag{30}$$

and the variable terms

$$b(h) = \|K\|_2^{2m}(T - i_m)h^{-m}, \quad c(h) = \sigma_K^4 h^4/4 \tag{31}$$

with $\|K\|_2^2 = \int K(u)^2du$ and $\sigma_K^2 = \int K(u)u^2du$. The sum of the expected variance and squared bias of the estimator just represents the asymptotic mean squared error. Note that if the vector of correct lags X_t^* is included in X_t, then $AFPE_a(h, \cdot)$ tends to A as both $b(h)B$ and $c(h)C_a$ tend to zero.

From (28) it is possible to determine the asymptotically optimal bandwidth h_{opt} by minimizing the asymptotic MISE, i.e. solving the variance-bias tradeoff between $b(h)B$ and $c(h)C$. The asymptotically optimal bandwidth is given by

$$h_{a,opt} = \left\{m\|K\|_2^{2m}B(T - i_m)^{-1}C_a^{-1}\sigma_K^{-4}\right\}^{1/(m+4)}. \tag{32}$$

Note that for a finite asymptotically optimal bandwidth to exist one has to assume that

(A7) C_a defined in (30) is positive and finite.

This requirement implies that in case of local linear estimation there does not exist a finite $h_{2,opt}$ for linear processes. This is because there does not exist an approximation bias and thus a larger bandwidth has no cost.

In order to obtain the plug-in bandwidth $\hat{h}_{a,opt}$ one has to estimate the unknown constants B and C_a. A local linear estimate of B (29) is obtained from

$$\hat{B}_2(h_B) = T^{-1} \sum_{t=1}^{T} \left\{ Y_t - \hat{f}_2(X_t, h_B) \right\}^2 w(X_{t,M}) / \hat{\mu}(X_t, h_B),$$

where $\hat{\mu}(\cdot)$ is the Gaussian kernel estimator (40) of the density $\mu(x)$. For estimating h_B one may use Silverman's (1986) rule-of-thumb bandwidth

$$\hat{h}_B = \hat{\sigma} \left(\frac{4}{T+2} \right)^{1/(m+4)} T^{-1/(m+4)} \tag{33}$$

with $\hat{\sigma} = \left(\prod_{j=1}^{m} \sqrt{Var(X_j)} \right)^{1/m}$ denoting the geometric mean of the standard deviation of the regressors.

For the local linear estimator (21), C_2 (30) can be consistently estimated by

$$\hat{C}_2(h_C) = \frac{1}{T} \sum_{t=i_m+1}^{T} \left[\sum_{j=1}^{m} \hat{f}^{(jj)}(X_t, h_C) \right]^2 w(X_{t,M}), \tag{34}$$

where $f^{(jj)}(\cdot)$ denotes the second direct derivative of the function $f(\cdot)$. It can be estimated using the partial local quadratic estimator

$$\{\hat{c}_0, \hat{c}_{11}, \ldots, \hat{c}_{1m}, \hat{c}_{21}, \ldots, \hat{c}_{2m}\} = \arg\min_{\{c_0, c_{11}, \ldots, c_{1m}, c_{21}, \ldots, c_{2m}\}} \tag{35}$$

$$\sum_{t=i_m+1}^{T} \{ Y_t - c_0 - c_{11}(X_{t1} - x_1) - \cdots - c_{1m}(X_{tm} - x_m) - c_{21}(X_{t1} - x_1)^2 - \cdots - \cdots - c_{2m}(X_{tm} - x_m)^2 \}^2 K_h(X_t - x).$$

The estimates of the direct second derivatives are then given by $\hat{f}^{(jj)}(x, h) = 2\hat{c}_{2j}$, $j = 1, \ldots, m$. Excluding all cross terms has no asymptotic effects while keeping the increase in the 'parameters' c_0, c_{1j}, c_{2j}, $j = 1, \ldots, m$ linear in the number of lags m. This approach is a simplification of the partial cubic

estimator proposed by Yang and Tschernig (1999) who also showed that the rule-of-thumb bandwidth

$$\hat{h}_C = 2\hat{\sigma}\left(\frac{4}{T+4}\right)^{1/(m+6)} T^{-1/(m+6)} \tag{36}$$

has the optimal rate. We note that for the estimation of C_1 of the Nadaraya-Watson estimator one has additionally to estimate the derivative of the density as it occurs in (23). Therefore, we exclusively use the local linear estimator (21). The direct second derivatives $f^{(jj)}(x)$ can be estimated with the quantlet `tp/capfe/fvllc`.

The plug-in bandwidth $\hat{h}_{2,opt}$ is then given by

$$\hat{h}_{2,opt} = \left\{ m\|K\|_2^{2m} \hat{B}_2(\hat{h}_B)(T - i_m)^{-1} \hat{C}_2(\hat{h}_C)^{-1} \sigma_K^{-4} \right\}^{1/(m+4)}. \tag{37}$$

It now turns out that when taking into account the estimation bias of A, the local linear estimator of $AFPE_2(h, \cdot)$ (28) becomes

$$AFPE_2 = \hat{A}_2(h_{2,opt}) + 2K(0)^m (T - i_m)^{-1} h_{2,opt}^{-m} \hat{B}_2(h_B) \tag{38}$$

and the expected squared bias of estimation drops out. In practice, $h_{2,opt}$ is replaced by the plug-in bandwidth (37). Note that one can interpret the second term in (38) as a penalty term to punish overfitting or choosing superfluous lags. This penalty term decreases with sample size as $h_{2,opt}$ is of order $T^{-1/(m+4)}$. The final prediction error for the true function A (27) is estimated by taking the sample average

$$\hat{A}_2(h) = T^{-1} \sum_{t=1}^{T} \left\{ y_t - \hat{f}_2(X_t, h) \right\}^2 w(X_{t,M})$$

of the residuals from the local linear estimator $\hat{f}_2(X_t, h)$. The asymptotic properties of the lag selection method rely on the fact that the argument of $w(\cdot)$ is the full lag vector $X_{t,M}$.

In order to select the adequate lag vector, one computes (38) for all possible lag combinations with $m \leq M$ and chooses the lag vector with the smallest $AFPE_2$. Given Assumptions (A1) to (A7) and a further technical condition, Tschernig and Yang (2000, Theorem 3.2) showed that this procedure is weakly consistent, i.e. the probability of choosing the correct lag vector if it is included in the set of lags considered approaches one with increasing sample size. This

consistency result may look surprising since the linear FPE is known to be inconsistent. However, in the present case the rate of the penalty term in (38) depends on the number of lags m. Thus, if one includes l lags in addition to m^* correct ones, the rate of the penalty term becomes slower which implies that too large models are ruled out asymptotically. Note that this feature is intrinsic to the local estimation approach since the number of lags influence the rate of convergence, see (22). We remark that the consistency result breaks down if Assumption (A7) is violated e.g. if the stochastic process is linear. In this case overfitting (including superfluous lags in addition to the correct ones) is more likely. The breakdown of consistency can be avoided if one uses the Nadaraya-Watson instead of the local linear estimator since the former is also biased in case of linear processes.

Furthermore, Tschernig and Yang (2000) show that asymptotically it is more likely to overfit than to underfit (miss some correct lags). In order to reduce overfitting and therefore increase correct fitting, they suggest to correct the AFPE and estimate the Corrected Asymptotic FPE

$$CAFPE_a = AFPE_a \left\{ 1 + m(T - i_m)^{-4/(m+4)} \right\}, \quad a = 1, 2. \tag{39}$$

The correction does not affect consistency under the stated assumptions while additional lags are punished more heavily in finite samples. One chooses the lag vector with the smallest $CAFPE_a$, $a = 1, 2$.

We note that if one allows the maximal lag M to grow with sample size, then one has a doubled nonparametric problem of nonparametric function estimation and nonparametric lag selection.

The nonparametric lag selection criterion $CAFPE_2$ can be computed using the quantlet `tp/cafpe/cafpe`. The quantlet `tp/cafpe/cafpefull` also allows to use $AFPE_a$. Both are part of the third party quantlib `tp/cafpe/cafpe` which contains various quantlets for lag and bandwidth selection for nonlinear autoregressive models (16). The quantlet `tp/cafpe/cafpe` is called as

```
{crmin, crpro} = cafpe(y, truedat, xdataln, xdatadif,
                       xdatastand, lagmax, searchmethod, dmax)
```

with the input variables:

`y`
: $T \times 1$ matrix of the observed time series or set to zero if `truedat` is used,

`truedat`
character variable that contains path and name of ascii data file if `y=0`,

`xdataln`
character variable where `"yes"` takes natural logs, `"no"` doesn't,

`xdatadif`
character variable where the value `"yes"` takes first differences of data, `"no"` doesn't,

`xdatastand`
character variable where `"yes"` standardizes data, `"no"` doesn't,

`lagmax`
scalar variable, largest lag to be considered,

`searchmethod`
character variable where `"full"` considers all possible lag combinations, `"directed"` does directed search (recommended if `lagmax` > 10),

`dmax`
scalar variable with maximum number of possible lags,

and output variables

`crmin`
(`dmax`+1)×1 vector that stores for all considered lag combinations in the first `dmax` columns the selected lag vector, in the `dmax`+1 column the estimated $CAFPE_2$, in the `dmax`+2 column $\widehat{A}$, in the `dmax`+3 column the bias corrected estimate of A, see TY (equation 3.3),

`crpro`
(`dmax`+1)×(`dmax`+6) matrix that stores for each number of lags $(0, 1, \ldots, \texttt{dmax})$ in the first `dmax` columns the selected lag vector, in the `dmax`+1 column the plug-in bandwidth $\widehat{h}_{2,opt}$ for estimating the final prediction error for the true function A and $CAFPE_2$, in the `dmax`+2 column the bandwidth $\widehat{h}_B$ for estimating the constant B which is used for computing $CAFPE_2$ and the plug-in bandwidth $\widehat{h}_{2,opt}$, in the `dmax`+3 column the bandwidth $\widehat{h}_C$ for estimating the constant C which is used for computing the plug-in bandwidth $\widehat{h}_{2,opt}$, in the `dmax`+4 column the estimated $CAFPE_2$, in the `dmax`+5 column $\widehat{A}$, in the `dmax`+6 column the bias corrected estimate of A, see TY (equation 3.3).

Some comments may be appropriate. The weight function $w(\cdot)$ is the indicator function on the range of the observed data. If M is large or the time series is long, then conducting a full search over all possible lag combinations may take extraordinarily long. In this case, one should use the directed search suggested by Tjøstheim and Auestad (1994): lags are added as long as they reduce the selection criterion and one adds that lag from the remaining ones which delivers the largest reduction.

For computing $CAFPE_2$ TY follow Tjøstheim and Auestad (1994) and implement two additional features for robustification. For estimating $\mu(x,h)$ the kernel estimator

$$\hat{\mu}(x,h) = (T - i_m + i_1)^{-1} \sum_{i=i_m+1}^{T+i_1} K_h(X_i - x) \tag{40}$$

is used where the vectors X_i, $i = T+1,\ldots,T+i_1$ are all available from the observations Y_t, $t = 1,\ldots,T$. For example, X_{T+i_1} is given by $(Y_T,\ldots,Y_{T+i_1-i_m})^T$. This robustification is switched off if the sum stops at T. Furthermore, 5% of those observations whose density values $\hat{\mu}(\cdot)$ are the lowest, are screened off. These features can be easily switched off or modified in the quantlet `tp/cafpe/cafpefull`. This quantlet also allows to select the lags of the conditional standard deviation $\sigma(\cdot)$ and is therefore discussed in detail in Subsection 2.4.

If one is only interested in computing the plug-in bandwidth $\hat{h}_{2,opt}$, then one can directly use the quantlet `tp/cafpe/hoptest`. However, before it can be called it requires to prepare the time series accordingly so that it is easier to run the lag selection which automatically delivers the plug-in bandwidth for the chosen lag vector as well. For the definition of its variables the reader is referred to the helpfile of `tp/cafpe/hoptest`.

We are now ready to run the quantlet `tp/cafpe/cafpe` on the lynx data set. The following quantlet conducts a full search among the first six lags

```
pathcafpe       = "tp/cafpe/" ; path of CAFPE quantlets
;       load required quantlibs
library("xplore")
library("times")
func(pathcafpe + "cafpeload") ; load  XploRe files of CAFPE
cafpeload(pathcafpe)

setenv("outheadline","")   ; no header for each output file
```

```
  setenv("outlineno","")          ; no numbering of output lines
;       set parameters
  truedat          = "lynx.dat"  ; name of data file
  y                = 0
  xdataln          = "yes";      ; take logarithms
  xdatadif         = "no";       ; don't take first differences
  xdatastand       = "no";       ; don't standardize data
  lagmax           = 6         ; the largest lag considered is 6
  searchmethod     = "full"    ; consider all possible lag comb.
  dmax             = 6           ; consider at most 6 lags
;       conduct lag selection
  { crmin,crpro } = cafpe(y,truedat,xdataln,xdatadif,xdatastand,
                          lagmax,searchmethod,dmax)
  "selected lag vector,                 estimated CAFPE "
  crmin[,1:dmax+1]
  "number of lags, chosen lag vector, estimated CAFPE,
                                                plug-in bandwidth"
  (0:dmax)~crpro[,1:dmax|(dmax+4)|(dmax+1)]
```

flts09.xpl

A screenshot of the output which shows the criteria for all other number of lags is contained in Figure 11. The selected lags are 1 to 4 with plug-in bandwidth $\hat{h}_{2,opt} = 0.90975$ and $CAFPE_2 = 0.2163$. However, the largest decrease in $CAFPE_2$ occurs if one allows for two lags instead of one and lag 2 is added. In this case, $CAFPE_2$ drops from 0.64125 to 0.24936. Therefore lag 2 seems to capture the autocorrelation in the residuals of the NAR(1) model which was estimated in Subsections 1.1 to 1.3. For this reason a NAR(2) model could be sufficient for the lynx data. Its graphical representation is discussed in the next section.

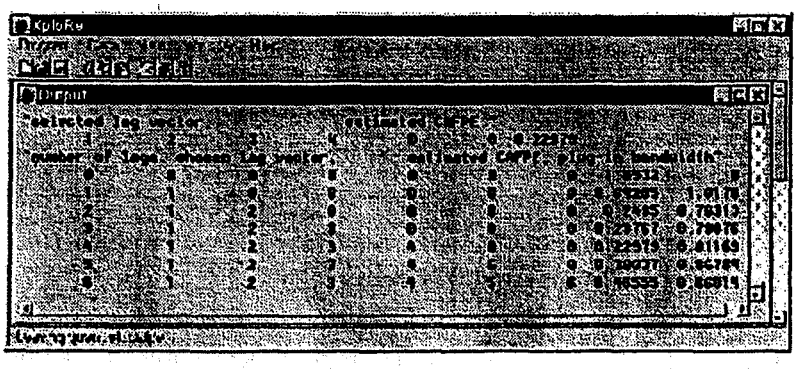


Figure 11: Results of the lag selection procedure using $CAFPE_2$ for lynx data

2.3 Plotting and Diagnostics

```
{hplugin, hB, hC, xs, resid} = plotloclin(xdata, xresid,
          xdataln, xdatadif, xdatastand, volat, lags, h,
                    xsconst, gridnum, gridmax, gridmin)
```

computes 1- or 2-dimensional plot of regression function of a nonlinear autoregressive process for a given lag vector on the range of the data; if more than 2 lags are used, then only two lags are allowed to vary, the others have to be fixed

Once the relevant lags and an appropriate bandwidth are determined, one would like to have a closer look at the implied conditional mean function as well as checking the residuals for potential model misspecification as discussed in Subsection 1.3. The latter may be done by inspecting the autocorrelation function and testing the normality of the residuals. The quantlet `tp/cafpe/plotloclin` of the quantlib `tp/cafpe/cafpe` allows to do both. It generates two- or three-dimensional plots of the autoregression function on a grid that covers the range of data and computes the residuals for the given time series. Both is done either with a bandwidth specified by the user or the plug-in bandwidth $\widehat{h}_{2,opt}$ which is automatically computed if required. The quantlet `tp/capfe/plotloclin` also allows to compute three-dimensional plots of functions with more than two lags by keeping $m-2$ lags fixed at user-selected values. It is called by

```
{hplugin,hB,hC,xs,resid} = plotloclin(xdata,xresid,xdataln,
                    xdatadif,xdatastand,volat,lags,h,xsconst,
                    gridnum,gridmax,gridmin)
```

with the input variables

`xdata`
: $T \times 1$ vector of the observed time series

`xresid`
: $T' \times 1$ vector of residuals or observations for plotting conditional volatility function, if not needed set `xresid` = 0.

`xdataln`
: character variable, `"yes"` takes natural logs, `"no"` doesn't,

`xdatadif`
: character variable, `"yes"` takes first differences of data, `"no"` doesn't,

`xdatastand`
: character variable, `"yes"` standardizes data, `"no"` doesn't,

`volat`
: character variable, `"no"` plots conditional mean function, `"resid"` plots conditional volatility function, the residuals of fitting a conditional mean function have to be contained in `xresid`,

`lags`
: $m \times 1$ vector of lags,

`h`
: scalar bandwidth for which if set to zero a scalar plug-in bandwidth using `hoptest` is computed or a $m \times 1$ vector bandwidth

`xsconst`
: $m \times 1$ vector (only needed if $m > 2$) indicates which lags vary and which are kept fixed for those keeping fixed, the entry in the correponding row contains the value at which it is fixed for those to be varied, the entry in the corresponding row is 1e-100,

`gridnum`
: scalar, number of grid points in one direction,

`gridmax`
scalar, maximum of grid,

`gridmin`
scalar, minimum of grid,

and output variables

`hplugin`
scalar plug-in bandwidth $\widehat{h}_{2,opt}$ (37) or chosen scalar or vector bandwidth,

`hB`
scalar, rule-of-thumb bandwidth (33) for nonparametrically estimating the constant B in $CAFPE_2$ and for computing the plug-in bandwidth,

`hC`
scalar, rule-of-thumb bandwidth (36) for nonparametrically estimating the constant C for computing the plug-in bandwidth,

`xs`
$T' \times m$ matrix with lagged values of time series which are used to compute plug-in bandwidth and residuals for potential diagnostics,

`resid`
$T' \times 1$ vector with residuals after fitting a local linear regression at `xs`.

Figure 12 shows the plot of the conditional mean function for an NAR(2) model of the lynx data on a grid covering all observations. The autocorrelation function of the residuals is shown in Figure 13. These graphs and a plot of the standardized residuals are computed with the following quantlet. It also returns the Jarcque-Bera test statistic of 2.31 with p-value of 0.32.

```
  pathcafpe   = "tp/cafpe/"    ; path of CAFPE quantlets
;       load required quantlibs
  library("xplore")
  library("times")
  func("jarber")
  func(pathcafpe + "cafpeload"); load XploRe files of CAFPE
  cafpeload(pathcafpe)

  setenv("outheadline","")     ; no header for each output file
```

```
  setenv("outlineno","")      ; no numbering of output lines
;       set parameters
  lynx        = read("lynx.dat");
  xresid      = 0
  xdataln     = "yes";        ; take logarithms
  xdatadif    = "no";         ; don't take first differences
  xdatastand  = "no";         ; don't standardize data
  lags        = 1|2     ; lag vector for regression function
  h           = 0
  xsconst     = 1e-100|1e-100 ; 1e-100 for the lags which are
                              ; varied for those kept fixed it
                              ; includes the chosen constant
  gridnum     = 30            ; number of gridpoints in one dir.
  gridmax     = 9             ; maximum of grid
  gridmin     = 4             ; minimum of grid
; compute opt. bandwidth and plot regression fct. for given lags
  { hplugin,hB,hC,xs,resid } = plotloclin(lynx,xresid,xdataln,
                               xdatadif,xdatastand,volat,lags,h,
                               xsconst,gridnum,gridmax,gridmin)
  "plug-in bandwidth" hplugin
;       diagnostics
  acfplot(resid) ; compute and plot acf of residuals
  {jb,probjb,sk,k} = jarber(resid,1)
        ; compute Jarque-Bera test for normality of residuals
```

flts10.xpl

From inspecting Figure 13 one can conclude that a NAR(2) model captures most of the linear correlation structure. However, the autocorrelation at lags 3 and 4 is close to the boundaries of the confidence intervals of white noise and explains why the CAFPE procedure suggests lags one to four. The regression surface in Figure 12 nicely shows the nonlinearity in the conditional mean function which may be difficult to capture with standard parametric nonlinear models.

2.4 Estimation of the Conditional Volatility

So far we have considered the estimation and lag selection for the conditional mean function $f(x)$. Finally, we turn our attention to modelling the function of the conditional standard deviation $\sigma(x)$. The conditional standard deviation

plays an important role in financial modelling, e.g. for computing option prices. As an example we consider 300 logged observations dmus58-300 of a 20 minutes spaced sample of the Deutschemark/US-Dollar exchange rate. Figures 14 and 15 display the logged observations and its first differences. The figures are generated with the quantlet

```
library("plot")
library("times")
setsize(640,480)
fx        = read("dmus58-300.dat"); read data
d1        = createdisplay(1,1)
x1        = #(1:300)~fx
setmaskl (x1, (1:rows(x1))', 0, 1)
show(d1,1,1,x1)           ; plot data
setgopt(d1,1,1,"title",
                  "20 min. spaced sample of DM/US-Dollar rate")
setgopt(d1,1,1,"xlabel","Periods","ylabel","levels")

d2        = createdisplay(1,1)
x2        = #(2:300)~tdiff(fx)
setmaskl (x2, (1:rows(x2))', 0, 1)
show(d2,1,1,x2)           ; plot data
setgopt(d2,1,1,"title","20 min. spaced sample of
                        DM/US-Dollar rate - first differences")
setgopt(d2,1,1,"xlabel","Periods","ylabel","first differences")
```

fltsl1.xpl

In the following we assume that the conditional mean function $f(\cdot)$ is known and subtracted from Y_t. Thus, we obtain $\tilde{Y}_t = Y_t - f(X_t)$. After squaring (16) and rearranging we have

$$\tilde{Y}_t^2 = \sigma^2(X_t) + \sigma^2(X_t)(\xi_t^2 - 1). \tag{41}$$

Since $\sigma^2(X_t)(\xi_t^2 - 1)$ has expectation zero, the stochastic process (41) can be modelled with the methods described in Subsections 2.1 and 2.2 by simply replacing the dependent variable Y_t by its squares. However, we have to remark that the existence of the expectation $E\left[\left(\tilde{Y}_t^2 - \sigma^2(X_t)\right)^2\right]$ is a necessary condition for applying $CAFPE_2$. Otherwise, the FPE cannot be finite. We

note that if $f(x)$ has to be estimated, the asymptotic properties of $CAFPE_2$ are expected to remain the same. Therefore, it may be used in practice, however, after replacing $\tilde{Y}_t$ by the residuals $Y_t - \hat{f}_2(X_t)$. This is possible with the quantlet `tp/capfe/cafpefull` which extends the functionality of the quantlet `tp/cafpe/cafpe` and allows the user to change additional tuning parameters. The quantlet `tp/cafpe/cafpefull` is called by

```
{crmin,crpro,crstore,crstoreadd,hstore,hstoretest} =
cafpefull(y,truedat,xresid,trueres,xdataln,xdatadif,xdatastand,
          lagmax,volat,searchmethod,dmax,selcrit,robden,perA,
          perB,startval,noutputf,outpath)
```

and has input variables

`y`
: $T \times 1$ vector of univariate time series,

`truedat`
: character variable that contains path and name of ascii data file if y=0,

`xresid`
: $T' \times 1$ vector of residuals or observations for selecting lags of conditional volatility function, if not needed set `xresid` = 0,

`trueres`
: character variable, `"yes"` takes natural logs, `"no"` doesn't,

`xdatadif`
: character variable, `"yes"` takes first differences of data, `"no"` doesn't,

`xdatastand`
: character variable, `"yes"` standardizes data, `"no"` doesn't,

`lagmax`
: scalar, largest lag to be considered,

`volat`
: character variable, `"no"` conducts lag selection for conditional mean function, `"resid"` conducts lag selection for conditional volatility function, the residuals of fitting a conditional mean function have to be contained in `xresid` or a file name has to be given in `trueres`,

searchmethod
character variable for determining search method, **"full"** conducts full search over all possible input variable combinations, **"directed"** does directed search,

dmax
scalar, maximal number of lags

selcrit
character variable to select lag selection criterion, **"lqafpe"** estimates the asymptotic Final Prediction Error $AFPE_2$ (38) using local linear estimation and the plug-in bandwidth $\widehat{h}_{2,opt}$ (37), **"lqcafpe"** estimates the corrected asymptotic Final Prediction Error $CAFPE_2$ (39) using local linear estimation and the plug-in bandwidth $\widehat{h}_{2,opt}$ (37)

robden
character variable, **"yes"** and **"no"** switch on and off robustification in density estimation (40),

perA
scalar, parameter used for screening off a fraction of $0 \leq$ **perA** ≤ 1 observations with the lowest density in computing $\widehat{A}_2$

perB
scalar, parameter like **perA** but for screening off a fraction of **perB** observations with lowest density in computing $\widehat{B}_2$,

startval
character variable to control treatment of starting values, **"different"** uses for each lag vector as few starting values as necessary, **"same"** uses for each lag vector the same starting value which is determined by the largest lag used in the lag selection quantlet **tp/cafpe/xorigxe**,

noutputf
character variable, name of output file,

outpath
character variable, path for output file.

The output variables are

`crmin`
: vector that stores for all considered lag combinations in the first `dmax` rows the selected lag vector, in the `dmax+1` row the estimated criterion, in the `dmax+2` row $\hat{A}_2$, in the `dmax+3` row the bias corrected estimate of A,

`crpro`
: matrix that stores for each number of lags in the first `dmax` rows the selected lag vector, in the `dmax+1` row the plug-in bandwidth $\hat{h}_{2,opt}$ for estimating A and $(C)AFPE$, in the `dmax+2` row the bandwidth $\hat{h}_B$ used for estimating B, in the `dmax+3` row the bandwidth $\hat{h}_C$ for estimating C, in the `dmax+4` row the estimated criterion $AFPE_2$ or $CAFPE_2$, in the `dmax+5` row $\hat{A}_2$, in the `dmax+6` row the bias corrected estimate of A,

`crstore`
: matrix that stores lag vector and criterion value for all lag combinations and bandwidth values considered, in the first `dmax` rows all considered lag vector are stored, in the `dmax+1` rows the estimated criterion for each lag vector is stored,

`crstoreadd`
: matrix that stores those criteria that are evaluated in passing for all lag combinations where all values for one lag combination are stored in one column (see program for details),

`hstore`
: row vector that stores the bandwidths used in computing (C)AFPE for each lag vector

`hstoretest`
: matrix that stores for each lag vector in one column the plug-in bandwidth $\hat{h}_{2,opt}$, $\hat{h}_B$ and $\hat{h}_C$.

The quantlet `fllts12.xpl` (for brevity not shown) conducts a lag selection for the conditional mean function $f(x)$ and finds lag 1 and 3 with bandwidth $\hat{h}_{2,opt} = 0.000432$. If you run the quantlet, you will obtain the XploRe warning "quantlet fvllc: inversion in local linear estimator did not work because probably the bandwidth is too small". This means that for one of the checked combinations of lags, one of the rule-of-thumb bandwidths or the plug-in bandwidth was too small so that the matrix $Z_2^T W Z_2$ in the local linear estimator

(21) is near singular and the matrix inversion failed. In this case, the relevant bandwidth is doubled (at most 30 times) until the near singularity disappears. Therefore, lag selection for the conditional volatility function $\sigma(x)$ is done with replacing the observations Y_t in model (41) by the estimated residuals $Y_t - \hat{f}(X_t)$. The computations are carried out with the following quantlet which also generates a plot of the conditional mean function on the range $[-0.0015, 0.0015]$ displayed in Figure 16 and plots the autocorrelation function of the residuals (not shown). The latter plot does not show significant autocorrelation.

```
  pathcafpe     = "tp/cafpe/"   ; path of CAFPE quantlets
;   load required quantlibs
  library("xplore")
  library("times")
  func("jarber")
  func(pathcafpe + "cafpeload") ;load XploRe files of CAFPE
  cafpeload(pathcafpe)
;   set output format
  setenv("outheadline","")  ; no header for each output file
  setenv("outlineno","")    ; no numbering of output lines
;   load data
  x             = read("dmus58-300.dat")  ; name of data file
  y             = tdiff(x)  ; compute first differences
  xresid        = 0
  truedat       = ""        ; name of potential data file
  trueres       = ""        ; name of potential residuals file
  xdataln       = "no"      ; don't take logarithms
  xdatadif      = "no"      ; don't take first differences
  xdatastand    = "no"      ; don't standardize data
  lagmax        = 6         ; the largest lag considered is 6
  searchmethod  = "full"    ; consider all possible lag comb.
  dmax          = 6         ; consider at most 6 lags
  volat         = "no"      ; plot cond. mean function
  selcrit       = "lqcafpe" ; use CAFPE with plug-in bandwidth
  robden        = "yes"     ; robustify density estimation
  perA          = 0
  perB          = 0.05      ; screen off data with lowest density
  startval      = "different"
  noutputf      = ""        ; name of output file
  outpath       = "test"    ; path for output file
```

```
  lags          = 1|3       ; lag vector for regression function
  h             = 0
  xsconst       = 1e-100|1e-100 ; 1e-100 for the lags which are
                            ; varied for those kept fixed it
                            ; includes the chosen constant
  gridnum       = 30      ; number of gridpoints in one direction
  gridmax       = 0.0015    ; maximum of grid
  gridmin       = -0.0015   ; minimum of grid
; compute optimal bandwidth and plot cond. mean for given lags
 { hplugin,hB,hC,xs,resid } = plotloclin(y,xresid,xdataln,
              xdatadif,xdatastand,volat,lags,h,xsconst,gridnum,
                                                  gridmax,gridmin)
  "plug-in bandwidth for conditional mean" hplugin
;   diagnostics
  acfplot(resid); compute and plot acf of residuals
  {jb,probjb,sk,k} = jarber(resid,1)
          ; compute Jarque-Bera test for normality of residuals
;   conduct lag selection for cond. standard deviation
  xresid        = resid
  volat         = "resid" ; conduct lat selection for cond. vol.
  {crmin,crpro,crstore,crstoreadd,hstore,hstoretest}
                = cafpefull(y,truedat,xresid,trueres,xdataln,
                            xdatadif,xdatastand,lagmax,volat,
                            searchmethod,dmax,selcrit,robden,
                            perA,perB,startval,noutputf,outpath)
  "Lag selection for cond. standard deviation using residuals"
  "selected lag vector,              estimated CAFPE "
  crmin[,1:dmax+1]
  "number of lags, chosen lag vector,  estimated CAFPE,
                                              plug-in bandwidth"
  (0:dmax)~crpro[,1:dmax|(dmax+4)|(dmax+1)]
```

flts13.xpl

For the conditional standard deviation one obtains lags 2 and 6 with bandwidth $\widehat{h}_{2,opt} = 0.000456$. Figures 17, 18 and 19 display the plot of the estimated conditional standard deviation $\widehat{\sigma}_2(x)$, of the standardized residuals of the modified model (41) and of their autocorrelation. The plots are generated with the following quantlet

```
  pathcafpe = "tp/cafpe/"   ; path of CAFPE quantlets

;   load required quantlets
  library("xplore")
  library("times")
  func("jarber")
  func(pathcafpe + "cafpeload"); load XploRe files of CAFPE
  cafpeload(pathcafpe)

  setenv("outheadline","")  ; no header for each output file
  setenv("outlineno","")    ; no numbering of output lines

;   set parameters
  x          = read("dmus58-300.dat");
  y          = tdiff(x)
  xresid     = 0
  xdataln    = "no"       ; don't take logarithms
  xdatadif   = "no"       ; don't take first differences
  xdatastand= "no"        ; don't standardize data
  volat      = "no"       ; compute cond. standard deviation
  lags       = 1|3        ; lag vector for regression function
  h          = 0          ; compute plug-in bandwidths
  xsconst    = 1e-100|1e-100
                          ; 1e-100 for the lags which are varied
                          ; for those kept fixed it includes the
                          ; chosen constant
  gridnum    = 30         ; number of gridpoints in one direction
  gridmax    = 0.0015     ; maximum of grid
  gridmin    = -0.0015    ; minimum of grid

; compute optimal bandwidth and plot cond. mean for given lags
  { hplugin,hB,hC,xs,resid } = plotloclin(y,xresid,xdataln,
           xdatadif,xdatastand,volat,lags,h,xsconst,gridnum,
                                          gridmax,gridmin)
  "plug-in bandwidth for mean" hplugin

; compute plug-in bandwidth and
; plot cond. standard deviation for given lags
  lags       = 2|6        ; lags for cond. volatility
  xresid     = resid
```

```
volat     = "resid"
gridmax   = 0.0008      ; maximum of grid
gridmin   = -0.0008     ; minimum of grid

{ hplugin,hB,hC,xs,resid } = plotloclin(y,xresid,xdataln,
          xdatadif,xdatastand,volat,lags,h,xsconst,gridnum,
                                              gridmax,gridmin)
"plug-in bandwidth for conditional volatility" hplugin

;   diagnostics
acfplot(resid); compute and plot acf of residuals
{jb,probjb,sk,k} = jarber(resid,1)
      ; compute Jarque-Bera test for normality of residuals
```

flts14.xpl

The surface plot of the conditional standard deviation is computed on the range $[-0.0008, 0.0008]$ in order to avoid boundary effects. Inspecting the range of the standardized residuals in Figure 18 indicates that the analysis may be strongly influenced by outliers which also may explain the extreme increase of the conditional standard deviation in Figure 17 in one corner. Moreover, Figure 19 shows some significant autocorrelation in the residuals. One explanation for this finding could be the presence of long memory in the squared observations. This topic is treated in detail in Chapter ??. Therefore, one should continue to improve the current function estimates by excluding extreme observations and using models that allow for many lags in the function of the conditional standard deviation such as, for example, Yang, Härdle and Nielsen (1999).

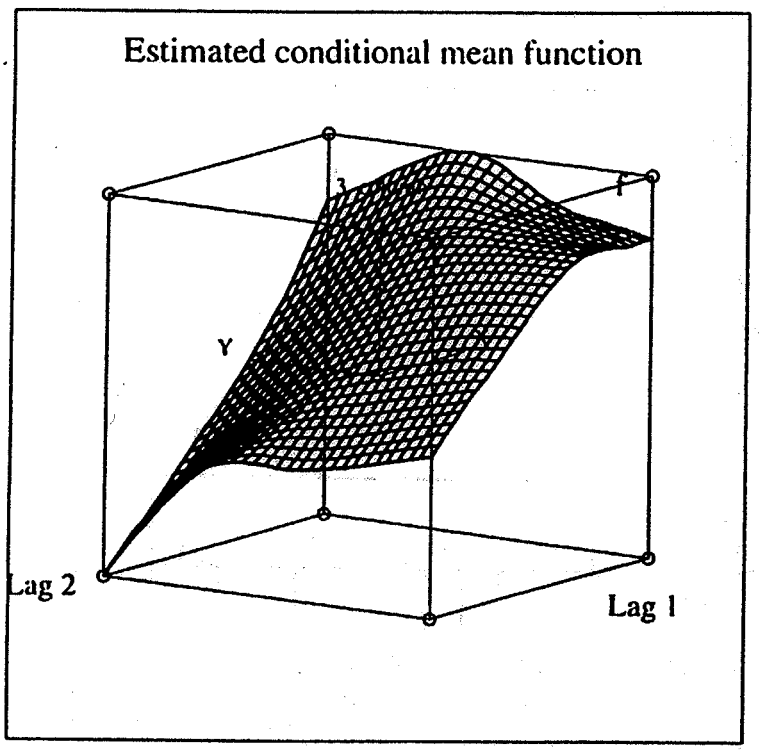


Figure 12: Plot of the conditional mean function of a NAR(2) model for the logged lynx data

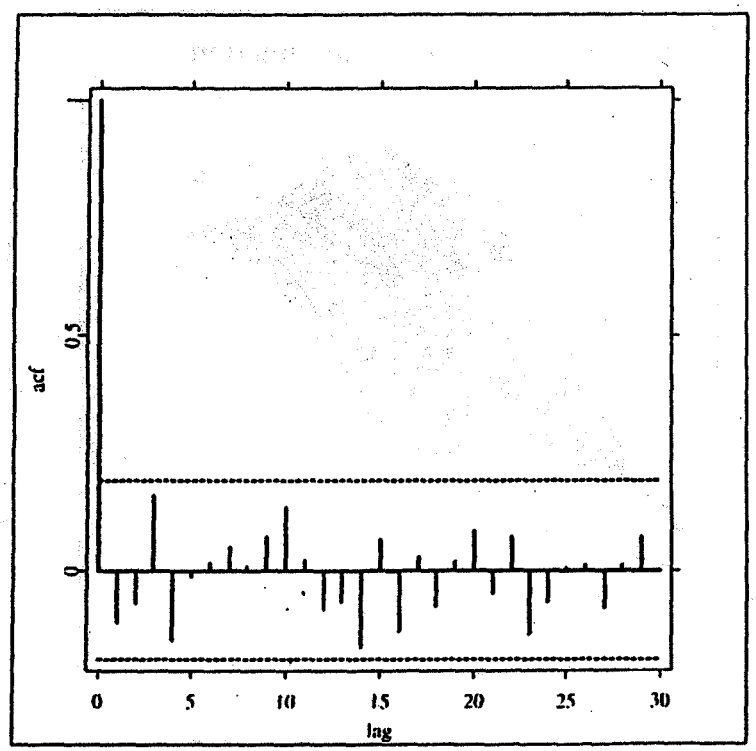


Figure 13: Plot of the autocorrelation function of the residuals of a NAR(2) model for the logged lynx data

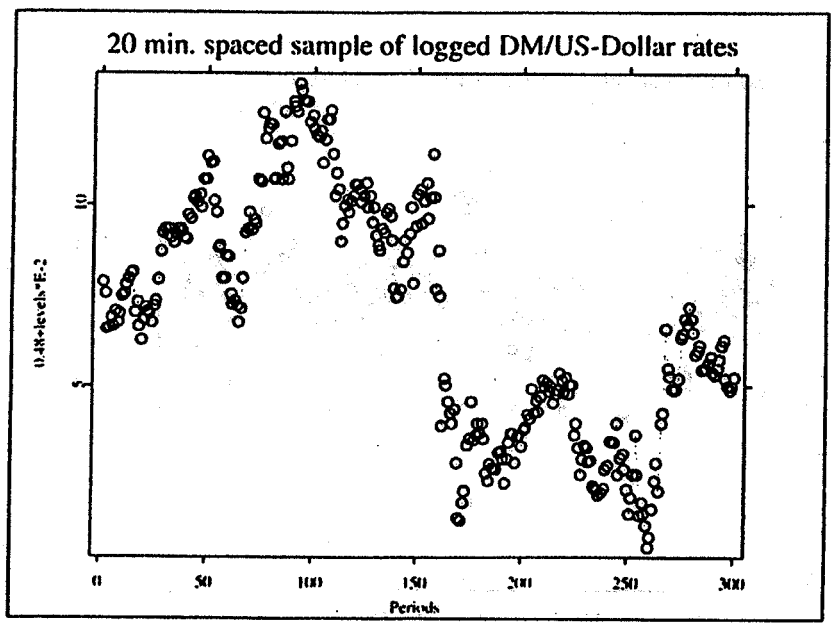


Figure 14: Time series of logarithm of 20 minutes spaced sample of DM/US-Dollar rate

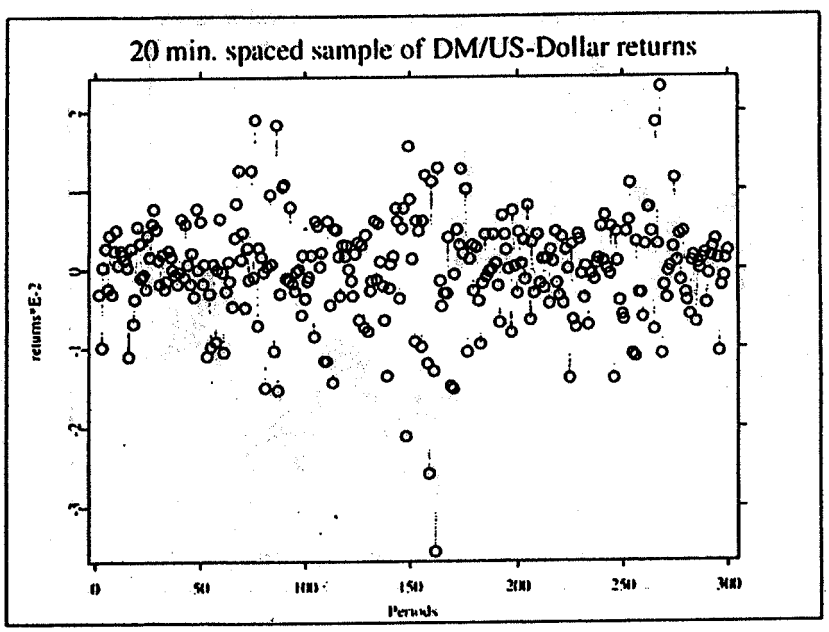


Figure 15: Time series of 20 minutes spaced sample of exchange rate returns

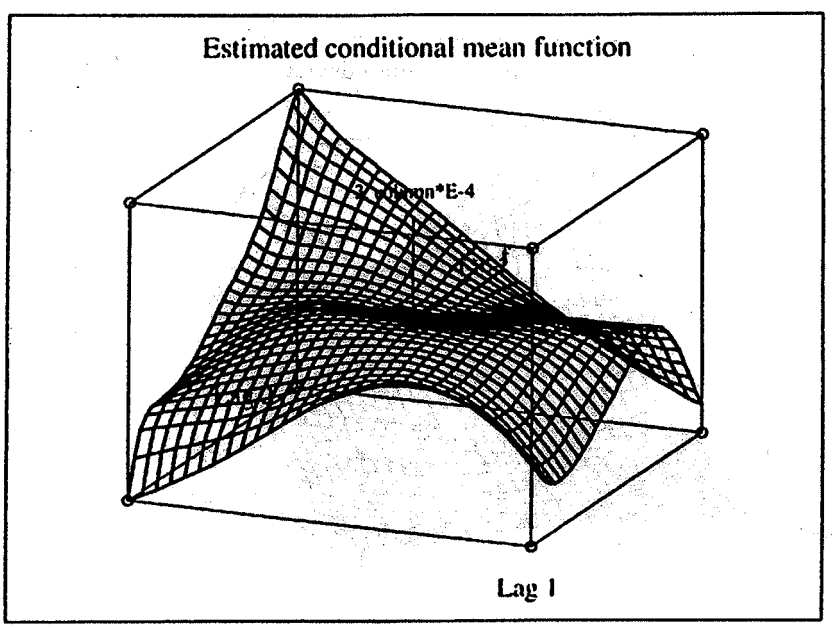


Figure 16: Plot of the conditional mean function of a NAR model with lags 1 and 3 for the returns of the Deutschemark/US-Dollar exchange rate

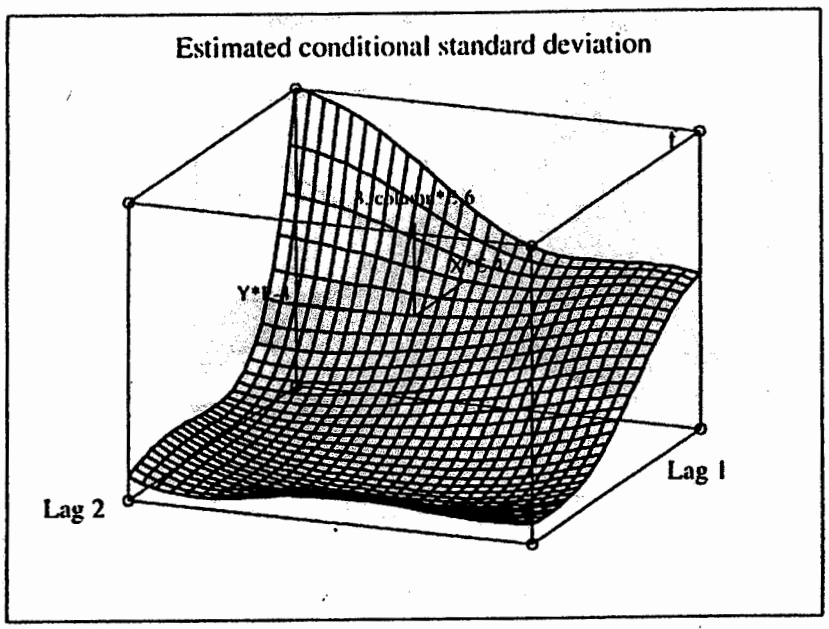


Figure 17: Plot of the conditional standard deviation of a NAR model with lags 2 and 6 for the returns of the Deutschemark/US-Dollar exchange rate

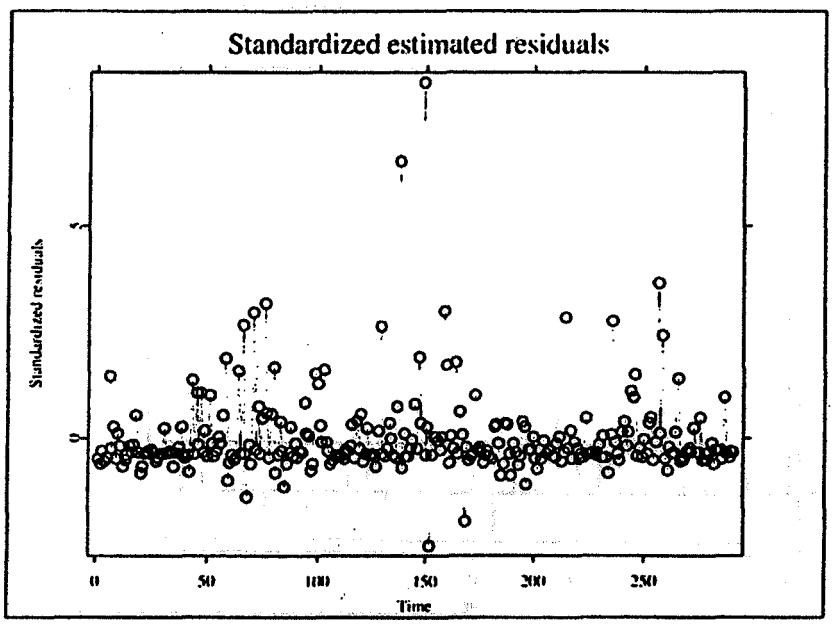


Figure 18: Plot of the standardized residuals of the modified model (41)

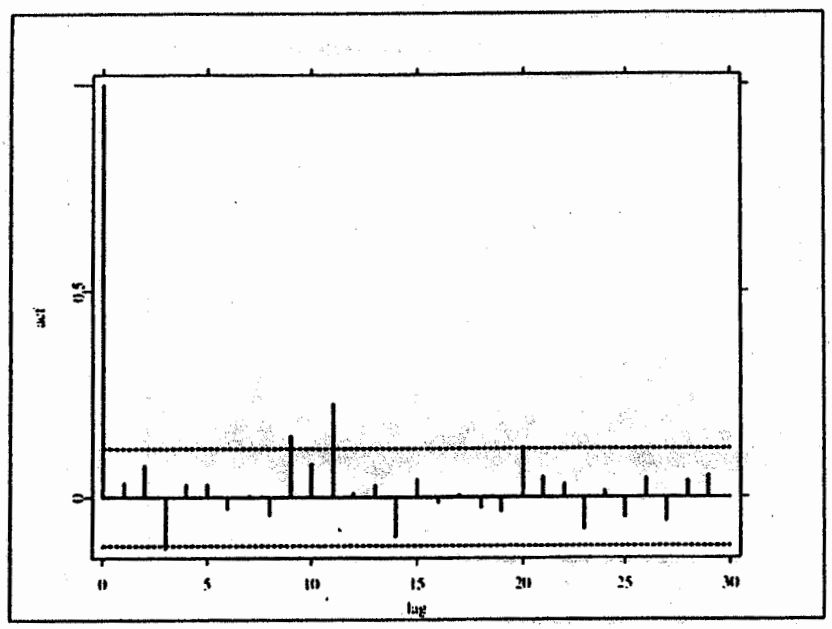


Figure 19: Plot of the autocorrelation function of residuals of the modified model (41)

References

Auestad, B. and Tjøstheim, D. (1990). 'Identification of nonlinear time series: first order characterization and order determination', *Biometrika* **77**: 669–687.

Bera, A.K. and Jarque, C.M. (1982). Model Specification Tests: a Simultaneous Approach, *Journal of Econometrics*, **20**: 59–82.

Billingsley, P. (1968). *Convergence of Probability Measures*, New York: Wiley.

Brockwell, P.J. and Davis, R.A. (1991). *Time Series: Theory and Methods*, Springer, New York.

Doukhan, P. (1994). *Mixing. Properties and Examples*, Springer-Verlag, New York et al.

Fan, J. and Gijbels, I. (1995). 'Data-driven bandwidth selection in local polynomial fitting: Variable bandwidth and spatial adaption', *Journal of the Royal Statistical Society B* **57**(2): 371–394.

Fan and Marron (1994). 'Fast implementations of nonparametric curve estimators', *Journal of Computation and Graphical Statistics* **3**, 35 – 56.

Franke, J., Kreiss, J.-P., Mammen, E., and Neumann, M. H. (1998). 'Properties of the nonparametric autoregressive bootstrap', Discussion Paper 54/98, SFB 373, Humboldt University, Berlin.

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

Härdle, W. and Tsybakov, A. (1997). 'Local polynomial estimators of the volatility function in nonparametric autoregression', *Journal of Econometrics* **81**, 223–242.

Härdle, W. and Vieu, P. (1992). ' Kernel regression smoothing of time series', *Journal of Time Series Analysis* **13**: 209–232.

Härdle, W., Klinke, S., and Müller, M. (2000), *XploRe — The Statistical Computing Environment*, Springer, New York.

Härdle, W., Lütkepohl, H., and Chen, R. (1997). 'A review of nonparametric time series analysis', *International Statistical Review* **65**(1): 49–72.

Lu, Z. (1998). 'On the geometric ergodicity of a non-linear autoregressive model with an autoregressive conditional heteroscedastic term', *Statistica Sinica* **8**: 1205–1217.

Robinson, P. M. (1983). 'Non-parametric estimation for time series models', *Journal of Time Series Analysis* **4**: 185–208.

Silverman, B. (1986). *Density estimation for Statistics and Data Analysis*, Chapman and Hall, London.

Tjøstheim, D. (1994). 'Nonlinear time-series - a selective review', *Scandinavian Journal of Statistics* **21**(2): 97–130.

Tjøstheim, D. and Auestad, B. (1994). 'Nonparametric identification of nonlinear time-series - selecting significant lags', *Journal of the American Statistical Association* **428**: 1410–1419.

Tschernig, R. and Yang, L. (2000). 'Nonparametric lag selection for time series', *Journal of Time Series Analysis* , forthcoming.

Vieu, P. (1994). 'Order choice in nonlinear autoregressive models', *Statistics* **24**: 1–22.

Yang, L., Härdle, W. and Nielsen, J.P. (1999). 'Nonparametric autoregression with multiplicative volatility and additive mean', *Journal of Time Series Analysis* **20**: 579–604.

Yang, L. and Tschernig, R. (1999). 'Multivariate bandwidth selection for local linear regression', *Journal of the Royal Statistical Society, Series B*, **61**, 793–815.

Yao, Q. and Tong, H. (1994). 'On subset selection in non-parametric stochastic regression', *Statistica Sinica* **4**: 51–70.

51 W.Härdle, R.Tschernig: *Flexible Time Series Analysis*

50 S.Sperlich, J.Zelinka: *Generalized Additive Models*

49 H.Mucha, H.Sofyan: *Cluster Analysis*

48 W.Härdle, W.Kim, G.Tripathi: *Nonparametric Estimation of Additive Models with Homogeneous Components*

47 E.Neuwirth: *Spreadsheets as Tools for Statistical Computing and Statistics Education*

46 B.Rönz, M.Müller, U.Ziegenhagen: *The Multimedia Project MM Stat for Teaching Statistics*

45 S.Huck, W.Müller: *Absent-minded drivers in the lab: Testing Gilboa's model*

44 M.Chinn, G.Meredith: *Interest Parity at Short and Long Horizons*

43 M.Chinn: *The Empirical Determinants of the Euro: Short and Long Run Perspectives*

42 A.Desdoigts: *Neoclassical Convergence Versus Technological Catch-Up: A Contribution for Reaching a Consensus*

41 J.Amendinger, D.Becherer, M.Schweizer: *Quantifying the Value of Initial Investment Information*

40 Y.Lengwiler, E.Wolfstetter: *Auctions and Corruption*

39 S.Huck, W.Müller, H.Normann: *Strategic Delegation in Experimental Markets*

38 D.Kübler: *On the Regulation of Social Norms*

37 R.Brüggemann, H.Lütkepohl: *Lag Selection in Subset VAR Models with an Application to a U.S. Monetary System*

36 M.Dufwenberg, U.Gneezy, W.Güth, E.van Damme: *An Experimental Test of Direct and Indirect Reciprocity in Case of Complete and Incomplete Information*

35 C.Müller, E.Hahn: *Money Demand in Europe: Evidence from the Past*

34 R.Krutchenko, A.Melnikov: *Quantile hedging for a jump-diffusion financial market model*

33 H.Karlsen, T.Myklebust, D.Tjostheim: *Nonparametric Estimation in a Nonlinear Cointegration Type Model*

DISCUSSION PAPERS are available via anonymous ftp on amadeus.wiwi.hu-berlin.de (141.20.100.2)
and on WWW (world wide web) URL: http://sfb.wiwi.hu-berlin.de
in the subdirectory pub/papers/sfb373

Computational Statistics, 12, 153-172

Computer–assisted Semiparametric Generalized Linear Models

Marlene Müller, Bernd Rönz, Wolfgang Härdle

Institut für Statistik und Ökonometrie, Humboldt-Universität zu Berlin, Spandauer Str. 1, D–10178 Berlin, Germany

Summary

The paper gives an overview on generalized linear models and its application in different branches of science. We introduce semiparametric extensions of the generalized linear model.
One particular model of interest is the generalized partially linear model which allows a nonparametric modelling of the influence of the continuous covariables. The estimation procedure is introduced and a test on the correct specification of this model (vs. a parametric generalized linear model) is presented. The application to a data set on East–West German migration illustrates the use of this technique.
Semiparametric methods are hightly demanding on software. Flexibility for extensions, tools for efficient computation on the user level as well as interactive graphics to display the resulting curves/surfaces are such requirements. We thus complete our presentation by indicating the practical implementation in new version of the statistical computing environment XploRe.

Keywords: generalized linear models, generalized partially linear models, semiparametric modelling, statistical software

(1997) Müller, M., Rönz, B., Härdle, W.
Computer assisted Semiparametric Generalized Linear Models.

Computational Statistics, 12, 153-172

1 Introduction

Lasting for decades the statistical analysis of the dependency of a response variable Y on a vector of covariables or explanatory variables $z = (z_1, \ldots, z_d)^T$ was dominated by the classical linear normal model $Y = Z^T\theta + \varepsilon$ with the assumptions

- the linear predictor $\eta = z^T\theta$ with the parameter vector $\theta = (\theta_1, \ldots, \theta_d)^T$ equals the conditional expectation $E(Y|z) = \mu$ of the continuous response Y, i.e. $\eta = \mu$;
- the error terms ε are independent and identically $N(0, \sigma^2)$ distributed and thus the responses Y have a $N(\mu, \sigma^2)$ distribution.

The wide usage of such linear models is obvious in the sense, that the underlying statistical methods for estimation of the unknown parameters are theoretically well investigated and understood, a variety of diagnostic tools have been developed for models of this type and the results are easy to interpret.

Although extensive research has been done to relax the stringent assumptions of the normal linear model the major impetus to a more flexible statistical modelling came in the 1970's when Nelder & Wedderburn (1972) introduced the concept of *generalized linear models* (GLM). The generalization concerns two aspects:

⋆ it is still asumed that the responses Y are independent and identically distributed, however, not necessarily normal but with a distribution from the exponential family with conditional expectation $E(Y|z) = \mu$;

⋆ the structural form of the model is extended in the sense that the linear predictor η is equal to some function of the conditional expectation μ of Y, i.e.

$$\eta = H(\mu) = z^T\theta,$$

where H, called the link function, is a known monotone, differentiable function. Or equivalently, the conditional expectation μ of Y is not directly related to the explanatory variables $\mu = z^T\theta$ but via a monotone, differentiable response function

$$\mu = G(\eta) = G(z^T\theta)$$

with G being the inverse of H.

With the introduction of generalized linear models the application of linear models was considerably extended to many practical data situations, especially in economic and social sciences where normally distributed response variables are in fact hardly found. The advantages of generalized linear models are due to the fact that a wide range of responses measured on nominal

(1997) Müller, M., Rönz, B., Härdle, W.
Computer assisted Semiparametric Generalized Linear Models.

Computational Statistics, 12, 153-172

and ordinal scales can be handled within this methodology. For example, in many economic, sociological, psychological, medical and biological applications the response variable is binary, the two possible outcomes generally labelled "success" and "failure", for references see the monographs Collett (1991), Cramer (1991), Fahrmeir & Hamerle (1984), Hosmer & Lemeshow (1989), Kleinbaum (1994).

Usually information on several other variables related to the response are available and summarized in the vector Z. The number of successes in n independent "trials" under the same conditions $S = \sum_i Y_i$ is binomial distributed with $E(S|z) = n\pi$. In estimating the effects of the covariables on the probability π one has to ensure that π is in the intervall $[0, 1]$. Thus, π is related to the linear predictor by a monotonous cumulative distribution function F, i.e. $\pi = G(\eta) = F(\eta)$. Choosing $F(\bullet)$ as the logistic distribution function results in the logit model with the response function

$$\pi = G(\eta) = \frac{1}{1 + \exp(-\eta)}$$

and the link function

$$\eta = H(\pi) = \log\left(\frac{\pi}{1-\pi}\right).$$

The probit model assumes that F is the standard normal cumulative distribution function, i.e. $G(\eta) = \Phi(\eta)$. Here the link function is given by the inverse standard normal distribution function. If $H(\pi) = \log\{-\log(1-\pi)\}$ is choosen as link function, the corresponding response function ist the extreme-minimal-value distribution function $G(\eta) = 1 - \exp\{-\exp(\eta)\}$ and a complementary log-log model ist fitted to the data.

In other research situations, such as demographic, consumer, market or other surveys, respondents are faced with several choices in answering questions. As a consequence, the response variable is categorical implying either nominal or ordinal scales with more than two categories. Again, additional information is available which characterizes the individuals and/or the responses.

Assume these variables are also measured on nominal or ordinal scales or they are categorized versions of continuous variables. If n independent repetitions (e.g. from n questioned individuals) are given then the observations are counts or frequencies in the cells of a contingency table. The cell frequencies, denoted by S_j, $j = 1, \ldots, r$ with r as the number of cells in the contingency table, are now considered as response variables and are multinomially distributed with expectation $E(S_j|z_j) = n\pi_j$, π_j being the cell probabilities. Interest focuses on estimating the main effects and interaction effects of the cross-classified variables on the cell probabilities. This leads to the log-linear model with link function $\eta_j = H(\pi_j) = \log(\pi_j) = z_j^T\theta$ and response function $\pi_j = G(\eta_j) = \exp(\eta_j)$ connecting $\mu_j = \pi_j$ multiplicatively with the linear predictor η_j. For details see the monographs Bishop,

(1997) Müller, M., Rönz, B., Härdle, W.
Computer assisted Semiparametric Generalized Linear Models.

Fienberg & Holland (1975), Christensen (1990), Fahrmeir & Hamerle (1984), Langenheine (1989), Santner & Duffy (1989).

In all these models the predictor η is of the linear form $\eta = z^T\theta$ but linked in different ways to the expectation of the response. Generalized linear models have found many applications and attained considerable popularity, especially in social sciences, for analyzing qualitative data, count data and continuous data, that are constrained to positive–only values, and had a major influence on statistical modelling. This was made possible with the development of computer hardware and appropriate computer software since the underlying estimation procedures involve a large amount of computation in practical analyses and can only be carried out computer–assisted.

2 Semiparametric extensions of GLM

In recent years a good deal of work has been devoted to "generalize" the generalized linear models. These extensions concern other data situations (multivariate responses, multivariate correlated responses, repeated measurements, discrete time survival data, non-normal time series, state space situations, random effects), other techniques (quasi–likelihood approaches, semi– and nonparametric approaches), and other models (nonlinear and nonexponential family models). For instance a workable alternative to the aforementioned parametric generalized linear models are *single index models* (SIM), keeping the linear form of the index $\eta = z^T\theta$ but allowing G to be an arbitrary smooth function, and *generalized additive models* (GAM) that maintain $G(\bullet)$ to be a known function but allow the argument inside G to be a sum of unknown smooth functions. For an overview see Härdle & Turlach (1992). Problems here are discrete covariables and economic prestructure of models.

One of the reasons for the wide propagation of generalized linear models is the computational feasibility (in particular for discrete covariables) and the easy access to standard computational systems (SPSS, LIMDEP). Another aspect is the good interpretability of the index $z^T\theta$ in all fields of applied statistics. Especially the study of marginal effects is an easy task for this structure of the exogeneous covariables. Any generalization should take care of these properties, see Fahrmeir & Tutz (1994), Maddala (1983). However, recent studies have questioned the strict linear structure of the index or the functional form of the link function. We refer here in particular to Horowitz (1993a), Horowitz (1993b).

In generalizing generalized linear models, we would like to post a certain caveat. A simple replacement of the index by an arbitrary nonparametric function would not be acceptible by the reasons given above: (a) computational feasibility, (b) interpretation/study of marginal effects would be hindered by a too flexible form of the nonparametric transformation. The simplest (modest) generalizations have been successfully applied for low dimensional models (le Cessie & van Houwelingen 1991, Proença & Ritter 1994).

(1997) Müller, M., Rönz, B., Härdle, W.
Computer assisted Semiparametric Generalized Linear Models.

Computational Statistics, 12, 153-172

The technique of integration for generalized additive modelling (Linton & Härdle 1996) or average derivative estimation for single index models (Powell, Stock & Stoker 1989, Härdle & Stoker 1989) extends for arbitrary dimension but fails in the analysis of partially discrete covariables. Backfitting with local scoring for generalized additive models works with discrete covariables but is unfortunately not supported by theoretical statements on statistical properties.

It is therefore of interest to consider models with the following structure

$$E(Y|x,t) = G\{x^T\beta + m(t)\}, \tag{1}$$

where $\beta = (\beta_1, \ldots, \beta_p)^T$ is a finite dimensional parameter and $m(\bullet)$ is a smooth function. Here we assume a decomposition of the explanatory variables z into two vectors, x and t. In the following we refer to this model also as a *generalized partially linear model* (GPLM), see also Severini & Staniswalis (1994), Hunsberger (1994). Here x denotes a realization from a p–variate random vector X which usually covers discrete covariables. t results from a q–variate random vector T of continuous covariables.

The estimation of model (1) is computationally feasible by the idea that an estimate $\widehat{\beta}$ can be found for known m, and an estimate $\widehat{m}$ can be found for known β. We formulate the procedure in terms of quasi–likelihood estimation. However, note that if the distribution of Y belongs to an exponential family, using the quasi–likelihood function is the same as using the log–likelihood function. The quasi–likelihood function is defined as

$$Q(\mu; y) = \int_{\mu}^{y} \frac{(s-y)}{V(s)}\, ds$$

where μ is the (conditional) expectation of Y, i.e. $\mu = G\{x^T\beta + m(t)\}$. It is assumed here that the conditional variance of Y is $\sigma^2 V(\mu)$ where σ is an unknown scale parameter and $V(\bullet)$ is a known function.

Estimators for β and $m(\bullet)$ have been proposed by Severini & Wong (1992), Severini & Staniswalis (1994) and Carroll, Fan, Gijbels & Wand (1995). We follow the approach of Severini & Wong (1992) and Severini & Staniswalis (1994) which use two different likelihood functions for the estimation of the parametric and semiparametric components. The usual likelihood for n i.i.d. observations (x_i, t_i, y_i)

$$\mathcal{L}(\beta) = \sum_{i=1}^{n} Q\left\{\beta^T x_i + m_\beta(t_i); y_i\right\} \tag{2}$$

is used to obtain $\widehat{\beta}$ and a "smoothed" likelihood

$$\mathcal{L}^S(\eta) = \sum_{i=1}^{n} K_h(t - t_i)\, Q\left(\beta^T x_i + \eta; y_i\right) \tag{3}$$

(1997) Müller, M., Rönz, B., Härdle, W.
Computer assisted Semiparametric Generalized Linear Models.

for the nonparametric smooth function $\widehat{m}_\beta(t) = \eta$ at point t. We give a more detailed description of the algorithm in the appendix. Note that m is estimated as a function of the parametric component β which yields an asymptotically efficient estimate $\widehat{\beta}$ (Severini & Wong 1992). The computational alogrithm consists in searching maxima of both likelihoods simultaneously. A detailed description of the algorithm can be found in the Appendix. It turns out that the resulting estimator $\widehat{\beta}$ is $\sqrt{n}$–consistent and asymptotically normal, and that estimators $\widehat{m} = \widehat{m}_{\widehat{\beta}}$ are consistent in supremum norm, see Severini & Staniswalis (1994). As in generalized linear models a possible scale parameter σ can be estimated by

$$\widehat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}(Y_i - \widehat{\mu}_i)^2/V(\widehat{\mu}_i), \tag{4}$$

where $\widehat{\mu}_i = G\{x_i^T\widehat{\beta} + \widehat{m}(t_i)\}$.

For higher dimensions in t the possible nonlinearities in (1) cannot anymore be graphically displayed and face the above mentioned problems (interpretability). An additive structured partially linear index may be considered, see Härdle, Huet, Mammen & Sperlich (1996). However, additivity of $m(\bullet)$ does not need to hold. For simplicity of the presentation we concentrate for the technical ideas in estimation and testing on model (1). In particular, t will be one–dimensional in our following example and thus the generalized additive and the generalized partially linear model coincide here.

Having estimated the influence $m(\bullet)$ of the covariables T, it is naturally to ask, whether the estimate $\widehat{m}$ is significantly different from a linear function obtained by a parametric GLM fit. The test procedure in Härdle, Mammen & Müller (1996) for

$$\begin{aligned} H_0 &: m(t) = t^T\gamma, \\ H_1 &: m(\bullet) \text{ is an arbitrary smooth function,} \end{aligned}$$

is based on a comparison of the semiparametric estimates with the estimators $(\widetilde{\beta}, \widetilde{\gamma})$ in the parametric model

$$(\widetilde{\beta}, \widetilde{\gamma}) = \arg\min_{\beta,\gamma}\sum_{i=1}^{n} Q\left(x_i^T\beta + t_i^T\gamma; y_i\right). \tag{5}$$

A direct comparison of $\widehat{m}(t)$ and $t^T\widetilde{\gamma}$ can be misleading because $\widehat{m}$ has a non–negligible smoothing bias. This holds even under the linearity hypothesis. Hence, in analogy to Härdle & Mammen (1993), a bias–corrected parametric estimate $\widetilde{m}$ is used instead of $t^T\widetilde{\gamma}$. This estimate can be obtained from the following smoothing step:

$$\widetilde{m} = \arg\min_{m}\int\sum_{i=1}^{n} K_h(t - t_i)\, Q[G\{x_i^T\widetilde{\beta} + m(t)\}; G\{x_i^T\widetilde{\beta} + t_i^T\widetilde{\gamma}\}]\, dt. \tag{6}$$

(1997) Müller, M., Rönz, B., Härdle, W.
Computer assisted Semiparametric Generalized Linear Models.

Computational Statistics, 12, 153-172

Note that, in (6) the second argument of $Q(\bullet;\bullet)$ is the parametric estimate of $E(Y_i|x_i,t_i)$ instead of Y_i which means to apply the smoothing step according to (3) to the artificial data set $\{G(x_i^T\widetilde{\beta}+t_i^T\widetilde{\gamma}),x_i,t_i\}, i=1,\ldots,n$.

Using this bias-corrected $\widetilde{m}$, Härdle, Mammen & Müller (1996) propose the likelihood-ratio type test statistic

$$R=-2\sum_{i=1}^{n}Q(\widetilde{\mu}_i;\widehat{\mu}_i), \tag{7}$$

where $\widetilde{\mu}_i=G\{x_i^T\widetilde{\beta}+\widetilde{m}(t_i)\}$ is the bias corrected GLM fit and $\widehat{\mu}_i=G\{x_i^T\widehat{\beta}+\widehat{m}(t_i)\}$ is the semiparametric GPLM fit. Note that if the distribution of Y does not belong to an exponential family, the calculation of R involves evaluation of n integrals. In this case the following Taylor expansion of R is easier to compute:

$$\widetilde{R} \;=\; \sum_{i=1}^{n}\frac{[G'\{x_i^T\widehat{\beta}+\widehat{m}(t_i)\}]^2}{V[G\{x_i^T\widehat{\beta}+\widehat{m}(t_i)\}]}\left\{x_i^T(\widehat{\beta}-\widetilde{\beta})+\widehat{m}(t_i)-\widetilde{m}(t_i)\right\}^2. \tag{8}$$

Härdle, Mammen & Müller (1996) show that R and $\widetilde{R}$ are asympotically equivalent and have an asymptotic normal distribution

$$v_n^{-1}(R-e_n)\xrightarrow{D}N(0,1).$$

Here we denote

$$e_n=\frac{\lambda\cdot\int K(u)^2\,du}{h_1\cdot\ldots\cdot h_q},\quad v_n^2=\frac{2\lambda\int\{K\star K(u)\}^2\,du}{h_1\cdot\ldots\cdot h_q}$$

with $h=(h_1,\ldots,h_q)^T$ denoting the multivariate bandwidth vector, λ is the Lebesgue measure of the support of T and $K\star K$ is the convolution of K with itself.

The asymptotic expansion of R shows that it behaves approximately like a sum of $O(h_1^{-1}\cdot\ldots\cdot h_q^{-1})$ independent summands. This is typically not very large and indeed it turns out that the normal approximation needs not to work well for R (Härdle, Mammen & Müller 1996). Therefore, for the calculation of quantiles, it is recommended to use the the following bootstrap procedure:

1. Generate samples $\{Y_1^*,\ldots,Y_n^*\}$ with $E^*(Y_i^*)=G(x_i^T\widetilde{\beta}+t_i^T\widetilde{\gamma})$ and $\mathrm{Var}^*(Y_i^*)=\widehat{\sigma}^2V\{G(x_i^T\widetilde{\beta}+t_i^T\widetilde{\gamma})\}$. Here E^* and Var^* denote the conditional expectation or variance given $(x_1,t_1,\ldots,x_n,t_n)$.

2. Calculate estimates $\widehat{\beta}^*,\widehat{m}^*,\widetilde{\beta}^*,\widetilde{\gamma}^*,\widetilde{m}^*$ based on the bootstrap samples $(x_1,t_1,Y_1^*),\ldots,(x_n,t_n,Y_n^*)$. Furthermore, calculate test the statistic R^* (or $\widetilde{R}^*$). Repeat this n^* times. The quantiles of the distribution of R (or $\widetilde{R}$) can be estimated by the quantiles of the conditional distribution of R^* (or $\widetilde{R}^*$).

(1997) Müller, M., Rönz, B., Härdle, W.
Computer assisted Semiparametric Generalized Linear Models.

In a binary response model the distribution of Y is completely specified by $\mu = G(x^T\beta + t^T\gamma)$ (under the linearity hypothesis). Here, it is reasonable to resample from the Bernoulli distribution with parameters $\widetilde{\mu}_i = G(x_i^T\widetilde{\beta} + t_i^T\widetilde{\gamma})$ (the parametric GLM fit). If the distribution of Y cannot be specified (apart from the first two moments) it is recommended to use wild bootstrap (Härdle & Mammen 1993).

3 Example: East–West German Migration

Let us illustrate the semiparametric estimation and the test procedure with an example on East–West German migration. Our interest in this subject has been inspired by the considerations of Burda(1993, 1995). We consider a sample of East Germans, which have been surveyed in 1991 in the German Socio-Economic Panel, see GSOEP (1991). Among other questions the East German participants have been asked, if they can imagine to move to the western part of Germany or West Berlin. We give the value 1 for those who responded positive and 0 if not.

		Yes	No	(in %)	
Y	**migration intention**	38.5	61.5		
X_1	**family/friends in west**	85.6	11.2		
X_2	**unemployed/job loss certain**	19.7	78.9		
X_3	**city size 10,000-100,000**	29.3	64.2		
X_4	**female**	51.1	49.8		
		Min	Max	Mean	S.D.
X_5	**age (years)**	18	65	39.84	12.61
T	**household income (DM)**	200	4000	2194.30	752.45

Table 1: Descriptive statistics for migration data. $n = 3235$.

The economic model is based on the idea that a person will migrate if its utility (wage differential) exceeds the costs of migration. Of course neither of both variables, wage differential and costs, are directly available. It is obvious that age has an important influence on migration intention. Younger people will have a higher wage differential. Currently low household income and unemployment will also increase a possible gain in wage after migration. On the other hand, friends or family members in the Western part of Germany will reduce the costs of migration. We also consider a city size variable and gender as interesting variables.

Table 2 shows in the middle column the results of a parametric logit fit. The migration intention is definitely determined by age. However, also the unemployment, city size and household income variables are highly significant.

(1997) Müller, M., Rönz, B., Härdle, W.
Computer assisted Semiparametric Generalized Linear Models.

	Coeff.	(t–value)	Coeff.	(t–value)
const.	0.512	(2.39)	–	–
family/friends in west	0.599	(5.20)	0.598	(5.14)
unemployed/job loss	0.221	(2.31)	0.230	(2.39)
city size 10-100,000	0.311	(3.77)	0.302	(3.63)
female	-0.240	(3.15)	-0.249	(3.26)
age	$-4.69 \cdot 10^{-2}$	(14.56)	$-4.74 \cdot 10^{-2}$	(14.59)
household income	$1.42 \cdot 10^{-4}$	(2.73)	–	–
	Linear (logit)		Part. Linear	

Table 2: Logit coefficients and coefficients in a generalized partially linear model for migration data (t–values in parenthesis). $n = 3235$, $h = 20\%$ for the GPLM.

A further analysis of this data set by a generalized additive model (keeping the logit link, but generalizing the influence of the age and income variables to nonparametric functions) showed that the age has a nearly perfect linear influence. Because of this relation, we use a generalized partially linear model with a logistic link function and only the influence of household income modelled as a nonparametric function.

Since the question of an optimal bandwidth selection is still open for generalized partially linear models, we have carried out the analysis for different bandwidths. Note that we give all bandwidths as percentage of the range of household income. Hence a bandwidth $h = 10\%$ means a value of 380 DM and so on. The coefficients for the parametric covariables in the GPLM are similar for all four considered bandwidths $h = 10\%, 20\%, 30\%$, and 40%. As an example we compare the estimated coefficients for $h = 20\%$ in the right column of Table 2. In Figure 1 we show the estimated curves for all four bandwidths. For comparison, the resulting fits $\widehat{m}$ (thick black lines) for the function m are shown together with the linear fits (thin black dashed lines) and the bias corrected parametric fits $\widetilde{m}$ (thin grey dashed lines). Recall that the estimate $\widetilde{m}$ was an estimate for the sum of the linear function and the bias of $\widehat{m}$, see (6). Figure 1 shows clearly that the bias increases with the bandwidth.

The nonparametric estimates $\widehat{m}$ for the different bandwidths are obviously nonlinear functions. However, it is difficult to judge the significance of the nonlinearity. In general, it cannot be excluded that the difference between the nonparametric and the linear fit may be caused by boundary and bias problems of $\widehat{m}$. Additionally, in this example the covariable "age" (included in a linear way) has a dominant influence on the migration intention.

Hence, we applied the test developed in Härdle, Mammen & Müller (1996). Table 3 shows the observed significance levels for the different choices of the

(1997) Müller, M., Rönz, B., Härdle, W.
Computer assisted Semiparametric Generalized Linear Models.

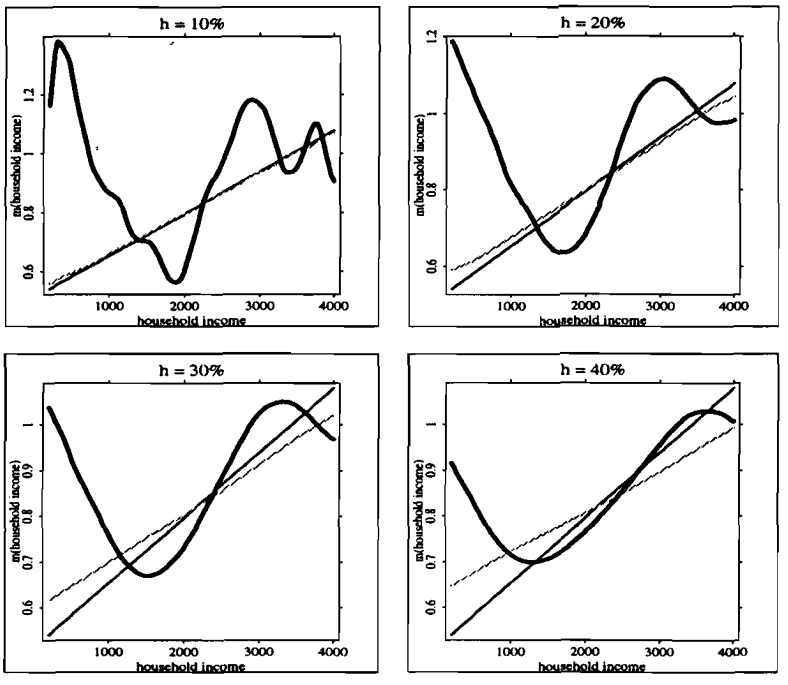


Figure 1: The influence $m(t)$ of household income on migration intention. Nonparametric fit (thick black lines), linear fit (thin black lines), and "biased" parametric estimate $\widetilde{m}$ (thin grey lines). $n = 3235$, bandwidths $h = 10\%, 20\%, 30\%, 40\%$.

bandwidth h, which have been obtained using the normal approximation of the test statistics R and $\widetilde{R}$. Linearity is clearly rejected for bandwidths 10% and 20%. The situation changes from $h = 30\%$ on.

This discrepancy is due to the bad approximation of the test statistic's distribution when h becomes large. To verify this we followed the bootstrap approach from Härdle, Mammen & Müller (1996). Figure 2 shows graphically the difference between the limiting normal distribution and the actual distribution of the bootstrapped test statistic R (estimated by kernel density estimates). The number of bootstrap samples has been chosen as $n^* = 200$ and nonlinearity turns out to be highly significant for all four bandwidths and both test statistics R and $\widetilde{R}$. We omit a table here, since all computed significance levels for rejection are below 0.01.

In consequence, we conclude that the global shape of m seems to be not well approximable by a linear function. Let us remark that the test results do not change much, when we test a GLM with quadratic influence of household income against the GPLM. The restriction on "true" linearity was just chosen to simplify the presentation. Thus, the semiparametric GPLM even

(1997) Müller, M., Rönz, B., Härdle, W.
Computer assisted Semiparametric Generalized Linear Models.

h	10%	20%	30%	40%
R	0.001	0.001	0.116	0.516
$\tilde{R}$	0.001	0.002	0.109	0.488

Table 3: Observed significance levels for linearity test for migration data, $n = 3235$.

outperforms the often modelled quadratic influence of income. However, a model which introduces a cubic influence of the income covariable fits well and is not rejected by the test anymore.

4 Implementation in XploRe 4

Generalizing the generalized linear model causes increasing complexity and thus demands for an efficient computational implementation. Speed and price have been important factors in the decision statistical software systems until a few years ago. Nowadays, the interactiveness, flexibility, extensibility, portability and an userfriendly interface are all important in statistical applications. Good reasons for interactivity are smoothing parameter selection and the ease of the (graphical) display of higher dimensional objects. Flexibility and extensibility allow to include own modifications to the predefined statistical procedures. Portability becomes increasingly important with distributedness of data and method banks on the internet, see Krishnan, Müller & Schmidt (1995).

We present here the implementaion of GLM and GPLM models in the *statistical computing environment* XploRe which is in its current version 4 fully internet capable, see Schmelzer, Klinke, Kötter & Härdle (1996). XploRe is an environment that has been designed for a large scale of statistical tasks ranging from data analysis to highly interactive operations. It combines the flexibility of multi–window desktops with standard operations and interactive user driven actions.

Let us point out that a statistical computing environment is – in contrast to a statistical system – a computing device that covers a wide range of data manipulations, problem solutions and graphical insights. This is meant not only over a wide class of statistical operations (horizontal coverage) but also over a set of user levels (vertical coverage) from first year students to graduates up to researchers. XploRe is used as a student front end tool for teaching elementary statistics as well as a research device in simulations for semiparametric analysis and bootstrapping.

The current XploRe 4 was developed on the basis of the experiences with XploRe 3 (Härdle, Klinke & Turlach 1995). It is simultaneously developed for Unix based systems and MS Windows systems. A particular advantage is the

(1997) Müller, M., Rönz, B., Härdle, W.
Computer assisted Semiparametric Generalized Linear Models.

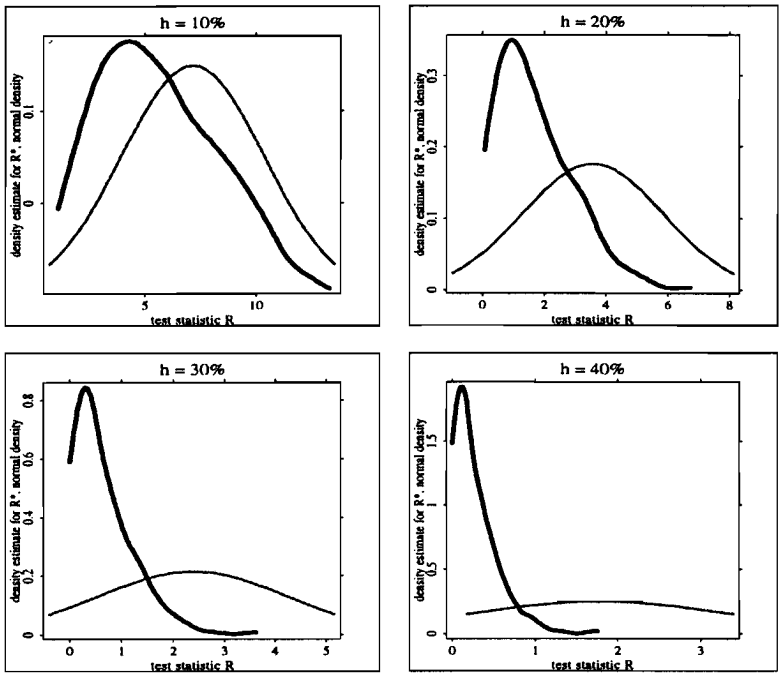


Figure 2: Density estimates of bootstrapped R (thick lines) and densities of limiting normal distribution (thin lines). $n^* = 200$ bootstrap replications, bandwidths $h = 10\%, 20\%, 30\%, 40\%$.

seamless integration of user written code into the software system. The well structured help system plays an important role here. The wide distribution, the high transparence and the free choice of HTML browsers have been the reasons for their use in the XploRe help system. User written macros can be transformed into help documents via a system internal processor. This feature is an important element of the design of XploRe, since it corresponds to the environment idea. The user customizes his interface to computational statistics by writing own macros and they become documents in the help system available to everybody.

Let us demonstrate the implementation of modules within XploRe 4 by means of the macro set for generalized linear models. This set of macros forms a *library* of routines completely written in the XploRe language. The library is named `glm` and contains macros like `glmbilo` for logit models (binomial distribution with logistic link) or `glmnoid` for the linear regression model (normal distribution with identity link). This `glm` library has been adopted in large parts from the `glm` library of the XploRe 3 version.

Figure 3 shows the header of the macro `glmbilo` and Figure 4 shows the

(1997) Müller, M., Rönz, B., Härdle, W.
Computer assisted Semiparametric Generalized Linear Models.

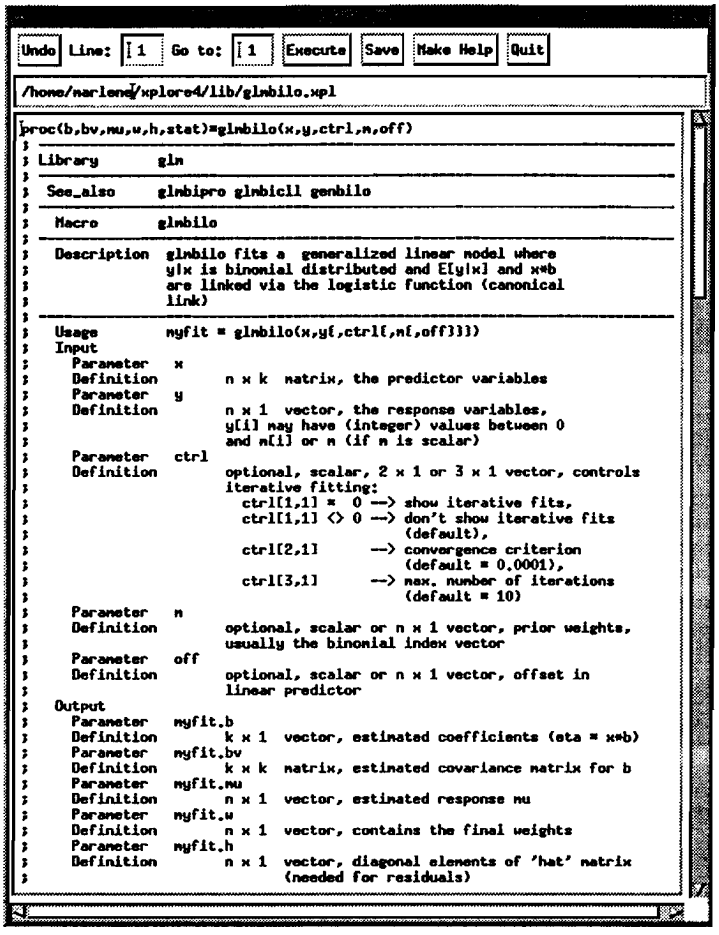


Figure 3: First lines of `glmbilo` macro

automatically created HTML help page within the XploRe 4 help system. The logit procedure `glmbilo` requires at least two input parameters, namely the design matrix `x` and the response vector `y`. Some further parameters can be given to control for algorithmic and run–time properties. The output consists of a bunch of objects. However, since XploRe supports lists, it is not necessary to write down all output objects when calling a procedure. The (minimal) XploRe code to compute the logit estimate for the migration model thus reduces to:

```
z=read("migall")              ; read data file migall.dat
x=matrix(rows(z))~z[,2:7]     ; column of ones, covariables
y=z[,1]                       ; responses
library("glm")                ; load GLM library
lf=glmbilo(x,y)               ; logit fit
```


(1997) Müller, M., Rönz, B., Härdle, W.
Computer assisted Semiparametric Generalized Linear Models.

The list `lf` contains all output of the `glmbilo` macro. For instance, the parameter estimates are in `lf.b` and the estimated covariance matrix is `lf.bv`. Some statistics like the log–likelihood, the deviance and the pseudo–R^2 are available from `lf.stat`.

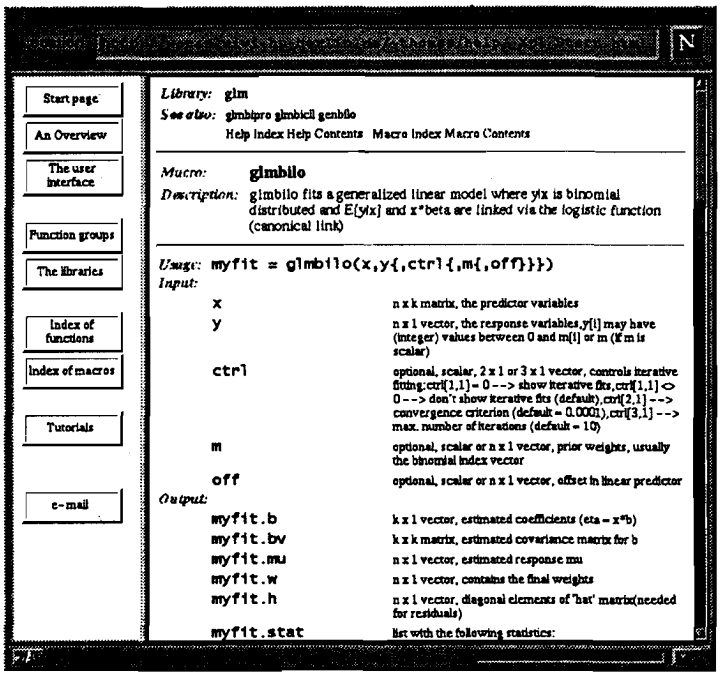


Figure 4: HTML help page of `glmbilo` macro

Generalized partially linear models in XploRe will be available in the library `gplm`. Since the development of this library is not yet finished, the following is still to a certain degree "work in progress". In the appendix, we give the algorithm for the GPLM and the test statistics in the case of a binary response. For other distributions of the responses, this algorithm can be easily adapted, see also Severini & Staniswalis (1994).

The algorithm for GPLM requires first an initialization step, this is naturally be done by a parametric GLM fit with the same link function. Next, the smoothing step for the nonparametric function $m(\bullet)$ has to be carried out. The updating step for $\eta_j(\beta) = m_\beta(t_j)$ requires a ratio with numerator and denominator of convolution type

$$\sum_{i=1}^{n} \ell_{i,j} K_h(t_i - t_j), \tag{9}$$

where $\ell_{i,j}$ is a derivative of the log–likelihood. Note, that this has to be done at least for all t_j $(j = 1, \ldots, n)$ since the updated values of m at all observation

(1997) Müller, M., Rönz, B., Härdle, W.
Computer assisted Semiparametric Generalized Linear Models.

points are required in the updating step for β. The evaluation of (9) is a standard procedure if $\ell_{i,j}$ is only dependent on i and t is one–dimensional (e.g. in Nadaraya–Watson kernel regression).

Let us start our excursion into the implementation of the GPLM by explaining a genuine XploRe implementation. For multi–dimensional t three–dimensional arrays can be used to avoid looping. XploRe 4 allows arrays and hence our genuine XploRe code to compute an updated function $\eta_j(\beta)$ would look like the following:

```
t=reshape("t",#(n,1,q))                    ; reshape t
h=reshape("h",#(1,1,q))                    ; reshape bandwidth vector h
w=prod(quartic((t-t').'/h),3)              ; matrix of kernel weights
tmp = exp(x*bold+etaold')
lll = y-tmp./(1+tmp)                       ; log-likelihood l'
ll2 = -tmp./(1+tmp)^2                      ; log-likelihood l''
etanew = (sum(lll.*w)./sum(ll2.*w))'
```

Note that we first store the columns of `t` in the third dimension of the array. The same has to be done for the bandwidth vector `h`. The Quartic kernel $K(u) = \frac{15}{16}(1-u^2)^2 \cdot I(|u| \leq 1)$ is used in product kernel form. `lll` and `ll2` compute the 1st and 2nd derivatives of the log–likelihood (logistic link), respectively. The resulting estimate is stored in `m`. Analogously, arrays could be used in the update of the design points $\widetilde{x}_j$. Here, a similar code as this above has to applied for each column of the design matrix. Finally, the updated parameter vector β is computed by a linear regression type procedure using the updated design matrix. We omit this code here.

This genuine XploRe code works quite well for small sample sizes n. As one can easily see, the evaluation of (9) requires $O(n^2)$ operations. The update $\widetilde{x}_j$ of the design points requires additional $O(p \cdot n^2)$ operations, where p denotes the dimension of the parametric covariables. So, obviously the use of arrays of this size is impossible for large data set as the migration data ($n = 3235$). Rewriting the code for the use with do–loops will increase computation time drastically, since XploRe code is interpreted. This is a particular disadvantage for the bootstrap test.

As a consequence, the GPLM is implemented in a hybrid fashion. To estimate a logit GPLM, the user calls the macro `gplmbilo` written in XploRe. (The naming convention is the same as for GLM, `gplmbilo` estimates a GPLM with binomial distribution and logistic link.) This macro itself calls two compiled functions `gplmbiloeta` and `gplmbiloxtilde` which perform the update of $\eta_j(\beta)$ and $\widetilde{x}_j$ in an efficient way. Both function are written in C and available from a shared library, which is dynamically linked to XploRe at runtime when the library `gplm` is loaded. The speed of operations in such compiled functions is comparable to that of XploRe internal commands. In contrast to internal commands, however, experienced users can modify the supplied C source code or add their own extensions. This allows the required flexibility and extensibility for the implementation of semiparametric extensions to the generalized linear model.

(1997) Müller, M., Rönz, B., Härdle, W.
Computer assisted Semiparametric Generalized Linear Models.

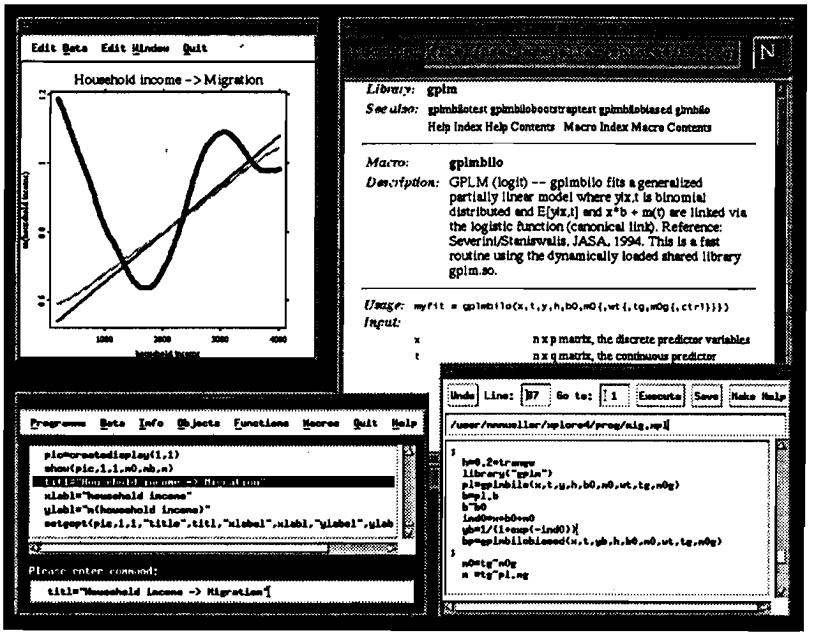


Figure 5: XploRe Session with GPLM estimation

Figure 5 shows a screen shot from the interactive GPLM fit of the migration data. In the upper left, the graphics display is shown, which yielded one part of Figure 1.

Acknowledgment

We thank Sigbert Klinke and Thomas Kötter from the XploRe programmers team for their engaged support in programming the external functions and the efforts to enable dynamical loading in XploRe.

References

Bishop, Y. M. M., Fienberg, S. E. & Holland, P. W. (1975). *Discrete Multivariate Analysis: Theory and Practice*, MIT Press, Cambridge (Mass.).

Burda, M. (1993). The determinants of East–West German migration, *European Economic Review* **37**: 452–461.

Burda, M. C. (1995). Migration and the option value of waiting, *Economic and Social Review* **27**: 1–19.

(1997) Müller, M., Rönz, B., Härdle, W.
Computer assisted Semiparametric Generalized Linear Models.

Carroll, R. J., Fan, J., Gijbels, I. & Wand, M. P. (1995). Generalized partially linear single–index models, *Discussion Paper 9506*, Institut de Statistique, Université Catholique, Louvain–La–Neuve.

Christensen, R. (1990). *Log–Linear Models*, Springer, New York.

Collett, D. (1991). *Modelling Binary Data*, Chapman and Hall, London.

Cramer, J. S. (1991). *The Logit Model*, Edward Arnold, London.

Fahrmeir, L. & Hamerle, A. (1984). *Multivariate Statistische Verfahren*, De Gruyter, Berlin.

Fahrmeir, L. & Tutz, G. (1994). *Multivariate Statistical Modelling Based on Generalized Linear Models*, Springer.

GSOEP (1991). *Das Sozio-ökonomische Panel (SOEP) im Jahre 1990/91*, Projektgruppe "Das Sozio-ökonomische Panel", Deutsches Institut für Wirtschaftsforschung. Vierteljahreshefte zur Wirtschaftsforschung, pp. 146–155.

Härdle, W., Huet, S., Mammen, E. & Sperlich, S. (1996). Semiparametric additive indices for binary response models, *Technical report*, Sonderforschungsbereich 373, Humboldt-Universität zu Berlin.

Härdle, W., Klinke, S. & Turlach, B. A. (eds) (1995). *XploRe – an interactive statistical computing environment*, Springer, New York.

Härdle, W. & Mammen, E. (1993). Testing parametric versus nonparametric regression, *Annals of Statistics* **21**: 1926–1947.

Härdle, W., Mammen, E. & Müller, M. (1996). Testing parametric versus semiparametric modelling in generalized linear models, *SFB 373 Discussion Paper 960028*, Institut für Statistik und Ökonometrie, Humboldt-Universität zu Berlin.

Härdle, W. & Stoker, T. (1989). Investigating smooth multiple regression by the method of average derivatives, *Journal of the American Statistical Association* **84**: 986–995.

Härdle, W. & Turlach, B. A. (1992). Nonparametric approaches to generalized linear models, *in* L. Fahrmeier, B. Francis, R. Gilchrist & G. Tutz (eds), *Advances in GLIM and Statistical Modelling*, Vol. 78 of *Lecture Notes in Statistics*, Springer-Verlag, New York, pp. 213–225.

Horowitz, J. L. (1993a). Semiparametric and nonparametric estimation of quantal response models, *in* G. S. Madala, C. R. Rao & H. D. Vinod (eds), *Handbook of Statistics*, Elsevier Science Publishers, pp. 45–72.

(1997) Müller, M., Rönz, B., Härdle, W.
Computer assisted Semiparametric Generalized Linear Models.

Horowitz, J. L. (1993b). Semiparametric estimation of a work-trip mode choice model, *Journal of Econometrics* **58**(1–2): 49–70.

Hosmer, D. W. & Lemeshow, S. (1989). *Applied Logistic Regression*, John Wiley & Sons, New York.

Hunsberger, S. (1994). Semiparametric regression in likelihood–based models, *Journal of the American Statistical Association* **89**: 1354–1365.

Kleinbaum, D. G. (1994). *Logistic Regression*, Springer, New York.

Krishnan, R., Müller, R. & Schmidt, P. (1995). Accessing "computable" information over the WWW: The MMM project, *SFB 373 Discussion Paper 950040*, Institut für Statistik und Ökonometrie, Humboldt-Universität zu Berlin.

Langenheine, R. (1989). *Log-Lineare Modelle zur multivariaten Analyse qualitativer Daten*, Oldenbourg, München.

le Cessie, S. & van Houwelingen, J. C. (1991). A goodness of fit test for binary regression models based on smoohting methods, *Biometrics* **47**: 1267–1282.

Linton, O. & Härdle, W. (1996). Estimation of additive regression models with known links, *Biometrika*. to appear.

Maddala, G. S. (1983). *Limited-dependent and qualitative variables in Econometrics*, Econometric Society Monographs No. 4, Cambridge University Press.

Nelder, J. A. & Wedderburn, R. W. M. (1972). Generalized linear models, *Journal of the Royal Statistical Society, Series A* **135**(3): 370–384.

Powell, J. L., Stock, J. H. & Stoker, T. M. (1989). Semiparametric estimation of index coefficients, *Econometrica* **57**(6): 1403–1430.

Proença, I. & Ritter, C. (1994). Semiparametric testing of the link function in models for binary outcomes, *SFB 373 Discussion Paper 940017*, Institut für Statistik und Ökonometrie, Humboldt-Universität zu Berlin.

Santner, T. J. & Duffy, D. E. (1989). *The Statistical Analysis of Discrete Data*, Springer, New York.

Schmelzer, S., Klinke, S., Kötter, T. & Härdle, W. (1996). A new generation of a statistical computing environment on the net, *in* A. Prat (ed.), *Proceedings of COMPSTAT Barcelona 1996*, Physica Verlag, Heidelberg, pp. 135–148.

(1997) Müller, M., Rönz, B., Härdle, W.
Computer assisted Semiparametric Generalized Linear Models.

Severini, T. A. & Staniswalis, J. G. (1994). Quasi-likelihood estimation in semiparametric models, *Journal of the American Statistical Association* **89**: 501–511.

Severini, T. A. & Wong, W. H. (1992). Generalized profile likelihood and conditionally parametric models, *Annals of Statistics* **20**: 1768–1802.

Appendix: Algorithm for GPLM

In this section we indicate how the estimates $\widehat{\beta}$, $\widehat{m}$, $\widetilde{m}$ and the test statistic can be numerically computed in a binary response model. The derivation of the algorithm for $\widehat{\beta}$, $\widehat{m}$ can be found in Härdle, Mammen & Müller (1996). The algorithm corresponds to that proposed in Severini & Staniswalis (1994) for the special case of a logistic link function. In order to avoid boundary effects, we used a weight function in the convergence criterion.

We put $\eta_j(\beta) = \widehat{m}_\beta(t_j)$ and $L_i(u) = Q\{G(u); y_i\}$. Note, that in a binary response model we have $L_i(u) = y_i \log G(u) + (1 - y_i) \log\{1 - G(u)\}$ and the derivatives $L_i'(u)$ and $L_i''(u)$ w.r.t. u can be easily determined in dependence of u and y_i. The maximization of the smoothed quasi-likelihood (3) requires to solve

$$0 = \sum_{i=1}^{n} L_i'\{x_i^T\beta + \eta_j(\beta)\}K_h(t_i - t_j). \tag{10}$$

Differentiation of (10) leads to an estimate for η_j' as a function of β

$$\eta_j'(\beta) = \frac{-\sum_{i=1}^{n} L_i''\{x_i^T\beta + \eta_j(\beta)\}K_h(t_i - t_j)x_i}{\sum_{i=1}^{n} L_i''\{x_i^T\beta + \eta_j(\beta)\}K_h(t_i - t_j)}. \tag{11}$$

For β we have to solve

$$0 = \sum_{i=1}^{n} L_i'\{x_i^T\beta + \eta_i(\beta)\}\{x_i + \eta_i'(\beta)\}. \tag{12}$$

Equations (10)–(12) suggest the following iterative Newton–Raphson type algorithm to find $\widehat{\beta}$ and $\widehat{m}(t_j) = \widehat{\eta}_j(\beta)$, $j = 1, \ldots, n$.

- *initialization*
  Start with $\widehat{\beta}^{(0)} = \widetilde{\beta}$, $\widehat{\eta}_j^{(0)} = t_j^T\widetilde{\gamma}$ from the parametric (GLM) fit.
- *updating step for $\eta_j(\beta) = m_\beta(t_j)$*
  The function $\eta_j(\beta)$ is updated by

$$\widehat{\eta}_j^{(k+1)} = \widehat{\eta}_j^{(k)} - \frac{\sum_{i=1}^{n} L_i'(x_i^T\widehat{\beta}^{(k)} + \widehat{\eta}_j^{(k)})\,K_h(t_i - t_j)}{\sum_{i=1}^{n} L_i''(x_i^T\widehat{\beta}^{(k)} + \widehat{\eta}_j^{(k)})\,K_h(t_i - t_j)}.$$

(1997) Müller, M., Rönz, B., Härdle, W.
Computer assisted Semiparametric Generalized Linear Models.

- *updating step for β*
  The parameter β is updated by

$$\widehat{\beta}^{(k+1)} = \widehat{\beta}^{(k)} - \mathcal{B}^{-1} \sum_{i=1}^{n} L_i'(x_i^T \widehat{\beta}^{(k)} + \widehat{\eta}_i^{(k)}) \widetilde{x}_i^{(k)}$$

  with a Hessian type matrix

$$\mathcal{B} = \sum_{i=1}^{n} L_i''(x_i^T \widehat{\beta}^{(k)} + \widehat{\eta}_i^{(k)})\, \widetilde{x}_i^{(k)} \widetilde{x}_i^{(k)T}$$

  and

$$\widetilde{x}_j^{(k)} = x_j - \frac{\sum_{i=1}^{n} L_i''(x_i^T \widehat{\beta}^{(k)} + \widehat{\eta}_j^{(k)})\, K_h(t_i - t_j)\, x_i}{\sum_{i=1}^{n} L_i''(x_i^T \widehat{\beta}^{(k)} + \widehat{\eta}_j^{(k)})\, K_h(t_i - t_j)}.$$

Alternatively, the functions $L_i''(u)$ can be replaced by their expectations (w.r.t. to Y) to obtain a Fisher scoring type procedure. To obtain the bias corrected parametric estimate $\widetilde{m}$, one has only to apply the updating step for $\eta_j(\beta) = m_\beta(t_j)$. Recall that instead of the observed responses y_i the fitted values $G(x_i^T \widetilde{\beta} + t_i^T \widetilde{\gamma})$ have to be used.

For the binary response y_i model the quasi–likelihood $Q\{G(u); y_i\}$ coincides with the log–likelihood, such that we used only this log–likelihood in the above algorithm. The test statistic R however does not contain binary arguments. Hence

$$\begin{aligned} R &= -2 \sum_{i=1}^{n} Q(\widetilde{\mu}_i; \widehat{\mu}_i) \\ &= -2 \sum_{i=1}^{n} \{\widehat{\mu}_i \log \widetilde{\mu}_i + (1 - \widehat{\mu}_i) \log (1 - \widetilde{\mu}_i) \\ &\qquad -\widehat{\mu}_i \log \widehat{\mu}_i - (1 - \widehat{\mu}_i) \log (1 - \widehat{\mu}_i)\}. \end{aligned}$$

is computed as a likelihood–ratio type statistic using the semiparametric fit $\widehat{\mu}_i = G\{x_i^T \widehat{\beta} + \widehat{m}(t_i)\}$ and the bias corrected parametric fit $\widetilde{\mu}_i = G\{x_i^T \widetilde{\beta} + \widetilde{m}(t_i)\}$.

(1997) Müller, M., Rönz, B., Härdle, W.
Computer assisted Semiparametric Generalized Linear Models.

International Statistical Review (1997), 65, 1, 49–72, Printed in Mexico

A Review of Nonparametric Time Series Analysis

Wolfgang Härdle[1] Helmut Lütkepohl[2] and Rong Chen[3]

[1 2]Institut für Statistik und Ökonometrie, Humboldt-Universität zu Berlin, 10178 Berlin, Germany
[3]Department of Statistics, Texas A&M University, College Station, TX 77843, USA

Summary

Various features of a given time series may be analyzed by nonparametric techniques. Generally the characteristic of interest is allowed to have a general form which is approximated increasingly precisely when the sample size goes to infinity. We review nonparametric methods of this type for estimating the spectral density, the conditional mean, higher order conditional moments or conditional densities. Moreover, density estimation with correlated data, bootstrap methods for time series and nonparametric trend analysis are described.

Key words: Kernel estimators; Smoothing techniques; Dependent observations; Bootstrap; Hermite expansions.

1 Introduction

The use of nonparametric techniques has a long tradition in time series analysis. As early as the late 19th century Schuster (1898) introduced the periodogram which may be regarded as the origin of spectral analysis. By now the latter technique is a classical nonparametric tool for analyzing time series. The increased data availability especially in finance and the explosion of computing power have made it possible to use a wide range of other modern nonparametric techniques in time series analysis recently. In this article we review some of these developments.

For a given time series $X_1, \ldots, X_n$, nonparametric techniques are used to analyze various features of interest. Generally, the idea underlying many of these techniques is that the characteristic of interest is allowed to have a general form which is approximated increasingly precisely with growing sample size. For example, if a process is assumed to be composed of periodic components, a general form of spectral density may be assumed which can be approximated with increasing precision when the sample size gets larger. Similarly, if the autocorrelation structure of a stationary process is of interest the spectral density may be estimated as a summary of the second moment properties. A brief review of this classical method of nonparametric time series analysis is given in Section 2.

Because the final objective of many time series analyses is prediction, it is often of interest to study the conditional means, conditional variances or complete conditional densities in some period, given the past of the process. When a point prediction is the final objective, an estimate of some conditional mean may be desired, while the conditional variances are needed if interval forecasts or assessments of future volatility are desired. Moreover, if higher order moments of a series are potentially important, the focus may be on estimating the complete conditional density.

In order to analyze the conditional mean nonparametrically one may, for instance, start from a

International Statistical Review, 12, 153-172

model of the form

$$X_t = f(X_{t-1}, X_{t-2}, \ldots) + \varepsilon_t \tag{1.1}$$

where ε_t is a series of innovations which is independent of past X_t. In this case $f(\cdot)$ represents the conditional expectation in period t, given past observations $X_{t-1}, X_{t-2}, \ldots$ and it is the minimum mean squared error (MSE) 1-step predictor for X_t. In parametric time series analysis the function $f(\cdot)$ is chosen from some parametric class so that the specific candidate is obtained by specifying a fixed finite number of parameters. Nonparametric approaches on the other hand allow $f(\cdot)$ to be from some flexible class of functions and they approximate $f(\cdot)$ in such a way that the approximation precision increases with the sample size. For this purpose several different techniques and procedures are available. For instance, local approaches approximate $f(\cdot)$ in the neighborhood of any given argument by letting the neighborhood decrease and thereby increase the approximation precision with growing sample size. For this purpose the number of lagged X_t used in the model is usually limited. In other words, $f(X_{t-1}, X_{t-2}, \ldots)$ is replaced by $f(X_{t-1}, \ldots, X_{t-p})$ for some fixed p. Alternatively, global approximators use parametric functions $f_n(\cdot)$, where the number of parameters and thereby the flexibility of the function may increase with the sample size n. The functions $f_n(\cdot)$ are chosen such that they approach $f(\cdot)$ in a certain norm when the sample size increases. This way it is also possible to let the number of lagged X_t's increase with the sample size n and thus avoid assuming a fixed number of lags at an early stage of the analysis. A number of methods for estimating the conditional mean function of a process are discussed in Section 3.

As mentioned earlier, in many situations point forecasting is too limited an objective and the future volatility and other higher order moments are of interest in addition to the conditional mean. Therefore the framework in (1.1) is often extended to a more general model

$$X_t = f(X_{t-1}, X_{t-2}, \ldots) + g(X_{t-1}, X_{t-2}, \ldots)\varepsilon_t \tag{1.2}$$

where $g(\cdot)$ is used to represent the conditional variance of the process in period t given the information from previous periods. Again various nonparametric approaches exist for joint estimation of $f(\cdot)$ and $g(\cdot)$. Of course, it is also possible to specify a parametric form of one of the two functions and treat the other one nonparametrically. Techniques for nonparametric analyses of model (1.2) are the subject of Section 4. More generally the complete predictive (conditional) density $h(X_t|X_{t-1}, X_{t-2}, \ldots)$ may be of interest when the shape of the conditional distribution and higher order moments are relevant to the analysis. For this case a number of different nonparametric approaches have been proposed as well. Some of them are also sketched in Section 4.

There are numerous other nonparametric procedures and techniques that have been used in time series analysis. For instance, when a parametric time series model such as (1.2) with parametric functions $f(\cdot)$ and $g(\cdot)$ is specified it may be of interest to estimate the distribution of the residuals by nonparametric methods in order to improve the parameter estimators or to assess the statistical properties of the estimators. More precisely, density estimation for the residuals and bootstrap methods based on the residuals have been used in this context. These methods are reviewed in Section 5. Another important characteristic of a time series is its trending behaviour. Deterministic trend functions have also been analyzed nonparametrically. In addition, there are a number of nonparametric tests for stochastic trends. They are also presented in Section 5.

If very general assumptions are made, a rich data set is usually necessary to obtain a good idea about the features of interest. Therefore, many of the nonparametric techniques reviewed in this article are typically used when long time series are available. Therefore, these methods have, for instance, been used for analyzing financial time series which are observed with a high frequency and are consequently relatively long. Other fields of applications include survey of riverflow, the analysis of encephalographic data and of sleep states. Although we provide a fairly broad survey of many nonparametric analysis techniques for time series we are aware that such a survey is necessarily

International Statistical Review, 12, 153-172

limited neglecting many interesting and potentially promising facets of research in this area. In particular, we are unable to give a complete listing of related publications because of the recent explosion in the literature due to the increase in data availability and computing power. We apologize for any omissions of relevant related work: Further references may be found in Györfi, Härdle, Sarda & Vieu (1989), Tjøstheim (1994) and Hart (1996).

2 Spectral Analysis

Suppose $\{X_t\}$ is a zero mean univariate stationary stochastic process with autocovariances $\gamma_k = E(X_t X_{t+k})$. Then the spectral density of $\{X_t\}$ is

$$f_X(\omega) = \frac{1}{2\pi} \sum_{k=-\infty}^{\infty} \gamma_k e^{-i\omega k} = \frac{1}{2\pi}(\gamma_0 + 2\sum_{k=1}^{\infty} \gamma_k \cos \omega k), \quad \omega \in [-\pi, \pi].$$

Here $i = +\sqrt{-1}$ as usual. Hence, the spectral density may be regarded as a weighted sum of cyclical components corresponding to frequencies ω in the interval $[-\pi, \pi]$. Since

$$\gamma_k = \int_{-\pi}^{\pi} e^{i\omega k} f_X(\omega) d\omega,$$

the second order characteristics of the process can be recovered if the spectral density is available. In particular, $\gamma_0 = \mathrm{Var}(X_t) = \int_{-\pi}^{\pi} f_X(\omega) d\omega$ and thus the spectral density represents the contributions of the frequencies to the variance of the process. Hence, the spectral density may be regarded as a summary of the cyclical components of the process or alternatively as a respresentation of the second order moments or autocovariance structure of the process.

Given a time series $X_1, \ldots, X_n$ the autocovariances of the generating process may be estimated as

$$\tilde{\gamma}_k = \frac{1}{n} \sum_{t=1}^{n-k} (X_t - \overline{X})(X_{t+k} - \overline{X}),$$

or by

$$\hat{\gamma}_k = \frac{1}{n-k} \sum_{t=1}^{n-k} (X_t - \overline{X})(X_{t+k} - \overline{X}),$$

$k = 1, \ldots, n-1$, where $\overline{X} = \sum_{t=1}^{n} X_t / n$ is the sample mean. An obvious estimator of the spectral density at frequency ω is the so called *periodogram*

$$\tilde{f}_X(\omega) = \frac{1}{2\pi} \sum_{k=-(n-1)}^{n-1} \tilde{\gamma}_k e^{-i\omega k},$$

or similarly with $\hat{\gamma}_k$ replacing $\tilde{\gamma}_k$. Unfortunately, this estimator is not consistent. The reason is that too many quantities are estimated from the sample.

To ensure consistency a smoothed estimator of the form

$$\hat{f}_X(\omega) = \frac{1}{2\pi} \sum_{k=-M}^{M} \lambda_k \tilde{\gamma}_k e^{-i\omega k}$$

is usually used. The weights $\lambda_{-M}, \ldots, \lambda_M$ represent the *spectral window* and M ($< n-1$) is the truncation point which depends on the sample size. A number of different windows has been

International Statistical Review, 12, 153-172

proposed in the literature. The following are examples:

$$\lambda_k = 1 - |k|/M \quad \text{(Bartlett, 1950)}$$

$$\lambda_k = 1 - 2a + 2a\cos\left(\frac{\pi k}{M}\right) \quad \text{(Tukey, 1949, Blackman \& Tukey, 1959)}$$

$$\lambda_k = \frac{1}{2}\left[1 + \cos\left(\frac{\pi k}{M}\right)\right] \quad \text{(Tukey, 1949)}$$

$$\lambda_k = \begin{cases} 1 - 6\left(\frac{k}{M}\right)^2 + 6\left(\frac{|k|}{M}\right)^3 & \text{for} \quad |k| \le \frac{M}{2} \\ 2\left(1 - \frac{|k|}{M}\right)^3 & \text{for} \quad \frac{M}{2} \le |k| \le M \end{cases} \quad \text{(Parzen, 1961).}$$

A number of other windows are discussed in Priestley (1981, Sec. 6.2.3.). It may be worth noting that, for frequencies $\omega_j = 2\pi j/n$, the resulting spectral density estimators may be obtained alternatively by averaging over the periodogram values of neighboring frequencies. Hence,

$$\hat{f}_X(\omega_j) = \frac{1}{2\pi}\sum_{m=-h}^{h} K(\omega_j, \omega_{j+m})\tilde{f}_X(\omega_{j+m}),$$

where $K(\cdot,\cdot)$ is a suitable kernel function and h is the bandwidth of frequencies used in the weighted average. In other words, $\hat{f}_X(\omega_j)$ may be obtained by kernel smoothing techniques which are discussed in more detail in the context of estimating the conditional mean (see Section 3.1). These ideas extend directly to the multivariate case where X_t is a vector of variables.

As mentioned in the introduction, spectral analysis of stationary processes is now a standard technique. It can be found in many time series textbooks and monographs. More recent developments in spectral analysis include nonstationary and nonlinear processes. For instance, Priestley (1981, Chapter 11) and Dahlhaus (1993) consider processes with time varying spectra. Priestley (1996) discusses the use of wavelets in this context. Nowadays spectral methods are used in various ways for analyzing time series both theoretically and empirically. Applications of these techniques include studies of seasonal behaviour of time series, approximation of the stationary part of more general processes, construction of testing and estimation procedures and examination of their properties (see, e.g., the chapters in Brillinger & Krishnaiah (1983) and in particular Robinson (1983a)). The related literature is too voluminous to be reviewed here. Hence, we regard our foregoing remarks on spectral analysis as a brief reminder that these techniques belong under the heading of this survey.

3 Estimation of the Conditional Mean

In this section we review some nonparametric methods for estimating the function $f(\cdot)$ in (1.1). We first present some smoothing approaches for locally approximating this function in the sense discussed in the introduction. For that purpose it is assumed that only a finite number of lagged X_t's enters $f(\cdot)$, that is, $f(X_{t-1}, X_{t-2}, \ldots) = f(X_{t-1}, \ldots, X_{t-p})$. Some of the methods discussed in this section impose further restrictions on $f(\cdot)$ by assuming e.g. additivity of the lags (see Section 3.2). We also consider the problem of choosing the lag length p. Moreover, in Section 3.3 global approximations are reviewed which, in principle, allow an infinite number of lags of X_t in $f(\cdot)$.

The parametric approach to estimation of the conditional mean of a time series is to formulate a parametric model for $f(\cdot)$. Many parametric structures proposed for $f(\cdot)$ have been successful in practice and have provided parsimonious models that capture the linearity or nonlinearity of the underlying process. The most common nonlinear structures are the threshold autoregressive (TAR) models of Tong (1983), the exponential autoregressive (EXPAR) models of Haggan & Ozaki (1981), the smooth-transition autoregressive (STAR) models of Chan & Tong (1986) and Granger & Teräsvirta (1993). In these models the structure for $f(\cdot)$ is supposed to be of threshold type where

International Statistical Review, 12, 153-172

the threshold functions are modeled in different ways. Many other related references can be found in Tong (1990) and Priestley (1988).

The nonparametric approach has the advantage of letting the data speak for themselves. Hence, it avoids the subjectivity of choosing a specific parametric model before looking at the data. However, there is the cost of more complicated mathematical arguments and difficulties in practical implementation, such as the selection of smoothing parameters. Also there is the cost of poor performance in high dimensions, often referred to as the 'curse of dimensionality'. Hence, the nonparametric approach often serves as a guidance for choosing appropriate lower dimensional parametric models and for deciding between competing classes of models. Powerful computers and easy-to-use interactive statistical and graphical softwares such as S (Becker, Chamber & Wilks, 1988) and XploRe (Härdle, Klinke & Turlach, 1995) provide solid platforms for these operations.

3.1 Unrestricted Local Smoothing Methods

Model (1.1) has the format of a nonlinear regression problem for which many smoothing methods exist when the observations are independent. Hart (1996) demonstrates that these methods can be 'borrowed' for time series analysis where observations are correlated by making use of the 'whitening by windowing principle'. This principle is introduced first. Then we list some common nonparametric smoothing methods for inference on the function $f(\cdot)$ in model (1.1).

The Whitening by Windowing Principle

Given an independent random sample $X_1, \ldots, X_n$, which is drawn from a distribution with density function $p(x)$, a popular method of estimating $p(x)$ is based on the kernel estimator

$$\hat{p}_h(x) = \frac{1}{nh} \sum_{i=1}^{n} K\left(\frac{x - X_i}{h}\right), \tag{3.1}$$

where $h > 0$ is the so-called bandwidth and $K(\cdot)$ is a kernel function, typically with finite support. The bandwidth is taken as a sequence $h = h_n$ tending to zero as $n \to \infty$. Note that, if the kernel function has support on $[-1, 1]$, the estimator only uses the observations in the interval $[x - h, x + h]$. This is an important feature when we extend this method to dependent observations. When the estimator is applied to dependent observations, it is affected only by the dependency of the observations in a small window, not that of the whole data set. Hence, if the dependency between the observations is of 'short memory' which makes the observations in small windows *almost independent*, then most of the techniques developed for independent observations apply in this situation. Hart (1996) calls this feature *the whitening by windowing principle.*

Various *mixing conditions* are the main tools for proving asymptotic properties of the smoothing techniques for dependent data. Basically these conditions try to control the dependence between X_i and X_j as the time distance $i - j$ increases. For example, a sequence is said to be α-mixing (strong mixing) (Robinson 1983b) if

$$\sup_{A \in \mathcal{F}_1^n, B \in \mathcal{F}_{n+k}^{\infty}} |P(A \cap B) - P(A)P(B)| \leq \alpha_k$$

where $\alpha_k \to 0$ and $\mathcal{F}_i^j$ is the σ-field generated by $X_i, \ldots, X_j$. A stronger condition is the ϕ-mixing (uniformly mixing) condition (Billingsley 1968) where

$$|P(A \cap B) - P(A)P(B)| \leq \phi_k P(A)$$

for any $A \in \mathcal{F}_1^n$, and $B \in \mathcal{F}_{n+k}^{\infty}$ and ϕ_k tends to zero for $k \to \infty$. The rate at which α_k and ϕ_k go to zero plays an important role in showing asymptotic properties of the nonparametric smoothing

International Statistical Review, 12, 153-172

procedures. We note that generally these conditions are difficult to check. However, if the process follows a stationary Markov chain, then geometric ergodicity implies absolute regularity, which in turn implies strong mixing conditions. Techniques exist for checking the geometric ergodicity, see Tweedie (1975), Tjøstheim (1990), Pham (1985), Diebolt & Guegan (1990).

Local Conditional Mean and Median

Consider the general nonlinear autoregressive process of order p

$$X_t = f(X_{t-1}, \ldots, X_{t-p}) + \varepsilon_t. \tag{3.2}$$

Let $Y_t = (X_{t-1}, \ldots, X_{t-p})$, and choose $\delta_n > 0$ as a function of the sample size n. For any $y = (x_1, \ldots, x_p) \in \mathbb{R}^p$, let $I_n(y) = \{i : 1 < i < n \text{ and } ||Y_i - y|| < \delta_n\}$ and $N_n(y) = \#I_n(y)$. Here $\|\cdot\|$ denotes the Euclidean norm. The local conditional mean function estimator is given by $\hat{f}(x_1, \ldots, x_p) = \hat{f}_n(y) = \{N_n(y)\}^{-1} \sum_{i \in I_n(y)} X_i$, that is, an average of all observations X_i corresponding to Y_i in a small neighborhood of the argument y is used as the estimator. Alternatively, the local conditional median estimator given by $\tilde{f}(x_1, \ldots, x_p) = median\{X_i,\ i \in I_n(y)\}$ may be used. Under strong mixing conditions, Truong (1993) proved strong consistency and asymptotic normality of these estimators, along with the optimal rate of convergence for suitable sequences $\delta_n \to 0$.

Nonparametric Kernel Estimation

Robinson (1983b), Auestad & Tjøstheim (1990), Härdle & Vieu (1992), and others used a kernel estimator (or robustified versions of it) to estimate the conditional mean function $f(X_{t-1}, \ldots, X_{t-p})$. For this purpose the Nadaraya–Watson estimator with product kernels

$$\hat{f}(x_1, \ldots, x_p) = \frac{\sum_{t=p+1}^{n} \prod_{i=1}^{p} K\{(x_i - X_{t-i})/h_i\} X_t}{\sum_{t=p+1}^{n} \prod_{i=1}^{p} K\{(x_i - X_{t-i})/h_i\}}, \tag{3.3}$$

is used where $K(\cdot)$ is again a kernel function with bounded support and the h_i's are the bandwidths. In other words, a weighted average of the observations is used as an estimator of $f(\cdot)$.

Robinson (1983b) and Masry & Tjøstheim (1995a) show strong consistency and asymptotic normality for α-mixing observations. Bierens (1983, 1987) and Collomb & Härdle (1986) proved the uniform consistency of the estimator under the assumption of a ϕ-mixing process. Singh & Ullah (1985) extend this approach to multiple time series, where X_t is a vector rather than a scalar random variable.

Local Polynomial Regression

Local polynomial regression techniques offer yet another alternative for estimating the conditional mean of time series nonparametrically. In this approach polynomials of a prespecified degree, say $l - 1$, are fitted locally in the neighborhood of a given argument of $f(\cdot)$, where the size of the neighborhood shrinks with increasing sample size n. To state this estimator formally, suppose for simplicity that $p = 1$, that is, the model is $X_t = f(X_{t-1}) + \varepsilon_t$. We wish to estimate $f(x)$. In this case the estimator is obtained by minimization of

$$c_n(x) = arg \min_{c \in \mathbb{R}^l} \sum_{t=1}^{n} (X_t - c^T U_{tn})^2 K\{(X_{t-1} - x)/h\}.$$

(1997) Härdle, W. , Lütkepohl, H. and Chen, R.
Nonparametric Time Series Analysis.

International Statistical Review, 12, 153-172

where $K(\cdot)$ is a kernel function, h is a positive bandwidth sequence, and

$$U_{tn} = F(u_{tn}), \quad F(u) = (1, u, \ldots, u^{l-1}/(l-1)!)^T, \quad u_{tn} = (X_{t-1} - x)/h.$$

The estimator $\hat{f}(x)$ is given by $\hat{f}(x) = c_n(x)^T F(0)$. This estimator was first developed by Stone (1977) and Katkovnik (1979). In the context of independent observations Fan (1993) studied minimax efficiency and made the technique popular to applied statisticians. Tsybakov (1986) and Härdle & Tsybakov (1997) proved asymptotic normality of these estimators under conditions satisfying the assumptions of Tweedie (1975) and Diebolt & Guegan (1990). A multivariate extension of this approach is given by Härdle, Tsybakov & Yang (1996).

Nonparametric Multi-step Prediction

All these methods estimate the conditional mean of a nonlinear AR process and thereby provide a one-step ahead predictor. Often forecasts for more than one step ahead are desired. Similar nonparametric techniques can be used for that purpose and we briefly mention some proposals here.

Consider the nonlinear AR(1) model $X_t = f(X_{t-1}) + \varepsilon_t$. Since the conditional mean $m_k(x) = E(X_{t+k} \mid X_t = x)$ is the least squares predictor for k-step ahead prediction, Auestad & Tjøstheim (1990), Härdle & Vieu (1992) and Härdle (1990) proposed using the ordinary Nadaraya–Watson estimator

$$\hat{m}_{h,k}(x) = \frac{\sum_{t=1}^{n-k} K\{(x - X_t)/h\} X_{t+k}}{\sum_{t=1}^{n-k} K\{(x - X_t)/h\}} \tag{3.4}$$

to estimate $E(X_{t+k} \mid X_t = x)$ directly.

Note, however, that the variables $X_{t+1}, \ldots, X_{t+k-1}$ may contain information about the conditional mean function $E(X_{t+k} \mid X_t)$. Therefore Chen (1996) and Chen & Hafner (1995) proposed a multistage kernel smoother which utilizes this information. For illustrative purposes consider two-step ahead forecasting. Due to the Markov property, we have

$$m_2(x) = E[X_{t+2} \mid X_t = x] = E[E(X_{t+2} \mid X_{t+1}, X_t) \mid X_t = x] = E[E(X_{t+2} \mid X_{t+1}) \mid X_t = x].$$

Define $f(y) = E(X_{t+2} \mid X_{t+1} = y)$. Ideally, if we knew $f(\cdot)$, we would use the pairs $(f(X_{t+1}), X_t)$, $t = 1, \ldots, (n-1)$ in estimating $E(X_{t+2} \mid X_t)$, whereas the direct estimator (3.4) uses the pairs (X_{t+2}, X_t). Since X_{t+2} is a noisy representative of $f(X_{t+1})$ with $O_p(1)$ error, we can improve the estimation by using an estimator $\hat{f}(X_{t+1})$ with $\hat{f}(X_{t+1}) - f(X_{t+1}) = o_p(1)$. This motivates the 'multistage smoother'

$$\hat{m}_{h_1,h_2}(x) = \frac{\sum_{t=1}^{n-1} K\{(x - X_t)/h_2\} \hat{f}_{h_1}(X_{t+1})}{\sum_{t=1}^{n-1} K\{(x - X_t)/h_2\}}$$

where

$$\hat{f}_{h_1}(y) = \frac{\sum_{j=1}^{n-1} K\{(y - X_j)/h_1\} X_{j+1}}{\sum_{j=1}^{n-1} K\{(y - X_j)/h_1\}}.$$

It can be shown that the new smoother has a smaller mean squared error than (3.4).

Implementation Issues

One of the important implementation issues of the nonparametric smoothing tools is the bandwidth selection in finite samples. There are many data-driven methods proposed for independent data, e.g. the cross-validation method of Rudemo (1982) and Bowman (1994) and the plug-in rules of Sheather (1983), Park & Marron (1990) and Park & Turlach (1992).

International Statistical Review, 12, 153-172

Again, for simplicity we assume a nonlinear AR(1) model $X_t = f(X_{t-1}) + \varepsilon_t$. For dependent data, one of the criteria for selecting the bandwidth is to minimize the averaged squared error

$$d_A(h) = \frac{1}{n}\sum_{t=1}^{n}\{f(X_t) - \hat{f}_h(X_t)\}^2 w(X_t),$$

which is an approximation of the integrated squared error

$$d_I(h) = \int \{f(x) - \hat{f}_h(x)\}^2 \eta(x) w(x) dx.$$

Here $\eta(\cdot)$ denotes the density of the stationary distribution and $w(\cdot)$ is a weight function with compact support. The measure of accuracy $d_A(h)$ involves the unknown autoregression function $f(\cdot)$, so it cannot be estimated by a plug-in type approach. For the nonparametric kernel estimator, Härdle & Vieu (1992) and Härdle (1990) proposed to use the leave-on-out cross-validation function

$$CV(h) = \frac{1}{n-1}\sum_{t=2}^{n}\{X_t - \hat{f}_{h,t}(X_{t-1})\}^2 w(X_{t-1}),$$

where

$$\hat{f}_{h,t}(x) = \frac{n^{-1}\sum_{j\neq t} K\{(x - X_{j-1})/h\}X_j}{n^{-1}\sum_{j\neq t} K\{(x - X_{j-1})/h\}}, \tag{3.5}$$

to select the bandwidth. Let $\hat{h}$ be the bandwidth that minimizes $CV(h)$. They proved that, under an α-mixing condition,

$$\frac{d_A(\hat{h})}{\inf_h d_A(h)} \to 1 \quad \text{in probability.}$$

Similar results for density estimation were obtained by Hart & Vieu (1990).

A Nonparametric Nonlinearity Test

Hjellvik & Tjøstheim (1995) proposed a nonlinearity test which may help in deciding whether to use a nonlinear model rather than a linear one. It is based on the distance between the best linear predictor $\rho_k X_{t-k}$ and the best nonlinear predictor $m_k(X_{t-k}) = E[X_t \mid X_{t-k}]$ of X_t based on X_{t-k}. The distance is defined as

$$L(m_k) = E[\{m_k(X_{t-k}) - \rho_k X_{t-k}\}^2 w(X_{t-k})]$$

where $w(x)$ is a weighting function with compact support and ρ_k is the autocorrelation between X_t and X_{t-k}, assuming X_t has zero mean. The function $m_k(\cdot)$ is estimated using the Nadaraya–Watson estimator.

Lag Selection and Order Determination

The lag selection and order determination problem is important for effective implementation of nonlinear time series modeling. Often the set of lagged variables and possibly additional exogenous variables is too large for an efficient application of nonparametric smoothing techniques. In that case one may wish to select the most significant components. For linear time series models, lag selection and order determination are usually done using information criteria as proposed by Akaike (1970, 1974), along with other model checking procedures such as residual analysis. In a fully nonparametric approach to time series analysis, Auestad & Tjøstheim (1990) and Tjøstheim & Auestad (1994b) proposed the FPE (final prediction error) criterion and Cheng & Tong (1992) suggested using cross

(1997) Härdle, W. , Lütkepohl, H. and Chen, R.
Nonparametric Time Series Analysis.

International Statistical Review, 12, 153-172

validation. More specifically, Tjøstheim & Auestad (1994b) proposed to use an estimated FPE criterion to select lag variables and to determine the model order of the general nonlinear AR model in (3.2). Let X_t be a stationary strong mixing nonlinear AR process and let $\boldsymbol{i} = (i_1, \ldots, i_p)$ and $Y_t(\boldsymbol{i}) = (X_{t-i_1}, \ldots, X_{t-i_p})^T$. Define

$$\widehat{FPE}(\boldsymbol{i}) = \frac{1}{n}\sum_t [X_t - \hat{f}\{Y_t(\boldsymbol{i})\}]^2 w\{Y_t(\boldsymbol{i})\} \frac{1 + (nh^p)^{-1} J^p B_p}{1 - (nh^p)^{-1}\{2K^p(0) - J^p\} B_p} \tag{3.6}$$

where

$$J = \int K(x)^2 dx, \quad B_p = n^{-1} \sum_t \frac{w\{Y_t(\boldsymbol{i})\}^2}{\hat{p}\{Y_t(\boldsymbol{i})\}}$$

and $\hat{f}\{Y_t(\boldsymbol{i})\}$ is the kernel conditional mean estimator in (3.3) based on the lags specified in $\boldsymbol{i}$ and $\hat{p}\{Y_t(\boldsymbol{i})\}$ is a multivariate kernel density estimator defined as in (3.1). Note that the $\widehat{FPE}$ is essentially a sum of squares of residuals (RSS) multiplied by a term in (3.6) that penalizes small bandwidths h and a large order p.

Cheng & Tong (1992) used a leave-one-out cross validation procedure to select the order of a general nonlinear AR model. Let $Y_t(p) = (X_{t-1}, \ldots, X_{t-p})$ and

$$CV(p) = \frac{1}{n - r + 1} \sum_t [X_t - \hat{f}_{h,t}\{Y_t(p)\}]^2 w\{Y_t(p)\}$$

where $\hat{f}_{h,t}$ is the kernel conditional mean estimator defined in (3.5) and $w(\cdot)$ is a weight function of finite support. They proved that, under regularity conditions,

$$CV(p) = RSS(p)\{1 + 2K(0)\gamma h^{-p}/n + o_p(1/h^p n)\}$$

where $\gamma = \int w(x)dx / \int w(x)p(x)dx$ and h is the bandwidth. Again, this can be viewed as a penalized sum of squares of residuals.

3.2 Restricted Autoregressive Approaches

Since the nonparametric general approach suffers from the 'curse of dimensionality', unless the AR order p is very small, restrictions on the function $f(\cdot)$ have been proposed. Common structural restrictions are additivity, single index restrictions and/or data dependent coefficients in a 'linear' model. These restrictions result in better convergence rates and are easier to interpret, especially with graphics supported from interactive statistical computing environments. This is important since nonparametric models are not the end of an analysis. They are rather an exploratory tool for a better understanding of the underlying dynamics of the process and a starting point for finding more parsimonious models.

Nonlinear Additive AR Models

A nonlinear additive autoregressive (NAAR) model is defined as

$$X_t = c + f_1(X_{t-i_1}) + f_2(X_{t-i_2}) + \cdots + f_p(X_{t-i_p}) + \varepsilon_t. \tag{3.7}$$

Additive models have been studied extensively in the regression context by Hastie & Tibshirani (1990). The NAAR model in (3.7) is a generalization of the first-order nonlinear AR model of Jones (1978). It is very flexible as it encompasses linear AR models and many interesting nonlinear models as special cases. These models naturally generalize the linear regression models and allow interpretation of marginal changes, i.e. the effect of one variable (or lagged variable) on the mean function. They are also interesting from a theoretical point of view since they combine flexible

International Statistical Review, 12, 153-172

nonparametric modeling of many variables with statistical precision that is typical for just one explanatory variable. Accurate estimation can be achieved with moderate sample sizes. Here we introduce three procedures for estimating the NAAR model. Order determination and lag selection problems are addressed as well.

Chen & Tsay (1993a) use *backfitting algorithms* such as the Alternating Conditional Expectation (ACE) algorithm and the BRUTO algorithm of Hastie & Tibshirani (1990) to fit the additive model (3.7). Note that the AVAS algorithm of Tibshirani (1988) can also be used here. The main idea of backfitting is that if the additive model is correct, then for any k we have $f_k(X_{t-i_k}) = E\{X_t - c - \sum_{j\neq k} f_j(X_{t-i_j}) \mid X_{t-i_k}\}$. Consequently, we can treat $X_t - c - \sum_{j\neq k} f_j(X_{t-i_j})$ as the conditional response variable and use nonparametric smoothers to estimate $f_k(\cdot)$. In practice, all $f_k(\cdot)$'s are unknown so that the estimates are iterated until they all converge. The effective hat matrix of this algorithm is computed in Härdle & Hall (1993), showing that the iteration results depend on the starting index.

One of the problems associated with the backfitting algorithms is that with highly correlated observations, the convergence can be slow, as noted in Chen & Tsay (1993a). Linton & Nielson (1995) and Chen *et al.* (1996) proposed an integration estimator for estimating the functions in additive regression models without using backfitting. At the same time, Tjøstheim & Auestad (1994a) and Masry & Tjøstheim (1995b) proposed the same estimator for NAAR models. Specifically, the 'integration idea' is based on the following observation. If the model is of the additive form (3.7), and $f(x_1, \ldots, x_p) = c + \sum_{j=1}^{p} f_j(x_j)$ is the conditional mean function, and $p_{-j}(\cdot)$ is the joint density of $X_{t-i_1}, \ldots, X_{t-i_{j-1}}, X_{t-i_{j+1}}, \ldots, X_{t-i_p}$, then for a fixed $x \in \mathbb{R}$,

$$f_j(x) + c = \int f(x_1, \ldots, x, \ldots, x_p) p_{-j}(x_1, \ldots, x_p) \prod_{l\neq j} dx_l,$$

provided $Ef_l(X_t) = 0$, $l = 1, \ldots, p$. Using the Nadaraya–Watson estimator to estimate the mean function $f(\cdot)$, we average over the observations to obtain the following estimator.

Let $K_h(\cdot) = h^{-1}K(\cdot/h)$, where $K(\cdot)$ is a kernel function. For $1 \leq j \leq p$ and any x in the domain of $f_j(\cdot)$, define, for $h_n > 0$, $h_n' > 0$,

$$\begin{aligned} \hat{f}_j(x) &= \frac{1}{n}\sum_{t=1}^{n} \hat{f}(X_{t-i_1}, \ldots, X_{t-i_{j-1}}, x, X_{t-i_{j+1}}, \ldots, X_{t-i_p}) \\ &= \frac{1}{n}\sum_{t=i_p+1}^{n} \left[\frac{\sum_{s=i_p+1}^{n} \left[\prod_{l\neq j} K_{h_n'}(X_{s-i_l} - X_{t-i_l})\right] K_{h_n}(X_{s-i_j} - x) X_s}{\sum_{s=i_p+1}^{n} \left[\prod_{l\neq j} K_{h_n'}(X_{s-i_l} - X_{t-i_j})\right] K_{h_n}(X_{s-i_j} - x)} \right]. \end{aligned} \tag{3.8}$$

The asymptotic normality of this estimator was established by Chen *et al.* (1996) for independent observations and by Masry & Tjøstheim (1995b) under strong mixing conditions for time series observations. The rate of convergence for estimating $f(\cdot)$ is $n^{2/5}$ which is typical for regression smoothing with just one explanatory variable. Hence, the estimator does not suffer from the 'curse of dimensionality'.

Wong & Kohn (1996) use spline nonparametric regression to estimate the components of a NAAR model. They adopt an equivalent Bayesian formulation of the spline smoothing and use a Gibbs sampler to estimate the components and the parameters of the model, through Monte Carlo simulation of the posterior distributions.

Chen, Liu & Tsay (1995) propose three nonparametric procedures for testing additivity in nonlinear time series analysis. For lag selection, Chen & Tsay (1993a) propose a procedure that is similar to the best subset procedure in linear regression analysis.

(1997) Härdle, W. , Lütkepohl, H. and Chen, R.
Nonparametric Time Series Analysis.

International Statistical Review, 12, 153-172

Functional Coefficient AR Model

A functional coefficient autoregressive (FAR) model can be written as

$$X_t = f_1(X_{t-d})X_{t-1} + f_2(X_{t-d})X_{t-2} + \cdots + f_p(X_{t-d})X_{t-p} + \varepsilon_t.$$

The model generalizes the linear AR models by allowing the coefficients to change according to a threshold lag variable X_{t-d}. The model can be extended to allow for multiple threshold variables in the coefficient functions. The model is general enough to include the threshold AR (TAR) models of Tong (1983) and Tsay (1989) (when the coefficient functions are step functions) and the exponential AR (EXPAR) models proposed by Haggan & Ozaki (1981) (when the coefficient functions are exponential functions) along with many other models (e.g., the STAR models of Granger & Teräsvirta (1993) and Teräsvirta (1994) and sine function models). Chen & Tsay (1993b) use an arranged local regression (ALR) procedure to roughly identify the nonlinear functional forms. For $x \in \mathbb{R}$ and $\delta_n > 0$, let $I_n(x) = \{t : 1 < t < n, |X_{t-d} - x| < \delta_n\}$. If we regress X_t on $X_{t-1}, \ldots, X_{t-p}$ using all the observations X_t for which $t \in I_n(x)$, then the estimated coefficients can be used as estimates of $f_i(x)$, $i = 1, \ldots, p$. One can then make inference directly or formulate parametric models based on the estimated nonlinear functional forms. Chen & Tsay (1993b) proved the consistency of the estimator under geometric ergodicity conditions. Note that the locally weighted regression of Cleveland & Devlin (1988) may be used for estimating FAR models as well.

Adaptive Spline Threshold AR Model

Lewis & Stevens (1991) propose the adaptive spline threshold autoregressive (ASTAR) model of the form $X_t = \sum_{j=1}^{s} c_j K_j(X_{t-1}, \ldots, X_{t-p}) + \varepsilon_t$, where $\{K_j(x)\}_{j=1}^{s}$ are product basis functions of truncated splines $T^-(x) = (t - x)_+$ and $T^+(x) = (x - t)_+$ associated with the subregions $\{R_j\}_{j=1}^{s}$ in the domain of the lag variables $(X_{t-1}, \ldots, X_{t-p})$. For example, Lewis & Stevens (1991) use the following ASTAR model for the famous sunspot numbers:

$$\begin{aligned}\hat{X}_t \;=\; & 2.711 + 0.96X_{t-1} + 0.332(47 - X_{t-5})_+ - 0.257(59.1 - X_{t-9})_+ \\ & -0.003X_{t-1}(X_{t-2} - 26.0)_+ + 0.017X_{t-1}(44.0 - X_{t-3})_+ \\ & -0.032X_{t-1}(17.1 - X_{t-4})_+ + 0.004X_{t-1}(26 - X_{t-2})_+(X_{t-5} - 41.0)_+\end{aligned}$$

where $(u)_+ = u$ if $u > 0$ and $(u)_+ = 0$ if $u \leq 0$. The modeling and estimation procedures follow the Multivariate Adaptive Regression Splines (MARS) algorithm of Friedman (1988). It is basically a regression tree procedure using truncated regression splines.

Index Models

Bierens (1994) discusses another way of imposing constraints on the general model (1.1). He shows that for a rational valued process the conditional expectation can be written as a function of an index, i.e. $E(X_t|X_{t-1}, X_{t-2}, \ldots) = f(\xi_t)$, where the index ξ_t is related to the past observations $X_{t-1}, X_{t-2}, \ldots$. For instance, the index may be of the form $\xi_t = \sum_{i=1}^{\infty} \eta^{i-1} X_{t-i}$ for some $\eta \in (-1, 1)$. Obviously, in this case $f(\cdot)$ is one dimensional and is therefore relatively easy to estimate by kernel methods. For practical purposes, assuming that X_t is rational is not restrictive because on a computer only a finite number of digits can be stored so that all observed time series are actually rational.

Bierens shows that there is a wide range of indices to choose from and suggests the following procedure for applied work. In a first step the best fitting linear ARMA model should be constructed. The optimal linear one-step-ahead predictor from that model is then used as an index ξ_t. If especially designed specification tests indicate remaining nonlinearity the function $f(\cdot)$ may be chosen either

International Statistical Review, 12, 153-172

from some parametric family or by using nonparametric smoothing techniques. Of course, a linear model is maintained if no nonlinearity is detected.

3.3 *Global Approximators*

As mentioned previously, a sequence of parametric functions can be used as global approximators to approximate the conditional mean function $f(\cdot)$ in (1.1). As the sample size increases, the dimension of the parameter space also increases to achieve greater approximation accuracy. Thereby it is possible to allow $f(\cdot)$ to depend on infinitely many lagged variables although only a finite number of lags is considered for any given finite sample size. The approaches of this type differ in the class of parametric functions used. We begin with simple linear functions where just the number of lags in the model grows with the sample size. For this class it is particularly easy to discuss the assumptions usually made for deriving asymptotic properties of estimators. Then we consider neural networks as an important general class of nonlinear approximators.

Linear Functions

Suppose $\{X_t\}$ is a zero mean purely nondeterministic causal stationary process, then it has an AR representation of potentially infinite order,

$$X_t = \sum_{i=1}^{\infty} \alpha_i X_{t-i} + \varepsilon_t.$$

If the second order moment properties of the process are of interest only it suffices to obtain the above representation which is linear in lagged X_t. Hence, the second order moment properties of the process may be estimated by approximating its infinite order AR representation. The simplest way to accomplish this is by fitting finite order AR(H_n) processes

$$X_t = \sum_{i=1}^{H_n} \alpha_i X_{t-i} + \varepsilon_{H_n,t},$$

where the order H_n is an increasing function of the sample size n. To obtain desirable properties of the resulting estimators and quantities derived from them we need to assume that the AR order H_n goes to infinity at a much smaller rate than n so that there is eventually enough information for estimating the parameters efficiently. On the other hand, the approximation quality must improve sufficiently rapidly so as to avoid large bias. Hence, there must be an appropriate lower bound on the rate of divergence of H_n. More precisely, it may be assumed that

(1.) H_n is $o(n^{1/3})$, and

(2.) $\sqrt{n} \sum_{i>H_n} |\alpha_i| \to 0$,

as $n \to \infty$. Here the two conditions are upper and lower bounds, respectively, on the rate at which the AR order goes to infinity with n. Under these conditions and mild assumptions for $\{\varepsilon_t\}$ the least squares estimators of the α_i are consistent and asymptotically normal. In fact, for consistency weaker conditions for H_n suffice.

Akaike (1969), Parzen (1974), Berk (1974) and Bhansali (1978) use this approach for spectral estimation and prediction of univariate processes. Parzen (1977), Lewis & Reinsel (1985), Lütkepohl (1991, Ch. 9) and Lütkepohl & Poskitt (1996) discuss multivariate extensions. They also consider estimation of other quantities derived from the autoregressive coefficients. Most of these results can be extended to nonstationary integrated and cointegrated processes (see Section 5.3).

International Statistical Review, 12, 153-172

Note that $\hat{X}_t = \sum_{i=1}^{\infty} \alpha_i X_{t-i}$ is the best (minimum MSE) *linear* 1-step predictor which may not be the conditional expectation and, hence, it may not be the optimal predictor in a more general class of nonlinear predictors. Consequently, it may be desirable to consider nonlinear functions $f_n(\cdot)$ to approximate the conditional mean function $f(\cdot)$. We will present one possible nonlinear approach next.

Neural Networks

Neural networks have been used in various fields to approximate complex nonlinear structures. Their name comes from the fact that they may be thought of as a network of neurons similar to (but of course much simpler as) the brain. The related computations may be extremely complex. Therefore neural network analysis nowadays represents a subfield of computer science or, more precisely, of artificial intelligence. Here we consider the *single hidden layer feedforeward network* which may be best thought of as a class of flexible nonlinear functions of the form

$$f_n(X_{t-1}, \ldots, X_{t-p}) = \beta_0 + \sum_{j=1}^{q} G(\gamma_{0j} + Y_t^T \gamma_j)\beta_j, \tag{3.9}$$

where $Y_t = (X_{t-1}, \ldots, X_{t-p})^T$ and the $\gamma_j = (\gamma_{1j}, \ldots, \gamma_{pj})^T$ are $(p \times 1)$ vectors for $j = 1, \ldots, q$, and $\beta_0, \beta_1, \ldots, \beta_q$ are scalar coefficients. The function $G : \mathbb{R} \to [0, 1]$ is a prespecified cumulative distribution function. Typical examples are the logistic function $G(x) = 1/(1 + e^{-x})$ and the hyperbolic function $G(x) = \tanh(x) = (e^x - e^{-x})/(e^x + e^{-x})$. Functions of the type (3.9) can approximate broad classes of functions if q is sufficiently large. Thus, if q increases with the sample size n, a good approximation of $f(X_{t-1}, \ldots, X_{t-p})$ will eventually result. The function in (3.9) may also be estimated without specifying $G(\cdot)$ by using the projection pursuit regression of Hutchinson, Lo & Poggio (1994). In the following we will, however, assume a given specific form of $G(\cdot)$.

For practical purposes it will be advantageous to obtain a good approximation with small or moderate values of q. Therefore adding a linear AR term in (3.9) is often useful. Thus, in practice, a possible approximating function is

$$f_n(X_{t-1}, \ldots, X_{t-p}) = \beta_0 + \sum_{i=1}^{p} \alpha_i X_{t-i} + \sum_{j=1}^{q} G(\gamma_{0j} + Y_t^T \gamma_j)\beta_j.$$

For given p and q, estimation of the parameters of this model is possible with LS procedures. Asymptotic properties of the resulting estimators are available both for fixed q and q increasing with the sample size. Kuan & White (1994) provide a comprehensive survey of neural network models and estimation results for the present situation. Also it is possible to let the number of lags p (i.e., the AR order) increase with the sample size. This, however, results in further complications of the asymptotic theory.

Since nonlinear optimization algorithms may be time consuming, it is undesirable to reestimate a model each time new observations become available. Therefore sequential estimation or learning procedures have been proposed which update the available estimates sequentially when new sample information becomes available. A prominent example is the backpropagation procedure (see Rumelhart, Hinton & Williams 1986). Kuan & White (1994) present asymptotic results for this procedure as well.

The network represented by (3.9) feeds the output of the neurons (the $G(\cdot)$) directly into the overall output and there is also no direct interaction between the neurons. There are various generalizations of this simple architecture. For instance, *multi-layer networks* may be considered. An example of a

International Statistical Review, 12, 153-172

2-layer network is

$$f_n(X_{t-1}, \ldots, X_{t-p}) = \beta_0 + \sum_{j=1}^{q_2} G_2\left(\sum_{l=1}^{q_1} G_1(\gamma_{0l} + Y_t^T \gamma_l)\beta_l\right)\delta_j,$$

where $G_1(\cdot)$ and $G_2(\cdot)$ are now prespecified cumulative distribution functions and the γ_{ij}, β_k and δ_j are unknown parameters which have to be estimated. Another possible extension would be to allow for feedback between the neurons. The following is an example of a *recurrent single hidden layer network*:

$$f_{n,t}(X_{t-1}, X_{t-2}, \ldots, X_0) = \beta_0 + \sum_{j=1}^{q} \phi_{tj}\beta_j, \quad t = 0, 1, 2, \ldots$$

where

$$\phi_{tj} = G(z_t^T \gamma_j + \sum_{l=1}^{q} \phi_{t-1,l}\delta_{lj}), \quad j = 1, 2, \ldots, q.$$

Although the simpler single hidden layer feedforward networks have quite general approximation properties it may be useful in practice to consider more sophisticated architectures to obtain a good approximation with fewer terms (or neurons) than that in (3.9). Also there may be information on the structure of a data generation mechanism that suggests multi-layer or feedback architectures.

In practice there will often be uncertainty regarding the most suitable architecture for a given time series and regarding the number of lags and neurons that guarantee a good approximation of the actual generation mechanism. Therefore methods have been proposed for model selection and for deciding on restrictions that may be imposed on a given neural network model. For instance, Murata, Yoshizawa & Amari (1994) proposed a model selection criterion which extends the ideas underlying the AIC criterion to the present situation. Specification tests are also reviewed by Kuan & White (1994).

As mentioned earlier, neural networks establish a subfield of computer science and are applied in many areas. Therefore it is impossible to provide a complete survey of the literature in a limited review of this type. Those interested in this fascinating tool for nonparametric time series analysis may find the survey article by Kuan & White (1994) a useful point of departure for further studies.

4 Estimating Higher Order Conditional Moments and Densities

Techniques similar to those discussed for estimating the conditional expectation of a process may also be used for approximating higher order conditional moments which are often of interest, as we have argued earlier. Here we summarize some of these extensions. We begin with methods for estimating conditional variances in addition to conditional means. Then some possibilities for approximating the complete conditional density are presented.

4.1 Conditional Variances

Nonparametric Kernel Estimation

Auestad & Tjøstheim (1990) and Tjøstheim & Auestad (1994a,b) use kernel estimation techniques for analyzing models like (1.2) assuming that both the conditional mean and the conditional variance function depend on at most p lagged X_t. The function $f(\cdot)$ may again be estimated by the Nadaraya–Watson estimator with product kernels as in Section 3.1,

$$\hat{f}(x_1, \ldots, x_p) = \frac{\sum_{t=p+1}^{n} \prod_{i=1}^{p} K\{(x_i - X_{t-i})/h_i\} X_t}{\sum_{t=p+1}^{n} \prod_{i=1}^{p} K\{(x_i - X_{t-i})/h_i\}},$$

(1997) Härdle, W. , Lütkepohl, H. and Chen, R.
Nonparametric Time Series Analysis.

International Statistical Review, 12, 153-172

and the conditional variance $g(\cdot)^2$ may be estimated by

$$\hat{g}(x_1, \ldots, x_p)^2 = \frac{\sum_{t=p+1}^{n} \prod_{i=1}^{p} K\{(x_i - X_{t-i})/h_i\} X_t^2}{\sum_{t=p+1}^{n} \prod_{i=1}^{p} K\{(x_i - X_{t-i})/h_i\}} - \{\hat{f}(x_1, \ldots, x_p)\}^2,$$

where again $K(\cdot)$ is a kernel function with bounded support and the h_i's are the bandwidths.

Masry & Tjøstheim (1995a) show strong consistency and asymptotic normality of these estimators for α-mixing observations and Tjøstheim & Auestad (1994a,b) consider model specification and lag selection in models of the form (1.2).

Local Polynomial Regression and Other Techniques

Local polynomial nonparametric regression techniques can be used in an analogous fashion to estimate the conditional mean and variance functions. Assume $p = 1$ so that the functions $f(\cdot)$ and $g(\cdot)$ depend on X_{t-1} only. Then they may be estimated by minimization of

$$c_n(x) = \arg\min_{c \in \mathbb{R}^l} \sum_{t=1}^{n} (X_t - c^T U_{tn})^2 K\{(X_{t-1} - x)/h\}$$

as in Section 3.1, and

$$s_n(x) = \arg\min_{s \in \mathbb{R}^l} \sum_{t=1}^{n} (X_t^2 - s^T U_{tn})^2 K\{(X_{t-1} - x)/h\}$$

where h is again a positive bandwidth, and

$$U_{tn} = F(u_{tn}), \quad F(u) = (1, u, \ldots, u^{l-1}/(l-1)!)^T, \quad u_{tn} = (X_{t-1} - x)/h.$$

Here the degree of the approximating polynomial is assumed to be $l - 1$. The estimators $\hat{f}(x)$ and $\hat{g}(x)$ are given by

$$\hat{f}(x) = c_n(x)^T F(0) \quad \text{and} \quad \hat{g}(x) = s_n(x)^T F(0) - \{c_n(x)^T F(0)\}^2.$$

Härdle & Tsybakov (1996) prove asymptotic normality of these estimators under similar conditions as in Section 3.1 where the conditional mean was estimated only.

An extension of this model to nonparametric vector autoregression is presented in Härdle, Tsybakov & Yang (1996) who consider the model

$$X_t = f(Y_t) + \Sigma^{1/2}(Y_t)\varepsilon_t, \quad t = p, p+1, \ldots,$$

where $X_t = (X_{t1}, X_{t2}, \ldots, X_{td})^T \in \mathbb{R}^d$, $\varepsilon_t = (\varepsilon_{t1}, \varepsilon_{t2}, \ldots, \varepsilon_{td})^T \in \mathbb{R}^d$ and $Y_t = (X_{t-1}, X_{t-2}, \ldots, X_{t-p}) \in \mathbb{R}^{d \times p}$ is a matrix of lagged variables.

Alternatively, conditional heteroscedasticity can also be modeled with neural network methods (Weigend & Nix 1994).

4.2 *Estimating the Predictive Density*

Kernel Techniques

For a stationary time series, Robinson (1983b) proposed a kernel estimator to estimate the one-step-ahead transition density $h(y \mid x)$. Note that $h(y \mid x) = p(x, y)/p(x)$, where $p(x, y)$ is the joint density of (X_t, X_{t+1}) and $p(x)$ is the marginal density of X_t. Replacing the terms on the right-hand

International Statistical Review, 12, 153-172

side with corresponding kernel estimators, we have

$$\hat{p}(y \mid x) = \frac{(nh^2)^{-1} \sum_{t=1}^{n-1} K_2[\{(x, y) - (X_t, X_{t+1})\}/h]}{(nh)^{-1} \sum_{t=1}^{n} K[(x - X_t)/h]}$$

where $K_2(\cdot)$ is a bivariate kernel function, commonly of the product form $K_2(u, v) = K(u)K(v)$. Note that the estimation of the transition density allows us to construct nonparametric multi-step-ahead prediction density functions as well. For extensions see Singh & Ullah (1985).

Hermite Expansion Approach

Gallant & Tauchen (1989) used Hermite expansions to approximate the one-step-ahead conditional density of the process given its past. This approach is based on the fact that a large class of density functions, $h(y)$ say, is proportional to $[P(z)]^2\phi(z)$, where $z = (y - \mu_y)/\sigma_y$, with μ_y and σ_y location and scale parameters of the distribution, respectively, $P(z) = 1 + \psi_1 z + \cdots + \psi_r z^r$ is a polynomial of possibly infinite degree r and $\phi(z) = (2\pi)^{-1} \exp(-z^2/2)$ is the standard normal density. Dividing $[P(z)]^2\phi(z)$ by a normalizing constant this is just the Hermite expansion of $h(y)$. Hence, the density may be written as the product of a standard normal density and the square of a polynomial.

In the present situation we are interested in the conditional density $h(x_t|x_{t-1}, x_{t-2}, \ldots)$. By the foregoing considerations we have

$$h(x_t|x_{t-1}, x_{t-2}, \ldots) \propto [P(z_t)]^2\phi(z_t)$$

where $z_t = (x_t - \mu_t)/\sigma_t$ with μ_t and σ_t being location and scale parameters, respectively, of the conditional distribution. The former is assumed to be a linear function of the past, $\mu_t = \nu + \alpha_1 X_{t-1} + \cdots + \alpha_p X_{t-p}$, and the latter may be modeled as

$$\sigma_t = \rho_0 + \rho_1|X_{t-1}| + \cdots + \rho_q|X_{t-q}|.$$

The specification of the conditional scale parameter σ_t is similar but not identical to an ARCH process as originally proposed by Engle (1982). Alternative specifications may be used here. At any rate, the location and scale parameters μ_t and σ_t are modeled parametrically whereas higher order moment terms are captured by the polynomial. Letting the polynomial degree increase with the sample size makes this approach nonparametric. Overall the approach has been termed semi nonparametric (SNP) because it combines parametric with nonparametric elements.

To achieve a flexible adjustment of the model to higher order dynamics the polynomial coefficients $\psi_1, \ldots, \psi_r$ may be made dependent on the past, that is,

$$\begin{aligned}\psi_j(x_{t-1}, \ldots, x_{t-K}) &= \psi_{j0} + \sum_{k=1}^{K} \psi_{j1}^{(k)} x_{t-k} + \sum_{k=1}^{K}\sum_{h=1}^{K} \psi_{j2}^{(kh)} x_{t-k}x_{t-h} + \cdots \\ &\quad + \sum_{k=1}^{K} \cdots \sum_{h=1}^{K} \psi_{jl}^{(\cdots)} x_{t-k} \cdots x_{t-h}\end{aligned}$$

where usually small values of K and l are sufficient to guarantee a rich dynamic structure. Of course, for $r = K = l = 0$ we get

$$h(x_t|x_{t-1}, x_{t-2}, \ldots) \propto \phi\left(\frac{x_t - \nu - \alpha_1 x_{t-1} - \cdots - \alpha_p x_{t-p}}{\sigma_t}\right)$$

so that we have a linear AR(p) process with conditionally heteroscedastic error term.

For given values of p, q, r, K and l the parameters of the model may be estimated by maximum

International Statistical Review, 12, 153-172

likelihood which is easily accomplished by minimizing the normalized negative log likelihood

$$L(\theta) = -\frac{1}{n}\sum_{t=1}^{n} \log h(X_t | X_{t-1}, \ldots, X_{t-p}; \theta).$$

Asymptotic properties of this estimation procedure are given by Gallant & Nychka (1987) who allow the order of the Hermite expansion to increase with the sample size. In principle, an extension of this approach to the multivariate case is possible (see Gallant & Tauchen 1989).

5 Other Nonparametric Techniques for Time Series

5.1 Density Estimation with Correlated Observations

Kernel Methods

There is a rich literature on density estimation for independent observations, see Silverman (1986) and the references therein. A popular method is the kernel estimator of the form (3.1) where the kernel function $K(\cdot)$ is typically a probability density function. The key in density estimation is the bandwidth selection. A number of different methods have been proposed, including the cross-validation (Rudemo 1982, Bowman 1984) and the plug-in rules of Sheather (1983), Park & Marron (1990) and Park & Turlach (1992).

The earliest work on density estimation for stationary processes is that of Roussas (1969) and Rosenblatt (1970). The properties of the kernel estimator for dependent observations were investigated by Robinson (1983b) and Hall & Hart (1990a). They found that the bias of the estimator is not affected by the serial correlation. However, the variance is affected. The cross-validation method for dependent observations is studied by Hart & Vieu (1990), under certain regularity conditions. Detailed information and references can be found in Györfi, Härdle, Sarda & Vieu (1989), Prakasa Rao (1983) and Hart (1996). Density estimation for long range dependent data was studied by Hall, Lahiri & Truong (1994) and Csörgő & Mielniczuk (1995a).

Testing for Serial Dependence

Kernel density estimation techniques may also be used to test for independence, for instance, in checking the residual behavior of an estimated nonlinear time series model. Skaug & Tjøstheim (1993) proposed a nonparametric test for independence between two variables which is suitable in this situation. They propose to estimate the quantity

$$I = \int \{p(x, y) - p_1(x)p_2(y)\}^2 p(x, y)w(x, y)dxdy$$

where $p(x, y)$ is the joint density and $p_1(\cdot)$, $p_2(\cdot)$ are the marginal densities while $w(\cdot, \cdot)$ is a weight function with compact support. Using kernel density estimators, we obtain

$$\hat{I} = \frac{1}{n}\sum_{t}\{\hat{p}(X_t, Y_t) - \hat{p}_1(X_t)\hat{p}_2(Y_t)\}^2 w(X_t, Y_t).$$

which should be small under the null hypothesis that X and Y are independent and which can therefore be the basis for an independence test.

5.2 Bootstrap Methods

The bootstrap method is an important nonparametric tool which has also been used for time series analysis in a number of different ways. For instance, it may be used for assessing and improving

International Statistical Review, 12, 153-172

the properties of estimators and forecasts. Originally it was proposed for independent observations (Efron & Tibshirani 1993). Therefore an obvious extension to time series analysis is to bootstrap the residuals of some model. This approach has been used in many applications. Efron & Tibshirani (1993) discuss estimating the standard errors of linear autoregressive parameter estimates using this approach. Bose (1988) evaluates the distribution of the parameter estimator of an AR(1) model by the bootstrap and Kreiss & Franke (1992) discuss its extensions to ARMA(p, q) processes. Furthermore, Franke & Härdle (1992) propose a bootstrap method for spectral estimation.

It is also possible to apply a bootstrap directly to the time series observations by sampling blocks of observations rather than individual ones. This method is known as the moving blocks bootstrap. Specifically, given a time series $X_1, \ldots, X_n$, all possible blocks of $l < n$ consecutive observations are considered and random samples of blocks are drawn and joint together to form a bootstrap time series of roughly length n. This process is repeated B times so that B bootstrap time series are obtained. These artificial series may be used to investigate the distributional properties of the original time series. The moving blocks bootstrap for time series was introduced by Künsch (1989) and Liu & Singh (1992). An introductory exposition is given by Efron & Tibshirani (1993, Sec. 8.6).

5.3 Trend Analysis

In much of the previous discussion we have assumed stationary processes. In practice many time series have trends and are therefore nonstationary. These trends may be removed prior to an analysis of the stationary part of the process if the trend function is known. In most cases it is unknown, however. In that situation nonparametric techniques may be used for trend estimation or trend elimination.

Estimating Trend Functions

Here we consider the case when the trend is characterized by a smooth deterministic function. Suppose $X_1, \ldots, X_n$ is a possibly nonstationary time series with trend $\mu(t) = E(X_t)$. Under the assumption that the trend is smooth, a traditional way of estimating the trend function is the running mean estimator described in Chatfield (1974). A more recent proposal is due to Hart (1991) who uses the kernel smoother of Gasser & Müller (1979) of the form

$$\hat{\mu}_{th} = \frac{1}{h} \sum_{i=1}^{n} X_i \int_{(i-1)/n}^{i/n} K\left(\frac{(t - 0.5)/n - u}{h}\right) du$$

for trend estimation. Hart (1994) proposed a method called time series cross-validation for selecting the bandwidth h. He noted that the ordinary leave-one-out cross-validation tends to select a bandwidth many orders of magnitude too small, if the data are highly positively correlated.

Nonparametric Regression with Dependent Errors

Consider the fixed-design regression model

$$X_{in} = m(z_{in}) + \varepsilon_{in}$$

where $z_{in} = i/n$ and the errors $\{\varepsilon_{in}\}$ are correlated, both the Gasser & Müller (1979) estimator

$$\hat{m}_n(z) = \frac{1}{h_n} \sum_{i=1}^{n} X_i \left\{ \int_{z_{(i-1)n}}^{z_{in}} K\left(\frac{z - t}{h_n}\right) dt \right\}$$

(1997) Härdle, W. , Lütkepohl, H. and Chen, R.
Nonparametric Time Series Analysis.

and the Nadaraya–Watson type estimator

$$\hat{m}_n(z) = \frac{1}{nh_n} \sum_{i=1}^{n} K\left(\frac{z - z_{in}}{h_n}\right) X_i$$

have been proposed and studied. See Hart & Wehrly (1986) and Härdle (1990). Hall & Hart (1990b) and Csörgő & Mielniczuk (1995b) studied the same problem with long-range dependent errors.

Truong & Patil (1996) propose to use wavelet methods to estimate possibly discontinuous trends. Wavelet estimators have been shown to have extraordinary adaptability in handling discontinuity of the underlying function with independent observations (Donoho & Johnstone 1992, Donoho *et al.* 1995, and Hall & Patil 1995). They may be equally powerful in time series analysis.

Nonparametric Unit Root and Cointegration Tests

As an alternative to a deterministic trend, a time series may have a stochastic trend which can be removed by differencing. A process is said to be integrated of order d, $I(d)$, if a stochastic trend can be removed by differencing d times. For example, a random walk $X_t = X_{t-1} + \varepsilon_t$ with white noise error process ε_t is $I(1)$ because $X_t - X_{t-1} =: \Delta X_t = \varepsilon_t$. Nonparametric tests can be used for checking the order of integration of a process.

The random walk is the simplest version of a stochastic trend. Fuller (1976) and Dickey & Fuller (1979) therefore consider an AR(1) model

$$X_t = \rho X_{t-1} + \varepsilon_t \tag{5.1}$$

and test $H_0 : \rho = 1$ against $H_1 : \rho < 1$. An obvious test statistic is the t-ratio based on the LS estimator $\hat{\rho}$ of ρ:

$$t_{\hat{\rho}} = \frac{\hat{\rho} - 1}{s_{\hat{\rho}}}$$

where $s_{\hat{\rho}}$ is the usual estimator of the standard error of $\hat{\rho}$. Equivalently, this statistic may be obtained as the t-ratio of the parameter estimator in the model

$$\Delta X_t = \alpha X_{t-1} + \varepsilon_t$$

where $\alpha = \rho - 1$. The resulting test is also known as Dickey–Fuller (DF) test. The t-statistic does not have the usual standard normal limiting distribution but it has a nonstandard distribution for which the relevant critical values have been tabulated in Fuller (1976).

In practice, the model (5.1) is often too limited to be a reasonable approximation to the underlying data generating process. Therefore more general assumptions are often made for the error process $\{\varepsilon_t\}$. For instance, it may be assumed to be a stationary process. Ignoring the dependency of the ε_t in that case in constructing the test statistic may result in a badly biased test. Therefore nonparametric techniques are often used to model the dependence of the ε_t. One possible approach fits autoregressions

$$\Delta X_t = \alpha X_{t-1} + \pi_1 \Delta X_{t-1} + \cdots + \pi_H \Delta X_{t-H} + \varepsilon_t \tag{5.2}$$

where H goes to infinity with the sample size (see Said & Dickey 1984). Alternatively, a correction for the t-statistic based on spectral techniques has been proposed by Phillips & Perron (1988).

Tests of the foregoing type are often referred to as unit root tests. There is an extensive literature on these tests. Extensions allow also for deterministic terms such as intercepts and linear time trends (see Hamilton 1994, Chapter 17, for details). Also tests of the null hypothesis of a stationary process against the alternative of a unit root have been proposed (see Kwiatkowski, Phillips, Schmidt & Shin 1992). Again spectral techniques are used in the latter variant of a unit root test to account for higher order dynamics of the data generating process.

International Statistical Review, 12, 153-172

Multivariate extensions of the DF tests were proposed by Johansen (1989, 1991). In a multivariate AR process, unit roots indicate that some or all of the components are integrated variables. There may be linear combinations of the variables, however, which are stationary or integrated of lower order. This phenomenon is known as cointegration. Therefore unit root tests in multivariate processes are treated under the heading of testing for cointegration. Nonparametric variants of the Johansen tests are considered by Saikkonen & Luukkonen (1997) who approximate the stationary part of the process by autoregressions of growing order when the sample size increases analogously to (5.2). Cointegration tests based on spectral techniques are discussed by Stock & Watson (1988).

Further nonparametric generalizations of unit root tests are obtained by assuming that there may be an AR unit root in some unknown nonlinear monotone transformation of the original variables. To check the existence of such a unit root in the data generating process, DF or other unit root tests based on the ranks of X_t may be used (see Granger & Hallman 1991, Campbell & Dufour 1993, Breitung & Gouriéroux 1997).

5.4 *Adaptive Estimation*

In a model with finite dimensional parameter vector of interest θ, say, and an infinite dimensional nuisance parameter vector ψ, say, the latter is often taken care of with nonparametric methods. If that is done in such a way that the estimator for θ is asymptotically efficient, it is said to be estimated adaptively. In time series models the conditional mean and variance functions are often of foremost interest. They are therefore often parameterized in a specific way, for instance, as a linear function of the past. The remaining parts of the data generating process may then be estimated nonparametrically. A number of authors have dicussed adaptive methods in this context (e.g., Linton 1993, Kreiss 1987, Robinson 1988, Steigerwald 1992, Engle & Gonzáles-Rivera 1991, Werker 1995, Drost, Klaassen & Werker 1994).

References

Akaike, H. (1969). Power spectrum estimation through autoregressive model fitting. *Annals of the Institute of Statistical Mathematics*, **21**, 407–419.

Akaike, H. (1970). Statistical predictor identification. *Ann. Inst. Statist. Math.* **22**, 203–217.

Akaike, H. (1974). A New Look at Statistical Model Identification, *IEEE Transactions on Automatic Control*, **AC-19**, 716–722.

Auestad, B. & Tjøstheim, D. (1990). Identification of nonlinear time series: First order characterization and order estimation, *Biometrika*, **77**, 669–687.

Bartlett, M.S. (1950). Periodogram analysis and continuous spectra. *Biometrika*, **37**, 1–16.

Becker, R.A., Chambers, J. M. & Wilks, A. R. (1988). *The New S Language*. New York: Chapman and Hall.

Berk, K.N. (1974). Consistent autoregressive spectral estimates. *Annals of Statistics*, **2**, 489–502.

Bhansali, R.J. (1978). Linear prediction by autoregressive model fitting in the time domain. *Annals of Statistics*, **6**, 224–231.

Bierens, H.J. (1983). Uniform consistency of kernel estimators of a regression function under generalized conditions. *Journal of the American Statistical Association*, **78**, 699–707.

Bierens, H.J. (1987). Kernel estimators of regression functions, in T.F. Bewley (ed.) *Advances in Econometrics: Fifth World Congress*, Vol. 1. Cambridge: Cambridge University Press.

Bierens, H.J. (1994). *Topics in Advanced Econometrics: Estimation, testing, and specification of cross-section and time series models*. Cambridge: Cambridge University Press.

Billingsley, P. (1968). *Convergence of Probability Measures*. New York: Wiley.

Blackman, R.B. & Tukey, J.W. (1959). *The Measurement of Power Spectrum from the Point of View of Communications Engineering*. New York: Dover.

Bose, A. (1988). Edgeworth correction by bootstrap in autoregressions. *Annals of Statistics*, **16**, 1709–1722.

Bowman, A.W. (1994). An alternative method of cross-validation for the smoothing of density estimates. *Biometrika*, **71**, 353–360.

Breitung, J. & Gouriéroux, C. (1997). Rank tests for unit roots. *Journal of Econometrics*, forthcoming.

Brillinger, D.R. & Krishnaiah, P.R. (Ed.) (1983). *Handbook of Statistics 3, Time Series in the Frequency Domain*. Amsterdam: North-Holland.

Campbell, B. & Dufour, J.-M. (1993). Exact nonparametric orthogonality and random walk tests. Working paper, C.R.D.E., Montreal.

(1997) Härdle, W. , Lütkepohl, H. and Chen, R.
Nonparametric Time Series Analysis.

Chan, K.S. & Tong, H. (1986). On estimating thresholds in autoregressive models. *Journal of Time Series Analysis*, **7**, 179–190.

Chatfield, C. (1984). *The Analysis of Time Series: An Introduction*, 3rd ed. London: Chapman and Hall.

Chen, R. (1996). A nonparametric multi-step prediction estimator in Markovian structures. *Statistica Sinica*, **6**, 603–615.

Chen, R. & Hafner, C. (1995). Nonlinear time series analysis, in *XploRe, an Interactive Statistical Computing Environment*, (ed. Härdle, W., Klinke, S. & Turlach, B.). Heidelberg: Springer Verlag.

Chen, R., Härdle, W., Linton, O.B. & Severance-Lossin, E. (1996). Estimation in additive nonparametric regression. In *COMPSTAT meeting Semmering*, Härdle, W. & Schimek, M. (eds), Physika Verlag.

Chen, R., Liu, J.S. & Tsay, R.S. (1995). Additivity tests for nonlinear autoregressive models. *Biometrika*, **82**, 369–383.

Chen, R. & Tsay, R. S. (1993a). Nonlinear additive ARX models. *Journal of the American Statistical Association*, **88**, 955–967.

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

Cheng, B. & Tong, H. (1992). On consistent non-parametric order determination and chaos (with discussion). *Journal of the Royal Statistical Society, Series B*, **54**, 427–474.

Cleveland, W. S. & Devlin, S. J. (1988). Locally weighted regression: An approach to regression analysis by local fitting. *Journal of the American Statistical Association*, **83**, 596–610.

Collomb, G. & Härdle, W. (1986). Strong uniform convergence rates in robust nonparametric time series analysis and prediction: Kernel regression estimation from dependent observations. *Stochastic Processes and their Applications*, **23**, 77–89.

Csörgő, S. & Mielniczuk, J. (1995a). Density estimation under long-range dependence. *Ann. Statist.* **23**, 990–999.

Csörgő, S. & Mielniczuk, J. (1995b). Nonparametric regression under long-range dependent normal errors. *Ann. Statist.* **23**, 1000–1014.

Dahlhans, R. (1993). *Fitting time series models to nonstationary processes*. Institut für Angewandte Mathematik, Universität Heidelberg.

Dickey, D.A. & Fuller, W.A. (1979). Distribution of the estimators for autoregressive time series with unit root. *Journal of the American Statistical Association*, **74**, 427–431.

Diebold, F. & Nason, J. (1990). Nonparametric exchange rate prediction. *Journal of International Economics*, **28**, 315–332.

Diebolt, J. & Guegan, D. (1990). Probabilistic properties of the general nonlinear autoregressive process of order one. *Technical report*, N^o 128, L.S.T.A., Université de Paris VI.

Donoho, D.L. & Johnstone, I.M. (1992). Minimax estimation via wavelet shrinkage. *Technical Report 402*, Dept. Stat., Stanford University.

Donoho, D.L., Johnstone, I.M., Kerkyacharian, G. & Picard, D. (1995). Wavelet Shrinkage: Asymptopia? (with discussion). *Journal of the Royal Statistical Society, Series B*, **57**, 301–369.

Drost, F.C., Klaassen, C.A.J. & Werker, B.J.M. (1994). Adaptive estimation in time series models. CentER Discussion Paper 9488, Tilburg University.

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

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

Engle, R.F. & Gonzáles-Rivera, G. (1991). Semiparametric ARCH models. *Journal of Business and Economic Statistics*, **9**, 345–359.

Fan, J. (1993). Local linear regression and their minimax efficiency. *Annals of Statistics*, **21**, 196–216.

Franke, J. & Härdle, W. (1992). On bootstrapping kernel spectral estimates. *Annals of Statistics*, **20**, 121–145.

Friedman, J.H. (1988). Multivariate adaptive regression splines (with discussion). *Ann. Statist.*, **19**, 1–141.

Fuller, W.A. (1976). *Introduction to Statistical Time Series*. New York: Wiley.

Gallant, A.R. & Nychka, D.W. (1987). Seminonparametric maximum likelihood estimation. *Econometrica*, **55**, 363–390.

Gallant, A.R. & Tauchen, G.E. (1989). Seminonparametric estimation of conditional constrained heterogeneous processes: Asset pricing applications, *Econometrica*, **57**, 1091–1120.

Gasser, T. & Müller, H.G. (1979). Kernel estimation of regression functions. In *Smoothing Techniques for Curve Estimation*, eds. T. Gasser & M. Rosenblatt, 23–68.

Granger, C.W.J. & Hallman, G. (1991). Nonlinear transformations of integrated time series. *Journal of Time Series Analysis*, **12**, 207–224.

Granger, C.W.J. & Teräsvirta, T. (1993). *Modeling Nonlinear Economic Relationships*. Oxford: Oxford University Press.

Györfi, L., Härdle, W., Sarda, P. & Vieu, P. (1989). *Nonparametric Curve Estimation from Time Series*. Lecture Notes in Statistics 60. Heidelberg: Springer-Verlag.

Haggan, V. & Ozaki, T. (1981). Modeling nonlinear vibrations using an amplitude-dependent autoregressive time series model. *Biometrika*, **68**, 189–196.

Hall, P. & Hart. J.D. (1990a). Convergence rates in density estimation for data from infinite-order moving average processes. *Probability Theory and Related Fields*, **87**, 253–274.

Hall, P. & Hart. J.D. (1990b). Nonparametric regression with long-range dependence. *Stochastic Processes and their Applications*, **36**, 339–351.

Hall, P., Lihiri, S.N. & Truong, Y.K. (1994). On bandwidth choice for density estimation with dependent data. *Manuscript.*

Hall, P. & Patil, P. (1995). Formulae for mean integrated squared error of nonlinear wavelet-based density estimators. *Ann. Statist.* **23**, 905–928.
Hamilton, J.D. (1994). *Time Series Analysis*. Princeton: Princeton University Press.
Härdle, W. (1990). *Applied Nonparametric Regression*. Cambridge: Cambridge University Press.
Härdle, W. & Hall, P. (1993). On the backfitting algorithm for additive regression models. *Statistica Neerlandica*, **47**, 43–57.
Härdle, W., Klinke, S. & Turlach, B. (1995). *XploRe, an Interactive Statistical Computing Environment*. Heidelberg: Springer-Verlag.
Härdle, W. & Tsybakov, A.B. (1997). Local polynomial estimators of the volatility function. *J. Econometrics*, to appear.
Härdle, W., Tsybakov, A.B. & Yang, L. (1996). Nonparametric vector autoregression. *Journal of Statistical Planning and Inference*, to appear.
Härdle, W. & Vieu, P. (1992). Kernel regression smoothing of time series. *Journal of Time Series Analysis*, **13**, 209–232.
Hart, J.D. (1991). Kernel regression estimation with time series errors. *Journal of the Royal Statistical Society, Series B*, **53**, 173–187.
Hart, J.D. (1994). Automated kernel smoothing of dependent data by using time series cross-validation. *Journal of the Royal Statistical Society, Series B*, **56**, 529–542.
Hart, J.D. (1996). Some automated methods of smoothing time-dependent data. *J. Nonparametric Statistics*, **6**, 115–142.
Hart, J.D. & Vieu, P. (1990). Data-driven bandwidth choice for density estimation based on dependent data. *Annals of Statistics*, **18**, 873–890.
Hart, J.D. & Wehrly, T.E. (1986). Kernel regression estimation using repeated measurement data. *Journal of the American Statistical Association*, **81**, 1080–1088.
Hastie, T.J. & Tibshirani, R.J. (1990). *Generalized Additive Models*, Vol. 43 of *Monographs on Statistics and Applied Probability*. London: Chapman and Hall.
Hjellvik, V. & Tjøstheim, D. (1995). Nonparametric tests of linearity for time series. *Biometrika*, **82**, 351–368.
Hutchinson, J.M., Lo, A.W. & Poggio, T. (1994). A nonparametric approach to pricing and hedging derivative securities via learning networks. *Journal of Finance*, **49**, 851–889.
Johansen, S. (1988). Statistical analysis of cointegration vectors. *Journal of Economic Dynamics and Control*, **12**, 231–254.
Johansen, S. (1991). Estimation and hypothesis testing of cointegration vectors in Gaussian vector autoregressive models. *Econometrica*, **59**, 1551–1580.
Jones, D. A. (1978). Non-linear autoregressive processes. *Journal of the Royal Statistical Society, Series A*, **360**, 71–95.
Katkovnik, V. Y. (1979). *Linear and nonlinear methods for nonparametric regression analysis (in Russian)*. Avtomatika i Telemehanika, 35–46.
Kreiss, J.-P. (1987). On adaptive estimation in stationary ARMA processes. *Annals of Statistics*, **15**, 112–133.
Kreiss, J.-P. and Franke, J. (1992). Bootstrapping stationary autoregressive moving-average models. *Journal of Time Series Analysis*, **13**, 297–317.
Kuan, C.-M. & White, H. (1994). Artificial neural networks: An econometric perspective. *Econometric Reviews*, **13**, 1–91.
Künsch, H.R. (1989). The jackknife and the bootstrap for general stationary observations. *Annals of Statistics*, **17**, 1217–1241.
Kwiatkowski, D., Phillips, P.C.B., Schmidt, P. & Shin, Y. (1992). Testing the null hypothesis of stationarity against the alternative of a unit root. *Journal of Econometrics*, **54**, 159–178.
Lewis, R. & Reinsel, G.C. (1985). Prediction of multivariate time series by autoregressive model fitting. *Journal of Multivariate Analysis*, **16**, 393–411.
Lewis, P.A.W. & Stevens, G. (1991). Nonlinear modeling of time series using multivariate adaptive regression splines (MARS), *Journal of the American Statistical Association*, **87**, 864–877.
Linton, O. (1993). Adaptive estimation in ARCH models. *Econometric Theory*, **9**, 539–569.
Linton, O. & Nielsen, J.P. (1995). A kernel method of estimating structured nonparametric regression based on marginal integration. *Biometrika*, **82**, 93–100.
Liu, R.Y. and Singh, K. (1992). Moving blocks jackknife and bootstrap capture weak dependence. In *Exploring the Limits of Bootstrap*. R. Lepage & L. Billard (eds.). New York: Wiley, 225–248.
Lütkepohl, H. (1991). *Introduction to Multiple Time Series Analysis*. Berlin: Springer-Verlag.
Lütkepohl, H. & Poskitt, D.S. (1996). Testing for causation using infinite order vector autoregressive processes. *Econometric Theory*, **12**, 61–87.
Masry, E. & Tjøstheim, D. (1995a). Nonparametric estimation and identification of nonlinear ARCH time series: strong convergence and asymptotic normality. *Econometric Theory*, **11**, 258–289.
Masry, E. & Tjøstheim, D. (1995b). Additive nonlinear ARX time series and projection estimates. *Econometric Theory*, to appear.
Murata, N., Yoshizawa, S. & Amari, S. (1994). Network information criterion – determining the number of hidden units for an artificial neural network model. *IEEE Transactions on Neural Networks*, **5**, 865–871.
Park, B.U. & Marron, J.S. (1990). Comparison of data-driven bandwidth selectors. *Journal of the American Statistical Association*, **85**, 66–72.
Park, B.U. & Turlach, B. (1992). Practical performance of several data driven bandwidth selectors (with discussion). *Computational Statistics*, **7**, 251–270.
Parzen, E. (1961). Mathematical considerations in the estimation of spectra. *Technometrics*, **3**, 167–190.
Parzen, E. (1974). Some recent advances in time series modeling. *IEEE Transactions on Automatic Control*, **AC-19**, 723–730.
Parzen, E. (1977). Multiple time series: Determining the order of approximating autoregressive schemes. In *Multivariate Analysis–IV*, P.R. Krishnaiah (ed.). Amsterdam: North-Holland, 389–409.
Pham, D.T. (1985). Bilinear Markovian representations and bilinear models. *Stochastic Process. Appl.* **20**, 295–306.
Phillips, P.C.B. & Perron, P. (1988). Testing for a unit root in time series analysis. *Biometrika*, **75**, 335–346.

Prakasa Rao, B.L.S. (1983). *Nonparametric Functional Estimation*. Orlando, FL: Academic Press.
Priestley, M.B. (1981). *Spectral Analysis and Time Series*. London: Academic Press.
Priestley, M. B. (1988). *Non-linear and Non-stationary Time Series Analysis*, New York: Academic Press.
Priestley, M.B. (1996). Wavelets and time-dependent spectral analysis. *Journal of Time Series Analysis*, **17**, 85–103.
Robinson, P.M. (1983a). Review of various approaches to power spectrum estimation. In *Handbook of Statistics, Vol. 3* (D.R. Brillinger and P.R. Krishnaiah eds.), pp. 343–368. Amsterdam: North-Holland.
Robinson, P.M. (1983b). Non-parametric estimation for time series models. *Journal of Time Series Analysis*, **4**, 185–208.
Robinson, P.M. (1988). Semiparametric econometrics: A survey. *Journal of Applied Econometrics*, **3**, 35–51.
Rosenblatt, M. (1970). Density estimation and Markov sequences. In *Nonparametric Techniques in Statistical Inference*, (M.L. Puri, ed.) 199–213. Cambridge University Press.
Roussas, G.G. (1969). Nonparametric estimation in Markov process. *Annals of the Institute of Statistical Mathematics*, **21**, 73–87.
Rudemo, M. (1982). Empirical choice of histograms and kernel density estimators. *Scandinavian J. of Statist.*, **9**, 65–78.
Rumelhart, D.E., Hinton, G.E. & Williams, R.J. (1986). Learning internal representations by error propagation. In *Parallel Distributed Processing: Explorations in the Microstructures of Cognition*. D.E. Rumelhart & J.L. McClelland eds. Cambridge: M.I.T. Press, 1, pp. 318–362.
Said, S.E. & Dickey, D.A. (1984). Testing for unit roots in autoregressive-moving average models of unknown order. *Biometrika*, **71**, 599–607.
Saikkonen, P. & Luukkonen, R. (1997). Testing cointegration in infinite order vector autoregressive processes. *Journal of Econometrics*, forthcoming.
Schuster, A. (1898). On the investigation of hidden periodicities with application to a supposed 26-day period of meteorological phenomena. *Terr. Mag. Atmos. Elect.*, **3**, 13–41.
Sheather, S.J. (1983). A data-based algorithm for choosing the window width when estimating the density at a point. *Computational Statistics and Data Analysis*, **1**, 229–238.
Silverman, B. W. (1986). *Density Estimation for Statistics and Data Analysis*. London: Chapman and Hall.
Singh, R.S. & Ullah, A. (1985). Nonparametric time series estimation of joint DGP, conditional DGP and vector autoregression. *Econometric Theory*, **1**, 27–52.
Skaug, H.J. & Tjøstheim, D. (1993). Non-parametric tests of serial independence. In *The M. Priestley Birthday Volume* (ed. T. Subba Rao), pp. 207–229.
Steigerwald, D.G. (1992). Adaptive estimation in time series regression models. *Journal of Econometrics*, **54**, 251–275.
Stock, J.H. & M.W. Watson (1988). Testing for common trends. *Journal of the American Statistical Association*, **83**, 1097–1107.
Stone, C.J. (1977). Consistent nonparametric regression. *Annals of Statistics* **5**, 595–635.
Teräsvirta, T. (1994). Specification, estimation, and evaluation of smooth transition autoregressive models. *Journal of the American Statistical Association*, **89**, 208–218.
Tibshirani, R. (1988). Estimating transformations for regression via additivity and variance stabilization. *Journal of the American Statistical Association*, **83**, 194–405.
Tjøstheim, D. (1990). Nonlinear time series and Markov chains. *Advances in Applied Probability*, **22**, 587–611.
Tjøstheim, D. (1994). Nonlinear time series, a selective review. *Scand. J. Statist.* **21**, 97–130.
Tjøstheim, D. & Auestad, B. (1994a). Non-parametric identification of non-linear time series: projection. *Journal of the American Statistical Association*, **89**, 1398–1409.
Tjøstheim, D. & Auestad, B. (1994b). Non-parametric identification of non-linear time series: selecting significant lags. *Journal of the American Statistical Association*, **89**, 1410–1419.
Tong, H. (1983). *Threshold Models in Nonlinear Time Series Analysis. Lecture Notes in Statistics*. **Vol. 21**, Heidelberg: Springer.
Tong, H. (1990). *Nonlinear Time Series Analysis: A Dynamic Approach*. Oxford: Oxford University Press.
Truong, Y.K. (1993). A nonparametric framework for time series analysis. *New Directions in Time Series Analysis*. New York: Springer.
Truong, Y.K. & Patil, P. (1996). On estimating possibly discontinous regression involving stationary time series. *Manuscript*.
Tsay, R. (1989). Testing and modeling threshold autoregressive processes. *Journal of the American Statistical Association*, **84**, 231–240.
Tsybakov, A.B. (1986). Robust reconstruction of functions by the local approximation method. *Problems of Information Transmission*, **22**, 133–146.
Tukey, J.W. (1949). The sampling theory of power spectrum estimates. *Proceedings of the Symposium on Applications of Autocorrelation Analysis to Physical Problems*, NAVEXOS-P-735, 47–67, Washington: Office of Naval Research.
Tweedie, R. L. (1975). Sufficient Conditions for Ergodicity and Recurrence of Markov Chain on a General State Space. *Stochastic Processes and their Applications*, **3**, 385–403.
Weigend, A.S. & Nix, D. (1994). Predictions with confidence intervals (local error bars). Discussion Paper No. 34, Sonderforschungsbereich 373, Humboldt-Universität zu Berlin.
Werker, B.J.M. (1995). *Statistical Methods in Financial Econometrics*. CentER, Tilburg University.
Wong, C.-M. & Kohn, R. (1996). A Bayesian approach to estimating and forecasting additive nonparametric autoregressive models. *Journal of Time Series Analysis*, **17**, 203–220.

International Statistical Review, 12, 153-172

Résumé

Beaucoup des elements des séries temporelles sont analysable par des methodes non-paramétriques. L'objet d'interêt a une forme generale qui est approximée plus et plus précisément le nombre d'obervations augmente. Cet article présente un survey des procédures non paramétriques en analyse des séries temporelles. Nous illustrons au moyen d'exemples portant sur l'estimation de densité, sur le bootstrap et l'estimation de tendence.

[*Received August 1996, accepted November 1996*]

(1997) Härdle, W. , Lütkepohl, H. and Chen, R.
Nonparametric Time Series Analysis.

Semiparametric Single Index Versus Fixed Link Function Modelling

W. Härdle V. Spokoiny
S. Sperlich

Humboldt Universität SFB 373, Spandauer Str. 1, 10178 Berlin
Weierstraß Institute for Applied Analysis and Stochastics, Mohrenstr. 39, 10117 Berlin

1997

Abstract

Discrete choice models are frequently used in statistical and econometric practice. Standard models such as logit models are based on exact knowledge of the form of the link and linear index function. Semiparametric models avoid possible misspecification but often introduce a computational burden. It is therefore interesting to decide between approaches. Here we propose a test of semiparametric versus parametric single index modelling. Our procedure allows that the (linear) index of the semiparametric alternative is different from that of the parametric hypothesis. The test is proved to be rate-optimal in the sense that it provides the (rate) minimal distance between hypothesis and alternative for a given power function. [1]

1 Introduction

Discrete choice models are frequently used in statistical and econometric applications. Among them binary response models, such as Probit or Logit regression, dominate the applied literature. A basic hypothesis made there is that the link and the index function have a known form, see McCullagh and Nelder (1989). The fixed form of the link function e.g. the logistic cdf is rarely justified by the context of the observed data but is often motivated by numerical convenience and by reference to "standard practice", say "accessible canned software".

Recent theoretical and practical studies have questioned this somewhat rigid approach and have proposed a more flexible semiparametric approach. Green and Silverman (1994) use the theory of penalizied likelihood to model nonparametric link functions with splines. Horowitz (1993) gives an excellent survey on single index methods and stresses economic applications. Staniswalis and Severini (1994) use kernel methods and keep a fixed link function but allow the index to be of partial linear form. Partial linear models are semiparametric models with a parametric linear

[1] We thank O. Lepski for helpful discussion and comments.
The research was supported by Deutsche Forschungsgemeinschaft, SFB 373.
AMS *1994 subject classifications.* Primary 62G10, 62H40; secondary 62G20, 62P20.
Key words and phrases: Semiparametric models, single index model, hypothesis testing.

and a nonparametric index and have been studied by Rice (1986), Speckman (1988) and Engle, Granger, Rice and Weiss (1986).

These models enhance the class of Generalized Linear Models (McCullagh and Nelder, 1989) in several ways. Here we concentrate on one generalization, the single index models with link function of unknown nonparametric form but (linear) index function. The advantage of this approach is that still an interpretable linear single index, a weighted sum of the predictor variables, is produced. The link function plays in theoretical justifications of single index models via stochastic utility functions an important role (Maddala, 1983): it is the cdf of the errors in a latent variable model. Our approach enables us to interpret the results still in terms of a stochastic utility model but enhances it by allowing for an unknown cdf of the errors.

Despite the gained flexibility in semiparametric regression modelling there is still an important gap between theory and practice, namely a device for testing between a parametric and semiparametric alternative. A first paper in bridging this gap is Horowitz and Härdle (1994). They considered for response Y and predictor X the parametric null hypothesis

$$H_0 : Y = F(X^\top \theta_0) + \varepsilon \tag{1}$$

where $x^\top \theta$ denotes the index and F is the fixed and known link function. The semiparametric alternative considered there is that the regression function has the form $f(x^\top \theta_0)$ with a nonparametric link function f and the same index $x^\top \theta_0$ as under H_0. The main drawback of that paper is that the index is supposed to be the same under the null and the alternative.

The goal of the present paper is to construct a test which has power for as large class of alternatives. We move to a full semiparametric alternative by considering alternatives of single index type

$$H_1 : Y = f(X^\top \beta) + \varepsilon \tag{2}$$

with β possibly different from θ_0. The situation of our test is illustrated in the following figures 1 and 2.

Figure 1: Parametric fitting

Figure 2: Semiparametric fitting

The data is a crosssection of 462 records on apprenticeship of the German Social Economic Panel from 1984 to 1992. The dependent variable is an indicator of unemployement, ($Y = 1$ =yes). Explanatory variables are X_1 gross monthly earnings as an apprentice, X_2 percentage of people apprenticed in a certain occupation, divided by the people employed in this occupation in the entire economy and X_3 unemployment rate in the state the respondent lived in during the year the apprenticeship was completed. The aim of the test is to decide between the logit model and

the semiparametric model with unknown link function and possibly different index. In Härdle, Klinke and Turlach (1995) this hypothesis is tested with the Horowitz-Härdle (HH) -test by Proenca and Werwatz who also prepared the dataset. They give a more detailled description of the HH-test procedure which does not reject.

We measure the quality of a test by the value of minimal distance between the regression function under the null and under the alternative which is sufficient to provide the desirable power of testing. The test proposed below is shown to be rate-optimal in this sense. The paper is organized as follows. The next section contains the main results then we present the test procedure. In Section 5 we present some simulation study. The proof of main results are given in Section 3 (Theorem 2.2) and in the Appendix (Theorem 2.1).

2 Main Results

We start with a brief historical background of the nonparametric hypothesis testing problem. The problem for the case of a simple hypothesis and univariate nonparametric alternative was considered by Ibragimov and Khasminskii (1977) and Ingster (1982). It was shown that the minimax rate for the distance between the null and the alternative set is of the order $n^{-2s/(4s+1)}$ where s is a measure of smoothness. Note that this rate differs from that of an estimation problem where we have $n^{-s/(2s+1)}$. In the multivariate case the corresponding rate changes to $n^{-2s/(4s+d)}$, as Ingster (1993) has shown. The problem of testing a parametric hypothesis versus a nonparametric alternative was discussed also in Härdle and Mammen (1993). Their results allow to extract the above minimax rate.

The results of Friedman and Stuetzle (1981), Huber (1985), Hall (1989) and Golubev (1992) show that estimation of the function f under (2) can be made with the rate corresponding to the univariate case. Below we will see though that for the problem of hypothesis testing the situation is slightly different. The rate for this additive alternative of single index type differs from that of a univariate alternative ($d = 1$) by an extra log-factor. Nevertheless, we have almost a univariate rate and we can therefore still expect efficiency of the test for practical applications.

We will come back to the introductory example in section 5. Suppose we are given independent observations (X_i, Y_i), $X_i \in I\!R^d$, $Y_i \in I\!R^1$, $i = 1, \ldots, n$, that follow the regression

$$Y_i = F(X_i) + \varepsilon_i, \qquad i = 1, \ldots, n. \tag{3}$$

Here $\varepsilon_i = Y_i - F(X_i)$ are mean zero error variables,

$$\mathbf{E}\varepsilon_i = 0, \qquad i = 1, \ldots, n,$$

with conditional variance

$$\sigma_i^2 = \mathbf{E}\left[\varepsilon_i^2 \mid X_i\right], \qquad i = 1, \ldots, n. \tag{4}$$

Example **2.1** As a first example take the above single index binary choice model. The observed response variables Y_i take two values $0, 1$ and

$$\begin{aligned} \mathbf{P}\left(Y_i = 1 \mid X_i\right) &= F(X_i), \\ \mathbf{P}\left(Y_i = 0 \mid X_i\right) &= 1 - F(X_i). \end{aligned}$$

In this case $\sigma_i^2 = F(X_i)\left\{1 - F(X_i)\right\}$.

Example **2.2** A second example is a nonlinear regression model with unknown transformation. An excellent introduction into nonlinear regression can be found in Huet, Jolivet and Messeau (1993). The model takes the same form as (1) but the response Y is not necessarily binary and the variance σ_i^2 may be an unknown function of the $F(X_i)$'s. Carroll and Ruppert (1988) use this kind of error structure to model fan shaped residual structure.

We wish to test the hypothesis H_0 that the regression function $F(x)$ belongs to a prescribed parametric family $(F_\theta(x),\, \theta \in \Theta)$, where Θ is a subset in a finite-dimensional space $\mathbb{R}^m$. This hypothesis is tested versus the semiparametric alternative H_1 that the regression function $F(\cdot)$ is of the form

$$F(x) = f(x^\top \beta) \tag{5}$$

where β is a vector in $\mathbb{R}^d$ with $|\beta| = 1$, and $f(\cdot)$ is a univariate function.

Example **2.3** Let the parametric family $(F_\theta(x), \theta \in \Theta)$ be of the form

$$F_\theta(x) = \frac{1}{1 + \exp(-x^\top \theta)} \tag{6}$$

and let otherwise (X, Y) have stochastic structure as in Example 2.1. This form of parametrization leads to a binary choice logit regression model. Probit or complementary log-log models have a different parametrization but still have this single index form.

Let $\mathcal{F}_0$ be the set of functions $(F_\theta(x),\, \theta \in \Theta)$ and let $\mathcal{F}_1$ be a set of alternatives of the form (5). We measure the power of a test φ_n by its power function on the sets $\mathcal{F}_0$ and $\mathcal{F}_1$: if $\varphi_n = 0$ then we accept the hypothesis H_0 and if $\varphi_n = 1$, then we accept H_1. The corresponding first and second type error probabilities are defined as usual:

$$\alpha_0(\varphi_n) = \sup_{F \in \mathcal{F}_0} \mathbf{P}_F\left(\varphi_n = 1\right),$$

$$\alpha_1(\varphi_n) = \sup_{F \in \mathcal{F}_1} \mathbf{P}_F\left(\varphi_n = 0\right).$$

Here $\mathbf{P}_F$ means the distributions of observations (X_i, Y_i) given the regression function $F(\cdot)$. When there is no risk of confusion we write $\mathbf{P}$ instead of $\mathbf{P}_F$. Our goal is to construct a test φ_n that has power over a wide class of alternatives. The assumptions needed are made precise below. We start with assumptions on the error distribution.

$(E1)$ The errors ε_i are bounded by a universal constant C_ε

$$\varepsilon_i \le C_\varepsilon\,, \qquad i = 1, \ldots, n.$$

$(E2)$ The conditional distributions of errors ε_i given X_i depend only on values of the regression function $F(X_i)$,

$$\mathcal{L}\left(\varepsilon_i \mid X_i\right) = \mathcal{L}\left(\varepsilon_i \mid F(X_i)\right) = P_{F(X_i)}$$

where (P_z) is a prescribed distribution family of one-dimensional parameter z;

$(E3)$ The variance function $\sigma^2(z) = \mathbf{E}\left[\varepsilon_i^2 \mid F(X_i) = z\right]$ and the fourth central moment function $\kappa^4(z) = \mathbf{E}\left[\left(\varepsilon_i^2 - \mathbf{E}\varepsilon_i^2\right)^2 \mid F(X_i) = z\right]$ are separated away from zero and infinity i.e.

$$0 < \sigma_* \le \sigma(z) \le \sigma^* < \infty$$

$$0 < \kappa_* \le \kappa(z) \le \kappa^* < \infty$$

with some prescribed $\sigma_*, \sigma^*, \kappa_*, \kappa^*$, and this function is uniformly continuous: for some positive constants C_σ and C_κ one has

$$|\sigma(z) - \sigma(z')| \le C_\sigma |z - z'|,$$
$$|\kappa(z) - \kappa(z')| \le C_\kappa |z - z'|.$$

Note that $(E1)$ is obviously fulfilled for the single index model in Examples 2.1 and 2.3. In the more general situation of Example 2.2 this assumption can be weakened to the existence of exponential moments for ε_i.

The assumption $(E3)$ restricts the set of X-observations to a bounded set. It is made more precise in the following assumption on the design X.

(D) The predictor variables X have a design density $\pi(x)$ which is supported on the compact convex set $\mathcal{X}$ in $\mathbb{R}^d$ and is separated from zero and infinity on $\mathcal{X}$;

Assumption (D) is quite common in nonparametric regression analysis. It is apparently fulfilled for the above example on apprenticeship and youth unemployment. We now specify the hypothesis and alternative.

$(H0)$ The parameter set Θ is a compact subset in $\mathbb{R}^m$.

For some universal constant C_Θ the following holds

$$|F_\theta(x) - F_{\theta'}(x)| \le C_\Theta |\theta - \theta'|, \qquad \forall x \in \mathcal{X},\ \theta, \theta' \in \Theta;$$

All functions $F_\theta(\cdot)$ belong to the Hölder class $\Sigma_d(s, L)$ of functions in $\mathbb{R}^d$.

$(H1)$ The univariate link function $f(\cdot)$ from (5) belongs to the Hölder class $\Sigma(s, L)$. The function $F(x) = f(x^\top \beta)$ is separated away from the parametric family $\mathcal{F}_0$ i.e.

$$\inf_{\theta \in \Theta} \|F - F_\theta\| \ge c_n \tag{7}$$

with a given $c_n > 0$. Here $\|F - F_\theta\| = \int |F(x) - F_\theta(x)|^2 \pi(x)\, dx$.

For the definition of a Hölder smoothness class in the context of statistical nonparametric problems we refer e.g. to Ibragimov and Khasminskii (1981). Assumption $(H0)$ is certainly fulfilled for Example 2.3 but also in Probit and other generalized linear regression models such as the log linear models.

The main results are given below. We compute first the optimal rate of convergence of the distance c_n distinguishing the null from the alternative. The second theorem states the existence of an optimal test. The test will be given more explicitly in the next section where we also apply it to the above concrete examples. Theorem 2.2 is proved in Section 4 and the proof of Theorem 2.1 is given in the appendix.

Theorem 2.1 *Let* $c_n = \left(a \frac{\sqrt{\ln n}}{n}\right)^{\frac{2s}{4s+1}}$. *If* a *is small enough then for any sequence of tests* φ_n *one has*

$$\liminf_{n \to \infty} \alpha_0(\varphi_n) + \alpha_1(\varphi_n) \ge 1.$$

Theorem 2.2 *For any constant a^* large enough there is a sequence of tests φ_n^* which distinguish consistently the hypothesis H_0 versus alternative $H_1 = H_1(c_n^*)$ with $c_n^* = \left(a^* \frac{\sqrt{\ln n}}{n}\right)^{\frac{2s}{4s+1}}$ i.e.*

$$\lim_{n\to\infty} \alpha_0(\varphi_n^*) = 0$$

and

$$\lim_{n\to\infty} \alpha_1(\varphi_n^*) = 0.$$

3 The test procedure

Before we describe the test procedure let us introduce some notation. Given functions $F(x)$ and $G(x)$ we denote by

$$\langle F, G\rangle = \frac{1}{n}\sum_{i=1}^{n} F(X_i)\, G(X_i). \tag{8}$$

the scalar product of the functions F and G. We write also $\langle F\rangle$ instead of $\langle F, F\rangle$ and identify the sequences $(Y_i), (\varepsilon_i)$ with the functions $Y(X_i)$ and $\varepsilon(X_i)$. We construct the tests φ_n^* from Theorem 2.2 in several steps.

First we shall do a preliminary pilot estimation $\tilde{F}_0$ under the null. Second we estimate the d-dimensional nonparametric regression $\tilde{F}_1$ necessary to construct estimators of expected value and the variance of the proposed test statistics. In the third step we estimate for each feasible value of β the corresponding link function f under $(H1)$ as in (2). Finally we compute the test statistic based on comparison of residuals under H_0 and H_1.

3.1 Parametric pilot estimation

Let Θ_n be a grid in the parametric set Θ with the step $\frac{\ln n}{\sqrt{n}}$. Put

$$\tilde{\theta}_n = \operatorname*{arginf}_{\theta\,\in\,\Theta_n} \langle Y - F_\theta\rangle = \operatorname*{arginf}_{\theta\,\in\,\Theta_n} \frac{1}{n}\sum_{i=1}^{n} |Y_i - F_\theta(X_i)|^2 . \tag{9}$$

Denote also

$$\tilde{F}_0(\cdot) = F_{\tilde{\theta}_n}(\cdot). \tag{10}$$

Note that $\tilde{\theta}_n$ is not necessarily an efficient estimator under the null since we do not correct for the variance function.

3.2 Nonparametric pilot estimation

For the nonparametric estimation of the expected value and the variance of the test statistic we shall use the standard kernel technique, see e.g. Härdle (1990) or Müller (1987). More precisely we use a one dimensional kernel satisfying the conditions

$(K1)$ $K(\cdot)$ is compactly supported;

$(K2)$ $K(\cdot)$ is symmetric;

$(K3)$ $K(\cdot)$ has s continuous derivatives;

$(K4)$ $\int K(t)\,dt = 1$;

$(K5)$ $\int K(t)\,t^k dt = 0, \quad k = 1, \ldots, s-1.$

Recall from $(H0)$ and $(H1)$ that s denotes the degree of smoothness of the regression function. Note also that $(K5)$ ensures that K is orthogonal to polynomials of order 1 to $s-1$. For a list of kernels satisfying $(K1)-(K5)$ we refer to Müller (1987). A d-dimensional product kernel K_1 is defined as

$$K_1(u_1, \ldots, u_d) = \prod_{j=1}^{d} K(u_j). \tag{11}$$

Take now

$$h_1 = n^{-\frac{1}{2s+d}}, \tag{12}$$

the optimal smoothing bandwidth in d-dimensions, and put

$$\tilde{F}_1(x) = \frac{\sum_{i=1}^n Y_i K_1\left(\frac{x-X_i}{h_1}\right)}{\sum_{i=1}^n K_1\left(\frac{x-X_i}{h_1}\right)}. \tag{13}$$

The nonparametric kernel smoother $\tilde{F}_1$ is the well known multidimensional Nadaraya-Watson kernel estimator.

3.3 Estimation under H_1

Set

$$h = \left(\frac{\sqrt{\ln n}}{n}\right)^{\frac{2}{4s+1}}. \tag{14}$$

We will use this bandwidth for estimation in the semiparametric model. Note that in (12) for the nonparametric estimation problem another rate, namely $n^{-1/(2s+d)}$ was used. Here we have almost this bandwidth except for the extra log-term.

Let S_d be the unit sphere in $\mathbb{R}^d$. Denote by $S_{n,d}$ a discrete grid in S_d with the step $b_n = h^{2s+2}$. Let N be the cardinality of $S_{n,d}$

$$N = \#S_{n,d}. \tag{15}$$

For each $\beta \in S_{n,d}$ define

$$K_{h,\beta}(x) = K\left(\frac{x^\top \beta}{h}\right), \quad x \in \mathbb{R}^d, \tag{16}$$

and introduce the smoothing operator $\mathcal{K}_\beta$ with

$$\mathcal{K}_\beta Y(X_i) = \Pi_\beta(X_i) \sum_{j \neq i} Y_j K_{h,\beta}(X_i - X_j) \tag{17}$$

where

$$\Pi_\beta(X_i) = \left(\sum_{j \neq i} K_{h,\beta}(X_i - X_j)\right)^{-1}. \tag{18}$$

Similarly we define $\mathcal{K}_\beta \varepsilon$ and $\mathcal{K}_\beta F$. Note that given β the values $\mathcal{K}_\beta Y$ estimate f in (2).

3.4 The test statistic

Now for each β we calculate a statistic T_β as follows:

$$T_\beta = \frac{n\sqrt{h}}{\tilde{V}_\beta}\left[2\left\langle Y-\tilde{F}_0, \mathcal{K}_\beta Y-\tilde{F}_0\right\rangle - \left\langle \mathcal{K}_\beta Y-\tilde{F}_0\right\rangle + \tilde{E}_\beta\right]. \tag{19}$$

Here $\langle\cdot\rangle$ is defined by (8), h by (14), $\tilde{F}_0$ by (10). We use the following notation

$$\tilde{E}_\beta = \frac{1}{n}\sum_i \sum_{j\neq i} \tilde{\sigma}_j^2\, \Pi_\beta^2(X_i)\, K_{h,\beta}^2(X_i - X_j) \tag{20}$$

where $\Pi_\beta(X_i)$ is from (18),

$$\tilde{\sigma}_j^2 = \sigma^2\left(\tilde{F}_1(X_j)\right), \quad j=1,\ldots,n, \tag{21}$$

the function $\sigma^2(\cdot)$ being defined in the model assumptions and $\tilde{F}_1(x)$ being the nonparametric pilot estimator. Finally,

$$\begin{aligned}\tilde{V}_\beta^2 &= h\sum_i\sum_{j\neq i}\tilde{\sigma}_i^2\,\tilde{\sigma}_j^2\,\Pi_\beta^2(X_i)\left|2K_{h,\beta}(X_i-X_j) - K_{h,\beta}^{(2)}(X_i,X_j)\right|^2 + \\ &\quad + h\sum_i \tilde{\kappa}_i^4\left|\sum_{j\neq i}\Pi_\beta^2(X_j)K_{h,\beta}^2(X_i-X_j)\right|^2\end{aligned}$$

with $\tilde{\kappa}_i = \kappa\left(\tilde{F}_1(X_i)\right), \quad i=1,\ldots,n,$ $\kappa(\cdot)$ being from $(E3)$ and

$$K_{h,\beta}^{(2)}(X_i,X_j) = \frac{1}{\Pi_\beta(X_i)}\sum_{k\neq i,j}\Pi_\beta^2(X_k)\,K_{h,\beta}(X_k-X_i)\,K_{h,\beta}(X_k-X_j). \tag{22}$$

Put now

$$T_n^* = \sup_{\beta\in S_{n,d}} T_\beta \tag{23}$$

and

$$\varphi_n^* = \mathbf{1}\left(T_n^* > \sqrt{(2+\delta)\log N}\right). \tag{24}$$

Here $\mathbf{1}(\cdot)$ is the indicator function of the corresponding event, δ is an arbitrary small positive number and N is the cardinality of $S_{n,d}$, see (15).

4 Proof of Theorem 2

We start with the decomposition of the test statistics T_β. Denote by $B_\beta(x)$ the bias function for the smoothing operator $\mathcal{K}_\beta$ from (17):

$$B_\beta(X_i) = \mathcal{K}_\beta F(X_i) - F(X_i), \quad i=1,\ldots,n. \tag{25}$$

Fix some $\beta\in S_{n,d}$ and $F\in\mathcal{F}_0\cup\mathcal{F}_1$.

Lemma 4.1

$$
\begin{aligned}
T_\beta &= \frac{n\sqrt{h}}{\tilde{V}_\beta}\Big[\Big\langle F-\tilde{F}_0\Big\rangle - \langle B_\beta\rangle + \\
&\quad + 2\langle \mathcal{K}_\beta\varepsilon,\varepsilon\rangle - \langle \mathcal{K}_\beta\varepsilon\rangle + \tilde{E}_\beta + \\
&\quad + 2\Big\langle F-\tilde{F}_0,\varepsilon\Big\rangle + 2\langle B_\beta,\varepsilon\rangle - 2\langle B_\beta,\mathcal{K}_\beta\varepsilon\rangle\Big]. \qquad (26)
\end{aligned}
$$

Proof. By definition $Y = F + \varepsilon$ and therefore

$$\mathcal{K}_\beta Y = \mathcal{K}_\beta F + \mathcal{K}_\beta\varepsilon = F + B_\beta + \mathcal{K}_\beta\varepsilon.$$

Now

$$
\begin{aligned}
2\Big\langle Y-\tilde{F}_0, \mathcal{K}_\beta Y-\tilde{F}_0\Big\rangle &= 2\Big\langle F-\tilde{F}_0+\varepsilon, F-\tilde{F}_0+B_\beta+\mathcal{K}_\beta\varepsilon\Big\rangle = \\
&= 2\Big\langle F-\tilde{F}_0\Big\rangle + 2\Big\langle F-\tilde{F}_0, B_\beta\Big\rangle + 2\Big\langle F-\tilde{F}_0,\mathcal{K}_\beta\varepsilon\Big\rangle + \\
&\quad + 2\Big\langle \varepsilon, F-\tilde{F}_0\Big\rangle + 2\langle\varepsilon, B_\beta\rangle + 2\langle\varepsilon,\mathcal{K}_\beta\varepsilon\rangle
\end{aligned}
$$

and

$$
\begin{aligned}
\Big\langle \mathcal{K}_\beta Y-\tilde{F}_0\Big\rangle &= \Big\langle F-\tilde{F}_0+B_\beta+\mathcal{K}_\beta\varepsilon\Big\rangle = \\
&= \Big\langle F-\tilde{F}_0\Big\rangle + \langle B_\beta\rangle + \langle\mathcal{K}_\beta\varepsilon\rangle + \\
&\quad + 2\Big\langle F-\tilde{F}_0, B_\beta\Big\rangle + 2\Big\langle F-\tilde{F}_0,\mathcal{K}_\beta\varepsilon\Big\rangle + 2\langle B_\beta,\mathcal{K}_\beta\varepsilon\rangle.
\end{aligned}
$$

Substituting this in the definition of T_β we obtain the assertion of the lemma.

The next step is to show that the expansion (26) for the statistic T_β can be simplified by discarding lower order terms. Indeed we shall see below that the last three terms are relatively small and can be omitted. The terms $\tilde{E}_\beta$ and $\tilde{V}_\beta$ can be substituted by similar expressions E_β and V_β which use "true" values σ_i and κ_i instead of estimated values $\tilde{\sigma}_i$ and $\tilde{\kappa}_i$ and finally, the parametric estimator $\tilde{\theta}_n$ can be replaced by θ_n defined by

$$\theta_n = \operatorname*{arginf}_{\theta\in\Theta_n}\langle F-F_\theta\rangle \qquad (27)$$

where F is a "true" regression function from (3). Suppose that all these replacements can be done. Define now

$$
\begin{aligned}
T'_\beta &= \frac{n\sqrt{h}}{V_\beta}\left[\langle F-F_{\theta_n}\rangle - \langle B_\beta\rangle + \right. \\
&\quad \left. + \ 2\langle\mathcal{K}_\beta\varepsilon,\varepsilon\rangle - \langle\mathcal{K}_\beta\varepsilon\rangle + E_\beta\right]
\end{aligned}
$$

with

$$
\begin{aligned}
E_\beta &= \frac{1}{n}\sum_i\sum_{j\neq i}\sigma_j^2\,\Pi_\beta^2(X_i)\,K_{h,\beta}^2(X_i-X_j), \\
V_\beta^2 &= h\sum_i\sum_{j\neq i}\sigma_i^2\,\sigma_j^2\,\Pi_\beta^2(X_i)\left|2K_{h,\beta}(X_i-X_j)-K_{h,\beta}^{(2)}(X_i,X_j)\right|^2 + \\
&\quad + h\sum_i\kappa_i^4\left|\sum_{j\neq i}\Pi_\beta^2(X_j)K_{h,\beta}^2(X_i-X_j)\right|^2.
\end{aligned}
$$

Below we show that the tests φ_n^{**} based on the statistics T_n^{**} with

$$T_n^{**} = \sup_{\beta \in S_{n,d}} T'_\beta \tag{28}$$

have the same asymptotic behavior as φ_n^*. For the moment we only consider the tests φ_n^{**}. Note that they are not tests in the usual sense since they use the non-observable values $E_\beta, V_\beta, \theta_n$. Central to our proof is the analysis of the asymptotic behavior of the random variables

$$\xi_\beta = n\sqrt{h}\,\left[2\,\langle \mathcal{K}_\beta \varepsilon, \varepsilon\rangle - \langle \mathcal{K}_\beta \varepsilon\rangle + E_\beta\right]. \tag{29}$$

Lemma 4.2 *The following assertions hold*

$$\mathbf{E}\,\xi_\beta = 0, \tag{30}$$

$$\mathbf{E}\,\xi_\beta^2 = V_\beta^2\,, \tag{31}$$

and uniformly in $F \in \mathcal{F}_0 \cup \mathcal{F}_1\,,\ \beta \in S_{n,d}$ *and* $t \in [-\ln n, \ln n]$

$$\frac{\mathbf{P}\left(\frac{\xi_\beta}{V_\beta} > t\right)}{1-\Phi(t)} \to 1, \quad n \to \infty, \tag{32}$$

$\Phi(\cdot)$ *being the standard normal distribution.*

Proof. The first two statements are derived by direct calculation. In fact, by definition and (22)

$$\begin{aligned}
\xi_\beta &= 2\sqrt{h}\sum_i \varepsilon_i\,\Pi_\beta(X_i)\sum_{j\neq i}\varepsilon_j\ K_{h,\beta}(X_i - X_j) - \\
&\quad -\sqrt{h}\sum_i\ \Pi_\beta^2(X_i)\left|\sum_{j\neq i}\varepsilon_j\ K_{h,\beta}(X_i-X_j)\right|^2 + \\
&\quad + \sqrt{h}\sum_i\sum_{j\neq i}\sigma_j^2\,\Pi_\beta^2(X_i)\,K_{h,\beta}^2(X_i-X_j) = \\
&= \sqrt{h}\sum_i\sum_{j\neq i}\varepsilon_i\,\varepsilon_j\,\Pi_\beta(X_i)\left[2K_{h,\beta}(X_i-X_j) - K_{h,\beta}^{(2)}(X_i,X_j)\right] + \\
&\quad + \sqrt{h}\sum_i\sum_{j\neq i}(\sigma_j^2-\varepsilon_j^2)\,\Pi_\beta^2(X_i)\,K_{h,\beta}^2(X_i-X_j).
\end{aligned}$$

Since the errors ε_i are independent and $\mathbf{E}\,\varepsilon_i = 0$, $\mathbf{E}\,\varepsilon_i^2 = \sigma_i^2$, we immediately obtain (30) and (31). The last statement (32) is a particular case of the general central limit theorem for quadratic forms of independent random variables and can be obtained in a standard way by calculation of the corresponding cumulants. We omit the details, see e.g. Härdle and Mammen (1993).

The assertion (32) of Lemma 4.2 straightforwardly implies the following corollary.

Lemma 4.3 *Uniformly in* $F \in \mathcal{F}_0 \cup \mathcal{F}_1$ *one has*

$$\mathbf{P}\left(\sup_{\beta\in S_{n,d}}\frac{\xi_\beta}{V_\beta} > \sqrt{(2+\delta)\ln N}\right) \to 0, \quad n\to\infty. \tag{33}$$

Proof. For any t one gets

$$\mathbf{P}\left(\sup_{\beta\in S_{n,d}}\frac{\xi_\beta}{V_\beta}>t\right)\le\sum_{\beta\in S_{n,d}}\mathbf{P}\left(\frac{\xi_\beta}{V_\beta}>t\right)\le N\sup_{\beta\in S_{n,d}}\mathbf{P}\left(\frac{\xi_\beta}{V_\beta}>t\right). \tag{34}$$

But through (32) for n large enough

$$\mathbf{P}\left(\frac{\xi_\beta}{V_\beta}>\sqrt{(2+\delta)\ln N}\right)\le 2\left(1-\Phi\left(\sqrt{(2+\delta)\ln N}\right)\right)\le$$
$$\le\quad\exp\left\{-\frac{1}{2}\left|\sqrt{(2+\delta)\ln N}\right|^2\right\}=N^{-1-\delta/2}$$

that implies (33) through (34).

Now we come to the calculation of the error probabilities for the tests φ_n^{**} based on T_n^{**}. Under the hypothesis H_0 one has $F=F_\theta\,,\theta\in\Theta$. This does not automatically yield $\langle F-F_{\theta_n}\rangle=0$ since $\theta_n\in\Theta_n$, see (27), and θ can be outside Θ_n. But the assumptions $(H0)$ on the parametric family guarantee that this value is small enough.

Lemma 4.4 *Let $F=F_\theta\,,\theta\in\Theta$. Then*

$$\langle F_\theta-F_{\theta_n}\rangle\le C_\theta^2\frac{\ln^2 n}{n}.$$

Proof. Let

$$\theta_n'=\operatorname*{arginf}_{\theta'\,\in\,\Theta_n}\left|\theta-\theta'\right|.$$

The definition of the grid Θ_n provides $|\theta-\theta_n'|^2\le\frac{\ln^2 n}{n}$. Now from the definition of θ_n and the assumptions $(H0)$ on the parametric family we obtain

$$\langle F_\theta-F_{\theta_n}\rangle\le\langle F_\theta-F_{\theta_n'}\rangle=\frac{1}{n}\sum_i\left|F_\theta(X_i)-F_{\theta_n'}(X_i)\right|^2\le C_\Theta^2\left|\theta-\theta'\right|^2\le C_\Theta^2\frac{\ln^2 n}{n}.$$

Using this result we have for $F=F_\theta$ by Lemma 4.3

$$\mathbf{P}\left(T_n^{**}>\sqrt{(2+\delta)\ln N}\right)\le$$
$$\le\quad\mathbf{P}\left(\sup_{\beta\in S_{n,d}}\frac{\xi_\beta}{V_\beta}>\sqrt{(2+\delta)\ln N}-C_\theta^2\frac{\ln^2 n}{n}n\sqrt{h}\right)\to 0,\quad n\to\infty,$$

i.e.

$$\alpha_0(\varphi_n^{**})=\sup_{F\in\mathcal{F}_0}\mathbf{P}_F\left(\varphi_n^{**}=1\right)\to 0,\quad n\to\infty.$$

Next we evaluate the error probability of the second type .

Lemma 4.5 *Let $F\in\mathcal{F}_1$. Then for n large enough*

$$\langle F-F_{\theta_n}\rangle\ge c_n/2\,.$$

Proof. Let $F \in \mathcal{F}_1$ be fixed and

$$\theta_F = \operatorname*{arginf}_{\theta \in \Theta} \|F - F_\theta\| .$$

By the triangle inequality and Lemma 4.4 one has

$$\langle F - F_{\theta_F} \rangle \le \langle F - F_{\theta_n} \rangle + \langle F_{\theta_n} - F_{\theta_F} \rangle \le \langle F - F_{\theta_n} \rangle + C_\theta^2 \frac{\ln^2 n}{n}.$$

It remains to check that the inequality $\|F - F_{\theta_F}\| \ge c_n$ implies $\langle F - F_{\theta_F} \rangle \ge c_n/2$. For n large enough that is obviously the case.

The following Lemma is a direct consequence of assumptions $(E3)$ and (D).

Lemma 4.6 *There exist constants $C_\pi\,, \sigma^*$ and V^* such that*

$$|\Pi_\beta(X_i) K_{h,\beta}(X_i - X_j)| \le C_\pi\, |\Pi_\beta(X_j) K_{h,\beta}(X_i - X_j)| \qquad \forall \beta, X_i, X_j \tag{35}$$

$$\sup_i \sigma_i \le \sigma^*. \tag{36}$$

and

$$\sup_\beta V_\beta \le V^*. \tag{37}$$

Recall now that each function $F(\cdot)$ from $\mathcal{F}_1$ is of the form $F(x) = f(x^\top \beta_0)$ with some $\beta_0 \in S_d$. As a consequence $F(\cdot)$ should be well approximated by the smoothing operator $\mathcal{K}_\beta$ with β coinciding or close to β_0. More precise, the following can be stated.

Lemma 4.7 *There is a positive constant C_b such that for each $F(\cdot) \in \mathcal{F}_1$, $F(x) = f(x^\top \beta_0)$,*

$$\langle B_{\beta_n} \rangle \le C_b h^{2s} \tag{38}$$

with

$$\beta_n = \operatorname*{\mathit{arginf}}_{\beta \in S_{n,d}} |\beta - \beta_0| . \tag{39}$$

Proof. The definition of the grid $S_{n,d}$ provides $|\beta_n - \beta_0| \le h^{2s+2}$. Then, it is well known, e.g. from Ibragimov and Khasminskii (1981), that for $F(x) = f(x^\top \beta_0)$ with $f \in \Sigma(s, L)$ one has

$$\langle B_{\beta_0} \rangle = \langle \mathcal{K}_{\beta_0} F - F \rangle \le L' h^{2s+1} \tag{40}$$

with $L' = L\|K\|/(s-1)!$ But

$$\begin{aligned} |\langle B_{\beta_n} \rangle - \langle B_{\beta_0} \rangle| &\le \langle B_{\beta_n} - B_{\beta_0} \rangle \le \\ &\le \langle \mathcal{K}_{\beta_0} F - \mathcal{K}_{\beta_n} F \rangle \le \\ &\le \frac{1}{n} \sum_i \Bigg| \Pi_{\beta_n}(X_i) \sum_{j \ne i} F(X_j)\, K_{h,\beta_n}(X_i - X_j) - \\ &\qquad - \Pi_{\beta_0}(X_i) \sum_{j \ne i} F(X_j)\, K_{h,\beta_0}(X_i - X_j) \Bigg| . \end{aligned}$$

Now using assumptions (D) and $(K1)-(K5)$ we obtain

$$\begin{aligned}\left|\Pi_{\beta_n}^{-1}(X_i)-\Pi_{\beta_0}^{-1}(X_i)\right| &\leq \sum_{j\neq i}|K_{h,\beta_n}(X_i-X_j)-K_{h,\beta_0}(X_i-X_j)|\leq \\ &\leq C\Pi_{\beta_0}(X_i)\frac{|\beta_n-\beta_0|}{h} \qquad (41)\end{aligned}$$

and similarly

$$\sum_{j\neq i}|F(X_j)K_{h,\beta_n}(X_i-X_j)-F(X_j)K_{h,\beta_0}(X_i-X_j)|\leq C\Pi_{\beta_0}(X_i)\frac{|\beta_n-\beta_0|}{h}. \qquad (42)$$

Putting together (41) and (42) we conclude that

$$|\langle B_{\beta_n}\rangle-\langle B_{\beta_0}\rangle|\leq C\frac{|\beta_n-\beta_0|}{h}\leq Ch^{2s+1}$$

and the lemma follows with $C_b=L'+1$.

To complete the proof for the tests φ_n^{**} it remains to note that for each $F\in\mathcal{F}_1$

$$T_n^{**}\geq\frac{n\sqrt{h}}{V_{\beta n}}|\langle F-F_{\theta_n}\rangle-\langle B_{\beta_n}\rangle|+\frac{\xi_{\beta_n}}{V_{\beta_n}}$$

and that if

$$\langle F-F_{\theta_n}\rangle\geq C_bh^{2s}+\frac{2V^*}{n\sqrt{h}}\sqrt{(2+\delta)\ln N}, \qquad (43)$$

with V^* from Lemma 4.6, then by Lemma 4.3 we obtain

$$\begin{aligned}&\mathbf{P}\left(T_n^{**}<\sqrt{(2+\delta)\ln N}\right)\leq \\ &\leq \mathbf{P}\left(\frac{n\sqrt{h}}{V_{\beta n}}\frac{2V^*}{n\sqrt{h}}\sqrt{(2+\delta)\ln N}+\frac{\xi_{\beta_n}}{V_{\beta_n}}<\sqrt{(2+\delta)\ln N}\right)\leq \\ &\leq \mathbf{P}\left(\left|\frac{\xi_{\beta_n}}{V_{\beta_n}}\right|>\sqrt{(2+\delta)\ln N}\right)\to 0,\quad n\to\infty.\end{aligned}$$

Finally we remark that $\ln N\leq C\ln n$ and the choice of h by (8) yields

$$C_bh^{2s}+\frac{2V^*}{n\sqrt{h}}\sqrt{(2+\delta)\ln N}\leq C'\left(\frac{\sqrt{\ln n}}{n}\right)^{\frac{4s}{4s+1}}=C'h^{2s}$$

i.e. (43) holds true if c_n in the definition of the alternative H_1 is taken with $c_n^2\geq 2C'h^{2s}$. This completes the proof for the tests φ_n^{**}

Now we explain why the statistics T_n^{**} can be considered in place of T_n^*. The idea is to show that the difference $T_n^{**}-T_n^*$ is relatively small (being compared with the test level $\sqrt{2\ln N}$ or deviation $\langle F-F_{\theta_n}\rangle$). First we treat the preliminary parametric estimator $\tilde{\theta}_n$. Denote for given $F\in\mathcal{F}_0\cup\mathcal{F}_1$

$$d_n(F)=\langle F-F_{\theta_n}\rangle+\frac{\ln^2 n}{n},$$

θ_n being from (27)

Lemma 4.8 *Uniformly in $F \in \mathcal{F}_0 \cup \mathcal{F}_1$ we have for each $\delta > 0$*

$$\mathbf{P}\left(\frac{1}{d_n(F)}\left|\left\langle F - \tilde{F}_0\right\rangle - \langle F - F_{\theta_n}\rangle\right| > \delta\right) \to 0, \tag{44}$$

$$\mathbf{P}\left(\frac{1}{d_n(F)}\left|\left\langle F - \tilde{F}_0, \varepsilon\right\rangle\right| > \delta\right) \to 0.$$

Proof. Let us fix some $\delta > 0$ and some $\theta \in \Theta_n$. First we show that the probability of the event

$$\left\{|\langle F - F_\theta, \varepsilon\rangle| > \delta\left(\langle F - F_\theta\rangle + \frac{\ln^2 n}{n}\right)\right\}$$

is asymptotically small. More precise, we state the following assertion:

$$\sum_{\theta\in\Theta_n} \mathbf{P}\left(|\langle F - F_\theta, \varepsilon\rangle| > \delta\left(\langle F - F_\theta\rangle + \frac{\ln^2 n}{n}\right)\right) \to 0, \quad n \to \infty. \tag{45}$$

In fact, if we put $d_\theta^2 = \mathbf{E}\,|\langle F - F_\theta, \varepsilon\rangle|^2$ then we have

$$\begin{aligned} d_\theta^2 &= \mathbf{E}\left|\frac{1}{n}\sum_i \varepsilon_i\left[F(X_i) - F_\theta(X_i)\right]\right|^2 = \\ &= \frac{1}{n^2}\sum_i \sigma_i^2\left|F(X_i) - F_\theta(X_i)\right|^2. \end{aligned}$$

Using Lemma 4.6 we have

$$d_\theta^2 \le \frac{\sigma^{*2}}{n^2}\sum_i |F(X_i) - F_\theta(X_i)|^2 = \frac{\sigma^{*2}}{n}\langle F - F_\theta\rangle.$$

Further,

$$\frac{1}{2}\left(\langle F - F_\theta\rangle + \frac{\ln^2 n}{n}\right) \ge \sqrt{\langle F - F_\theta\rangle \frac{\ln^2 n}{n}}$$

and

$$\begin{aligned} &\mathbf{P}\left(|\langle F - F_\theta, \varepsilon\rangle| > \delta\left(\langle F - F_\theta\rangle + \frac{\ln^2 n}{n}\right)\right) \le \\ &\le \mathbf{P}\left(\frac{1}{d_\theta}|\langle F - F_\theta, \varepsilon\rangle| > \frac{\delta}{d_\theta}\ln n\sqrt{\langle F - F_\theta\rangle / n}\right) \le \\ &\le \mathbf{P}\left(\frac{1}{d_\theta}|\langle F - F_\theta, \varepsilon\rangle| > \frac{\delta}{\sigma^*}\ln n\right). \end{aligned}$$

Now we use an estimate of the large deviation probability for the centered and normalized random variables $\frac{1}{d_\theta}\langle F - F_\theta, \varepsilon\rangle$, see Lemma 4.11 below. Indeed, for n large enough

$$\begin{aligned} &\sum_{\theta\in\Theta_n} \mathbf{P}\left(\frac{1}{d_\theta}\langle F - F_\theta, \varepsilon\rangle > \frac{\delta}{\sigma^*}\ln n\right) \le \\ &\le \sum_{\theta\in\Theta_n} \exp\left\{-\frac{\delta^2}{2\sigma^{*2}}\ln^2 n\right\} \le n^d \exp\left\{-(d+1)\ln n\right\} \le n^{-1} \end{aligned}$$

which implies (45). Here we used that the cardinality of Θ_n is of order n^d. Let $\theta \in \Theta_n$ be such that

$$\langle F - F_\theta \rangle - \langle F - F_{\theta_n} \rangle > 2\delta d_n(F). \tag{46}$$

For δ small enough this yields

$$\langle F - F_\theta \rangle - \langle F - F_{\theta_n} \rangle > \delta \left(\langle F - F_\theta \rangle + \langle F - F_{\theta_n} \rangle \right). \tag{47}$$

Now by definition of $\tilde{\theta}_n$ we obtain through (46) and (47)

$$\begin{aligned} \left\{ \tilde{\theta}_n = \theta \right\} &\subseteq \{ \langle Y - F_\theta \rangle \le \langle Y - F_{\theta_n} \rangle \} = \\ &= \{ \langle F - F_\theta + \varepsilon \rangle \le \langle F - F_{\theta_n} + \varepsilon \rangle \} = \\ &= \{ \langle F - F_\theta \rangle - \langle F - F_{\theta_n} \rangle \le 2 \langle F - F_\theta, \varepsilon \rangle + 2 \langle F - F_{\theta_n}, \varepsilon \rangle \} \subseteq \\ &\subseteq \left\{ \langle F - F_\theta, \varepsilon \rangle > \frac{\delta}{2} \langle F - F_\theta \rangle \right\} \cup \left\{ \langle F - F_{\theta_n}, \varepsilon \rangle > \frac{\delta}{2} \langle F - F_{\theta_n} \rangle \right\}. \end{aligned}$$

Using this relation and (45) we deduce

$$\begin{aligned} &\mathbf{P}\left(\left| \left\langle F - F_{\tilde{\theta}_n} \right\rangle - \langle F - F_{\theta_n} \rangle \right| > 2\delta d_n(F) \right) \le \\ &\le \sum_{\theta \in \Theta_n} \mathbf{1}\left(|\langle F - F_\theta \rangle - \langle F - F_{\theta_n} \rangle| > 2\delta d_n(F) \right) \mathbf{P}\left(\tilde{\theta}_n = \theta \right) \le \\ &\le \sum_{\theta \in \Theta_n} \mathbf{P}\left(\langle F - F_\theta, \varepsilon \rangle > \frac{\delta}{2} \langle F - F_\theta \rangle \right) \to 0, \quad n \to \infty, \end{aligned}$$

that proves (44). The second statement of the lemma follows directly from (45).

The next step is to show that the last two terms in the expansion (29) are vanishing.

Lemma 4.9 *Given F let*

$$b_\beta = \langle B_\beta \rangle + \frac{\ln^2 n}{n}.$$

Then uniformly in $F \in \mathcal{F}_0 \cup \mathcal{F}_1$ for each $\delta > 0$ the following assertions hold:

$$\begin{aligned} \sum_{\beta \in S_{n,d}} \mathbf{P}\left(\langle B_\beta, \varepsilon \rangle > \delta b_\beta \right) &\to 0, \\ \sum_{\beta \in S_{n,d}} \mathbf{P}\left(\langle B_\beta, \mathcal{K}_\beta \varepsilon \rangle > \delta b_\beta \right) &\to 0. \end{aligned}$$

Remark **4.1** The statements of this lemma yield immediately that

$$\mathbf{P}\left(\langle B_\beta, \varepsilon \rangle \le \delta b_\beta \,, \quad \forall \beta \in S_{n,d} \right) \to 1$$

and similarly for $\langle B_\beta, \mathcal{K}_\beta \varepsilon \rangle$.

Proof. The statements of the lemma are proved in the same manner as in the last part of the proof of Lemma 4.8. For the second statement we use in addition the fact that

$$\operatorname{Var} \langle B_\beta, \mathcal{K}_\beta \varepsilon \rangle \le \frac{C}{n} \langle B_\beta \rangle. \tag{48}$$

Indeed, using assumptions $(E1)$-$(E3)$ and $(K1)$-$(K5)$, Lemma 4.6 and Jensen's inequality we have

$$\begin{aligned}
\mathbf{E}\left|\langle B_\beta, \mathcal{K}_\beta\varepsilon\rangle\right|^2 &= \frac{1}{n^2}\mathbf{E}\left|\sum_i B_\beta(X_i)\Pi_\beta(X_i)\sum_{j\neq i}\varepsilon_j K_{h,\beta}(X_i-X_j)\right|^2 = \\
&= \frac{1}{n^2}\mathbf{E}\left|\sum_j \varepsilon_j \sum_{i\neq j} B_\beta(X_i)\Pi_\beta(X_i)K_{h,\beta}(X_i-X_j)\right|^2 = \\
&= \frac{1}{n^2}\sum_j \sigma_j^2\left|\sum_{i\neq j} B_\beta(X_i)\Pi_\beta(X_i)K_{h,\beta}(X_i-X_j)\right|^2 \leq \\
&\leq \frac{1}{n^2}\sigma^{*2}C_\pi^2\sum_j \Pi_\beta^2(X_j)\left|\sum_{i\neq j} B_\beta(X_i)K_{h,\beta}(X_i-X_j)\right|^2 \leq \\
&\leq \frac{1}{n^2}\sigma^{*2}C_\pi^2\sum_j \left|\frac{\sum_{i\neq j} B_\beta(X_i)K_{h,\beta}(X_i-X_j)}{\sum_{i\neq j} K_{h,\beta}(X_i-X_j)}\right|^2 \leq \\
&\leq \frac{1}{n^2}\sigma^{*2}C_\pi^2 C\left\langle B_\beta\right\rangle.
\end{aligned}$$

Next we show that the quantities $\tilde{E}_\beta$ and $\tilde{V}_\beta$ estimate E_β and V_β good enough.

Lemma 4.10 *For each $\delta > 0$ and uniformly in $F \in \mathcal{F}_0 \cup \mathcal{F}_1$*

$$\begin{aligned}
\mathbf{P}\left(\sup_{\beta\in S_{n,d}}\left|\tilde{E}_\beta - E_\beta\right| > \frac{1}{n\sqrt{h}\ln n}\right) &\to 0, \\
\mathbf{P}\left(\sup_{\beta\in S_{n,d}}\left|\frac{\tilde{V}_\beta}{V_\beta} - 1\right| > \delta\right) &\to 0.
\end{aligned}$$

Proof. The assumption $(E3)$ implies for each $j = 1, \ldots, n$

$$\left|\sigma_j^2 - \tilde{\sigma}_j^2\right| \leq C_\sigma\left|\tilde{F}_1(X_j) - F(X_j)\right|$$

and hence

$$\left|\tilde{E}_\beta - E_\beta\right| \leq \frac{1}{n}\sum_i\sum_{j\neq i}\left|\sigma_j^2 - \tilde{\sigma}_j^2\right|\Pi_\beta^2(X_i)K_{h,\beta}^2(X_i-X_j).$$

Now by the design and kernel properties we derive for each $j = 1, \ldots, n$

$$\sum_{j\neq i}\Pi_\beta^2(X_i)K_{h,\beta}^2(X_i-X_j) \leq \frac{C}{nh}$$

and using Cauchy-Schwarz inequality we obtain

$$\left|\tilde{E}_\beta - E_\beta\right| \leq \frac{C}{n^2 h}\sum_i\left|\tilde{F}_1(X_j) - F(X_j)\right| \leq \frac{C}{n^2 h}\left[\frac{1}{n}\sum_i\left|\tilde{F}_1(X_j) - F(X_j)\right|^2\right]^{1/2}.$$

The pilot estimator $\tilde{F}_1$ fulfills with high probability

$$\left\langle \tilde{F}_1 - F \right\rangle \leq C n^{-\frac{2s}{2s+d}}.$$

Hence using the inequality $\frac{2s}{2s+d} > \frac{1}{4s+1}$ and the definition of h we arrive to the conclusion that

$$n\sqrt{h}\left|\tilde{E}_\beta - E_\beta\right| \leq \frac{C}{\sqrt{h}} n^{-\frac{s}{2s+d}} = o\left(\frac{1}{\ln n}\right).$$

Lemmas 4.8–4.10 together imply the asymptotic equivalence of the tests based on T_β and T'_β. We finish the proof of the theorem with a result on probabilities of deviations of centered and normalized sums of independent errors ε_i over the logarithmic level. The following lemma was already used in the proof of Lemma 4.8.

Lemma 4.11 *For each positive constants r, a the following relation holds uniformly in functions F from the Lipschitz class $\Sigma_d(1, L)$ of functions in $\mathbb{R}^d$:*

$$n^r \mathbf{P}\left(\xi(F) > a \ln n\right) \to 0, \quad n \to \infty,$$

where

$$\xi(F) = \frac{\langle F, \varepsilon \rangle}{\sqrt{\mathbf{E}\langle F, \varepsilon \rangle^2}}.$$

Proof. We proceed in a standard way using the exponential inequality and boundedness of errors ε_i due to $(E1)$. The details are omitted.

5 A simulation and an application

The purpose of our simulation experiments was to study the quantiles of the test statistic T_n^* and the power of the test in finite samples. All calculations have been performed in the languages GAUSS and XploRe (Härdle, Klinke and Turlach (1995)). The observations were generated according to a binary response model. The explanatory variables were identically independent uniform distributed on $[-1, 1]$. We took the parameter $\theta = \binom{1}{1}\frac{1}{\sqrt{2}}$ and considered the functions

$$f_0(u) = \frac{1}{1 + \exp^{-u}} \tag{49}$$

$$f_1(u) = f_0(u) + \eta \cdot \varphi'(u) \tag{50}$$

$$f_2(u) = 1 - \exp(-\exp(u)) \tag{51}$$

for different $0 < \eta \leq 1$, where φ is the density function of the standard normal distribution. While f_0 is a logit function, f_1 consists of a logit disturbed by a bump (figure 3). The response Y under H_0 was generated such that $P(Y = 1 | x^T\theta_0 = u) = f_0(u)$. We are thus interested in the hypothesis H_0

$$H_0 : F_\theta(x) = \mathbf{E}[Y | u(x, \theta) = u] = f_0(u) \qquad , \ \theta \in S_2$$

In a first step we calculated empirically the 90 and 95 percent quantiles of T_n^* for $n = 100$ and 200 observations generated by f_0. They were used then as rejection boundaries, defined as

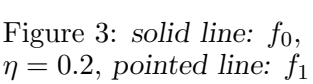


Figure 3: *solid line:* f_0, *dashed line:* f_1 with $\eta = 0.2$, *pointed line:* f_1 with $\eta = 0.6$

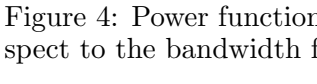


Figure 4: Power function of the test with respect to the bandwidth for funtion f_{1c}

$\sqrt{(2+\delta)\ln N}$, see (24). We calculated T_n^* by optimizing T_β over a grid, see (23), with $N = 50$ gridpoints. As kernel function K we used always the quartic kernel

$$K(u) = \frac{15}{16}(1-u^2)^2 \mathbf{1}_{\{|u|<1\}}$$

In the second step we analyzed the effect of increasing sample size on the power. In table 1 we show the power of the test when the data were generated with functions f_{1a}, that is f_1 for $\eta = 0.2$, f_{1c}, where $\eta = 0.6$ and f_2. In order not to oversmooth we used the bandwidth $h_1 = h = 0.5$ for $n = 100,\ 200$ and $h_1 = h = 0.25$ for $n = 350,\ 500$. Although we substituted for speed reasons in the cases $n = 350$ and 500 $\tilde{V}_\beta$ by $\tilde{V}_{\hat{\theta}}$ for all β, the power increases very fast with n. Therefore, it could be of interest to compare the power with regard to the bump η in the logit model. In table 2 we show for $n = 200$ and 350 the power of the test as a function of η. We see that for $\eta > 0.4$ this test procedure works very well.

Table 1: *Power and rejection boundaries for different alternatives.*

$n\ ,\ h\ =$	100 , 0.5		200 , 0.5		350 , 0.25		500 , 0.25	
level	5%	10%	5%	10%	5%	10%	5%	10%
rejection boundary	4.00	3.35	3.30	3.25	3.75	2.90	3.20	2.76
f_{1a}	0.056	0.096	0.112	0.215	0.133	0.207	0.150	0.200
f_{1c}	0.224	0.294	0.530	0.690	0.798	0.856	0.900	0.960
f_2	0.316	0.376	0.946	0.991	0.995	1.000	0.995	1.000

Table 2: *Power for different bumps η .*

	$\eta =$	0.2		0.4		0.6		1.0	
	level	5%	10%	5%	10%	5%	10%	5%	10%
$n,\ h$	200, 0.50	0.112	0.215	0.227	0.419	0.530	0.690	0.687	0.801
	350, 0.25	0.133	0.207	0.321	0.478	0.798	0.856	0.889	0.926

The last step of the simulation experiment was the study of bandwidth choice. For the sake of simplicity we set $h_1 = h$ as above. First we always have had to determine numerically the rejection boundaries for the special bandwidth h. Here we observed shrinking boundaries, when h grew from 0.25 up to 2.25 . In figure 4 we plot the bandwidth vs the power of the test with observations generated by f_{1c}. Obviously for this kind of alternative we get better power for larger bandwidths.

In the introductory example we dealt with youth unemployement. The question is, can we explain the youth unemployement with the aforementioned predictor variables X in a single index model with logit link? In the application of this dataset, we used a slightly modified numerical procedure as described in Proenca and Ritter (1995). Further we rescaled the explanatory variables of each dimension to $[-1, 1]$. Since there are three dimensions ($d = 3$) for a sample size of $n = 462$, we chose the bandwidth h_1 large, definitively 1.5, whereas $h = 0.3$. By Monte Carlo studies described above we determined the 90 and 95% one side quantiles of T^*_{462} and got 1.74 respectively 2.38 . Now we ran the test for our data and got the statistic value $T^*_{462} = 3.076$ for $\beta = (-0.18010, -0.10725, 0.97778)$. For purpose of comparison in table 3 we switch the norm of β and set his first component equal to the corresponding one of θ, the parameter of the logit fit in figure 1.

Table 3: *Comparison of θ and β.*

explanatory variables	intercept	earnings as an apprentice	percentage of apprentices divided by employees	unemployed rate
θ	-2.40996	-0.07999	-0.17989	0.95113
β	-	-0.07999	-0.04763	0.43422

Appendix

Proof of Theorem 1. To simplify our exposition and to emphasize the main idea we consider the case when the parametric family consists of one point, namely, a zero regression function, and errors ε_i are independent and standard Gaussian. Moreover, we assume random design with a design density $\pi(x)$ in $\mathbb{R}^d$ of the form $\pi(x) = \pi_1(|x|)$ where a univariate function $\pi_1(\cdot)$ is compactly supported on $[-1, 1]$, symmetric, twice continuously differentiable and satisfies $\pi_1(t) = 3/4$ for $|t| \le 1/2$.

The idea of the proof is standard. We replace the minimax problem by a Bayes one where we consider instead of the set $\mathcal{F}_1$ of alternatives one Bayes alternative corresponding to a prior ν concentrated on $\mathcal{F}_1$. We try to choose this prior ν in such a way that the likelihood $Z_\nu = d\mathbf{P}_\nu/d\mathbf{P}_0$ is close to 1 where the measure $\mathbf{P}_\nu$ is the Bayes measure for the prior ν and $\mathbf{P}_0$ corresponds to the case of zero regression function. The Neyman-Pearson Lemma yields that the hypothesis $H_0 : \mathbf{P} = \mathbf{P}_0$ can not be consistently distinguished versus the Bayes alternative $H_\nu : \mathbf{P} = P_\nu$ and hence versus the composite alternative $H_1 : \mathbf{P} \in \mathcal{F}_1$.

Now we describe the structure of the prior ν. Let $g(\cdot)$ be some function from the Hölder class $\Sigma(s, L)$, supported on $[-1, 1]$ and satisfying the conditions

$$\int g(t)\,dt = 0, \qquad \|g\|^2 = \int g^2(t)\,dt > 0. \tag{52}$$

Set

$$h = \left(\frac{a\sqrt{\ln n}}{n}\right)^{\frac{2}{4s+1}} \tag{53}$$

where a constant a will be chosen later. Denote by $\mathcal{I}_n$ the partition of the interval $\left[-\frac{1}{2}, \frac{1}{2}\right]$ into intervals of length h. Without loss of generality we assume that the cardinality of the set $\mathcal{I}_n$ coincides with $1/h$

$$\#\mathcal{I}_n = \frac{1}{h}. \tag{54}$$

For each interval $I \in \mathcal{I}_n$ introduce a function $g_I(t)$ of the form

$$g_I(t) = h^s g\left(\frac{t - t_I}{h}\right), \tag{55}$$

t_I being the center of I. Evidently $g_I(\cdot)$ is supported on I, $g_I \in \Sigma(s, L)$ and the followings hold for h small enough:

$$\int g_I(t)\, dt = 0, \qquad \int g_I^2(t)\, dt = h^{2s+1} \|g\|^2. \tag{56}$$

Let now μ be a set of binary values $\{\mu_I\,, I \in \mathcal{I}_n\}$ i.e. $\mu_I = \pm 1$. Define a function $G_\mu(t)$ with

$$G_\mu(t) = \sum_{I \in \mathcal{I}_n} \mu_I\, g_I(t). \tag{57}$$

This function $G_\mu \in \Sigma(s, L)$ vanishes outside $\left[-\frac{1}{2}, \frac{1}{2}\right]$ and by (56)

$$\int G_I^2(t)\, dt = \sum_{I \in \mathcal{I}_n} \int g_I^2(t)\, dt = \frac{1}{h} h^{2s+1} \int g^2(t)\, dt = h^{2s} \|g\|^2. \tag{58}$$

Taking into account (53) we see that the distance between zero function and each G_μ is just of the rate c_n^2 from Theorems 2.1 and 2.2.

Denote by $\mathcal{M}_n$ the set of all possible collections $\{\mu_I\,, I \in \mathcal{I}_n\}$ with binaries $\mu_I = \pm 1$, and let $m(d\mu)$ be the uniform measure on $\mathcal{M}_n$. This measure can be represented as the direct product of binary measures $m_I(d\mu_I)$ with $m_I(\mu_I = \pm 1) = 1/2$.

Now we pass to the semiparametric model. Let S_n be a grid on the unit sphere S_d with the step b_n,

$$b_n = h^{1/8}, \tag{59}$$

h being from (53). This means that $|\beta - \beta'| \geq b_n = h^{1/8}$ for each $\beta, \beta' \in S_n$, $\beta \neq \beta'$. Below we will use that for some $\alpha > 0$

$$N = \#S_n \asymp n^\alpha \tag{60}$$

and for n large enough

$$\frac{h \ln n}{|\beta - \beta'|^4} \leq \frac{h \ln n}{b_n^4} \leq h^{1/4} \quad \forall \beta, \beta' \in S_n,\, \beta \neq \beta'. \tag{61}$$

For each $\beta \in S_n$ and each $\mu \in \mathcal{M}_n$ define the multivariate function $G_{\beta,\mu}(x)$ on $\mathbb{R}^d$ with

$$G_{\beta,\mu}(x) = G_\mu(x^\top \beta).$$

It is clear that the function $G_{\beta,\mu}(x)$ is Hölder, $G_{\beta,\mu}(x) \in \Sigma_d(s, L)$, and by (58) we get

$$\int G_{\beta,\mu}^2(x)\, \pi(x)\, dx = \int G_\mu^2(x^\top \beta)\, \pi_1(|x|)\, dx = \int G_\mu^2(t)\, \pi_2(t)\, dt = C_0 h^{2s} \tag{62}$$

with $\pi_2(t) = \frac{d}{dt} \int \mathbf{1}(x^\top \beta \le t)\, \pi_1(|x|)\, dx$ and $C_0 \in \left[\frac{1}{2} \|g\|^2, \|g\|^2 \right]$.

Finally we take the prior ν as the uniform measure on the set of functions $\{G_{\beta,\mu}\}$, $\beta \in S_n$, $\mu \in \mathcal{M}_n$, and

$$\mathbf{P}_\nu = \frac{1}{N} \sum_{\beta \in S_n} \frac{1}{M} \sum_{\mu \in \mathcal{M}_n} \mathbf{P}_{G_{\beta,\mu}} . \tag{63}$$

Here $M = \#\mathcal{M}_n = 2^{1/h}$, N being from (60). Denote also $Z_\nu = \frac{d\mathbf{P}_\nu}{d\mathbf{P}_0}$ and notice that this likelihood can be represented in the form $Z_\nu = \frac{1}{N} \sum_{\beta \in S_n} Z_\beta$ with

$$Z_\beta = \frac{1}{M} \sum_{\mu \in \mathcal{M}_n} Z_{\beta,\mu} = \frac{1}{M} \sum_{\mu \in \mathcal{M}_n} d\mathbf{P}_{G_{\beta,\mu}} / d\mathbf{P}_0. \tag{64}$$

Our goal is to prove that for a small enough in (53) one has

$$Z_\nu \to 1 \tag{65}$$

under the measure $\mathbf{P}_0$.

We start from a decomposition and an asymptotic expansion for each Z_β from (64). For that we need some more notation. Fix some $\beta \in S_n$ and put

$$\sigma^2_{\beta,I} = \sum_i g^2_I(X_i^\top \beta), \quad I \in \mathcal{I}_n , \tag{66}$$

$$\xi_{\beta,I} = \frac{1}{\sigma_{\beta,I}} \sum_i g_I(X_i^\top \beta)\, \varepsilon_I, \quad I \in \mathcal{I}_n . \tag{67}$$

We see that $\xi_{\beta,I}$ are standard normal and independent for different $I \in \mathcal{I}_n$, and

$$\sum_i G^2_{\beta,\mu}(X_i) = \sum_i G^2_\mu(X_i^\top \beta) = \sum_{I \in \mathcal{I}_n} \sigma^2_{\beta,I} .$$

Recall that we assume the random design and

$$\mathbf{E} \sum_i G^2_{\beta,\mu}(X_i) = n \int G^2_{\beta,\mu}(x)\, \pi(x)\, dx = nC_0 h^{2s}. \tag{68}$$

Similarly for each $\sigma^2_{\beta,I}$

$$\mathbf{E}\sigma^2_{\beta,I} = n \int g^2_I(x^\top \beta)\, \pi(x)\, dx = n \int g^2_I(x^\top \beta)\, \pi_1(|x|)\, dx = nC_I h^{2s+1} \tag{69}$$

where C_I does not depends on β and $C_I \in \left[C_0/\sqrt{2}, \sqrt{2} C_0 \right]$.

Lemma 5.1

$$Z_\beta = \prod_{I \in \mathcal{I}_n} ch(\sigma_{\beta,I}\, \xi_{\beta,I}) e^{-\frac{1}{2}\sigma^2_{\beta,I}}$$

where $ch(z) = \frac{1}{2}\,(e^z + e^{-z})$.

Proof. By Girsanov formulae and (66)-(67)

$$\begin{aligned} Z_{\beta,\mu} &= \exp\left\{\sum_i G_{\beta,\mu}(X_i)\,\varepsilon_i - \frac{1}{2}\sum_i G^2_{\beta,\mu}(X_i)\right\} = \\ &= \exp\left\{\sum_{I\in\mathcal{I}_n} \mu_I \sigma_{\beta,I}\xi_{\beta,I} - \frac{1}{2}\sum_{I\in\mathcal{I}_n}\sigma^2_{\beta,I}\right\} = \\ &= \prod_{I\in\mathcal{I}_n} \exp\left\{\mu_I\sigma_{\beta,I}\xi_{\beta,I} - \frac{1}{2}\sigma^2_{\beta,I}\right\}. \end{aligned}$$

Now the lemma's assertion follows from the direct product structure of the measure $m(d\mu)$.

Denote also

(70) $$v^2_\beta = \frac{1}{2}\sum_{I\in\mathcal{I}_n}\sigma^4_{\beta,I}\,,$$

(71) $$\zeta_\beta = \frac{1}{v_\beta}\sum_{I\in\mathcal{I}_n}\sigma^2_{\beta,I}\left(\xi^2_{\beta,I}-1\right).$$

Lemma 5.2 *The following statements hold:*

(i) $\mathbf{E}\zeta_\beta = 0$;

(ii) $\mathbf{E}\zeta^2_\beta = 1$;

(iii) $v^2_\beta = C_1 n^2 h^{4s+1} = C_1 \ln n$ *with* $C_1 \le a$;

(iv) There exists an independent standard normal r.v. $\tilde{\zeta}_\beta$ *that*

$$\ln n \sup_{\beta\in S_n} \mathbf{E}_0\left(\tilde{\zeta}_\beta - \zeta_\beta\right)^2 \to 0.$$

Proof. The first two statements are obvious. (iii) follows from (69). Finally, (iv) is the application of the Strassen type invariance principle (see, e.g. ??).

The next step is the asymptotic expansion for each Z_β.

Lemma 5.3 *The following statements are satisfied uniformly in* $\beta \in S_n$*: for each* $\delta > 0$

(i)
$$\mathbf{P}_0\left(\left|Z_\beta - \exp\left\{v_\beta\zeta_\beta - \frac{1}{2}v^2_\beta\right\}\right| > \delta\right) \to 0;$$

(ii)
$$\mathbf{P}_0\left(\left|\tilde{Z}_\beta - \exp\left\{v_\beta\tilde{\zeta}_\beta - \frac{1}{2}v^2_\beta\right\}\right| > \delta\right) \to 0;$$

Proof. The first statement is equivalent to the following one:

$$\mathbf{P}_0\left(\left|\ln Z_\beta - v_\beta\zeta_\beta + \frac{1}{2}v_\beta^2\right| > \delta\right) \to 0.$$

But the latter can be obtained using Taylor expansion for $\ln Z_\beta$

$$\begin{aligned}\ln Z_\beta &= \sum_{I\in\mathcal{I}_n} \ln ch(\sigma_{\beta,I}\xi_{\beta,I}) - \frac{1}{2}\sigma_{\beta,I}^2 = \\ &= \sum_{I\in\mathcal{I}_n}\left[\frac{1}{2}\sigma_{\beta,I}^2\left(\xi_{\beta,I}^2-1\right) - \frac{1}{12}\sigma_{\beta,I}^4\xi_{\beta,I}^4 + O(\sigma_{\beta,I}^6\xi_{\beta,I}^6)\right]\end{aligned}$$

and the following asymptotic relations which hold uniformly in β

$$\begin{aligned}\mathbf{P}_0\left(\left|\sum_{I\in\mathcal{I}_n}\sigma_{\beta,I}^4(\xi_{\beta,I}^4-3)\right| > \delta\right) &\to 0;\\ \mathbf{P}_0\left(\left|\sum_{I\in\mathcal{I}_n}\sigma_{\beta,I}^6\xi_{\beta,I}^6\right| > \delta\right) &\to 0;\end{aligned}$$

for details we refer to Ingster(1993).

The second statement of the lemma follows directly from (iii) of Lemma 5.2.

Now we arrive at the central point of the proof. Actually we prove that "submodels" corresponding to different β are in some sense asymptotically independent. That is why we have to pay with the extra log-term for the choice of "direction" β.

Lemma 5.4 *There exist a universal constant R such that for any $\beta,\beta'\in S_n$, $\beta\neq\beta'$,*

$$|\mathbf{E}\,\zeta_\beta\zeta_{\beta'}| \le \frac{Rh}{|\beta-\beta'|^4}. \tag{72}$$

Proof. Let us fix some β,β' from S_n. Denote by ρ their scalar product,

$$\rho = (\beta,\beta').$$

Now fix also some I, I' from $\mathcal{I}_n$ and set

$$r = r(\beta, I, \beta', I') = \mathbf{E}\,\xi_{\beta,I}\xi_{\beta',I'}\,.$$

Using normality of $\xi_{\beta,I}$ and $\xi_{\beta',I'}$ we calculate easily

$$\mathbf{E}\left(\xi_{\beta,I}^2-1\right)\left(\xi_{\beta',I'}^2-1\right) = 4r^2-2r. \tag{73}$$

Below we state that r satisfies the condition

$$|r| \le Ch^2/(1-\rho)^2 \tag{74}$$

with some universal constant C and now we show that this implies (72). In fact, through (73) one has

$$\begin{aligned}\mathbf{E}\,\zeta_\beta\zeta_{\beta'} &= \mathbf{E}\frac{1}{v_\beta}\sum_{I\in\mathcal{I}_n}\sigma^2_{\beta,I}\left(\xi^2_{\beta,I}-1\right)\frac{1}{v_{\beta'}}\sum_{I'\in\mathcal{I}_n}\sigma^2_{\beta',I'}\left(\xi^2_{\beta',I'}-1\right)= \\ &= \frac{1}{v_\beta}\frac{1}{v_{\beta'}}\sum_{I\in\mathcal{I}_n}\sum_{I'\in\mathcal{I}_n}\sigma^2_{\beta,I}\sigma^2_{\beta',I'}\left[4r^2(\beta,I,\beta',I')-2r(\beta,I,\beta',I')\right]\end{aligned}$$

and hence by (69) and (iii) of Lemma 5.2 we obtain

$$|\mathbf{E}\,\zeta_\beta\zeta_{\beta'}| \le \frac{Ch^2}{(1-\rho)^2}\frac{1}{v_\beta}\frac{1}{v_{\beta'}}\sum_{I\in\mathcal{I}_n}\sum_{I'\in\mathcal{I}_n}\sigma^2_{\beta,I}\sigma^2_{\beta',I'} \le \frac{Ch}{(1-\rho)^2}$$

and (72) follows.

To prove (74) we note that

$$\begin{aligned}r &= \mathbf{E}\,\xi_{\beta,I}\xi_{\beta',I'} = \\ &= \mathbf{E}\frac{1}{\sigma_{\beta,I}\sigma_{\beta',I'}}\sum_i g_I(X_i^\top\beta)\,g_{I'}(X_i^\top\beta') = \\ &= \frac{n}{\sigma_{\beta,I}\sigma_{\beta',I'}}\int g_I(x^\top\beta)\,g_{I'}(x^\top\beta')\,\pi(x)\,dx.\end{aligned}$$

Introduce new variables y_1 and y_2 with $x^\top\beta = t_I + hy_1,\ x^\top\beta' = t_{I'} + hy_2$. We have

$$|x|^2 = (t_I+hy_1)^2 + \left|\frac{t_{I'}+hy_2-\rho\,(t_I+hy_1)}{1-\rho}\right|^2, \tag{75}$$

$$r = \frac{nh^{2s+2}}{\sigma_{\beta,I}\sigma_{\beta',I'}(1-\rho)}\int g(y_1)\,g(y_2)\,\pi_1(|x|^2)\,dy_1\,dy_2.$$

Now we use the Taylor expansion for the function $p(y_1,y_2) = \pi_1\left(|x|^2\right)$ with $|x|^2$ due to (75). This function is continuous differentiable and all first derivatives are bounded by $Ch/(1-\rho)$ with some constant C depending only on the function π_1. Using the equality $\int g(t)\,dt = 0$ and (69) we get

$$|r| \le \frac{Cnh^{2s+2}h}{nh^{2s+1}(1-\rho)^2} = \frac{Ch^2}{(1-\rho)^2}.$$

Now everything is prepared to complete the proof of (65). The results of Lemmas 5.2 and 5.3 reduce this assertion to the following one:

$$\frac{1}{N}\sum_{\beta\in S_n}\left[\exp\left\{v_\beta\tilde{\zeta}_\beta-\frac{1}{2}v_\beta^2\right\}-1\right]\to 0 \tag{76}$$

under the measure $\mathbf{P}_0$. It suffices to check that

$$\frac{1}{N^2}\mathbf{E}_0\left|\sum_{\beta\in S_n}\left(\tilde{Z}_\beta-1\right)\right|^2\to 0$$

with

$$\tilde{Z}_\beta = \exp\left\{v_\beta \tilde{\zeta}_\beta - \frac{1}{2}v_\beta^2\right\}.$$

Using normality of $\tilde{\zeta}_\beta$ and (iii) of Lemma 5.2 one derives

$$\mathbf{E}_0\left[\exp\left\{v_\beta \tilde{\zeta}_\beta - \frac{1}{2}v_\beta^2\right\} - 1\right]^2 = \exp\left\{v_\beta^2\right\} \le n^a.$$

For different $\beta, \beta' \in S_n$ denote $r = \mathbf{E}_0 \tilde{\zeta}_\beta \tilde{\zeta}_{\beta'}$. Then $\tilde{\zeta}_{\beta'}$ can be represented in the form $\tilde{\zeta}_{\beta'} = r\tilde{\zeta}_\beta + (1-r)\zeta'$ with ζ' independent of $\tilde{\zeta}_\beta$. Now

$$\begin{aligned}
\mathbf{E}_0 \tilde{Z}_\beta \tilde{Z}_{\beta'} &= \mathbf{E}_0 \exp\left\{(v_\beta + r v_{\beta'})\tilde{\zeta}_\beta - \frac{1}{2}v_\beta^2\right\} \exp\left\{(1-r)v_{\beta'}\zeta' - \frac{1}{2}v_{\beta'}^2\right\} = \\
&= \exp\left\{\frac{1}{2}(v_\beta + r v_{\beta'})^2 - \frac{1}{2}v_\beta^2 + (1-r)^2 v_{\beta'}^2 - \frac{1}{2}v_{\beta'}^2\right\} = \\
&= \exp\left\{r v_\beta v_{\beta'} - r v_{\beta'}^2 + \frac{1}{2}r^2(v_\beta^2 + v_{\beta'}^2)\right\}.
\end{aligned}$$

The results of Lemma 5.4 and (iv) of Lemma 5.2 allow us to obtain

$$\mathbf{E}_0\left(\tilde{Z}_\beta - 1\right)\left(\tilde{Z}_{\beta'} - 1\right) = \mathbf{E}_0 \tilde{Z}_\beta \tilde{Z}_{\beta'} + 1 \le Cr \ln n.$$

Finally, by (61), Lemma 5.4 and (iii),(iv) of Lemma 5.2 we derive

$$\begin{aligned}
&\frac{1}{N^2}\mathbf{E}_0\left|\sum_{\beta\in S_n}\left(\tilde{Z}_\beta - 1\right)\right|^2 = \\
&= \frac{1}{N^2}\sum_{\beta\in S_n}\mathbf{E}_0\left(\tilde{Z}_\beta - 1\right)^2 + \frac{1}{N^2}\sum_{\beta\in S_n}\sum_{\beta'\in S_n, \beta'\neq\beta}\mathbf{E}_0\left(\tilde{Z}_\beta - 1\right)\left(\tilde{Z}_{\beta'} - 1\right) \le \\
&\le \frac{1}{N^2}\sum_{\beta\in S_n} e^{v_\beta^2} + \frac{1}{N^2}\sum_{\beta\in S_n}\sum_{\beta'\in S_n, \beta'\neq\beta}\frac{Ch\ln n}{|\beta - \beta'|^4} \le \\
&\le \frac{1}{N^2}n^a + \frac{Ch\ln n}{b_n^4} \to 0
\end{aligned}$$

if a is small enough.

References

[1] Carroll, R. J. and Ruppert, D. (1988). Transformation and weighting in regression, *Chapman and Hall, New York.*

[2] Engle, R. F., Granger, W. J., Rice, J. and Weiss, A. (1986). Semiparametric estimates of the relation between weather and electricity sales, *Journal of the American Statistical Association,* **81**, pp. 310-20

[3] Friedman, F. H. and Stuetzle, W. (1981). A projection pursuit regression, *Journal of the American Statistical Association*, **79**, pp. 599-608

[4] Golubev, G. (1992). Asymptotic minimax regression estimation in additive model, *Problems of Information Transmission,* **28**, pp. 3-15 (in russian)

[5] Green, P. and Silverman, B. W. (1994). The penalized likelihood approach, *Chapman and Hall, London.*

[6] Härdle, W. (1990). Applied nonparametric regression, *Econometric Society Monographs No. 19, Cambridge University Press.*

[7] Härdle, W., Klinke, S. and Turlach, T. A. (1995). XploRe-An interactive statistical computing environment, *Springer Verlag.*

[8] Härdle, W. and Mammen, E. (1993). Comparing nonparametric versus parametric regression fits, *Annals of Statistic,* **4**, pp. 1926-47

[9] Härdle, W. and Horowitz, J. (1994). Testing a parametric model against a semiparametric alternative, *Econometric Theory.*

[10] Hall, P. (1989). On projection pursuit regression, *Annals of Statistic,* **17**, pp. 573-8

[11] Horowitz, J. (1993). Semiparametric and nonparametric estimation of quantal response models, in *G. S. Maddala, C. R. Rao and H. D. Vinod (eds), Handbook of Statistics, Elsevier Science Publishers,* pp. 45-72

[12] Huber, P. J. (1985). Projection pursuit, *Annals of Statistic,* **13**, pp. 435-75

[13] Huet, S., Jolivet, E. and Mésseau, A. (1993). La regression non-lineaire: methodes et applications en biologie, *INRA, Paris,* Chapter 1,3.

[14] Ibragimov, I. A. and Khasminski, R.Z. (1977). One problem of statistical estimation in Gaussian white noise, *Soviet Math. Dokl.,* **236**, pp. 1351-4

[15] Ibragimov, I. A. and Khasminski, R.Z. (1981). Statistical Estimation; Asymptotic Theory, *Springer Verlag.*

[16] Ingster, Yu. I. (1982). Minimax nonparametric detection of signals in white Gaussian noise, *Problems of Information Transmission,* **18**, pp. 130-40

[17] Ingster, Yu. I. (1993). Asymptotically minimax hypothesis testing for nonparametric alternatives I, II, III, *Mathematical Methods of Statistics.* **2**, pp. 85-114

[18] Maddala, G. (1983). Limited-dependent and qualitative variables in econometrics, *Cambridge University Press.*

[19] McCullagh, P. and Nelder, J. A. (1989). Generalized Linear Models, *Monographs on Statistics and Applied Probability, 2 edn,* **37**, *Chapman and Hall, London.*

[20] Müller, H. G. (1987). Weighted local regression and kernel methods for nonparametric curve fitting, *Journal of the American Statistical Association,* **82**, pp. 231-8

[21] Proenca, I. and Ritter, Chr. (1995). Negative bias in the H-H Statistik, *Computational Statistics,*

[22] Rice, J. A. (1986). Convergence rates for partially splined models, *Statistics and Probability Letters,* **4**, pp.203-8

[23] Severini, T. A. and Staniswalis, J. G. (1994). Quasi-likelihood Estimation in Semiparametric Models, *Journal of the American Statistical Association,* **89**, pp.501-11

[24] Speckman, P. (1988). Kernel smoothing in partial linear models, *Journal of the Royal Statistical Society, Series B,* **50**, pp. 413-46

Journal of Econometrics, 81, 223-242

ELSEVIER

Journal of Econometrics 81 (1997) 223–242

JOURNAL OF Econometrics

Local polynomial estimators of the volatility function in nonparametric autoregression

W. Härdle[a,*], A. Tsybakov[b]

[a] *Humboldt-Universität zu Berlin, Wirtschaftswissenschaftliche Fakultät, Institut für Statistik und Ökonometrie, Spandauer Strasse 1, D-10178 Berlin, Germany*

[b] *Université Paris VI, Laboratoire de Statistique Théorique et Appliquée, 4, pl. Jussieu, Tour 45-55, F-75252 Paris, France*

Abstract

In this paper we consider a class of dynamic models in which both the conditional mean and the conditional variance (volatility) are unknown functions of the past. We first derive probabilistic conditions under which nonparametric estimation of these functions is possible. We then construct an estimator based on local polynomial fitting. We examine the rates of convergence of these estimators and give a result on their asymptotic normality. The local polynomial fitting of the volatility function is applied to different foreign exchange rate series. We find an asymmetric U-shaped 'smiling face' form of the volatility function. © 1997 Elsevier Science S.A.

Key words: Local polynomials; Nonlinear time series; Nonlinear autoregression; Volatility

JEL classification: C14; C22

1. Nonparametric autoregression with unknown volatility

The time series literature has provided many new approaches for dynamic econometric modelling. For example, vector autoregressive models are now widely used as an alternative to structural models (Lütkepohl, 1992). However, this literature is mostly devoted to the (linear) conditional mean (given the past). The conditional variance is assumed to be fixed or of specific form. In the beginning of the eighties this drawback has been stressed by Engle (1982), Robinson (1983, 1984) in the econometric literature and by Collomb (1984), Vieu (1995) in the statistical literature. In the framework of the ARCH models (Engle, 1982) the conditional variance is often specified as a linear function of the squared values

* Corresponding author

PII S0304-4076(97)00044-4

Journal of Econometrics, 81, 223-242

of the part innovations, although nonparametric and semiparametric approaches (Gregory, 1989; Engle and Gonzales-Rivera, 1989) have also been proposed. To our knowledge the first paper that models both the conditional mean and the conditional variance in a flexible nonparametric way is by Gouriéroux and Monfort (1992). Their model is in the case of one lag:

$$Y_i = \sum_{j=1}^{J} \alpha_j I(Y_{i-1} \in A_j) + \sum_{j=1}^{J} \beta_j I(Y_{i-1} \in A_j)\xi_i \tag{1.1}$$

and is called a Qualitative threshold ARCH (QTARCH(1)) model. Here $\{A_j\}_{j=1}^{J}$ with fixed J denotes a partition of the set of values for Y, $(\alpha_j),(\beta_j)$ are unknown parameter vectors and matrices respectively, and ξ_i is white noise. It is a generalisation of the threshold models for the conditional mean (Tong, 1983).

In this paper we generalize model (1.1) to a wider class of conditional mean and variance functions. In a sense the following model can be seen as a limit of (1.1) for $J \to \infty$:

$$Y_i = f(Y_{i-1}) + s(Y_{i-1})\xi_i, \quad i = 1,2,\ldots \tag{1.2}$$

where ξ_i are i.i.d. random variables, $E(\xi_i)=0$, $E(\xi_i^2)=1$, f and s are unknown functions on $\mathbb{R}^1$, $s(y)>0$, $\forall y \in \mathbb{R}^1$, and Y_0 is a random variable independent of $\{\xi_i\}$. We study the problem of estimation of the volatility function $v(x)=s^2(x)$, given a sample $Y_1,\ldots,Y_n$.

The model (1.2) was widely studied in financial time series context, especially under the assumption of ARCH structure (Engle, 1982). We are interested in the nonparametric situation where the exact parametric form of $f(\bullet)$ and $s(\bullet)$ is not predefined. Interest in this approach has grown in the economics and statistics literature. The method of Gouriéroux and Monfort (1992), and the paper of McKeague and Zhang (1994) are based on histogram type estimators of volatility. The papers by Chen and Tsay (1993a,b) concentrate on additive modelling for the mean function f. Here we propose a general class of volatility function estimators based on local polynomial (LP) estimation. The advantage of such estimators is that they approximate the volatility function better when it is smoother.

The idea of local polynomial estimation goes up to Stone (1977), Cleveland (1979) and Katkovnik (1979, 1985), who applied it for nonparametric regression models. For the study of statistical properties of LP estimators in nonparametric regression (convergence, rate of convergence and pointwise asymptotic normality) we refer to Tsybakov (1986). For the references on more recent work in this area see Fan and Gijbels (1996). In the present paper we modify the original LP approach by considering the joint LP-estimation of conditional mean and volatility function. Also, we treat the time-series model (1.2), instead of the classical i.i.d. nonparametric regression model. The main result of this paper is pointwise joint asymptotic normality of LP-estimators of conditional mean and volatility.

Journal of Econometrics, 81, 223-242

Inspection of the proofs in Section 5 shows that this result also holds (with obvious reformulation) for the nonparametric regression model with heteroscedastic errors: $Y_i = f(X_i) + s(X_i)\xi_i$, where ξ_i are as in (1.2), (X_i, Y_i) are i.i.d., and the design points $\{X_i\}$ are independent of $\{\xi_i\}$.

Along with statistical studies of the model (1.2), we mention the work on probabilistic properties of the process (1.2): Doukhan and Ghindés (1980, 1981), Chan and Tong (1985), Mokkadem (1987), Diebolt and Guégan (1990) and Ango Nze (1992). In these papers the ergodicity, geometric ergodicity and mixing properties of the process $\{Y_i\}$ are derived under appropriate conditions.

2. The estimator

Note that, if $\{Y_i\}$ were a stationary process, we would have

$$v(x) = \mathrm{E}(Y_i^2 | Y_{i-1} = x) - \mathrm{E}^2(Y_i | Y_{i-1} = x). \tag{2.1}$$

In fact, $\{Y_i\}$ approaches a stationary process, as $i \to \infty$. Thus, we look for an estimator of v of the form

$$\hat{v}_n(x) = \hat{g}_n(x) - \hat{f}_n^2(x), \tag{2.2}$$

where $\hat{g}_n(x)$ is an estimator of

$$g(x) = f^2(x) + s^2(x),$$

and $\hat{f}_n(x)$ is an estimator of $f(x)$. In order to define $\hat{f}_n$ and $\hat{g}_n$ by the LP method, consider the following minimisation problems:

$$\begin{aligned} \tilde{c}_n(x) &= \arg\min_{c \in \mathbb{R}^l} \sum_{i=1}^{n} (Y_i^2 - c^{\mathrm{T}} U_{in})^2 K\left(\frac{Y_{i-1} - x}{h_n}\right), \\ c_n(x) &= \arg\min_{c \in \mathbb{R}^l} \sum_{i=1}^{n} (Y_i - c^{\mathrm{T}} U_{in})^2 K\left(\frac{Y_{i-1} - x}{h_n}\right), \end{aligned} \tag{2.3}$$

where $K : \mathbb{R}^1 \to \mathbb{R}^1$ is a kernel and h_n is a positive number (bandwidth), $h_n \to 0$, as $n \to \infty$,

$$U_{in} = F(u_{in}), \quad F(u) = \begin{pmatrix} 1 \\ u \\ \vdots \\ \dfrac{u^{l-1}}{(l-1)!} \end{pmatrix}, \quad u_{in} = \frac{Y_{i-1} - x}{h_n}. \tag{2.4}$$

The estimator $\hat{v}_n(x)$ of $v(x)$ is defined as

$$\hat{v}_n(x) = \tilde{c}_n(x)^{\mathrm{T}} F(0) - \{c_n(x)^{\mathrm{T}} F(0)\}^2. \tag{2.5}$$

Journal of Econometrics, 81, 223-242

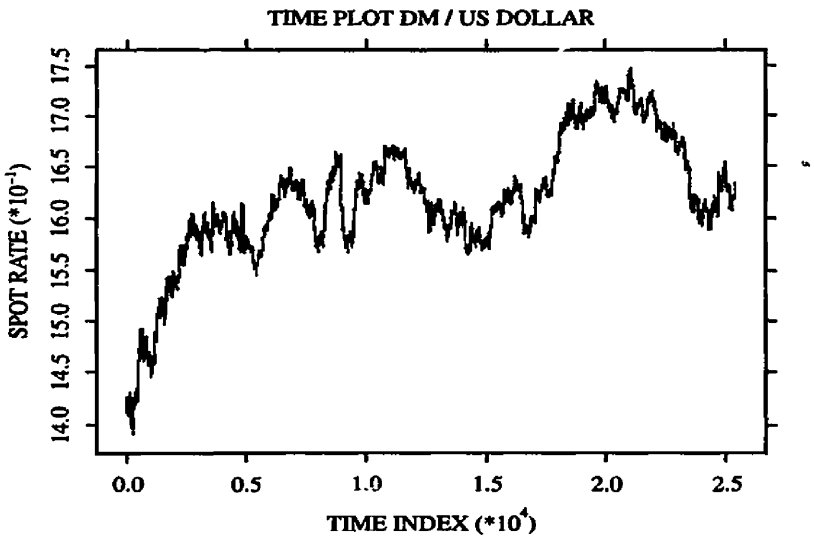


Fig. 1a. DM/US$ exchange rate series.

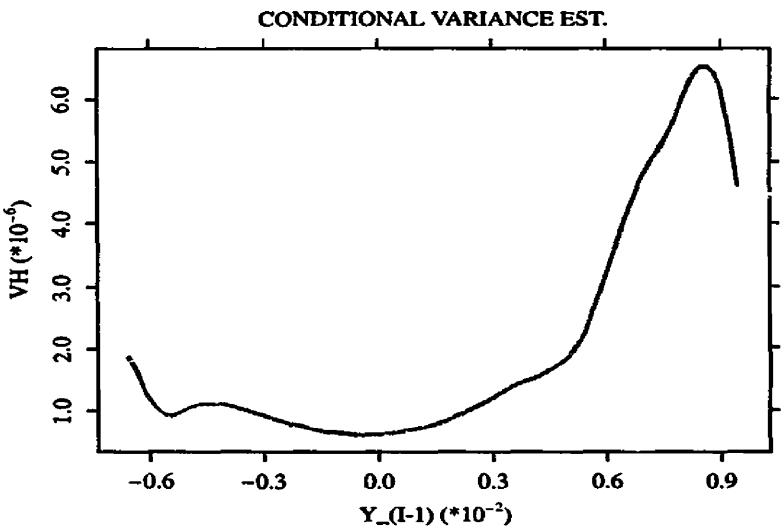


Fig. 1b. The estimated volatility function $\hat{v}(y_{i-1})$.

This is a straightforward modification of the local polynomial nonparametric regression estimator, as defined in Tsybakov (1986).

Fig. 1a shows us the DM/US-dollar foreign exchange rate from 1 October 1992 to 30 September 1993 in 20 min intervals. There are $n = 25,144$

Journal of Econometrics, 81, 223-242

observations. For a definition of 'time' in this series we refer to Bossaerts and Härdle (1995).

We formed the returns of this series and applied the estimator (2.5) to the time series of returns. The estimated volatility function as displayed in Fig. 1b shows a U-shaped structure, also called a 'smiling face'. It says that risks of returns are much higher for extreme values taken on the past day. There is a boundary effect on the right edge of the interval where the estimated volatility function decreases. This is due to only a few observations at this time end and is unavoidable in this context, see Müller (1988).

3. The result

Assume the following.

(A1) $\mathrm{E}(\xi_1^2)=1$, $\mathrm{E}(\xi_1)=\mathrm{E}(\xi_1^3)=0$, and $m_4=\mathrm{E}\{(\xi_1^2-1)^2\}<\infty$.

(A2) The density $p(\bullet)$ of ξ_1 exists and satisfies $\inf_{x\in\mathscr{X}} p(x)>0$ for any compact $\mathscr{X}\subset\mathbb{R}^1$.

(A3) There exist constants $C_1>0$, $C_2>0$ such that

$$|f(y)|\leqslant C_1(1+|y|), \tag{3.1}$$

$$|s(y)|\leqslant C_2(1+|y|), \quad y\in\mathbb{R}^1. \tag{3.2}$$

(A4) The function $s(\bullet)$ satisfies $\inf_{y\in\mathscr{X}} s(y)>0$, for any compact $\mathscr{X}\subset\mathbb{R}^1$.

(A5) $C_1+C_2\mathrm{E}|\xi_1|<1$.

Assumptions (A2), (A4) guarantee that the process (1.1) does not die out whereas (A3) and (A5) are conditions for $\{Y_i\}$ not to explode.

The following lemma given by Ango Nze (1992) guarantees ergodicity of the process $\{Y_i\}$. It is based on application of results of Nummeiin and Tuominen (1982) and Tweedie (1975).

Lemma 3.1. *Under the conditions* (A1)–(A5) *the Markov chain* $\{Y_i\}$ *is geometrically ergodic, i.e. it is ergodic, with stationary probability measure* $\pi(\bullet)$ *such that, for almost every* y,

$$\|\mathrm{P}^n(\bullet\,|\,y)-\pi(\bullet)\|_{\mathrm{TV}}=\mathrm{O}(\rho^n),$$

for some $0\leqslant\rho<1$. *Here* $\mathrm{P}^n(B|y)=\mathrm{P}\{Y_n\in B|Y_0=y\}$, *for a Borel subset* $B\subset\mathbb{R}^1$, *and* $\|\bullet\|_{\mathrm{TV}}$ *is the total variation distance.*

Now we state the conditions necessary to derive asymptotic normality of $\hat{v}_n(x)$ at a fixed point $x\in\mathbb{R}^1$.

(A6) The functions f and s are $(l-1)$ times continuously differentiable and there exist one-sided derivatives $f_{\pm}^{(l)}(x), s_{\pm}^{(l)}(x)$, at the point $x\in\mathbb{R}^1$.

Journal of Econometrics, 81, 223-242

(A7) The density $\mu(\bullet)$ of the stationary distribution $\pi(\bullet)$ exists, is bounded, continuous and strictly positive in a neighbourhood of the point x.

(A8) The kernel $K: \mathbb{R}^1 \to \mathbb{R}^+$ is a compactly supported bounded function, such that $K > 0$ on a set of positive Lebesgue measure.

(A9) $h_n = \beta n^{-1/(2l+1)}$, where $\beta > 0$.

(A10) The initial value Y_0 is a fixed number in $\mathbb{R}^1$.

Define the following matrices $A = \int F(u)F^{\mathrm{T}}(u)K(u)\,du$, $\Phi = \int F(u)F^{\mathrm{T}}(u) \times K^2(u)\,du$. Condition (A8) implies that A and Φ are positive definite, see Lemma 1 of Tsybakov (1986). Set $\mathscr{D} = A^{-1}\Phi A^{-1}$. Let

$$f^{(l)}(x;u) = \begin{cases} f_+^{(l)}(x), & u \geqslant 0, \\ f_-^{(l)}(x), & u < 0, \end{cases}$$

and define the asymptotic biases

$$b_f(x) = A^{-1}\frac{\beta^l}{l!}\int F(u)u^l K(u) f^{(l)}(x;u)\,du,$$

$$b_g(x) = A^{-1}\frac{\beta^l}{l!}\int F(u)u^l K(u) g^{(l)}(x;u)\,du.$$

Denote

$$c(x) = \begin{pmatrix} f(x) \\ f'(x)h_n \\ \vdots \\ f^{(l-1)}(x)h_n^{l-1} \end{pmatrix}, \qquad \bar{c}(x) = \begin{pmatrix} g(x) \\ g'(x)h_n \\ \vdots \\ g^{(l-1)}(x)h_n^{l-1} \end{pmatrix}.$$

Theorem 3.1. *Assume* (A1)–(A10). *Then*

$$\{\bar{c}_n(x) - \bar{c}(x)\}^{\mathrm{T}} F(0) \xrightarrow{\mathrm{P}} 0, \qquad \{c_n(x) - c(x)\}^{\mathrm{T}} F(0) \xrightarrow{\mathrm{P}} 0, \tag{3.3}$$

and

$$n^{l/(2l+1)} \begin{pmatrix} \bar{c}_n(x) - \bar{c}(x) \\ c_n(x) - c(x) \end{pmatrix} \xrightarrow{\mathscr{L}} \mathscr{N}\{b(x), \Sigma(x)\}, \tag{3.4}$$

as $n \to \infty$, *where*

$$b(x) = \begin{pmatrix} b_g(x) \\ b_f(x) \end{pmatrix},$$

$$\Sigma(x) = \frac{s^2(x)}{\beta\mu(x)} \begin{pmatrix} 4f^2(x) + s^2(x)m_4 & 2f(x) \\ 2f(x) & 1 \end{pmatrix} \otimes \mathscr{D}.$$

Journal of Econometrics, 81, 223-242

Here $\mathscr{D}' \otimes \mathscr{D}$ denotes the Kronecker product of matrices $\mathscr{D}'$ and $\mathscr{D}$. A simple consequence of this theorem is the following.

Theorem 3.2. *Assume* (A1)–(A10). *Then*

$$n^{l/(2l+1)}\{\hat{v}_n(x) - v(x)\} \xrightarrow{\mathscr{D}} \mathcal{N}\{b_v(x), \sigma_v^2(x)\},$$

as $n \to \infty$, *where*

$$b_v(x) = F^{\mathrm{T}}(0)\{b_g(x) - 2f(x)b_f(x)\},$$

$$\sigma_v^2(x) = \frac{s^4(x)m_4}{\beta\mu(x)} F^{\mathrm{T}}(0)\mathscr{D}F(0).$$

Consider the special case of $l = 2$, and assume that f and s are twice continuously differentiable, and that the kernel K satisfies $\int K(u)\,du = 1$, $K(u) = K(-u)$. Then

$$A = \begin{pmatrix} 1 & 0 \\ 0 & \sigma_K^2 \end{pmatrix} \quad \text{where } \sigma_K^2 = \int u^2 K(u)\,du,$$

$$\Phi = \begin{pmatrix} \int K^2(u)\,du & 0 \\ 0 & \int u^2 K^2(u)\,du \end{pmatrix},$$

$$b_f(x) = A^{-1}\frac{\beta^2 f''(x)}{2}\begin{pmatrix} \sigma_K^2 \\ 0 \end{pmatrix} = \begin{pmatrix} \sigma_K^2\beta^2 f''(x)/2 \\ 0 \end{pmatrix},$$

$$b_g(x) = A^{-1}\frac{\beta^2 g''(x)}{2}\begin{pmatrix} \sigma_K^2 \\ 0 \end{pmatrix} = \begin{pmatrix} \sigma_K^2\beta^2 g''(x)/2 \\ 0 \end{pmatrix},$$

$$\mathscr{D} = \begin{pmatrix} \int K^2(u)\,du & 0 \\ 0 & \dfrac{1}{\sigma_K^4}\int u^2 K^2(u)\,du \end{pmatrix},$$

$$b_v(x) = \frac{\sigma_K^2\beta^2}{2}[\{f^2(x) + s^2(x)\}'' - 2f(x)f''(x)]$$

$$= \frac{\sigma_K^2\beta^2}{2}[v''(x) + 2\{f'(x)\}^2],$$

$$\sigma_v^2(x) = \frac{s^4(x)m_4}{\beta\mu(x)}\int K^2(u)\,du = \frac{v^2(x)m_4}{\beta\mu(x)}\int K^2(u)\,du.$$

Journal of Econometrics, 81, 223-242

In particular, the normalized squared error of $\hat{v}_n(\bullet)$ calculated from the asymptotical distribution is

$$E\{n^{2l/2l+1}(\hat{v}_n(x) - v(x))^2\} \sim b_v^2(x) + \sigma_v^2(x)$$

$$= \frac{v^2(x)m_4}{\beta\mu(x)} \int K^2(u)\,du + \frac{\sigma_K^4\beta^4}{4}[v''(x) + 2\{f'(x)\}^2]^2.$$

Minimisation of this expression with respect to K and β leads to the Epanechnikov kernel $K(u) = K^*(u) = \frac{3}{4}(1-u^2)_+$ and to the following value of β:

$$\beta(K) = \left(\frac{v^2(x)m_4 \int K^2(u)\,du}{\mu(x)\sigma_K^4[v''(x) + 2\{f'(x)\}^2]^2}\right)^{1/5},$$

so that

$$\beta^* = \beta(K^*) = \left(\frac{15v^2(x)m_4}{\mu(x)[v''(x) + 2\{f'(x)\}^2]^2}\right)^{1/5}.$$

4. Simulations and finance applications

We did a simulation study of finite sample properties of the LP estimation method. We generated a process of the form (1.1) with the logistic mean function $f(x) = 1/\{1+\exp(-x)\}$, and volatility function $s(x) = \varphi(x+1.2) + 1.5\varphi(x-1.2)$, where φ denotes the pdf of standard normal distribution. The errors ξ_i were chosen to be uniformly distributed, the kernel was the quartic one, $K(u) = \frac{15}{16}(1-u^2)^2 I(|u| \leq 1)$, the local polynomial approximation was chosen to be linear ($l = 2$). The bandwidth was selected by cross-validation; $n = 1000$ observations were generated. The LP algorithm and all other computations were done in XploRe, see Härdle et al. (1995).

In Fig. 2a we show the time series as a scatterplot in the form (Y_{i-1}, Y_i) and the estimate $\hat{f}$ of the logistic mean function f. The little peak at the right is due to boundary effects. Fig. 2b gives the bimodal scaling function $s(x) = \{v(x)\}^{1/2}$, together with its estimate $\hat{s}(x) = \{\hat{v}(x)\}^{1/2}$. There are almost no observations on the right side as can be seen from Fig. 2a. The sparseness of the observations there is responsable for this jagged behaviour of $\hat{v}(x)$.

As an application, we report the YEN/DM foreign exchange rate. We used the same LP technique as described above but applied it to the returns time series as in the introductory example (Figs. 2a and b).

The YEN/DM series were computed in 20 min intervals as an average of the bid/ask spot rate. For details concerning this financial market application we refer to Bossaerts and Härdle (1995). The series is shown in Fig. 2c. The corresponding estimate $\hat{v}$ of the volatility function v is displayed in Fig. 2d. It shows again the

(1997) Härdle, W. and Tsybakov, A.
Local polynomial estimators of the volatility function in nonparametric autoregression.

Journal of Econometrics, 81, 223-242

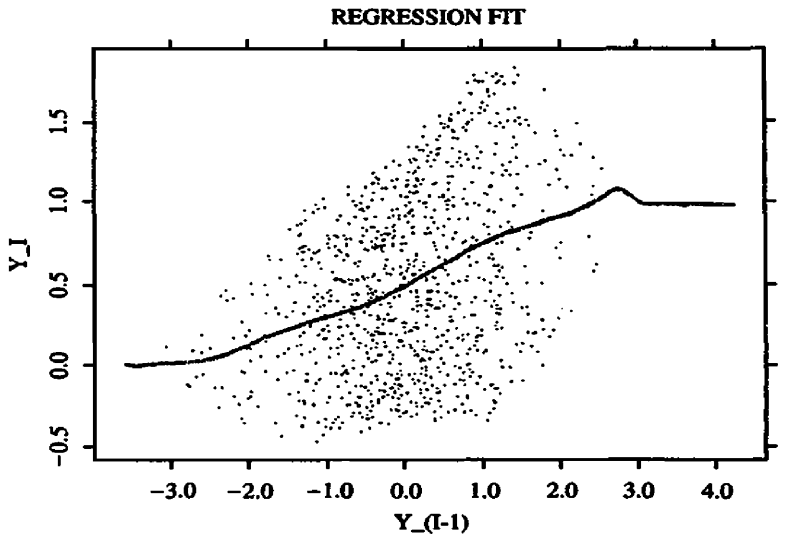


Fig. 2a. The time series and $\hat{f}$.

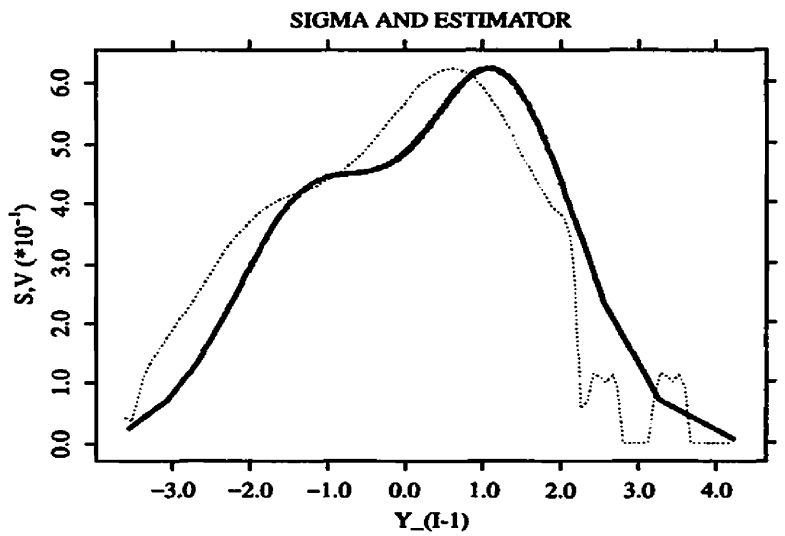


Fig. 2b. The true function $s(x)$ and the estimate $\{\hat{v}(x)\}^{1/2}$.

U-shaped form that we have seen already for the DM/US$ exchange rate. Note that the curve returns to zero at the boundaries. This effect is, as explained before, to be attributed to the small density $p(x)$ (see (A2)).

Journal of Econometrics, 81, 223-242

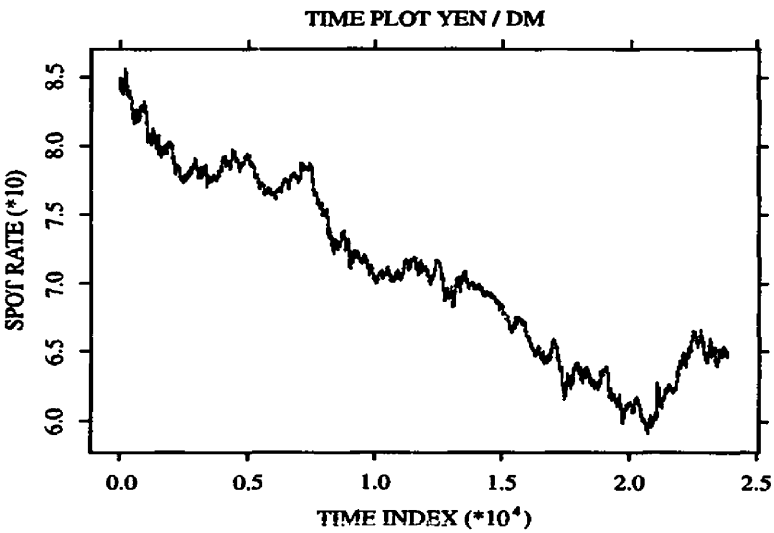


Fig. 2c. YEN/DM foreign exchange rate.

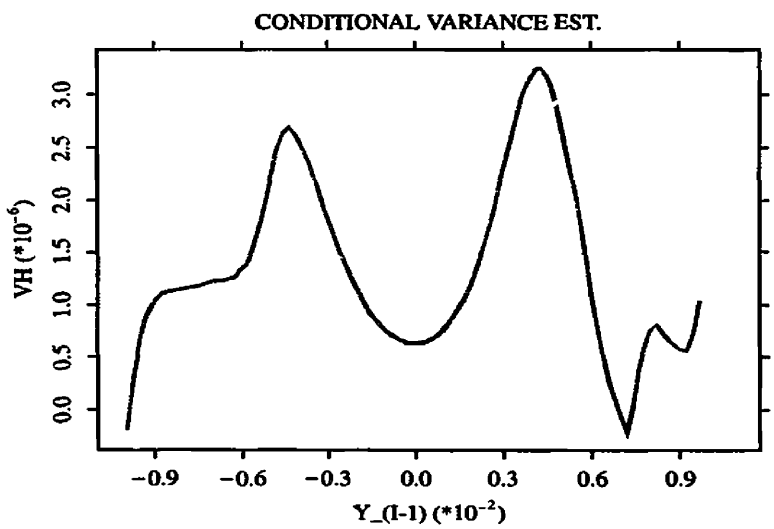


Fig. 2d. The volatility function for the YEN/DM exchange rate.

5. Proofs

Proof of Theorem 3.1. The normal equations for the first LS problem in (2.3) are

$$n^{l/(2l+1)}B_n\bar{c}_n(x) = n^{-l/(2l+1)}\sum_{i=1}^{n} Y_i^2 U_{in}K(u_{in}), \tag{5.1}$$

Journal of Econometrics, 81, 223-242

where the matrix

$$B_n = n^{-2l/(2l+1)} \sum_{i=1}^{n} U_{in} U_{in}^{\mathrm{T}} K(u_{in}).$$

On the other hand

$$n^{l/(2l+1)} B_n \bar{c}(x) = n^{-l/(2l+1)} \sum_{i=1}^{n} U_{in} U_{in}^{\mathrm{T}} \bar{c}(x) K(u_{in}), \tag{5.2}$$

and

$$Y_i^2 = g(Y_{i-1}) + 2f(Y_{i-1}) s(Y_{i-1}) \xi_i + s^2(Y_{i-1})(\xi_i^2 - 1). \tag{5.3}$$

By Taylor expansion of $g = f^2 + s^2$ we get

$$g(Y_{i-1}) - U_{in}^{\mathrm{T}} \bar{c}(x) = \frac{(Y_{i-1} - x)^l}{(l-1)!} \int_0^1 g(l)(x + t(Y_{i-1} - x))(1-t)^{l-1} \, \mathrm{d}t$$

$$\equiv r_a(Y_{i-1}, x). \tag{5.4}$$

From (5.1)–(5.4) we find

$$n^{l/(2l+1)} B_n \{\bar{c}_n(x) - \bar{c}(x)\} = n^{-l/(2l+1)} \sum_{i=1}^{n} \{Y_i^2 - U_{in}^{\mathrm{T}} \bar{c}(x)\} U_{in} K(u_{in})$$

$$= \bar{b}_n(x) + \bar{q}_n(x), \tag{5.5}$$

where

$$\bar{b}_n(x) = n^{-l/(2l+1)} \sum_{i=1}^{n} r_g(Y_{i-1}, x) U_{in} K(u_{in}),$$

$$\bar{q}_n(x) = n^{-l/(2l+1)} \sum_{i=1}^{n} \alpha_i U_{in} K(u_{in}),$$

and

$$\alpha_i = 2f(Y_{i-1}) s(Y_{i-1}) \xi_i + s^2(Y_{i-1})(\xi_i^2 - 1).$$

Calculations similar to (5.1)–(5.5) give

$$n^{l/(2l+1)} B_n \{c_n(x) - c(x)\} = b_n(x) + q_n(x), \tag{5.6}$$

where

$$b_n(x) = n^{-l/(2l+1)} \sum_{i=1}^{n} r_f(Y_{i-1}, x) U_{in} K(u_{in}),$$

$$q_n(x) = n^{-l/(2l+1)} \sum_{i=1}^{n} \beta_i U_{in} K(u_{in}),$$

Journal of Econometrics, 81, 223-242

with $\beta_i = s(Y_{i-1})\xi_i$. The proof of Theorem 3.1 will be based on the following steps.

First, we show that, elementwise,

$$B_n \xrightarrow{\mathrm{p}} B, \tag{5.7}$$

as $n \to \infty$, where $B = \beta\mu(x)A$ is a positive definite matrix. Next we show the relations

$$\bar{b}_n(x) \xrightarrow{\mathrm{p}} Bb_g(x) \quad \text{as } n \to \infty, \tag{5.8}$$

$$b_n(x) \xrightarrow{\mathrm{p}} Bb_f(x) \quad \text{as } n \to \infty, \tag{5.9}$$

Finally, we show that the compound random vector is asymptotically normal:

$$\begin{pmatrix} \bar{q}_n(x) \\ q_n(x) \end{pmatrix} \xrightarrow{\mathscr{L}} \mathcal{N}(0, \Sigma_0) \quad \text{as } n \to \infty, \tag{5.10}$$

where

$$\Sigma_0 = \{s^2(x)\beta\mu(x)\} \begin{pmatrix} 4f^2(x) + s^2(x)m_4 & 2f(x) \\ 2f(x) & 1 \end{pmatrix} \otimes \Phi.$$

Together (5.7)–(5.10) and the relation (cf. (5.5) and (5.6))

$$n^{l/(2l+1)} B_n \begin{pmatrix} \bar{c}_n(x) - \bar{c}(x) \\ c_n(x) - c(x) \end{pmatrix} = \begin{pmatrix} \bar{b}_n(x) \\ b_n(x) \end{pmatrix} + \begin{pmatrix} \bar{q}_n(x) \\ q_n(x) \end{pmatrix}$$

entail (3.4). To prove (3.3) it suffices to show that

$$n^{-l/(2l+1)} q_n^{\mathrm{T}}(x) F(0) \xrightarrow{\mathrm{p}} 0, \tag{5.11}$$

$$n^{-l/(2l+1)} \bar{q}_n^{\mathrm{T}}(x) F(0) \xrightarrow{\mathrm{p}} 0$$

as $n \to \infty$. In fact, combining (5.11) and (5.5)–(5.9) yields (3.3). It remains to prove (5.7) to (5.11).

We will need some auxiliary results.

Lemma 5.1 (Davydov, 1973). *Let $\{Y_i\}$ be a geometrically ergodic Markov chain, where Y_0 is distributed with its stationary distribution $\pi(\bullet)$. Then the chain is geometrically strongly mixing with the mixing coefficients satisfying $\alpha(n) \leqslant c_0 \rho_0^n$ for some $0 < \rho_0 < 1$, $c_0 > 0$.*

Denote $\mathscr{F}_k = \sigma(Y_k, Y_{k-1}, \ldots, Y_0)$ the σ-algebra generated by $Y_0, \ldots, Y_k$.

Lemma 5.2 (Liptser and Shirjaev, 1980, Corollary 6). *Let for every $n > 0$, the sequence $\eta^n = (\eta_{nk}, \mathscr{F}_k)$ be a square integrable martingale difference, i.e.*

$$\mathrm{E}(\eta_{nk} | \mathscr{F}_{k-1}) = 0, \quad \mathrm{E}(\eta_{nk}^2) < \infty, \quad 1 \leqslant k \leqslant n. \tag{5.12}$$

Journal of Econometrics, 81, 223-242

and let

$$\sum_{k=1}^{n} \mathrm{E}(\eta_{nk}^2) = 1, \quad \forall n \geqslant n_0 > 0, \tag{5.13}$$

The conditions

$$\sum_{k=1}^{n} \mathrm{E}(\eta_{nk}^2 | \mathscr{F}_{k-1}) \xrightarrow{\mathrm{p}} 1 \quad \text{as } n \to \infty, \tag{5.14}$$

$$\sum_{k=1}^{n} \mathrm{E}(\eta_{nk}^2 I(|\eta_{nk}| > \varepsilon) | \mathscr{F}_{k-1}) \xrightarrow{\mathrm{p}} 0 \quad \text{as } n \to \infty, \tag{5.15}$$

$(\forall \varepsilon > 0)$ *are sufficient for convergence*

$$\sum_{k=1}^{n} \eta_{nk} \xrightarrow{\mathscr{D}} \mathscr{N}(0,1) \quad \text{as } n \to \infty.$$

Now we proceed to the proof of (5.7)–(5.11). Introduce some more notation and define the matrices

$$\Phi_n^{11} = n^{-2l/(2l+1)} \sum_{i=1}^{n} \mathrm{E}(\alpha_i^2 | \mathscr{F}_{i-1}) U_{in} U_{in}^{\mathrm{T}} K^2(u_{in}),$$

$$\Phi_n^{12} = n^{-2l/(2l+1)} \sum_{i=1}^{n} \mathrm{E}(\alpha_i \beta_i | \mathscr{F}_{i-1}) U_{in} U_{in}^{\mathrm{T}} K^2(u_{in}),$$

$$\Phi_n^{22} = n^{-2l/(2l+1)} \sum_{i=1}^{n} \mathrm{E}(\beta_i^2 | \mathscr{F}_{i-1}) U_{in} U_{in}^{\mathrm{T}} K^2(u_{in}),$$

and the compound matrix

$$\Sigma_n = \begin{pmatrix} \Phi_n^{11} & \Phi_n^{12} \\ \Phi_n^{12} & \Phi_n^{22} \end{pmatrix}.$$

Lemma 5.3. *Under the conditions of Theorem* 3.1 *we have*

$$n^{-2l/(2l+1)} \sum_{i=1}^{n} \varphi_1(Y_{i-1}) \varphi_2(u_{in}) K(u_{in}) \xrightarrow{\mathrm{p}} \beta \mu(x) \varphi_1(x) \int \varphi_2(u) K(u)\, \mathrm{d}u$$

$$n^{-2l/(2l+1)} \sum_{i=1}^{n} \mathrm{E}\{\varphi_1(Y_{i-1}) \varphi_2(u_{in}) K(u_{in})\} \to \beta \mu(x) \varphi_1(x) \int \varphi_2(u) K(u)\, \mathrm{d}u, \tag{5.16}$$

as $n \to \infty$, *provided* $\varphi_1(\bullet)$ *is a bounded continuous function and* $\varphi_2(\bullet)$ *is a bounded function.*

Proof. Let $\{Y_i^*\}$ be a Markov chain satisfying (1.2), such that $Y_0 = Y_0^*$ has the stationary distribution $\pi(\bullet)$ introduced in Lemma 3.1. This chain is stationary

Journal of Econometrics, 81, 223-242

and by Lemma 5.1 it is geometrically strongly mixing (α-mixing). Therefore

$$n^{-2l/(2l+1)} \sum_{i=1}^{n} \varphi_1(Y^*_{i-1})\varphi_2(u^*_{in})K(u^*_{in}) - n^{1/(2l+1)} \times \mathrm{E}\{\varphi_1(Y^*_i)\varphi_2(u^*_{1n})K(u^*_{1n})\} \xrightarrow{\mathrm{p}} 0, \tag{5.17}$$

as $n \to \infty$, where $u^*_{in} = (Y^*_{i-1} - x)/h_n$. Now,

$$\begin{aligned} & n^{1/(2l+1)}\mathrm{E}\{\varphi_1(Y^*_i)\varphi_2(u^*_{1n})K(u^*_{1n})\} \\ &= \beta \frac{1}{h_n} \int \varphi_1(y)\varphi_2\left(\frac{y-x}{h_n}\right) K\left(\frac{y-x}{h_n}\right) \mu(y)\,\mathrm{d}y \\ &= \beta\mu(x)\varphi_1(x)\int \varphi_2(u)K(u)\,\mathrm{d}u\{1 + \mathrm{o}(1)\}, \end{aligned} \tag{5.18}$$

as $n \to \infty$. On the other hand, denoting for brevity

$$\zeta_i = \varphi_1(Y_{i-1})\varphi_2(u_{in})K(u_{in}),$$

$$\zeta^*_i = \varphi_1(Y^*_{i-1})\varphi_2(u^*_{in})K(u^*_{in}),$$

and choosing an integer $\gamma_n = \mathrm{o}(n^{(2l/(2l+1))})$, such that $\gamma_n \to \infty$ as $n \to \infty$, we get

$$\begin{aligned} & n^{-2l/(2l+1)} \sum_{i=1}^{n} |\mathrm{E}(\zeta_i - \zeta^*_i)| \\ &\leqslant n^{-2l/(2l+1)} \left[\sum_{i=1}^{\gamma_n - 1} |\mathrm{E}(\zeta_i - \zeta^*_i)| + \sum_{i=\gamma_n}^{n} |\mathrm{E}(\zeta_i - \zeta^*_i)| \right] \\ &\leqslant 2n^{-2l/(2l+1)}\gamma_n \|\varphi_1\varphi_2 K\|_\infty + n^{-2l/(2l+1)} \sum_{i=\gamma_n}^{n} |\mathrm{E}(\zeta_i - \zeta^*_i)| \\ &= n^{-2l/(2l+1)} \sum_{i=\gamma_n}^{n} |\mathrm{E}(\zeta_i - \zeta^*_i)| + \mathrm{o}(1) \quad \text{as } n \to \infty. \end{aligned} \tag{5.19}$$

In view of geometric ergodicity of $\{Y_i\}$ (Lemma 3.1) we have

$$\begin{aligned} & n^{-2l/(2l+1)} \sum_{i=\gamma_n}^{n} |\mathrm{E}(\zeta_i - \zeta^*_i)| \\ &= n^{-2l/(2l+1)} \sum_{i=\gamma_n}^{n} |\mathrm{E}\{\varphi_1(Y_{i-1})\varphi_2(u_{in})K(u_{in}) \\ &\quad - \varphi_1(Y^*_{i-1})\varphi_2(u^*_{in})K(u^*_{in})\}| \end{aligned}$$

Journal of Econometrics, 81, 223-242

$$\leqslant n^{-2l/(2l+1)} \sum_{i=\gamma_n}^{n} \|\varphi_1\varphi_2 K\|_\infty \int |\mu_i(y) - \mu(y)|\,dy$$

$$= O\left(n^{-2l/(2l+1)} \sum_{i=\gamma_n}^{n} \rho^i\right) = o(1) \quad \text{as } n \to \infty, \tag{5.20}$$

where $\mu_i(\bullet)$ is the density of Y_{i-1}. Applying Markov's inequality and combining (5.17)–(5.20) we get (5.16).

The correctness of (5.7) is shown in

Lemma 5.4. *Under the conditions of Theorem* 3.1 *we have elementwise*

$$B_n \xrightarrow{p} B = \beta\mu(x)A \quad \text{as } n \to \infty,$$

and

$$\Sigma_n \xrightarrow{p} \Sigma_0, \quad E(\Sigma_n) \to \Sigma_0 \quad \text{as } n \to \infty.$$

Proof. The elements of matrices B_n and Σ_n are of the form

$$n^{-2l/(2l+1)} \sum_{i=1}^{n} \varphi_1(Y_{i-1})\varphi_2(u_{in})K(u_{in})$$

where $\varphi_1(\bullet)$ is a bounded continuous function and $\varphi_2(\bullet)$ is a bounded function. Applying Lemma 5.3, we get the result. In particular, for Σ_n the functions $\varphi_1(Y_{i-1})$ are of the form

$$E(\alpha_i^2|\mathcal{F}_{i-1}) = 4f^2(Y_{i-1})s^2(Y_{i-1}) + s^4(Y_{i-1})m_4,$$
$$E(\alpha_i\beta_i|\mathcal{F}_{i-1}) = 2f(Y_{i-1})s^2(Y_{i-1}),$$
$$E(\beta_i^2|\mathcal{F}_{i-1}) = s^2(Y_{i-1}),$$

where we used \1). Note that, by (A6) these functions are continuous and bounded in any neighbourhood of x. Since K is compactly supported, it suffices to have bounde ness and continuity of φ_1 in a neighbourhood of x, and thus these particular examples of $\varphi_1(\bullet)$ satisfy the conditions of Lemma 5.3.

Let us now prove (5.8) and (5.9).

Lemma 5.5. *Under the conditions of Theorem* 3.1 *we have* (5.8) *and* (5.9).

Proof. Consider (5.8) only, since the proof of (5.9) is quite similar. Since s and f satisfy (A6), it is clear that $g = f^2 + s^2$ also does. Note that

$$r_g(Y_{i-1}, x) = u_{in}^l h_n^l \frac{1}{(l-1)!} \int_0^1 g^{(l)}\{x + t(Y_{i-1} - x)\}(1-t)^{l-1}\,dt$$
$$= u_{in}^l n^{-l/(2l+1)} \varphi_3(Y_{i-1}),$$

Journal of Econometrics, 81, 223-242

where

$$\varphi_3(Y_{i-1}) = \frac{\beta^l}{(l-1)!}\int_0^1 g^{(l)}\{x + t(Y_{i-1} - x)\}(1-t)^{l-1}\,dt,$$

and thus

$$\bar{b}_n(x) = n^{-2l/(2l+1)}\sum_{i=1}^n \varphi_3(Y_{i-1})u_{in}^l U_{in} K(u_{in}).$$

The elements of $\bar{b}_n(x)$ are of the form described in Lemma 5.3. Following (5.19) and (5.20), we get, with $U_{in}^* = F(u_{in}^*)$,

$$\bar{b}_n(x) - n^{-2l/(2l+1)}\sum_{i=1}^n \varphi_3(Y_{i-1}^*)(u_{in}^*)^l U_{in}^* K(u_{in}^*) \xrightarrow{p} 0, \tag{5.21}$$

as $n \to \infty$. Since $\{Y_i^*\}$ is α-mixing, we get, as in (5.17),

$$n^{-2l/(2l+1)}\sum_{i=1}^n \varphi_3(Y_{i-1}^*)(u_{in}^*)^l U_{in}^* K(u_{in}^*)$$

$$- n^{1/(2l+1)}\mathrm{E}\{\varphi_3(Y_1^*)(u_{1n}^*)^l U_{1n}^* K(u_{1n}^*)\} \xrightarrow{p} 0,$$

as $n \to \infty$.

To end the proof, note that

$$n^{1/(2l+1)}\mathrm{E}\{\varphi_3(Y_1^*)(u_{1n}^*)^l U_{1n}^* K(u_{1n}^*)\}$$

$$= \beta\int \varphi_3(x + uh_n)u^l F(u)K(u)\mu(x + uh_n)\,du,$$

and that

$$\lim_{n\to\infty} \varphi_3(x + uh_n) = \beta^l g^{(l)}(x;u)/l \tag{5.22}$$

for any $u \in \mathbb{R}^1$. Using (5.22) and (A7), we find

$$\lim_{n\to\infty} \beta\int \varphi_3(x + uh_n)u^l F(u)K(u)\mu(x + uh_n)\,du$$

$$= \frac{\beta^{l+1}}{l!}\left(\int F(u)u^l K(u)g^{(l)}(x;u)\,du\right)\mu(x)$$

$$= \{A\mu(x)\beta\}b_g(x) = Bb_g(x).$$

This proves the lemma. □

Lemma 5.6. *Under the conditions of Theorem* 3.1 *we have* (5.11).

Journal of Econometrics, 81, 223-242

Proof. For the sake of brevity, we show only that $n^{-l/(2l+1)}q_n^{\mathrm{T}}(x)F(0) \xrightarrow{\mathrm{P}} 0$, as $n \to \infty$. We have

$$n^{-l/(2l+1)}q_n^{\mathrm{T}}(x)F(0) = n^{-2l/(2l+1)} \sum_{i=1}^{n} \beta_i U_{in}^{\mathrm{T}} F(0) K(u_{in})$$

$$= n^{-2l/(2l+1)} \sum_{i=1}^{n} \beta_i K(u_{in}) = n^{-2l/(2l+1)} \sum_{i=1}^{n} (\beta_i - \mathrm{E}(\beta_i|\mathscr{F}_{i-1}))K(u_{in}).$$

By (A8), the value $d^* = \max\{|u| : u \in \operatorname{supp} K\} < \infty$ and K is bounded. Hence,

$$\mathrm{E}([n^{-l/(2l+1)}q_n^{\mathrm{T}}(x)F(0)]^2)$$

$$= n^{-4l/(2l+1)}\mathrm{E}\left(\left[\sum_{i=1}^{n}(\beta_i - \mathrm{E}(\beta_i|\mathscr{F}_{i-1}))K(u_{in})\right]^2\right)$$

$$\leqslant \kappa_0 n^{-4l/(2l+1)} \sum_{i=1}^{n} \mathrm{E}([\beta_i - \mathrm{E}(\beta_i|\mathscr{F}_{i-1})]^2 I\{|u_{in}| \leqslant d^*\}),$$

where $\kappa_0 > 0$ is a constant. Now,

$$\mathrm{E}([\beta_i - \mathrm{E}(\beta_i|\mathscr{F}_{i-1})]^2 I\{|u_{in}| \leqslant d^*\})$$

$$= \mathrm{E}\left(s^2(Y_{i-1})\xi_i^2 I\left\{\frac{|Y_{i-1} - x|}{h_n} \leqslant d^*\right\}\right)$$

$$= \mathrm{E}\left(s^2(Y_{i-1}) I\left\{\frac{|Y_{i-1} - x|}{h_n} \leqslant d^*\right\}\right) \leqslant \sup_{|y-x| \leqslant h_n d^*} s^2(y) < \infty,$$

if n is large enough. This yields the lemma. □

To prove Theorem 3.1 it remains to show (5.10).

Proof of (5.10). By the Cramér–Wold device it suffices to prove

$$a^{\mathrm{T}}\begin{pmatrix} \bar{q}_n(x) \\ q_n(x) \end{pmatrix} \xrightarrow{\mathscr{D}} \mathscr{N}(0, a^{\mathrm{T}}\Sigma_0 a) \quad n \to \infty \tag{5.23}$$

for any vector $a \in \mathbb{R}^{2l}$ with unit Euclidean norm: $|a| = 1$.

Fix such a vector a, and let n_0 be the integer such that $\mathrm{E}(\Sigma_n) > \frac{1}{2}\Sigma_0$ for all $n \geqslant n_0$. Such an integer n_0 exists since, by Lemma 5.4, $\mathrm{E}(\Sigma_n) \to \Sigma_0 > 0$ elementwise, as $n \to \infty$. From now on consider only $n \geqslant n_0$. Denote

$$\eta_{ni} = \frac{n^{-l/(2l+1)}}{\sqrt{a^{\mathrm{T}}\mathrm{E}(\Sigma_n)a}} a^{\mathrm{T}}\begin{pmatrix} \alpha_i U_{in} \\ \beta_i U_{in} \end{pmatrix} K(u_{in}).$$

Then

$$\sum_{i=1}^{n} \eta_{ni} = \frac{1}{\sqrt{a^{\mathrm{T}}\mathrm{E}(\Sigma_n)a}} a^{\mathrm{T}}\begin{pmatrix} \bar{q}_n(x) \\ q_n(x) \end{pmatrix},$$

Journal of Econometrics, 81, 223-242

and the convergence (5.23) is equivalent to the convergence

$$\sum_{k=1}^{n} \eta_{nk} \xrightarrow{\mathcal{L}} \mathcal{N}(0,1) \quad n \to \infty.$$

Let us prove this relation by using Lemma 5.2. Let us check the conditions (5.12)–(5.15) of Lemma 5.2. First, $\mathrm{E}(\alpha_i|\mathcal{F}_{i-1}) = 0, \mathrm{E}(\beta_i|\mathcal{F}_{i-1}) = 0$ (a.s.), and thus the equality in (5.12) holds. It is easy to check that

$$\sum_{k=1}^{n} \mathrm{E}(\eta_{nk}^2|\mathcal{F}_{k-1}) = \frac{a^{\mathrm{T}}\Sigma_n a}{a^{\mathrm{T}}\mathrm{E}(\Sigma_n)a}.$$

Hence (5.13) holds, and Lemma 5.4 gives (5.14). It remains to prove (5.15). Since $n \geqslant n_0$, we obtain

$$\eta_{nk}^2 \leqslant \frac{n^{-2l/(2l+1)}}{a^{\mathrm{T}}\mathrm{E}(\Sigma_n)a}(a^{\mathrm{T}}Z_{nk})^2 \leqslant \frac{2n^{-2l/(2l+1)}}{a^{\mathrm{T}}\Sigma_0 a}(a^{\mathrm{T}}Z_{nk})^2 \leqslant \kappa_1 n^{-2l/(2l+1)}|Z_{nk}|^2,$$

where $\kappa_1 > 0$ is a constant and

$$Z_{nk} = \begin{pmatrix} \alpha_k U_{kn} \\ \beta_k U_{kn} \end{pmatrix} K(u_{kn}).$$

Therefore, using the fact that K is bounded and compactly supported, and f and s are locally bounded, we find

$$\begin{aligned} \eta_{nk}^2 &\leqslant \kappa_1 n^{-2l/(2l+1)}(\alpha_k^2 + \beta_k^2)|U_{kn}|^2 K^2(u_{kn}) \\ &\leqslant \kappa_2 n^{-2l/(2l+1)}(1 + |\xi_k|^4)K(u_{kn}), \end{aligned}$$

where $\kappa_2 > 0$ is a constant. Hence

$$\begin{aligned} &\mathrm{E}(\eta_{nk}^2 I(|\eta_{nk}| \geqslant \varepsilon)\,|\,\mathcal{F}_{k-1}) \\ &\quad \leqslant \kappa_2 n^{-2l/(2l+1)} K(u_{kn}) \mathrm{E}[\{1+|\xi_1|^4\} I\{\sqrt{1+|\xi_1|^4} \geqslant \varepsilon n^{l/(2l+1)} \kappa_2^{-1/2} \|K\|_\infty^{-1/2}\}] \\ &\quad = \kappa_2 n^{-2l(2l+1)} K(u_{kn}) \cdot \mathrm{o}(1), \end{aligned}$$

as $n \to \infty$, where $\mathrm{o}(1)$ does not depend on k. This entails

$$\sum_{k=1}^{n} \mathrm{E}(\eta_{nk}^2 I(|\eta_{nk}| \geqslant \varepsilon)\,|\,\mathcal{F}_{k-1}) \leqslant \mathrm{o}(1) \sum_{k=1}^{n} n^{-2l/(2l+1)} K(u_{kn}) \quad \text{as } n \to \infty. \tag{5.24}$$

By Lemma 5.3

$$n^{-2l/(2l+1)} \sum_{k=1}^{n} K(u_{kn}) \xrightarrow{\mathrm{p}} \beta\mu(x) \int K(u)\,\mathrm{d}u, \quad \text{as } n \to \infty. \tag{5.25}$$

Using (5.24) and (5.25), we obtain (5.15). This proves the theorem. □

Journal of Econometrics, 81, 223-242

Proof of Theorem 3.2. We have

$$\{\hat{v}_n(x) - v(x)\} = \{c_n(x) - \bar{c}(x)\}^{\mathrm{T}} F(0) - [2c(x)^{\mathrm{T}} F(0) + \{c_n(x) - c(x)\}^{\mathrm{T}} F(0)][\{c_n(x) - c(x)\}^{\mathrm{T}} F(0)].$$

By (3.3),

$$(c_n(x) - c(x))^{\mathrm{T}} F(0) \xrightarrow{\mathrm{p}} 0 \quad \text{as } n \to \infty.$$

Hence,

$$n^{l/(2l+1)}\{\hat{v}_n(x) - v(x)\} = n^{l/(2l+1)}\{\bar{c}_n(x) - \bar{c}(x)\}^{\mathrm{T}} F(0) - [2c(x)^{\mathrm{T}} F(0) + o_{\mathrm{p}}(1)] n^{l/(2l+1)}(c_n(x) - c(x))^{\mathrm{T}} F(0),$$

as $n \to \infty$. It remains to note that $c(x)^{\mathrm{T}} F(0) = f(x)$ and to apply (3.4). □

Acknowledgements

We would like to thank Christian Hafner for computational help. The research was supported by Sonderforschungsbereich 'Quantifikation und Simulation Ökonomischer Prozesse' and INRA and INSEE, France.

References

Ango Nze, P., 1992. Critères d'ergodicité de quelques modèles à représentation markovienne. Comptes Rendus des Seances de l'Academie des Sciences Paris 315, sér 1, 1301–1304.

Bossaerts, P., Härdle, W., 1995. Foreign Exchange-rates have surprising volatility. SFB 373 Discussion Paper, available via FTP: amadeus.wiwi.hu.berlin.de (141.20.100.2).

Chan, K.S., Tong, H., 1985. On the use of deterministic Lyapunov functions for the ergodicity of stochastic difference equations. Advances in Applied Probability 17, 666–678.

Chen, R., Tsay, R.S., 1993a. Nonlinear additive ARX models. Journal of the American Statistical Association 88, 955–967.

Chen, R., Tsay, R.S., 1993b. Functional-coefficient autoregressive models. Journal of the American Statistical Association 88, 298–308.

Cleveland, W.S., 1979. Robust locally weighted regression and smoothing scatterplots. JASA 74, 829–836.

Collomb, 1984. Propriétés de convergence presque complète du prédicteur à noyau. Zeitschrift für Wahrscheinlichkeitstheorie und verwandte Gebiete 66, 441–460.

Davydov, Yu.A., 1973. Mixing conditions for Markov chains. Theory of Probability and its Applications 18, 312–328.

Diebolt, J., Guégan, D., 1990. Probabilistic properties of the general nonlinear autoregressive process of order one. Technical Report N°128, L.S.T.A., Université Paris VI.

Doukhan, P., Ghindès, M., 1980. Estimation dans le processus $X_{n+1} = f(X_n) + \varepsilon_{n+1}$. Comptes Rendus des Seances de l'Academie des Sciences Paris, Sér. A 297, 61–64.

Doukhan, P., Ghindès, M., 1981. Processus autorégressifs non-linéaires. Comptes Rendus des Seances de l'Academie des Sciences Paris, Sér. A 290, 921–923.

Journal of Econometrics, 81, 223-242

Doukhan, P., Tsybakov, A.B., 1993. Nonparametric robust estimation in nonlinear ARX models. Problems of Information Transmission 29, 24–34.

Engle, R.F., 1982. Autoregressive conditional heteroscedasticity with estimates of the variance of U.K. inflation. Econometrica 50, 987–1008.

Engle, R.F., Gonzales-Rivera, G., 1989. Semiparametric ARCH Models. U.C.S.D. Discussion Paper.

Engle, R.F., Granger, W.J., Rice, J., Weiss, A., 1986. Semiparametric estimates of the relation between weather and electricity sales. Journal of the American Statistical Association 81, 310–320.

Fan, J., Gijbels, I., 1996. Local Polynomial Modelling. Chapman and Hall, London.

Gouriéroux, Ch., Monfort, A., 1992. Qualitative threshold ARCH models. Journal of Econometrics 52, 159–199.

Gregory, A.W., 1989. A nonparametric test for autoregressive conditional heteroscedasticity: a Markov chain approach. Journal of Business and Economic Statistics 7, 107–115.

Härdle, W., Klinke, S., Turlach, B., 1995. XploRe – an Interactive Statistical Computing Environment. Springer, Heidelberg.

Katkovnik, V.Ya., 1979. Linear and nonlinear methods of nonparametric regression analysis, Automatika, 35–46.

Katkovnik, V.Ya., 1985. Nonparametric Identification and Data Smoothing. Nauka, Moscow (in Russian).

Liptser, R.Sh., Shirjaev, A.N., 1980. A functional central limit theorem for martingales. Theory of Probability and its Applications 25, 667–688.

Lütkepohl, H., 1992. Introduction to Multiple Time Series Analysis. Springer, Heidelberg.

McKeague, I.W., Zhang, M.J., 1994. Identification of nonlinear time series from first order cumulative characteristics. Annals of Statistics 22, 495–514.

Mokkadem, A., 1987. Sur un modèle autorégressif nonlinéaire. Ergodicité et ergodicité géometrique. Journal of Time Series Analysis 8, 195–204.

Müller, H.G., 1988. Nonparametric Regression Analysis of Longitudinal Data, Springer Lecture Notes in Statistics, vol. 46. Springer, New York.

Nummelin, E., Tuominen, P., 1982. Geometric ergodicity of Harris-recurrent Markov chains with application to renewal theory. Stochastic Processes and their Applications 12, 187–202.

Robinson, P.M., 1983. Nonparametric estimators for time series. J. Time Series Analysis 4, 185–207.

Robinson, P.M., 1984. Robust nonparametric autoregression, In: Franke, H., Martin (Eds.), Robust and Nonlinear Time Series Analysis. Springer, Heidelberg.

Stone, C.J., 1977. Consistent nonparametric regression. Ann. Statist. 5, 595–645.

Tong, H., 1983. Threshold Models in Nonlinear Time Series Analysis, Lecture Notes in Statistics, vol. 21. Springer, Heidelberg.

Tweedie, R.L., 1975. Sufficient conditions for ergodicity and geometric ergodicity of Markov chains on a general state space. Stochastic Process and their Applications 3, 385–403.

Tsybakov, A.B., 1986. Robust reconstruction of functions by the local-approximation method. Problems of Information Transmission 22, 133–146.

Vieu, P., 1995. Order choice in Nonlinear Autoregressive Models. Discussion Paper, Laboratoire de Statistique et Probabilités, Université Toulouse.

Computational Statistics, 13, 141-151
Computational Statistics (1998) 13:141–151

Teaching Wavelets in XploRe[1]

Sigbert Klinke[2], Yuri Golubev[3], Wolfgang Härdle[2] and Michael H. Neumann[4]

[2] Humboldt University of Berlin, Department of Economics, Institute of Statistics and Econometrics, Spandauer Strasse 1, D-10178 Berlin, Germany
[3] Institute for Problems of Information Transmission, Bolshoi Karetny 19, Moscow, Russia
[4] Weierstrass Institute for Applied Analysis and Stochastics, Mohrenstrasse 39, D-10117 Berlin, Germany

Summary

Teachware is a set of computer software tools for computeraided interactive teaching of certain knowledge elements. The construction of teachware for statistical knowledge is a rather young field since it heavily depends on data structures and graphical interactive possibilities. In this paper we present a teachware module for XploRe - a statistical computing environment. We focus on the situation of teaching wavelets, a technique for adaptation of spatial inhomogenuity.

Keywords: teachware, wavelets, interactive HMTL

[1] You can contact the authors and access the wavelet tutorial via WWW (`http://wotan.wiwi.hu-berlin.de`.

(1997) Klinke, S., Golubev, Y., Härdle, W. and Neumann, M. Teaching Wavelets in XploRe.

Computational Statistics, 13, 141-151

1 Introduction

1.1 Teachware

Over the past 15 years many programs have been written and developed which aim at assisting a teacher at conveying topics of statistical science to students or pupils. Koch & Haag (1996) list in their "Statistical Software Guide 95/96" some programs which name themselves teachware or learnware. We list some of them:

PRISTAT 1 an interactive program designed for both general purpose analysis and education. It is based on the book of Kolev (1993),

SchoolStat a MacIntosh shareware program on a spreadsheet basis,

Sila a tool for teaching students how the logic of inference in statistics works.

Many more teachware programs available for several platforms. One of the first programs is described ine Bowman & Robinson (1989), Bowman & Robinson (1990). Proenca (1995) describes a teachware for interactive linear regression and smoothing.

Many of these teachwares are based on systems like Toolbook which are more like a weakly dynamic textbook. We say weakly dynamic, since desirable links to the (underlying) statistical software are not an element of such teachware. We may change certain program elements but we are not able to exchange the dataset, for example.

If we allow for external effects like dataset exchange by the student we risk to have no self contained and self explained system. We must therefore introduce a second level where a teacher the explains also the handling of the underlying statistical computing environment. That makes it necessary to jump between direct software use and the explanation of how to use it. For this task we found the HTML technology the most suited one.

1.2 Wavelets

In order to demonstrate our concept, we present a tutorial on wavelets. The application of wavelet ideas to nonparametric statistics is relatively new and has drawn much attention by statisticians. Wavelets are also used in other fields like approximation theory, sound analysis and image compression. One of their basic properties is that they provide a sparse representation of many smooth functions, even if the degree of smoothness varies considerably over the domain of interest of if the function is only piecewise smooth. These favorable approximation properties, which are not shared by the classical

(1997) Klinke, S., Golubev, Y., Härdle, W. and Neumann, M.
Teaching Wavelets in XploRe.

Fourier basis, lead to a superior performance of estimators of functions with spatially inhomogeneous smoothness properties compared to classical linear estimators (kernel, spline).

Introductions to wavelets and applications may be found in the books by Härdle, Kerkyacharian, Picard & Tsybakov (1997) and Kaiser (1994). A first attempt for an interactive tool for wavelet smoothing was integrated in the teachware lessons of Proenca (1995) in XploRe 3 by Klinke (1997).

2 Structure of the System

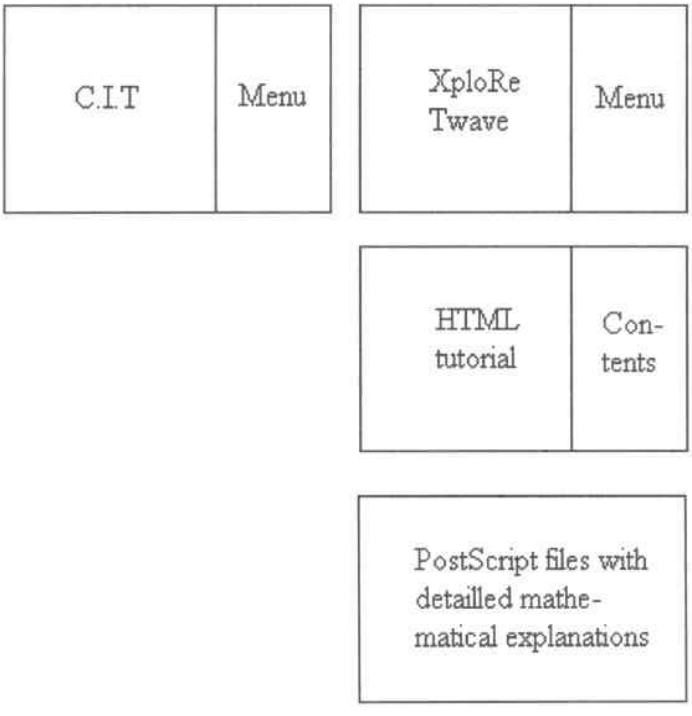


Figure 1: The Wavelet tutorial in XploRe consists of three parts: the XploRe library `twave`, the HTML tutorial and the Postscript files for detailled mathematical explanations. In contrast C.I.T just consists of the program and additional manuals.

Computational Statistics, 13, 141-151

One of the principles of teachware is the accessibility from everywhere. A student must be able to use the system in the class as well as at home. An internet link is therefore a necessity. We offer two parallel possibilties of access:

1. browsing through HTML-files in the internet or
2. javing XploRe.

By *javing* we mean the use of Java in XploRe. Both entries are found a direct internet link over the XploRe system which is availabe through

`http://wotan.wiwi.hu-berlin.de/xplore/xplore4.html`

The use of the teachware requires that the user has both processes started. We have explained above why this offers the necessary flexibility e.g. with user written datasets and procedures.

The second entry requires that the user has already installed XploRe and has started XploRe (see Figure 2). If we click with the mouse on `Help` in the menu bar then a WWW browser will be opened with a link to the tutorial.

The tutorial itself and the WWW-browser should be used in parallel such that the user can read the text and immediately execute the commands, see Figure 3. Besides loading the library `twave` the user has not to type anything and the whole dialog is now menu-driven.

A current drawback of HTML 2.0 as a basis language to code the tutorial is does not support the typesetting of formulas. However HTML 3.0 will support also formular typsetting.

But in practice the non-existence of formula typesetting has a big advantage: it forces us to describe the properties in words rather than in formulas. This is especially important if we are teaching to students which are not familiar with the mathematical notation, e.g. in economics.

2.2 The Developers View

From the developers view the whole teaching system can be decomposed into two parts:

1. Programming the single task and combining them to a system
2. Writing the HTML pages and the PostScript files

(1997) Klinke, S., Golubev, Y., Härdle, W. and Neumann, M.
Teaching Wavelets in XploRe.

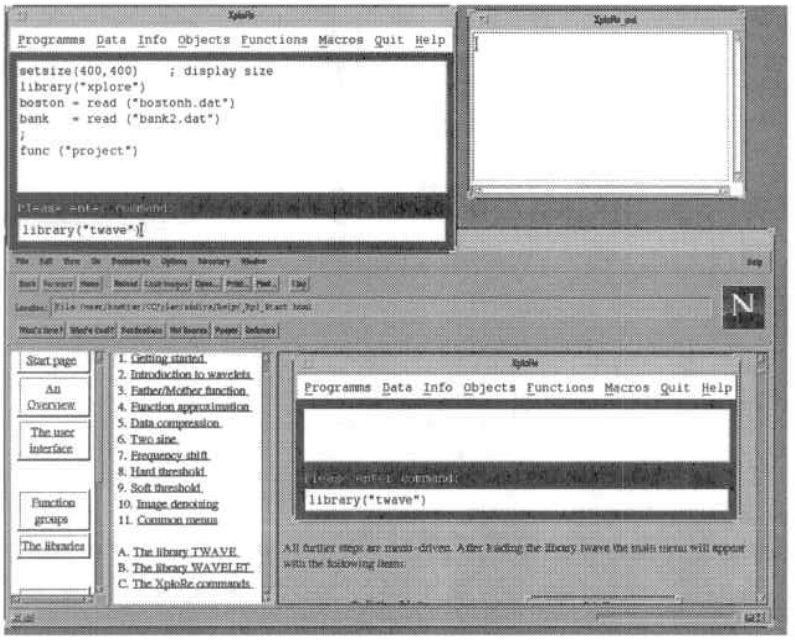


Figure 2: The screen shows in the left upper corner the XploRe command window. The window in the right upper corner is the XploRe output window. The window in the bottom of the screen is the WWW-browser (here: Netscape) with the Wavelet tutorial. The window consists of three frames: the XploRe help system frame (left), the wavelet tutorial overview (middle) and the wavelet tutorial itself (right).

The macros are based on two commands in XploRe 4:

`fwt`	the fast wavelet transform and
`invfwt`	the inverse fast wavelet transform.

The library **`wavelet`** computes the necessary constants to use the different wavelet bases. In the teaching system the Haar basis, the Daubechies-4 and the Symmlet-4 are used.

Analoguous to the generation of the wavelet bases in **`waveletmain`** when the library **`wavelet`** is called, the library **`twave`** executes during loading the macro **`twavemain`** which starts the teaching system. **`twavemain`** calls a macro named **`twlesson`** which either interactively starts a wavelet lesson (call **`twlesson(0)`**) or a specific lesson (call **`twlesson(i)`** with $i = 1, ..., 8$).

Computational Statistics, 13, 141-151

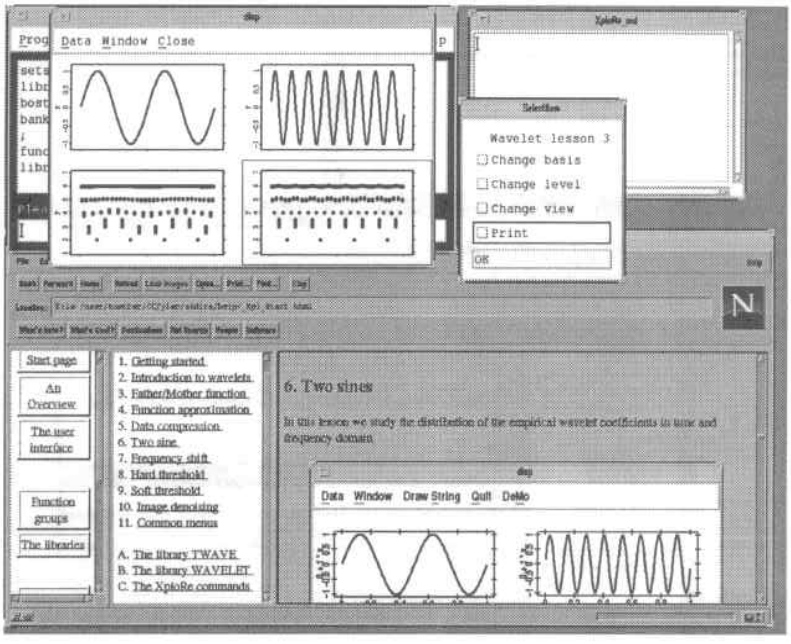


Figure 3: The screen shows in the left upper corner the XploRe command window overlayed by a display of the "two sines" lesson of **twave** library. The WWW-browser in the bottom of the screen shows the "two sines" lesson in the browser. In this lesson we look at the approximation of two sines with different frequencies by different mother wavelet coefficients. Whereas the browser shows the initial picture with the two sine functions and an approximation with the Daubechies-4 basis, we have changed already in XploRe to the Haar basis. In the right upper corner we see the XploRe output window and the actual menu for this lesson.

twlesson displays in an endless loop the basic menu which allows to select one or more lesson for execution. Then one or more of the macros **twles1**, ..., **twles8** is executed which itself displays a submenu which allows different manipulations.

In each lesson a display is shown with two windows vertically and a lesson dependent number of windows horizontally (usally two or three). The upper window always shows a plot of the problem and the lower window a representation of the mother wavelet coefficients which are appropriate for the problem.

We provide four basic views to the wavelet coefficients:

The standard view which shows the coefficients for each resolution level in a line as a vertical bar. The height of the bar is determined by the size of the coefficient.

The ordered coefficients which shows the coefficients as the absolute size of the coefficient. Not all coefficients will be shown.

The circle coefficient shows the coefficients as in the standard view, but uses circles instead of bars. The radius of the circle depends on the absolute size of the coefficient. The important fact is that the circle is drawn in red if the coefficient is used in construction the wavelet function and it is drawn in blue otherwise.

The partial sum shows the approximation of the wavelet function by adding sequentially one resolution level after another.

The views are produced by the macros `waveint1` to `waveint4`.

3 Content of the System

The teaching system is composed by 8 lessons which cover the most interesting facts about wavelets.

The tutorial itself is based on 14 topics, see Figure 2. **Front page** which allows the user to jump back to the beginning of the tutorial. **Getting started** describes how XploRe is started and how the library `twave` has to be called. **Introduction to wavelets** introduces the user of the system to wavelets in general. **Father/Mother function**, **Function approximation** **Data compression**, **Two sines**, **Frequency shift**, **Hard thresholding**, **Soft thresholding** and **Image denoising** correspond to the 8 lessons in the tutorial and will be described later. **Common menus** describes some common menus, e.g. choice of wavelet basis, choice of function, printing. **The library TWAVE**, **The library WAVELET** and **The XploRe commands** describes the macros of the two libraries and the basic XploRe commands.

In the following we describe in more detail the contents of the wavelet lessons.

Father/Mother function
In this lesson one is made familiar with the basis functions used in the wavelet analysis. These functions are basically obtained by dyadic translations and dilations of two specific functions, a so-called scaling function and a wavelet. The obtained basis functions are called father and mother functions, respectively. One can choose between functions from three different wavelet bases, the classical Haar basis, the Daubechies 4 basis and the Coiflet 2 basis. Since we use a periodic wavelet basis, the father wavelets may look a bit different

Computational Statistics, 13, 141-151

from commonly known father wavelets corresponding to a basis on the whole real line.

Function approximation
The ability of wavelet bases to provide parsimonious approximations for many smooth functions is the key to a favourable behaviour of statistical estimators based on a wavelet expansion. This lesson demonstrates how certain smoothness features of a function translate into sparsity in the space of coefficients. For example, for a piecewise constant function the coefficients corresponding to the mother wavelets are equal to zero, except for those coefficients which correspond to mother wavelets whose support contains a jump point of the function.

Data compression
This lesson describes the ability to compress certain functions into a small number of significant coefficients. For certain functions, some of them with spatially inhomogeneous smoothness properties, the ordered coefficients with respect to a wavelet basis are compared with the ordered coefficients with respect to the classical Fourier basis. It can be clearly seen that the wavelet bases have a superior ability to compress functions with inhomogeneous smoothness properties into a small number of coefficients.

Two sines
In this lesson we study the distribution of the empirical wavelet coefficients in time and frequency domain. For two sine functions with different frequencies the coefficients are displayed with respect to their spatial position and their location in resolution scale. The power of the wavelet coefficients moves to the finer resolution scales if the frequency of the function increases.

Frequency shift
The goal of this lesson is to demonstrate that the wavelet transform reflects the properties of the signal simultaneously in frequency and time domains. We consider the signal composed from two sine waves having different frequencies on the time intervals $[0, 0.5]$ and $(0.5, 1]$, respectively. It can be again seen how the power of the wavelet coefficients moves to the finer scales when the frequency of the sine becomes higher.

Hard thresholding
In statistical applications like nonparametric regression, only noisy information about the underlying function is given. Therefore, empirical versions of the wavelet coefficients are equal to the true coefficients plus some contribution by the noise. With nonparametric wavelet methods, the smoothing operation is usually performed in the domain of coefficients. Whereas a linear downweighting of the coefficients is appropriate in the case of functions functions with spatially homogeneous smoothness properties, functions with considerably inhomogeneous smoothness properties require different, essentially nonlinear regularization methods. Quite a popular method to "denoise"

(1997) Klinke, S., Golubev, Y., Härdle, W. and Neumann, M.
Teaching Wavelets in XploRe.

data is hard thresholding, that means all coefficients which are in absolute value above a certain threshold are untouched, whereas the other coefficients are set to zero. In this lesson the effect of hard thresholding can be studied for different choices of the threshold parameter.

Soft thresholding
Along with hard thresholding, soft thresholding procedures are used in many statistical applications. In this lesson we study the so-called wavelet shrinkage procedure for recovering the regression function from noisy data. The only difference between the hard and the soft thresholding procedure is the choice of the nonlinear transform on the empirical wavelet coefficients. For soft thresholding the following nonlinear transform is used:

$$S(x) = \text{sign}(x)(\text{abs}(x) - t)I(\text{abs}(x) > t),$$

where t is the threshold.

Image denoising
As an application of higher-dimensional wavelet methodology, we also included a lesson on image denoising. It was shown in Neumann & von Sachs (1995) that the commonly used isotropic d-dimensional basis does not lead to optimal estimators if the function to be estimated has different degrees of smoothness in different directions. In contrast, estimators based on a certain anisotropic basis can attain optimal rates of convergence in such anisotropic smoothness classes as well as if the "effective dimension" of the function is smaller than its nominal dimension; cf. Neumann & von Sachs (1995) and Neumann (1996). We applied nonlinear thresholding in a two-dimensional anisotropic wavelet basis to an image corrupted by some additive noise. A considerable improvement over the noisy image can be observed.

4 Future Work

Some future work consists in the integration of more interactivity in the system. For example, the visualization of the father and mother wavelet functions used, can be directed by the cursor keys instead of using a menu entry.

We need a stronger integration of the help system and tutorial with the software. We need help buttons in the software which immediately display some help text to the actual lesson, e.g. a part of the tutorial. Vice versa we need to start a specific lesson from the tutorial on mouseclick.

Kötter (1996) develops a JAVA-interface for XploRe. This interface will allow to start XploRe from a specific lesson and to perform all necessary operations. The software itself is already decomposed enough to allow such an operation.

Computational Statistics, 13, 141-151

Another extension is the inclusion of further wavelet lessons. For example, the stationary wavelet transform described in Coifman & Donoho (1995) and Nason & Silverman (1995) could be considered.
To improve the understanding of the local approximation ability of wavelets one could include an interactive manipulation of wavelet coefficients such that the student gets immediate feedback.

References

Bowman, A. & Robinson, D. (1989). C.I.T.: Introduction to Statistics, *Software*, IOP Publishing Ltd.

Bowman, A. & Robinson, D. (1990). C.I.T.: Regression & ANOVA, *Software*, IOP Publishing Ltd.

Coifman, R. & Donoho, D. (1995). Translation-invariant de-noising, *in* A. Antoniadis (ed.), *Lectures Notes in Statistics: Wavelets and Statistics*, Springer, pp. 125–150.

Härdle, W., Kerkyacharian, G., Picard, D. & Tsybakov, A. (1997). *Wavelets, approximation and statistical applications*, Springer.

Kaiser, G. (1994). *A friendly guide to wavlets*, Birkhäuser, Boston.

Klinke, S. (1997). *Data Structures in Computational Statistics*, Physica Verlag.

Koch, A. & Haag, U. (1996). The statistical software guide '95/96, *Computational Statistics & Data Analysis* **21**(2): 231–256.

Kolev, N. (1993). *Applied Statistics 1*, Stopanstvo, Sofia.

Kötter, T. (1996). *Entwicklung statistischer Software*, PhD thesis, Humboldt University of Berlin, Department of Economics, Institute of Statistics and Econometrics.

Nason, G. & Silverman, B. (1995). The stationary wavelet transform and some statistical applications, *in* A. Antoniadis (ed.), *Lectures Notes in Statistics: Wavelets and Statistics*, Springer, pp. 281–299.

Neumann, M. (1996). Multivariate wavelet thresholding: a remedy against the curse of dimensionality?, *Preprint no. 229*, Weierstrass Institute, Berlin.

Neumann, M. & von Sachs, R. (1995). Wavelet thresholding in anisotropic function classes and application to adaptive estimation of evolutionary spectra, *The Annals of Statistics* (to appear).

(1997) Klinke, S., Golubev, Y., Härdle, W. and Neumann, M.
Teaching Wavelets in XploRe.

Proenca, I. (1995). Interactive graphics for teaching simple statistics, *in* W. Härdle, S. Klinke & B. Turlach (eds), *XploRe - an interactive statistical computing environment*, Springer, pp. 113–140.

Journal of the American Statistical Association, 92, 1512-1521

An Analysis of Transformations for Additive Nonparametric Regression

Oliver B. LINTON, Rong CHEN, Naiysin WANG, and Wolfgang HÄRDLE

We consider a nonparametric regression model with a parametric family of dependent variable transformations, one of which induces additive covariate effects. We estimate the additive regression effects using the integration method and estimate the transformation parameter from a profiled instrumental variable and pseudolikelihood criterion. The asymptotic distributions of the parameter and regression estimates are given. The practical performance is investigated via an application.

KEY WORDS: Box–Cox transformation; Dimensionality reduction; Kernels.

1. INTRODUCTION

Taking transformations of the data has been an integral part of statistical practice for many years. Transformations have been used to aid interpretability as well as to improve statistical performance. An important contribution to this methodology was made by Box and Cox (1964), who proposed a parametric power family of transformations that nested the logarithm and the level. They suggested that the power transformation, when applied to the dependent variable in a linear regression setting, might induce normality, error variance homogeneity, and additivity of effects. They proposed estimation methods for the regression and transformation parameters. Carroll and Ruppert (1984) suggested applying this and other transformations to both dependent and independent variables. A number of other dependent variable transformations have been suggested—for example, the Zellner–Revankar transform (see Zellner and Revankar 1969). The transformation methodology has been quite successful, and a large literature now exists on this subject for parametric models (see Carroll and Ruppert 1988). There are also a number of applications to economics data (see Ehrlich 1977; Heckman and Polachek 1974; Hulten and Wykoff 1981; Zarembka 1968; Zellner and Revankar 1969).

We work with transformations inside a regression setting. For many data, the linearity of covariate effect after transformation may be too strong. For example, a respected study of the effects of schooling and experience on earnings (Heckman and Polachek 1974, p. 350) found that although their data supported the logarithmic transformation of their dependent variable earnings, it was "somewhat less clear on the functional form for the independent variables." We thus consider a more general specification, allowing for nonparametric covariate effects. Let (X, Y) be a random variable with $X \in \mathbb{R}^d$ and $Y \in \mathbb{R}$, and let $\{(X_i, Y_i)\}_{i=1}^n$ be an iid sample from this population. Consider the estimation of the regression function $m(x) = E\ (Y|X = x)$. Ibragimov and Hasminskii (1980) and Stone (1980, 1982) showed that the optimal rate for estimating m is $n^{-l/(2l+d)}$, with l a measure of the smoothness of m. This rate of convergence can be very slow for large dimensions d. One way of achieving better rates of convergence is through additive modeling. An additive structure for m is a regression function of the form $m(x) = c + \sum_{\alpha=1}^d m_\alpha(x_\alpha)$, where $x = (x_1, \ldots, x_d)'$ are the d-dimensional predictor variables and m_α are one-dimensional nonparametric functions with $E\{m_\alpha(X_\alpha)\} = 0$. Stone (1986) showed that for such regression curves, the optimal rate for estimating m is the one-dimensional rate of convergence $n^{-l/(2l+1)}$. Thus one speaks of dimensionality reduction through additive modeling.

We examine a semiparametric model that combines a parametric transformation with the flexibility of an additive nonparametric regression function. For a parametric family of transforms $\theta_\lambda(\cdot), \lambda \in \Lambda \subset \mathbb{R}$, define the regression functions $m_\lambda(x) = E\ \{\theta_\lambda(Y)|X_1 = x_1, \ldots, X_d = x_d\}$, and suppose that for some unique $\lambda_0 \in \Lambda$,

$$m_{\lambda_0}(x) = c + \sum_{\alpha=1}^d g_\alpha(x_\alpha), \tag{1}$$

where $g_\alpha(\cdot)$ are of unknown form with $E\{g_\alpha(X_\alpha)\} = 0, \alpha = 1, \ldots, d$. This model was previously addressed by Hastie and Tibshirani (1990, p. 187). They suggested estimation procedures based on the iterative backfitting method; however, they did not provide many results about the statistical properties of their procedures. Breiman and Friedman (1985) suggested a generalization of (1), called alternating conditional expectation (ACE), in which the transformation is not restricted to be parametric. Again, plausible estimation procedures are available, but little is known about their statistical properties. Finally, Tibshirani (1988) suggested a modification of the ACE estimation algorithm, called additivity and variance stabilization (AVAS). This has several advantages over the ACE algorithm: in particular, it reproduces model transformations and is equivariant under monotone transformations.

Oliver Linton is Associate Professor of Economics, Cowles Foundation for Research in Economics, Yale University, New Haven, CT 06520. Rong Chen is Associate Professor of Statistics and Naiysin Wang is Assistant Professor of Statistics, Department of Statistics, Texas A&M University, College Station, TX 77843. Wolfgang Härdle is Professor at the Institut für Statistik und Ökonometrie, Humboldt-Universität zu Berlin, Germany. The authors would like to thank Jens Perch Nielsen and Stefan Sperlich for helpful comments, and Nick Hengartner for sharpening their view on Section 2. They thank the National Science Foundation, the North Atlantic Treaty Organization, and the Sonderforschungsbereich 373 for financial support. GAUSS software is available from the first author on request.

Journal of the American Statistical Association
December 1997, Vol. 92, No. 440, Theory and Methods

Journal of the American Statistical Association, 92, 1512-1521

Independently, Linton and Nielsen (1995), Newey (1994), and Tjøstheim and Auestad (1991, 1994), introduced a new method for estimating additive nonparametric models (when the transformation is known) that is based on direct integration of an initial pilot smoother of the regression function. It has been possible to prove a central limit theorem for this estimator because of its mathematical tractability. We extend their theory to the semiparametric model (1). We use the semiparametric profile method described by Bickel, Klaassen, Ritov, and Wellner (1993) to estimate the parameters $\lambda_0 \in \Lambda \subset \mathbb{R}$ in (1). This uses a preliminary integration estimate of the additive components of m_λ. We also show how to estimate the constant c, the univariate component functions $g_\alpha(\cdot)$, and the nonparametric regression function $m_\lambda(\cdot)$. We derive the asymptotic distributions of our estimators under standard regularity conditions. The estimator of λ_0 is root-n consistent, whereas the estimates of $g_\alpha(x_\alpha)$ and $m_\lambda(x)$ are consistent at the one-dimensional rate of $n^{2/5}$.

The article is organized as follows. In Section 2 we give an outline of the integration method and explain how it can circumvent the curse of dimensionality. In Section 3 we define the estimation procedures for both parametric and nonparametric parts. In Section 4 we give the asymptotic properties of the procedures. In Section 5 we illustrate our methods in an application and on simulated data. The Appendixes contain the proofs of all results.

2. THE INTEGRATION METHOD

Suppose that we have some pilot estimator $\hat{f}(x_1, x_2)$ of a function $f(x_1, x_2)$ of the scalar x_1 and the $d-1$ variables x_2. Define the integration estimator

$$\tilde{\varphi}_1(x_1) = \int \hat{f}(x_1, x_2)\, dQ(x_2) \tag{2}$$

for any $d-1$-dimensional signed measure Q. Usually, Q will be a probability measure. If the estimated function were additive—that is, $f(x_1, x_2) = f_1(x_1) + f_2(x_2)$—then $\tilde{\varphi}_1(\cdot)$ would consistently estimate $f_1(\cdot)$ plus a constant that is the same for all x_1. Likewise, if the function were multiplicative—that is, $f(x_1, x_2) = f_1(x_1) f_2(x_2)$—then $\tilde{\varphi}_1(\cdot)$ would consistently estimate $f_1(\cdot)$ times a constant that is the same for all x_1. This is the basic idea behind the integration method. Of course, it is also important to know whether one can also obtain the one-dimensional rate of convergence for $\tilde{\varphi}_1(\cdot)$. Intuitively, this will hold because integration is averaging and so reduces variance; a proof of this was first given by Newey (1994). However, the dependency of $\tilde{\varphi}_1(x_1)$ on the high-dimensional pilot smoother $\hat{f}(x_1, x_2)$ suggests that the $\tilde{\varphi}_1(x_1)$ may suffer somewhat from the curse of dimensionality. This viewpoint is supported by the fact that in the currently available central limit theorem proofs (see Chen, Härdle, Linton, and Severance-Lossin 1996; Linton and Härdle 1996; Newey 1994), it has been necessary to use bias reduction when the dimensions were greater than four. As usual, the conditions given there were sufficient but not necessary; in our opinion, they can be weakened considerably. To support this viewpoint, we give a different interpretation of the integration estimator.

To focus ideas, we look at density estimation. Let

$$\hat{f}(x_1, x_2) = \frac{1}{n h_1 h_2^{d-1}} \sum_{i=1}^{n} K_1\left(\frac{x_1 - X_{1i}}{h_1}\right) K_2\left(\frac{x_2 - X_{2i}}{h_2}\right) \tag{3}$$

be a product kernel density estimate based on an iid sample $\{X_i\}_{i=1}^n$ partitioned as earlier. When the dimensions d are large, $\hat{f}(x_1, x_2)$ has very poor performance. Now suppose that we integrate $\hat{f}(x_1, x_2)$ with respect to Lebesgue measure, that is, take $dQ(x_2) = dx_2$ in (2), and let

$$\begin{aligned}\tilde{\varphi}_1(x_1) &= \int \hat{f}(x_1, x_2)\, dx_2 \\ &= \frac{1}{n h_1 h_2^{d-1}} \sum_{i=1}^{n} K_1\left(\frac{x_1 - X_{1i}}{h_1}\right) \\ &\quad \times \int K_2\left(\frac{x_2 - X_{2i}}{h_2}\right) dx_2. \end{aligned} \tag{4}$$

By the usual change of variables, we obtain exactly

$$\tilde{\varphi}_1(x_1) = \frac{1}{n h_1} \sum_{i=1}^{n} K_1\left(\frac{x_1 - X_{1i}}{h_1}\right), \tag{5}$$

which is the one-dimensional kernel density estimate of $f_1(x_1) = \int f(x_1, x_2)\, dx_2$. Thus integration has taken a d-dimensional smoother into a one-dimensional smooth and completely eliminated the second bandwidth h_2.

In regression the pilot estimators typically used are more complicated than (3), and we cannot go from step (4) to (5) exactly; some approximation argument is necessary. However, we would like to suggest that even in this case the integration estimator can be interpreted as a one-dimensional smooth. Recent work by Hengartner (1996) confirms this view.

3. ESTIMATION

To estimate the quantities of interest, we use a multistage procedure. In the first part we estimate the transformation using the semiparametric profile method, discussed by Bickel et al. (1993). This requires an estimate of the regression functions for each parameter value. In the second part we estimate the additive regression function using the estimated transformation.

1. For each λ, m_λ is estimated by a multidimensional kernel smoother $\hat{m}_\lambda$.

2. For each direction α, the pilot estimate is then integrated to obtain an estimate of the individual effect of X_α on Y in the λ scale. The individual effect estimates are combined to form an "additive reconstruction" $\tilde{m}_\lambda$. A detailed description of the integration method is given in Section 3.1.

(1997) Linton, O., Chen, R., Wang, N. and Härdle, W.
An analysis of transformations for additive nonparametric regression.

Journal of the American Statistical Association, 92, 1512-1521

3. We define a criterion function for λ depending on the profiled estimate $\tilde{m}_\lambda$. We choose $\hat{\lambda}$ to optimize this criterion.

4. Finally, we construct a new kernel smoother $\hat{m}_{\hat{\lambda}}$ using the estimated $\hat{\lambda}$, and obtain the final additive estimate $\tilde{m}_{\hat{\lambda}}$ by integration.

Note that in Steps 1–3, we are concerned only with the properties of $\hat{\lambda}$.

3.1 Nonparametric Estimation

Step 1 involves estimating a regression function for the sample $\{X_i, \theta_\lambda(Y_i)\}_{i=1}^n$ for any λ. We estimate $m_\lambda(x)$ using the multidimensional Nadaraya–Watson estimator,

$$\hat{m}_\lambda(x) = \frac{\sum_{j=1}^n \prod_{\alpha=1}^d K_\alpha\left(\frac{x_\alpha - X_{\alpha j}}{h_\alpha}\right)\theta_\lambda(Y_j)}{\sum_{l=1}^n \prod_{\alpha=1}^d K_\alpha\left(\frac{x_\alpha - X_{\alpha l}}{h_\alpha}\right)}, \tag{6}$$

where $K_1, \ldots, K_d$ are univariate kernels and $h_1(n), \ldots, h_d(n)$ are bandwidth sequences, one for each direction. When our ultimate goal is to estimate λ, we take common bandwidth and kernel for notational simplicity; otherwise, we allow both kernel and bandwidth to vary with direction. For each $\alpha = 1, \ldots, d$, partition $x = (x_\alpha, x_{\underline{\alpha}})$, where x_α is a one-dimensional direction of interest and $x_{\underline{\alpha}}$ is a $d-1$-dimensional nuisance direction; do likewise with $X = (X_\alpha, X_{\underline{\alpha}})$ and $X_i = (X_{\alpha i}, X_{\underline{\alpha} i})$. Now define

$$\hat{\gamma}_\alpha(x_\alpha; \lambda) = n^{-1}\sum_{i=1}^n \hat{m}_\lambda(x_\alpha, X_{\underline{\alpha} i}). \tag{7}$$

This is the empirical integration estimator of Chen et al. (1996). Let p be the joint density of $X_1, \ldots, X_d$, let p_α be the marginal density of X_α, and let $p_{\underline{\alpha}}$ be the joint density of $X_{\underline{\alpha}}$. Then $\hat{\gamma}_\alpha(x_\alpha; \lambda)$ consistently estimates the population quantity

$$\gamma_\alpha(x_\alpha; \lambda) = \int m_\lambda(x_\alpha, x_{\underline{\alpha}}) p_{\underline{\alpha}}(x_{\underline{\alpha}})\, dx_{\underline{\alpha}}.$$

In fact, under the additive model (1), $\gamma_\alpha(x_\alpha; \lambda_0) = c + g_\alpha(x_\alpha)$. More generally, one could interpret $\gamma_\alpha(x_\alpha; \lambda)$ as one aspect of the univariate effect of the covariate on the transformed dependent variable. This is the point of view expressed by Newey (1994) in connection with econometric applications. Now, let

$$\tilde{m}_\lambda(x) = \sum_{\alpha=1}^d \hat{\gamma}_\alpha(x_\alpha; \lambda) - (d-1)\hat{c}_\lambda, \tag{8}$$

where $\hat{c}_\lambda = n^{-1}\sum_{i=1}^n \theta_\lambda(Y_i)$. Then $\tilde{\mathbf{m}}_\lambda(x)$ consistently estimates

$$\bar{m}_\lambda(x) = \sum_{\alpha=1}^d \gamma_\alpha(x_\alpha; \lambda) - (d-1)c_\lambda,$$

where $c_\lambda = E\{\theta_\lambda(Y)\}$. Note that $\bar{m}_{\lambda_0}(x) = m_{\lambda_0}(x)$; that is, combining the covariate effects in the λ_0 scale gives us the regression function. However, $\bar{m}_\lambda(x) \neq m_\lambda(x)$ for $\lambda \neq \lambda_0$. This information is used to identify λ_0. We implement (6), and hence (7) and (8), in several different ways according to our purpose.

3.2 Estimation of λ

To estimate λ, we use a nonlinear instrumental variable procedure. The assumed properties of the mean are used to generate first-order conditions that identify the parameter λ. This method has been widely used in estimating simultaneous equation systems and in certain dynamic models, and was used by Amemiya and Powell (1981) in the parametric Box–Cox model. [See Newey and McFadden (1994 p. 2116) for background references and discussion. See also Angrist, Imbens, and Rubin (1996) for a connection between instrumental variables and the Rubin causal model used in causal inference.]

Let $\mathbf{Z}_i, i = 1, \ldots, n$, be a J by 1 vector of iid instruments with the property that

$$E[\{\theta_{\lambda_0}(Y_i) - \bar{m}_{\lambda_0}(X_i)\}|\mathbf{Z}_i] = 0 \quad \text{a.s.}$$

for a unique λ_0. The instruments must satisfy an additional identification condition, which we discuss in Section 4. Examples of valid instruments include functions of the X's themselves. Let $\mathbf{Z} = (\mathbf{Z}_1, \ldots, \mathbf{Z}_n)', \boldsymbol{\Theta}_\lambda = (\theta_\lambda(Y_1), \ldots, \theta_\lambda(Y_n))'$, and $\tilde{\mathbf{M}}_\lambda = (\tilde{m}_\lambda(X_1), \ldots, \tilde{m}_\lambda(X_n))'$, and define $\hat{\lambda}$ to be any minimizer of

$$Q_n(\lambda) = \{(\boldsymbol{\Theta}_\lambda - \tilde{\mathbf{M}}_\lambda)'\mathbf{Z}/n\}\mathbf{W}_n\{\mathbf{Z}'(\boldsymbol{\Theta}_\lambda - \tilde{\mathbf{M}}_\lambda)/n\}, \tag{9}$$

where $\mathbf{W}_n$ is a sequence of J by J weighting matrices satisfying $\mathbf{W}_n \xrightarrow{p} \mathbf{W}$ a positive definite matrix. A simple feasible choice for $\mathbf{W}_n$ is the identity matrix. We discuss how more efficiency can be obtained by wiser choice of $\mathbf{W}_n$. In our application we computed $\hat{\lambda}$ by grid search, but iterative techniques such as Newton–Raphson also work well here.

An alternative method for estimating λ is the full Gaussian pseudolikelihood (see Box and Cox 1964). After profiling out a variance parameter, this amounts to choosing $\tilde{\lambda}$ that minimizes

$$L_n(\lambda) = n^{-1}\sum_{i=1}^n \ln J_\lambda(Y_i) - \ln\{n^{-1}(\boldsymbol{\Theta}_\lambda - \tilde{\mathbf{M}}_\lambda)'(\boldsymbol{\Theta}_\lambda - \tilde{\mathbf{M}}_\lambda)\},$$

where $J_\lambda(\cdot)$ is the Jacobian of the transformation $y \to \theta_\lambda(y)$. (See Carroll and Ruppert 1988, pp. 124–127, for more discussion of this method in the parametric regression problem. A robust version of this procedure was given in Carroll 1980.)

An alternative way to proceed is to profile both c and λ out by using $\tilde{m}_{\lambda,c}(x) = \sum_{\alpha=1}^d \hat{\gamma}_\alpha(x_\alpha; \lambda) - (d-1)c$ in place of $\tilde{m}_{\lambda,c}(x)$, and to construct profiled criteria $Q_n(\lambda, c)$ or $L_n(\lambda, c)$, which are then optimized with respect to both c and λ. We restrict attention to the simpler procedure based on profiling only λ. In this case, given estimates of λ, we define estimators of $c, g_\alpha(\cdot)$, and $m_{\lambda_0}(\cdot)$ as in Section 3.1.

Journal of the American Statistical Association, 92, 1512-1521

3.3 Discussion

Among the many examples of interest, the following ones are used most commonly:

Box–Cox: $\theta_\lambda(y) = (y^\lambda - 1)/\lambda; J_\lambda(y) = y^{\lambda-1}$
Zellner–Revankar: $\theta_\lambda(y) = \ln y + \lambda y^2; J_\lambda(y) = y^{-1} + 2\lambda y$
arcsinh: $\theta_\lambda(y) = \sinh^{-1}(\lambda y)/\lambda; J_\lambda(y) = (1 + \lambda y^2)^{-1/2}$.

The arcsinh transform was discussed by Johnson (1949) and more recently by Robinson (1991). The main advantage of the arcsinh transform is that it works for y taking any value, whereas the Box–Cox and the Zellner–Revankar transforms are defined only if y is positive. For these transformations, the error term cannot be normally distributed except for a few isolated parameters, and so the Gaussian likelihood is misspecified. In fact, as Amemiya and Powell (1981) pointed out, the resulting estimators (in the parametric case) are inconsistent only when $n \to \infty$. This is the main advantage of the instrumental variable criterion; it is consistent even for these transformations under general sampling conditions. However, we note that Bickel and Doksum (1981) established consistency of the Gaussian pseudolikelihood procedure when both $n \to \infty$ and $\sigma \to 0$, where σ is the scale of the error term.

We subsequently focus on the instrumental variable procedure in our treatment of the asymptotics.

4. ASYMPTOTIC PROPERTIES

We develop asymptotic approximations for the instrumental variable estimator $\hat{\lambda}$, as $n \to \infty$. Under similar conditions, $\tilde{\lambda}$ is also consistent for some transformations (although not the Box–Cox as discussed earlier). Finally, we establish the asymptotic distribution of the covariate effects. We give three theorems. As far as the properties of $\hat{\lambda}$ are concerned, we use a common kernel K and bandwidth h in (6). For consistency of $\hat{\lambda}$, second-order kernels suffice, whereas to achieve root-n consistency we must use bias reduction in all directions when the dimensions exceed 3. This type of condition is common in other semiparametric situations, see for example Robinson (1988). In the last theorem, which is about the properties of $\hat{\gamma}_\alpha(\cdot)$, we allow for bias reduction in directions other than α when the dimensions exceed four (as in Linton and Härdle 1996).

4.1 Consistency of $\hat{\lambda}$

Our proof uses primarily that $\tilde{m}_\lambda(x)$ is uniformly consistent as an estimator of $\bar{m}_\lambda(x)$; no rates are needed. For this reason, we take the standard kernel estimator (6) with common bandwidth h and kernel K for each direction. We work with iid instruments $\mathbf{Z}_i$ that are mean 0 (this is without loss of generality and can be achieved by subtracting off sample means) and make some additional technical assumptions, stated in Appendix A.

Theorem 1. Suppose that conditions A1–A6 given in Appendix A hold. Suppose further that the following condition holds:

A7: For all $\lambda \neq \lambda_0$ in a neighborhood $\mathcal{N}(\lambda_0)$ of λ_0, $E[\{\bar{m}_\lambda(X_i) - m_\lambda(X_i)\}|\mathbf{Z}_i] \neq 0$. Then $\hat{\lambda}$ is locally consistent.

Remark. The identification condition A7 is difficult to verify from primitive conditions, as is true in related situations (see Newey and McFadden 1994, p. 2127). Nevertheless, we expect that condition A7 is obtained through additivity. For $\lambda \neq \lambda_0$, the regression function m_λ is not additive, and so the difference between the additive reconstruction $\tilde{m}_\lambda$ and m_λ should, in the limit, be correlated with the instruments.

4.2 Asymptotic Normality of $\hat{\lambda}$

We use a common kernel K and bandwidth $h(n)$ for each direction α. We need some additional conditions given in Appendix A. Specifically, to obtain $n^{1/2}$ consistency for $\hat{\lambda}$, it is necessary to use bias-reducing kernels of order $q \geq 2$ to estimate the regression function when the dimensions $d > 4$. Define the mean 0 iid random variables $\varepsilon_\lambda(i) = \theta_\lambda(Y_i) - m_\lambda(X_i), i = 1, \ldots, n$, and also let $\tilde{\varepsilon}_\lambda(i) = \theta_\lambda(Y_i) - \tilde{m}_\lambda(X_i)$ for any λ. Also, let $r_j(\lambda) = E[\partial^j/\partial\lambda^j\{m_\lambda(X_i) - \tilde{m}_\lambda(X_i)\}\mathbf{Z}_i], j = 0, 1, 2$, and set $r_{j0} = r_j(\lambda_0)$. Finally, let

$$\Omega = E[\varepsilon^2_{\lambda_0}(i)\{\mathbf{Z}_i - \mathbf{Z}_i^*\}\{\mathbf{Z}_i - \mathbf{Z}_i^*\}'],$$

where

$$\mathbf{Z}_i^* = \sum_{\alpha=1}^{d} E(\mathbf{Z}_i|X_{\alpha i})\,\frac{p_\alpha(X_{\alpha i})p_{\underline{\alpha}}(X_{\underline{\alpha} i})}{p(X_i)}.$$

Theorem 2. Suppose that conditions A1–A7 and B1–B6 given in the Appendix hold. Then

$$n^{1/2}(\hat{\lambda} - \lambda_0) \Rightarrow \mathrm{N}\{0, (r_{10}'\mathbf{W}r_{10})^{-2}r_{10}'\mathbf{W}\Omega\mathbf{W}r_{10}\}. \quad (10)$$

We can consistently estimate r_{10} by $\hat{r}_1 = n^{-1}\sum_{i=1}^n \mathbf{Z}_i(\partial/\partial\lambda)\{\hat{m}_{\hat{\lambda}}(X_i) - \tilde{m}_{\hat{\lambda}}(X_i)\}$ and Ω by $\hat{\Omega} = n^{-1}\sum_{i=1}^n \mathbf{Z}_i\mathbf{Z}_i'\tilde{\varepsilon}^2_{\hat{\lambda}}(i)$, and thereby consistently estimate the asymptotic variance V by

$$\hat{V} = (\hat{r}_1'\mathbf{W}_n\hat{r}_1)^{-2}\hat{r}_1'\mathbf{W}_n\hat{\Omega}^{-1}\mathbf{W}_n\hat{r}_1.$$

Carroll and Wand (1991) and Wang and Ruppert (1996) obtained similar results for their semiparametric estimators.

By straightforward calculations one can show that the optimal weighting matrix is $\mathbf{W}_n = \Omega^{-1}$ or any consistent estimate thereof (see Newey and McFadden 1994), in which case the asymptotic variance in (10) becomes $V_{\text{opt}} = (r_{10}'\Omega^{-1}r_{10})^{-1}$. If we take $\mathbf{W}_n = \hat{\Omega}^{-1}$, then the efficient variance is achieved.

Consider the infeasible procedure $\bar{\lambda}$ that replaces $\tilde{m}_\lambda$ by $\bar{m}_\lambda$ in (6). This has asymptotic variance the same as (10) with Ω replaced by $\Omega_0 = E[\varepsilon^2_{\lambda_0}(i)\mathbf{Z}_i\mathbf{Z}_i']$. Now work with the special case that $\varepsilon^2_{\lambda_0}(i)$ is independent of $\mathbf{Z}_i, E(\mathbf{Z}_i|X_i)$ is additive and the components of X are mutually independent. In this case $\mathbf{Z}_i^* = E(\mathbf{Z}_i|X_i)$, so that $\Omega \leq \Omega_0$; that is, the asymptotic variance of $\hat{\lambda}$ can be smaller than that for $\bar{\lambda}$. Although this appears anomalous, it has been found by

Journal of the American Statistical Association, 92, 1512-1521

other authors (see Gutierrez and Carroll 1995; Robins, Rotnizky, and Zhao 1994). It is mainly due to the fact that the instrumental variable method may not reach the semiparametric efficiency bound as defined by Newey (1990). Even though $\hat{\lambda}$ is not efficient, it does not impose either normality of the errors or homoscedasticity in the conditional variance, and it is easy to implement. In this sense it is robust and practical. One can impose additional restrictions, such as constant variance, within this framework by adding an additional moment condition to (9). This results in more efficient estimates of λ when the restrictions are true.

4.3 Asymptotic Normality of Regression Function Estimates

We now consider the final stage in which the root-n consistent estimate $\hat{\lambda}$ is used to estimate the one-dimensional functions. We use a procedure that has bandwidth $h = \beta n^{-1/5}$ and second-order kernel K for the direction of interest and $(d-1)$-dimensional kernel L of order q with common bandwidth g for the remaining directions (as in Chen et al. 1996 for additive nonparametric regression). As pointed out there, if $d \leq 4$, then the theory is consistent with using second-order kernels. When $d > 4$, one must use higher-order kernels to satisfy the conditions of the theorem.

Let $\|K\|_2^2 = \int K^2(u)\,du$ and $\mu_2(K) = \int u^2 K(u)\,du$.

Theorem 3. Suppose that the condition C1 given in the Appendix holds. Then

$$n^{2/5}\{\hat{\gamma}_\alpha(x_\alpha;\hat{\lambda}) - \gamma_\alpha(x_\alpha;\lambda_0)\} \Rightarrow \mathrm{N}\{b_\alpha(x_\alpha), v_\alpha(x_\alpha)\},$$

where

$$b_\alpha(x_\alpha) = \beta^2\mu_2(K)\left\{\frac{1}{2}\,g''_\alpha(x_\alpha) + g'_\alpha(x_\alpha) \times \int \frac{\partial \ln p}{\partial x_\alpha}\,(x)p_{\underline{\alpha}}(x_{\underline{\alpha}})\,dx_{\underline{\alpha}}\right\}$$

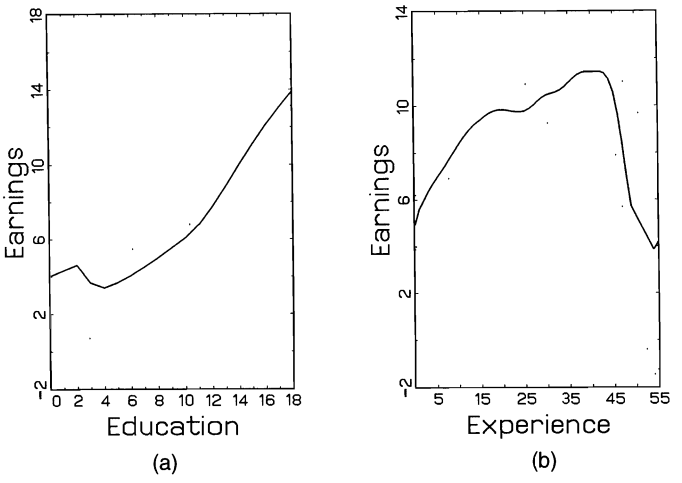


Figure 2. The Estimated Covariate Effects (a) Education and (b) Earnings on Earnings (λ = 1) Shown With Pointwise Symmetric 95% Confidence Intervals.

and

$$v_\alpha(x_\alpha) = \beta^{-1}\|K\|_2^2 \int \sigma^2(x)\,\frac{p^2_{\underline{\alpha}}(x_{\underline{\alpha}})}{p(x)}\,dx_{\underline{\alpha}},$$

where $\sigma^2(x) = \mathrm{var}\{\theta_{\lambda_0}(Y)|X = x\}$. Finally, $v_\alpha(x_\alpha)$ can be consistently estimated by $\sum_{j=1}^n w_j^2\tilde{\varepsilon}^2_{\hat{\lambda}}(j)$, where $\{w_j\}_{j=1}^n$ are the weights in $\hat{\gamma}_\alpha(x_\alpha;\hat{\lambda}) = \sum_{j=1}^n w_j\theta_{\hat{\lambda}}(y_j)$. Similar results can be obtained for local linear regression smoothers (Fan 1992), with the bias function given by the simpler form $b_\alpha(x_\alpha) = \beta^2\mu_2(K)g''_\alpha(x_\alpha)/2$.

The estimated regression function has the one-dimensional convergence rate and is unaffected by the estimation of the transformation parameter. Bandwidth selection and order selection can be addressed exactly as for the regression problem. We do not advocate using higher-order kernels in practice, even when the dimensions are high, because of their well-known poor small-sample performance and because we think that with better proof technology, one can prove Theorem 3 without bias reduction. (See Fan, Härdle,

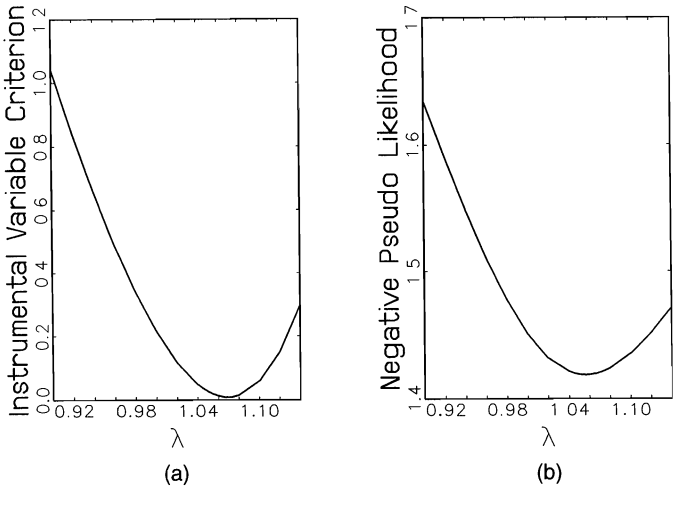


Figure 1. The Criterion Functions (a) Instrumental Variables ($\times 10^{-4}$) and (b) Negative Pseudolikelihood Function ($\times 10^{-3}$) Against λ for a Grid of Values Between [.90, 1.14] for the Earnings/Education Data.

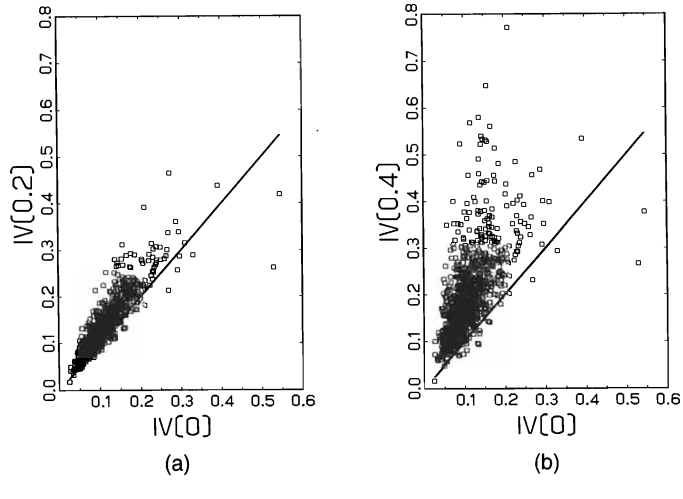


Figure 3. The Relative Magnitudes of the Criterion Functions for the Simulated Data at Different Parameter Values (a) λ = .2 Versus λ = 0 and (b) λ = .4 Versus λ = 0. The 45-degree line is shown for comparison. Sample size n = 100.

(1997) Linton, O., Chen, R., Wang, N. and Härdle, W.
An analysis of transformations for additive nonparametric regression.

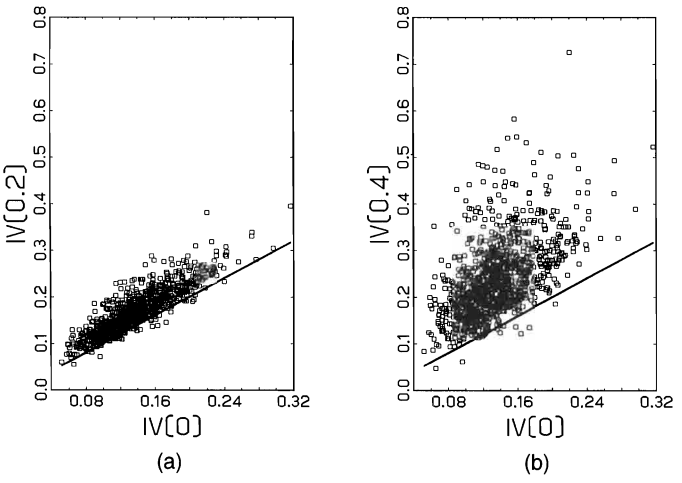


Figure 4. The Relative Magnitudes of the Criterion Functions for the Simulated Data at Different Parameter Values (a) $\lambda = .2$ Versus $\lambda = 0$ and (b) $\lambda = .4$ Versus $\lambda = 0$. The 45-degree line is shown for comparison. Sample size $n = 250$.

and Mammen 1996 and Hengartner 1996 for further discussion.)

5. NUMERICAL RESULTS

5.1 Application

We examine again the dataset used by Linton and Nielsen (1995) obtained from a random sample of 534 individuals from the 1985 Current Population Survey conducted by the U.S. Department of Commerce. (Details of this dataset can be found in Berndt 1991, chap. 5.) We examine the relationship between wages (y) and the covariates education in years (x_1) and experience in years (x_2). Linton and Nielsen (1995) used a logarithmic transformation of the dependent variable, as suggested by much other work. We use the Box–Cox transformation on the dependent variable that nests both logarithm and level. We implemented the additive regression procedure using a Gaussian kernel and rule of thumb bandwidth selection, following Linton and Neilson (1995). Figure 1 shows the instrumental variable (IV) criterion and the pseudolikelihood (LIKE) criterion computed at a grid of λ values. For instruments we took 1, x_1, x_1^2, x_2, and x_2^2, and we used $\mathbf{W}_n = I$ as weighting. The IV criterion was maximized at $\lambda = 1.069$, and LIKE was minimized at $\lambda = 1.056$. (The optima were found by grid search to a precision of $\pm.0005$.) The standard error was .0749 for the IV estimator. Thus the optimal transformation here is not far from linear; that is, the effects of education and experience on earnings itself appear to be additive.

Figure 2 plots the fitted additive regression for the untransformed y; that is, $\lambda = 1$. The effect of education on earnings is mildly convex, whereas that of experience is somewhat concave, with rapidly increasing returns to the first 10 years of experience followed by a slow but steady increase through 40 years, followed by a decline in later years. This is consistent with other studies, although some have found a similar dip in the returns to education, (see Mukarjee and Stern 1994).

5.2 Simulations

We first investigate by simulation how well the IV criterion function does at discriminating between rival models for small samples. This bears on how well $\hat{\lambda}$ behaves in small samples.

We generated 1,000 samples of sizes 100 and 250 from

$$\ln Y = X_1 + \left(X_2^2 - \frac{1}{12}\right) + \varepsilon,$$

where $\varepsilon \sim$ N(0, .1), while X_1 and X_2 were mutually independent uniforms on $[-.5, .5]$. We work again with the Box–Cox model for which the foregoing data-generation process is equivalent to $\lambda_0 = 0$. Figure 3a plots $Q_n(.2)$ against $Q_n(0)$, and Figure 3b shows $Q_n(.4)$ against $Q_n(0)$, both for the smaller sample size of 100. We used the same instruments as in the application. The median values were .101, .129, and .182. In 927 cases the criterion preferred $\lambda = 0$ to $\lambda = .2$, and in 973 cases it preferred $\lambda = 0$ to $\lambda = .4$. Figure 4 shows the same plots as in Figure 3, for the larger sample size of 250. In this case the criterion preferred $\lambda = 0$ to $\lambda = .2$ in 961 cases, and preferred $\lambda = 0$ to $\lambda = .4$ in 989 cases. Thus as sample size increases, the evidence in favor of the true value mounts.

Secondly, we check whether the integration method breaks down when applied to high-dimensional data. We generated data from the following model:

$$Y = \sum_{\alpha=1}^{d} X_\alpha^3 + \varepsilon,$$

where X_α are independent uniforms on $[-.5, .5]$ and $\varepsilon \sim$ N(0, .1). We examine our integration estimate $\tilde{m}_1(0)$ of $m_1(0)$, assuming that $\lambda = 1$ is known, for the cases $d = 1, 2, \ldots, 10$. Note that the estimate of $m_1(0)$ for the case $d = 1$ is the one-dimensional smooth of Y on X_1; its asymptotic variance cannot be bettered when $d \geq 2$, even by the backfitting method. We used a uniform kernel and a single bandwidth of size precisely $n^{-1/5}$. Our results are given in Table 1.

There is a deterioration in performance as dimension increases but less so at the larger sample size. Even in the case where $n = 100$, there is only about a 50% increase in the root mean squared error for $d = 10$.

6. FINAL REMARKS AND CONCLUSIONS

There was some controversy about how to conduct infer-

Table 1. Root Mean Squared Error of $\tilde{m}_1(0)$ Based on 500 Replications

	$d = 1$	$d = 2$	$d = 3$	$d = 4$	$d = 5$	$d = 6$	$d = 7$	$d = 8$	$d = 9$	$d = 10$
$n = 100$	.0121	.0122	.0130	.0138	.0150	.0158	.0167	.0172	.0181	.0188
$n = 200$	.0098	.0101	.0104	.0103	.0107	.0115	.0120	.0126	.0129	.0133

Journal of the American Statistical Association, 92, 1512-1521

ence in fully parametric transformation models; in particular, whether one should take into account the uncertainty due to not knowing the transformation when evaluating estimates of the regression coefficients (see Bickel and Doksum 1981; Hinkley and Runger 1984). In our case such questions are moot, because the uncertainty in measuring the transformation is of smaller order than that in estimating the regression function. Thus the asymptotic distributions that we have given are appropriate from both points of view. However, controversy would reemerge if one were interested in the constant c or some other root-n consistently estimable functional of the covariate effects.

APPENDIX A: PROOFS OF THEOREMS

The proofs of our theorems are based on several lemmas established in Appendix B. We make use of the uniform weak laws of large numbers (ULLN) established by Andrews (1995) and Newey (1991). We use the following notations:

$$M_\lambda = \begin{pmatrix} m_\lambda(X_1) \\ \vdots \\ m_\lambda(X_n) \end{pmatrix};\ \bar{M}_\lambda = \begin{pmatrix} \bar{m}_\lambda(X_1) \\ \vdots \\ \bar{m}_\lambda(X_n) \end{pmatrix};\ \Xi_\lambda = \begin{pmatrix} \varepsilon_\lambda(1) \\ \vdots \\ \varepsilon_\lambda(n) \end{pmatrix}.$$

We also use the following regularity conditions:

Assumption A

1. The kernel function $K(\cdot)$ is bounded, nonnegative, compactly supported, and Lipschitz continuous, and $\int K(u)\,du = 1$.
2. The sequence of bandwidths satisfies $h \to 0$ and $nh^{d+1} \to \infty$.
3. The densities $p_\alpha, p_{\underline{\alpha}}$, and p are bounded away from 0 and infinity and are Lipschitz continuous on their compact supports $\mathcal{X}_\alpha, \mathcal{X}_{\underline{\alpha}}$, and $\mathcal{X}$. Furthermore, $m_\lambda(x)$ is Lipschitz continuous on $\mathcal{X} \times \mathcal{N}(\lambda_0)$.
4. The random variables Z_i have a positive definite second moment matrix.
5. The stochastic process $\{Z_i\varepsilon_\lambda(i)\}_{\lambda\in\mathcal{N}(\lambda_0)}$ has a uniformly bounded first absolute moment and is also Lipschitz continuous in the sense that

$$|Z_i\varepsilon_{\lambda_1}(i) - Z_i\varepsilon_{\lambda_2}(i)| \le |\lambda_1 - \lambda_2|A$$

for all $\lambda_1, \lambda_2 \in \mathcal{N}(\lambda_0)$, where $E(A) < \infty$. Furthermore, $E[\varepsilon_\lambda(i)|Z_i] = 0$ for $\lambda \in \mathcal{N}(\lambda_0)$.
6. The stochastic process $\{[\bar{m}_\lambda(X_i) - m_\lambda(X_i)]Z_i\}_{\lambda\in\mathcal{N}(\lambda_0)}$ has a uniformly bounded first absolute moment.

These conditions are standard. For asymptotic normality, we require these additional conditions:

Assumption B

1. The density p (and hence p_α and $p_{\underline{\alpha}}$) and the regression function m_{λ_0} are q times (Lipschitz) continuously differentiable on $\mathcal{X}$.
2. $\int u^j K(u)\,du = 0, j = 1, \ldots, q$, and $K(\cdot)$ has a Lipschitz-continuous derivative.
3. The sequence of bandwidths satisfies $nh^{d+1} \to \infty, nh^{2q} \to 0$.
4. The stochastic processes $(Z_i\{[\partial\varepsilon_\lambda(i)]/\partial\lambda\})_{\lambda\in\mathcal{N}(\lambda_0)}$ and $(Z_i\{[\partial^2\varepsilon_\lambda(i)]/\partial\lambda^2\})_{\lambda\in\mathcal{N}(\lambda_0)}$ have uniformly bounded first absolute moments and are Lipschitz continuous with respect to λ on $\mathcal{N}(\lambda_0)$.
5. The stochastic processes $\{Z_i(\partial^j/\partial\lambda^j)[m_\lambda(X_i) - \bar{m}_\lambda(X_i)]\}_{\lambda\in\mathcal{N}(\lambda_0)}, j = 0, 1, 2$, have uniformly bounded first absolute moments and are also Lipschitz continuous with respect to λ on $\mathcal{N}(\lambda_0)$. Let $r_j(\lambda) = E[Z_i(\partial^j/\partial\lambda^j)\{m_\lambda(X_i) - \bar{m}_\lambda(X_i)\}], j = 0, 1, 2$, and assume that $r_1(\lambda_0) \ne 0$.
6. $E\{Z_iZ_i'\varepsilon^2_{\lambda_0}(i)\} < \infty$.

Note that conditions A5 and B4 imply that $E(\{[\partial\varepsilon_\lambda(i)]/\partial\lambda\}|Z_i)$ and $E(\{[\partial^2\varepsilon_\lambda(i)]/\partial\lambda^2\}|Z_i)$ are both 0.

Finally, for Theorem 3 we use the following conditions:

Assumption C

1. The conditions of Chen et al. (1996) hold. Suppose also that the family of random variables $\partial\theta_\lambda(Y)/\partial\lambda$ is tight in λ. Finally, suppose that $n^{1/2}(\hat{\lambda} - \lambda) = O_p(1)$.

Proof of Theorem 1

Write

$$\boldsymbol{\Theta}_\lambda - \tilde{\mathbf{M}}_\lambda = (\boldsymbol{\Theta}_\lambda - M_\lambda) + (M_\lambda + \bar{M}_\lambda) + (\bar{M}_\lambda - \tilde{\mathbf{M}}_\lambda).$$

Condition A5 is sufficient to guarantee that the ULLN holds for the first term; that is,

$$\sup_{\lambda\in\mathcal{N}(\lambda_0)} |n^{-1}\mathbf{Z}'(\boldsymbol{\Theta}_\lambda - M_\lambda)| = o_p(1).$$

Furthermore,

$$\sup_{\lambda\in\mathcal{N}(\lambda_0)} |n^{-1}\mathbf{Z}'(\bar{M}_\lambda - \tilde{\mathbf{M}}_\lambda)| = o_p(1)$$

by Cauchy–Schwarz, a weak law of large numbers applied to $\mathbf{Z}'\mathbf{Z}/n$, and Lemma 1. Hence, using Cauchy–Schwarz once again, we have

$$\begin{aligned} Q_n(\lambda) &= n^{-2}(M_\lambda - \bar{M}_\lambda)'\mathbf{Z}\mathbf{W}_n\mathbf{Z}'(M_\lambda - \bar{M}_\lambda) + o_p(1) \\ &= Q(\lambda) + o_p(1), \end{aligned}$$

where $Q(\lambda) = r_0'(\lambda)\mathbf{W}r_0(\lambda)$. The last equality is due to ULLN applied to $n^{-1}\mathbf{Z}'(M_\lambda - \bar{M}_\lambda)$, which holds under conditions A3 and A6. Because $m_{\lambda_0} - \bar{m}_{\lambda_0} = 0$, we have $Q(\lambda_0) = 0$.

Using the condition A7 and the fact that $\mathbf{W}$ is positive definite, we have $Q(\lambda) > 0$ for $\lambda \ne \lambda_0$ and $\lambda \in \mathcal{N}(\lambda_0)$. Hence λ_0 uniquely minimizes $Q(\lambda)$.

Proof of Theorem 2

By a Taylor expansion,

$$\begin{aligned} 0 = n^{1/2}s_n(\hat{\lambda}) &= n^{1/2}s_n(\lambda_0) + H_n(\lambda_0)n^{1/2}(\hat{\lambda} - \lambda_0) \\ &\quad + \{H_n(\lambda^*) - H_n(\lambda_0)\}n^{1/2}(\hat{\lambda} - \lambda_0), \quad \text{(A.1)} \end{aligned}$$

where $s_n(\lambda) = \partial Q_n/\partial\lambda$ and $H_n(\lambda) = \partial^2 Q_n/\partial\lambda^2$, whereas λ^* is between $\hat{\lambda}$ and λ_0. We have

$$s_n(\lambda) = 2n^{-2}(\boldsymbol{\Theta}_\lambda - \tilde{\mathbf{M}}_\lambda)'\mathbf{Z}\mathbf{W}_n\mathbf{Z}'\left(\frac{\partial\boldsymbol{\Theta}_\lambda}{\partial\lambda} - \frac{\partial\tilde{\mathbf{M}}_\lambda}{\partial\lambda}\right)$$

and

$$\begin{aligned} n^2 H_n(\lambda) &= 2(\boldsymbol{\Theta}_\lambda - \tilde{\mathbf{M}}_\lambda)\mathbf{Z}\mathbf{W}_n\mathbf{Z}'\left(\frac{\partial^2\boldsymbol{\Theta}_\lambda}{\partial\lambda^2} - \frac{\partial^2\tilde{\mathbf{M}}_\lambda}{\partial\lambda^2}\right) \\ &\quad + 2\left(\frac{\partial\boldsymbol{\Theta}_\lambda}{\partial\lambda} - \frac{\partial\tilde{\mathbf{M}}_\lambda}{\partial\lambda}\right)'\mathbf{Z}\mathbf{W}_n\mathbf{Z}'\left(\frac{\partial\boldsymbol{\Theta}_\lambda}{\partial\lambda} - \frac{\partial\tilde{\mathbf{M}}_\lambda}{\partial\lambda}\right). \end{aligned}$$

We show that

(a) $n^{1/2}s_n(\lambda_0) \Rightarrow \mathrm{N}(0, 4r_{10}'\mathbf{W}\Omega\mathbf{W}r_{10})$,
(b) $H_n(\lambda_0) = 2r_{10}'\mathbf{W}r_{10} + o_p(1)$, and
(c) $\{H_n(\lambda^*) - H_n(\lambda_0)\}n^{1/2}(\hat{\lambda} - \lambda_0)$ is of smaller order than the other terms.

(1997) Linton, O., Chen, R., Wang, N. and Härdle, W.
An analysis of transformations for additive nonparametric regression.

Journal of the American Statistical Association, 92, 1512-1521

We first establish (a). We have

$$n^{1/2}s_n(\lambda_0) = 2\{n^{-1/2}(\boldsymbol{\Theta}_{\lambda_0} - \tilde{\mathbf{M}}_{\lambda_0})'\mathbf{Z}\}\mathbf{W}_n \times \left\{n^{-1}\mathbf{Z}'\left(\frac{\partial\boldsymbol{\Theta}_{\lambda_0}}{\partial\lambda} - \frac{\partial\tilde{\mathbf{M}}_{\lambda_0}}{\partial\lambda}\right)\right\} \equiv 2T_{1n}T_{2n}T_{3n},$$

where $T_{2n} = \mathbf{W}_n = \mathbf{W} + o_p(1)$. An application of Lemma 2 yields

$$T_{3n} = r_1(\lambda_0) + o_p(1).$$

Finally, by Lemma 4, $T_{1n} \Rightarrow \mathrm{N}(0, \Omega)$. Therefore,

$$n^{1/2}s_n(\lambda_0) \Rightarrow \mathrm{N}(0, 4r_{10}'\mathbf{W}\Omega\mathbf{W}r_{10}).$$

We now turn to (b) and (c). These are proved by establishing that

$$\sup_{\lambda\in\mathcal{N}(\lambda_0)} |H_n(\lambda) - H(\lambda)| = o_p(1), \tag{A.2}$$

where

$$H(\lambda) = 2r_1(\lambda)'\mathbf{W}r_1(\lambda) + 2r_0(\lambda)'\mathbf{W}r_2(\lambda),$$

which implies (c). Then note that

$$H(\lambda_0) = 2r_1(\lambda_0)'\mathbf{W}r_1(\lambda_0),$$

because $r_0(\lambda_0) = 0$ by condition A7. This gives (b). Using the results in Theorem 1 and Lemmas 1, 2, and 3, (A.2) follows by the Cauchy–Schwarz inequality.

Using (a), (b), and (c), (10) follows. The consistency of the standard errors follows from the uniform consistency results used in the foregoing argument.

Proof of Theorem 3

By a Taylor expansion,

$$n^{2/5}\{\hat{\gamma}_\alpha(x_\alpha;\hat{\lambda}) - \gamma_\alpha(x_\alpha;\lambda_0)\} = n^{2/5}\{\hat{\gamma}_\alpha(x_\alpha;\lambda_0) - \gamma_\alpha(x_\alpha;\lambda_0)\} + n^{-1/10}\left\{\frac{\partial\hat{\gamma}_\alpha}{\partial\lambda}(x_\alpha;\lambda^*)n^{1/2}(\hat{\lambda} - \lambda_0)\right\}, \tag{A.3}$$

where λ^* is an intermediate point between $\hat{\lambda}$ and λ_0. Then

$$\frac{\partial\hat{\gamma}_\alpha}{\partial\lambda}(x_\alpha;\lambda) = n^{-1}\sum_{i=1}^n \frac{\sum_{j=1}^n K\left(\frac{x_\alpha - X_{\alpha j}}{h}\right)L\left(\frac{X_{\underline{\alpha}i} - X_{\underline{\alpha}j}}{g}\right)\frac{\partial\theta_\lambda}{\partial\lambda}(Y_j)}{\sum_{l=1}^n K\left(\frac{x_\alpha - X_{\alpha l}}{h}\right)L\left(\frac{X_{\underline{\alpha}i} - X_{\underline{\alpha}l}}{g}\right)} = O_p(1),$$

by C1. Therefore, the second term in (A.3) is $o_p(1)$. The result then follows from theorem 1 of Chen et al. (1996).

APPENDIX B: LEMMAS

Lemma 1. Suppose that conditions A1–A6 hold. Then

$$\sup_{x\in\mathcal{X}}\sup_{\lambda\in\mathcal{N}(\lambda_0)} |\tilde{m}_\lambda(x) - \bar{m}_\lambda(x)| = o_p(1).$$

Proof. We first show that

$$\sup_{x\in\mathcal{X}}\sup_{\lambda\in\mathcal{N}(\lambda_0)} |\hat{\gamma}_\alpha(x_\alpha;\lambda) - \gamma_\alpha(x_\alpha;\lambda)| = o_p(1).$$

We have

$$|\hat{\gamma}_\alpha(x_\alpha;\lambda) - \gamma_\alpha(x_\alpha;\lambda)| \le \left|n^{-1}\sum_{i=1}^n \{\hat{m}_\lambda(x_\alpha, X_{\underline{\alpha}i}) - m_\lambda(x_\alpha, X_{\underline{\alpha}i})\}\right| + \left|n^{-1}\sum_{i=1}^n \{m_\lambda(x_\alpha, X_{\underline{\alpha}i}) - \gamma_\alpha(x_\alpha;\lambda)\}\right|$$

by the triangle inequality. We can bound the first term on the right side by $\sup_{x_{\underline{\alpha}}} |\hat{m}_\lambda(x) - m_\lambda(x)|$, where

$$\sup_{x\in\mathcal{X}}\sup_{\lambda\in\mathcal{N}(\lambda_0)} |\hat{m}_\lambda(x) - m_\lambda(x)| = o_p(1)$$

(by Andrews 1995, thm. 3). Finally,

$$\sup_{x\in\mathcal{X}}\sup_{\lambda\in\mathcal{N}(\lambda_0)} \left|n^{-1}\sum_{i=1}^n m_\lambda(x_\alpha, X_{\underline{\alpha}i}) - \gamma_\alpha(x_\alpha;\lambda)\right| = o_p(1)$$

(by Newey 1991, cor. 3.1). Therefore, $\hat{\gamma}_\alpha(x_\alpha;\lambda)$ is uniformly consistent.

Now, because $\bar{m}_\lambda(x) = \sum_{\alpha=1}^d \gamma_\alpha(x_\alpha;\lambda) - (d-1)c_\lambda$ and $\tilde{m}_\lambda(x) = \sum_{\alpha=1}^d \hat{\gamma}_\alpha(x_\alpha;\lambda) - (d-1)\hat{c}_\lambda$, the conclusion of Lemma 2 follows by the triangle inequality and the fact that $\sup_{\lambda\in\mathcal{N}(\lambda_0)} |\hat{c}_\lambda - c_\lambda| = o_p(1)$.

Lemma 2. Suppose that conditions A1–A6 and B4 and B5 hold. Then

$$\sup_{\lambda\in\mathcal{N}(\lambda_0)} \left|n^{-1}\mathbf{Z}'\left(\frac{\partial\boldsymbol{\Theta}_\lambda}{\partial\lambda} - \frac{\partial\tilde{\mathbf{M}}_\lambda}{\partial\lambda}\right) - r_1(\lambda)\right| = o_p(1).$$

Proof. We have

$$n^{-1}\mathbf{Z}'\frac{\partial}{\partial\lambda}(\boldsymbol{\Theta}_\lambda - \tilde{\mathbf{M}}_\lambda) = n^{-1}\mathbf{Z}'\frac{\partial}{\partial\lambda}\{(\boldsymbol{\Theta}_\lambda - M_\lambda) + (M_\lambda - \bar{M}_\lambda) + (\bar{M}_\lambda + \tilde{\mathbf{M}}_\lambda)\}.$$

The first term is uniformly $o_p(1)$, by the ULLN guaranteed by condition B4. The second term converges in probability to its mean constant vector $r_1(\lambda)$ by the same reasoning with condition B5. The last term can be proven to be $o_p(1)$ by noting that

$$\sup_{x\in\mathcal{X}}\sup_{\lambda\in\mathcal{N}(\lambda_0)} \left|\frac{\partial\tilde{m}_\lambda}{\partial\lambda}(x) - \frac{\partial\bar{m}_\lambda}{\partial\lambda}(x)\right| = o_p(1),$$

by the same arguments as in Lemma 1.

Lemma 3. Suppose that conditions A1–A6 and B4 and B5 hold. Then

$$\sup_{\lambda\in\mathcal{N}(\lambda_0)} \left|n^{-1}\mathbf{Z}'\left(\frac{\partial^2\boldsymbol{\Theta}_\lambda}{\partial\lambda^2} - \frac{\partial^2\tilde{\mathbf{M}}_\lambda}{\partial\lambda^2}\right) - r_2(\lambda)\right| = o_p(1).$$

Proof. Similar to Lemma 2.

Lemma 4. Suppose that conditions A1, A3–A5, A7, B1–B3, and B6 hold. Then

$$n^{-1/2}(\boldsymbol{\Theta}_{\lambda_0} - \tilde{\mathbf{M}}_{\lambda_0})'\mathbf{Z} \Rightarrow \mathrm{N}(0, \Omega).$$

(1997) Linton, O., Chen, R., Wang, N. and Härdle, W.
An analysis of transformations for additive nonparametric regression.

Journal of the American Statistical Association, 92, 1512-1521

Proof. Write

$$\begin{aligned} T_{1n} &= n^{-1/2}\{(\Theta_{\lambda_0} - \bar{M}_{\lambda_0}) + (M_{\lambda_0} - \bar{M}_{\lambda_0}) \\ &\quad + (\bar{M}_{\lambda_0} - \tilde{\mathbf{M}}_{\lambda_0})\}'\mathbf{Z} \\ &= n^{-1/2}\Xi'_{\lambda_0}\mathbf{Z} + n^{-1/2}(\bar{M}_{\lambda_0} - \tilde{\mathbf{M}}_{\lambda_0})'\mathbf{Z}. \end{aligned}$$

because $M_{\lambda_0} - \bar{M}_{\lambda_0} = 0$. We have

$$n^{-1/2}(\tilde{\mathbf{M}}_{\lambda_0} - \bar{M}_{\lambda_0})'\mathbf{Z} = \mathrm{I} + \mathrm{II} + \mathrm{III},$$

where

$$\mathrm{I} = n^{-3/2}\sum_{\alpha=1}^{d}\sum_{i=1}^{n}\sum_{j=1}^{n} Z_i\{\hat{m}_{\lambda_0}(X_{\alpha i}, X_{\underline{\alpha} j}) - m_{\lambda_0}(X_{\alpha i}, X_{\underline{\alpha} j})\},$$

$$\mathrm{II} = -(d-1)n^{-1}\sum_{i=1}^{n} Z_i \times n^{-1/2}\sum_{j=1}^{n}\varepsilon_{\lambda_0}(j),$$

and

$$\mathrm{III} = n^{-3/2}\sum_{\alpha=1}^{d}\sum_{i=1}^{n}\sum_{j=1}^{n} Z_i\left\{m_{\lambda_0}(X_{\alpha i}, X_{\underline{\alpha} j}) - \int m_{\lambda_0}(X_{\alpha i}, x_{\underline{\alpha}})p_{\underline{\alpha}}(x_{\underline{\alpha}})\,dx_{\underline{\alpha}}\right\}.$$

because Z_i is mean 0, $\mathrm{II} = o_p(1)$. Second,

$$\mathrm{III} = n^{-1}\sum_{i=1}^{n} Z_i \times (d-1)n^{-1/2}\sum_{\alpha=1}^{d}\sum_{j=1}^{n} g_\alpha(X_{\alpha j}) = o_p(1),$$

because $E\{g_\alpha(X_{\alpha j})\} = 0$. Finally, we show that

$$\mathrm{I} = n^{-1/2}\sum_{j=1}^{n}\sum_{\alpha=1}^{d} w_{\alpha j}\varepsilon_{\lambda_0}(j) + o_p(1), \tag{B.1}$$

where $\omega_{\alpha j} = E\,(Z_i|X_\alpha = X_{\alpha j})\{[p_\alpha(x_{\alpha j})p_{\underline{\alpha}}(X_{\underline{\alpha} j})]/[p(X_j)]\}$.

To prove (B.1), let $L_h(t_{\underline{\alpha}}) = \prod_{\beta\neq\alpha} K_h(t_\beta)$ for any vector $\mathbf{t} = (t_\alpha, t_{\underline{\alpha}})$, and write

$$\mathrm{I} = n^{-1/2}\sum_{i=1}^{n}\sum_{\alpha=1}^{d}\left\{\frac{1}{n}\sum_{j=1}^{n}\frac{\hat{a}(X_{\alpha i}, X_{\underline{\alpha} j})}{\hat{p}(X_{\alpha i}, X_{\underline{\alpha} j})}\right\} Z_i, \tag{B.2}$$

where

$$\hat{a}(x) = \hat{r}(x) - \hat{p}(x)m_{\lambda_0}(x),$$

$$\hat{r}(x) = \frac{1}{n}\sum_{l=1}^{n} L_h(x_{\underline{\alpha}} - X_{\underline{\alpha} l})K_h(x_\alpha - X_{\alpha l})\theta_{\lambda_0}(Y_l),$$

and

$$\hat{p}(x) = \frac{1}{n}\sum_{l=1}^{n} L_h(x_{\underline{\alpha}} - X_{\underline{\alpha} l})K_h(x_\alpha - X_{\alpha l}).$$

Using the arguments of Chen et al. (1996), $\mathrm{I} = (T_{4n} + T_{5n})\{1 + o_p(1)\}$, where

$$T_{4n} = n^{-1/2}\sum_{i=1}^{n}\sum_{\alpha=1}^{d}\frac{1}{n}\sum_{j=1}^{n}\frac{\hat{a}(X_{\alpha i}, X_{\underline{\alpha} j}) - E_*\{\hat{a}(X_{\alpha i}, X_{\underline{\alpha} j})\}}{p(X_{\alpha i}, X_{\underline{\alpha} j})} Z_i$$

and

$$T_{5n} = n^{-1/2}\sum_{i=1}^{n}\sum_{\alpha=1}^{d}\frac{1}{n}\sum_{j=1}^{n}\frac{E_*\{\hat{a}(X_{\alpha i}, X_{\underline{\alpha} j})\}}{p(X_{\alpha i}, X_{\underline{\alpha} j})} Z_i,$$

where E_* denotes expectation conditional on $X_1, \ldots, X_n$. Note that

$$\begin{aligned} &\hat{a}(X_{\alpha i}, X_{\underline{\alpha} j}) - E_*\{\hat{a}(X_{\alpha i}, X_{\underline{\alpha} j})\} \\ &\qquad = \frac{1}{n}\sum_{l=1}^{n} L_h(X_{\underline{\alpha} j} - X_{\underline{\alpha} l})K_h(X_{\alpha i} - X_{\alpha l})\varepsilon_{\lambda_0}(l), \end{aligned}$$

so

$$T_{4n} = n^{-1/2}\sum_{\alpha=1}^{d}\sum_{l=1}^{n}\varepsilon_{\lambda_0}(l)\left\{\frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}\omega_{\alpha l i j}\right\},$$

where $\omega_{\alpha l i j} = L_h(X_{\underline{\alpha} j} - X_{\underline{\alpha} l})K_h(X_{\alpha i} - X_{\alpha l})Z_i/p(X_{\alpha i}, X_{\underline{\alpha} j})$. By the same integration arguments of Chen et al. (1996, thm. 1), we can replace $1/n^2\sum_i\sum_j \omega_{\alpha l i j}$ by $\omega_{\alpha l}$. Thus

$$T_{4n} = n^{-1/2}\sum_{l=1}^{n}\sum_{\alpha=1}^{d}\omega_{\alpha l}\varepsilon_{\lambda_0}(l) + o_p(1),$$

as required. By direct calculation, the bias term T_{5n} is $O_p(n^{1/2}h^q)$ by conditions B1 and B2 and is $o_p(1)$ by B3.

In conclusion, we have shown that

$$n^{-1/2}(\tilde{\mathbf{M}}_{\lambda_0} - \bar{M}_{\lambda_0})'\mathbf{Z} = n^{-1/2}\sum_{j=1}^{n} Z_j^*\varepsilon_{\lambda_0}(j) + o_p(1).$$

By B4 and B6, we can apply the central limit theorem to

$$n^{-1/2}\sum_{j=1}^{n}\{Z_j - Z_j^*\}\varepsilon_{\lambda_0}(j),$$

as required.

[Received September 1995. Revised December 1996.]

REFERENCES

Amemiya, T., and Powell, J. L. (1981), "A Comparison of the Box–Cox Maximum Likelihood Estimator and the Non-Linear Two-Stage Least Squares Estimator," *Journal of Econometrics*, 17, 351–381.

Andrews, D. W. K. (1995), "Nonparametric Kernel Estimation for Semiparametric Models," *Econometric Theory*, 11, 560–596.

Angrist, J. D., Imbens, G. W., and Rubin, D. B. (1996), "Identification of Causal Effects Using Instrumental Variables," (with discussion) *Journal of the American Statistical Association*, 91, 444–472.

Auestad, B., and Tjøstheim, D. (1991), "Functional Identification in Nonlinear Time Series," in *Nonparametric Functional Estimation and Related Topics*, ed. G. Roussas, Dordrecht: Kluwer, pp. 493–507.

Berndt, E. (1991), *The Practice of Econometrics*, Reading, MA: Addison-Wesley.

Bickel, P. J., and Doksum, K. A. (1981), "An Analysis of Transformations Revisited," *Journal of the American Statistical Association*, 76, 296–311.

Bickel, P. J., Klaassen, C. A. J., Ritov, Y., and Wellner, J. A. (1993), *Efficient and Adaptive Inference in Semiparametric Models*. Baltimore: Johns Hopkins University, Press.

Box, G. E. P., and Cox, D. R. (1964), "An Analysis of Transformations," *Journal of the Royal Statistical Society*, Ser. B, 26, 211–252.

Breiman, L., and Friedman, J. H. (1985), "Estimating Optimal Transformations for Multiple Regression and Correlation," (with discussion) *Journal of the American Statistical Association*, 80, 580–619.

Burbidge, J. B., Magee, L., and Robb, A. L. (1988), "Alternative Transformations to Handle Extreme Values of the Dependent Variable," *Journal of the American Statistical Association*, 83, 123–127.

Buja, A., Hastie, T., and Tibshirani, R. (1989), "Linear Smoothers and Additive Models," (with discussion), *The Annals of Statistics*, 17, 453–555.

Carroll, R. J. (1980), "A Robust Method of Testing Transformations to Achieve Approximate Normality," *Journal of the Royal Statistical Society*, Ser. B, 42, 71–78.

(1997) Linton, O., Chen, R., Wang, N. and Härdle, W.
An analysis of transformations for additive nonparametric regression.

——— (1982), "Prediction and Power Transformations When the Choice of Power is Restricted to a Finite Set," *Journal of the American Statistical Association*, 77, 908–915.

Carroll, R. J., and Ruppert, D. (1984), "Power Transformation When Fitting Theoretical Models to Data," *Journal of the American Statistical Association*, 79, 321–328.

——— (1988), *Transformation and Weighting in Regression*, New York: Chapman and Hall.

Carroll, R. J., and Wand, M. P. (1991), "Semiparametric Estimation in Logistic Measurement Error Models," *Journal of the Royal Statistical Society*, Ser. B, 53, 573–585.

Chen, R., Härdle, W., Linton, O. B., and Severance-Lossin, E. (1996), "Estimation in Additive Nonparametric Regression," *Proceedings of the COMPSTAT Conference Semmering*, eds. W. Härdle and M. Schimek, Heidelberg: Physika Verlag.

Ehrlich, I. (1977), "Capital Punishment and Deterrence: Some Further Thoughts and Additional Evidence," *Journal of Political Economy*, 85, 741–788.

Engle, R. F., Granger, C. W. J., Rice, J., and Weiss, A. (1986), "Semiparametric Estimates of the Relation Between Weather and Electricity Sales," *Journal of the American Statistical Association*, 81, 310–320.

Fan, J. (1992), "Design-Adaptive Nonparametric Regression," *Journal of the American Statistical Association*, 87, 998–1004.

Fan, J., Härdle, W., and Mammen, E. (in press), "Direct Estimation of Low-Dimensional Components in Additive Models," submitted to *The Annals of Statistics*.

Gutierrez, R., and Carroll, R. J. (1995), "Plug-In Semiparametric Estimating Equations," unpublished manuscript, Texas A&M University.

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

Hastie, T. J., and Tibshirani, R. J. (1990), *Generalized Additive Models*, London: Chapman and Hall.

Heckman, J. J., and Polachek, S. (1974), "Empirical Evidence on the Functional Form of the Earnings–Schooling Relationship," *Journal of the American Statistical Association*, 69, 350–354.

Hengartner, N. W. (1996), "Rate-Optimal Estimation of Additive Regression via the Integration Method in the Presence of Many Covariates," preprint, Yale University, Dept. of Statistics.

Hinkley, D. V., and Runger, G. (1984), "The Analysis of Transformed Data" (with discussion), *Journal of the American Statistical Association*, 79, 302–320.

Hulten, C. R., and Wykoff, F. C. (1981), "The Estimation of Economic Depreciation Using Vintage Asset Prices," *Journal of Econometrics*, 15, 367–396.

Ibragimov, I. A., and Hasminskii, R. Z. (1980), "On Nonparametric Estimation of Regression," *Soviet Mathematics. Doklady*, 21, 810–4.

Johnson, N. L. (1949), "Systems of Frequency Curves Generated by Methods of Translation," *Biometrika*, 36, 149–176.

Linton, O. B., and Härdle, W. (1996), "Estimating Additive Regression With Known Links," *Biometrika*, 83, 529–540.

Linton, O. B., and Nielsen, J. P. (1995), "Estimating Structured Nonparametric Regression of the Kernel Method," *Biometrika*, 82, 93–101.

Mukarjee, H., and Stern, S. (1994), "Feasible Nonparametric Estimation of Multiargument Monotone Functions," *Journal of the American Statistical Association*, 89, 77–80.

Newey, W. K. (1990), "Semiparametric Efficiency Bounds," *Journal of Applied Econometrics*, 5, 99–135.

——— (1991), "Uniform Convergence in Probability and Stochastic Equicontinuity," *Econometrica*, 59, 1161–1168.

——— (1994), "Kernel Estimation of Partial Means," *Econometric Theory*, 10, 233–253.

Newey, W. K., and McFadden, D. (1994), "Large Sample Estimation and Hypothesis Testing," in *Handbook of Econometrics*, Vol. 4, eds. R. F. Engle and D. L. McFadden, Amsterdam: Elsevier.

Robins, J., Rotnizky, A., and Zhao, L. (1994), "Estimation of Regression Coefficients When Some Regressors are not Always Observed," *Journal of the American Statistical Association*, 89, 846–857.

Robinson, P. M. (1988), "Root-n-Consistent Semiparametric Regression," *Econometrica*, 56, 931–954.

——— (1991), "Best Nonlinear Three-Stage Least Squares Estimation of Certain Econometric Models," *Econometrica*, 59, 755–786.

Stone, C. J. (1980), "Optimal Rates of Convergence for Nonparametric Estimators," *The Annals of Statistics*, 8, 1348–1360.

——— (1982), "Optimal Global Rates of Convergence for Nonparametric Regression," *The Annals of Statistics*, 8, 1040–1053.

——— (1986), "The Dimensionality Reduction Principle for Generalized Additive Models," *The Annals of Statistics*, 14, 592–606.

Tibshirani, R. (1988), "Estimating Optimal Transformations for Regression via Additivity and Variance Stabilization," *Journal of the American Statistical Association*, 83, 394–405.

Tjøstheim, D., and Auestad, B. (1994), "Nonparametric Identification of Nonlinear Time Series: Projections," *Journal of the American Statistical Association*, 89, 1398–1409.

Wang, N., and Ruppert, D. (1996), "Estimation of Regression Parameters in a Semiparametric Transformation Model," *Journal of Statistical Planning and Inference*, 52, 331–351.

Zarembka, P. (1968), "Functional Form in the Demand for Money," *Journal of the American Statistical Association*, 63, 502–511.

——— (1974), "Transformations of Variables in Econometrics," in *Frontiers in Econometrics*, ed. P. Zarembka, Boston: Academic Press.

Zellner, A., and Revankar, N. (1969), "Generalized Production Functions," *Review of Economic Studies*, 37, 241–250.

J. Stat. Planning. Inference, 68, 221-245

ELSEVIER

Journal of Statistical Planning and Inference 68 (1998) 221–245

journal of statistical planning and inference

Nonparametric vector autoregression

W. Härdle [a], A. Tsybakov [b], L. Yang [a,*]

[a] *Humboldt-Universität zu Berlin, Wirtschaftswissenschaftliche Fakultät, Institut für Statistik und Ökonometrie, Spandauer Strasse 1, D-10178 Berlin, Germany*

[b] *Université Paris VI, Laboratoire de Statistique Théorique et Appliquée, 4, pl. Jussieu, Tour 45-55, F-75252 Paris, France*

Received 1 March 1996; received in revised form 17 December 1996; accepted 6 January 1997

Abstract

We consider a vector conditional heteroscedastic autoregressive nonlinear (CHARN) model in which both the conditional mean and the conditional variance (volatility) matrix are unknown functions of the past. Nonparametric estimators of these functions are constructed based on local polynomial fitting. We examine the rates of convergence of these estimators and give a result on their asymptotic normality. These results are applied to estimation of volatility matrices in foreign exchange markets. Estimation of the conditional covariance surface for the Deutsche Mark/US Dollar (DEM/USD) and Deutsche Mark/British Pound (DEM/GBP) daily returns show negative correlation when the two series have opposite lagged values and positive correlation elsewhere. The relation of our findings to the capital asset pricing model is discussed.

1. Nonparametric vector autoregression

Multivariate time series occur in many scientific disciplines. Their analysis helps in modeling dynamics over time as well as explaining interdependence among variables. A common model in this context is vector autoregression where the dynamics over time are modeled via a linear operation on the past values of the vector time series, see Lütkepohl (1991). Typically, in these models the conditional covariance is assumed to be either fixed or of specific form. Since the beginning of the 1980s the drawback of fixed linear structures has been stressed by Engle (1982), Robinson (1983, 1984) and Teräsvirta (1994) in the econometric literature and by Collomb (1984), Tjøstheim (1994), McKeague and Zhang (1994), and Vieu (1994) in the statistical literature. Nonlinear time-series models that have been proposed are, e.g., threshold autoregressive (TAR) models of Tong (1978, 1983), the exponential autoregressive (EXPAR) models of Haggan and Ozaki (1981), the smooth-transition autoregressive (STAR) models of Chan and Tong (1986) and Granger and Teräsvirta (1992).

* Corresponding author. E-mail: yang@wiwi.hu-berlin.de.

0378-3758/98/$19.00
PII S0378-3758(97)00143-2

J. Stat. Planning. Inference, 68, 221-245

In the analysis of financial time series, e.g., exchange rates, models for conditional heteroscedasticity are an important feature. Meese and Rose (1991) state that "*it is now recognized that empirical exchange rate models of the post-Bretton Woods era are characterized by parameter instability and dismal forecast performance...*" This pessimism about the quality of exchange-rate models became generally accepted after the publication of the influential papers by Meese and Rogoff (1983) and Diebold and Nason (1990).

The nonparametric modeling of the mean function and the volatility matrix offers a way out of this pessimism. It does not depend on specific structures of these quantities and may thus lead to valuable suggestions. In the framework of ARCH models (Engle, 1982), non- and semi-parametric approaches (Gregory, 1989; Engle and Gonzalez-Rivera, 1991) have been proposed. Engle and Ng (1993) measured the impact of news on volatility and found asymmetric volatility functions. Gouriéroux and Monfort (1992) models both the conditional mean and the conditional variance in a flexible nonparametric way

$$Y_i = \sum_{j=1}^{J} \alpha_j \, I(X_i \in A_j) + \sum_{j=1}^{J} \beta_j \, I(X_i \in A_j)\xi_i, \quad i = 1, 2, \ldots,$$

$$X_i = (Y_{i-1}^{\mathrm{T}}, Y_{i-2}^{\mathrm{T}}, \ldots, Y_{i-m}^{\mathrm{T}})^{\mathrm{T}} \in \mathbb{R}^{md}, \quad Y_i \in \mathbb{R}^d \tag{1.1}$$

is called a qualitative threshold ARCH model. Here $\{A_j\}_{j=1}^{J}$ with fixed J denotes a partition of the set of lagged values for Y, (α_j), and (β_j) are unknown parameter vectors and matrices, respectively, and ξ_i is the white noise. It is a generalization of the threshold model (Tong, 1983), for the conditional mean but shares with it the drawback of a fixed number J of threshold points.

A generalization of model (1.1) to a wider class of conditional mean and variance functions can be seen as a limit of (1.1) for $J \to \infty$, thus allowing J to be unknown

$$Y_i = f(X_i) + \Sigma^{1/2}(X_i)\xi_i, \quad i = 1, 2, \ldots,$$

$$X_i = (Y_{i-1}^{\mathrm{T}}, Y_{i-2}^{\mathrm{T}}, \ldots, Y_{i-m}^{\mathrm{T}})^{\mathrm{T}} \in \mathbb{R}^{md}, \quad Y_i \in \mathbb{R}^d. \tag{1.2}$$

We call (1.2) a conditional heteroskedastic autoregressive nonlinear (CHARN) model. It is a generalization of an ARCH structure.

The use of CHARN modeling is motivated by several examples. It has been found that GARCH(1, 1) processes fit daily and weekly FX (foreign exchange) rates well in most cases. The situation for intra-daily data is different though, see Guillaume et al. (1994).

Drost and Nijman (1993) argued that the specific GARCH structure would not allow arbitrary combinations of conditional heteroskedasticity, and leptokurtocity, for example. Typically, for intra-daily data the deviation of the unconditional return density from normality increases when the sampling interval is decreased. The model (1.2) will not suffer from these effects since it neither makes structural assumptions on f and Σ nor distributional assumptions on ξ. The situation for the CHARN model is

J. Stat. Planning. Inference, 68, 221-245

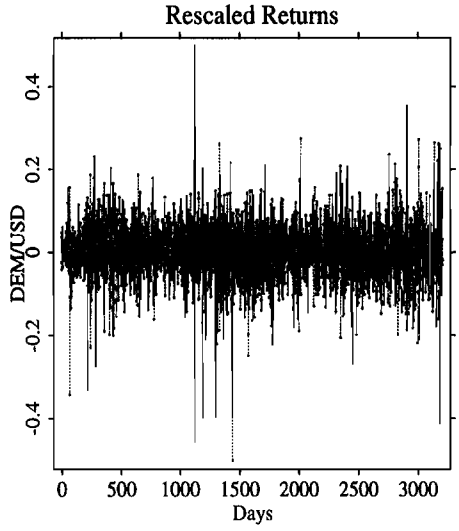


Fig. 1. The daily returns of the exchange rates of DEM/USD from 2 January 1980 to 30 October 1992.

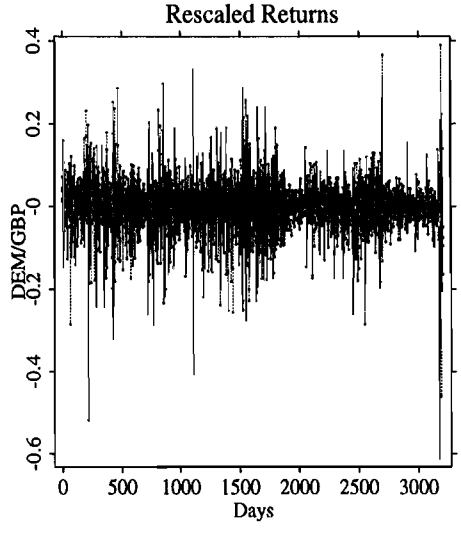


Fig. 2. The daily returns of the exchange rates of DEM/GBP from 2 January 1980 to 30 October 1992.

depicted in Figs. 1–3. All computations and graphics are done in XploRe, see Härdle et al. (1995).

Figs. 1 and 2 show the daily returns (differences of log spot rates) of $Y_{i1} = \text{DEM/USD}$ (Deutsche Mark/US Dollar) and of $Y_{i2} = \text{DEM/GBP}$ (Deutsche Mark/British Pound) for the period from 2 January 1980 to 30 October 1992, a total of 3212 observations: both are rescaled so that the range always has length 1. Fig. 3 shows that the two returns are highly correlated, the correlation equals 0.34, and the squared returns (i.e.

J. Stat. Planning. Inference, 68, 221-245

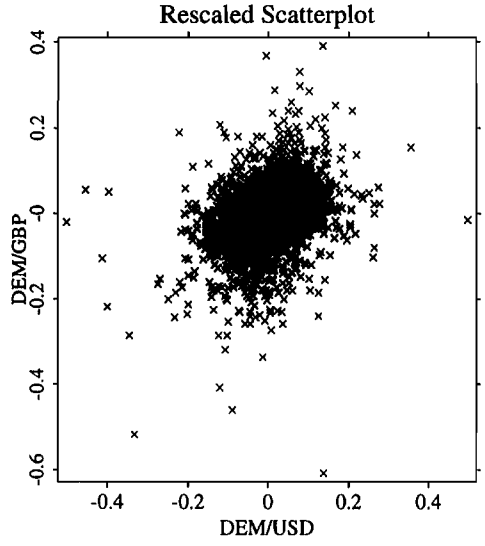


Fig. 3. The daily returns of the exchange rates of both DEM/USD and DEM/GBP from 2 January 1980 to 30 October 1992.

Y_{i1}^2 and Y_{i2}^2) also have a correlation of 0.17. Both are statistically significantly different from zero, for a sample size of 3212.

Figs. 4 and 5 display the conditional covariance function as dependent on one lag. Thus, in (1.2) we have $d=2$, $m=1$ and the task is to estimate

$$f(x)=(f_1,f_2)^{\mathrm{T}}(x) \quad \text{and} \quad \Sigma(x)=\begin{pmatrix}\sigma_{11} & \sigma_{12}\\ \sigma_{21} & \sigma_{22}\end{pmatrix}(x).$$

There exists a negative correlation when the two returns have opposite lagged values, which correspond to the upper left and the lower right corners of the contour plot or the lowest contour level at about 15.76% below which are the negative values, while positive correlations are everywhere else. Both the computation and graphics are done in XploRe, using the WARPing technique (Härdle et al., 1995), subsequent work in Section 4 is done in the same fashion and uses the same single bandwidth.

Härdle and Tsybakov (1996) proposed a general class of joint mean and volatility-function estimators based on the local polynomial (LP) method in the case of one-lag-dependence model (1.2) with one-dimensional Y_i. The LP estimator was chosen in favor of the Nadaraya–Watson (NW) estimator, since the NW estimator does not achieve good asymptotic convergence rates, unless the marginal (stationary) density of X_i is sufficiently many times differentiable. Sufficient conditions for such a property to hold in the model (1.2) are not known. The LP method avoids this difficulty, since it needs only the continuity of the density of X_i. A more practical reason to use the LP method is that it corresponds to a local least-squares problem, and for this problem easy and efficient algorithms are available. Bossaerts et al. (1996) used this method to study foreign exchange rates. For large dimension d and many lags m, however, the

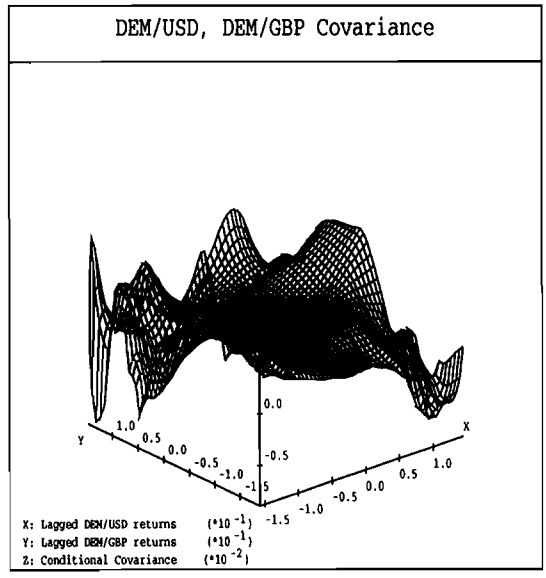


Fig. 4. The conditional covariance, using bandwidth $h = 0.0536531$.

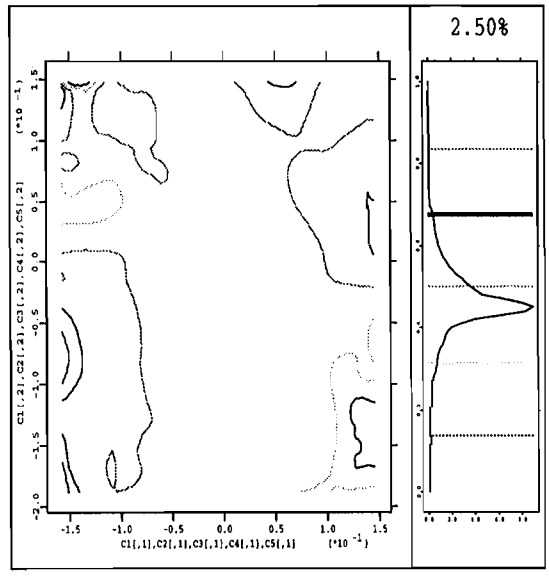


Fig. 5. The contours of the conditional covariance.

precision of the estimators of both f and Σ will decrease. A structured modelling based on additive assumptions has therefore been proposed by Chen and Tsay (1993a, b).

The idea of local polynomial estimation goes back to Stone (1977), Cleveland (1979) and Katkovnik (1979, 1985). The statistical properties of LP estimators in

J. Stat. Planning. Inference, 68, 221-245

nonparametric regression (convergence, minimax rate of convergence and pointwise asymptotic normality) were studied by Tsybakov (1986). The LP estimation method was later discussed by several authors (see Fan and Gijbels, 1996, for references). For the multidimensional case, we refer to the work of Ruppert and Wand (1994) who studied the multivariate local linear regression estimation.

This paper is devoted to estimation of the $f(\cdot)$ and $\Sigma(\cdot)$ functions for the multivariate CHARN model. We generalize to the vector case the result of Härdle and Tsybakov (1996) on asymptotic normality of LP estimators. We restrict the study, however, to the local linear case. This is motivated by the fact that higher-order polynomial estimation in higher dimension is less attractive computationally, while the expressions for asymptotic bias and variance are much more technical, and they do not seem to be of practical use.

Inspection of the proofs in Section 5 shows that the result of the present paper also holds (with obvious reformulation) for the multivariate nonparametric regression model with heteroskedastic errors: $Y_i = f(X_i) + \Sigma^{1/2}(X_i)\xi_i$, where ξ_i are as in (1.1), (X_i, Y_i) are i.i.d., and the design points $\{X_i\}$ are independent of $\{\xi_i\}$.

We shall use the work on probabilistic properties of the process (1.2): Doukhan and Ghindès (1980, 1981), Chan and Tong (1985), Mokkadem (1987), Diebolt and Guégan (1990), Ango Nze (1992). In these papers the ergodicity, geometric ergodicity and mixing properties of the process $\{Y_i\}$ are derived under appropriate conditions.

The paper is organized as follows. In Section 2, we present the estimator and in Section 3 we study the asymptotic properties of this LP technique. In Section 4 we give an application based on the two-dimensional data of DEM/USD and DEM/GBP returns. In Section 5, proofs of theorems are given.

2. The estimators

The model we consider is

$$Y_i = f(X_i) + \Sigma^{1/2}(X_i)\xi_i, \quad i = m, m+1, \ldots, \tag{2.1}$$

where $Y_i = (Y_{i1}, Y_{i2}, \ldots, Y_{id})^{\mathrm{T}} \in \mathbb{R}^d$, $\xi_i = (\xi_{i1}, \xi_{i2}, \ldots, \xi_{id})^{\mathrm{T}} \in \mathbb{R}^d$, $i = m, m+1, \ldots, n$, and $X_i = (Y_{i-1}^{\mathrm{T}}, Y_{i-2}^{\mathrm{T}}, \ldots, Y_{i-m}^{\mathrm{T}})^{\mathrm{T}} \in \mathbb{R}^{md}$ are random vector variables; ξ_i are i.i.d. with $E(\xi_{1j}) = 0$, for any $1 \leqslant j \leqslant d$, $E(\xi_{1j}^2) = 1$. The mean vector function $f : \mathbb{R}^{md} \to \mathbb{R}^d$ and volatility matrix function $\Sigma : \mathbb{R}^{md} \to \mathbb{R}^d \times \mathbb{R}^d$ are unknown, $\Sigma(x)$ is symmetric and positive definite for any $x \in \mathbb{R}^{md}$, and the initial value $X_m = (Y_{m-1}^{\mathrm{T}}, Y_{m-2}^{\mathrm{T}}, \ldots, Y_0^{\mathrm{T}})^{\mathrm{T}}$ is a random vector variable independent of $\{\xi_i\}$. We study the problem of estimating the conditional volatility matrix function $\Sigma(x)$ and the conditional mean vector function $f(x)$, given a time series $Y_0, \ldots, Y_n$.

The technique we employ here is typical in multivariate problems. Instead of Σ and f, we can equivalently estimate the following functions:

- The mean function of $v^{\mathrm{T}}Y$, which is $f(x; v) = v^{\mathrm{T}} f(x)$, where $v \in \mathbb{R}^d$ has unit length and $x \in \mathbb{R}^{md}$;

(1998) Härdle, W., Tsybakov, A.B. and Yang, L.
Nonparametric Vector Auto-regression.

J. Stat. Planning. Inference, 68, 221-245

- The covariance function of $v^{\mathrm{T}}Y$ and $s^{\mathrm{T}}Y$, which is $v^{\mathrm{T}}\Sigma(x)s$, where $v,s\in\mathbb{R}^d$ both have unit length and $x\in\mathbb{R}^{md}$.

For the moment we are implicitly assuming stationarity of $\{Y_i\}$. In fact, only an approximation is true: $\{X_i\}$ approaches a stationary process, for $i\to\infty$ as we shall see later in Lemma 3.1. The LP method solves the following minimization problems:

$$\begin{aligned} c_n(x;v,s) &= \arg\min_{c\in\mathbb{R}^{md+1}} \sum_{i=m}^{n} (v^{\mathrm{T}}Y_iY_i^{\mathrm{T}}s - c^{\mathrm{T}}U_{in})^2 K_h(X_i - x), \\ c_n(x;v) &= \arg\min_{c\in\mathbb{R}^{md+1}} \sum_{i=m}^{n} (v^{\mathrm{T}}Y_i - c^{\mathrm{T}}U_{in})^2 K_h(X_i - x), \end{aligned} \tag{2.2}$$

where $K:\mathbb{R}^{md}\to\mathbb{R}^1$ is a kernel $K_h(u)=1/h^{md}K(u/h)$, $h=h_n$ is a positive number (bandwidth), $h_n\to 0$, as $n\to\infty$ and

$$U_{in}=F(u_{in}), \qquad u_{in}=\frac{X_i-x}{h}, \tag{2.3}$$

where $F(u)=\binom{1}{u}\in\mathbb{R}^{md+1}$, for $u\in\mathbb{R}^{md}$. The estimator of $f(x;v)$ is defined as

$$\hat{f}(x;v)=c_n(x;v)^{\mathrm{T}}F(0).$$

The estimator of the function $\sigma(x;v,s)=v^{\mathrm{T}}\Sigma(x)s$ is defined as

$$\hat{\sigma}(x;v,s)=c_n(x;v,s)^{\mathrm{T}}F(0) - \{c_n(x;v)^{\mathrm{T}}F(0)\}\{c_n(x;s)^{\mathrm{T}}F(0)\}. \tag{2.4}$$

We have dropped reference to the sample size n in $\hat{f}(x;v)$ and $\hat{\sigma}(x;v,s)$ for notational simplicity, we will keep this convention in similar situations hereafter. Another simplification of notation is the use of one single bandwidth in all coordinates of X. The asymptotic results in the next section are easily extendable to the case of different bandwidth in each direction, e.g., in a product kernel

$$K_h(u)=\left(\prod_{j=1}^{md}h_j\right)^{-1}\prod_{j=1}^{md}K\left(\frac{u_j}{h_j}\right),$$

where $h=(h_1,\ldots,h_{md})\in\mathbb{R}_+^{md}$, see Wand and Jones (1995).

3. The asymptotic results

Let $|\cdot|$ denote the L_1-norm when it is applied to a vector, and the usual matrix norm

$$|A| = \sup_{|x|=1}|Ax|,$$

when it is applied to a matrix A. Assume the following:

(A1) The error variables ξ_{1j}, $1\leqslant j\leqslant d$, are i.i.d. The density $p(\cdot)$ of ξ_1 exists and satisfies

$$\inf_{x\in\mathcal{K}} p(x)>0$$

J. Stat. Planning. Inference, 68, 221-245

for any compact $\mathscr{K} \subset \mathbb{R}^d$. Also $E(\xi_{1j}) = E(\xi_{1j}^3) = 0$, $E(\xi_{1j}^2) = 1$, and $E(\xi_{1j}^4) = 1 + m_4 < \infty$.

(A2) There exist constants $C_1 \geqslant 0$, $C_2 \geqslant 0$, $r > 0$ such that for $|x| \geqslant r$

$$|f(x)| \leqslant C_1(1 + |x|), \tag{3.1}$$

$$|\Sigma^{1/2}(x)| \leqslant C_2(1 + |x|). \tag{3.2}$$

(A3) The matrix function $\Sigma(x)$ is symmetric for any $x \in \mathbb{R}^{md}$, and satisfies

$$\inf_{x \in \mathscr{K}} \lambda_{\min}\{\Sigma(x)\} > \lambda_{\mathscr{K}} > 0,$$

for any compact $\mathscr{K} \subset \mathbb{R}^{md}$, where $\lambda_{\min}(\Sigma)$ denotes the minimal eigenvalue of a real symmetric matrix Σ.

(A4) $C_1 + C_2 E|\xi_1| < 1/m$.

Assumption (A1) is needed for identifiability of the estimation procedure. Assumptions (A1) and (A3) guarantee that the process $\{X_i\}$ does not die out whereas (A2) and (A4) are conditions for $\{X_i\}$ not to explode. The following lemma given by Ango Nze (1992) guarantees ergodicity of the process $\{X_i\}$. It is based on the application of the results of Nummelin and Tuominen (1982) and Tweedie (1975). Note that (A4) becomes redundant when both $f(x)$ and $\Sigma(x)$ are bounded, in which case $C_1 = C_2 = 0$.

Lemma 3.1. *Under the conditions* (A1)–(A4) *the Markov chain* $\{X_i\}$ *is geometrically ergodic, i.e. it is ergodic, with stationary probability measure* $\pi(\cdot)$ *such that, for almost every* x, *as* $k \to \infty$

$$\|P^k(\cdot \mid x) - \pi(\cdot)\|_{TV} = O(\rho^k)$$

for some $0 \leqslant \rho < 1$. *Here*

$$P^k(B \mid x) = P\{X_k \in B \mid X_m = x\}$$

for a Borel subset $B \subset \mathbb{R}^{md}$, *and* $\|\cdot\|_{TV}$ *is the total variation distance.*

Now we state the conditions necessary to derive joint asymptotic normality of $\hat{f}(x; v)$ and $\hat{\sigma}(x; v, s)$ at a fixed point $x \in \mathbb{R}^{md}$.

(A5) The functions f and Σ are componentwise twice continuously differentiable at the point $x \in \mathbb{R}^{md}$.

(A6) The density $\mu(\cdot)$ of the stationary distribution $\pi(\cdot)$ exists, is bounded, continuous and strictly positive in a neighborhood of the point x.

(A7) The kernel K is a compactly supported bounded nonnegative function on $\mathbb{R}^{md}$, such that

$$\int K(u)\,du = 1, \quad \int uK(u)\,du = 0, \quad \int uu^{\mathrm{T}}K(u)\,du = \sigma_K^2 I_{md},$$

where $\sigma_K^2 > 0$, and I_{md} denotes the identity matrix of dimension md.

J. Stat. Planning. Inference, 68, 221-245

(A8) $h_n = \beta n^{-1/(4+md)}$, where $\beta > 0$.
(A9) The initial value X_m is a fixed vector in $\mathbb{R}^{md}$.

Condition (A5) is a smoothness condition for the functions f and Σ. Note that it is related to (A8), the optimal speed of bandwidth. Condition (A8) guarantees a balance between bias and variance. A faster speed of h would lead to undersmoothing, a slower rate would increase the bias over the standard deviation of the estimator by oversmoothing. Both situations are undesirable since they result in less precise estimation. Condition (A6) is necessary to compute asymptotic bias and variance, (A7) is a typical assumption for kernels. Assumption (A9) supposes that the CHARN model is started at some fixed vector.

Let $f_j(x)$ and $\sigma_{jk}(x)$, $j,k = 1,2,\ldots,d$, be the components of the vector function $f(x)$ and the matrix function $\Sigma(x)$, respectively. Denote $\|K\|_2^2 = \int K^2(u)\,du$. Asymptotic normality results are presented in the following theorems.

Theorem 1. *Under the assumptions* (A1)–(A9)

$$n^{\frac{2}{4+md}}\{\hat{f}(x;v) - v^{\mathrm{T}} f(x)\} \xrightarrow{\mathscr{L}} \mathcal{N}\{b(x;v), V(x;v)\} \tag{3.3}$$

as $n \to \infty$ *with*

$$b(x;v) = \beta^2 \frac{\sigma_K^2}{2} \operatorname{Tr}[\nabla^2 (v^{\mathrm{T}} f(x))]$$

and

$$V(x;v) = \beta^{-md} \frac{v^{\mathrm{T}} \Sigma(x) v}{\mu(x)} \|K\|_2^2.$$

In particular, if one let v *be the* j*th or the* k*th coordinate vector of* $\mathbb{R}^d$*, one gets the following joint asymptotic distribution*:

$$n^{\frac{2}{4+md}} \begin{pmatrix} \hat{f}_j(x) - f_j(x) \\ \hat{f}_k(x) - f_k(x) \end{pmatrix} \xrightarrow{\mathscr{L}} \mathcal{N}\left\{ \begin{pmatrix} b_j(x) \\ b_k(x) \end{pmatrix}, \begin{pmatrix} V_j(x) & c_{jk}(x) \\ c_{jk}(x) & V_k(x) \end{pmatrix} \right\} \tag{3.4}$$

as $n \to \infty$ *with*

$$b_j(x) = \beta^2 \frac{\sigma_K^2}{2} [\operatorname{Tr}(\nabla^2 f_j(x))]$$

and

$$V_j(x) = \beta^{-md} \frac{\sigma_{jj}(x)}{\mu(x)} \|K\|_2^2, \qquad c_{jk}(x) = \beta^{-md} \frac{\sigma_{jk}(x)}{\mu(x)} \|K\|_2^2.$$

Denote

$$\operatorname{diag}(a) = \begin{bmatrix} a_1 & 0 & \cdots & 0 \\ 0 & a_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & a_d \end{bmatrix}$$

for any vector

$$a=\begin{pmatrix} a_1 \\ \vdots \\ a_d \end{pmatrix}\in\mathbb{R}^d.$$

Theorem 2. *Under the assumptions* (A1)–(A9)

$$n^{\frac{2}{4+md}}\{\hat{\sigma}(x;v,s)-v^{\mathrm{T}}\Sigma(x)s\}\xrightarrow{\mathscr{D}}\mathcal{N}\{b(x;v,s),V(x;v,s)\} \tag{3.5}$$

as $n\to\infty$ *with*

$$b(x;v,s)=\beta^2\frac{\sigma_K^2}{2}[\operatorname{Tr}\{\nabla^2 g(x)\}-\{s^{\mathrm{T}}f(x)\}\operatorname{Tr}\{\nabla^2 f^{\mathrm{T}}(x)v\}]$$
$$-\beta^2\frac{\sigma_K^2}{2}[\{v^{\mathrm{T}}f(x)\}\operatorname{Tr}\{\nabla^2 f^{\mathrm{T}}(x)s\}]$$

and

$$V(x;v,s)=\beta^{-md}\frac{\|K\|_2^2}{\mu(x)}[(m_4-2)T^*(x)+\{v^{\mathrm{T}}\Sigma(x)s\}^2]$$
$$+\beta^{-md}\frac{\|K\|_2^2}{\mu(x)}\{v^{\mathrm{T}}\Sigma(x)v\}\{s^{\mathrm{T}}\Sigma(x)s\},$$

where

$$g(x)=g(x;v,s)=\{v^{\mathrm{T}}f(x)\}\{s^{\mathrm{T}}f(x)\}+\{v^{\mathrm{T}}\Sigma(x)s\},$$
$$T^*(x)=T^*(x;v,s)=\operatorname{Tr}[\operatorname{diag}^2\{v^{\mathrm{T}}\Sigma^{1/2}(x)\}\operatorname{diag}^2\{\Sigma^{1/2}(x)s\}].$$

The covariance of $\hat{\sigma}(x;v,s)$ *and* $\hat{\sigma}(x;v',s')$ *is*

$$\beta^{-md}\frac{\|K\|_2^2}{\mu(x)}(m_4-2)$$
$$\times\operatorname{Tr}[\operatorname{diag}\{v^{\mathrm{T}}\Sigma^{1/2}(x)\}\operatorname{diag}\{\Sigma^{1/2}(x)s\}\operatorname{diag}\{v'^{T}\Sigma^{1/2}(x)\}\operatorname{diag}\{\Sigma^{1/2}(x)s'\}]$$
$$+\beta^{-md}\frac{\|K\|_2^2}{\mu(x)}[\{v^{\mathrm{T}}\Sigma(x)v'\}\{s^{\mathrm{T}}\Sigma(x)s'\}+\{v^{\mathrm{T}}\Sigma(x)s'\}\{s^{\mathrm{T}}\Sigma(x)v'\}].$$

In particular, if one let v *and* s *be the* j*th and* k*th coordinate vectors of* $\mathbb{R}^d$ *or the* j'*th and* k'*th coordinate vectors, one gets*

$$n^{\frac{2}{4+md}}\begin{pmatrix}\hat{\sigma}_{jk}(x)-\sigma_{jk}(x)\\ \hat{\sigma}_{j'k'}(x)-\sigma_{j'k'}(x)\end{pmatrix}\xrightarrow{\mathscr{D}}\mathcal{N}\left\{\begin{pmatrix}b_{jk}(x)\\ b_{j'k'}(x)\end{pmatrix},\begin{pmatrix}V_{jk}(x) & c_{jk,j'k'}(x)\\ c_{jk,j'k'}(x) & V_{j'k'}(x)\end{pmatrix}\right\} \tag{3.6}$$

as $n\to\infty$ *with*

$$b_{jk}(x)=\beta^2\frac{\sigma_K^2}{2}[\operatorname{Tr}\{\nabla^2\sigma_{jk}(x)+2\nabla^{\mathrm{T}}f_j(x)\nabla f_k(x)\}],\quad V_{jk}(x)=c_{jk,jk}(x),$$

where

$$c_{jk,j'k'}(x)=\beta^{-md}\frac{\|K\|_2^2}{\mu(x)}\{(m_4-2)T^*_{jk,j'k'}(x)+\sigma_{jj'}(x)\sigma_{kk'}(x)+\sigma_{jk'}(x)\sigma_{kj'}(x)\}$$

and

$$T^*_{jk,j'k'}(x)=\sum_{l=1}^{d}s_{jl}(x)s_{j'l}(x)s_{kl}(x)s_{k'l}(x)$$

in which $s_{jl}(x)$ *denotes the* (j,l)*th entry of the matrix* $\Sigma^{1/2}(x)$. *Finally, as* $n\to\infty$

$$n^{\frac{2}{4+md}}\begin{pmatrix}\hat\sigma_{jk}(x)-\sigma_{jk}(x)\\ \hat f_{j'}(x)-f_{j'}(x)\end{pmatrix}\xrightarrow{\mathscr{D}}\mathscr{N}\left\{\begin{pmatrix}b_{jk}(x)\\ b_{j'}(x)\end{pmatrix},\begin{pmatrix}V_{jk}(x)&0\\0&V_{j'}(x)\end{pmatrix}\right\}. \tag{3.7}$$

The practical use of these results lies in the possibility to check the form of the mean and volatility functions. For instance, at each point x we can construct a confidence interval for $\sigma_{jk}(x)$ based on plug-in estimates for $b_{jk}(x)$ and $V_{jk}(x)$. The bias conceivably can be estimated from a local cubic estimate. The variance can be estimated by first calculating the stochastic innovation term $\hat\xi_{ij}^2=\{Y_{ij}-\hat f_j(X_i)\}^2/\hat\sigma_{jj}(X_i)$ and then setting $\hat m_4=d^{-1}\sum_{j=1}^{d}n^{-1}\sum_{i=m}^{n}(\hat\xi_{ij}^2-1)^2$. The marginal density μ can be estimated as usual by a kernel estimator. Since the bias formula is slightly more involved than the variance formula, some undersmoothing might be recommended.

4. Application

The importance of the CHARN model for financial data has been pointed out in the introduction. In this section we come back to the introductory example of DEM/USD and DEM/GBP exchange rates. Figs. 6 and 7 show the estimated conditional mean functions $\hat f_1(x)$ and $\hat f_2(x)$ as functions of the lagged values $x_i=(y_{1,i-1},y_{2,i-1})^{\mathrm T}$. The surface and the contour plots all show that the mean functions are rather flat and are around zero. In fact, 80% of the $\hat f_1(x)$ values are in an interval around 0 whose length is only 0.11 times of the range of $y_{1,i}$, while 80% of the $\hat f_2(x)$ values are in an interval around 0 whose length is only 0.1557 times of the range of $y_{2,i}$. The pattern of the conditional covariance function $\hat\sigma_{12}(x)$ is different though, it changes from negative to positive as shown in Figs. 4 and 5.

Bollerslev et al. (1988, 1992), studied the *capital asset pricing model* (CAPM) by means of the multivariate GARCH model. To illustrate the connection between our vector CHARN model and their model, consider a random vector Y_t of excess asset returns with $E(Y_t\,|\,\mathscr{F}_{t-1})\equiv\mu_t$ and $\mathrm{Var}(Y_t\,|\,\mathscr{F}_{t-1})\equiv\Sigma_t$, where $\mathscr{F}_{t-1}$ is the information set generated by Y_{t-i}, $i=1,2,\ldots,$. If for nonnegative weight vector w_t whose elements add to 1, $w_t^{\mathrm T}Y_t$ is a mean–variance efficient portfolio, then the CAPM is

$$Y_t=\beta_t\mu_t^m+\varepsilon_t,$$

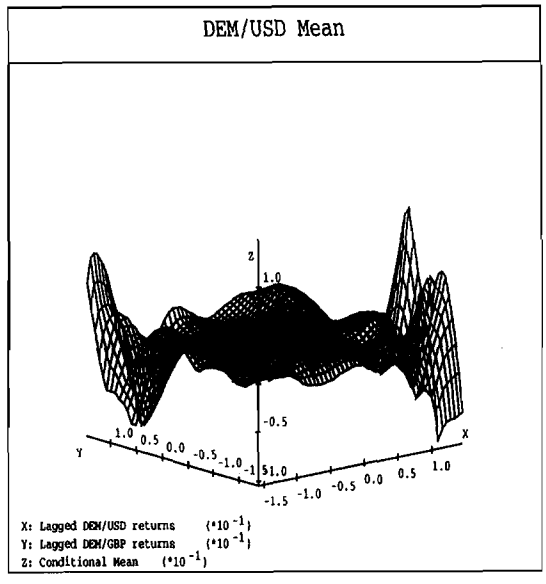


Fig. 6. The conditional mean function of the DEM/USD daily returns.

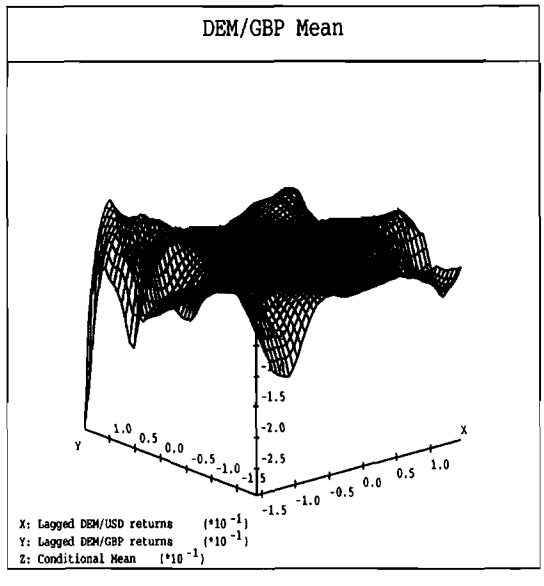


Fig. 7. The conditional mean function of the DEM/GBP daily returns.

where

$$\beta_t \equiv \Sigma_t w_t / w_t^{\mathrm{T}} \Sigma_t w_t,$$

with $E(\varepsilon_t \mid \mathscr{F}_{t-1}) \equiv 0$, $\mathrm{Var}(\varepsilon_t \mid \mathscr{F}_{t-1}) \equiv \Sigma_t$, and $\mu_t^m = w_t^{\mathrm{T}} \mu_t$. This is more general than ordinary CAPM which restricts Σ_t to be constant. While our CHARN model would

J. Stat. Planning. Inference, 68, 221-245

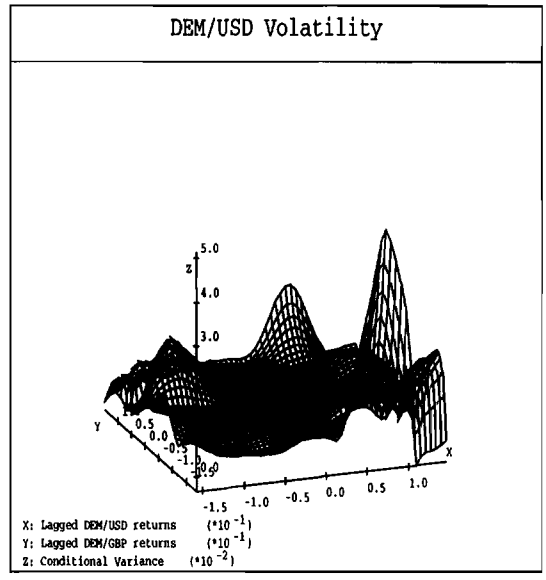


Fig. 8. The conditional variance function of the DEM/USD daily returns.

stipulate that Σ_t depends nonparametrically on a finite number of past observations, Bollerslev et al. (1988) used the multivariate GARCH model which allows Σ_t to depend on infinite number of past values, but only parametrically. A special form of the multivariate GARCH model is

$$\Sigma_t = \sigma Y_{t-1} Y_{t-1}^{\mathrm{T}}$$

for some constant $\sigma > 0$ in which case

$$\sigma_{12}(Y_{t-1}) = \sigma Y_{t-1,1} Y_{t-1,2}.$$

This is a hyperbolic function which exhibits the pattern visible in Figs. 4 and 5. For such a case, our CHARN model and the multivariate GARCH model would yield similar results.

Figs. 8 and 9 show the estimated conditional variance functions $\hat{\sigma}_{11}(x)$ and $\hat{\sigma}_{22}(x)$ as functions of the lagged values $x_i = (y_{1,i-1}, y_{2,i-1})^{\mathrm{T}}$. One can see that the variance function for the DEM/USD returns has a parabolic shape while that for DEM/GBP is roughly flat and positive.

5. Proofs

The proofs of Theorems 1 and 2 proceed in the following steps. First the normal equations of the LS problems (2.3) for the mean- and second-moment functions are solved. All estimators are split into a stochastic part and a systematic bias

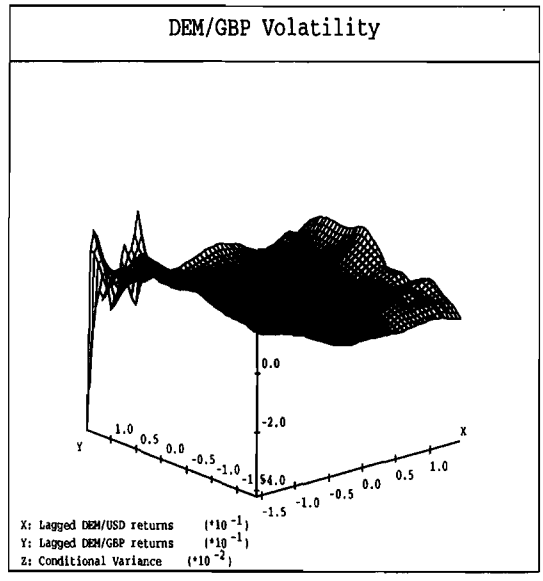


Fig. 9. The conditional variance function of the DEM/GBP daily returns.

part. Lemma 3.1 is essential in controling the stochastic part. Lemma 5.1 guarantees the strong mixing property of the recursive scheme (1.2). In combination with Lemmas 5.2–5.5 we then prove the joint asymptotic normality of the mean estimation as stated in Theorem 1 and that of volatility as stated in Theorem 2.

Set the matrices $W=\operatorname{diag}\{\frac{1}{n}K_h(X_i-x)\}_{i=m}^{n}$ and

$$Z=\begin{pmatrix} 1 & \cdots & 1 \\ \frac{X_m-x}{h} & \cdots & \frac{X_n-x}{h} \end{pmatrix}.$$

Define

$$v^{\mathrm{T}}Y=\begin{pmatrix} v^{\mathrm{T}}Y_m \\ \vdots \\ v^{\mathrm{T}}Y_n \end{pmatrix}=\begin{pmatrix} v^{\mathrm{T}}f(X_m)+v^{\mathrm{T}}\Sigma^{1/2}(X_m)\xi_m \\ \vdots \\ v^{\mathrm{T}}f(X_n)+v^{\mathrm{T}}\Sigma^{1/2}(X_n)\xi_n \end{pmatrix}$$

and also

$$v^{\mathrm{T}}YY^{\mathrm{T}}s=\begin{pmatrix} v^{\mathrm{T}}Y_mY_m^{\mathrm{T}}s \\ \vdots \\ v^{\mathrm{T}}Y_nY_n^{\mathrm{T}}s \end{pmatrix}$$

$$=\begin{pmatrix} (v^{\mathrm{T}}f(X_m)+v^{\mathrm{T}}\Sigma^{1/2}(X_m)\xi_m)(s^{\mathrm{T}}f(X_m)+s^{\mathrm{T}}\Sigma^{1/2}(X_m)\xi_m) \\ \vdots \\ (v^{\mathrm{T}}f(X_n)+v^{\mathrm{T}}\Sigma^{1/2}(X_n)\xi_n)(s^{\mathrm{T}}f(X_n)+s^{\mathrm{T}}\Sigma^{1/2}(X_n)\xi_n) \end{pmatrix}.$$

J. Stat. Planning. Inference, 68, 221-245

Then

$$\hat{f}(x;v) = F(0)^{\mathrm{T}}(ZWZ^{\mathrm{T}})^{-1}ZW[v^{\mathrm{T}}Y] \tag{5.1}$$

and

$$\hat{\sigma}(x;v,s) = F(0)^{\mathrm{T}}(ZWZ^{\mathrm{T}})^{-1}ZW[v^{\mathrm{T}}YY^{\mathrm{T}}s] - \hat{f}(x;v)\hat{f}(x;s) \tag{5.2}$$

by direct calculations.

First, to have the limit of $(ZWZ^{\mathrm{T}})^{-1}$, we need an auxiliary result based on Lemma 3.1.

Lemma 5.1 (Davydov, 1973). *A geometrically ergodic Markov chain whose initial variable is distributed with its stationary distribution is geometrically strongly mixing with the mixing coefficients satisfying* $\alpha(n) \leqslant c_0 \rho_0^n$ *for some* $0 < \rho_0 < 1$, $c_0 > 0$.

Having Lemma 5.1, the next lemma follows:

Lemma 5.2. *Under the conditions of Theorem* 1 *we have*

$$n^{-\frac{4}{4+md}} \sum_{i=m}^{n} \varphi_1(X_i)\,\varphi_2(u_{in})\,K(u_{in}) \xrightarrow{\mathrm{p}} \beta^{md}\mu(x)\,\varphi_1(x) \int \varphi_2(u)\,K(u)\,\mathrm{d}u, \tag{5.3}$$

$$n^{-\frac{4}{4+md}} \sum_{i=m}^{n} E\{\varphi_1(X_i)\,\varphi_2(u_{in})\,K(u_{in})\} \longrightarrow \beta^{md}\mu(x)\,\varphi_1(x) \int \varphi_2(u)\,K(u)\,\mathrm{d}u$$

as $n \to \infty$, *provided* $\varphi_1(\cdot)$ *is a bounded continuous function in a neighborhood of* x *and* $\varphi_2(\cdot)$ *is a bounded measurable function.*

Proof. See Härdle and Tsybakov (1996, Lemma 4.3). □

Lemma 5.3. *As* $n \to \infty$,

$$(ZWZ^{\mathrm{T}})^{-1} = \frac{1}{\mu(x)} \begin{bmatrix} 1 & 0_{1\times md} \\ 0_{md\times 1} & \sigma_K^{-2} I_{md} \end{bmatrix} \{1 + \mathrm{o_p}(1)\} \tag{5.4}$$

uniformly in a compact neighborhood of x.

Proof. The elements of ZWZ^{T} are all in the form of the left-hand side of (5.3). Using assumption (A7) and then taking matrix inverse, one gets (5.4). □

Now notice that, in view of Lemma 5.3

$$\begin{aligned} \hat{f}(x;v) - v^{\mathrm{T}}f(x) &= F(0)^{\mathrm{T}}(ZWZ^{\mathrm{T}})^{-1}ZW[v^{\mathrm{T}}Y] - v^{\mathrm{T}}f(x) \\ &= F(0)^{\mathrm{T}}(ZWZ^{\mathrm{T}})^{-1}ZW[v^{\mathrm{T}}Y] \\ &\quad - F(0)^{\mathrm{T}}(ZWZ^{\mathrm{T}})^{-1}(ZWZ^{\mathrm{T}}) \begin{bmatrix} v^{\mathrm{T}}f(x) \\ h\nabla(v^{\mathrm{T}}f(x)) \end{bmatrix} \end{aligned}$$

$$= F(0)^{\mathrm{T}}(ZWZ^{\mathrm{T}})^{-1}ZW\left[v^{\mathrm{T}}Y - Z^{\mathrm{T}}\begin{bmatrix} v^{\mathrm{T}}f(x) \\ h\nabla(v^{\mathrm{T}}f(x)) \end{bmatrix}\right]$$

$$= \frac{1}{\mu(x)n}\{1 + o_p(1)\}$$

$$\times \sum_{i=m}^{n} K_h(X_i - x)[v^{\mathrm{T}}f(X_i) - v^{\mathrm{T}}f(x) - (X_i - x)^{\mathrm{T}}\nabla\{v^{\mathrm{T}}f(x)\}]$$

$$+ \frac{1}{\mu(x)n}\{1 + o_p(1)\}\sum_{i=m}^{n} K_h(X_i - x)\{v^{\mathrm{T}}\Sigma^{1/2}(X_i)\xi_i\}. \qquad \square$$

(5.5)

To prove Theorem 1, one separates (5.5) into a bias part and a stochastic part as usual. The bias part is handled by the following lemma:

Lemma 5.4. *Let $g: \mathbb{R}^{md} \to \mathbb{R}^1$ be a twice continuously differentiable function. Then, under the assumptions of Theorem* 1

$$\frac{1}{\mu(x)n}\sum_{i=m}^{n} K_h(X_i - x)[g(X_i) - g(x) - (X_i - x)^{\mathrm{T}}\nabla g(x)]$$

$$= h^2\frac{\sigma_K^2}{2}\,\mathrm{Tr}[\nabla^2 g(x)] + o_p(h^2).$$

Proof. Using the Taylor expansion of $g(x)$, we get

$$\frac{1}{\mu(x)n}\sum_{i=m}^{n} K_h(X_i - x)[g(X_i) - g(x) - (X_i - x)^{\mathrm{T}}\nabla g(x)]$$

$$= \frac{1}{\mu(x)nh}\sum_{i=m}^{n} K(u_{in})[g(X_i) - g(x) - hu_{in}^{\mathrm{T}}\nabla g(x)]$$

$$= \frac{1}{2\mu(x)nh}\sum_{i=m}^{n} K(u_{in})[h^2 u_{in}^{\mathrm{T}}\nabla^2 g(x)u_{in}] + R,$$

where

$$|R| \leqslant \frac{1}{\mu(x)nh}\sum_{i=m}^{n} K(u_{in})\left[h^2\mathscr{D}^2 \sup_{|w|\leqslant\mathscr{D}} |\nabla^2 g(x + hw) - \nabla^2 g(x)|\right]$$

$$= \frac{o(h^2)}{nh}\sum_{i=m}^{n} K(u_{in}) = o_p(h^2) \qquad (5.6)$$

as $n \to \infty$, where $\mathscr{D} = \max\{|w|: w \in \mathrm{supp}\,K\}$ and the last equality in (5.6) is due to Lemma 5.2. Again, by Lemma 5.2 one has, as $n \to \infty$

$$\frac{1}{2\mu(x)nh}\sum_{i=m}^{n} K(u_{in})[u_{in}^{\mathrm{T}}\nabla^2 g(x)u_{in}] = \frac{1}{2\mu(x)nh}\sum_{i=m}^{n} \mathrm{Tr}[K(u_{in})u_{in}u_{in}^{\mathrm{T}}\nabla^2 g(x)]$$

$$\xrightarrow{\mathrm{p}} \frac{1}{2}\int \mathrm{Tr}[K(u)uu^{\mathrm{T}}\nabla^2 g(x)]\,du$$

$$= \frac{1}{2}\operatorname{Tr}\left[\int K(u)uu^{\mathrm{T}}\,du\nabla^2 g(x)\right]$$

$$= \frac{\sigma_K^2}{2}\operatorname{Tr}[\nabla^2 g(x)].$$

Combining this with (5.6) we get the lemma.

In particular, if $g(x) = v^{\mathrm{T}} f(x)$, one gets from Lemma 5.4

$$\frac{1}{\mu(x)n}\sum_{i=m}^{n} K_h(X_i - x)[v^{\mathrm{T}} f(X_i) - v^{\mathrm{T}} f(x) - (X_i - x)^{\mathrm{T}}\nabla\{v^{\mathrm{T}} f(x)\}]$$

$$= b(x; v)n^{-2/(4+md)} + o_p(n^{-2/(4+md)}) \tag{5.7}$$

as $n \to \infty$, where $b(x; v)$ is as given in Theorem 1. This yields the asymptotics of the bias term in (5.5).

To work out the asymptotics of the variance term, we need another lemma. Denote $\mathcal{F}_{k-1} = \sigma(X_k, X_{k-1}, \ldots, X_m)$ the σ-algebra generated by $X_m, \ldots, X_k$. □

Lemma 5.5 (Liptser and Shirjaev, 1980, Corollary 6). *Let m be a fixed integer and for every $n \geqslant m$, let the sequence $\eta^n = (\eta_{nk}, \mathcal{F}_k)$ be a square integrable martingale difference, i.e.*

$$E(\eta_{nk} \mid \mathcal{F}_{k-1}) = 0, \quad E(\eta_{nk}^2) < \infty, \quad m \leqslant k \leqslant n, \tag{5.8}$$

and let

$$\sum_{k=m}^{n} E(\eta_{nk}^2) = 1, \quad \forall n \geqslant n_0 \geqslant m. \tag{5.9}$$

The conditions

$$\sum_{k=m}^{n} E(\eta_{nk}^2 \mid \mathcal{F}_{k-1}) \xrightarrow{\mathrm{p}} 1, \quad \text{as } n \to \infty, \tag{5.10}$$

$$\sum_{k=m}^{n} E(\eta_{nk}^2 I(|\eta_{nk}| > \varepsilon) \mid \mathcal{F}_{k-1}) \xrightarrow{\mathrm{p}} 0, \quad \text{as } n \to \infty, \tag{5.11}$$

are sufficient for convergence

$$\sum_{k=m}^{n} \eta_{nk} \xrightarrow{\mathcal{D}} \mathcal{N}(0, 1), \quad \text{as } n \to \infty.$$

Proof of Theorem 1. Now we apply Lemma 5.5 to the following stochastic term of (5.5)

$$\sum_{i=m}^{n} \frac{1}{\mu(x)n} K_h(X_i - x)v^{\mathrm{T}}\Sigma^{1/2}(X_i)\xi_i \tag{5.12}$$

and observe that (in view of Lemma 5.2)

$$G_n = \sum_{i=m}^{n} E\left\{\left(\frac{1}{\mu(x)n}K_h(X_i - x)v^{\mathrm{T}}\Sigma^{1/2}(X_i)\xi_i\right)^2\right\} \tag{5.13}$$

$$= \frac{1}{\mu(x)nh^{md}}\{1 + o(1)\}\int K^2(u)v^{\mathrm{T}}\Sigma(x)v\,du$$

$$= n^{-4/(4+md)}V(x;v)\{1 + o(1)\}.$$

Define

$$\eta_{nk} = K_h(X_k - x)v^{\mathrm{T}}\Sigma^{1/2}(X_k)\xi_k\frac{1}{\mu(x)n\sqrt{G_n}}.$$

It is clear from (5.13) and (5.3) that (5.8)–(5.10) hold. It remains to check (5.11) in order to show that

$$\sum_{k=m}^{n}\eta_{nk} \xrightarrow{\mathscr{D}} \mathcal{N}(0,1), \quad \text{as } n \to \infty. \tag{5.14}$$

We have

$$\eta_{nk}^2 \leqslant Q_{nk}|\xi_k|^2, \tag{5.15}$$

where

$$Q_{nk} = \frac{h^{md}\beta^{-md}}{V(x;v)\mu^2(x)n}K_h(X_k - x)^2|v^{\mathrm{T}}\Sigma^{1/2}(X_k)|^2\{1 + o(1)\}.$$

Note that for some constant $C(x,v)$ depending only on x and v

$$Q_{nk} \leqslant C(x,v)\frac{1}{nh^{md}}K\left(\frac{X_k - x}{h}\right),$$

because of the fact that K is compactly supported and that Σ is bounded in a shrinking neighborhood of x. This entails

$$E\left\{|\xi_k|^2 I\left(|\xi_k|^2 \geqslant \frac{\varepsilon^2}{Q_{nk}}\right)\right\} \leqslant C_n(x,v),$$

where

$$C_n(x,v) = E\left\{|\xi_1|^2 I\left(|\xi_1|^2 \geqslant \frac{\varepsilon^2 nh^{md}}{C(x,v)\|K\|_\infty}\right)\right\} \to 0,$$

independent of k. This and (5.15) yield

$$\sum_{k=m}^{n} E\left\{\eta_{nk}^2 I(|\eta_{nk}| \geqslant \varepsilon)\,|\,\mathscr{F}_{k-1}\right\} = \sum_{k=m}^{n} E\left\{\eta_{nk}^2 I(\eta_{nk}^2 \geqslant \varepsilon^2)\,|\,\mathscr{F}_{k-1}\right\}$$

$$\leqslant \sum_{k=m}^{n} Q_{nk}E\left\{|\xi_k|^2 I\left(|\xi_k|^2 \geqslant \frac{\varepsilon^2}{Q_{nk}}\right)\right\}$$

$$\leqslant C_n(x,v)C(x,v)\sum_{k=m}^{n}\frac{1}{nh^{md}}K\left(\frac{X_k - x}{h}\right),$$

J. Stat. Planning. Inference, 68, 221-245

while

$$\sum_{k=m}^{n} \frac{1}{nh^{md}} K\left(\frac{X_k - x}{h}\right) \xrightarrow{\mathrm{p}} \mu(x), \quad \text{as } n \to \infty,$$

by Lemma 5.2. Thus we have proved (5.14). Now (3.3) is a consequence of (5.5), (5.7), (5.13) and (5.14). To prove the joint asymptotic normality (3.4), note that, in view of (5.5) and (5.7),

$$n^{\frac{2}{4+md}} \begin{pmatrix} \hat{f}_j(x) - f_j(x) \\ \hat{f}_k(x) - f_k(x) \end{pmatrix} = \begin{pmatrix} b_j(x) \\ b_k(x) \end{pmatrix} \{1 + \mathrm{o_p}(1)\} + n^{\frac{2}{4+md}} \begin{pmatrix} \zeta_{jn} \\ \zeta_{kn} \end{pmatrix} \{1 + \mathrm{o_p}(1)\}$$

as $n \to \infty$, where

$$\zeta_{jn} = \frac{1}{\mu(x)n} \sum_{i=m}^{n} K_h(X_i - x) v_j^{\mathrm{T}} \Sigma^{1/2}(X_i) \xi_i$$

and v_j is the jth coordinate vector in $\mathbb{R}^d$.

By the Cramér–Wold device, the joint asymptotic normality of ζ_{jn} and ζ_{kn} is proved if one shows that linear combinations of these random variables satisfy

$$n^{\frac{2}{4+md}} (\alpha_j \zeta_{jn} + \alpha_k \zeta_{kn}) \xrightarrow{\mathscr{D}} \mathscr{N}(0, \alpha_j^2 V_j(x) + \alpha_k^2 V_k(x) + 2\alpha_j \alpha_k c_{jk}(x)) \tag{5.16}$$

as $n \to \infty$, $\forall \alpha_j, \alpha_k \in \mathbb{R}^1$.

The proof of (5.16) is quite similar to that of (5.14), and it is based again on the application of Lemma 5.5. The difference is that instead of G_n, one should use now

$$G_n' = \sum_{i=m}^{n} E\left\{ \left(\frac{1}{\mu(x)n} K_h(X_i - x)(\alpha_j v_j + \alpha_k v_k)^{\mathrm{T}} \Sigma^{1/2}(X_i) \xi_i \right)^2 \right\}$$

$$= n^{-4/(4+md)} [\alpha_j^2 V_j(x) + \alpha_k^2 V_k(x) + 2\alpha_j \alpha_k c_{jk}(x)] \{1 + \mathrm{o}(1)\},$$

where the last equality follows from Lemma 5.2 (cf. (5.13)). □

Proof of Theorem 2. Similar to (5.5), we write

$$\begin{aligned} \hat{\sigma}(x; v, s) - \sigma(x; v, s) &= (v^{\mathrm{T}} f(x))(s^{\mathrm{T}} f(x)) - \hat{f}(x; v)\hat{f}(x; s) \\ &\quad + F(0)^{\mathrm{T}} (ZWZ^{\mathrm{T}})^{-1} ZW[v^{\mathrm{T}} Y Y^{\mathrm{T}} s] \\ &\quad - (v^{\mathrm{T}} f(x))(s^{\mathrm{T}} f(x)) - v^{\mathrm{T}} \Sigma(x) s \\ &= (v^{\mathrm{T}} f(x))(s^{\mathrm{T}} f(x)) - \hat{f}(x; v)\hat{f}(x; s) \\ &\quad + F(0)^{\mathrm{T}} (ZWZ^{\mathrm{T}})^{-1} ZW \Bigg[v^{\mathrm{T}} Y Y^{\mathrm{T}} s \\ &\quad - Z^{\mathrm{T}} \begin{pmatrix} (v^{\mathrm{T}} f(x))(s^{\mathrm{T}} f(x)) + v^{\mathrm{T}} \Sigma(x) s \\ h\nabla((v^{\mathrm{T}} f(x))(s^{\mathrm{T}} f(x)) + v^{\mathrm{T}} \Sigma(x) s) \end{pmatrix} \Bigg] \end{aligned}$$

$$
\begin{aligned}
&= (v^{\mathrm{T}} f(x))(s^{\mathrm{T}} f(x)) - \hat{f}(x;v)\hat{f}(x;s) \\
&\quad + \frac{1}{\mu(x)n}\{1 + o_p(1)\} \sum_{i=m}^{n} K_h(X_i - x)[v^{\mathrm{T}} f(X_i) f(X_i)^{\mathrm{T}} s \\
&\quad - v^{\mathrm{T}} f(x) f(x)^{\mathrm{T}} s - (X_i - x)^{\mathrm{T}} \nabla\{(v^{\mathrm{T}} f(x))(s^{\mathrm{T}} f(x))\}] \\
&\quad + \frac{1}{\mu(x)n}\{1 + o_p(1)\} \sum_{i=m}^{n} K_h(X_i - x)[v^{\mathrm{T}} \Sigma(X_i) s - v^{\mathrm{T}} \Sigma(x) s \\
&\quad - (X_i - x)^{\mathrm{T}} \nabla\{v^{\mathrm{T}} \Sigma(x) s\}] \\
&\quad + \frac{1}{\mu(x)n}\{1 + o_p(1)\} \sum_{i=m}^{n} K_h(X_i - x) v^{\mathrm{T}} \Sigma^{1/2}(X_i) \\
&\quad \times (\xi_i \xi_i^{\mathrm{T}} - I_d) \Sigma^{1/2}(X_i) s \\
&\quad + \frac{1}{\mu(x)n}\{1 + o_p(1)\} \sum_{i=m}^{n} K_h(X_i - x) \\
&\quad \times \{s^{\mathrm{T}} f(X_i) v^{\mathrm{T}} + v^{\mathrm{T}} f(X_i) s^{\mathrm{T}}\} \Sigma^{1/2}(X_i) \xi_i,
\end{aligned}
$$

which after plugging in the formula for $\hat{f}(x;v) - v^{\mathrm{T}} f(x)$ and $\hat{f}(x;s) - s^{\mathrm{T}} f(x)$ (cf. (5.5)) yields

$$
\hat{\sigma}(x;v,s) - \sigma(x;v,s) = \{1 + o_p(1)\} \sum_{j=1}^{8} T_j, \tag{5.17}
$$

where

$$
\begin{aligned}
T_1 &= \frac{1}{\mu(x)n} \sum_{i=m}^{n} K_h(X_i - x)[v^{\mathrm{T}} f(X_i) f(X_i)^{\mathrm{T}} s - v^{\mathrm{T}} f(x) f(x)^{\mathrm{T}} s - (X_i - x)^{\mathrm{T}} \\
&\quad \times \nabla\{(v^{\mathrm{T}} f(x))(s^{\mathrm{T}} f(x))\}], \\
T_2 &= \frac{1}{\mu(x)n} \sum_{i=m}^{n} K_h(X_i - x)[v^{\mathrm{T}} \Sigma(X_i) s - v^{\mathrm{T}} \Sigma(x) s - (X_i - x)^{\mathrm{T}} \nabla\{v^{\mathrm{T}} \Sigma(x) s\}], \\
T_3 &= -\frac{1}{\mu(x)n} \sum_{i=m}^{n} K_h(X_i - x) s^{\mathrm{T}} f(x)[v^{\mathrm{T}} f(X_i) - v^{\mathrm{T}} f(x) - (X_i - x)^{\mathrm{T}} \nabla\{v^{\mathrm{T}} f(x)\}], \\
T_4 &= -\frac{1}{\mu(x)n} \sum_{i=m}^{n} K_h(X_i - x) v^{\mathrm{T}} f(x)[s^{\mathrm{T}} f(X_i) - s^{\mathrm{T}} f(x) - (X_i - x)^{\mathrm{T}} \nabla\{s^{\mathrm{T}} f(x)\}], \\
T_5 &= \frac{1}{\mu(x)n} \sum_{i=m}^{n} K_h(X_i - x) v^{\mathrm{T}} \Sigma^{1/2}(X_i)(\xi_i \xi_i^{\mathrm{T}} - I_d) \Sigma^{1/2}(X_i) s, \\
T_6 &= \frac{1}{\mu(x)n} \sum_{i=m}^{n} K_h(X_i - x)\{s^{\mathrm{T}} f(X_i) - s^{\mathrm{T}} f(x)\}\{v^{\mathrm{T}} \Sigma^{1/2}(X_i) \xi_i\}, \\
T_7 &= \frac{1}{\mu(x)n} \sum_{i=m}^{n} K_h(X_i - x)\{v^{\mathrm{T}} f(X_i) - v^{\mathrm{T}} f(x)\}\{s^{\mathrm{T}} \Sigma^{1/2}(X_i) \xi_i\},
\end{aligned}
$$

J. Stat. Planning. Inference, 68, 221-245

$$T_8 = -[\hat{f}(x;v) - v^{\mathrm{T}} f(x)][\hat{f}(x;s) - s^{\mathrm{T}} f(x)]$$
$$= -n^{-4/(4+md)}[n^{2/(4+md)}(\hat{f}(x;v) - v^{\mathrm{T}} f(x))][n^{2/(4+md)}(\hat{f}(x;s) - s^{\mathrm{T}} f(x))]. \tag{5.18}$$

Using Lemma 5.4, one derives

$$T_1 = \beta^2 \frac{\sigma_K^2}{2}[\mathrm{Tr}\{\nabla^2((v^{\mathrm{T}} f(x))(s^{\mathrm{T}} f(x)))\}]n^{-2/(4+md)} + o_p(n^{-2/(4+md)}),$$

$$T_2 = \beta^2 \frac{\sigma_K^2}{2}[\mathrm{Tr}\{\nabla^2(v^{\mathrm{T}} \Sigma(x) s)\}]n^{-2/(4+md)} + o_p(n^{-2/(4+md)}),$$

$$T_3 = -\beta^2 \frac{\sigma_K^2}{2}\{s^{\mathrm{T}} f(x)\}\,\mathrm{Tr}\{\nabla^2 f^{\mathrm{T}}(x) v\}n^{-2/(4+md)} + o_p(n^{-2/(4+md)}),$$

$$T_4 = -\beta^2 \frac{\sigma_K^2}{2}\{v^{\mathrm{T}} f(x)\}\,\mathrm{Tr}\{\nabla^2 f^{\mathrm{T}}(x) s\}n^{-2/(4+md)} + o_p(n^{-2/(4+md)}), \tag{5.19}$$

and thus

$$T_1 + T_2 + T_3 + T_4 = b(x;v,s)n^{-2/(4+md)} + o_p(n^{-2/(4+md)}). \tag{5.20}$$

Now we calculate T_6. Note that

$$|K_h(z-x)(s^{\mathrm{T}} f(z) - s^{\mathrm{T}} f(x))|$$
$$\leqslant K_h(z-x) \sup_{|w| \leqslant \mathscr{D}} |f(x+wh) - f(x)| \leqslant ChK_h(z-x),$$

since K is compactly supported (here $C>0$ is a constant). Thus,

$$E(T_6^2) = \frac{1}{\mu(x)^2 n^2} \sum_{i=m}^{n} E[K_h^2(X_i - x)\{s^{\mathrm{T}} f(X_i) - s^{\mathrm{T}} f(x)\}^2\{v^{\mathrm{T}} \Sigma(X_i) v\}]$$
$$\leqslant \frac{C^2 h^2}{\mu(x)^2 n^2} \sum_{i=m}^{n} E[K_h^2(X_i - x) v^{\mathrm{T}} \Sigma(X_i) v].$$

By Lemma 5.2

$$\frac{h^2 \beta^{md}}{nh^{md}} \sum_{i=m}^{n} E[K_h^2(X_i - x) v^{\mathrm{T}} \Sigma(X_i) v]$$
$$\to \beta^{md} \mu(x) v^{\mathrm{T}} \Sigma(x) v \int K^2(u)\,\mathrm{d}u = \mathrm{O}(1), \quad \text{as } n \to \infty,$$

and therefore

$$E(T_6^2) = \mathrm{O}\left(\frac{h^2}{nh^{md}}\right) = o(n^{-4/(4+md)}), \quad \text{as } n \to \infty.$$

The evaluation of T_7 is quite analogous and, hence, we get

$$T_6 + T_7 = o_p(n^{-2/(4+md)}), \quad \text{as } n \to \infty. \tag{5.21}$$

Also, in view of Theorem 1,

$$n^{2/(4+md)}T_8 \xrightarrow{\mathscr{D}} 0, \quad \text{as } n \to \infty. \tag{5.22}$$

The relations (5.20)–(5.22) show that the sum $\sum_{j=1}^{4} T_j$ in (5.17) yields the correct asymptotic bias, while the terms T_6, T_7, and T_8 are asymptotically negligible. It remains to show the asymptotic normality of the term T_5:

$$n^{2/(4+md)}T_5 \xrightarrow{\mathscr{D}} \mathcal{N}(0, V(x;v,s)), \quad \text{as } n \to \infty.$$

Again, to prove this, we use Lemma 5.5. We leave out the verification of the conditions (5.10) and (5.11) of Lemma 5.5, since it is done as in the proof of Theorem 1. We only deduce the asymptotic expression for the variance of T_5, which is given, analogous to G_n of the proof of Theorem 1, by the asymptotics of

$$G_n'' = \frac{1}{\mu(x)^2 n^2} \sum_{i=m}^{n} E[(K_h(X_i - x)v^{\mathrm{T}}\Sigma^{1/2}(X_i)(\xi_i\xi_i^{\mathrm{T}} - I_d)\Sigma^{1/2}(X_i)s)^2]. \tag{5.23}$$

To study this expression, use the following lemma.

Lemma 5.6. . *Let* $a=(a_1,\dots,a_d)^{\mathrm{T}}$, $\tilde{a}=(\tilde{a}_1,\dots,\tilde{a}_d)^{\mathrm{T}}$, $b=(b_1,\dots,b_d)^{\mathrm{T}}$, *and* $\tilde{b}=(\tilde{b}_1,\dots,\tilde{b}_d)^{\mathrm{T}}$ *be vectors in* $\mathbb{R}^d$. *Then under* (A1),

$$E[(a^{\mathrm{T}}(\xi_1\xi_1^{\mathrm{T}} - I_d)b)(\tilde{a}^{\mathrm{T}}(\xi_1\xi_1^{\mathrm{T}} - I_d)\tilde{b})]$$
$$= (m_4 - 2)\,\mathrm{Tr}[\,\mathrm{diag}(a)\,\mathrm{diag}(b)\,\mathrm{diag}(\tilde{a})\,\mathrm{diag}(\tilde{b})] + (a^{\mathrm{T}}\tilde{a})(b^{\mathrm{T}}\tilde{b}) + (a^{\mathrm{T}}\tilde{b})(\tilde{a}^{\mathrm{T}}b).$$

Proof. Denoting by δ_{jk} the Kronecker delta and using (A1), we get

$$E[(a^{\mathrm{T}}(\xi_1\xi_1^{\mathrm{T}} - I_d)b)(\tilde{a}^{\mathrm{T}}(\xi_1\xi_1^{\mathrm{T}} - I_d)\tilde{b})]$$
$$= E\left[\sum_{k,j=1}^{d} a_j(\xi_{1j}\xi_{1k} - \delta_{jk})b_k \sum_{l,m=1}^{d} \tilde{a}_l(\xi_{1l}\xi_{1m} - \delta_{lm})\tilde{b}_m\right]$$
$$= E\left[\sum_{j=1}^{d} a_j(\xi_{1j}^2 - 1)b_j \sum_{k=1}^{d} \tilde{a}_k(\xi_{1k}^2 - 1)\tilde{b}_k\right]$$
$$+ E\left[\sum_{1\leqslant j<k\leqslant d} (a_jb_k + a_kb_j)\xi_{1j}\xi_{1k} \sum_{1\leqslant l<m\leqslant d} (\tilde{a}_l\tilde{b}_m + \tilde{a}_m\tilde{b}_l)\xi_{1l}\xi_{1m}\right]$$
$$= m_4 \sum_{j=1}^{d} a_jb_j\tilde{a}_j\tilde{b}_j + \sum_{1\leqslant j<k\leqslant d} (a_jb_k + a_kb_j)(\tilde{a}_j\tilde{b}_k + \tilde{a}_k\tilde{b}_j)$$
$$= (m_4 - 2)\sum_{j=1}^{d} a_jb_j\tilde{a}_j\tilde{b}_j + \frac{1}{2}\sum_{j,k=1}^{d} (a_jb_k + a_kb_j)(\tilde{a}_j\tilde{b}_k + \tilde{a}_k\tilde{b}_j),$$

which yields the lemma.

Applying Lemma 5.6 with $a = \Sigma^{1/2}(X_i)v$ and $b = \Sigma^{1/2}(X_i)s$, we find

$$E[(v^{\mathrm{T}}\Sigma^{1/2}(X_i)(\xi_i\xi_i^{\mathrm{T}} - I_d)\Sigma^{1/2}(X_i)s)^2 \mid X_i]$$

$$= (m_4 - 2)T^*(X_i) + \{v^{\mathrm{T}}\Sigma(X_i)s\}^2 + \{v^{\mathrm{T}}\Sigma(X_i)v\}\{s^{\mathrm{T}}\Sigma(X_i)s\}.$$

This and (5.23) yield

$$G_n'' = \frac{1}{\mu(x)^2 n^2}\sum_{i=m}^{n} E[K_h^2(X_i - x)(m_4 - 2)T^*(X_i)]$$

$$+ \frac{1}{\mu(x)^2 n^2}\sum_{i=m}^{n} E[K_h^2(X_i - x)(\{v^{\mathrm{T}}\Sigma(X_i)s\}^2 + \{v^{\mathrm{T}}\Sigma(X_i)v\}\{s^{\mathrm{T}}\Sigma(X_i)s\})]$$

and, in view of Lemma 5.2,

$$n^{4/(4+md)}G_n''$$

$$= \beta^{-md}\frac{\|K\|_2^2}{\mu(x)}[(m_4 - 2)T^*(x) + \{v^{\mathrm{T}}\Sigma(x)s\}^2 + \{v^{\mathrm{T}}\Sigma(x)v\}\{s^{\mathrm{T}}\Sigma(x)s\}]$$

$$\times(1 + o(1))$$

$$= V(x; v, s) + o(1), \quad \text{as } n \to \infty,$$

which is the expression for asymptotic variance given in Theorem 2. □

To show the joint asymptotic normality (3.6) and (3.7) one proceeds as in the proof of Theorem 1, by using the Cramér–Wold device and checking the conditions of Lemma 5.5. The calculations of covariance terms in (3.6) are based on Lemma 5.6 as well.

Acknowledgements

We would like to thank Christian Hafner, Helmut Lütkepohl, and Rolf Tschernig for helpful discussions. We also thank our referees for pointing out several technical errors to us. This research was supported by Sonderforschungsbereich 373 'Quantifikation und Simulation Ökonomischer Prozesse' Deutsche Forschungsgemeinschaft.

References

Ango Nze, P., 1992. Critères d'ergodicité de quelques modèles à représentation markovienne. C.R. Acad. Sci. Paris Sér. I 315, 1301–1304.

Bollerslev, T., Chou, R., Kroner, K., 1992. ARCH modeling in finance: a review of the theory and empirical evidence. J. Econometrics 52, 5–59.

Bollerslev, T., Engle, R., Wooldridge, J., 1988. A capital asset pricing model with time-varying covariances. J. Political Economy 96, 116–131.

Bossaerts, P., Härdle, W., Hafner, C., 1996. Foreign exchange-rates have surprising volatility. In: Robinson, P.M. (Ed.), Athens Conf. on Applied Probability and Time Series, vol. 2, Lecture Notes in Statistics, vol. 115, pp. 55–72. Springer, Berlin.

Chan, K.S., Tong, H., 1985. On the use of deterministic Lyapunov functions for the ergodicity of stochastic difference equations. Adv. Appl. Probab. 17, 666–678.

Chan, K.S., Tong, H., 1986. On estimating thresholds in autoregressive models. J. Time Ser. Anal. 7, 179–190.

Chen, R., Tsay, R.S., 1993a. Nonlinear additive ARX models. J. Amer. Statist. Assoc. 88, 955–967.

Chen, R., Tsay, R.S., 1993b. Functional-coefficient autoregressive models. J. Amer. Statist. Assoc. 88, 298–308.

Cleveland, W.S., 1979. Robust locally weighted regression and smoothing scatterplots. J. Amer. Statist. Assoc. 74, 829–836.

Collomb, 1984. Propriétés de convergence presque complète du prédicteur à noyau. Z. Wahrscheinlichkeitstheorie verwandte Gebiete 66, 441–460.

Davydov, Yu.A., 1973. Mixing conditions for Markov chains. Theory Probab. Appl. 18, 312–328.

Diebolt, J., Guégan, D., 1993. Tail behaviour of the stationary density of general nonlinear autoregressive processes of order one. J. Appl. Probab. 30, 315–329.

Diebold, F., Nason, J., 1990. Nonparametric exchange rate prediction. J. Internat. Econom. 28, 315–332.

Doukhan, P., Ghindès, M., 1980. Estimation dans le processus $X_{n+1} = f(X_n) + \varepsilon_{n+1}$. C.R. Acad. Sci. Paris Sér. A 297, 61–64.

Doukhan, P., Ghindès, M., 1981. Processus autorégressifs non-linéaires. C.R. Acad. Sci. Paris Sér. A 290, 921–923.

Drost, F.C., Nijman, T.E., 1993. Temporal aggregation of GARCH processes. Econometrica 61, 909–927.

Engle, R.F., 1982. Autoregressive conditional heteroscedasticity with estimates of the variance of UK inflation. Econometrica 50, 987–1008.

Engle, R.F., Gonzalez-Rivera, G., 1991. Semiparametric ARCH models. J. Bus. Econom. Statist. 9, 345–360.

Engle, R.F., Ng, V., 1993. Measuring and testing the impact of news on volatility. J. Finance 48, 1749–1778.

Fan, J., Gijbels, I., 1996. Local Polynomial Modelling and its Applications. Chapman & Hall, London.

Gouriéroux, Ch., Monfort, A., 1992. Qualitative threshold ARCH models. J. Econometrics 52, 159–199.

Granger, C., Teräsvirta, T., 1993. Modelling Nonlinear Dynamic Relationships. Oxford University Press, Oxford.

Gregory, A.W., 1989. A nonparametric test for autoregressive conditional heteroscedasticity: a Markov chain approach. J. Bus. Econom. Statist. 7, 107–115.

Guillaume, D.M., Dacorogna, M.M., Davé, R.R., Müller, U.A., Olsen, R.B., Pictet, O.V., 1994. From the bird's eye to the microscope: a survey of new stylized facts of the intra-daily foreign exchange market, Olsen Associates Working Paper.

Haggan, V., Ozaki, T., 1981. Modelling nonlinear vibrations using an amplitude-dependent autoregressive time series model. Biometrika 68, 189–196.

Härdle, W., Klinke, S., Turlach, B., 1995. XploRe – an interactive statistical computing environment. Springer, Heidelberg.

Härdle, W., Tsybakov, A.B., 1996. Local polynomial estimators of the volatility function in nonparametric autoregression. J. Econometrics, to appear.

Katkovnik, V.Ya., 1979. Linear and nonlinear methods of nonparametric regression analysis. Automatika 35–46.

Katkovnik, V.Ya., 1985. Nonparametric Identification and Data Smoothing. Nauka, Moscow (in Russian).

Liptser, R.Sh., Shirjaev, A.N., 1980. A functional central limit theorem for martingales. Theory Probab. Appl. 25, 667–688.

Lütkepohl, H., 1991. Introduction to Multiple Time Series Analysis. Springer, Heidelberg.

McKeague, I.W., Zhang, M.J., 1994. Identification of nonlinear time series from first order cumulative characteristics. Ann. Statist. 22, 495–514.

Meese, R.A., Rogoff, K., 1983. Empirical exchange rate models of the seventies, do they fit out of sample? J. Internat. Econom. 14, 3–24.

Meese, R.A., Rose, A., 1991. An empirical assessment of non-linearities in models of exchange rate determination. Rev. Econom. Studies 58, 601–619.

Mokkadem, A., 1987. Sur un modèle autorégressif nonlinéaire. Ergodicité et ergodicité géometrique. J. Time Ser. Anal. 8, 195–204.

Nummelin, E., Tuominen, P., 1982. Geometric ergodicity of Harris-recurrent Markov chains with application to renewal theory. Stochastic Process. Appl. 12, 187–202.

Robinson, P.M., 1983. Nonparametric Estimators for Time Series. J. Time Ser. Anal. 4, 185–207.

Robinson, P.M., 1984. Robust nonparametric autoregression. In: Franke, Härdle, Martin (Eds.), Robust and Nonlinear Time Series Analysis, Lecture Notes in Statistics, vol. 26. Springer, Heidelberg.

Ruppert, D., Wand, M.P., 1994. Multivariate locally weighted least squares regression. Ann. Stat. 22, 1346–1370.

Stone, C.J., 1977. Consistent nonparametric regression. Ann. Statist. 5, 595–645.

Teräsvirta, T., 1994. Specification, estimation and evaluation of smooth transition autoregressive models. J. Amer. Statist. Assoc. 89, 208–218.

Tjøstheim, D., 1994. Non-linear time series analysis: a selective review. Scand. J. Statist. 21, 97–130.

Tong, H., 1978. On a threshold model. In: Chen, C.H. (Ed.), Pattern Recognition and Signal Processing. Sijthoff & Noordholf, Alphen a/d Rijn, The Netherlands.

Tong, H., 1983. Threshold Models in Nonlinear Time Series Analysis, Lecture Notes in Statistics, vol. 21. Springer, Heidelberg.

Tsybakov, A.B., 1986. Robust reconstruction of functions by the local approximation method. Prob. Inform. Transm. 22, 133–146.

Tweedie, R.L., 1975. Sufficient conditions for ergodicity and recurrence of Markov chains on a general state space. Stochastic Process. Appl. 3, 385–403.

Vieu, P., 1994. Order choice in nonlinear autoregressive models. Statistics 24, 1–22.

Wand, M.P., Jones, M.C., 1995. Kernel Smoothing. Chapman & Hall, London.

JOURNAL OF APPLIED ECONOMETRICS
J. Appl. Econ., **13**, 525–541 (1998)

SEMIPARAMETRIC ANALYSIS OF GERMAN EAST–WEST MIGRATION INTENTIONS: FACTS AND THEORY

MICHAEL C. BURDA,[a] WOLFGANG HÄRDLE,[b] MARLENE MÜLLER[b]* AND AXEL WERWATZ[b]

[a]*Institut für Wirtschaftstheorie II, Wirtschaftswissenschaftliche Fakultät Humboldt–Universität zu Berlin, Germany*
[b]*Institut für Statistik und Ökonometrie, Wirtschaftswissenschaftliche Fakultät Humboldt–Universität zu Berlin, Germany*

SUMMARY

East–West migration in Germany peaked at the beginning of the 1990s although the average wage gap between Eastern and Western Germany continues to average about 25%. We analyse the propensity to migrate using microdata from the German Socioeconomic Panel. Fitting a parametric Generalized Linear Model (GLM) yields non-linear residual behavior. This finding is not compatible with classical Marshallian theory of migration and motivates the semiparametric analysis. We estimate a Generalized Partial Linear Model (GPLM) where some components of the index of explanatory variables enter non-parametrically. We find the estimate of the non-parametric influence in concordance with a number of alternative migration theories, including the recently proposed option-value-of-waiting theory. © 1998 John Wiley & Sons, Ltd.

1. INTRODUCTION

German East–West migration has been the subject of several recent papers. Using microdata from the German Socio-Economic Panel, Burda (1993), Büchel and Schwarze (1994) and Schwarze (1996) have investigated this issue empirically. Especially interesting is the fact that, although migration peaked in the early 1990s following unification, the gap between average Eastern and Western wages remains about 25% as of 1997.

We take the empirical findings of Burda (1993) as our point of departure. We re-analyse the data by estimating a Generalized Linear Model (GLM) but find that the GLM does not provide a satisfactory fit. Estimating a semiparametric Generalized Partial Linear Model (GPLM) reveals a non-linear, non-monotonic influence of household income on the propensity to migrate from East to West. This non-linear influence of income, while difficult to reconcile with classical economic theory of migration, is compatible with a number of alternative models of the migration decision including the option value approach proposed by Dixit and Pindyck (1994) and applied recently to the migration decision by Burda (1995) and O'Connell (1997). It is also consistent with unobserved heterogeneity and misspecification of the estimation equation.

In the following section we present a brief discussion of the classical (Marshallian) theory of migration behaviour. In Section 3 we introduce the data and discuss how facts and theory play together. Results from fitting a parametric GLM to the data are presented in Section 4. As we shall see, standard logit analysis does not sufficiently capture the phenomenon underlying the observations. We therefore turn to a more flexible setting by allowing some components to

* Correspondence to: Marlene Müller, Institut für Statistik und Ökonometrie, Humboldt Univeristät, Spandauer Str. 1, D-10178 Berlin, Germany. E-mail: marlene @wiwi-hu.berlin.de

Contract grant sponsor: Deutsche Forschungsgemeinschaft.

CCC 0883–7252/98/050525–17$17·50

Received 15 October 1997
Revised 27 April 1998

Journal Applied Econometrics, 13, 525-541

take a non-parametric form. These semiparametric Generalized Partial Linear Models (GPLM) are described and estimated in Section 5. In Section 6 we discuss our findings and speculate on theoretical explanations for our results. Section 7 concludes the paper.

2. SOME THEORETICAL CONSIDERATIONS

Since Ravenstein's pathbreaking work on the determinants of migration more than a century ago, income has been the focus of economists' attempts to explain spatial mobility. More precisely, the difference between income at home (W^E) and the attainable income upon migration (W^W) has been singled out as the key explanatory variable (Sjaastad, 1962). Some migration is an investment, a forward-looking agent will care not only about the current income differential but also about future income differentials. That is, he will consider the net expected present value of future additional income earned if he decides to migrate.

Yet even if this expected present value is positive, an agent may not migrate if the fixed costs of migrating are sufficiently high. Such fixed costs include pecuniary components associated with physically moving a household from one place to another. In addition, moving away means leaving behind a familiar environment as well as friends and family members. Following classical ('Marshallian') economic theory, we may therefore say that a rational, forward-looking agent will migrate if the expected present value of the income stream from migrating exceeds monetary valuation of the associated fixed costs, or if the expected net present value from migrating (net of fixed costs) is positive. Incorporating risk aversion will change the trigger rule, but at most by a constant amount which would depend on the relative riskiness of the options and individual preferences.

Under a number of weak assumptions about the stochastic process generating relative income, the expected present value of future gains from migration will be a function of the current observed income differential, and for plausible assumptions this relationship will be linear. To consider an extreme but simple example, if the absolute per-period income differential $\Omega_t = W_t^W - W_t^E$ follows an arithmetic Brownian process with negative drift ν, then the expected present value of migration in time $t = 0$ is given by $V^m = (\Omega_0 - \nu/\delta)/\delta$, where δ denotes the discount rate.

Let the fixed costs of migration (including monetary equivalent of utility loss from moving) be given by F and denote the migration decision by the binary variable Y ($Y = 1 \rightarrow$ migration). Then the decision rule for a rational agent can be formally written as:

$$Y = \begin{cases} 1 & \text{if } V^m = \frac{1}{\delta}(\Omega_0 - \nu/\delta) - F > 0 \\ 0 & \text{otherwise} \end{cases} \tag{1}$$

This theory delivers the clear prediction that an increase in period t income by reducing Ω_t will decrease migration propensity, holding alternatives available in the West constant.

3. THE DATA

In the empirical analysis we use data drawn from the German Socio-Economic Panel (GSOEP). The GSOEP is a representative survey of German households that was extended to the former East in 1990. We use 3367 observations from the GSOEP's second East German wave which was collected in the spring of 1991 (time $t = 0$). All calculations were carried out with the statistical computing environment XploRe (1998).

Journal Applied Econometrics, 13, 525-541

Because very few actual migrants were observed in this wave of the GSOEP, we use migration propensity ('intention') as the dependent variable Y.

At the outset, it is important to state that this variable—as is the case with all intentions variables—is somewhat problematic for a number of reasons. (For an extensive discussion of this problem as well as a plea for not disregarding such information, see Manski, 1990.) First, agents may be simply irrational and have little idea of what their future behaviour may be or of the probability distribution of future events conditioning future decisions. Second, even if agents are rational in the sense that they can forecast their own future decision-making process and have rational expectations of future forcing variables, future decisions (realizations) may be correlated across individuals due to systematic intervening shocks. In this paper we simply take the position that 'intentions' are a monotonic function of the underlying driving variables which motivate migration.

The theoretical discussion of the previous section has focused on the income differential between host region and home region and the fixed cost of migrating as the key explanatory variables. Yet measuring both quantities poses a challenge. Regarding the income differential, we are faced with the problem that the potential income in the West is not observable. Hence, some imputation is generally necessary. Since Germany shares the same institutions and language one could assume that upon migration eastern Germans are able to employ at least some component of their human capital, earning 'western returns' for their attributes, at least up to a (macroeconomic) constant. A natural approach to estimate W_0^W would be to imply estimates of a traditional earnings equation of the Mincer type, which attributes observed wages to either market 'returns' multiplied by observable measures of human capital endowment (education, experience, training, tenure) or to attributes unobservable to the econometrician modeled as a random disturbance. Estimating this relation on a sample of Westerners, however, will most likely produce biased estimates of returns to Easterners (Burda and Schmidt, 1997). Moreover, it is unclear how to use these estimates to calculate an imputed Western wage for those Easterners who are registered as unemployed or out of the labour force. Rather than producing spurious findings based on biased estimates of the West–East income differential (Dunn, Kreyenfeld and Lovely, 1997), we decide to include income in the East only. We shall return to this point when discussing our results in Section 6.

The GSOEP data provides a multitude of variables that arguably are related to the intention to migrate from the East to West. Starting from a set of roughly 30 potential explanatory variables considered in the empirical analysis of Burda (1993) we used economic intuition and statistical selection criteria to limit the number of explanatory variables. This was done merely for better exposition of the facts. The proposed statistical method is valid for any dimension of the vector of explanatory variables.

Summary statistics for Y and the explanatory variables are given in Table I. Presence of a partner, home ownership and increasing age are expected to increase the fixed cost of migrating whereas relatives or friends in the West supposedly have the opposite effect. Age will also influence the migration decision via the discount rate. The variable *environmental satisfaction* is measured on a scale from 1 ('very unhappy with environmental conditions') to 10 ('very happy') and can therefore be expected to have a negative influence on migration propensity. The sign of the coefficients of the gender, city size and education variables is rather unclear apriori.

We have separated *age* and *household income* from the remaining explanatory variables in the table as—for the purposes of this study—they can be regarded as *continuous* explanatory variables.

(1998) Burda, M., Härdle, W., Müller, M. and Werwatz, A.
Semiparametric Analysis of German East-West Migration Intentions: Facts and Theory.

Journal Applied Econometrics, 13, 525-541

Table I. Summary statistics

		Mean	S.D.	Expected effect
Y	*migration intention*	0·394	0·489	
X_1	*female*	0·511	0·500	
X_2	*partner*	0·854	0·353	−
X_3	*owner*	0·322	0·467	−
X_4	*family/friends in west*	0·855	0·352	+
X_5	*unemployed/jobloss certain*	0·196	0·397	+
X_6	*environmental satisfaction*	3·9	2·4	−
X_7	*city size < 10,000*	0·522	0·499	
X_8	*city size 10–100,000*	0·342	0·474	
X_9	*university degree*	0·085	0·278	
X_{10}	*age* min. 18, max. 65	39·4	12·8	−
X_{11}	*household income* min. 200, max. 4000	2189·5	754·7	

4. PARAMETRIC ESTIMATION RESULTS

Collect the explanatory variables described in the previous section into the vector x. The goal of the empirical analysis is to estimate the probability of migration intention, i.e. $E(Y \mid x) = \text{Prob}(Y = 1 \mid x)$. A natural starting point for estimating this probability is fitting a parametric GLM. More precisely, we estimated a logit model.

This parametric model is based on two assumptions. First, the underlying latent variable Y is a sum of a linear index of the explanatory variables x and an individual error term u. Second, the cumulative distribution function (cdf) of u conditional on x is the logistic distribution function. Combining both assumptions gives

$$E(Y \mid x) = \text{Prob}(Y = 1 \mid x) = \{1 + \exp(-x^T\beta)\}^{-1} \tag{2}$$

As usual, $G(u) = \{1 + \exp(-u)\}^{-1}$ is called the (inverse) link function.

Table II gives the Maximum Likelihood logit estimates of β. Most coefficients have the expected sign: age, a partner, home ownership and environmental satisfaction reduce migration propensity whereas family or friends in the West and poor labour market prospects in the East have the opposite effect.

The estimated coefficient of the linear logit specification suggests that migration propensity significantly increases with household income. Figure 1 reflects the actual dependence of the response Y on the variables age and income.We have plotted each variable versus the logits $\log(\hat{p}/1 - \hat{p})$ where $\hat{p}$ are the relative frequencies for $Y = 1$ (migration intention). Essentially, these logits are obtained from classes of neighboured realizations (where the range of either age or income has been divided into 50 equidistant intervals). In case that $\hat{p}$ was 0 or 1, several classes were merged. Thicker bullets correspond to move observations in a class. Figure 1 shows that age has an almost linear influence on migration intention, whereas the relationship between income and migration intention exhibits a U-shaped curve.

If we include the square of household income as an additional regressor then both income coefficients are individually insignificant. This finding may lead an analyst to conclude that income does not have a non-linear influence. Yet, if we add income cubed as a regressor to the

(1998) Burda, M., Härdle, W., Müller, M. and Werwatz, A.
Semiparametric Analysis of German East-West Migration Intentions: Facts and Theory.

Journal Applied Econometrics, 13, 525-541

Table II. Logit estimates

Dependent variable: migration intention		
Variable	Estim. coeff.	*t*-ratio
constant	1·864	7·74
female	−0·233	−3·03
partner	−0·325	−2·87
owner	−0·576	5·79
family/friends in west	0·647	5·61
unemployed	0·217	2·24
environmental satisfaction	−0·057	−3·52
city size < 10,000	−0·718	−5·69
city size 10–100,000	−0·347	−2·91
university degree	0·481	3·56
age	−0·050	−14·89
household income	0·0001202	2·22
Sample size: 3367, log likelihood: −1992·7		

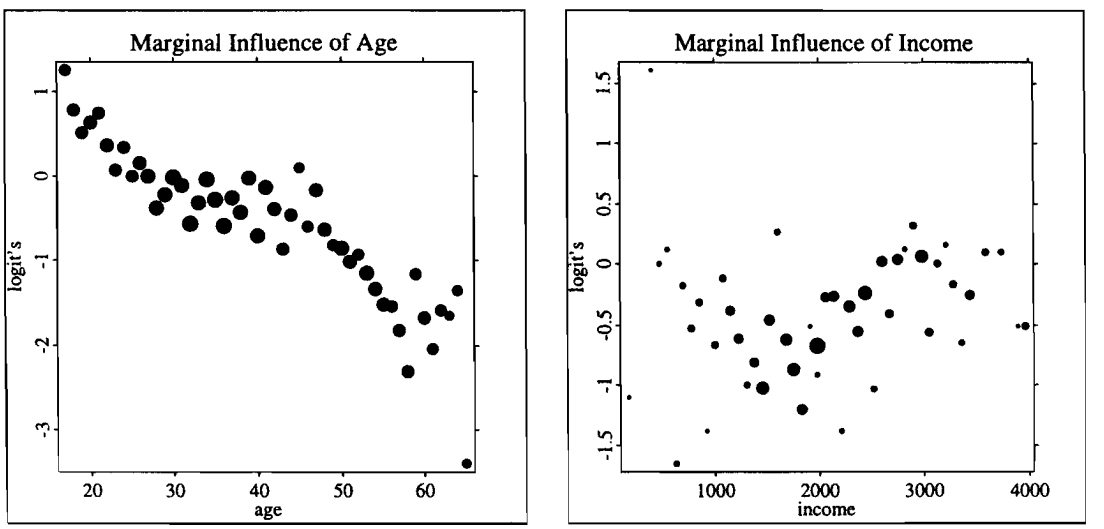


Figure 1. Marginal influence of age (left) and income (right) on migration intention, visualized by logits on classes

model that already includes income and income squared then all three income coefficients are individually as well as jointly significant. These findings are summarized in Table III.

Rather than continuing with the refinement of this parametric specification we decided to estimate a semiparametric Generalized Partial Linear Model which allows the data to freely determine the shape of the influence of income on migration propensity. By means of generalized additive modelling (Hastie and Tibshirani, 1990) this can be extended to the variable age as well. An analysis of this model yielded a linear dependence of migration propensity on age (as in Figure 1). We therefore included only income as a possible non-linear candidate.

(1998) Burda, M., Härdle, W., Müller, M. and Werwatz, A.
Semiparametric Analysis of German East-West Migration Intentions: Facts and Theory.

Journal Applied Econometrics, 13, 525-541

Table III. Parametric specification search

Variable	Estim. coeff.	*t*-ratio
'Quadratic' model		
household income	−0·0001288	−0·507
household income^2	5·46e-08	1·002
'Cubic' model		
household income	−0·0016491	−2·130
household income^2	8·08e-07	2·206
household income^3	−1·12e-10	−2·080
Dependent variable: migration intention Same regressors as above besides income		

5. SEMIPARAMETRIC ESTIMATION RESULTS

Before turning to estimates, we will briefly introduce the generalized partially linear model (GPLM). As before, the GPLM assumes that the mean of Y is related to an index of explanatory variables via the known link function G. Contrary to the logit model of the previous section the index of explanatory variables is composed of a linear parametric component and a non-parametric component. That is, the GPLM assumes that

$$E(Y \mid x, t) = G\{x^T\beta + m(t)\} \tag{3}$$

where—in a slight abuse of notation—we have collected the explanatory variables that enter the argument of $G(\cdot)$ linearly in the $p \times 1$ vector x, and those that enter non-linearly in the $q \times 1$ vector t. The unknown quantities that need to be estimated are the parameter vector β and the unknown function $m(\cdot)$. Note that there is no intercept parameter since it can be absorbed into the non-parametric part $m(t)$. In the empirical analysis x will—with the exception of age—be made up of discrete (categorical) variables while t contains solely household income.

The estimation methods for model (3) are based on the idea that an estimate $\hat{\beta}$ can be found for known $m(\cdot)$, and an estimate $\hat{m}(\cdot)$ can be found for known β. In what follows we will concentrate on *profile likelihood* estimation which goes back to Severini and Wong (1992) and Severini and Staniswalis (1994). Denote by $L(\mu, y)$ the individual log-likelihood, where $\mu = E(Y \mid x, t) = G\{x^T\beta + m(t)\}$. The profile likelihood uses two different likelihood functions for the estimation of the parametric and semiparametric components. The usual likelihood for n i.i.d. observations (x_i, t_i, y_i)

$$\mathcal{L}(\beta) = \sum_{i=1}^{n} L\{\beta^T x_i + m_\beta(t_i); y_i\} \tag{4}$$

is used to obtain $\hat{\beta}$ and a 'smoothed' likelihood

$$\mathcal{L}_h(\eta) = \sum_{i=1}^{n} K_h(t - t_i)L(\beta^T x_i + \eta; y_i) \tag{5}$$

(1998) Burda, M., Härdle, W., Müller, M. and Werwatz, A.
Semiparametric Analysis of German East-West Migration Intentions: Facts and Theory.

Journal Applied Econometrics, 13, 525-541

Table IV. GPLM estimates

Dependent variable: *migration intention*				
	GPLM estimates		Logit estimates	
Variable	Coeff.	t-ratio	Coeff.	t-ratio
female	−0·238	−3·09	−0·233	−3·03
partner	−0·282	−2·44	−0·325	−2·87
owner	−0·569	−5·71	−0·576	−5·79
family/friends in west	0·640	5·54	0·647	5·61
unemployed	0·216	2·23	0·217	2·24
environmental satisfaction	0·056	−3·47	−0·057	−3·52
city size < 10,000	−0·689	−5·43	−0·718	−5·69
city size 10–100,000	−0·323	−2·71	−0·347	−2·91
university degree	0·471	3·48	0·481	3·56
age	−0·050	−14·89	−0·050	−14·89
Sample size: 3367, log likelihood: −1989·8, GPLM bandwidth: 0·3				

for the non-parametric smooth function $\hat{m}_\beta(t) = \eta$ at point t and $K_h(u) = h^{-1}K(u/h)$ a kernel function with bandwidth h (Severini and Staniswalis, 1994) belongs to an exponential family using the

The computational algorithm consists of searching maxima of both likelihoods simultaneously. A detailed description of the algorithm can be found in the Appendix. It turns out that the resulting estimator $\hat{\beta}$ is $\sqrt{n}$-consistent and asymptotically normal, and that estimators $\hat{m} = \hat{m}_{\hat{\beta}}$ are consistent in supremum norm (see Severini and Staniswalis, 1994).

Table IV gives the GPLM estimates of β in a model that includes the same explanatory variables as the logit fit of Table II. The logit estimates and their t-ratios are also reported to conveniently compare results across the different approaches. In general, the GPLM estimates are very close to their logit counterparts. In terms of the GPLM, income plays the role of the variable t in equation (3). The estimated influence of income is depicted in Figure 2, with income on the horizontal axis and the estimate of $m(t)$ on the vertical axis. The highly non-linear estimate of $m(t)$ strongly contrasts with the linear influence of income implied by the logit model which we have also included in Figure 2.

The GPLM fit suggests an S-shaped effect of income, or a U-shaped influence over the range of income values that carry most of the mass of the income distribution. The bandwidth h underlying the estimate of $m(t)$ was set equal to 30% of the range of household income. The U-shaped estimate is obtained for a range of values of h, though. Note that the decreasing part of $\hat{m}(t)$ above $t = 3000$ may be attributed to random fluctuations for this bandwidth size. Above this income level, we have only a small number of observations (see Figure 1).

The visual impression of Figure 2 suggests that the estimate of $m(t)$ significantly deviates from the estimated linear influence of the parametric GLM fit. We use a test procedure to formally test that $m(t)$ is a linear function:

$$H_0 : m(t) = \alpha t + \alpha_0$$

$$H_1 : m(t) \text{ is an arbitrary smooth function}$$

(1998) Burda, M., Härdle, W., Müller, M. and Werwatz, A.
Semiparametric Analysis of German East-West Migration Intentions: Facts and Theory.

Journal Applied Econometrics, 13, 525-541

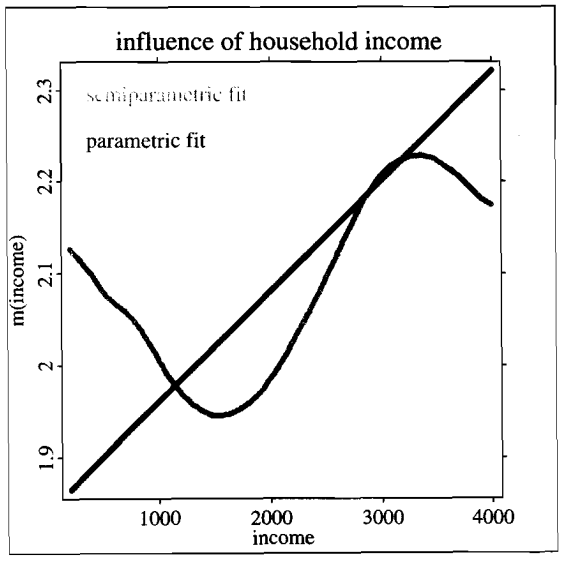


Figure 2. Influence of the net household income on migration propensity

This test is based on comparing the semiparametric estimates with the parametric estimates

$$(\tilde{\beta}, \tilde{\alpha}, \tilde{\alpha}_0) = \arg\min_{\beta,\alpha,\alpha_0} \sum_{i=1}^{n} L[G\{x_i^T\beta + \alpha t_i + \alpha_0\}; y_i] \tag{6}$$

where α denotes the coefficient of income and α_0 the constant in the parametric fit.

A test of the hypothesis GLM (logit model) against the alternative of a GPLM may be based on the likelihood ratio statistic. Denote by $\tilde{\mu}_i = G(x_i^T\tilde{\beta}t + \tilde{\alpha}t + \tilde{\alpha}_0)$ the parametric GLM fit and by $\hat{\mu}_i = G\{x_i^T\hat{\beta} + \hat{m}(t)\}$ the GPLM fit. Hastie and Tibshirani (1990) propose using

$$R = 2\sum_{i=1}^{n}\{L(\hat{\mu}_i, y_i) - L(\tilde{\mu}_i, y_i)\} \tag{7}$$

which has heuristically a distribution that is similar to a χ^2 distribution. However, the degrees of freedom for the GPLM need to be replaced by an approximate value and theoretic distribution of R is unknown.

Härdle, Mammen and Müller (1996) propose a modification of the test statistic R. This modification is based on the fact that a direct comparison of $\hat{m}(t)$ and $\tilde{\alpha}t + \tilde{\alpha}_0$ can be misleading because $\hat{m}$ has a non-negligible smoothing bias. this holds even under the linearity hypothesis. Hence, a bias-corrected parametric estimate $\bar{m}(t)$ is used instead of $\tilde{\alpha}t + \tilde{\alpha}_0$.

Using this bias-corrected $\bar{m}(t)$ the following modified likelihood ratio test statistic is computed:

$$R^M = 2\sum_{i=1}^{n}\{L(\hat{\mu}_i, \hat{\mu}_i) - L(\bar{\mu}_i, \hat{\mu}_i)\} \tag{8}$$

(1998) Burda, M., Härdle, W., Müller, M. and Werwatz, A.
Semiparametric Analysis of German East-West Migration Intentions: Facts and Theory.

Journal Applied Econometrics, 13, 525-541

where $\tilde{\mu}_i = G\{x_i^T\tilde{\beta} + \tilde{m}(t_i)\}$ is the bias-corrected GLM fit and $\hat{\mu}_i$ the GPLM fit as before. Härdle *et al.* (1996) show asymptotic normality of R^M. The proof of this result is based on showing that the asymptotic expansion of R^M behaves approximately like a sum of $O(h)$ independent summands. This is typically not very large and indeed simulations show that the normal approximation need not work well for R^M (Müller, 1997). Therefore, for the calculations of quantiles, it is recommended to use the following bootstrap procedure:

(1) Generate samples $\{Y_1^*, \ldots, Y_n^*\}$ under the parametric hypothesis with $E^*(Y_i^*) = G(x_i^T\tilde{\beta} + \tilde{\alpha}t_i)$. Here E^* denotes the conditional expectation given $(x_1, t_1, \ldots, x_n, t_n)$.
(2) Calculate estimates $\hat{\beta}^*$, $\hat{m}^*$, $\tilde{\beta}^*$, $\tilde{\alpha}^*$, $\tilde{m}^*$ based on the bootstrap samples $\{(x_1, t_1, Y_1^*), \ldots, (x_n, t_n, Y_n^*)\}$. Furthermore, calculate test the statistic R^{M^*}. Repeat this n^* times. The quantiles of the distribution of R^M can be estimated by the quantiles of the conditional distribution of R^{M^*}.

Since in our case the distribution of Y is completely specified by $EY = \mu = G(x^T\beta + \alpha t + \alpha_0)$ (under the hypothesis of linearity) we resample from the Bernoulli distribution with parameters $\tilde{\mu}_i = G(x_i^T\tilde{\beta} + \tilde{\alpha}t_i + \tilde{\alpha}_0)$ (the parametric GLM fit).

Table V shows the result of both test procedures for the GLM versus the GPLM. With R^M we denote the test using test statistic (8), where the rest has been carried out using the normal approximation. R^{M^*} bootstrap denotes the results for the bootstrapped quantiles of R^M. Since an optimal bandwidth choice for the GPLM is not known, all tests were performed for a sequence of bandwidths. However, we can recognize a clear rejection of the linearity hypothesis across all bandwidths for the R and the bootstrapped R^{M^*}. The normal approximation for R^M works poorly for higher bandwidth levels, as indicated above.

6. INTERPRETING THE RESULTS: ALTERNATIVE EXPLANATIONS

In the previous section we found a significant non-linear relationship between migration intensity and household income which appears non-monotonic. That is, for certain intervals the migration propensity is increasing in household income. This is at variance with the linear relationship implied by the classical theory of migration outlined in Section 2. In this section we will briefly outline theoretic models of migration that may give rise to non-linearities in income and/or non-monotonic relationships and which therefore could aid in the interpretation of the shape of the estimate presented in Figure 2.

6.1 Option Value of Migration

One limiting aspect of the Marshallian theory of migration of Section 2 is its 'all-or-nothing' aspect; either migration occurs now or never. The work of Dixit and Pindyck (1994) and others

Table V. Observed significance level for linearity test for migration data, $n = 3367$

h	0·1	0·2	0·25	0·3	0·4
R	0·028	0·021	0·019	0·017	0·016
R^M	0·035	0·069	0·130	0·269	0·602
R^{M^*}	0·015	0·005	0·005	0·005	0·010

Note: 200 bootstrap replications. Bandwidth h in % of range of household income

(1998) Burda, M., Härdle, W., Müller, M. and Werwatz, A.
Semiparametric Analysis of German East-West Migration Intentions: Facts and Theory.

Journal Applied Econometrics, 13, 525-541

has shown that postponement of the decision without forsaking it can be a valuable option under a large class of irrevocable investment problems. future. In terms of equation (1), migrating today not only means incurring a fixed cost F and forgoing the current and future income in the sending region. It also means forgoing the opportunity to postpone migration on the basis of new, currently unanticipated information. This opportunity has positive (expected) value today because waiting brings more information about the future, which may evolve against migration in an unexpected way. Assuming no loss of opportunity is implied, postponement leaves open the possibility of migrating at a later date, saving the fixed cost over the interval.

This opportunity cost of migrating today — in addition to the expected present value of future income gains from migration net of migration costs — is referred to as the *option value of waiting* and we will denote it as V^o. V^o is equal to what one is willing to pay for the option to postpone the migration decision rather than having to decide 'now or never'. It can be calculated as the difference between the expected net present value from postponing migration, V^p, and the expected net present value from migrating today, V^m. V^o — which is a function of current household income, among other things — can be derived as the solution to a dynamic programming problem under a variety of assumptions (see Dixit and Pindyck, 1994).

Figure 3 graphs V^o (kinked curve in the lower panel), V^p (the positively sloped curve in the upper panel) and V^m (the dashed straight line in the upper panel) as functions of the current income differential. If the current wage differential is below MT (the 'marshallian trigger') immediate migrating does not have positive net value ($V^m < 0$). Hence V^o is just equal to V^p.

If the current wage differential is between MT and OT ('option-value trigger') then immediate migration has positive expected value and hence $V^o = V^p - V^m$. We have displayed the values V^o as vertical bars in the upper panel for selected values of the current wage differential. If the current wage differential is above OT then V^o is zero: the current wage differential is so large that any further postponement of migration has zero value.

It appears from Figure 3 that V^o has the opposite shape as the estimate relationship of the previous section. But V^o is the option value of *postponing* migration. That is, high values of V^o imply a low propensity to migrate and vice versa. This is clearly evident if we rewrite the 'classical' decision rule (1) to incorporate the option value of waiting:

$$Y = \begin{cases} 1 & \text{if } \frac{1}{\delta}(\Omega_0 + \nu/\delta) - F - V^o(\Omega_0) > 0 \\ 0 & \text{otherwise} \end{cases} \tag{9}$$

As a result, the option value theory applied to the migration decision requires mirroring V^o around the x-axis producing a U-shaped relationship. In principle, this shape is consistent with the empirical findings of Figure 2.

Arguably, a superior strategy is to estimate the option value of migration directly, as has been done recently by a number of researchers in other applications. For example, Pakes (1986) estimates the option value of patents in this spirit, Rust (1987) has used similar methods to estimate deep parameters of a dichotomous investment problem (optimal replacement of bus engines), and Rothwell and Rust (1995) have examined optimal response of the nuclear industry to regulatory changes. (A special issue of the *Journal of Applied Econometrics* in 1995 highlighted the considerable breadth of potential applications in this area.) In principle, this approach would be possible but difficult to implement for two reasons. First, explicit modelling of the dynamic programming problem would require more detailed information on the individual's characteristics than are available in the GSOEP (e.g. household wealth), although Rust (1987) has shown

Journal Applied Econometrics, 13, 525-541

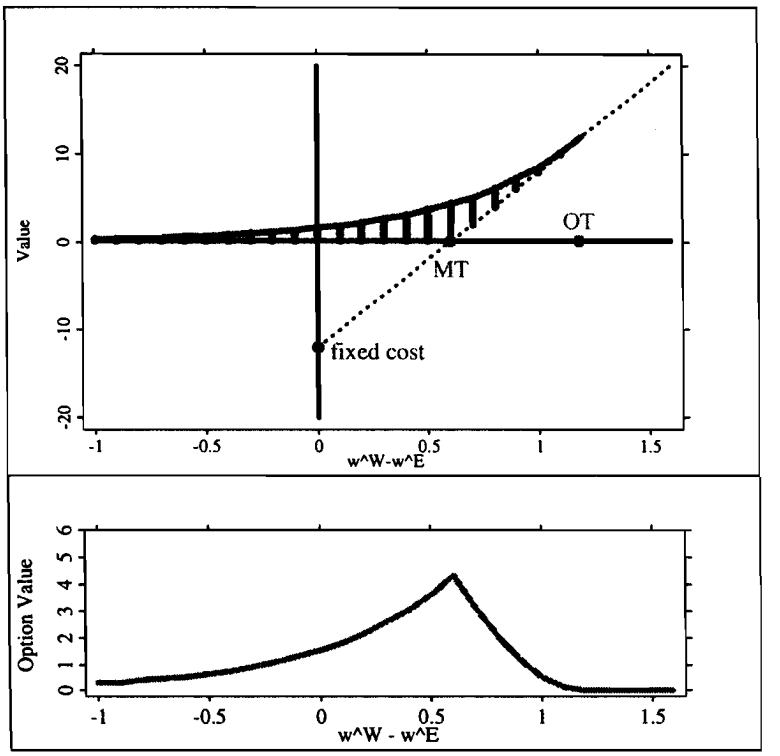


Figure 3. The option value of waiting

that problems with unobservable state variables can be overcome with sufficient assumptions concerning the structure of costs, etc., in a state-space model. Second, it would be necessary to impute a statistical process to wages at home and in the East, which would be highly speculative, given the limited data observations currently available since unification. Finally and perhaps most importantly, a number of plausible competing models exist (see below) which would give rise to confounding effects for individual decision makers. The derivation of a model nesting classical, option-value and other models of migration is beyond the scope of this exploratory paper, and is left for future research.

6.2 Risk-aversion, Income Effects and the Demand for Immobility

The previous discussion assumed risk neutrality and the absence of preferences for living at home or abroad. In fact, both risk aversion and the 'demand for immobility' might affect the propensity to migrate. The latter hypothesis has been put forward and investigated by, among others, Faini and Venturini (1993, 1994). Under the assumption that current place of residence is a normal 'good', the income effect of higher absolute wages at home implies a lower propensity to migrate. Alternatively, wealthier individuals might seek to escape impoverishment or reduce dependence on relatives by moving to the wealthier West, where public goods infrastructure is better and better-paying job opportunities are more plentiful. In the end, the effect of income is an empirical proposition and will depend on preferences of individual agents, but theory predicts that, given

(1998) Burda, M., Härdle, W., Müller, M. and Werwatz, A.
Semiparametric Analysis of German East-West Migration Intentions: Facts and Theory.

Journal Applied Econometrics, 13, 525-541

opportunities, wealthier families should exhibit a greater demand for immobility, and that the demand for immobility would play a subservient role for poorer to middle-income families. Indeed, this prediction is consistent with the shape of Figure 2 in for household income exceeding 1500 DM/month.

In the case of risk aversion, the influence of uncertainty is also ambiguous. In general, curvature in the utility function (as opposed to strict linearity in previous sections) will attenuate the attractiveness of migration if uncertainty impinges primarily on income abroad. An exception is Stark (1989) who shows that in some cases migration may serve a function of risk diversification or reduction. Below we show an example of how introducing curvature in the utility function (risk aversion, decreasing marginal utility) could affect the valuation of the migration decision without considering any option value. This line of reasoning is therefore also consistent with either a negative or a positive effect of absolute home income on migration propensities. The underlying presumption in the current application is, of course, that income at home is riskier, so that normal patterns of risk aversion imply demand for migration which is increasing in household income.

6.3 Borrowing Constraints and Liquidity Effects

In addition to aspects of preferences addressed in the previous section, it seems likely that capital markets are imperfect. Realistically, poorer segments of the population are likely to be liquidity-strapped and therefore unable to finance the migration investment, even if it has positive expected present value. Suppose that a component of moving costs, F, must be paid in cash, and thus cannot be financed out of future earnings in the host country. In such a situation, the absolute value of current income (and not relative to abroad) matters for some range—as long as assets are inadequate to finance the move. When the wage rises, some households which may have been willing to migrate for some time can do so, financing the move out of current income. This reasoning predicts a positive effect of home wage/income on migration propensity for some range of current income. To the extent that the probability of being faced with credit constraints depends negatively on income and wealth, borrowing constraints seem a good candidate explanation for the negative branch found in the household income range 0–1500 DM/month.

6.4 Potential Misspecification

One important potential explanation of our results is related to misspecification of the estimation equation, i.e. the arguments of the link function in equation (3). For example, one might raise the objection that migration models are based on income *differentials* while our empirical analysis employs income in the East only. In Figures 4 and 5 we try to clarify this point.

The top panel of Figure 4 is a repetition of the lower panel of Figure 3. The middle panel of Figure 4 plots a (hypothetical) Western income (vertical axis) versus Eastern income. The former was imputed using a Mincer regression on the Western half of the 1990–93 waves of the GSOEP. The lower straight line is the 45° reference line whereas the upper straight line corresponds to the Western. Now suppose that the option value of postponing migration is depending on the income differential. Then, in the situation indicated in the middle panel, the option value of postponing migration plotted as a function of the income in the East (lower panel) has the same shape as if it is plotted as a function of the income differential.

(1998) Burda, M., Härdle, W., Müller, M. and Werwatz, A.
Semiparametric Analysis of German East-West Migration Intentions: Facts and Theory.

Journal Applied Econometrics, 13, 525-541

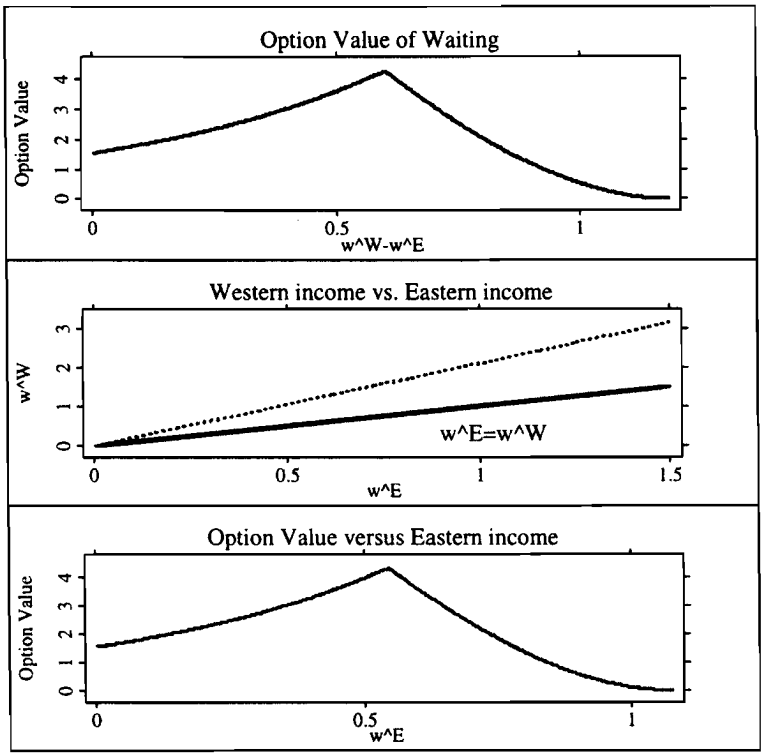


Figure 4. Eastern income versus Western income

Similarly, Figure 5 shows that different hypotheses about the relationship between Eastern and Western income still preserve the non-linearity of the option value — regardless whether it is plotted as a function of the income differential or income in the East. Specifically, the parabola in the middle panel of this figure reflects the hypothesis that Easterners with a low income (expect to) receive a relatively high Western income, those with a mid-range income receive a rather small increase in the West and individuals with a high Eastern income expect a relatively strong increase in income by moving to the West. Under this assumption about the relationship between income in the East and income in the West, and under the assumption that the option value of waiting depends on the current West–East income differential as depicted in the top panel of Figure 5, we obtain the non-linear relationship between the option value and income in the East as shown in the lower panel of Figure 5.

A more serious problem which potentially confounds all attempts to estimate a more tightly parameterized version of the model is unobserved heterogeneity. Individual-specific determinants of migration normally attributed to the unsystematic error u and unobserved to the econometrician may be correlated with included covariates, in particular with income. Suppose that the characteristic 'entrepreneurship' was rewarded somewhat in eastern Germany but even more so in the West, so that his factor will elicit a positive migration intention. A misspecified equation excluding 'entrepreneurship' would therefore result in upward-biased

(1998) Burda, M., Härdle, W., Müller, M. and Werwatz, A.
Semiparametric Analysis of German East-West Migration Intentions: Facts and Theory.

Journal Applied Econometrics, 13, 525-541

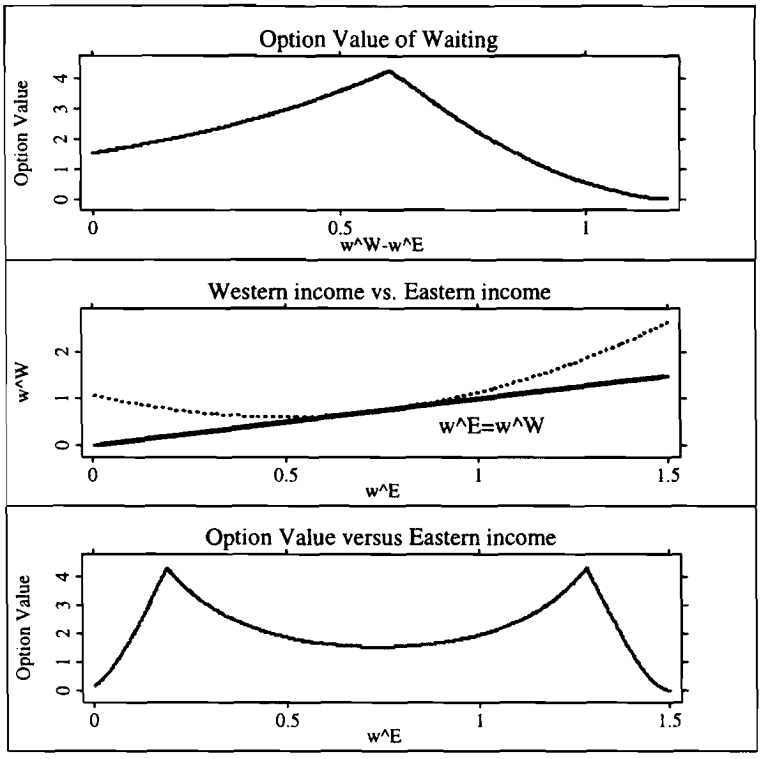


Figure 5. Relationship between Eastern and Western income

estimates of the effect of household income. More generally, in this first round of analysis we omitted a number of other variables including the labour market status of partners and other family members which might also bias our results in ways which are highly dependent on the particular effect of income assumed to be relevant (see discussion in the previous sections above).

7. CONCLUSIONS

In this paper we explored empirically the intention to migrate using microdata from the German Socio-Economic Panel. Fitting a parametric Generalized Linear Model (GLM) produced an unsatisfactory estimate of the influence of income. By estimating a Generalized Partial Linear Model (GPLM) we found an S- or U-shaped relation between income and (the systematic part of) migration propensity. This functional form was not detected by a specification search within the framework of a parametric GLM. The nonmonotone influence is also estimated for individual states in eastern Germany, so it appears to be robust phenomenon begging for explanation.

This paper is primarily exploratory, but we conclude by pointing out that economic theory suggest a number of alternative explanations in addition to the traditional 'Marshallian'

(1998) Burda, M., Härdle, W., Müller, M. and Werwatz, A.
Semiparametric Analysis of German East-West Migration Intentions: Facts and Theory.

Journal Applied Econometrics, 13, 525-541

arguments: the recently proposed option value of waiting theory, liquidity constraints, wealth-conditioned immobility, as well as unobservable heterogeneity. Future work will be directed at a tighter parameterization and use more sample information in estimation in order to identify which of these forces is operative and for which individuals.

APPENDIX: ALGORITHM FOR GPLM

In this section we indicate how the estimates $\hat{\beta}$, $\hat{m}$, $\bar{m}$ and the test statistic can be numerically computed. The algorithm can be motivated as follows. Consider the parametric (profile) likelihood function

$$\mathcal{L}(\beta) = \sum_{i=1}^{n} L(\mu_{i,\beta}, y_i) \tag{A1}$$

$\mu_{i,\beta} = G\{x_i^T\beta + m_\beta(t_i)\}$. This function is optimized to obtain an estimate for β. The *smoothed* or *local* likelihood

$$\mathcal{L}^h(m_\beta(t)) = \sum_{i=1}^{n} \mathcal{K}_h(t - t_i) L\{\mu_{i,m_\beta(t)}, y_i\} \tag{A2}$$

$\mu_i, m_\beta(t) = G\{x_i^T\beta + m_\beta(t)\}$ is optimized to estimate the smooth function $m_\beta(t)$ at point t. The local weights $\mathcal{K}_h(t - t_i)$ here denote kernel weights with $\mathcal{K}$ denoting a kernel function and h the bandwidth.

Abbreviate now $m_j = m_\beta(t_j)$ and the individual log-likelihood in y_i by $\ell_i(\eta) = L\{G(\eta), y_i\}$. In the following, ℓ_i' and ℓ_i'' denote the derivatives of $\ell_i(\eta)$ with respect to η. The maximization of the local likelihood (A2) requires to solve

$$0 = \sum_{i=1}^{n} \ell_i'(x_i^T\beta + m_j)\mathcal{K}_h(t_i - t_j) \tag{A3}$$

For β we have from equation (A1) to solve

$$0 = \sum_{i=1}^{n} \ell_i'(x_i^T\beta + m_i)\{x_i + m_i'\} \tag{A4}$$

A further differentiation of equation (A3) leads to an expression for the derivative m_j' of m_j with respect to β

$$m_j' = -\frac{\sum_{i=1}^{n} \ell_i''(x_i^T\beta + m_j)\mathcal{K}_h(t_i - t_j)x_i}{\sum_{i=1}^{n} \ell_i''(x_i^T\beta + m_j)\mathcal{K}_h(t_i - t_j)} \tag{A5}$$

(1998) Burda, M., Härdle, W., Müller, M. and Werwatz, A.
Semiparametric Analysis of German East-West Migration Intentions: Facts and Theory.

Journal Applied Econometrics, 13, 525-541

Equations (A3) and (A4) imply the following iterative Newton–Raphson type algorithm. Alternatively, the functions ℓ_i'' can be replaced by their expectations (w.r.t. to y_i) to obtain a Fisher scoring-type procedure.

Profile likelihood algorithm

- *Updating step for* β

$$\beta^{\text{new}} = \beta - \mathcal{B}^{-1} \sum_{i=1}^{n} \ell_i' \left(x_i^T \beta + m_i \right) \tilde{x}_i$$

with a Hessian-type matrix

$$\mathcal{B} = \sum_{i=1}^{n} \ell_i''(x_i^T \beta + m_i) \tilde{x}_i \tilde{x}_i^T$$

and

$$\tilde{x}_j = x_j - \frac{\sum_{i=1}^{n} \ell_i''(x_i^T \beta + m_j) \mathcal{K}_h(t_i - t_j) x_i}{\sum_{i=1}^{n} \ell_i''(x_i^T \beta + m_j) \mathcal{K}_h(t_i - t_j)}$$

- *Updating step for* m_j

$$m_j^{\text{new}} = m_j - \frac{\sum_{i=1}^{n} \ell_i'(x_i^T \beta + m_j) \mathcal{K}_h(t_i - t_j)}{\sum_{i=1}^{n} \ell_i''(x_i^T \beta + m_j) \mathcal{K}_h(t_i - t_j)}$$

The updating step for m_j is of quite complex structure. In some models (in particular, for identity and exponential link functions G) equation (A3) can be solved explicitly for m_j. For more details on this algorithm and possible simplifications we refer to Müller (1997).

To obtain the bias corrected parametric estimate $\bar{m}$, one needs to apply the updating step for $m_j = m_\beta(t_j)$, keeping $\tilde{\beta}$ fixed.

ACKNOWLEDGEMENTS

The research in this paper was supported by Sonderforschungsbereich 373 at Humboldt–Universität zu Berlin, http://sfb.wiwi.hu-berlin.de. The paper is printed using funds made available by the Deutsche Forschungsgemeinschaft.

We would like to thank Alan Kirman, Joel Horowitz, Richard Blundell and participants of the Tilburg Conference and the Berlin–Paris seminar for very useful comments and suggestions. Swetlana Schmelzer's help in designing the graphs in XploRe is gratefully acknowledged.

Journal Applied Econometrics, 13, 525-541

REFERENCES

Büchel, F. and J. Schwarze (1994), 'Die Migration von Ost- nach Westdeutschland — Absicht und Realisierung', *Mitteilungen aus der Arbeitsmarkt- und Berufsforschung*, **27**(1), 43–52.

Burda, M. (1993), 'The determinants of east–west German migration — some first results', *European Economic Review*, **37**, 452–461.

Burda, M. C. (1995), 'Migration and the option value of waiting', *Economic and Social Review*, **27**, 1–19.

Burda, M. and C. Schmidt (1997), 'Getting behind the east–west wage differential: Theory and evidence', in R. Pohl (ed.), *Wandeln oder Weichen — Herausforderungen der wirtschaftlichen Integration für Deutschland*, Sonderheft 3/97, Wirtschaft im Wandel.

Dixit, A. K. and R. S. Pindyck (1994), *Investment Under Uncertainty*, Princeton University Press, Princeton.

Dunn, T., M. Kreyenfeld and M. Lovely (1997), 'Communist human capital in a capitalist labor market — the experience of East German and ethnic German immigrants to West Germany', *Vierteljahrshefte zur Wirtschaftsforschung*, **66**, 151–158.

Faini, R. and A. Venturini (1993), 'Trade, aid and migrations', *European Economic Review*, **37**, 435–442.

Faini, R. and A. Venturini (1994), 'Migration and growth: The experience of Southern Europe', Discussion Paper 964, CEPR.

Härdle, W., E. Mammen and M. Müller (1996), 'Testing parametric versus semiparametric modelling in generalized linear models, SFB 373 Discussion Paper 28, Institut für Statistik und Ökonometrie, Humboldt–Universität zu Berlin. To appear in *Journal of the American Statistical Association*.

Hastie, T. J. and R. J. Tibshirani (1990), *Generalized Additive Models*, Vol. 43 of *Monographs on Statistics and Applied Probability*, Chapman and Hall, London.

Manski, C. (1990), 'The use of intentions data to predict behavior: A best-case analysis', *Journal of the American Statistical Association*, **85**, 934–940.

Müller, M. (1997), 'Computer-assisted generalized partial linear models', *Interface '97 Proceedings*, Houston, Texas.

O'Connell, P. (1997), 'Migration under uncertainty: "Try your luck" or "Wait and see"', *Journal of Regional Science*, **47**, 331–347.

Pakes, A. (1986), 'Patents as options: Some estimates of the value of holding European patent stocks', *Econometrica*, **54**, 755–784.

Rothwell, G. and J. Rust (1995), 'Optimal response to a shift in regulatory regime: the case of the US nuclear power industry', *Journal of Applied Econometrics*, **10**, 75–118.

Rust, J. (1987), 'Optimal replacement of GMC bus engines: an empirical model of Harold Zurcher', *Econometrica*, **55**, 999–1033.

Schwarze, J. (1996), Beeinflußt das Lohngefälle zwischen Ost- und Westdeutschland das Migrationsverhalten der Ostdeutschen? *Allgemeines Statistisches Archiv*, **80**(1), 50–68.

Severini, T. A. and J. G. Staniswalis (1994), 'Quasi-likelihood estimation in semiparametric models', *Journal of the American Statistical Association*, **89**, 501–511.

Severini, T. A. and W. H. Wong (1992), 'Generalized profile likelihood and conditionally parametric models', *Annals of Statistics*, **20**, 1768–1802.

Sjaastad, L. (1962), 'The costs and returns of human migration', *Journal of Political Economy*, **70**, 80–93.

Stark, O. (1989), *The Migration of Labor*, Basil Blackwell, Oxford.

XploRe (1998), *XploRe — the interactive statistical computing environment*, WWW: http://www.xplore-stat.de

(1998) Burda, M., Härdle, W., Müller, M. and Werwatz, A.
Semiparametric Analysis of German East-West Migration Intentions: Facts and Theory.

JOURNAL OF APPLIED ECONOMETRICS
J. Appl. Econ., **13**, 525–541 (1998)

SEMIPARAMETRIC ANALYSIS OF GERMAN EAST–WEST MIGRATION INTENTIONS: FACTS AND THEORY

MICHAEL C. BURDA,[a] WOLFGANG HÄRDLE,[b] MARLENE MÜLLER[b]* AND AXEL WERWATZ[b]

[a]*Institut für Wirtschaftstheorie II, Wirtschaftswissenschaftliche Fakultät Humboldt–Universität zu Berlin, Germany*
[b]*Institut für Statistik und Ökonometrie, Wirtschaftswissenschaftliche Fakultät Humboldt–Universität zu Berlin, Germany*

SUMMARY

East–West migration in Germany peaked at the beginning of the 1990s although the average wage gap between Eastern and Western Germany continues to average about 25%. We analyse the propensity to migrate using microdata from the German Socioeconomic Panel. Fitting a parametric Generalized Linear Model (GLM) yields non-linear residual behavior. This finding is not compatible with classical Marshallian theory of migration and motivates the semiparametric analysis. We estimate a Generalized Partial Linear Model (GPLM) where some components of the index of explanatory variables enter non-parametrically. We find the estimate of the non-parametric influence in concordance with a number of alternative migration theories, including the recently proposed option-value-of-waiting theory. © 1998 John Wiley & Sons, Ltd.

1. INTRODUCTION

German East–West migration has been the subject of several recent papers. Using microdata from the German Socio-Economic Panel, Burda (1993), Büchel and Schwarze (1994) and Schwarze (1996) have investigated this issue empirically. Especially interesting is the fact that, although migration peaked in the early 1990s following unification, the gap between average Eastern and Western wages remains about 25% as of 1997.

We take the empirical findings of Burda (1993) as our point of departure. We re-analyse the data by estimating a Generalized Linear Model (GLM) but find that the GLM does not provide a satisfactory fit. Estimating a semiparametric Generalized Partial Linear Model (GPLM) reveals a non-linear, non-monotonic influence of household income on the propensity to migrate from East to West. This non-linear influence of income, while difficult to reconcile with classical economic theory of migration, is compatible with a number of alternative models of the migration decision including the option value approach proposed by Dixit and Pindyck (1994) and applied recently to the migration decision by Burda (1995) and O'Connell (1997). It is also consistent with unobserved heterogeneity and misspecification of the estimation equation.

In the following section we present a brief discussion of the classical (Marshallian) theory of migration behaviour. In Section 3 we introduce the data and discuss how facts and theory play together. Results from fitting a parametric GLM to the data are presented in Section 4. As we shall see, standard logit analysis does not sufficiently capture the phenomenon underlying the observations. We therefore turn to a more flexible setting by allowing some components to

* Correspondence to: Marlene Müller, Institut für Statistik und Ökonometrie, Humboldt Univeristät, Spandauer Str. 1, D-10178 Berlin, Germany. E-mail: marlene @wiwi-hu.berlin.de

Contract grant sponsor: Deutsche Forschungsgemeinschaft.

CCC 0883–7252/98/050525–17$17·50

Received 15 October 1997
Revised 27 April 1998

Journal Applied Econometrics, 13, 525-541

take a non-parametric form. These semiparametric Generalized Partial Linear Models (GPLM) are described and estimated in Section 5. In Section 6 we discuss our findings and speculate on theoretical explanations for our results. Section 7 concludes the paper.

2. SOME THEORETICAL CONSIDERATIONS

Since Ravenstein's pathbreaking work on the determinants of migration more than a century ago, income has been the focus of economists' attempts to explain spatial mobility. More precisely, the difference between income at home (W^E) and the attainable income upon migration (W^W) has been singled out as the key explanatory variable (Sjaastad, 1962). Some migration is an investment, a forward-looking agent will care not only about the current income differential but also about future income differentials. That is, he will consider the net expected present value of future additional income earned if he decides to migrate.

Yet even if this expected present value is positive, an agent may not migrate if the fixed costs of migrating are sufficiently high. Such fixed costs include pecuniary components associated with physically moving a household from one place to another. In addition, moving away means leaving behind a familiar environment as well as friends and family members. Following classical ('Marshallian') economic theory, we may therefore say that a rational, forward-looking agent will migrate if the expected present value of the income stream from migrating exceeds monetary valuation of the associated fixed costs, or if the expected net present value from migrating (net of fixed costs) is positive. Incorporating risk aversion will change the trigger rule, but at most by a constant amount which would depend on the relative riskiness of the options and individual preferences.

Under a number of weak assumptions about the stochastic process generating relative income, the expected present value of future gains from migration will be a function of the current observed income differential, and for plausible assumptions this relationship will be linear. To consider an extreme but simple example, if the absolute per-period income differential $\Omega_t = W_t^W - W_t^E$ follows an arithmetic Brownian process with negative drift ν, then the expected present value of migration in time $t = 0$ is given by $V^m = (\Omega_0 - \nu/\delta)/\delta$, where δ denotes the discount rate.

Let the fixed costs of migration (including monetary equivalent of utility loss from moving) be given by F and denote the migration decision by the binary variable Y ($Y = 1 \rightarrow$ migration). Then the decision rule for a rational agent can be formally written as:

$$Y = \begin{cases} 1 & \text{if } V^m = \frac{1}{\delta}(\Omega_0 - \nu/\delta) - F > 0 \\ 0 & \text{otherwise} \end{cases} \tag{1}$$

This theory delivers the clear prediction that an increase in period t income by reducing Ω_t will decrease migration propensity, holding alternatives available in the West constant.

3. THE DATA

In the empirical analysis we use data drawn from the German Socio-Economic Panel (GSOEP). The GSOEP is a representative survey of German households that was extended to the former East in 1990. We use 3367 observations from the GSOEP's second East German wave which was collected in the spring of 1991 (time $t = 0$). All calculations were carried out with the statistical computing environment XploRe (1998).

Journal Applied Econometrics, 13, 525-541

Because very few actual migrants were observed in this wave of the GSOEP, we use migration propensity ('intention') as the dependent variable Y.

At the outset, it is important to state that this variable—as is the case with all intentions variables—is somewhat problematic for a number of reasons. (For an extensive discussion of this problem as well as a plea for not disregarding such information, see Manski, 1990.) First, agents may be simply irrational and have little idea of what their future behaviour may be or of the probability distribution of future events conditioning future decisions. Second, even if agents are rational in the sense that they can forecast their own future decision-making process and have rational expectations of future forcing variables, future decisions (realizations) may be correlated across individuals due to systematic intervening shocks. In this paper we simply take the position that 'intentions' are a monotonic function of the underlying driving variables which motivate migration.

The theoretical discussion of the previous section has focused on the income differential between host region and home region and the fixed cost of migrating as the key explanatory variables. Yet measuring both quantities poses a challenge. Regarding the income differential, we are faced with the problem that the potential income in the West is not observable. Hence, some imputation is generally necessary. Since Germany shares the same institutions and language one could assume that upon migration eastern Germans are able to employ at least some component of their human capital, earning 'western returns' for their attributes, at least up to a (macroeconomic) constant. A natural approach to estimate W_0^W would be to imply estimates of a traditional earnings equation of the Mincer type, which attributes observed wages to either market 'returns' multiplied by observable measures of human capital endowment (education, experience, training, tenure) or to attributes unobservable to the econometrician modeled as a random disturbance. Estimating this relation on a sample of Westerners, however, will most likely produce biased estimates of returns to Easterners (Burda and Schmidt, 1997). Moreover, it is unclear how to use these estimates to calculate an imputed Western wage for those Easterners who are registered as unemployed or out of the labour force. Rather than producing spurious findings based on biased estimates of the West–East income differential (Dunn, Kreyenfeld and Lovely, 1997), we decide to include income in the East only. We shall return to this point when discussing our results in Section 6.

The GSOEP data provides a multitude of variables that arguably are related to the intention to migrate from the East to West. Starting from a set of roughly 30 potential explanatory variables considered in the empirical analysis of Burda (1993) we used economic intuition and statistical selection criteria to limit the number of explanatory variables. This was done merely for better exposition of the facts. The proposed statistical method is valid for any dimension of the vector of explanatory variables.

Summary statistics for Y and the explanatory variables are given in Table I. Presence of a partner, home ownership and increasing age are expected to increase the fixed cost of migrating whereas relatives or friends in the West supposedly have the opposite effect. Age will also influence the migration decision via the discount rate. The variable *environmental satisfaction* is measured on a scale from 1 ('very unhappy with environmental conditions') to 10 ('very happy') and can therefore be expected to have a negative influence on migration propensity. The sign of the coefficients of the gender, city size and education variables is rather unclear apriori.

We have separated *age* and *household income* from the remaining explanatory variables in the table as—for the purposes of this study—they can be regarded as *continuous* explanatory variables.

(1998) Burda, M., Härdle, W., Müller, M. and Werwatz, A.
Semiparametric Analysis of German East-West Migration Intentions: Facts and Theory.

Journal Applied Econometrics, 13, 525-541

Table I. Summary statistics

		Mean	S.D.	Expected effect
Y	*migration intention*	0·394	0·489	
X_1	*female*	0·511	0·500	
X_2	*partner*	0·854	0·353	−
X_3	*owner*	0·322	0·467	−
X_4	*family/friends in west*	0·855	0·352	+
X_5	*unemployed/jobloss certain*	0·196	0·397	+
X_6	*environmental satisfaction*	3·9	2·4	−
X_7	*city size < 10,000*	0·522	0·499	
X_8	*city size 10–100,000*	0·342	0·474	
X_9	*university degree*	0·085	0·278	
X_{10}	*age* min. 18, max. 65	39·4	12·8	−
X_{11}	*household income* min. 200, max. 4000	2189·5	754·7	

4. PARAMETRIC ESTIMATION RESULTS

Collect the explanatory variables described in the previous section into the vector x. The goal of the empirical analysis is to estimate the probability of migration intention, i.e. $E(Y \mid x) = \text{Prob}(Y = 1 \mid x)$. A natural starting point for estimating this probability is fitting a parametric GLM. More precisely, we estimated a logit model.

This parametric model is based on two assumptions. First, the underlying latent variable Y is a sum of a linear index of the explanatory variables x and an individual error term u. Second, the cumulative distribution function (cdf) of u conditional on x is the logistic distribution function. Combining both assumptions gives

$$E(Y \mid x) = \text{Prob}(Y = 1 \mid x) = \{1 + \exp(-x^T\beta)\}^{-1} \tag{2}$$

As usual, $G(u) = \{1 + \exp(-u)\}^{-1}$ is called the (inverse) link function.

Table II gives the Maximum Likelihood logit estimates of β. Most coefficients have the expected sign: age, a partner, home ownership and environmental satisfaction reduce migration propensity whereas family or friends in the West and poor labour market prospects in the East have the opposite effect.

The estimated coefficient of the linear logit specification suggests that migration propensity significantly increases with household income. Figure 1 reflects the actual dependence of the response Y on the variables age and income.We have plotted each variable versus the logits $\log(\hat{p}/1 - \hat{p})$ where $\hat{p}$ are the relative frequencies for $Y = 1$ (migration intention). Essentially, these logits are obtained from classes of neighboured realizations (where the range of either age or income has been divided into 50 equidistant intervals). In case that $\hat{p}$ was 0 or 1, several classes were merged. Thicker bullets correspond to move observations in a class. Figure 1 shows that age has an almost linear influence on migration intention, whereas the relationship between income and migration intention exhibits a U-shaped curve.

If we include the square of household income as an additional regressor then both income coefficients are individually insignificant. This finding may lead an analyst to conclude that income does not have a non-linear influence. Yet, if we add income cubed as a regressor to the

(1998) Burda, M., Härdle, W., Müller, M. and Werwatz, A.
Semiparametric Analysis of German East-West Migration Intentions: Facts and Theory.

Journal Applied Econometrics, 13, 525-541

Table II. Logit estimates

Dependent variable: migration intention		
Variable	Estim. coeff.	*t*-ratio
constant	1·864	7·74
female	−0·233	−3·03
partner	−0·325	−2·87
owner	−0·576	5·79
family/friends in west	0·647	5·61
unemployed	0·217	2·24
environmental satisfaction	−0·057	−3·52
city size < 10,000	−0·718	−5·69
city size 10–100,000	−0·347	−2·91
university degree	0·481	3·56
age	−0·050	−14·89
household income	0·0001202	2·22
Sample size: 3367, log likelihood: −1992·7		

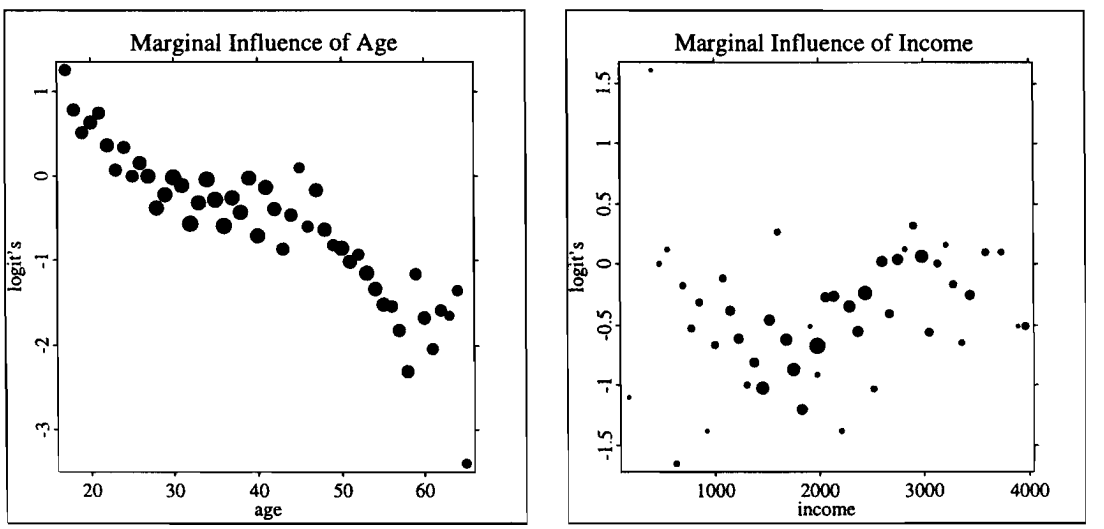


Figure 1. Marginal influence of age (left) and income (right) on migration intention, visualized by logits on classes

model that already includes income and income squared then all three income coefficients are individually as well as jointly significant. These findings are summarized in Table III.

Rather than continuing with the refinement of this parametric specification we decided to estimate a semiparametric Generalized Partial Linear Model which allows the data to freely determine the shape of the influence of income on migration propensity. By means of generalized additive modelling (Hastie and Tibshirani, 1990) this can be extended to the variable age as well. An analysis of this model yielded a linear dependence of migration propensity on age (as in Figure 1). We therefore included only income as a possible non-linear candidate.

(1998) Burda, M., Härdle, W., Müller, M. and Werwatz, A.
Semiparametric Analysis of German East-West Migration Intentions: Facts and Theory.

Journal Applied Econometrics, 13, 525-541

Table III. Parametric specification search

Variable	Estim. coeff.	*t*-ratio
'Quadratic' model		
household income	−0·0001288	−0·507
household income^2	5·46e-08	1·002
'Cubic' model		
household income	−0·0016491	−2·130
household income^2	8·08e-07	2·206
household income^3	−1·12e-10	−2·080
Dependent variable: migration intention Same regressors as above besides income		

5. SEMIPARAMETRIC ESTIMATION RESULTS

Before turning to estimates, we will briefly introduce the generalized partially linear model (GPLM). As before, the GPLM assumes that the mean of Y is related to an index of explanatory variables via the known link function G. Contrary to the logit model of the previous section the index of explanatory variables is composed of a linear parametric component and a non-parametric component. That is, the GPLM assumes that

$$E(Y \mid x, t) = G\{x^T\beta + m(t)\} \tag{3}$$

where—in a slight abuse of notation—we have collected the explanatory variables that enter the argument of $G(\cdot)$ linearly in the $p \times 1$ vector x, and those that enter non-linearly in the $q \times 1$ vector t. The unknown quantities that need to be estimated are the parameter vector β and the unknown function $m(\cdot)$. Note that there is no intercept parameter since it can be absorbed into the non-parametric part $m(t)$. In the empirical analysis x will—with the exception of age—be made up of discrete (categorical) variables while t contains solely household income.

The estimation methods for model (3) are based on the idea that an estimate $\hat{\beta}$ can be found for known $m(\cdot)$, and an estimate $\hat{m}(\cdot)$ can be found for known β. In what follows we will concentrate on *profile likelihood* estimation which goes back to Severini and Wong (1992) and Severini and Staniswalis (1994). Denote by $L(\mu, y)$ the individual log-likelihood, where $\mu = E(Y \mid x, t) = G\{x^T\beta + m(t)\}$. The profile likelihood uses two different likelihood functions for the estimation of the parametric and semiparametric components. The usual likelihood for n i.i.d. observations (x_i, t_i, y_i)

$$\mathcal{L}(\beta) = \sum_{i=1}^{n} L\{\beta^T x_i + m_\beta(t_i); y_i\} \tag{4}$$

is used to obtain $\hat{\beta}$ and a 'smoothed' likelihood

$$\mathcal{L}_h(\eta) = \sum_{i=1}^{n} K_h(t - t_i)L(\beta^T x_i + \eta; y_i) \tag{5}$$

(1998) Burda, M., Härdle, W., Müller, M. and Werwatz, A.
Semiparametric Analysis of German East-West Migration Intentions: Facts and Theory.

Journal Applied Econometrics, 13, 525-541

Table IV. GPLM estimates

Dependent variable: *migration intention*				
	GPLM estimates		Logit estimates	
Variable	Coeff.	t-ratio	Coeff.	t-ratio
female	−0·238	−3·09	−0·233	−3·03
partner	−0·282	−2·44	−0·325	−2·87
owner	−0·569	−5·71	−0·576	−5·79
family/friends in west	0·640	5·54	0·647	5·61
unemployed	0·216	2·23	0·217	2·24
environmental satisfaction	0·056	−3·47	−0·057	−3·52
city size < 10,000	−0·689	−5·43	−0·718	−5·69
city size 10–100,000	−0·323	−2·71	−0·347	−2·91
university degree	0·471	3·48	0·481	3·56
age	−0·050	−14·89	−0·050	−14·89
Sample size: 3367, log likelihood: −1989·8, GPLM bandwidth: 0·3				

for the non-parametric smooth function $\hat{m}_\beta(t) = \eta$ at point t and $K_h(u) = h^{-1}K(u/h)$ a kernel function with bandwidth h (Severini and Staniswalis, 1994) belongs to an exponential family using the

The computational algorithm consists of searching maxima of both likelihoods simultaneously. A detailed description of the algorithm can be found in the Appendix. It turns out that the resulting estimator $\hat{\beta}$ is $\sqrt{n}$-consistent and asymptotically normal, and that estimators $\hat{m} = \hat{m}_{\hat{\beta}}$ are consistent in supremum norm (see Severini and Staniswalis, 1994).

Table IV gives the GPLM estimates of β in a model that includes the same explanatory variables as the logit fit of Table II. The logit estimates and their t-ratios are also reported to conveniently compare results across the different approaches. In general, the GPLM estimates are very close to their logit counterparts. In terms of the GPLM, income plays the role of the variable t in equation (3). The estimated influence of income is depicted in Figure 2, with income on the horizontal axis and the estimate of $m(t)$ on the vertical axis. The highly non-linear estimate of $m(t)$ strongly contrasts with the linear influence of income implied by the logit model which we have also included in Figure 2.

The GPLM fit suggests an S-shaped effect of income, or a U-shaped influence over the range of income values that carry most of the mass of the income distribution. The bandwidth h underlying the estimate of $m(t)$ was set equal to 30% of the range of household income. The U-shaped estimate is obtained for a range of values of h, though. Note that the decreasing part of $\hat{m}(t)$ above $t = 3000$ may be attributed to random fluctuations for this bandwidth size. Above this income level, we have only a small number of observations (see Figure 1).

The visual impression of Figure 2 suggests that the estimate of $m(t)$ significantly deviates from the estimated linear influence of the parametric GLM fit. We use a test procedure to formally test that $m(t)$ is a linear function:

$$H_0 : m(t) = \alpha t + \alpha_0$$

$$H_1 : m(t) \text{ is an arbitrary smooth function}$$

(1998) Burda, M., Härdle, W., Müller, M. and Werwatz, A.
Semiparametric Analysis of German East-West Migration Intentions: Facts and Theory.

Journal Applied Econometrics, 13, 525-541

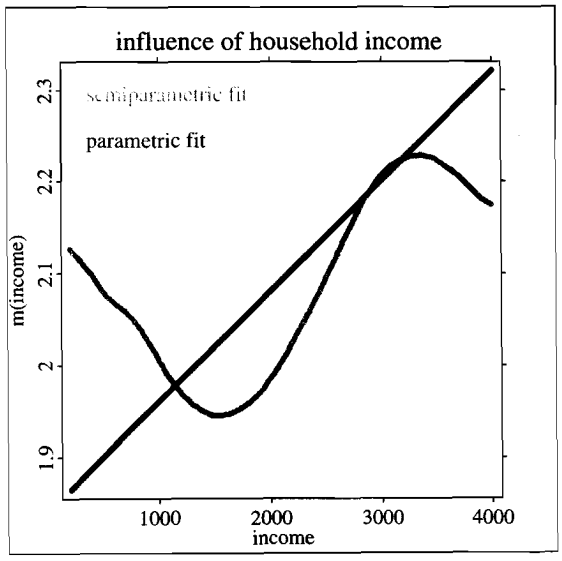


Figure 2. Influence of the net household income on migration propensity

This test is based on comparing the semiparametric estimates with the parametric estimates

$$(\tilde{\beta}, \tilde{\alpha}, \tilde{\alpha}_0) = \arg\min_{\beta,\alpha,\alpha_0} \sum_{i=1}^{n} L[G\{x_i^T\beta + \alpha t_i + \alpha_0\}; y_i] \tag{6}$$

where α denotes the coefficient of income and α_0 the constant in the parametric fit.

A test of the hypothesis GLM (logit model) against the alternative of a GPLM may be based on the likelihood ratio statistic. Denote by $\tilde{\mu}_i = G(x_i^T\tilde{\beta}t + \tilde{\alpha}t + \tilde{\alpha}_0)$ the parametric GLM fit and by $\hat{\mu}_i = G\{x_i^T\hat{\beta} + \hat{m}(t)\}$ the GPLM fit. Hastie and Tibshirani (1990) propose using

$$R = 2\sum_{i=1}^{n}\{L(\hat{\mu}_i, y_i) - L(\tilde{\mu}_i, y_i)\} \tag{7}$$

which has heuristically a distribution that is similar to a χ^2 distribution. However, the degrees of freedom for the GPLM need to be replaced by an approximate value and theoretic distribution of R is unknown.

Härdle, Mammen and Müller (1996) propose a modification of the test statistic R. This modification is based on the fact that a direct comparison of $\hat{m}(t)$ and $\tilde{\alpha}t + \tilde{\alpha}_0$ can be misleading because $\hat{m}$ has a non-negligible smoothing bias. this holds even under the linearity hypothesis. Hence, a bias-corrected parametric estimate $\bar{m}(t)$ is used instead of $\tilde{\alpha}t + \tilde{\alpha}_0$.

Using this bias-corrected $\bar{m}(t)$ the following modified likelihood ratio test statistic is computed:

$$R^M = 2\sum_{i=1}^{n}\{L(\hat{\mu}_i, \hat{\mu}_i) - L(\bar{\mu}_i, \hat{\mu}_i)\} \tag{8}$$

(1998) Burda, M., Härdle, W., Müller, M. and Werwatz, A.
Semiparametric Analysis of German East-West Migration Intentions: Facts and Theory.

Journal Applied Econometrics, 13, 525-541

where $\tilde{\mu}_i = G\{x_i^T\tilde{\beta} + \tilde{m}(t_i)\}$ is the bias-corrected GLM fit and $\hat{\mu}_i$ the GPLM fit as before. Härdle *et al.* (1996) show asymptotic normality of R^M. The proof of this result is based on showing that the asymptotic expansion of R^M behaves approximately like a sum of $O(h)$ independent summands. This is typically not very large and indeed simulations show that the normal approximation need not work well for R^M (Müller, 1997). Therefore, for the calculations of quantiles, it is recommended to use the following bootstrap procedure:

(1) Generate samples $\{Y_1^*, \ldots, Y_n^*\}$ under the parametric hypothesis with $E^*(Y_i^*) = G(x_i^T\tilde{\beta} + \tilde{\alpha}t_i)$. Here E^* denotes the conditional expectation given $(x_1, t_1, \ldots, x_n, t_n)$.
(2) Calculate estimates $\hat{\beta}^*$, $\hat{m}^*$, $\tilde{\beta}^*$, $\tilde{\alpha}^*$, $\tilde{m}^*$ based on the bootstrap samples $\{(x_1, t_1, Y_1^*), \ldots, (x_n, t_n, Y_n^*)\}$. Furthermore, calculate test the statistic R^{M^*}. Repeat this n^* times. The quantiles of the distribution of R^M can be estimated by the quantiles of the conditional distribution of R^{M^*}.

Since in our case the distribution of Y is completely specified by $EY = \mu = G(x^T\beta + \alpha t + \alpha_0)$ (under the hypothesis of linearity) we resample from the Bernoulli distribution with parameters $\tilde{\mu}_i = G(x_i^T\tilde{\beta} + \tilde{\alpha}t_i + \tilde{\alpha}_0)$ (the parametric GLM fit).

Table V shows the result of both test procedures for the GLM versus the GPLM. With R^M we denote the test using test statistic (8), where the rest has been carried out using the normal approximation. R^{M^*} bootstrap denotes the results for the bootstrapped quantiles of R^M. Since an optimal bandwidth choice for the GPLM is not known, all tests were performed for a sequence of bandwidths. However, we can recognize a clear rejection of the linearity hypothesis across all bandwidths for the R and the bootstrapped R^{M^*}. The normal approximation for R^M works poorly for higher bandwidth levels, as indicated above.

6. INTERPRETING THE RESULTS: ALTERNATIVE EXPLANATIONS

In the previous section we found a significant non-linear relationship between migration intensity and household income which appears non-monotonic. That is, for certain intervals the migration propensity is increasing in household income. This is at variance with the linear relationship implied by the classical theory of migration outlined in Section 2. In this section we will briefly outline theoretic models of migration that may give rise to non-linearities in income and/or non-monotonic relationships and which therefore could aid in the interpretation of the shape of the estimate presented in Figure 2.

6.1 Option Value of Migration

One limiting aspect of the Marshallian theory of migration of Section 2 is its 'all-or-nothing' aspect; either migration occurs now or never. The work of Dixit and Pindyck (1994) and others

Table V. Observed significance level for linearity test for migration data, $n = 3367$

h	0·1	0·2	0·25	0·3	0·4
R	0·028	0·021	0·019	0·017	0·016
R^M	0·035	0·069	0·130	0·269	0·602
R^{M^*}	0·015	0·005	0·005	0·005	0·010

Note: 200 bootstrap replications. Bandwidth h in % of range of household income

(1998) Burda, M., Härdle, W., Müller, M. and Werwatz, A.
Semiparametric Analysis of German East-West Migration Intentions: Facts and Theory.

Journal Applied Econometrics, 13, 525-541

has shown that postponement of the decision without forsaking it can be a valuable option under a large class of irrevocable investment problems. future. In terms of equation (1), migrating today not only means incurring a fixed cost F and forgoing the current and future income in the sending region. It also means forgoing the opportunity to postpone migration on the basis of new, currently unanticipated information. This opportunity has positive (expected) value today because waiting brings more information about the future, which may evolve against migration in an unexpected way. Assuming no loss of opportunity is implied, postponement leaves open the possibility of migrating at a later date, saving the fixed cost over the interval.

This opportunity cost of migrating today — in addition to the expected present value of future income gains from migration net of migration costs — is referred to as the *option value of waiting* and we will denote it as V^o. V^o is equal to what one is willing to pay for the option to postpone the migration decision rather than having to decide 'now or never'. It can be calculated as the difference between the expected net present value from postponing migration, V^p, and the expected net present value from migrating today, V^m. V^o — which is a function of current household income, among other things — can be derived as the solution to a dynamic programming problem under a variety of assumptions (see Dixit and Pindyck, 1994).

Figure 3 graphs V^o (kinked curve in the lower panel), V^p (the positively sloped curve in the upper panel) and V^m (the dashed straight line in the upper panel) as functions of the current income differential. If the current wage differential is below MT (the 'marshallian trigger') immediate migrating does not have positive net value ($V^m < 0$). Hence V^o is just equal to V^p.

If the current wage differential is between MT and OT ('option-value trigger') then immediate migration has positive expected value and hence $V^o = V^p - V^m$. We have displayed the values V^o as vertical bars in the upper panel for selected values of the current wage differential. If the current wage differential is above OT then V^o is zero: the current wage differential is so large that any further postponement of migration has zero value.

It appears from Figure 3 that V^o has the opposite shape as the estimate relationship of the previous section. But V^o is the option value of *postponing* migration. That is, high values of V^o imply a low propensity to migrate and vice versa. This is clearly evident if we rewrite the 'classical' decision rule (1) to incorporate the option value of waiting:

$$Y = \begin{cases} 1 & \text{if } \frac{1}{\delta}(\Omega_0 + \nu/\delta) - F - V^o(\Omega_0) > 0 \\ 0 & \text{otherwise} \end{cases} \tag{9}$$

As a result, the option value theory applied to the migration decision requires mirroring V^o around the x-axis producing a U-shaped relationship. In principle, this shape is consistent with the empirical findings of Figure 2.

Arguably, a superior strategy is to estimate the option value of migration directly, as has been done recently by a number of researchers in other applications. For example, Pakes (1986) estimates the option value of patents in this spirit, Rust (1987) has used similar methods to estimate deep parameters of a dichotomous investment problem (optimal replacement of bus engines), and Rothwell and Rust (1995) have examined optimal response of the nuclear industry to regulatory changes. (A special issue of the *Journal of Applied Econometrics* in 1995 highlighted the considerable breadth of potential applications in this area.) In principle, this approach would be possible but difficult to implement for two reasons. First, explicit modelling of the dynamic programming problem would require more detailed information on the individual's characteristics than are available in the GSOEP (e.g. household wealth), although Rust (1987) has shown

Journal Applied Econometrics, 13, 525-541

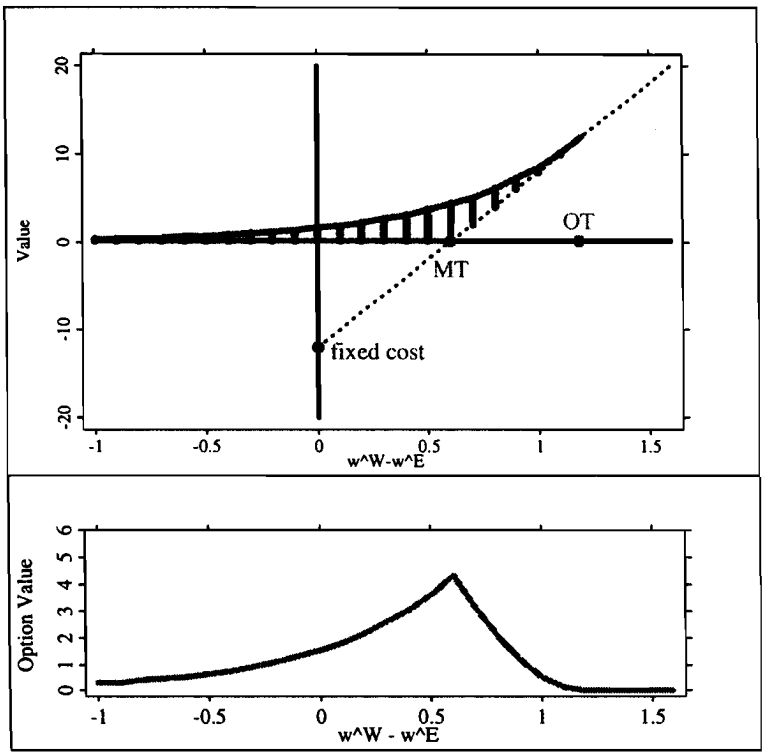


Figure 3. The option value of waiting

that problems with unobservable state variables can be overcome with sufficient assumptions concerning the structure of costs, etc., in a state-space model. Second, it would be necessary to impute a statistical process to wages at home and in the East, which would be highly speculative, given the limited data observations currently available since unification. Finally and perhaps most importantly, a number of plausible competing models exist (see below) which would give rise to confounding effects for individual decision makers. The derivation of a model nesting classical, option-value and other models of migration is beyond the scope of this exploratory paper, and is left for future research.

6.2 Risk-aversion, Income Effects and the Demand for Immobility

The previous discussion assumed risk neutrality and the absence of preferences for living at home or abroad. In fact, both risk aversion and the 'demand for immobility' might affect the propensity to migrate. The latter hypothesis has been put forward and investigated by, among others, Faini and Venturini (1993, 1994). Under the assumption that current place of residence is a normal 'good', the income effect of higher absolute wages at home implies a lower propensity to migrate. Alternatively, wealthier individuals might seek to escape impoverishment or reduce dependence on relatives by moving to the wealthier West, where public goods infrastructure is better and better-paying job opportunities are more plentiful. In the end, the effect of income is an empirical proposition and will depend on preferences of individual agents, but theory predicts that, given

(1998) Burda, M., Härdle, W., Müller, M. and Werwatz, A.
Semiparametric Analysis of German East-West Migration Intentions: Facts and Theory.

Journal Applied Econometrics, 13, 525-541

opportunities, wealthier families should exhibit a greater demand for immobility, and that the demand for immobility would play a subservient role for poorer to middle-income families. Indeed, this prediction is consistent with the shape of Figure 2 in for household income exceeding 1500 DM/month.

In the case of risk aversion, the influence of uncertainty is also ambiguous. In general, curvature in the utility function (as opposed to strict linearity in previous sections) will attenuate the attractiveness of migration if uncertainty impinges primarily on income abroad. An exception is Stark (1989) who shows that in some cases migration may serve a function of risk diversification or reduction. Below we show an example of how introducing curvature in the utility function (risk aversion, decreasing marginal utility) could affect the valuation of the migration decision without considering any option value. This line of reasoning is therefore also consistent with either a negative or a positive effect of absolute home income on migration propensities. The underlying presumption in the current application is, of course, that income at home is riskier, so that normal patterns of risk aversion imply demand for migration which is increasing in household income.

6.3 Borrowing Constraints and Liquidity Effects

In addition to aspects of preferences addressed in the previous section, it seems likely that capital markets are imperfect. Realistically, poorer segments of the population are likely to be liquidity-strapped and therefore unable to finance the migration investment, even if it has positive expected present value. Suppose that a component of moving costs, F, must be paid in cash, and thus cannot be financed out of future earnings in the host country. In such a situation, the absolute value of current income (and not relative to abroad) matters for some range—as long as assets are inadequate to finance the move. When the wage rises, some households which may have been willing to migrate for some time can do so, financing the move out of current income. This reasoning predicts a positive effect of home wage/income on migration propensity for some range of current income. To the extent that the probability of being faced with credit constraints depends negatively on income and wealth, borrowing constraints seem a good candidate explanation for the negative branch found in the household income range 0–1500 DM/month.

6.4 Potential Misspecification

One important potential explanation of our results is related to misspecification of the estimation equation, i.e. the arguments of the link function in equation (3). For example, one might raise the objection that migration models are based on income *differentials* while our empirical analysis employs income in the East only. In Figures 4 and 5 we try to clarify this point.

The top panel of Figure 4 is a repetition of the lower panel of Figure 3. The middle panel of Figure 4 plots a (hypothetical) Western income (vertical axis) versus Eastern income. The former was imputed using a Mincer regression on the Western half of the 1990–93 waves of the GSOEP. The lower straight line is the 45° reference line whereas the upper straight line corresponds to the Western. Now suppose that the option value of postponing migration is depending on the income differential. Then, in the situation indicated in the middle panel, the option value of postponing migration plotted as a function of the income in the East (lower panel) has the same shape as if it is plotted as a function of the income differential.

(1998) Burda, M., Härdle, W., Müller, M. and Werwatz, A.
Semiparametric Analysis of German East-West Migration Intentions: Facts and Theory.

Journal Applied Econometrics, 13, 525-541

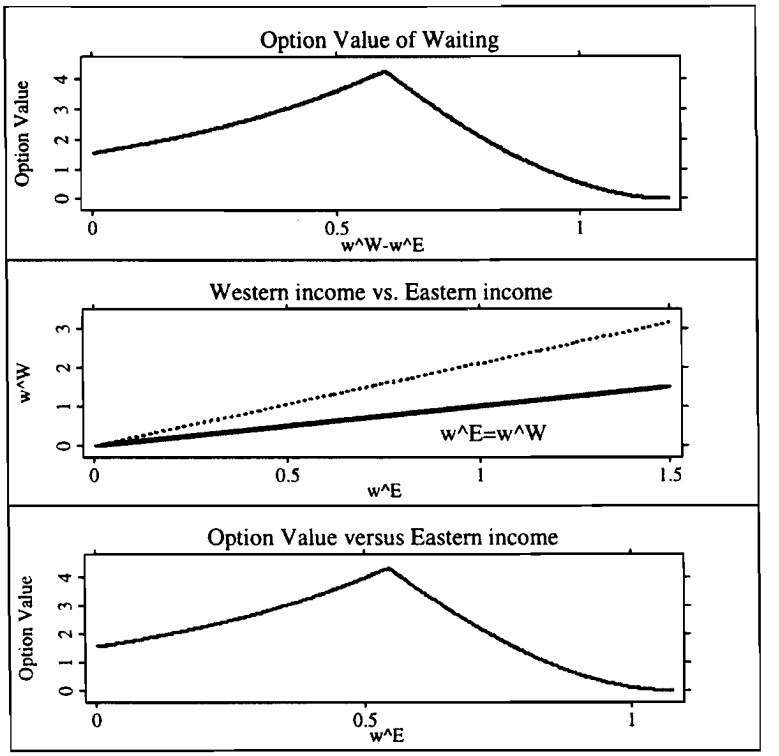


Figure 4. Eastern income versus Western income

Similarly, Figure 5 shows that different hypotheses about the relationship between Eastern and Western income still preserve the non-linearity of the option value — regardless whether it is plotted as a function of the income differential or income in the East. Specifically, the parabola in the middle panel of this figure reflects the hypothesis that Easterners with a low income (expect to) receive a relatively high Western income, those with a mid-range income receive a rather small increase in the West and individuals with a high Eastern income expect a relatively strong increase in income by moving to the West. Under this assumption about the relationship between income in the East and income in the West, and under the assumption that the option value of waiting depends on the current West–East income differential as depicted in the top panel of Figure 5, we obtain the non-linear relationship between the option value and income in the East as shown in the lower panel of Figure 5.

A more serious problem which potentially confounds all attempts to estimate a more tightly parameterized version of the model is unobserved heterogeneity. Individual-specific determinants of migration normally attributed to the unsystematic error u and unobserved to the econometrician may be correlated with included covariates, in particular with income. Suppose that the characteristic 'entrepreneurship' was rewarded somewhat in eastern Germany but even more so in the West, so that his factor will elicit a positive migration intention. A misspecified equation excluding 'entrepreneurship' would therefore result in upward-biased

(1998) Burda, M., Härdle, W., Müller, M. and Werwatz, A.
Semiparametric Analysis of German East-West Migration Intentions: Facts and Theory.

Journal Applied Econometrics, 13, 525-541

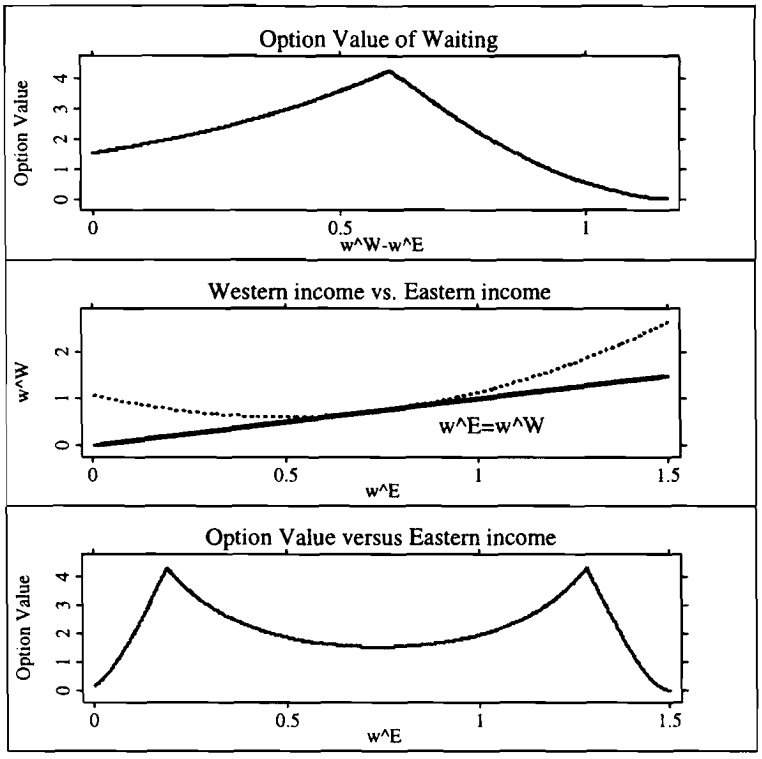


Figure 5. Relationship between Eastern and Western income

estimates of the effect of household income. More generally, in this first round of analysis we omitted a number of other variables including the labour market status of partners and other family members which might also bias our results in ways which are highly dependent on the particular effect of income assumed to be relevant (see discussion in the previous sections above).

7. CONCLUSIONS

In this paper we explored empirically the intention to migrate using microdata from the German Socio-Economic Panel. Fitting a parametric Generalized Linear Model (GLM) produced an unsatisfactory estimate of the influence of income. By estimating a Generalized Partial Linear Model (GPLM) we found an S- or U-shaped relation between income and (the systematic part of) migration propensity. This functional form was not detected by a specification search within the framework of a parametric GLM. The nonmonotone influence is also estimated for individual states in eastern Germany, so it appears to be robust phenomenon begging for explanation.

This paper is primarily exploratory, but we conclude by pointing out that economic theory suggest a number of alternative explanations in addition to the traditional 'Marshallian'

(1998) Burda, M., Härdle, W., Müller, M. and Werwatz, A.
Semiparametric Analysis of German East-West Migration Intentions: Facts and Theory.

Journal Applied Econometrics, 13, 525-541

arguments: the recently proposed option value of waiting theory, liquidity constraints, wealth-conditioned immobility, as well as unobservable heterogeneity. Future work will be directed at a tighter parameterization and use more sample information in estimation in order to identify which of these forces is operative and for which individuals.

APPENDIX: ALGORITHM FOR GPLM

In this section we indicate how the estimates $\hat{\beta}$, $\hat{m}$, $\bar{m}$ and the test statistic can be numerically computed. The algorithm can be motivated as follows. Consider the parametric (profile) likelihood function

$$\mathcal{L}(\beta) = \sum_{i=1}^{n} L(\mu_{i,\beta}, y_i) \tag{A1}$$

$\mu_{i,\beta} = G\{x_i^T\beta + m_\beta(t_i)\}$. This function is optimized to obtain an estimate for β. The *smoothed* or *local* likelihood

$$\mathcal{L}^h(m_\beta(t)) = \sum_{i=1}^{n} \mathcal{K}_h(t - t_i) L\{\mu_{i,m_\beta(t)}, y_i\} \tag{A2}$$

$\mu_i, m_\beta(t) = G\{x_i^T\beta + m_\beta(t)\}$ is optimized to estimate the smooth function $m_\beta(t)$ at point t. The local weights $\mathcal{K}_h(t - t_i)$ here denote kernel weights with $\mathcal{K}$ denoting a kernel function and h the bandwidth.

Abbreviate now $m_j = m_\beta(t_j)$ and the individual log-likelihood in y_i by $\ell_i(\eta) = L\{G(\eta), y_i\}$. In the following, ℓ_i' and ℓ_i'' denote the derivatives of $\ell_i(\eta)$ with respect to η. The maximization of the local likelihood (A2) requires to solve

$$0 = \sum_{i=1}^{n} \ell_i'(x_i^T\beta + m_j)\mathcal{K}_h(t_i - t_j) \tag{A3}$$

For β we have from equation (A1) to solve

$$0 = \sum_{i=1}^{n} \ell_i'(x_i^T\beta + m_i)\{x_i + m_i'\} \tag{A4}$$

A further differentiation of equation (A3) leads to an expression for the derivative m_j' of m_j with respect to β

$$m_j' = -\frac{\sum_{i=1}^{n} \ell_i''(x_i^T\beta + m_j)\mathcal{K}_h(t_i - t_j)x_i}{\sum_{i=1}^{n} \ell_i''(x_i^T\beta + m_j)\mathcal{K}_h(t_i - t_j)} \tag{A5}$$

(1998) Burda, M., Härdle, W., Müller, M. and Werwatz, A.
Semiparametric Analysis of German East-West Migration Intentions: Facts and Theory.

Journal Applied Econometrics, 13, 525-541

Equations (A3) and (A4) imply the following iterative Newton–Raphson type algorithm. Alternatively, the functions ℓ_i'' can be replaced by their expectations (w.r.t. to y_i) to obtain a Fisher scoring-type procedure.

Profile likelihood algorithm

- *Updating step for* β

$$\beta^{\text{new}} = \beta - \mathcal{B}^{-1} \sum_{i=1}^{n} \ell_i' \left(x_i^T \beta + m_i \right) \tilde{x}_i$$

with a Hessian-type matrix

$$\mathcal{B} = \sum_{i=1}^{n} \ell_i''(x_i^T \beta + m_i) \tilde{x}_i \tilde{x}_i^T$$

and

$$\tilde{x}_j = x_j - \frac{\sum_{i=1}^{n} \ell_i''(x_i^T \beta + m_j) \mathcal{K}_h(t_i - t_j) x_i}{\sum_{i=1}^{n} \ell_i''(x_i^T \beta + m_j) \mathcal{K}_h(t_i - t_j)}$$

- *Updating step for* m_j

$$m_j^{\text{new}} = m_j - \frac{\sum_{i=1}^{n} \ell_i'(x_i^T \beta + m_j) \mathcal{K}_h(t_i - t_j)}{\sum_{i=1}^{n} \ell_i''(x_i^T \beta + m_j) \mathcal{K}_h(t_i - t_j)}$$

The updating step for m_j is of quite complex structure. In some models (in particular, for identity and exponential link functions G) equation (A3) can be solved explicitly for m_j. For more details on this algorithm and possible simplifications we refer to Müller (1997).

To obtain the bias corrected parametric estimate $\bar{m}$, one needs to apply the updating step for $m_j = m_\beta(t_j)$, keeping $\tilde{\beta}$ fixed.

ACKNOWLEDGEMENTS

The research in this paper was supported by Sonderforschungsbereich 373 at Humboldt–Universität zu Berlin, http://sfb.wiwi.hu-berlin.de. The paper is printed using funds made available by the Deutsche Forschungsgemeinschaft.

We would like to thank Alan Kirman, Joel Horowitz, Richard Blundell and participants of the Tilburg Conference and the Berlin–Paris seminar for very useful comments and suggestions. Swetlana Schmelzer's help in designing the graphs in XploRe is gratefully acknowledged.

Journal Applied Econometrics, 13, 525-541

REFERENCES

Büchel, F. and J. Schwarze (1994), 'Die Migration von Ost- nach Westdeutschland — Absicht und Realisierung', *Mitteilungen aus der Arbeitsmarkt- und Berufsforschung*, **27**(1), 43–52.

Burda, M. (1993), 'The determinants of east–west German migration — some first results', *European Economic Review*, **37**, 452–461.

Burda, M. C. (1995), 'Migration and the option value of waiting', *Economic and Social Review*, **27**, 1–19.

Burda, M. and C. Schmidt (1997), 'Getting behind the east–west wage differential: Theory and evidence', in R. Pohl (ed.), *Wandeln oder Weichen — Herausforderungen der wirtschaftlichen Integration für Deutschland*, Sonderheft 3/97, Wirtschaft im Wandel.

Dixit, A. K. and R. S. Pindyck (1994), *Investment Under Uncertainty*, Princeton University Press, Princeton.

Dunn, T., M. Kreyenfeld and M. Lovely (1997), 'Communist human capital in a capitalist labor market — the experience of East German and ethnic German immigrants to West Germany', *Vierteljahrshefte zur Wirtschaftsforschung*, **66**, 151–158.

Faini, R. and A. Venturini (1993), 'Trade, aid and migrations', *European Economic Review*, **37**, 435–442.

Faini, R. and A. Venturini (1994), 'Migration and growth: The experience of Southern Europe', Discussion Paper 964, CEPR.

Härdle, W., E. Mammen and M. Müller (1996), 'Testing parametric versus semiparametric modelling in generalized linear models, SFB 373 Discussion Paper 28, Institut für Statistik und Ökonometrie, Humboldt–Universität zu Berlin. To appear in *Journal of the American Statistical Association*.

Hastie, T. J. and R. J. Tibshirani (1990), *Generalized Additive Models*, Vol. 43 of *Monographs on Statistics and Applied Probability*, Chapman and Hall, London.

Manski, C. (1990), 'The use of intentions data to predict behavior: A best-case analysis', *Journal of the American Statistical Association*, **85**, 934–940.

Müller, M. (1997), 'Computer-assisted generalized partial linear models', *Interface '97 Proceedings*, Houston, Texas.

O'Connell, P. (1997), 'Migration under uncertainty: "Try your luck" or "Wait and see"', *Journal of Regional Science*, **47**, 331–347.

Pakes, A. (1986), 'Patents as options: Some estimates of the value of holding European patent stocks', *Econometrica*, **54**, 755–784.

Rothwell, G. and J. Rust (1995), 'Optimal response to a shift in regulatory regime: the case of the US nuclear power industry', *Journal of Applied Econometrics*, **10**, 75–118.

Rust, J. (1987), 'Optimal replacement of GMC bus engines: an empirical model of Harold Zurcher', *Econometrica*, **55**, 999–1033.

Schwarze, J. (1996), Beeinflußt das Lohngefälle zwischen Ost- und Westdeutschland das Migrationsverhalten der Ostdeutschen? *Allgemeines Statistisches Archiv*, **80**(1), 50–68.

Severini, T. A. and J. G. Staniswalis (1994), 'Quasi-likelihood estimation in semiparametric models', *Journal of the American Statistical Association*, **89**, 501–511.

Severini, T. A. and W. H. Wong (1992), 'Generalized profile likelihood and conditionally parametric models', *Annals of Statistics*, **20**, 1768–1802.

Sjaastad, L. (1962), 'The costs and returns of human migration', *Journal of Political Economy*, **70**, 80–93.

Stark, O. (1989), *The Migration of Labor*, Basil Blackwell, Oxford.

XploRe (1998), *XploRe — the interactive statistical computing environment*, WWW: http://www.xplore-stat.de

(1998) Burda, M., Härdle, W., Müller, M. and Werwatz, A.
Semiparametric Analysis of German East-West Migration Intentions: Facts and Theory.

The Annals of Statistics
1998, Vol. 26, No. 3, 943–971

DIRECT ESTIMATION OF LOW-DIMENSIONAL COMPONENTS IN ADDITIVE MODELS[1]

BY JIANQING FAN,[2] WOLFGANG HÄRDLE AND ENNO MAMMEN

University of North Carolina, Humboldt-Universität zu Berlin and Ruprecht-Karls-Universität Heidelberg

Additive regression models have turned out to be a useful statistical tool in analyses of high-dimensional data sets. Recently, an estimator of additive components has been introduced by Linton and Nielsen which is based on marginal integration. The explicit definition of this estimator makes possible a fast computation and allows an asymptotic distribution theory. In this paper an asymptotic treatment of this estimate is offered for several models. A modification of this procedure is introduced. We consider weighted marginal integration for local linear fits and we show that this estimate has the following advantages.

(i) With an appropriate choice of the weight function, the additive components can be efficiently estimated: An additive component can be estimated with the same asymptotic bias and variance as if the other components were known.

(ii) Application of local linear fits reduces the design related bias.

1. Introduction. In this paper we consider the multivariate regression model

$$E(Y \mid X = x) = \mu + f_1(x_1) + f_{23}(x_2, x_3), \tag{1.1}$$

where Y is a real-valued dependent variable, $X = (X_1, X_2, X_3)$ is a vector of explanatory variables and μ is a constant. The variables X_1 and X_2 are continuous with values in $\mathbb{R}^p$ or $\mathbb{R}^q$, respectively, and X_3 is discrete and takes values in $\mathbb{R}^r$. For identifiability, we assume $Ef_1(X_1) = Ef_{23}(X_2, X_3) = 0$. The novelty of this paper is to directly estimate $f_1(x)$ at the usual nonparametric rate with good sampling properties. Our model includes the additive nonparametric regression model:

$$E(Y \mid U = u) = \mu + g_1(u_1) + \cdots + g_p(u_p), \tag{1.2}$$

where now $U = (U_1, \ldots, U_p)$ is a vector of explanatory variables. A discussion of this model can be found in Buja, Hastie and Tibshirani (1989) and Hastie and Tibshirani (1990). Model (1.2) is easy to interpret and is much more flexible than a linear model. Furthermore, the additive components g_j can be estimated with the one-dimensional nonparametric rate [Stone (1985, 1986)].

Received December 1995; revised August 1997.

[1] This research was supported by the Deutsche Forschungsgemeinschaft, Sonderforschungsbereich 373 "Quantifikation und Simulation ökonomischer Prozesse," Humboldt-Universität zu Berlin and NATO Collaborative Research Grant 931 363.

[2] Supported in part by NSF Grant DMS-95-04414 and NSA grant 96-1-0015.

Annals of Statistics, 26, 943-971

The main conclusion of this paper is somewhat surprising: The component g_j can be estimated with the same asymptotic bias and variance as the one-dimensional smoother, as if the other components were known. This kind of adaptivity result appears to be new in the literature. It provides foundational insights into additive modeling: Unknown components in the additive model, although increasing the effective number of parameters, do not add any extra difficulty of estimation, at least asymptotically.

In most papers, for the calculation of the additive components, algorithms have been proposed which are based on iterative procedures using backfitting. Recently, asymptotic properties of backfitting estimates have been analyzed in Opsomer and Ruppert (1997), Opsomer (1997) and Linton, Mammen and Nielsen (1997). Because of the implicit definition of these estimates, their behavior is difficult to understand. For this reason, in Linton and Nielsen (1995), Tjøstheim and Auestad (1994) and Chen, Härdle, Linton and Severance-Lossin (1996) a direct method has been proposed that is based on "marginal integration." This procedure is based on the fact that, up to a constant, $g_j(u_j)$ is equal to

$$EW(U_1, \ldots, U_{j-1}, U_{j+1}, \ldots, U_s)m(U_1, \ldots, U_{j-1}, u_j, U_{j+1}, \ldots, U_s),$$

where $m(u) = E(Y \mid U = u)$. Here W is a weight function with

$$EW(U_1, \ldots, U_{j-1}, U_{j+1}, \ldots, U_s) = 1.$$

The estimate of g_j is achieved by (weighted) marginal integration of an estimate of m. In particular, this method does not use iterations. Fast computation can be implemented. Furthermore, the explicit definition allows a detailed asymptotic analysis.

The present paper extends this idea in two directions:

(i) It introduces a weighting scheme W, which leads to efficient estimation [for another proposal for efficient estimation based on marginal integration, see Linton (1997)].

(ii) It allows a more flexible model, which can be incorporated with discrete data.

Our asymptotic analysis can be extended to the case that model (1.1) does not hold (see Remark 3). Then in the case of the additive model (1.2) the marginal integration estimate gives a consistent estimate of

$$\bar{g}_j(u_j) = EW(U_1, \ldots, U_{j-1}, U_{j+1}, \ldots, U_s)m(U_1, \ldots, U_{j-1}, u_j, U_{j+1}, \ldots, U_s).$$

This can be interpreted as an average effect of the jth component and is the best additive approximation under some specific L_2-norm [see Fan (1997)]. The backfitting estimate behaves quite differently. Under appropriate conditions it is a consistent estimate of g_j^* where $\mu + g_1^*(u_1) + \cdots + g_p^*(u_p)$ is the orthogonal projection in the Hilbert Space $L_2(p)$ onto the subspace of additive functions. Here p is the joint density of $(U_1, \ldots, U_p)$ (design density). For identifiability, g_j^* is normed s.t. $Eg_j^*(U_j) = 0$. This statement follows from the results of Linton, Mammen and Nielsen (1997). So, if model (1.2) is only

(1998) Fan, J., Härdle, W. and Mammen, E.
Direct Estimation of Low Dimensional Components in Additive Models.

Annals of Statistics, 26, 943-971

approximately true, we conjecture that backfitting will lead to a more accurate estimate of the full-dimensional regression function m. This would be preferable if one is interested in prediction. Furthermore, the application of marginal integration requires consistency of a full-dimensional smoother. This puts restrictions on the dimension that may not be shared by the backfitting estimate; see Linton, Mammen and Nielsen (1997). On the other hand, (in the case of model misspecification), the average effect $\bar{g}_j$ is always easy to interpret and it may be argued that marginal integration is preferable as a data analytic tool.

Our model includes additive partial linear models. With $X = (U_1, \ldots, U_p, X_3)$, $x = (u_1, \ldots, u_p, x_3)$ we write

$$E(Y \mid X = x) = \mu + g_1(u_1) + \cdots + g_p(u_p) + x_3^T \beta. \tag{1.3}$$

In this case, each nonparametric additive component can be estimated with optimal rate by our direct estimate $\hat{g}_j$, $j = 1, \ldots, p$. Furthermore, we will show that a least-squares estimate

$$\begin{pmatrix} \hat{\mu} \\ \hat{\beta} \end{pmatrix} = (Z^T Z)^{-1} Z^T \left(Y - \hat{g}_1 - \cdots - \hat{g}_p\right)$$

possesses root-n consistency. Here, for n observations $Y_1, \ldots, Y_n$ and design vectors $X^i = (U_{1i}, \ldots, U_{pi}, X_{3i})$, $i = 1, \ldots, n$, the vectors Y and $\hat{g}_j$ have elements Y_i and $\hat{g}_j(U_{ij})$, respectively, $i = 1, \ldots, n$; $j = 1, \ldots, p$. The design matrix Z has rows $(1, X_{3i}^T)$.

Another application of our model consists of partial interaction models

$$E(Y \mid U = u) = \mu + g_{12}(u_1, u_2) + g_3(u_3) + \cdots + g_s(u_s). \tag{1.4}$$

Our method directly applies interactions such as g_{12} by treating the rest of the variables as X_2-vectors and/or X_3-vectors [see (1.1)].

This paper is organized as follows. In Section 2, we introduce our estimation procedure. Section 3 presents asymptotic results. A further discussion of additive models (1.2), additive partially linear models (1.3) and partial interaction models (1.4) can be found in Section 4. In Section 5 our methodology is applied to a data set on female labor supply in East Germany. Furthermore, there a small simulation study can be found. Assumptions and proofs are postponed to Section 6.

2. Estimation procedure. Let $m(x_1, x_2, x_3) = E(Y \mid X_1 = x_1, X_2 = x_2, X_3 = x_3)$ be the regression function and let $W: \mathbb{R}^{q+r} \to \mathbb{R}$ be a known function with $EW(X_2, X_3) = 1$. Observe that under (1.1)

$$\begin{aligned} Em(x_1, X_2, X_3) W(X_2, X_3) &= \mu + f_1(x_1) + Ef_{23}(X_2, X_3) W(X_2, X_3) \\ &= \mu_1 + f_1(x_1) \\ &\equiv f_1^*(x_1), \end{aligned} \tag{2.1}$$

where

$$\mu_1 = \mu + Ef_{23}(X_2, X_3) W(X_2, X_3). \tag{2.2}$$

Annals of Statistics, 26, 943-971

Thus, f_1 can be directly estimated within a constant factor. This can be done by averaging out a nonparametric estimator of m with respect to other variables X_2, X_3. Since, in practice, $f_1(\cdot)$ will be normalized to have sample mean 0, the constant fact μ_1 is irrelevant to the final estimated curve. This kind of integration idea was studied in the additive model (1.1) by Tjøstheim and Auestad (1994), Linton and Nielsen (1995) and Chen, Härdle, Linton and Severance-Lossin (1996).

To utilize (2.1), we consider the local linear approximation near a fixed point x_1:

$$f_1(v_1) \approx a(x_1) + b^T(x_1)(v_1 - x_1),$$

where v_1 lies in a neighborhood of x_1. Further, the local constant approximation for f_{23} at a fixed point x_2 and x_3 is employed:

$$f_2(v_2, x_3) \approx c(x_2, x_3) \quad \text{for } v_2 \approx x_2.$$

Thus, in a neighborhood of (x_1, x_2) and for the given value of x_3, we can approximate the regression function as

$$\begin{aligned} m(v_1, v_2, x_3) &\approx \mu + a(x_1) + b^T(x_1)(v_1 - x_1) + c(x_2, x_3) \\ &\equiv \alpha + \beta^T(v_1 - x_1). \end{aligned} \tag{2.3}$$

Note that $f_{23}(\cdot; x_3)$ is locally approximated by a constant. This is because:

(i) the function $c(x_2, x_3)$ will be averaged out by an integration via (2.1);
(ii) the higher-order approximation will increase the number of local parameters and hence is harder to implement in higher dimensions.

In principle, we can approximate $f_1(\cdot)$ to a higher order. We opt not to do this for simplicity. Furthermore, the higher-order approximation rarely takes effect for the finite amount of data—the size of the local neighborhood plays a more crucial role [see, e.g., Fan and Gijbels (1996)].

Consider now that we have an i.i.d. data set $(Y_i, X_{1i}, X_{2i}, X_{3i})$, $i = 1, \ldots, n$, for model (1.1). The local model (2.3) leads to the following regression problem: Minimize

$$\sum_{i=1}^{n} \left(Y_i - \alpha - \beta^T(X_{1i} - x_1)\right)^2 K_{h_1}(X_{1i} - x_1) L_{h_2}(X_{2i} - x_2) I\{X_{3i} = x_3\}. \tag{2.4}$$

Here K and L are kernel functions and for bandwidths h_1 and h_2 we put

$$K_{h_1}(t) = \frac{1}{h_1^p} K\left(\frac{t}{h_1}\right) \quad \text{and} \quad L_{h_2}(t) = \frac{1}{h_2^q} L\left(\frac{t}{h_2}\right).$$

Note that the factor $K_{h_1} L_{h_2} I$ in (2.4) is just to confine our localization idea. Let $\hat{\alpha}(x)$ and $\hat{\beta}(x)$ be the solution to (2.4). Then, from (2.3) by setting $(v_1, v_2, x_3) = x$, we can easily see that $m(x) \approx \alpha$. Thus, our partial local

Annals of Statistics, 26, 943-971

linear estimator is $\hat{m}(x) = \hat{\alpha}$. By (2.1), we propose the following estimator:

$$\hat{f}_1^*(x_1) = \frac{1}{n}\sum_{i=1}^{n} \hat{m}(x_1, X_{2i}, X_{3i}) W(X_{2i}, X_{3i}) \tag{2.5}$$

and

$$\hat{f}_1(x_1) = \hat{f}_1^*(x_1) - \bar{f}_1, \qquad \bar{f}_1 = \frac{1}{n}\sum_{i=1}^{n} \hat{f}_1^*(X_{1i}). \tag{2.6}$$

Note that when the local constant fit is employed (i.e., $\beta = 0$) in (2.3), the resulting estimate $\hat{\alpha}$ is basically the multivariate kernel regression estimator.

Let X be the design matrix and let A be the diagonal weight matrix to the least-squares problem (2.4). Then

$$\begin{pmatrix} \hat{\alpha} \\ \hat{\beta} \end{pmatrix} = (X^T A X)^{-1} X^T A Y,$$

where $Y = (Y_1, \ldots, Y_n)^T$, and simple algebra shows that $\hat{m}(x) = \hat{\alpha}$ can be expressed as

$$\hat{m}(x) = \sum_{i=1}^{n} K_n(X_i - x) Y_i, \tag{2.7}$$

where, with $S_n(x) = (X^T A X)$ and $e_1^T = (1, 0, \ldots, 0)$,

$$K_n(t_1, t_2, t_3) = e_1^T S_n^{-1} \begin{pmatrix} 1 \\ t_1 \end{pmatrix} K_{h_1}(t_1) L_{h_2}(t_2) I\{t_3 = 0\}. \tag{2.8}$$

Note that it follows from least-squares theory that

$$\sum_{i=1}^{n} K_n(X_i - x) = 1 \quad \text{and} \quad \sum_{i=1}^{n} K_n(X_i - x)(X_{1i} - x_1) = 0. \tag{2.9}$$

3. Main results. Let us begin by introducing some notation. Let $p_1(x_1)$ and $p_{1,2}(x_1, x_2)$ be respectively the density of X_1 and (X_1, X_2) and let $p_{1,2|3}(x_1, x_2, | x_3)$, $p_{2|3}(x_2 | x_3)$ be respectively the conditional density of (X_1, X_2) given X_3 and of X_2 given X_3. Set $p_3(x_3) = P(X_3 = x_3)$. The conditional variance of $\varepsilon = Y - E(Y | X)$ is denoted by

$$\sigma^2(x) = E(\varepsilon^2 | X = x) = \operatorname{var}(Y | X = x),$$

where $X = (X_1, X_2, X_3)$. Let

$$\|K\|^2 = \int K^2 \quad \text{and} \quad \mu_2(K) = \int t t^T K(t)\, dt.$$

Then, under Condition A in Section 6, we have the following theorem that generalizes the main result in Linton and Nielsen (1995).

Annals of Statistics, 26, 943-971

THEOREM 1. *Under Condition* A *for a point* $x_1 \in \mathbb{R}^p$, *if the bandwidths are chosen such that* $nh_1^p h_2^q/\log n \to \infty$, $h_1 \to 0$, $h_2 \to 0$ *in such a way that* $h_2^d/h_1^2 \to 0$, *then*

$$\sqrt{nh_1^p}\left\{\hat{f}_1^*(x_1) - f_1(x_1) - \mu_1 - b(x_1) + o(h_1^2)\right\} \to N(0, v(x_1)), \tag{3.1}$$

where

$$b(x_1) = \tfrac{1}{2}h_1^2 \operatorname{tr}(f_1''(x_1)\mu_2(K)) \tag{3.2}$$

and

$$\begin{aligned} v(x_1) &= \|K\|^2 p_1(x_1) \\ &\quad \times E\left\{\sigma^2(X_1, X_2, X_3)\frac{p_{2|3}^2(X_2 \mid X_3)W^2(X_2, X_3)}{p_{1,2|3}^2(X_1, X_2 \mid X_3)}\middle| X_1 = x_1\right\}. \end{aligned} \tag{3.3}$$

REMARK 1. Condition A(vi) is also not a necessary condition for Theorem 1. It is imposed to simplify the technical proof. In the proof we approximate the matrix S_n^{-1} by a deterministic sequence. If we used a higher-order stochastic expansion of S_n^{-1}, Condition A(vi) could be weakened. Note that if the local polynomial of order d is used to approximate the function f_2 instead of using the local constant fit with a higher-order kernel, then the result of Theorem 1 continues to hold without imposing Condition A(vi) and the derivative conditions on $p_{1,2|3}(x_1, x_2 \mid x_3)$. In other words, these conditions are not essential to our estimation problem.

REMARK 2. Under the additional assumptions that X_1 has compact support $\mathscr{X}$ and that Condition A holds uniformly for $x_1 \in \mathscr{X}$ [i.e., the infimum in A(iii) is uniformly bounded from below and the derivatives considered in A(iii) and A(iv) are uniformly bounded], it is easy to show that

$$\bar{f}_1 = \frac{1}{n}\sum_{i=1}^n \hat{f}_1^*(X_{1i}) = \bar{b} + o(h_1^2) + o_P\left(\frac{1}{\sqrt{nh_1^p}}\right),$$

where $\bar{b} = \frac{1}{2}h_1^2 E \operatorname{tr}[f_1''(X_1)\mu_2(K)]$. So it follows from Theorem 1 for $\hat{f}_1 = \hat{f}_1^* - \bar{f}_1$ that

$$\sqrt{nh_1^p}\left\{\hat{f}_1(x_1) - f_1(x_1) + \bar{b} - b(x_1) + o(h_1^2)\right\} \to N(0, v(x_1)).$$

Note that the "additional bias" term $\bar{b}$ can be dropped in the preceding expression if a different bandwidth (smaller than h_1) is used to construct $\bar{f}_1$. If one can only assume that Condition A holds uniformly over a subset $\mathscr{X}'$ of $\mathscr{X}$, then one could consider $\hat{f}_1 = \hat{f}_1^* - \bar{f}_1$ with $\bar{f}_1 = \sum_{i=1}^n \gamma(X_{1i})\hat{f}_1^*(X_{1i})/\sum_{i=1}^n \gamma(X_{1i})$, where γ is a weight function that vanishes outside of $\mathscr{X}'$. Then $\hat{f}_1$ is a consistent estimate of $f_1(x_1) - E\gamma(X_1)f_1(X_1)/E\gamma(X_1)$ and its asymptotic distribution can be easily seen from Theorem 1. Our following results have similar implications. For brevity we will not mention them.

(1998) Fan, J., Härdle, W. and Mammen, E.
Direct Estimation of Low Dimensional Components in Additive Models.

Annals of Statistics, 26, 943-971

REMARK 3. An analogous result can be proved for the case that model (1.1) does not hold, that is, that the regression function $m(x) = E[Y \mid X = x]$ is not of the form $\mu + f_1(x_1) + f_{23}(x_2, x_3)$. If Condition A(iv) is replaced by the assumption that $m(u_1, u_2, u_2)$ has bounded partial derivatives up to order 2 with respect to u_1 and up to order d with respect to u_2 for u_1 in a neighborhood of x_1 and for (u_2, u_3) in the support of the weight function W, one can show that (3.1) holds with

$$b_1(x_1) = \frac{1}{2}h_1^2\mu_2(K)E\left\{\operatorname{tr}\left[\frac{\partial^2}{(\partial x_1)^2}m(x_1, X_2, X_3)\right]W(X_2, X_3)\right\}$$

and with $\mu_1 + f_1(x_1)$ replaced by $E[m(x_1, X_2, X_3)W(X_2, X_3)]$. In this case $\hat{f}_1^*$ is a consistent estimate of a weighted average effect of the covariable X_1.

REMARK 4. Due to the local linear fitting, the resulting estimate $\hat{f}_1^*(x_1)$ is automatically adapted to the boundary of the design density of X_1. This can be seen from our proof. The theoretical formulation of boundary properties of a nonparametric estimator can be found in Gasser and Müller (1979) and its applications to the local polynomial fitting is given by Fan and Gijbels (1992, 1996), and Ruppert and Wand (1994).

We now consider the optimal weight function $W(\cdot)$. This is equivalent to minimizing

$$\min_W E\left\{\sigma^2(X)\frac{p_{2\mid 3}^2(X_2 \mid X_3)W^2(X_2, X_3)}{p_{1,2\mid 3}^2(X_1, X_2 \mid X_3)} \mid X_1 = x_1\right\} \tag{3.4}$$

subject to $EW(X_2, X_3) = 1$.

We first state a simple lemma.

LEMMA 1. *The minimization problem* $\min_W \int W^2(x)g^2(x)\,dx$ *subject to* $\int W(x)h(x) = 1$ *is obtained at*

$$W = \frac{h(x)}{g^2(x)}\Big/\int\frac{h^2}{g^2}$$

and the minimum value is $\{\int h^2(x)/g^2(x)\,dx\}^{-1}$.

PROOF. Using the Language multiplier method, we have to minimize $\int W^2g^2 - \theta Wh$. This is equivalent to minimizing $W^2(x)g^2(x) - \theta W(x)h(x)$, yielding the solution

$$W(x) = \frac{\theta}{2}\frac{h(x)}{g^2(x)}.$$

The constraint $\int Wh = 1$ gives

$$W(x) = \frac{h(x)}{g^2(x)}\Big/\int\frac{h^2}{g^2}.$$

This completes the proof. □

Annals of Statistics, 26, 943-971

Applying Lemma 1 to problem (3.4), we obtain the optimal solution

$$W(X_2, X_3) = c^{-1} \frac{p_{2,3}(X_2, X_3)\, p^2_{1,2\mid 3}(x_1, X_2 \mid X_3)}{\sigma^2(x_1, X_2, X_3)\, p^2_{2\mid 3}(X_2 \mid X_3)\, p_{2,3\mid 1}(X_2, X_3 \mid x_1)} = c^{-1} \frac{p(x_1, X_2, X_3)\, p_1(x_1)}{\sigma^2(x_1, X_2, X_3)\, p_{2,3}(X_2, X_3)}, \tag{3.5}$$

where $p(x) = p_{1,2\mid 3}(x_1, x_2 \mid x_3) p_3(x_3)$ and $p_{2,3}(x) = p_{2\mid 3}(x_2 \mid x_3) p_3(x_3)$ are respectively the joint "density" of $X = (X_1, X_2, X_3)$ and (X_2, X_3) and where $c = p_1(x_1)^2 E\{\sigma^{-2}(X) \mid X_1 = x_1\}$. The minimal variance is

$$\min_W v(x_1) = \frac{\|K\|^2}{p_1(x_1)} \left[E\{\sigma^{-2}(X) \mid X_2 = x_1\} \right]^{-1}. \tag{3.6}$$

REMARK 5. The optimal weight function W depends on x_1. When it is used, the constant μ_1 [see (2.2)] depends on x_1. So in this case the estimate $\hat{f}_1^*(x_1)$ no longer estimates a function that is parallel to f_1. Nevertheless the estimate $\hat{f}_1$ is a consistent estimate of f_1. Note that for the calculation of $\hat{f}_1(x_1)$ the same weight function (depending on x_1) is used for $\hat{f}_1^*(X_{1i})$ in (2.6). Therefore the term $\mu_1 = \mu_1(x_1)$ cancels. See also Remark 2. Furthermore, as noted in Remark 2, the extra term of bias can be completely eliminated if a different bandwidth is applied to construct $\bar{f}$.

REMARK 6. Typically, the design densities $p(X)$, $p_1(X_1)$, $p_{2,3}(X_2, X_3)$ are not known. A theoretically satisfactory way out consists of dividing our sample into a relatively small first subsample and a relatively large second subsample. Then, under our smoothness assumptions, the design densities can be consistently estimated by the first subsample. The regression functions can be estimated in a second step using the other subsample. This shows that the optimal variance can be achieved, at least theoretically. The practically more relevant procedure, using the full data set for the estimation of the design densities and of the regression function, is not covered by our theory.

REMARK 7. When $f_{23}(x_2, x_3)$ is known and $\sigma^2(x) \equiv \sigma^2$, one can directly smooth $Y - f_{23}(X_2, X_3)$ on X_1 to obtain an estimate of $f_1(x_1)$ and this estimate is optimal in an asymptotic minimax sense [cf. Fan (1993)]. The variance of this estimate is $\sigma^2\|K\|^2/p_1(x_1)$, which is the same as (3.6). In other words, our direct estimator (2.6) shares the same optimality as this ideal estimator and has the same ability of estimating the additive component even if f_{23} is unknown.

(1998) Fan, J., Härdle, W. and Mammen, E.
Direct Estimation of Low Dimensional Components in Additive Models.

Annals of Statistics, 26, 943-971

REMARK 8. In the case that X_1 is independent of (X_2, X_3) and $\sigma^2(x) \equiv \sigma^2$, one can directly smooth Y on X_1 to obtain an estimate of $f_1(x_1)$ [cf. Härdle and Tsybakov (1995)]. This estimator has the asymptotic variance

$$\frac{\|K\|^2}{p_1(x_1)}\left[\sigma^2 + \operatorname{var}\{f_{23}(X_2, X_3)\}\right],$$

which is larger than our direct estimator (2.6) with the optimal weight (3.5).

To summarize, we have

THEOREM 2. *Under the assumptions of Theorem* 1, *if the ideal weight* (3.5) *is used*, *we have*

$$\sqrt{nh_1^p}\left\{\hat{f}_1^*(x_1) - f_1(x_1) - \mu_1 - b(x_1) + o\left(h_1^2\right)\right\} \to N\left(0, \frac{\|K\|^2}{p_1(x_1)}\left[E\left\{\sigma^{-2}(X) \mid X_1 = x_1\right\}\right]^{-1}\right),$$

where $b(x_1)$ *was defined in* (3.2).

4. Applications to special models.

4.1. *Additive model*. We now assume the following additive model:

$$Y = \mu + g_1(U_1) + \cdots + g_p(U_p) + \varepsilon, \tag{4.1}$$

where $g_1(\cdot), \ldots, g_p(\cdot)$ are univariate functions satisfying the identifiability condition

$$E g_1(U_1) = 0, \ldots, E g_p(U_p) = 0$$

and $U_1, \ldots, U_p$ are continuous variables having a joint density p. Now, for each variable U_α, we can form directly $\hat{g}_\alpha^*$ as in (2.6), using now $h_1 = h_{1\alpha}$ and $h_2 = h_{2\alpha}$.

THEOREM 3. *If the conditions of Theorem* 1 *hold for each component* α, *then we have the following joint asymptotic normality*:

$$\begin{pmatrix} \sqrt{nh_{11}}\left\{\hat{g}_1^*(u_1) - g_1(u_1) - \mu_{11} - \frac{1}{2}h_{11}^2\mu_2(K)g_1''(u_1) + o\left(h_{11}^2\right)\right\} \\ \vdots \\ \sqrt{nh_{1p}}\left\{\hat{g}_p^*(u_p) - g_p(u_p) - \mu_{1p} - \frac{1}{2}h_{1p}^2\mu_2(K)g_p''(u_1) + o\left(h_{1p}^2\right)\right\} \end{pmatrix} \to_d N(0, \Sigma), \tag{4.2}$$

where $\mu_{1\alpha}$ *is analogous to that defined in* (2.1) *and*

$$\Sigma = \|K\|^2 \operatorname{diag}\left(\sigma_1^2, \ldots, \sigma_p^2\right)$$

and

$$\sigma_\alpha^2(u) = E\left\{\frac{\sigma^2(U)p_\alpha(u_\alpha)p_{-\alpha}^2(U_{-\alpha})W_\alpha^2(U_{-\alpha})}{p^2(U)}\middle| U_\alpha = u_\alpha\right\},$$

with $U_{-\alpha} = (U_1, \ldots, U_{\alpha-1}, U_{\alpha+1}, \ldots, U_p)$ *and* $p_{-\alpha}$ *is its joint density.*

(1998) Fan, J., Härdle, W. and Mammen, E.
Direct Estimation of Low Dimensional Components in Additive Models.

Annals of Statistics, 26, 943-971

REMARK 9. When $W \equiv 1$, the variance matrix Σ is the same as that obtained by Chen, Härdle, Linton and Severance-Lossin (1996). However, since we employ the local linear fit, our bias has a nicer expression. Put another way, the local linear fit (2.6) uses one extra local parameter without increasing the variance. See Fan and Gijbels (1996) for further discussion on the advantages of using local polynomial fits.

REMARK 10. Under the standard assumption that all components are only two times continuously differentiable (i.e., $d = 2$) and smoothing of optimal order is done for α (i.e., $h_{1\alpha}$ is of order $n^{-1/5}$), then the conditions $nh_{1\alpha}h_{2\alpha}^{p-1}/\log n \to \infty$, $h_{2\alpha}/h_{1\alpha} \to 0$ imply $p \le 4$. [Furthermore, Condition A(vi) implies $p \le 2$. However, this condition can be weakened; see Remark 1.] So for $p \ge 5$ two times differentiable component rates of order $n^{-2/5}$ cannot be achieved by the marginal integration estimate. However, with a modification given by Hengartner (1996), the marginal integration estimate can still achieve the optimal rate of convergence.

If the ideal weight scheme (3.5) is applied to each additive component, the weight function should be

$$W_\alpha = \frac{p(U)p_\alpha(U_\alpha)}{\sigma^2(U)p_{-\alpha}(U_{-\alpha})} \Bigg/ \int \frac{p(U)p_\alpha(U_\alpha)}{\sigma^2(U)}\, dU_{-\alpha} \tag{4.3}$$

and the ideal variance is $\|K\|^2\sigma^2/p_\alpha(U_\alpha)$ if $\sigma^2(U) \equiv \sigma^2$.

4.2. *Additive partially linear model.* Consider the additive partially linear model (1.3), which possesses the flexibility to model a part of covariates (in particular, discrete variables) linearly. In this model, one can form the estimate of $g_\alpha(\cdot)$ via $\hat{g}_\alpha^*(\cdot)$ as in Section 4.1 [by treating the additional discrete variable X_3 as in Section 3]. Let

$$\theta = \sum_{\alpha=1}^{p} \mu_{1\alpha}. \tag{4.4}$$

Then $\hat{g}_1^* + \cdots + \hat{g}_p^*$ overestimates $g_1 + \cdots + g_p$ by an amount of θ. Since model (1.3) involves an intercept term, this will only affect the estimate of μ, not the slope β. Since the grand mean $\mu = EY - EX_2^T\beta$ can be estimated as

$$\hat{\mu} = \frac{1}{n}\sum_{i=1}^{n} Y_i - \frac{1}{n}\sum_{i=1}^{n} X_{3i}^T\hat{\beta}, \tag{4.5}$$

the actual value of θ is not a concern to us.

The quality of the estimator $\hat{g}_\alpha^*$ is not high at the region where the data are sparse. To eliminate such deficiencies used in the parametric estimation, we use the ith data point if $X^i = (U_{1,i}, \ldots, U_{p,i}, X_{3,i}) \in A$, where A is a

Annals of Statistics, 26, 943-971

prescribed set (usually a rectangle) in $\mathbb{R}^{p+r}$. Now consider the following least-squares problem:

$$\min_{\mu,\beta} \sum_{i=1}^{n} \left\{Y_i - \hat{g}_1^*(U_{1i}) - \cdots - \hat{g}_p^*(U_{pi}) - \mu - X_{3i}^T\beta\right\}^2 I\{X^i \in A\}. \tag{4.6}$$

Let $Z_i = \begin{pmatrix} 1 \\ X_{3i} \end{pmatrix}$ and $\tilde{Z} = (Z_1, \ldots, Z_n)^T$ be the design matrix.

Put $\Delta = \operatorname{diag}\{I\{X^1 \in A\}, \ldots, I\{X^n \in A\}\}$ and $\hat{\beta}^* = \begin{pmatrix} \hat{\mu} \\ \hat{\beta} \end{pmatrix}$. Then

$$\hat{\beta}^* = \left(\tilde{Z}^T \Delta \tilde{Z}\right)^{-1} \tilde{Z}^T \Delta\left(Y - \hat{g}_1^* - \cdots - \hat{g}_p^*\right),$$

where $Y = (Y_1, \ldots, Y_n)^T$ and $\hat{g}_\alpha^* = (g_\alpha^*(U_{\alpha 1}), \ldots, g_\alpha^*(U_{\alpha n}))^T$. To state the asymptotic normality of $\hat{\beta}^*$, we use the notation introduced in Section 4.1. Additionally, we need the following notation.

Let $p_\alpha(\cdot)$ be the marginal density of U_α and let $p_{-\alpha,3}(\cdot)$ be the marginal density of $(U_{-\alpha}, X_3)$, $\alpha = 1, \ldots, p$,

$$Z_A = Z I(X \in A) - \sum_{\alpha=1}^{p} \frac{W_\alpha(U_{-\alpha}, X_3)\, p_{-\alpha,3}(U_{-\alpha}, X_3)}{p(X)} p_\alpha(U_\alpha) E\{Z I(X \in A) \mid U_\alpha\}.$$

Put $\beta^* = \begin{pmatrix} \mu + \theta \\ \beta \end{pmatrix}$. For simplicity of discussion, we assume that $W_\alpha(\cdot)$ is independent of u_α. (Otherwise, the root-n of $\hat{\beta}^*$ holds, but the covariance is more complicated. Set

$$\begin{aligned} V_\alpha = {} & \{W_\alpha(U_{-\alpha}, X_3) - 1\} E\left[g_\alpha(U_\alpha) I\{X \in A\} Z\right] \\ & + \left\{\left(g_{-\alpha}(U_{-\alpha}) + X_3^T\beta\right) W_\alpha(U_{-\alpha}, X_3) \right. \\ & \left. - E\left[\left(g_{-\alpha}(U_{-\alpha}) + X_3^T\beta\right) W_\alpha(U_{-\alpha}, X_3)\right]\right\} E\left[I\{X \in A\} Z\right], \end{aligned}$$

where

$$g_{-\alpha}(U_{-\alpha}) = g_1(U_1) + \cdots + g_{\alpha-1}(U_{\alpha-1}) + g_{\alpha+1}(U_{\alpha+1}) + \cdots + g_p(U_p).$$

THEOREM 4. *Under the assumptions of Theorem* 3, *if* $\|X_3\|$ *has a bounded fourth moment*, $nh_{1\alpha}^2 h_{2\alpha}^{2(p-1)}/(\log n)^2 \to \infty$ *and* $h_{1\alpha} = o(n^{-1/4})$, *we have*

$$\sqrt{n}\left(\hat{\beta}^* - \beta^*\right) \to N\left(0, B_1^{-1} B_2 B_1^{-1}\right),$$

where

$$B_1 = E I(X \in A) Z Z^T$$

and

$$B_2 = E\sigma^2(X) Z_A Z_A^T + \operatorname{var}\left(\sum_{\alpha=1}^{p} V_\alpha\right).$$

When X_3 contains quite a few binary variables, the estimator (2.6) can be hard to use, since few data points are available in (2.4). For the additive

(1998) Fan, J., Härdle, W. and Mammen, E.
Direct Estimation of Low Dimensional Components in Additive Models.

Annals of Statistics, 26, 943-971

partially linear model (1.3), special care is needed. In the local step, we can replace (2.4) by

$$(4.7)\quad \sum_{i=1}^{n} \big(Y_i - a - b(U_{1i} - u_1) - X_{3i}^T \beta\big)^2 K_{h_1}(U_{1i} - u_1) L_{h_2}(X_{2i} - x_2),$$

where $X_{2i} = (U_{2i}, \ldots, U_{pi})^T$ and $x_2 = (u_2, \ldots, u_p)^T$. Note that (4.7) is obtained via the local regression model in a neighborhood of (u_1, x_2). This kind of idea appears already in Carroll, Fan, Gijbels and Wand (1997). We denote $g(u) = g_1(u_1) + \cdots + g_p(u_p)$. Let $\hat{a}$, $\hat{b}$ and $\hat{\beta}$ minimize (4.7). Then

$$\hat{g}^*(u_1, x_2) = \hat{a}$$

is a nonparametric estimator of g. Let $W(x_2)$ be a function such that $EW(X_2) = 1$ and

$$g_1^+(u_1) = \mu + g_1(u_1) + EW(X_2) f_2(X_2) = g_1(u_1) + \mu_1^*,$$

where $f_2 = g_2 + \cdots + g_p$. Then

$$(4.8)\quad \hat{g}_1^+(u_1) = n^{-1} \sum_{i=1}^{n} \hat{g}^*(u_1, X_{2i}) W(X_{2i})$$

is an estimator of $g_1^+(u_1)$, with the following asymptotic properties.

THEOREM 5. *Suppose that Condition* B *holds for* $\alpha = 1$. *Then, if* $nh_1 h_2^{p-1}/\log n \to \infty$ *and* $h_1 \to 0$ *and* $h_2^d/h_1^2 \to 0$,

$$\sqrt{nh_1}\big\{\hat{g}_1^+(u_1) - g_1(u_1) - \mu_1^* - \tfrac{1}{2}h_1^2 \mu_2(K) g_1''(u_1) + o(h_1^2)\big\} \to N(0, v^*(u_1)),$$

with $p_1, p_2, p_{1,2}$ *being the densities of* U_1, X_2 *and* $(U_1; X_2)$, *respectively*,

$$v^*(u_1) = p_1(u_1)\|K\|^2 E\left\{\sigma^2(X) \frac{W_2^2(X_2) p_2(X_2)}{p_{1,2}^2(U_1, X_2)} e_1^T \Sigma_1^{-1} \Sigma_2 \Sigma_1^{-1} e_1 \mid U_1 = u_1\right\},$$

$$\Sigma_1 = E\left\{\begin{pmatrix} 1 & X_3^T \\ X_3 & X_3 X_3^T \end{pmatrix} \middle| U_1, X_2\right\},$$

$$\Sigma_2 = \begin{pmatrix} 1 & X_3^T \\ X_3 & X_3 X_3^T \end{pmatrix}.$$

REMARK 11. If we apply the estimating procedure to each additive component of model (1.3), then the resulting estimators are asymptotically independent and normal.

Next, we estimate the parameter β. Let $\hat{\mu}^*$ and $\hat{\beta}^{**}$ minimize (4.6) with $\hat{g}_\alpha^*$ replaced by $\hat{g}_\alpha^+$. Then we can compute explicitly the asymptotic variance of $\hat{\beta}^{**}$ in a similar fashion to Theorem 4. Since the notation gets very complicated, we only state a simpler version of it.

(1998) Fan, J., Härdle, W. and Mammen, E.
Direct Estimation of Low Dimensional Components in Additive Models.

Annals of Statistics, 26, 943-971

THEOREM 6. *Under the assumptions of Theorem* 4, *we have*

$$\sqrt{n}\,(\,\hat{\beta}^{**} - \beta\,) \to N(0, B_3)$$

for some positive definite matrix B_3.

The proof of this theorem is similar to that of Theorem 4 and is omitted.

REMARK 12. Theorems 5 and 6 can be extended to the case that X_3 is continuous.

REMARK 13. For the case of one nonparametric component ($p = 1$) and continuous X_3, Speckman (1988) has shown that another method leads to an unbiased estimate of β. The approach of Speckman does not require undersmoothing [i.e., $h_{1\alpha} = o(n^{-1/4})$]. The estimate is based on the regression of $(I - M_S)Y$ onto $(I - M_S)X_3$, where M_S denotes a smoothing matrix. It is not clear to us how this approach generalizes to the case with more than one additive components. Efficient estimation of β for $p = 1$ has been considered in Bhattacharya and Zhao (1997).

4.3. *Exploring possible interactions*. Suppose one is interested in validating the additive model (1.2) by checking whether there is a nonnegligible interaction term such as $g_{12}(u_1, u_2)$. One can embed the additive model (1.2) into the model (1.4) or more generally model (1.1) with $p = 2$. Now, estimate the function $\hat{g}_{12}$ using our method. Plot $\hat{g}_{12}(\cdot; x_2)$ for a few different values of x_2. The parallelism of the plot suggests the additivity contributions of x_1 and x_2. This provides a quick and informal model diagnostic tool.

5. Simulations and an application. In a small simulation study we have compared the "indicator method" [see (2.4)] and the "linear approach" where the linear parametric part has been incorporated in the local linear smoothing [see (4.7)]. In our simulation and in the following data example we have not studied estimation of the optimal weight function W. First experience suggests that a practically working adaptation of this idea needs some further research.

We have generated 100 samples of 200 normal observations Y. Four covariates have been generated: U_1 and U_2 are normal with mean 0, variance 1 and covariance 0.4; Z_1 takes values 1, 2, 3 and 4 with probability 0.25, 0.35, 0.25 or 0.15, respectively; Z_2 takes values 0 or 1 with probability 0.2 or 0.8, respectively. The (conditional) variance of Y is $\{1 + (U_1^2 + U_2^2 + Z_1^2 + Z_2^2)^{1/2}\}/4$. The simulated regression function is $1.5 + g_1(u_1) + g_2(u_2) + \beta_1 z_1 - \beta_2 z_2$ with $g_1(u_1) = 1 - u_1^2$, $g_2(u_2) = \sin(-u_2)$, $\beta_1 = 0.3$ and $\beta_2 = -0.5$. In the estimation of the parametric components only observations have been used with $|U_1| \le 1.5$ and $|U_2| \le 1.5$; see (4.6). Bandwidths 0.3 and 0.4 have been used for the smoothing of the estimated or the nuisance nonparametric component, respectively. Table 1 shows the simulated MASE (i.e., the

(1998) Fan, J., Härdle, W. and Mammen, E.
Direct Estimation of Low Dimensional Components in Additive Models.

Annals of Statistics, 26, 943-971

TABLE 1

*Results from a small simulation study comparing the "linear approach" and the "indicator method." Two nonparametric additive components g_1 and g_2; two linear parameters β_1 and β_2; sample size $n = 200$**

		$\hat{g}_1$	$\hat{g}_2$	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{m}$
	Indicator	0.1857	0.1775	0.0096	0.0409	0.2647
	Method	(0.0609)	(0.518)			
MASE	Linear	0.2739	0.3207	0.0075	0.0393	0.5081
	Approach	(0.1450)	(0.1549)			

*In parentheses the MASE are given for the nonparametric components with summation region truncated by the 2.5% and 97.5% quantiles of the covariates.

squared error averaged over the design points). The values in parentheses are the MASE for the nonparametric components with the summation region truncated by the 2.5% and 97.5% quantiles of the covariates. These values have been added because they reflect better the behavior of the curve estimates in the middle region.

In this simulation the "indicator method" clearly shows a better performance. We conjecture that the "indicator method" may be outperformed by the "linear method" only in cases where the discrete variables take on a rather large number of different values. In the following data example we used the "indicator method."

Figure 1 contains the resulting plots from a study on the female labor supply in East Germany. A sample of 607 women with a job who live together with a partner were asked their weekly number Y of working hours. Furthermore, the following information was recorded: if the woman has children less than 16 years old (Z_1), the unemployment rate Z_2 in the "land" of the Federal Republic of Germany where she lives, the age U_1 of the woman, her wage per hour U_2, the "Treiman prestige index" of her job U_3 [see Treiman (1978)], her years U_4 of education (introduction of this covariate makes sense because of the strongly regulated system of education in the former East Germany), her rent or redemption U_5, and the monthly net income U_6 of her husband. A partial linear model for these data has been fitted. The fit has been chosen linearly in Z_1 and Z_2. The covariate Z_2 takes only five values. (There are five "lands" in the eastern part of Germany.) The other six additive components have been estimated nonparametrically. For this data set a constant weight function W has been used. Bandwidths 0.4 and 0.6 times the empirical standard deviation of the covariable have been used for the smoothing of the estimated or the nuisance nonparametric component, respectively. The resulting parametric estimates are $\beta_1 = -1.46$ and $\beta_2 = 0.52$. The resulting nonparametric fits can be found in the left frames of Figure 1. Dashed lines have been added for indicating the pointwise variance of the curve estimates. These lines differ from the curve estimates by 1.64

Annals of Statistics, 26, 943-971

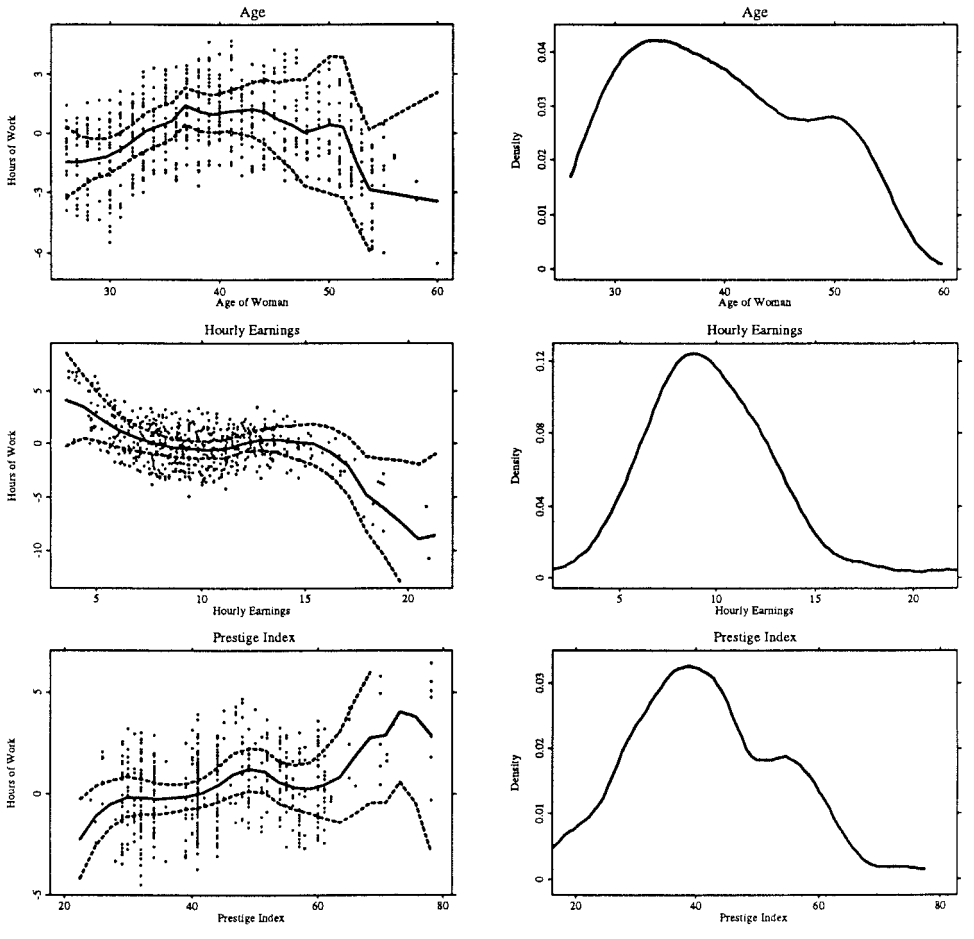


FIG. 1. *Female labor supply in East Germany. Left frames show nonparametric estimates of additive components with approximate* 90% *confidence intervals. Right frames give kernel density estimates of the covariates.*

times the (estimated) pointwise standard deviation of the curve estimates; that is, this corresponds to an approximate 90% confidence interval (without bias correction). The estimation of the pointwise standard deviation of the curve estimates has been done under the additional assumption that the conditional variance of the errors is constant. Note that all curve estimates at a fixed point are averages $\sum w_i Y_i$ of the observations Y_i. The variance of this estimate can be estimated by $\sum w_i^2 \hat{\sigma}_2$, where $\hat{\sigma}^2$ is the empirical variance of the residuals $\hat{\varepsilon} = Y - \hat{\mu} - \sum_j \hat{f}_j(U_j) - \sum_j \hat{\beta}_j Z_j$. Another estimate of the variance of $\sum w_i Y_i$ is $\sum w_i^2 \hat{\varepsilon}_i^2$. This estimate does not require the additional assumption that the conditional variance of the errors is constant. Plots of

Annals of Statistics, 26, 943-971

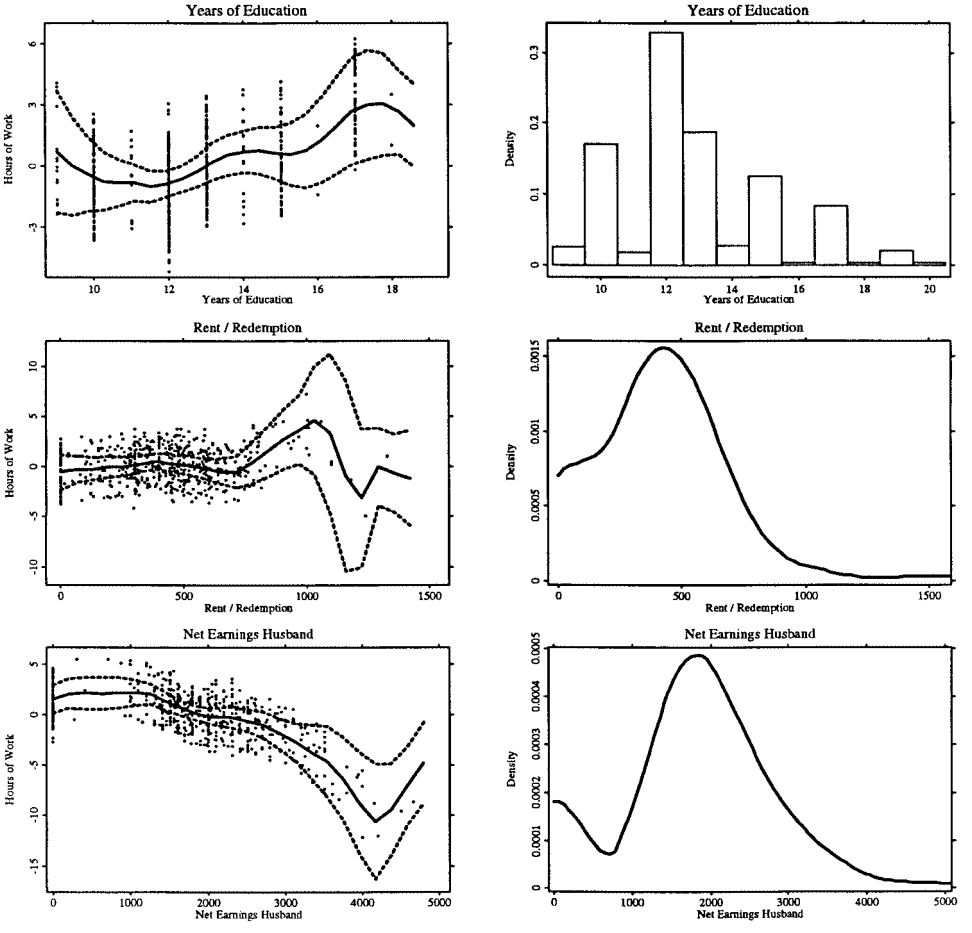


FIG. 1. *Continued*

pointwise confidence intervals based on this estimate are of similar size but of rougher shape. They are not shown here. In each plot of Figure 1 the covariable has been plotted against the estimated function plus the logarithm of the residual [i.e., $\hat{f}_\alpha(U_\alpha) + \text{sgn}(\hat{\varepsilon})\log|\hat{\varepsilon}|$; the logarithmic transform has been used to show all data]. The right frames show the density estimates of the covariates.

The plots show some clear nonlinearities. In particular, one sees a flat part in the lower range for rent and prestige index and in the middle range of hourly earnings, whereas the relation is monotone elsewhere. The results quantify the extent to which each variable affects the female labor supply. Using the dynamic ranges of the plots as a criterion to assess the practical

(1998) Fan, J., Härdle, W. and Mammen, E.
Direct Estimation of Low Dimensional Components in Additive Models.

Annals of Statistics, 26, 943-971

importance of a variable, the key factors that affect the labor supply are hourly earnings U_2 and monthly net income of husbands U_6. Slightly less influential covariates are age of the woman U_1 and prestige of the job U_2. Table 2 shows the results of a parametric least-squares analysis.

The covariates $U_1^2 = (\text{AGE W.})^2$ and $U_2^2 = (\text{WAGE P. H.})^2$ have been introduced in the parametric model. The presence of these quadratic terms is highly significant. The introduction of U_1^2 is motivated partially by the shape of the nonparametric estimate of g_1. There is no significant change in the values of the parameters β_1 and β_2. Otherwise, there are some differences between the parametric and the semiparametric analysis. Clearly, the piecewise linear shape of g_2, g_3, and g_5 cannot be recovered in the parametric model. For g_5 the sign of the estimated parameter agrees with the slope of the nonparametric estimate in the upper part. Note that for g_2 the parametric analysis with covariates U_2 and U_2^2 differs strongly for the upper part of g_2. At the boundaries of the functions g_4, g_5 and g_6 we see some differences between the parametric analysis and the semiparametric analysis. Clearly, the boundary behavior of the nonparametric estimates depends on a relatively small fraction of the observations. For example, the monotone decreasing part at the beginning of g_4, is caused by only 15 women with 9 years of education and an introduction of a covariate U_4^2 in the parametric analysis is not significant.

It seems to be difficult to verify the data analytic findings of a semiparametric analysis. A first step is to consider test statistics which are based on the comparison of parametric and nonparametric fits; see, for instance,

TABLE 2
Female labor supply in East Germany. Results of an ordinary least-squares analysis

Source	Sum of squares	Degrees of freedom	Mean square	F-ratio
Regression	6526.3	10	652.6	9.24
Residual	42,101.1	596	70.6	
R squared	= 13.4%	R squared (adjusted)	= 12.0%	

Variable	Estimate	Standard error	t-value	Probability $> \lvert t \rvert$
CONSTANT	1.36	8.95	0.15	0.8797
CHILD	−2.63	1.09	−2.41	0.0163
UNEMPLOYMENT	0.48	0.22	2.13	0.0333
AGE W.	1.63	0.43	3.75	0.0002
(AGE W.)^2	−0.021	0.0054	−3.82	0.0001
WAGE P. H.	−1.07	0.18	−6.11	≤ 0.0001
(WAGE P. H.)^2	0.0017	0.0033	4.96	≤ 0.0001
PRESTIGE	0.13	0.034	3.69	0.0002
YEARS EDUC.	0.66	0.19	3.58	0.0004
RENT/RED.	0.0018	0.0012	1.56	0.1198
NET INC. H.	−0.0016	0.0003	−4.75	≤ 0.0001

Annals of Statistics, 26, 943-971

Härdle and Mammen (1993) and Härdle, Mammen and Müller (1995). The second paper discusses also extensions to generalized regression.

6. Conditions and proofs.

CONDITION A. (i) We suppose that the functions W and f_2 are bounded on the support S of W. The weight function $W(x_2, x_3)$ is uniformly continous with respect to x_2.

(ii) The kernel functions K and L are symmetric and have bounded supports. Furthermore, L is an order-d kernel.

(iii) The support of the discrete variable X_3 is finite and

$$\inf_{\substack{u_1 \in x_1 \pm \delta \\ (x_2, x_3) \in S}} p_3(x_3)\, p_{1,2\,|\,3}(u_1, x_2 \mid x_3) > 0 \quad \text{for some } \delta > 0.$$

For u_1 in a neighborhood of x_1 and for (u_2, u_3) in S, the conditional density $p_{1,2\,|\,3}(u_1, u_2 \mid u_3)$ has bounded partial derivatives up to order 2 with respect to u_1 and up to order d with respect to u_2.

(iv) f_1 has a bounded second derivative in a neighborhood of x_1 and $f(x_2, x_3)$ has a bounded dth-order derivative with respect to x_2.

(v) $E\varepsilon^4$ is finite and $\sigma^2(x) = E(\varepsilon^2 \mid X = x)$ is continuous, where $\varepsilon = Y - E(Y \mid X)$. Furthermore, for a $\delta > 0$, the conditional absolute moment $E(|\varepsilon|^{2+\delta} | X_1 = u_1)$ is bounded for u_1 in a neighborhood of x_1.

(vi) $nh_1^p h_2^{2q} / \log^2 n \to \infty$ and $h_1^4 \log n / h_2^q \to 0$.

CONDITION B. (i) The functions $g_{-\alpha}$ and W_α are bounded on the support S_α of W_α. The weight function W_α is uniformly continuous.

(ii) The same as Condition A(ii).

(iii) $\inf p(u_1, \ldots, u_p) > 0$, where the infimum runs over $u_\alpha \in x_\alpha \pm \delta$ and $(u_1, \ldots, u_{\alpha-1}, u_{\alpha+1}, \ldots, u_p) \in S_\alpha$. For u_1 in a neighborhood of x_1 and for $(u_1, \ldots, u_{\alpha-1}, u_{\alpha+1}, \ldots, u_p) \in S_\alpha$, the density p has bounded partial derivatives up to order 2 with respect to u_α and up to order d with respect to u_β, $\beta \neq \alpha$.

(iv) g_α has bounded and continous derivatives up to order 2 and g_β, $\beta \neq \alpha$, have bounded and continous derivatives up to order d.

(v) The same as Condition A(v).

(vi) $nh_1 h_2^{2(p-1)} / \log^2 n \to \infty$ and $h_1^4 \log n / h_2^{p-1} \to 0$.

PROOF OF THEOREM 1. Let $x^i = (x_1, X_{2i}, X_{3i})$ and let E_i denote the conditional expectation given $X_i = (X_{1i}, X_{2i}, X_{3i})$. Denote by $p(x) = p_3(x_3) p_{1,2\,|\,3}(x_1, x_2 \mid x_3)$. Then, by (2.1) and Condition A(i), we have

$$n^{-1} \sum_{i=1}^{n} m(x^i) W(X_{2i}, X_{3i}) = f_1^*(x_1) + O_p(n^{-1/2}). \tag{6.1}$$

(1998) Fan, J., Härdle, W. and Mammen, E.
Direct Estimation of Low Dimensional Components in Additive Models.

Annals of Statistics, 26, 943-971

Thus,

$$\hat{f}_1^*(x_1) - f_1^*(x_1) = n^{-1}\sum_{i=1}^{n}\{\hat{m}(x^i) - m(x^i)\}W(X_{2i}, X_{3i}) + O_p(n^{-1/2}). \tag{6.2}$$

Let $\hat{r}_{ij} = m(X_j) - m(x^i) - f_1'(x_1)^T(X_{1j} - x_1)$ and let $\hat{r}_i$ be the resulting $n \times 1$ vector. Then, by (1.1) and the definition of K_n, it follows that

$$\hat{m}(x^i) - m(x^i) = e_1^T S_n^{-1}(x^i)\begin{pmatrix} 1 & \cdots & 1 \\ X_{11} - x_1 & \cdots & X_{1n} - x_1 \end{pmatrix} A(x^i)(\hat{r}_i + \tilde{\varepsilon}), \tag{6.3}$$

where $A(x)$ is a diagonal matrix with diagonal elements $A_i(x) = K_{h_1}(X_{1i} - x_1)L_{h_2}(X_{2i} - x_2)I\{X_{3i} = x_3\}$ and $\tilde{\varepsilon} = (\varepsilon_1, \cdots, \varepsilon_n)^T$ with $\varepsilon_i = Y_i - m(X_i)$. Let $H = \operatorname{diag}(1, h_1^{-1}, \ldots, h_1^{-1})$ be a $(p+1) \times (p+1)$ diagonal matrix and $a_n = \{\log n/(nh_1^p h_2^q)\}^{1/2}$. Then, owing to the uniform convergence of the kernel density estimator [cf. Stone (1993)], we have

$$\begin{aligned} n^{-1}HS_n(x)H &= n^{-1}\sum_{i=1}^{n} A_i(x)\begin{pmatrix} 1 \\ (X_{1i} - x_1)/h_1 \end{pmatrix}\begin{pmatrix} 1 \\ (X_{1i} - x_1)/h_1 \end{pmatrix}^T \\ &= EA_i(x)\begin{pmatrix} 1 \\ (X_{1i} - x_1)/h_1 \end{pmatrix}\begin{pmatrix} 1 \\ (X_{1i} - x_1)/h_1 \end{pmatrix}^T + O_p(a_n) \\ &= \begin{pmatrix} p(x) & h_1 p^{(1,0)}(x)^T\mu_2(K) \\ h_1\mu_2(K)p^{(1,0)}(x) & p(x)\mu_2(K) \end{pmatrix} + O_p(c_n) \\ &= \begin{pmatrix} p(x) & 0 \\ 0 & p(x)\mu_2(K) \end{pmatrix} + o_p(c_n), \end{aligned} \tag{6.4}$$

where $c_n = h_1^2 + h_2^d + a_n$ and where $p^{(1,0)}$ denotes the vector of partial derivatives of p with respect to x_1. Now note that

$$\begin{aligned} &\begin{pmatrix} p(x) & h_1 p^{(1,0)}(x)^T\mu_2(K) \\ h_1\mu_2(K)p^{(1,0)}(x) & p(x)\mu_2(K) \end{pmatrix}^{-1} \\ &\quad = \begin{pmatrix} p(x) & 0 \\ 0 & p(x)\mu_2(K) \end{pmatrix}^{-1} \\ &\qquad + \frac{h_1}{p(x)}\begin{pmatrix} 0 & p^{(1,0)}(x)^T\mu_2(K) \\ p^{(1,0)}(x)\mu_2(K) & 0 \end{pmatrix} + O_p(h_1^2). \end{aligned}$$

Annals of Statistics, 26, 943-971

A similar argument to the above leads to the following uniform results:

$$\begin{aligned} & n^{-1}H\begin{pmatrix} 1 & \cdots & 1 \\ X_{11}-x_1 & \cdots & X_{1n}-x_1 \end{pmatrix} A(x^i)\hat{r}_i \\ &= n^{-1}\sum_{j=1}^{n} A_j(x^i)\hat{r}_{ij}\begin{pmatrix} 1 \\ (X_{1j}-x_1)/h_1 \end{pmatrix} \\ &= E_i A_j(X^i)\hat{r}_{ij}\begin{pmatrix} 1 \\ (X_{1j}-x_1)/h_1 \end{pmatrix} + O_p(a_n) \\ &= O_p(c_n), \end{aligned}$$

where in the third expression j is an arbitrary index with $j \neq i$ and

$$\begin{aligned} & n^{-1}H\begin{pmatrix} 1 & \cdots & 1 \\ X_{11}-x_1 & \cdots & X_{1n}-x_1 \end{pmatrix} A(x^i)\tilde{\varepsilon} \\ &= n^{-1}\sum_{j=1}^{n} A_j(x^i)\varepsilon_j\begin{pmatrix} 1 \\ (X_{1j}-x_1)/h_1 \end{pmatrix} \\ &= O_p(a_n). \end{aligned}$$

Substituting all of the above expressions into (6.3), after some algebra, we obtain

$$\begin{aligned} & \hat{m}(x^i) - m(x^i) \\ &= e_1^T\left[\begin{pmatrix} p(x^i) & 0 \\ 0 & p(x^i)\mu_2(K) \end{pmatrix}^{-1}\right. \\ &\qquad \left. + \frac{h_1}{p(x^i)}\begin{pmatrix} 0 & p^{(1,0)}(x^i)^T\mu_2(K) \\ p^{(1,0)}(x^i)\mu_2(K) & 0 \end{pmatrix} + O_p(c_n)\right] \\ &\qquad \times n^{-1}\sum_{j=1}^{n} A_j(x^i)\big(\hat{r}_{ij}+\varepsilon_j\big)\begin{pmatrix} 2 \\ (X_{1j}-x_1)/h_1 \end{pmatrix} \\ &= n^{-1}\sum_{\substack{j=1 \\ j\neq i}}^{n} A_j(x^i)\big(\hat{r}_{ij}+\varepsilon_j\big)/p(x^i) \\ &\qquad + n^{-1}\sum_{\substack{j=1 \\ j\neq i}}^{n} A_j(x^i)\big(\hat{r}_{ij}+\varepsilon_j\big)p^{-1}(x^i)p^{(1,0)}(x^i)^T\mu_2(K)(X_{1j}-x_1) \\ &\qquad + O_p\big(c_n^2\big). \end{aligned}$$

Clearly, the second term will be smaller than the first one by an order of $O(h_1)$. When the above O_p-term is averaged in (6.2), it is still of the order

$$O_p\big(c_n^2\big) = o_p\big((nh_1^p)^{-1/2}\big)$$

Annals of Statistics, 26, 943-971

by the conditions on the bandwidths. Furthermore, by calculation of the first two moments one can show that

$$n^{-2}\sum_{j\neq i} W(X_{2i}, X_{3i})A_j(x^i)\hat{r}_{ij}p^{-2}(x^i)p^{(1,0)}(x^i)^T\mu_2(K)(X_{1j}-x_1)$$
$$= o(h_1^2) + O(h_2^d) + o_p\big((nh_1^p)^{-1/2}\big)$$

and

$$n^{-2}\sum_{j\neq i} W(X_{2i}, X_{3i})A_j(x^i)\varepsilon_j p^{-2}(x^i)p^{(1,0)}(x^i)^T\mu_2(K)(X_{1j}-x_1)$$
$$= o_p\big((nh_1^p)^{-1/2}\big).$$

In other words, the approximation error from (6.4) is negligible.

Note that, for $j \neq i$,

$$E_i A_j(x^i)\hat{r}_{ij} = \tfrac{1}{2}h_1^2\mathrm{tr}\{f_1''(x_1)\mu_2(K)\}p(x^i) + o(h_1^2) + O(h_2^d).$$

Let $\tilde{r}_{ij} = A_j(x^i)\hat{r}_{ij} - E_iA_j(x^i)\hat{r}_{ij}$ for $j \neq i$ and $\tilde{r}_{ij} = 0$ for $j = i$. Thus, by (6.2), we have

$$\hat{f}_1^*(x_1) - f_1^*(x_1) = \tfrac{1}{2}h_1^2\mathrm{tr}\{f_1''(x_1)\mu_2(K)\} + o(h_1^2) + T_{n,1} + T_{n,2} + o_p\big\{(nh_1^p)^{-1/2}\big\}, \tag{6.5}$$

where

$$T_{n,1} = n^{-1}\sum_{j\neq i}\varepsilon_j K_{h_1}(X_{1j}-x_1)\Gamma(X_{2i}, X_{3i})L_{h_2}(X_{2j}-X_{2i})I\{X_{3j}=X_{3i}\}$$

and

$$T_{n,2} = n^{-2}\sum_{j\neq i}\Gamma(X_{2i}, X_{3i})\tilde{r}_{ij},$$

with

$$\Gamma(X_{2i}, X_{3i}) = W(X_{2i}, X_{3i})/p(x^i).$$

We will show that with $\varepsilon_j^* = G(X_{2j}, X_{3j})\varepsilon_j$, $G(X_{2j}, X_{3j}) = \Gamma(X_{2j}, X_{3j})p_{2,3}(X_{2j}, X_{3j})$,

$$T_{n,1} = n^{-1}\sum_{j=1}^{n} K_{h_1}(X_{1j}-x_i)\varepsilon_j^* + o_p\big((nh_1^p)^{-1/2}\big) \tag{6.6}$$

and

$$T_{n,2} = o_p\big((nh_1^p)^{-1/2}\big). \tag{6.7}$$

Combination of (6.5)–(6.7) leads to

$$\hat{f}_1^*(x_1) - f_1^*(x_1) = \tfrac{1}{2}h_1^2\mathrm{tr}\{f_1''(x_1)\mu_2(K)\} + n^{-1}\sum_{j=1}^{n}\varepsilon_j^* K_{h_1}(X_{1j}-x_1) + o_p\big\{(nh_1^p)^{-1/2}\big\}. \tag{6.8}$$

Annals of Statistics, 26, 943-971

It is easy to show that

$$\sqrt{nh_1^p}\, n^{-1} \sum_{j=1}^{n} \varepsilon_j^* K_{h_1}(X_{1j} - x_1) \to \mathcal{N}(0, v(x_1)) \tag{6.9}$$

by checking the Lyapounov condition. By using (6.8) and (6.9), we establish Theorem 1. It remains to verify (6.7) and (6.8).

PROOF OF (6.6). Let

$$V_{i,j} = \Gamma(X_{2i}, X_{3i}) L_{h_2}(X_{2j} - X_{2i}) I\{X_{3i} = X_{3j}\} - p_3(X_{3j}) p_{2|3}(X_{2j} \mid X_{3j}) \Gamma(X_{2j}, X_{3j}).$$

Note that, for $i \neq j$,

$$E_j V_{i,j} = p_3(X_{3j}) \int \Gamma(x_2, X_{3j}) L_{h_2}(x_2 - X_{2j}) p_{2|3}(x_2 \mid X_{3j})\, dx_2 - p_3(X_{3j}) p_{2|3}(X_{2j} \mid X_{3j}) \Gamma(X_{2j}, X_{3j}).$$

Thus,

$$|E_j V_{i,j}| \leq \int \left| \Gamma(X_{2j} + h_2 u, X_{3j}) p_{2|3}(X_{2j} + h_2 u \mid X_{3j}) - \Gamma(X_{2j}, X_{3j}) p_{2|3}(X_{2j} \mid X_{3j}) \right| |L(u)|\, du \to 0.$$

Note also that the difference between the left-hand side of (6.6) and the main term on the right-hand side of (6.6) can be expressed as

$$D_{n,1} = n^{-2} \sum_{j \neq i} \varepsilon_j K_{h_1}(X_{1j} - x_1) V_{i,j}.$$

To prove (6.6), it suffices to show

$$ED_{n,1}^2 = o\left((nh_1^p)^{-1}\right).$$

It follows from direct expansion that

$$ED_{n,1}^2 = n^{-4} \sum_{i \neq j;\, k \neq l} E\varepsilon_j K_{h_1}(X_{1j} - x_1) V_{i,j} \varepsilon_l K_{h_1}(X_{1l} - x_1) V_{k,l}.$$

Because of $E\{\varepsilon_j \mid X_j\} = 0$ we have

$$ED_{n,1}^2 = n^{-4} \sum_{i \neq j;\, k \neq j} E\varepsilon_j^2 K_{h_1}(X_{1j} - x_1) V_{i,j} \varepsilon_l K_{h_1}(X_{1l} - x_1) V_{k,j}.$$

For $i = k$ the order of summands on the right-hand side is at most $O(h_1^{-p} h_2^{-q})$. Because of $n^{-2} h_1^{-p} h_2^{-q} = o(n^{-1})$, we have

$$\begin{aligned} ED_{n,1}^2 &= n^{-4} \sum_{i \neq j \neq k \neq i} E\varepsilon_j^2 K_{h_1}^2(X_{1j} - x_1) V_{i,j} V_{k,j} + o(n^{-1}) \\ &= n^{-4} \sum_{i \neq j \neq k \neq i} E\varepsilon_j^2 K_{h_1}^2(X_{1j} - x_1) E_j V_{i,j} E_j V_{k,j} + o(n^{-1}) \\ &= o\left(n^{-1} h_1^{-p}\right). \end{aligned}$$

Annals of Statistics, 26, 943-971

PROOF OF (6.7). The claim follows from $ET_{n,2}^2 = o((nh_1^p)^{-1})$. Note that $E_i\tilde{r}_{i,j} = 0$. Therefore, for the calculation of $ET_{n,2}^2$ we need only consider terms of the form

$$n^{-4}E\Gamma(X_{2i}, X_{3i})\tilde{r}_{i,j}\Gamma(X_{2k}, X_{3k})\tilde{r}_{k,l},$$

where $i \neq j, k \neq l$, $j \in \{k, l\}$ and $l \in \{i, j\}$. It is easy to bound the summands for two different indices. For three different indices we have $j = l$ and $i \neq j \neq k \neq i$. Note now that for this case

$$E\Gamma(X_{2i}, X_{3i})\tilde{r}_{i,j}\Gamma(X_{2k}, X_{3k})\tilde{r}_{k,j} = O\big(h_1^{-p}\big[h_1^4 + h_2^2\big]\big).$$

Here we have used that the random variables $\hat{r}_{i,j}$ are always bounded by a constant which is of order $O(h_1^2 + h_2)$. Thus,

$$ET_{n,2}^2 = O\big(n^{-1}h_1^{-p}\big[h_1^4 + h_2^2\big]\big) = o\big((nh_1^p)^{-1}\big),$$

verifying (6.7).

PROOF OF THEOREM 3. By (6.5)–(6.7), each component of $\hat{g}_\alpha^*(u_\alpha)$ has the following stochastic representation:

$$\begin{aligned}\hat{g}_\alpha^*(u_\alpha) - g_\alpha(u_\alpha) - \mu_{1\alpha} &= \tfrac{1}{2}h_{1\alpha}^2\mu_2(K)g_\alpha''(u_\alpha) + o\big(h_{1\alpha}^2\big) \\ &\quad + n^{-1}\sum_{j=1}^n K_{h_{1\alpha}}(U_{\alpha j} - U_\alpha)G_\alpha(U_{-\alpha j})\varepsilon_j + o_p\big((nh_{1\alpha})^{-1/2}\big),\end{aligned} \tag{6.10}$$

where

$$G_\alpha(U_{-\alpha j}) = \frac{W_\alpha(U_{-\alpha j})p_{-\alpha}(U_{-\alpha j})}{p(U_\alpha^j)},$$

with $U_\alpha^j = (U_{1j}, \ldots, U_{\alpha-1,j}, u_\alpha, U_{\alpha+1,j}, \ldots, U_{pj})$. For $\alpha \neq \beta$, the covariance for the stochastic terms in (6.10) is

$$\begin{aligned}&\operatorname{cov}\left(n^{-1}\sum_{j=1}^n K_{h_{1\alpha}}(U_{\alpha j} - u_\alpha)G_\alpha(U_{-\alpha j})\varepsilon_j, n^{-1}\sum_{j=1}^n K_{h_{1\beta}}(U_{\beta j} - u_\beta)G_\beta(U_{-\beta j})\varepsilon_j\right) \\ &\quad = n^{-1}E\big[K_{h_{1\alpha}}(U_\alpha - u_\alpha)G_\alpha(U_{-\alpha})K_{h_{1\beta}}(U_\beta - u_\beta)G_\beta(U_{-\beta})\varepsilon^2\big] \\ &\quad = O(n^{-1}) = o\big((nh_{1\alpha})^{-1/2}(nh_{1\beta})^{-1/2}\big).\end{aligned}$$

Therefore, the asymptotic covariance should be 0.

Annals of Statistics, 26, 943-971

PROOF OF THEOREM 4. We only outline the key steps of the proof. Proceeding as in the proof of Theorem 1, one shows first that

$$\begin{aligned}\frac{1}{n}\sum_{i=1}^{n} Z_i I(X^i \in A)\big[\hat{g}_\alpha^*(U_{\alpha i}) - g_\alpha(U_{\alpha i}) - \mu_{1\alpha}\big] &= \frac{1}{n}\sum_{i=1}^{n} Z_i I(X^i \in A)\frac{1}{n}\sum_{j\neq k} K_{h_{1\alpha}}(U_{\alpha j} - U_{\alpha i}) \\ &\quad \times L_{h_{2\alpha}}(U_{-\alpha j} - U_{-\alpha k}) I(X_{3j} = X_{3k}) \\ &\quad \times \big[m(X^j) - m(U_{\alpha i}, U_{-\alpha k}, X_{3k}) \\ &\qquad - g'_\alpha(U_{\alpha i})(U_{\alpha j} - U_{\alpha i}) + \varepsilon_k\big] \\ &\quad \times \frac{1}{p(U_{\alpha i}, U_{-\alpha k}, X_{3k})} + O_P(c_n^2) + o_p(n^{-1/2}),\end{aligned}$$

where now $c_n = h_{1\alpha}^2 + h_{2\alpha}^d + \{\log n/(nh_{1\alpha}h_{2\alpha}^{p-1})\}^{1/2}$. Note that under our assumptions we have $c_n^2 = o(n^{-1/2})$. By considering the first two moments of the difference, one can show that (see also the asymptotic treatment of $T_{n,1}$ and $T_{n,2}$ in the proof of Theorem 1)

$$\begin{aligned}\frac{1}{n}\sum_{i=1}^{n} Z_i I(X^i \in A)\big[\hat{g}_\alpha^*(U_{\alpha i}) - g_\alpha(U_{\alpha i}) - \mu_{1\alpha}\big] &= n^{-2}\sum_{i=1}^{n}\sum_{j=1}^{n} Z_i I(X^i \in A) K_{h_{1\alpha}}(U_{\alpha j} - U_{\alpha i}) \\ &\quad \times G_\alpha(U_{\alpha i}, U_{-\alpha j}, X_{3,j})\varepsilon_j + o_p(n^{-1/2}),\end{aligned} \tag{6.11}$$

where

$$G_\alpha(U_{\alpha i}, U_{-\alpha j}, X_{3,j}) = \frac{W_\alpha(U_{-\alpha j}, X_{3,j}) p_{-\alpha,3}(U_{-\alpha j}, X_{3,j})}{p(U_{\alpha i}, U_{-\alpha j}, X_{3,j})}.$$

Thus, the main term of $\hat{\beta}^*$ is

$$\hat{\beta}^* = \beta^* + (\tilde{Z}^T \Delta \tilde{Z})^{-1} \tilde{Z}^T \Delta \bar{\varepsilon} + o_p(n^{-1/2}), \tag{6.12}$$

where the ith element of $\bar{\varepsilon}$ is

$$\begin{aligned}\bar{\varepsilon}_i = \varepsilon_i - n^{-1}\sum_{\alpha=1}^{p}\sum_{j=1}^{n} K_{h_{1\alpha}}(U_{\alpha j} - U_{\alpha i}) \\ \times G_\alpha(U_{\alpha i}, U_{-\alpha j}, X_{3,j})\varepsilon_j - \sum_{\alpha=1}^{p} \varepsilon_{\alpha i}^*.\end{aligned} \tag{6.13}$$

Obviously, by the law of large numbers,

$$n^{-1}\tilde{Z}^T \Delta \tilde{Z} = E\big[I(X \in)ZZ^T\big] + o_p(1) = B_1 + o_p(1). \tag{6.14}$$

Annals of Statistics, 26, 943-971

Let $\varepsilon^*_{\alpha i} = g_\alpha(U_{\alpha i})a_{n,\alpha} + b_{n,\alpha}$ be the approximation error in (6.1), where

$$a_{n,\alpha} = n^{-1}\sum_{j=1}^{n} W_\alpha(U_{-\alpha j}, X_{3j}) - 1$$

and

$$b_{n,\alpha} = n^{-1}\sum_{j=1}^{n}\left\{g_{-\alpha}(U_{-\alpha}) + X_{3j}^T\beta\right\}W_\alpha(U_{-\alpha j}, X_{3j}) - E\left[\left\{g_{-\alpha}(U_{-\alpha j}) + X_2^T\beta\right\}W_\alpha(U_{-\alpha}, X_3)\right].$$

We need only consider the term

$$\begin{aligned} n^{-1}\tilde{Z}^T\Delta\bar{\varepsilon} &= n^{-1}\sum_{i=1}^{n} Z_i\bar{\varepsilon}_i I\{X^i \in A\} \\ &= n^{-1}\sum_{i=1}^{n} Z_i\varepsilon_i I\{X^i \in A\} \\ &\quad - n^{-1}\sum_{\alpha=1}^{p}\sum_{i=1}^{n} Z_i I\{X^i \in A\}(g_\alpha(U_{\alpha i})a_{n,\alpha} + b_{n,\alpha}) \\ &\quad - n^{-1}\sum_{i=1}^{n}\varepsilon_i n^{-1}\sum_{\alpha=1}^{p}\sum_{j=1}^{n} G_\alpha(U_{\alpha j}, U_{-\alpha i}, X_{3,i}) \\ &\qquad \times K_{h_{1\alpha}}(U_{\alpha j} - U_{\alpha i})I\{X^j \in A\}. \end{aligned} \tag{6.15}$$

By using the same argument as in the proof of (6.6), we can show that

$$\begin{aligned} &n^{-1}\sum_{i=1}^{n}\varepsilon_i n^{-1}\sum_{j=1}^{n} G_\alpha(U_{\alpha j}, U_{-\alpha i}, X_{3,i})K_{h_{1\alpha}}(U_{\alpha j} - U_{\alpha i})Z_j I\{X^j \in A\} \\ &= n^{-1}\sum_{i=1}^{n}\varepsilon_i G_\alpha(U_{\alpha i}, U_{-\alpha i}, X_{3,i}) \\ &\qquad \times E\{Z_i I(X^i \in A) \mid U_{\alpha i}\}p_\alpha(U_{\alpha i}) + o_p(n^{-1/2}). \end{aligned} \tag{6.16}$$

Let $Z_{i,A} = Z_i I\{X^i \in A\} - \sum_{\alpha=1}^{p} G_\alpha(U_{\alpha i}, U_{-\alpha i}, X_{3,i})E\{Z_i I(X^i \in A) \mid U_{\alpha i}\}p_\alpha(U_{\alpha i})$. Then, by combining (6.15) and (6.16), we obtain

$$\begin{aligned} n^{-1/2}\tilde{Z}^T\Delta\bar{\varepsilon} &= n^{-1/2}\sum_{i=1}^{n} Z_{i,A}\varepsilon_i \\ &\quad - n^{-1/2}\sum_{\alpha=1}^{p} E\left[Z I\{X \in A\}g_\alpha(U_\alpha)\right]a_{n,\alpha} \\ &\quad + E\left[Z I\{X \in A\}\right]b_{n,\alpha} + o_p(1) \\ &\to N\left(0, E\left[\varepsilon_i^2 Z_{i,A}Z_{i,A}^T\right] + \operatorname{var}\left(\sum_{\alpha=1}^{p} V_\alpha\right)\right). \end{aligned} \tag{6.17}$$

By conditioning on X^i, one can easily see that the covariance matrix in (6.17) is B_2. Combination of (6.12), (6.14) and (6.17) shows the statement of Theorem 4.

Annals of Statistics, 26, 943-971

PROOF OF THEOREM 5. The main ideas of the proof are the same as those of Theorem 1. Thus, we only indicate the main steps. Let $x^i = (u_1, X_{2i})$. Then we have

$$n^{-1}\sum_{i=1}^{n} m(x^i)W(X_{2i}) = g_1^+(u_1) + O_p(n^{-1/2}) \tag{6.18}$$

and

$$\hat{g}_1^+(u_1) - g_1^+(u_1) = n^{-1}\sum_{i=1}^{n}\{\hat{m}^*(x^i) - m(x^i)\}W(X_{2i}) + O_p(n^{-1/2}). \tag{6.19}$$

Set $A_j(u) = K_{h_1}(U_{1j} - u_1)L_{h_2}(X_{2j} - x_2)$. Let X be the design matrix of (4.7) and $A(u) = \operatorname{diag}(A_1(u), \ldots, A_n(u))$ be the corresponding weight matrix.

Denote by

$$\hat{r}_{ij} = g_1(U_{1j}) - g_1(u_1) - g_1'(u_1)(U_{1j} - u_1) + f_2(X_{2j}) - f_2(X_{2i}),$$

where $f_2(x_2) = g_2(u_2) + \cdots + g_p(u_p)$. Let $\hat{r}_i$ be the resulting $(n \times 1)$ vector. Then

$$\hat{g}^*(x^i) - g(x^i) = e_1^T S_n^{-1}(x^i)X^T A(x^i)(\hat{r}_i + \tilde{\varepsilon}), \tag{6.20}$$

where $S_n(x) = X^T A(x^i)X$. For $u = (u_1, x_2)$ let

$$S(u) = E\left\{\begin{pmatrix} 1 & 0 & X_3^T \\ 0 & \mu_2(K) & 0 \\ X_3 & 0 & X_3X_3^T \end{pmatrix}\middle| U_1 = u_1,\ X_2 = x_2\right\}$$

and

$$H = \begin{pmatrix} 1 & & & & \\ & h_1^{-1} & & & \\ & & 1 & & \\ & & & \ddots & \\ & & & & 1 \end{pmatrix}.$$

With the same ideas as in the proof of Theorem 1, one gets an expansion of $[n^{-1}HS_n(u)H]^{-1}$ up to error terms of order $O_P(c_n)$ where, as in the proof of Theorem 4, $c_n = h_{1\alpha}^2 + h_{2\alpha}^d + \{\log n/(nh_{1\alpha}h_{2\alpha}^{p-1})\}^{1/2}$. In particular, we have that

$$n^{-1}HS_n(u)H = p(u)S(u) + o_p(1) \tag{6.21}$$

uniformly in u. Direct calculation yields

$$n^{-1}E_iHX^TA(x^i)\hat{r}_i = \left\{\tfrac{1}{2}h_1^2 g_1''(u_1)\mu_2(K) + o(h_1^2) + O(h_2^d)\right\}p(x^i)\begin{pmatrix} 1 \\ 0 \\ E(X_{3i} \mid X^i) \end{pmatrix}. \tag{6.22}$$

Annals of Statistics, 26, 943-971

Note now that

$$S(x^i)\begin{pmatrix}1\\0\\0\end{pmatrix} = \begin{pmatrix}1\\0\\E(X_{3i}\mid x^i)\end{pmatrix}.$$

Therefore,

$$e_1^T S^{-1}(x^i)\begin{pmatrix}1\\0\\E(X_{3i}\mid x^i)\end{pmatrix} = 1. \tag{6.23}$$

Substituting the higher-order expansion of $[n^{-1}HS_n(u)H]^{-1}$ and (6.22) into (6.20), we obtain with (6.23) that

$$\begin{aligned}\hat g^*(x^i) - g(x^i) &= \tfrac{1}{2}h_1^2 g_1''(u_1)\mu_2(K) + o(h_1^2) + O(h_2^d)\\ &\quad + n^{-1}\sum_{j=1}^n p(x^i)^{-1} e_1^T S^{-1}(x^i) A_j(x^i)\begin{pmatrix}1\\(U_{j1}-u_1)/h_1\\X_{3j}\end{pmatrix}\varepsilon_j \qquad (6.24)\\ &\quad + n^{-1}\sum_{j=1}^n p(x^i)^{-1} e_1^T S^{-1}(x^i)\tilde r_{ij} + O_p(n^{-1/2}),\end{aligned}$$

where

$$\tilde r_{ij} = A_j(x^i)\hat r_{ij}\begin{pmatrix}1\\(U_{1j}-u_1)/h_1\\X_{3j}\end{pmatrix} - E_i A_j(x^i)\hat r_{ij}\begin{pmatrix}1\\(U_{1j}-u_1)/h_1\\X_{3j}\end{pmatrix}.$$

Note that again we obtain that the expansion of the estimate depends only on the first-order approximation (6.21) $n^{-1}HS_n(u)H$.

Using the same argument as in the proof of Theorem 1, the average of the last term in (6.24) over i is of order $o_p(n^{-1/2})$. Thus, by (6.19) and (6.24), we have

$$\begin{aligned}\hat g_1^+(u_1) - g_1^+(u_1) &= \tfrac{1}{2}h_1^2 g_1''(u_1)\mu_2(K) + o(h_1^2) + o(h_2^d)\\ &\quad + n^{-2}\sum_{i=1}^n\sum_{j=1}^n p(x^i)^{-1} e_1^T S^{-1}(x^i) A_j(x^i) W(X_{2i})\\ &\qquad \times\begin{pmatrix}1\\(U_{j1}-u_1)/h_1\\X_{3j}\end{pmatrix}\varepsilon_j + o_p(n^{-1/2}).\end{aligned}$$

Annals of Statistics, 26, 943-971

By the projection argument which we used when treating $T_{n,1}$ in the proof of Theorem 1, we obtain

(6.25)
$$\hat{g}_1^+(u_1) - g_1^+(u_1) = \tfrac{1}{2}h_1^2 g_1''(u_1)\mu_2(K) + n^{-1}\sum_{j=1}^{n} K_{h_1}(U_{1j} - u_1)\varepsilon_j^* + o_p\big((nh_1)^{-1/2}\big),$$

where

$$\varepsilon_j^* = \frac{\varepsilon_j p_2(X_{2j}) e_1^T S^{-1}(x^j) W(X_{2j})}{p(x^j)} \begin{pmatrix} 1 \\ (U_{j1} - u_1)/h_1 \\ X_{3j} \end{pmatrix}.$$

Therefore, by checking the Lyapounov condition, we can establish Theorem 5, where the variance is obtained from (6.25) along with some algebra.

Acknowledgments. The manuscript was completed while Fan was visiting the Department of Statistics, the Chinese University of Hong Kong, and he is grateful for their hospitality. Furthermore, the authors would like to thank S. Sperlich for computational assistance and S. Profit for helpful remarks.

REFERENCES

BERNDT, E. R. (1991). *The Practice of Econometrics: Classic and Contemporary*. Addison-Wesley, Reading, MA.

BHATTACHARYA, P. K. and ZHAO, P.-L. (1997). Semiparametric inference in a partial linear model. *Ann. Statist.* **25** 244–262.

BUJA, A., HASTIE, T. J. and TIBSHIRANI, R. J. (1989). Linear smoothers and additive models (with discussion). *Ann. Statist.* **17** 453–510.

CARROLL, R. J., FAN, J., GIJBELS, I. and WAND, M. P. (1997). Generalized partially linear single-index models. *J. Amer. Statist. Assoc.* **92** 477–489.

CHEN, R., HÄRDLE, W., LINTON, O. and SEVERANCE-LOSSIN, E. (1996). Estimation and variable selection in additive nonparametric regression models. In *Statistical Theory and Computational Aspects of Smoothing* (W. Härdle and M. Schimek, eds.). Physika, Heidelberg.

FAN, J. (1993). Local linear regression smoothers and their minimax efficiency. *Ann. Statist.* **21** 196–216.

FAN, J. (1997). Comments on "Polynomial splines and their tensor product in the extended linear models" by C. J. Stone, M. H. Hansen, C. Kooperberg and Y. U. Troung. *Ann. Statist.* **25** 1425–1432.

FAN, J. and GIBELS, I. (1992). Variable bandwidth and local linear regression smoothers. *Ann. Statist.* **20** 2008–2036.

FAN, J. and GIJBELS, I. (1996). *Local Polynomial Modeling and Its Applications*. Chapman and Hall, London.

FRANZ, W. (1991). *Arbeitsökonomik*. Springer, Berlin.

GASSER, T. and MÜLLER, H.-G. (1979). Kernel estimation of regression functions. *Smoothing Techniques for Curve Estimation. Lecture Notes in Math.* **757** 23–68. Springer, New York.

HÄRDLE, W. and MAMMEN, E. (1993). Testing parametric versus nonparametric regression. *Ann. Statist.* **21** 1926–1947.

Annals of Statistics, 26, 943-971

HÄRDLE, W., MAMMEN, E. and MÜLLER, M. (1995). Testing parametric versus semiparametric modelling in generalized linear models. Technical Report.

HÄRDLE, W. and TSYBAKOV, A. B. (1995). Additive nonparametric regression on principal components, *J. Nonparametr. Statist.* **5** 157–184.

HASTIE, T. J. and TIBSHIRANI, R. J. (1990). *Generalized Additive Models*. Chapman and Hall, London.

HENGARTNER, N. W. (1996). Rate optimal estimation of additive regression via the integration method in the presence of many covariates. Unpublished manuscript.

LINTON, O. B. (1997). Efficient estimation of additive nonparametric regression models. *Biometrika* **84** 469–473.

LINTON, O. B., MAMMEN, E. and NIELSEN, J. P. (1997). The existence and asymptotic properties of a backfitting projection algorithm under weak conditions. Preprint.

LINTON, O. B. and NIELSEN, J. P. (1995). A kernel method of estimating structured nonparametric regression based on marginal integration. *Biometrika* **82** 93–101.

MACK, Y. P. and SILVERMAN, B. W. (1982). Weak and strong uniform consistency of kernel regression estimates. *Z. Wahrsch. Verw. Gebiete* **61** 405–415.

OPSOMER, J. D. (1997). On the existence and asymptotic properties of backfitting estimators. Preprint.

OPSOMER, J. D. and RUPPERT, D. (1997). Fitting a bivariate additive model by local polynomial regression. *Ann. Statist.* **25** 186–211.

RUPPERT, D. and WAND, M. P. (1994). Multivariate weighted least squares regression. *Ann. Statist.* **22** 1346–1370.

SPECKMAN, P. (1988). Kernel smoothing in partial linear models. *J. Roy. Statist. Soc. Ser. B* **50** 413–436.

STONE, C. J. (1983). Optimal uniform rate of convergence for nonparametric estimators of a density function or its derivatives. In *Recent Advances in Statistics: Papers Presented in Honor of Herman Chernoff's Sixtieth Birthday* (M. H. Rizvi, J. S. Rustagi and D. Siegmund, eds.). Academic Press, New York.

STONE, C. J. (1985). Additive regression and other nonparametric models. *Ann. Statist.* **13** 685–705.

STONE, C. J. (1986). The dimensionality reduction principle for generalized additive models. *Ann. Statist.* **14** 592–606.

TJØSTHEIM, D. and AUESTAD, B. H. (1994). Nonparametric identification of nonlinear time series: projections. *J. Amer. Statist. Assoc.* **89** 1398–1409.

TREIMAN, D. J. (1978). Probleme der Begriffsbildung und Operationalisierung in der international vergleichenden Mobilitätsforschung. In *Sozialstrukturanalysen mit Umfragedaten* (F. U. Pappi, ed.). Athenäum, Kronberg im Taunus.

DEPARTMENT OF STATISTICS
UNIVERSITY OF NORTH CAROLINA
CHAPEL HILL, NORTH CAROLINA 27599-3260

INSTITUT FÜR STATISTIK UND ÖKONOMETRIE
WIRTSCHAFTSWISSENSCHAFTLICHE FAKULTÄT
HUMBOLDT-UNIVERSITÄT ZU BERLIN
SPANDAUER STRASSE 1
10178 BERLIN
GERMANY

INSTITUT FÜR ANGEWANDTE MATHEMATIK
RUPRECHT-KARLS-UNIVERSITÄT HEIDELBERG
IM NEUENHEIMER FELD 294
69120 HEIDELBERG
GERMANY
E-MAIL: mammen@statlab.uni-heidelberg.de

Chapter in the Encyclopedia of Statistical Science, Volume X

NONPARAMETRIC REGRESSION[1]

WOLFGANG HÄRDLE[2]
Wirtschaftswissenschaftliche Fakultät
Humboldt-Universität zu Berlin
D 10178 Berlin, Germany

OLIVER LINTON
Department of Economics
Yale University
New Haven, CT 06511, U.S.A

May 2, 1995

SUMMARY

We review different approaches to nonparametric regression estimation. Kernel estimators are compared to k-NN estimators, orthogonal series and splines. Pointwise and uniform confidence bands are described, and the choice of smoothing parameter is discussed. Finally, the method is applied to nonparametric prediction of time series and to additive modelling for dimension reduction.

[1] Forthcoming in the *Encyclopedia of Statistical Science,* Volume X, eds A. Kotz, C.B. Read and D. Banks.

[2] We would like to thank Rong Chen and Jens Perch Nielsen for helpful comments.

(1998) Linton, O. and Härdle, W. Nonparametric Regression.

List of Figures

1. Nonparametric Regression. The nonparametric approach to regression is based on the belief that parametric regression models are frequently misspecified and may result in incorrect inferences. By not restricting the functional form one obtains valid inferences for a much wider range of circumstances. Perhaps the primary use of the nonparametric method is to provide exploratory information that helps in model building. For this reason, the flexibility and robustness of this method is desirable.

We observe a bivariate dataset $\{(X_i, Y_i)\}_{i=1}^n$ generated from

$$Y_i = m(X_i) + \epsilon_i, \qquad i = 1, \ldots, n, \tag{1}$$

where ϵ_i is a random error independent over observations that satisfies $E(\epsilon_i \mid X_i = x) = 0$ and $\mathrm{Var}(\epsilon_i \mid X_i = x) = \sigma^2(x)$. Then $m(\bullet)$ is the regression function of Y on X. Usually, it is of interest to estimate m at a grid of points covering some subset $\mathcal{X}$ of the support of X. The smoothness of m on this set determines how well it can be estimated.

DEFINITION. *Let $\mathcal{M}_r$ be the class of all functions that possess r derivatives with Taylor expansion remainder that is Hölder continuous on a set $\mathcal{X}$.*

We concentrate on the special case $\mathcal{M}_2$, corresponding to two continuous derivatives, about which most is written, see Müller (1988). We discuss a number of estimators of $m(x)$ for $x \in \mathcal{X}$; these are all linear "smoothers" of the form $\sum_{i=1}^n W_{ni}(x)Y_i$, for some weighting sequence $\{W_{ni}(x)\}_{i=1}^n$ depending only on $X_1, .., X_n$, but arise from different motivations and possess different statistical properties. The methods we consider are appropriate for both *random design*, where (X_i, Y_i) are i.i.d, and *fixed design,* where X_i are fixed in repeated samples. In the random design case, X is an ancillary statistic, and standard statistical practice, see Cox and Hinkley (1974), is to conduct inference conditional on the sample $\{X_i\}_{i=1}^n$, Stute (1986). However, many papers in the literature

(1998) Linton, O. and Härdle, W. Nonparametric Regression.

prove theoretical properties unconditionally, and we shall, for ease of exposition, present results in this form. We also quote most results only for the case where X is scalar, although in section 4 we discuss the extension to multivariate data. In some cases, it is convenient to restrict attention to the equispaced design sequence $X_i = i/n$, $i = 1, .., n$. We restrict our attention to independent sampling, but some extensions to the dependent sampling case are given in Section 3.

Smoothing techniques have a long history starting at least in 1857 when the Saxonian economist E.Engel (1857, p.169) found the law named after him. He analyzed Belgian data on household expenditure, using what we would now call the regressogram. Whittaker (1923) used a graduation method for regression curve estimation which one would now call spline smoothing. Nadaraya (1964) and Watson (1964) provided an extension for general random design based on kernel methods. In time series, Daniell (1946) introduced the smoothed periodogram for consistent estimation of the spectral density. Fix and Hodges (1951) extended this for the estimation of a probability density and used in classification. Rosenblatt (1956) proved asymptotic consistency of the kernel density estimator. Schuster (1972) provided the proofs of consistency and asymptotic normality of the Wadaraya-Watson regression smoother.

These methods have developed considerably in the last ten years, and are now frequently used by applied statisticians. The massive increase in computing power as well as the increased availability of large cross-sectional and high-frequency financial time-series datasets are partly responsible for the popularity of these methods. They are typically simple to implement in software like GAUSS or XploRe, see Härdle, Klinke and Turlach (1995).

1.1. Kernel Estimators. Recall that

$$m(x) = \frac{\int y f(x,y) dy}{\int f(x,y) dy}, \tag{2}$$

where $f(x,y)$ is the joint density of (X,Y). A natural way to estimate $m(\bullet)$ is first to compute an estimate of $f(x,y)$ and then to integrate it according to this formula. A

kernel density estimate $\widehat{f}_h(x,y)$ of $f(x,y)$ is

$$\widehat{f}_h(x,y) = n^{-1}\sum_{i=1}^{n} K_h(x-X_i)K_h(y-Y_i),$$

where $K(\bullet)$ is any function (kernel) satisfying $\int K(u)du = 1$ and $K_h(\bullet) = h^{-1}K(h^{-1}\bullet)$, see Silverman (1986), Jones (1989), Jones and Foster (1993).

DEFINITION. *Let $\mathcal{K}_q$ be the class of all kernels of order q for which $\int u^j K(u)du = 0$, $j = 1,..,q-1$, and $\int u^q K(u)du < \infty$.*

Frequently, attention is restricted to K a probability density function symmetric about zero for which $q = 2$. For a list of kernels we refer the reader to Gasser and Müller (1984), Härdle and Linton (1994). We have:

$$\int \widehat{f}_h(x,y)dy = n^{-1}\sum_{i=1}^{n} K_h(x-X_i) \; ; \int y\widehat{f}_h(x,y)dy = n^{-1}\sum_{i=1}^{n} K_h(x-X_i)Y_i.$$

Plugging these into numerator and denominator of (2) we obtain the Nadaraya–Watson kernel estimate

$$\widehat{m}_h(x) = \frac{\sum_{i=1}^{n} K_h(x-X_i)Y_i}{\sum_{i=1}^{n} K_h(x-X_i)}.$$

The bandwidth h determines the degree of smoothness of $\widehat{m}_h$. This can be immediately seen by considering the limits for h tending to zero or to infinity, respectively. Indeed, at an observation X_i, $\widehat{m}_h(X_i) \to Y_i$, as $h \to 0$, while at an arbitrary point x, $\widehat{m}_h(x) \to n^{-1}\sum_{i=1}^{n} Y_i$, as $h \to \infty$. These two limit considerations make it clear that the smoothing parameter h in relation to the sample size n should not converge to zero too rapidly nor too slowly. Under only continuity conditions on m, f, and σ^2, Schuster (1972) established consistency of $\widehat{m}_h$. Under further conditions, it is asymptotically normal, as was first shown in Schuster (1972):

Theorem 1. *Suppose that* $K \in \mathcal{K}_2$ *satisfies* $\int |K(u)|du \leq \infty$, $\lim_{|u|\to\infty} uK(u) = 0$, *and* $\int |K(u)|^{2+\eta}du < \infty$, *for some* $\eta > 0$. *Suppose also that* $m(x)$ *and* $f(x) \in \mathcal{M}_2$, *that* $f(x) > 0$, *and that* $E(|Y|^{2+\eta} \mid x)$ *exists and is continuous at* x. *Finally, suppose that* $h = h(n) \to 0$ *and* $\overline{\lim}\, h^5 n < \infty$. *Then*

$$(nh)^{1/2}\left[\widehat{m}_h(x) - m(x) - h^2 B_{nw}(x)\right] \Rightarrow N(0, V_{nw}(x)),$$

where

$$B_{nw}(x) = \tfrac{1}{2}\int u^2 K(u)du\left[m''(x) + 2m'(x)\tfrac{f'}{f}(x)\right]$$
$$V_{nw}(x) = \left[\int K^2(u)du\right]\sigma^2(x)/f(x).$$

Note this theorem only applies to interior points; for boundary points, the bias is typically of order h unless some modifications are made to the kernel, see Müller (1987) for details.

1.2. k-Nearest Neighbor Estimators.

1.2.1. Ordinary k-NN Estimators. The kernel estimate was defined as a weighted average of the response variables in a fixed neighborhood of x. The k-nearest neighbor (k-NN) estimate is defined as a weighted average of the response variables in a varying neighborhood. This neighborhood is defined through those X-variables which are among the k-nearest neighbors of a point x.

Let $\mathcal{N}(x) = \{i : X_i$ is one of the k-NN to $x\}$ be the set of indices of the k-nearest neighbors of x. The k-NN estimate is the average of Y's with index in $\mathcal{N}(x)$,

$$\widehat{m}_k(x) = k^{-1}\sum_{i\in\mathcal{N}(x)} Y_i. \tag{3}$$

Connections to kernel smoothing can be made by considering (3) as a kernel smoother with uniform kernel $K(u) = \frac{1}{2}\mathrm{I}(|u| \leq 1)$ and variable bandwidth $h = R(k)$, the distance between x and its furthest k-NN,

$$\widehat{m}_k(x) = \frac{\sum_{i=1}^n K_R(x - X_i)\,Y_i}{\sum_{i=1}^n K_R(x - X_i)}. \tag{4}$$

Note that in (4), for this specific kernel, the denominator is equal to k/nR the k-NN density estimate of $f(x)$. Formula (4) provides sensible estimators for arbitrary kernels. The bias and variance of this more general k-NN estimator is given in a theorem by Mack (1981).

Theorem 2. *Let the conditions of Theorem 1 hold, except instead that* $k \to \infty$, $k/n \to 0$ *and* $\overline{\lim}\, k^5/n^4 < \infty$ *as* $n \to \infty$. *Then*

$$k^{1/2}\left[\widehat{m}_k(x) - m(x) - (k/n)^2 B_{nn}(x)\right] \Rightarrow N(0, V_{nn}(x)),$$

where,

$$B_{nn}(x) = \int u^2 K(u)du \left[\frac{m''(x)+2m'(x)\frac{f'}{f}(x)}{8f^2(x)}\right]$$
$$V_{nn}(x) = 2\sigma^2(x)\int K^2(u)du.$$

In contrast to kernel smoothing, the variance of the k-NN regression smoother does not depend on f, the density of X. This makes sense since the k-NN estimator always averages over exactly k observations independently of the distribution of the X-variables. The bias constant $B_{nn}(x)$ is also different from the one for kernel estimators given in Theorem 1. An approximate identity between k-NN and kernel smoothers can be obtained by setting

$$k = 2nhf(x), \tag{5}$$

or equivalently $h = k/2nf(x)$. For this choice of k or h respectively, the asymptotic mean squared error formulas of Theorem 1 and Theorem 2 are identical.

1.2.2. Symmetrized k-NN Estimators. A computationally useful modification of $\widehat{m}_k$ is to restrict the k-nearest neighbors always to symmetric neighborhoods, i.e., one takes $k/2$ neighbors to the left and $k/2$ neighbors to the right. In this case, weight-updating formulas can be given, see Härdle (1990, Section 3.2), Härdle (1991). The bias formulas are slightly different, see Härdle and Carroll (1989), but (5) remains true.

1.3. Local Polynomial Estimators. The Nadaraya-Watson estimator can be regarded as the solution of the minimization problem

$$\widehat{m}_h(x) = \arg\min_{\theta} \sum_{i=1}^{n} K_h(x - X_i) \{Y_i - \theta\}^2 .$$

This motivates the local polynomial class of estimators. Let $\widehat{\theta}_0, .., \widehat{\theta}_p$ minimize

$$\sum_{i=1}^{n} K_h(x - X_i) \left\{ Y_i - \theta_0 - \theta_1(X_i - x) - .. - \theta_p \frac{(X_i - x)^p}{p!} \right\}^2 . \tag{6}$$

Then $\widehat{m}_{h,p}(x) = \widehat{\theta}_0$ consistently estimates $m(x)$, while $\widehat{\theta}_j$ estimates the $j'th$ derivative of m. A variation on these estimators called *LOWESS* was first considered in Cleveland (1979) who employed a nearest neighbor window. Fan (1992) establishes an asymptotic approximation for the case where $p = 1$, which he calls the local linear estimator $\widehat{m}_{h,1}(x)$.

Theorem 3. *Let the conditions of Theorem 1 hold. Then*

$$(nh)^{1/2} \left[\widehat{m}_{h,1}(x) - m(x) - h^2 B_1(x) \right] \Rightarrow N(0, V_1(x)),$$

where

$$B_1(x) = \tfrac{1}{2} \left[\int u^2 K(u) du \right] m''(x)$$
$$V_1(x) = \left[\int K^2(u) du \right] \sigma^2(x)/f(x).$$

Higher order polynomials can achieve bias reduction for general regression functions, see Fan and Gijbels (1992) and Ruppert and Wand (1995). A general property here is that $\widehat{m}_{h,p}(x)$ is exactly unbiased when m is a p'th order, or less, polynomial.

The principle underlying the local polynomial estimator can be generalized in a number of ways. Tibshirani (1984) introduced the local likelihood procedure in which an arbitrary parametric regression function $g(x;\theta)$ substitutes the polynomial in (6). Fan, Heckman and Wand (1995) develop theory for a nonparametric estimator in a Generalized Linear Model (GLIM) in which, for example, a probit likelihood function replaces the polynomial in (6).

1.4. Spline Estimators. For any estimate $\widehat{m}$ of m, the residual sum of squares (RSS) is defined as $\sum_{i=1}^{n} \{Y_i - \widehat{m}(X_i)\}^2$, which is a widely used criterion, in parametric contexts, for generating estimators of regression functions. However, the RSS is minimized by an $\widehat{m}$ interpolating the data, assuming no ties in the $X's$. To avoid this problem it is necessary to add a penalty for lack of smoothness called the stabilizer. Most work is based on the stabilizer $\Omega(\widehat{m}) = \int \{\widehat{m}''(u)\}^2 \, du$, although see Ansley, Kohn and Wong (1993) and Koenker, Ng and Portnoy (1993) for alternatives. The cubic spline estimator $\widehat{m}_\lambda$ is the (unique) minimizer of

$$R_\lambda(\widehat{m}, m) = \sum_{i=1}^{n} \{Y_i - \widehat{m}(X_i)\}^2 + \lambda \int \{\widehat{m}''(u)\}^2 \, du. \tag{7}$$

The spline $\widehat{m}_\lambda$ has the following properties: It is a cubic polynomial between two successive X-values; at the observation points $\widehat{m}_\lambda(\bullet)$ and its first two derivatives are continuous; at the boundary of the observation interval the spline is linear. This characterization of the solution to (7) allows the integral term on the right hand side to be replaced by a quadratic form, see Eubank (1988) and Wahba (1990), and computation of the estimator proceeds by standard, although computationally intensive, matrix techniques.

The smoothing parameter λ controls the degree of smoothness of the estimator $\widehat{m}_\lambda$. As $\lambda \to 0$, $\widehat{m}_\lambda$ interpolates the observations, while if $\lambda \to \infty$, $\widehat{m}_\lambda$ tends to a least squares regression line. Although $\widehat{m}_\lambda$ is linear in the Y data, see Härdle (1990, p58-59), its dependency on the design and on the smoothing parameter is rather complicated. This has resulted in rather less treatment of the statistical properties of these estimators, except in rather simple settings, although see Wahba (1990) – in fact, the extension to multivariate design is not straightforward. However, splines are asymptotically equivalent to kernel smoothers as Silverman (1984) showed. The equivalent kernel is

$$K(u) = \frac{1}{2} \exp\left(-\frac{|u|}{\sqrt{2}}\right) \sin\left(\frac{|u|}{\sqrt{2}} + \frac{\pi}{4}\right),$$

which is of fourth order, since its first three moments are zero, while the equivalent

bandwidth $h = h(\lambda; X_i)$ is

$$h(\lambda; X_i) = \lambda^{1/4} n^{-1/4} \, f(X_i)^{-1/4}. \tag{8}$$

One advantage of spline estimators over kernels is that global inequality and equality constraints can be imposed more conveniently: for example, it may be desirable to restrict the smooth to pass through a particular point, see Jones (1985). Silverman (1985) discusses a Bayesian interpretation of the spline procedure.

1.5. Series Estimators. Series estimators have received considerable attention in the econometrics literature, following Elbadawi, Gallant and Souza (1983). This theory is very much tied to the structure of Hilbert space. Suppose that m has an expansion for all x:

$$m(x) = \sum_{j=0}^{\infty} \beta_j \varphi_j(x), \tag{9}$$

in terms of the orthogonal basis functions $\{\varphi_j\}_{j=0}^{\infty}$ and their coefficients $\{\beta_j\}_{j=0}^{\infty}$. Suitable basis systems include the *Legendre* polynomials described in Härdle (1990), the *Fourier* series used in Gallant and Souza (1991), and the recently developed *wavelet* basis, see Härdle, Kerkyacharian, Picard and Tsybakov (1995).

A simple method of estimating $m(x)$ involves firstly selecting a basis system and a truncation sequence $\tau(n)$, where $\tau(n)$ is an integer less than n, and then regressing Y_i on $\varphi_{ti} = (\varphi_0(X_i), .., \varphi_\tau(X_i))^T$. Let $\left\{\widehat{\beta}_j\right\}_{j=0}^{\tau(n)}$ be the least squares "parameter" estimates, then

$$\widehat{m}_\tau(x) = \sum_{j=0}^{\tau(n)} \widehat{\beta}_j \varphi_j(x) = \sum_{i=1}^{n} W_{ni}(x) Y_i, \tag{10}$$

where $W_n(x) = (W_{n1}, .., W_{nn})^T$, with

$$W_n(x) = \varphi_{\tau x}^T (\Phi_\tau^T \Phi_\tau)^{-1} \Phi_\tau^T, \tag{11}$$

where $\varphi_{\tau x} = (\varphi_0(x), .., \varphi_\tau(x))^T$ and $\Phi_\tau = (\varphi_{\tau 1}, .., \varphi_{\tau n})^T$.

These estimators are typically very easy to compute. In addition, the extension to additive structures and semiparametric models is convenient, see Andrews and Whang (1990) and Andrews (1991). Finally, series estimators can adapt to the smoothness of m: provided $\tau(n)$ grows at a sufficiently fast rate, the optimal rate of convergence, for the smoothness class of m, can be established, see Stone (1982), while fixed window order q kernel estimators achieve at best a rate of convergence of $n^{2q/2q+1}$. However, the same effect can be achieved by using a kernel estimator whose order changes with n in such a way as to produce bias reduction of the desired degree, see Müller (1987). In any case, the evidence of Marron and Wand (1992) cautions against the application of bias reduction techniques unless quite large sample sizes are available. Finally, a major disadvantage with the series method is that there is relatively little theory about how to select the basis system and the smoothing parameter $\tau(n)$.

1.6. Kernels, $k-$NN, splines, and series. Splines and series are both "global" methods in the sense that they try to approximate the whole curve at once, while kernel and nearest neighbor methods work separately on each estimation point. Nevertheless, when X is uniformly distributed, kernels and nearest neighbor estimators of $m(x)$ are identical, while spline estimators are roughly equivalent to a kernel estimator of order 4. Only when the design is not equispaced do substantial differences appear.

We apply kernel, k-NN, orthogonal series (we used the Legendre system of orthogonal polynomials), and splines to the car data set (Table 7, p. 352–355 in Chambers, Cleveland, Kleiner and Tukey (1983)).

In each plot, we give a scatterplot of the data x = price in dollars of car (in 1979) versus y = miles per US gallon of that car, and one of the nonparametric estimators. The sample size is $n = 74$ observations. In Figure 1a we have plotted together with the raw data a kernel smoother $\widehat{m}_h$ for which a quartic kernel was used with $h = 2000$. Very similar to this is the spline smoother shown in Figure 1c ($\lambda = 10^9$). In this example, the

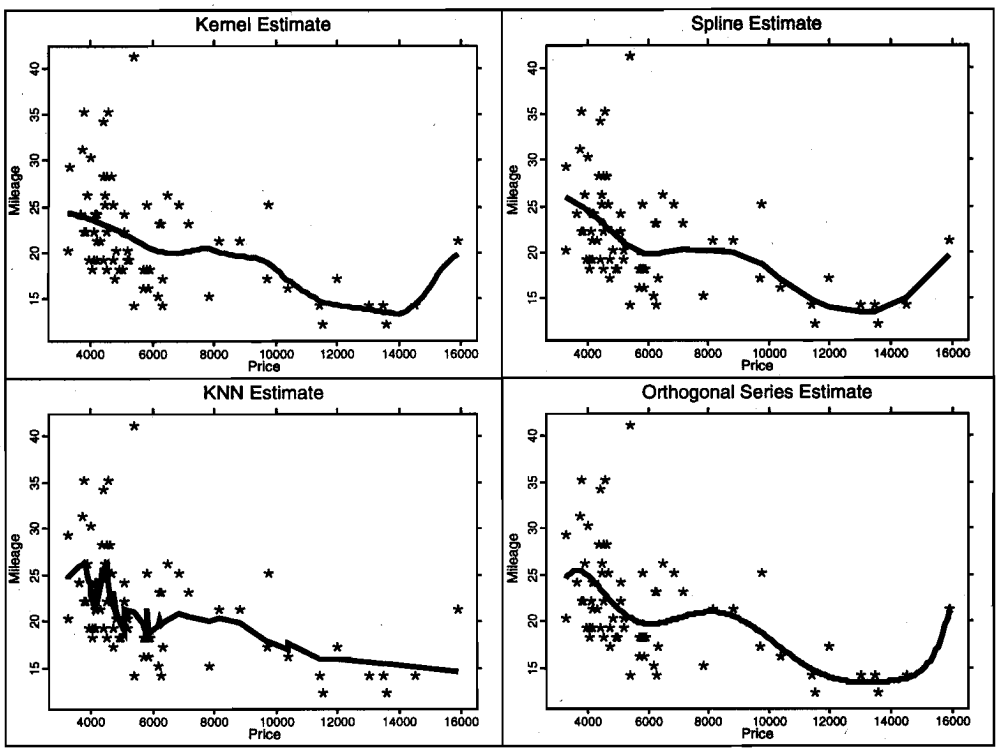


FIG. 1. *Scatterplot of car price (x) and miles per gallon (y) with four different smooth approximations ($n = 74$, $h = 2000$, $k = 11$, $\lambda = 10^9$, $\tau = 8$). Standard deviation of car price is* 2918.

X's are not too far from uniform. The effective local bandwidth for the spline smoother from (8) is a function of $f^{-1/4}$ only, which does not vary that much. Of course at the right end with the isolated observation at $x = 15906$ and $y = 21$ (Cadillac Seville) both kernel and splines must have difficulties. Both work essentially with a window of fixed width. The series estimator (Figure 1d) with $\tau = 8$ is quite close to the spline estimator.

In contrast to these regression estimators stands the k-NN smoother ($k = 11$) in Figure 1b. We used the symmetrized k-NN estimator for this plot see Härdle and Müller (1993). By formula (5) the dependence of k on f is much stronger than for the spline. At the right end of the price scale no local effect from the outlier described above is visible. By contrast in the main body of the data where the density is high this k-NN smoother tends to be wiggly.

1.7. Confidence Intervals. The asymptotic distribution results contained in Theorems 1-3 can be used to calculate pointwise confidence intervals for the estimators described above. In practice it is usual to ignore the bias term, since this is rather complicated, depending on higher derivatives of the regression function and perhaps on the derivatives of the density of X. This approach can be justified when a bandwidth is chosen that makes the bias relatively small.

We restrict our attention to the Nadaraya-Watson regression estimator. In this case, we suppose that $nh^5 \to 0$, which ensures that the bias term does not appear in the limiting distribution. Let

$$CLO(x) = \widehat{m}_h(x) - c_{\alpha/2}\widehat{s} \quad ; \quad CUP(x) = \widehat{m}_h(x) + c_{\alpha/2}\widehat{s},$$

where $\Phi(c_\alpha) = (1-\alpha)$ with $\Phi(\bullet)$ the standard normal distribution, while $\widehat{s}^2$ is any consistent estimate of the asymptotic variance of $\widehat{m}_h(x)$. For example: $\widehat{s}_1^2 = \sum_{i=1}^n W_{ni}^2(x)\widehat{\epsilon}_i^2$, $\widehat{s}_2^2 = \widehat{\sigma}_h^2(x)\sum_{i=1}^n W_{ni}^2(x)$, or $\widehat{s}_3^2 = n^{-1}h^{-1}[\int K^2(u)du]\widehat{\sigma}_h^2(x)/\widehat{f}_h(x)$, where $\widehat{f}_h(x)$ is the kernel density estimator, $\widehat{\epsilon}_i = Y_i - \widehat{m}_h(X_i)$ are the nonparametric residuals, and $\widehat{\sigma}_h^2(x) = \sum_{i=1}^n W_{ni}(x)\widehat{\epsilon}_i^2$ is a nonparametric estimator of $\sigma^2(x)$. With these definitions,

$$\Pr\{m(x) \in [CLO(x), CUP(x)]\} \to 1-\alpha.$$

The pointwise approach is relevant if the behavior of the regression function at a single point is under consideration. Usually, however, its behavior over an interval is under study. In this case, pointwise confidence intervals do not take account of the joint nature of the implicit null hypothesis. We now consider uniform confidence bands for the function m, over some compact subset χ of the support of X. Without loss of generality we take $\chi = [0,1]$. We require functions $CLO^*(x)$ and $CUP^*(x)$ such that

$$\Pr\{m(x) \in [CLO^*(x), CUP^*(x)] \ \textit{for all}\ x \in \chi\} \to 1-\alpha, \tag{12}$$

Let

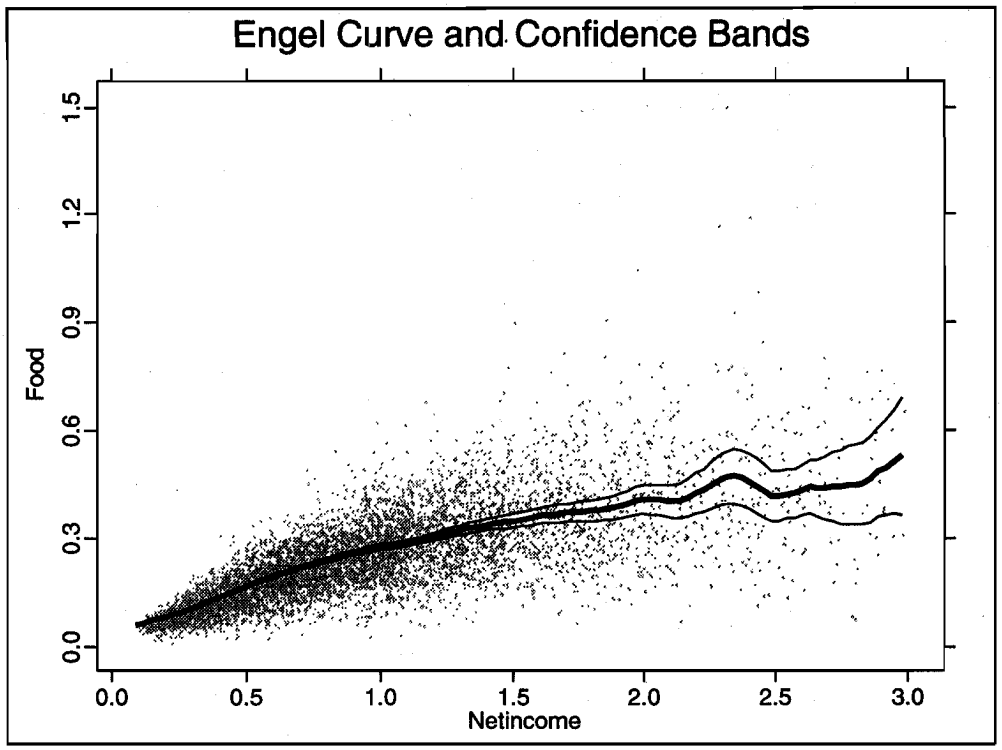


FIG. 2. *Uniform confidence bands for the income data. Food versus net income. Calculated using XploRe macro* **reguncb**. *Family Expenditure Survey. (1968 - 1983)*

$$CLO^*(x) = \widehat{m}_h(x) - \left[\frac{c_\alpha^*}{\delta} + \delta + \frac{1}{2\delta}\ln\left\{\frac{\int K'^2(u)du}{4\pi^2[\int K^2(u)du]}\right\}\right]\widehat{s}_1$$

$$CUP^*(x) = \widehat{m}_h(x) + \left[\frac{c_\alpha^*}{\delta} + \delta + \frac{1}{2\delta}\ln\left\{\frac{\int K'^2(u)du}{4\pi^2[\int K^2(u)du]}\right\}\right]\widehat{s}_1,$$

where $\delta = [2\log(1/h)]^{1/2}$, and $\exp\{-2\exp(-c_\alpha^*)\} = (1-\alpha)$. Then (12) is satisfied under the conditions given in Härdle (1990, Theorem 4.3.1).

In the figure below we show the uniform confidence bands for a data set described in Härdle and Linton (1994), Härdle and Jerison (1991).

1.8. Bootstrap Confidence Intervals. The bootstrap can be used to construct pointwise and uniform confidence intervals for both fixed and random design. The bootstrap can have a significant advantage here as was pointed out by Hall (1993): the error in (12) is $O(\log^{-1} n)$, while the error for a correct bootstrap procedure can be

$O((\log h^{-1})^3/nh)$ in the random design case. We outline the bootstrap procedure for the two sampling schemes.

1.8.1. Fixed Design with iid errors. The following steps are carried out

Step 1. Calculate residuals: $\widehat{\varepsilon}_i = Y_i - \widehat{m}_h(X_i)$, $i = 1, .., n$.

Step 2. Centering: $\widetilde{\varepsilon}_i = \widehat{\varepsilon}_i - n^{-1}\sum_{j=1}^n \widehat{\varepsilon}_j$, $i = 1, .., n$

Step 3. Resampling: Draw randomly $\varepsilon_1^*, .., \varepsilon_n^*$ from $\{\widetilde{\varepsilon}_1, .., \widetilde{\varepsilon}_n\}$

Step 4. Create bootstrap observations: $Y_i^* = \widehat{m}_g^*(X_i) + \varepsilon_i^*$, $i = 1, .., n$, where $\widehat{m}_g^*(\bullet)$ is a kernel estimate of $m(\bullet)$ using bandwidth g.

With the bootstrap data one calculates a kernel estimate

$$\widehat{m}_h^*(x) = \frac{\sum_{i=1}^n K_h(x - X_i^*)Y_i^*}{\sum_{i=1}^n K_h(x - X_i^*)}.$$

To evaluate the variability of $(nh)^{1/2}[\widehat{m}_h(x) - m(x)]$ one uses the conditional distribution of $(nh)^{1/2}[\widehat{m}_h^*(x) - \widehat{m}_h(x)]$ given the sample. Provided $h \sim n^{-1/5}$, $g \to 0$ and $g/h \to \infty$, the bootstrap works in this case, i.e. the two distributions are asymptotically the same, see Hall (1992).

1.8.2. Random Design. It would appear natural to resample from the joint empirical of the sample; unfortunately this will tend to underestimate the bias, see Härdle and Mammen (1991). One can either provide simultaneously a bias correction or one can resample from the modified empirical distribution

$$\widehat{F}_n(x, y) = n^{-1}\sum_{i=1}^n \mathrm{I}(Y_i \le y)\int_{-\infty}^x K_g(z - X_i)dz$$

for some alternative bandwidth g as in Gonzalez–Manteiga, Prada–Sanchez, Fiestras–Janeiro and Garcia-Jurado (1990).

2. Optimality and Bandwidth Choice.

2.1. Optimality. We say that a bandwidth sequence h^* is asymptotically optimal relative to a performance criterion $Q(h)$ if

$$\frac{Q(h^*)}{\inf_{h\in H_n} Q(h)} \xrightarrow{P} 1,$$

as $n \to \infty$, where H_n is the range of permissible bandwidths. There are a number of alternative optimality criteria in use. Firstly, we may be interested in the quadratic loss of the estimator at a single point x, which is measured by the *Mean Squared Error*, $MSE\{\widehat{m}_h(x)\}$. Secondly, we may be only concerned with a global measure of performance. In this case, we may consider the *Integrated Mean Squared Error*, $IMSE = \int MSE\{\widehat{m}_h(x)\}\pi(x)f(x)dx$ for some weighting function $\pi(\bullet)$. An alternative is the in-sample version of this, the *Average Squared Error*

$$d_A(h) = n^{-1}\sum_{i=1}^{n}\{\widehat{m}_h(X_i) - m(X_i)\}^2\pi(X_i).$$

The purpose of $\pi(\bullet)$ may be to down weight observations in the tail of $X's$ distribution, and thereby to eliminate boundary effects. When $h = O(n^{-1/5})$, the squared bias and the variance of the kernel smoother have the same magnitude; this is the optimal order of magnitude for h with respect to all three criteria, and the corresponding performance measures are all $O(n^{-4/5})$ in this case.

Now let $h = \gamma n^{-1/5}$, where γ is a constant. The optimal constant balances the contributions to MSE from the squared bias and the variance respectively. From Theorem 1 we obtain an approximate mean squared error expansion,

$$MSE[\widehat{m}_h(x)] \approx n^{-1}h^{-1}V(x) + h^4B^2(x), \tag{13}$$

and the bandwidth minimizing (13) is

$$h_0(x) = \left\{\frac{V(x)}{4B^2(x)}\right\}^{1/5} n^{-1/5}. \tag{14}$$

Similarly, the optimal bandwidth with respect to $IMSE$ is the same as (14) with $V = \int V(x)\pi(x)f(x)dx$ and $B^2 = \int B^2(x)\pi(x)f(x)dx$ replacing $V(x)$ and $B^2(x)$. Unfortunately, in either case the optimal bandwidth depends on the unknown regression function and design density. We discuss in Section 2.2 below how one can obtain empirical versions of (14).

By substituting h_0 in (13), we find that the optimal MSE and $IMSE$ depend on K only through

$$T(K) = [\int K^2(u)du]^2 / \int u^2 K(u)du. \tag{15}$$

This functional can be minimized with respect to K using the calculus of variations, although it is necessary to first adopt a scale standardization of K; for details, see Gasser, Müller, and Mammitzsch (1985). A kernel is said to be optimal if it minimizes (15). The optimal kernel of order 2 is the Epanechnikov kernel $K(u) = 0.75 * (1 - u^2)\mathrm{I}(|u| \leq 1)$. However, over a wide class of kernel estimators, the loss in efficiency is not that drastic; more important is the choice of h than the choice of K, see Marron and Nolan (1989).

2.2. Choice of Smoothing Parameter. For each nonparametric regression method, one has to choose how much to smooth for the given dataset. In Section 1 we saw that k-NN, series, and spline estimation are asymptotically equivalent to the kernel method, so we describe here only the selection of bandwidth h for kernel regression smoothing.

2.2.1. Plug-in. The asymptotic approximation given in (14) can be used to determine an optimal local bandwidth. We can calculate an estimated optimal bandwidth $\widehat{h}_{pl}$ in which the consistent estimators $\widehat{m}''_{h^*}(x)$, $\widehat{\sigma}^2_{h^*}(x)$, $\widehat{f}_{h^*}(x)$ and $\widehat{f}'_{h^*}(x)$ replace the unknown functions. We then use $\widehat{m}_{\widehat{h}_{pl}}(x)$ to estimate $m(x)$. Likewise, if a globally optimal bandwidth is required, one must substitute estimators of the appropriate average functionals. This procedure is generally fast and simple to implement. Its properties are examined in Härdle, Hall, and Marron (1992). However, this method fails to provide pointwise optimal bandwidths, when $m(x)$ possesses less than two continuous derivatives. Finally, a major disadvantage of this procedure is that a preliminary bandwidth

h^* must be chosen for estimation of $m''(x)$ and the other quantities.

2.2.2. Crossvalidation. Crossvalidation is a convenient method of global bandwidth choice for many problems, and relies on the well established principle of out-of-sample predictive validation. Suppose that optimality with respect to $d_A(h)$ is the aim. We must first replace $d_A(h)$ by a computable approximation to it. A naive estimate would be to just replace the unknown values $m(X_i)$ by the observations Y_i:

$$p(h) = n^{-1} \sum_{i=1}^{n} \{\widehat{m}_h(X_i) - Y_i\}^2 \pi(X_i),$$

which is called the resubstitution estimate. However, this quantity makes use of the each observation twice – the response variable Y_i is used in $\widehat{m}_h(X_i)$ to predict itself. Therefore, $p(h)$ can be made arbitrarily small by taking $h \to 0$. Alternatively, note that conditional on $X_1, .., X_n$, we have

$$E\left[p(h)\right] = E\left[d_A(h)\right] + n^{-1} \sum_{i=1}^{n} \sigma^2(X_i)\pi(X_i) - 2n^{-1} \sum_{i=1}^{n} W_{ni}(X_i)\sigma^2(X_i)\pi(X_i),$$

and the third term is of the same order of magnitude as $E\left[d_A(h)\right]$, but with negative sign. Therefore, d_A is wrongly underestimated, and the selected bandwidth will be downward biased.

The simplest way to avoid this problem is to remove the i-th observation from $\widehat{m}_h(X_i)$, and define

$$\widehat{m}_{h,i}(X_i) = \frac{\sum_{j \neq i} K_h(X_j - X_i) Y_i}{\sum_{j \neq i} K_h(X_j - X_i)}.$$

This leave-one-out estimate is used to form the so-called crossvalidation function

$$CV(h) = n^{-1} \sum_{i=1}^{n} \{\widehat{m}_{h,i}(X_i) - Y_i\}^2 \pi(X_i),$$

which is to be minimized with respect to h. For technical reasons, the infimum must be taken only over a restricted set of bandwidths such as $H_n = [n^{-(1/5-\zeta)}, n^{-(1/5+\zeta)}]$, for some $\zeta > 0$. The following theorem is proved in Härdle and Marron (1985):

Theorem 4. *Assume that the conditions given in Härdle (1990, Theorem 5.1.1) hold. Then the bandwidth selection rule, "Choose $\widehat{h}$ to minimize $CV(h)$" is asymptotically optimal with respect to $d_A(h)$ and $IMSE$.*

The conditions include the restriction that $f > 0$ on the compact support of π, moment conditions on ϵ, and a Lipschitz condition on K. However, unlike for the plug-in procedure, m and f need not be differentiable (a Lipschitz condition is required, however).

2.2.3. Other data driven selectors. There are a number of different automatic bandwidth selectors that produce asymptotically optimal kernel smoothers. They are based on various ways of correcting the downwards bias of the resubstitution estimate of $d_A(h)$. The function $p(h)$ is multiplied by a correction factor that in a sense penalizes the too small h's. The general form of this selector is

$$G(h) = n^{-1} \sum_{i=1}^{n} \{\widehat{m}_h(X_i) - Y_i\}^2 \pi(X_i) \Xi \{W_{ni}(X_i)\},$$

where Ξ is the correction function with first-order Taylor expansion

$$\Xi(u) = 1 + 2u + O(u^2),$$

as $u \to 0$. Some well known example are:

(a) *Generalized Cross-validation* (Craven & Wahba 1979): $\Xi_{GCV}(u) = (1-u)^{-2}$;

(b) *Akaike's Information Criterion* (Akaike 1970): $\Xi_{AIC}(u) = \exp(2u)$;

(c) *Finite Prediction Error* (Akaike 1974): $\Xi_{FPE}(u) = (1+u)/(1-u)$;

(d) *Shibata's (1981) model selector*: $\Xi_S(u) = 1 + 2u$;

(e) *Rice's (1984) bandwidth selector*: $\Xi_T(u) = (1-2u)^{-1}$.

Härdle, Hall, and Marron (1988) show that the general criterion $G(h)$ works in producing asymptotically optimal bandwidth selection, although they present their results for the equispaced design case only.

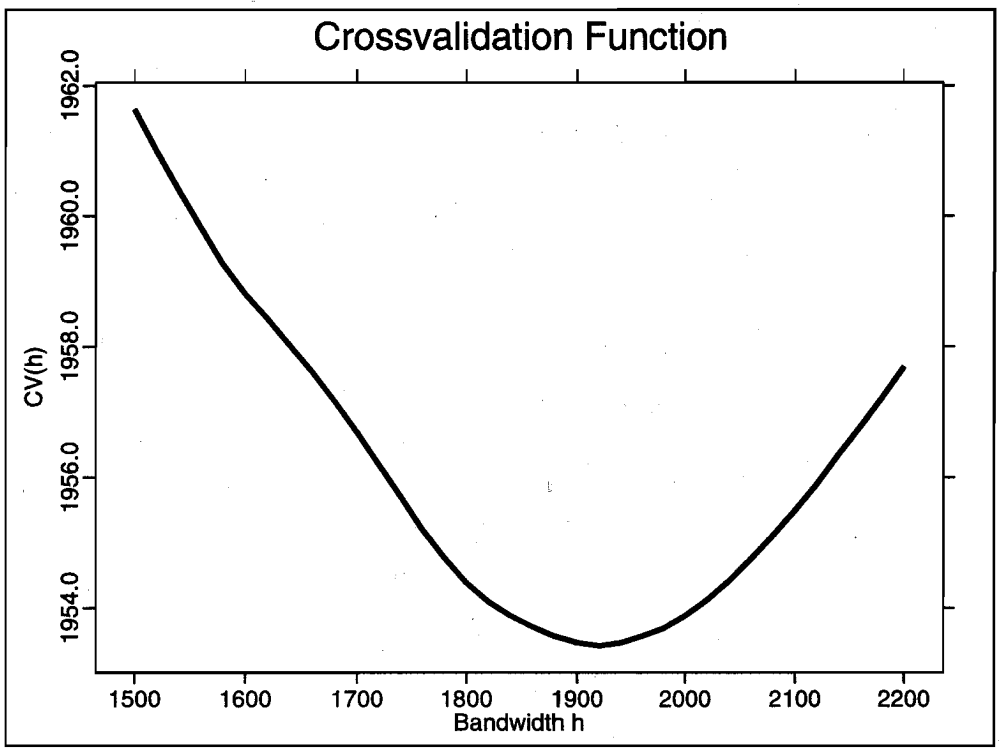


FIG. 3. *The crossvalidation function $CV(h)$ for the car data. Quartic kernel. Computation made with XploRe macro* `regcvl`.

The method of crossvalidation was applied to the car data set to find the optimal smoothing parameter h. A plot of the crossvalidation function is given in Figure 3. The computation is for the quartic kernel $K(u) = \frac{15}{16}(1-u^2)^2 \mathrm{I}(|u| \leq 1)$ using the WARPing method, see Härdle and Scott (1992). The minimal $\widehat{h} = \arg\min CV(h)$ is at 1800 which shows that in Figure 5a we used a slightly too large bandwidth.

Härdle, Hall and Marron (1988) investigate how far the crossvalidation optimal $\widehat{h}$ is from the true optimum $\widehat{h}_0$ (that minimizes $d_A(h)$). They show that for each optimization method,

$$n^{1/10}\left(\frac{\widehat{h}-\widehat{h}_0}{\widehat{h}_0}\right) \Rightarrow N(0,\sigma^2) \tag{16}$$

$$n\left\{d_A(\widehat{h}) - d_A(\widehat{h}_0)\right\} \Rightarrow C_1\chi^2_1,$$

where σ^2 and C_1 are both positive. To this higher order of approximation, the above methods are all asymptotically equivalent. Another interesting result is that the estimated $\widehat{h}$ and optimum $\widehat{h}_0$ are actually negatively correlated! Hall and Johnstone (1992) show how to correct for this effect in density estimation and in regression with uniform X's. It is still an open question how to improve this for the general regression setting we are considering here.

There has been considerable research into finding improved methods of bandwidth selection, that give faster rates of convergence in (16). Most of this work is in density estimation – see the recent review of Jones, Marron and Sheather (1992) for references. In this case, various $n^{1/2}$ consistent bandwidth selectors have been suggested. The finite sample properties of these procedures are not well established, although Park and Turlach (1992) contains some preliminary simulation evidence. Härdle, Hall and Marron (1992) construct a $n^{1/2}$ consistent bandwidth selector for regression based on a bias reduction technique.

3. Application to Time Series. In the theoretical development described up to this point, we have restricted our attention to independent sampling. However, smoothing methods can also be applied to dependent data. We focus on the issue of functional form, rather than that of correlation structure – this latter issue is treated, from a nonparametric point of view, in Brillinger (1980). Suppose that we observe the vector time series $\{(X_i, Y_i)\}_{i=1}^n$. It is convenient to assume that the process is stationary and mixing as defined in Chanda (1974), Garodetskii (1977), Gallant and White (1988), although extensions to certain types of nonstationarity can also be permitted. We consider two distinct problems. Firstly, we want to predict Y_i from its own past which we call autoregression. Secondly, when we want to predict Y_i from X_i which problem we call regression with correlated errors.

3.1. Autoregression. For convenience we restrict attention to the problem of predicting the scalar Y_{i+k} given Y_i for some $k > 0$. The best predictor is provided by

the autoregression function

$$M_k(y) = \mathrm{E}(Y_{i+k} \mid Y_i = y). \tag{17}$$

More generally, one may wish to estimate the conditional variance of Y_{i+k} from lagged values,

$$V_k(y) = \mathrm{Var}(Y_{i+k} \mid Y_i = y),$$

and even the predictive density $f_{Y_{i+k}|Y_i}$. These quantities can be estimated using any of the smoothing methods described in this chapter. See Robinson (1983) and Bierens (1987) for some theoretical results including convergence rates and asymptotic distributions.

A scientific basis can also be made for choosing bandwidth in this sampling scheme. Härdle and Vieu (1991) showed that crossvalidation also works in the autoregression problem – "choose" $\widehat{h} = \mathrm{argmin}\, CV(h)$ gives asymptotically optimal estimates.

To illustrate this result we simulated an autoregressive process $Y_i = M(Y_{i-1}) + \epsilon_i$ with

$$M(y) = y \exp(-y^2),$$

where the innovations ϵ_i were uniformly distributed over the interval $(-1/2, 1/2)$. Such a process is α-mixing with geometrically decreasing $\alpha(n)$ as shown by Doukhan and Ghindès (1980) and Györfi et al. (1990, Section III.4.4). The sample size investigated was $n = 100$. The quartic kernel function was used. The minimum of $CV(h)$ was $\widehat{h} = 0.43$, while the optimum of $d_A(h)$ is at $h = 0.52$. The curve $d_A(h)$ is very flat for this example, since there is very little bias present. In Figure 4 we compare the estimated curve with the autoregression function and find good coincidence.

3.2. Correlated Errors. We now consider the regression model

$$Y_i = m(X_i) + \epsilon_i,$$

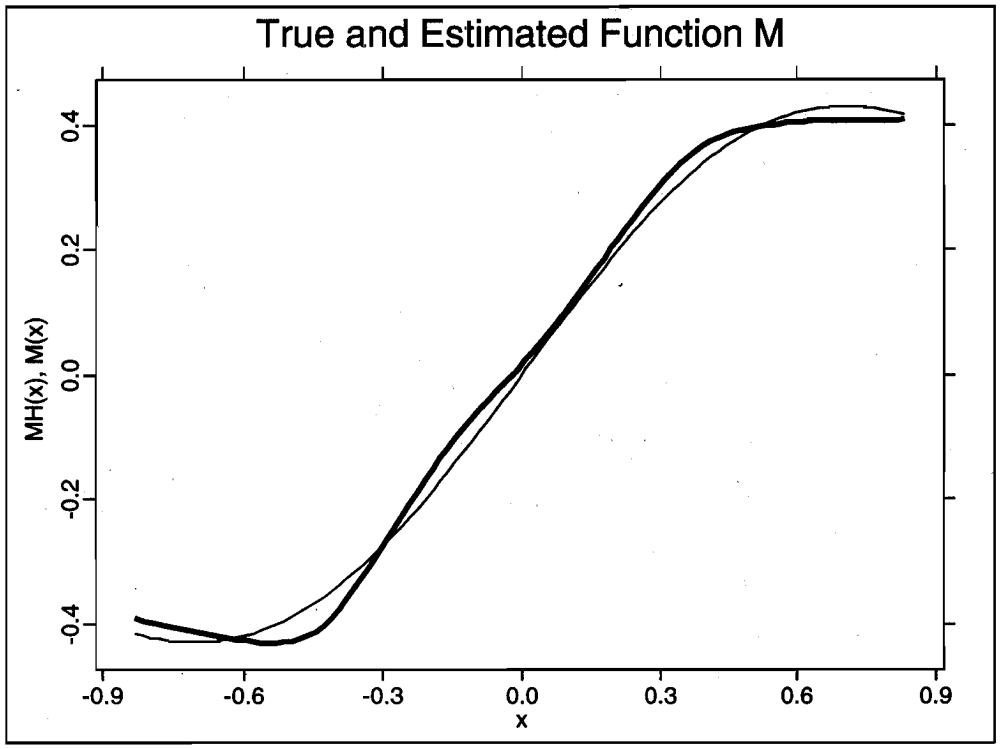


FIG. 4. *The time regression function* $M(y) = y\exp(-y^2)$ *for the simulated example (thick line) and the kernel smoother (thin line).*

where X_i is fixed in repeated samples and the errors ϵ_i satisfy $E(\epsilon_i|X_i) = 0$, but are autocorrelated. The kernel estimator $\widehat{m}_h(x)$ of $m(x)$ is consistent under quite general conditions. In fact, its bias is the same as when ϵ_i are independent. However, the variance is generally affected by the dependency structure. Suppose that the error process is $MA(1)$, i.e.

$$\epsilon_i = u_i + \theta u_{i-1},$$

where u_i are i.i.d with zero mean and variance σ^2. In this case,

$$\text{Var}\,[\widehat{m}_h(x)] = \sigma^2 \left\{ (1+\theta^2) \sum_{i=1}^{n} W_{ni}^2 + 2\theta \sum_{i=1}^{n-1} W_{ni} W_{ni+1} \right\} \tag{18}$$

which is $O(n^{-1}h^{-1})$, but differs from Theorem 1. If the explanatory variable were time itself (i.e. $X_i = i/n$, $i = 1, .., n$), then a further approximation is possible:

$$\text{Var}\,[\widehat{m}_h(x)] \approx n^{-1}h^{-1}\sigma^2(1+\theta^2+2\theta)[\int K^2(u)du].$$

Hart and Wehrly (1986) develop MSE approximations in a regression model in which the error correlation is a general function $\rho(\bullet)$ of the time between observations.

Unfortunately, crossvalidation fails in this case. The error process tends to stay too long on one side of the mean curve. Therefore, the bandwidth selection procedure gives undersmoothed estimates, since it interprets the little bumps of the error process as part of the regression curve. An example is given in Härdle (1990, Figures 7.6, 7.7). The effect of correlation on the crossvalidation criterion may be mitigated by leaving out more than just one observation. For the $MA(1)$ process, leaving out the 3 contiguous (in time) observations works. This "leave-out-some" technique is sometimes appealing also in the independent setting, see the discussion of Härdle, Hall and Marron (1988), and Hart and Vieu (1991). It may also be possible to correct for this effect by "whitening" the residuals, although this has yet to be shown.

4. Multidimensional Design. Now suppose that X is d-dimensional with $d > 1$ and let $X_i = (X_{1i}, .., X_{di})^T$ and $x = (x_1, .., x_d)^T$. A product kernel estimator of $m(x)$ is given by

$$\widehat{m}_h(x) = \frac{\sum_{i=1}^n \prod_{\alpha=1}^d K_h(X_{\alpha i} - x_\alpha)Y_i}{\sum_{i=1}^n \prod_{\alpha=1}^d K_h(X_{\alpha i} - x_\alpha)},$$

where $\widehat{m}(x)$ is consistent provided $h \to 0$ and $nh^d \to \infty$, see Härdle (1990). When $m \in \mathcal{M}_2$, the bias of $\widehat{m}_h(x)$ with $q = 2$ is $O(h^2)$ just as for $d = 1$, but the variance is $O(n^{-1}h^{-d})$ and increases with d. Thus the optimal rate of convergence of $\widehat{m}(x)$ is the slower $n^{2/d+4}$; this is often called the curse of dimensionality. An additional problem is that simple plots are not available to aid model selection. There are a number of simplifying structures that have been used to avoid these problems. These include

single index models as in Härdle and Stoker (1986), the regression tree structure of Gordon & Olshen (1980), the projection pursuit model of Friedman & Stuetzle (1981), semiparametric models such as considered in Engle, Granger, Rice & Weiss (1986), and the additive structure of Buja, Hastie & Tibshirani (1989), see Härdle (1990, p257-287) for further discussion. We briefly discuss some recent work on additive models.

4.1. Additive Models. Suppose that

$$m(x) = c + \sum_{\alpha=1}^{d} m_\alpha(x_\alpha),$$

where without loss of generality $E[m_\alpha(X_{\alpha i})] = 0$. Stone (1985) shows that m_α, $\alpha = 1, .., d$ can be estimated with the one-dimensional convergence rate of $n^{2/5}$. In practice, the Hastie & Tibshirani (1990) estimation procedures are widely used. These involve multiple iterations, where the additive structure is used in each step, to obtain estimates of m_α, $\alpha = 1, .., d$. A major disadvantage of this method is that its statistical properties are not well understood. Recently, Linton and Nielsen (1995) have proposed an alternative method based on integration. Let Q be some $d-1$ probability measure, and define

$$\widehat{m}_\alpha(x_\alpha) = \int \widehat{m}(x) dQ(x_1, .., x_{\alpha-1}, x_{\alpha+1}, .., x_d).$$

Then $\widehat{m}_\alpha$ estimates m_α up to a constant. This constant is zero if Q is the joint distribution of $X_{1i}, .., X_{\alpha-1 i}, X_{\alpha+1 i}, .., X_{di}$ or a consistent estimate of it as provided by the empirical distribution. Chen, Härdle, Linton and Severance-Lossin (1995) show that with Q this empirical distribution, $\widehat{m}_\alpha(x_\alpha) - m_\alpha(x_\alpha) = O_p(n^{-2/5})$ under appropriate conditions.

REFERENCES

[1] AKAIKE, H. (1970): "Statistical predictor information," *Annals of the Institute of Statistical Mathematics* 22, 203–17.

[2] AKAIKE, H. (1974): "A new look at the statistical model identification." *IEEE Transactions of Automatic Control AC* 19, 716–23.

[3] ANDREWS, D.W.K., (1991): "Asymptotic Normality of Series Estimators for Nonparametric and Semiparametric Regression Models." *Econometrica* 59, 307-346.

[4] ANDREWS, D.W.K., AND Y-J. WHANG. (1990): "Additive and Interactive Regression Models: Circumvention of the Curse of Dimensionality," *Econometric Theory* 6, 466-479.

[5] ANSLEY, C.F., R.KOHN, AND C.WONG (1993): "Nonparametric spline regression with prior information," *Biometrika* 80, 75-88.

[6] BIERENS, H.J., (1987): "Kernel Estimators of Regression Functions." in *Advances in Econometrics: Fifth World Congress,* Vol 1, ed. by T.F. Bewley. Cambridge University Press.

[7] BRILLINGER, D.R., (1980) *Time Series, Data analysis and Theory.* Holden-Day.

[8] BUJA, A., HASTIE, T. AND R. TIBSHIRANI (1989): "Linear smoothers and additive models (with discussion)" *Annals of Statistics* 17, 453-555.

[9] CHAMBERS, J.M., CLEVELAND, W.S., KLEINER, B., AND P.A. TUKEY (1983). *Graphical Methods for Data Analysis.* Duxburry Press.

[10] CHANDA, K.C. (1974): "Strong mixing properties of linear stochastic process." *Journal of Applied Probabilities* 11, 401–408.

[11] CHEN, R.; W. HÄRDLE; O. LINTON AND E. SEVERANCE-LOSSIN (1995): "Estimation and variable selection in additive nonparametric regression models."

[12] CLEVELAND, W.S., (1979): "Robust Locally Weighted Regression and Smoothing Scatterplots." *Journal of the American Statistical Association* 74, 829-836.

[13] COX, D.R., AND D.V. HINKLEY (1974): *Theoretical Statistics.* Chapman and Hall.

[14] CRAVEN, P. AND WAHBA, G. (1979): "Smoothing noisy data with spline functions," *Numer. Math.* 31, 377–403.

[15] DANIELL, P.J., (1946): "Discussion of paper by M.S.Bartlett," *Journal of the Royal Statistical Society Supplement* 8:27.

[16] DOUKHAN, P. AND GHINDES, M. (1980): "Estimation dans le processus $X_n = f(X_{n-1}) + \epsilon_n$," *Comptes Rendus, Académie des Sciences de Paris* 297, Série A, 61–4.

[17] ELBADAWI, I., A.R.GALLANT, AND G.SOUZA, (1983): "An elasticity can be estimated consistently without a priori knowledge of functional form," *Econometrica* 51, 1731-1751.

[18] ENGEL, E. (1857): "Die vorherrschenden Gewerbszweige in den Gerichtsämtern mit Beziehung auf die Productions- und Consumptionsverhältnisse des Königreichs Sachsen." *Zeitschrift des*

*Statistischen Bureaus des Königlichen Sächsischen Ministerium des Innern*8,9, 153-182.

[19] ENGLE, R.F. GRANGER, C.W.J. RICE, J., AND A. WEISS (1986): "Semiparametric estimates of the relation between weather and electricity sales." *Journal of the American Statistical Association* 81, 310-320.

[20] EUBANK, R.L., (1988): *Smoothing Splines and Nonparametric Regression.* Marcel Dekker.

[21] FAMILY EXPENDITURE SURVEY, Annual Base Tapes (1968-1983). Department of Employment, Statistics Division, Her Majesty's Stationary Office, London 1968-1983.

[22] FAN, J. (1992): "Design-Adaptive Nonparametric Regression," *Journal of the American Statistical Association* 87, 998-1004.

[23] FAN, J. and Gijbels, I. (1992) "Variable bandwidth and local linear regression smooths." *Am. Statistics* 20,2008 – 2036.

[24] FAN, J., N.E. HECKMAN, AND M.P. WAND, (1995): "Local Polynomial Kernel Regression for Generalized Linear Models and Quasi-Likelihood Functions," *Journal of the American Statistical Association* 90, 141-150.

[25] FIX, E. AND J.L.HODGES (1951): "Discriminatory analysis, nonparametric estimation: consistency properties," *Report no 4, Project no 21-49-004,* USAF School of Aviation Medicine, Randolph Field, Texas.

[26] FRIEDMAN, J. AND W. STUETZLE (1981): "Projection pursuit regression," *Journal of the American Statistical Association* 76, 817-823.

[27] GALLANT, A.R., AND G.SOUZA, (1991): "On the asymptotic normality of Fourier flexible form estimates," *Journal of Econometrics* 50, 329-353.

[28] GALLANT, A.R., AND H. WHITE, (1988): "A Unified Theory of Estimation and Inference for Nonlinear Dynamic Models", *Blackwell: Oxford*

[29] GARODETSKII, V.V. (1977): "On the strong mixing condition for linear process," *Theory of Probability and its Applications* 22, 411–413.

[30] GASSER, T. AND H. G. MÜLLER (1984): "Estimating regression functions and their derivatives by the kernel method," *Scandinavian Journal of Statistics* 11, 171–85.

[31] GONZALEZ–MANTEIGA, W., PRADA–SANCHEZ, J.M., FIESTRAS–JANEIRO, M.G. AND GARCIA–JURADO, I. (1990) : "Dependence between fusion temperatures and chemical components of a certain type of coal using classical, non–parametric and bootstrap techiques", *Journal of Chemometrics* 4, 429–439.

[32] GASSER, T., MÜLLER, H. G., AND V. MAMMITZSCH (1985): "Kernels for nonparametric curve estimation," *Journal of the Royal Statistical Society Series B* 47, 238–52.

[33] GORDON, L. AND R.A. OLSHEN (1980): "Consistent nonparametric regression from recursive partitioning schemes." *Journal of Multivariate Analysis* 10, 611-627.

[34] GYÖRFI, L., HÄRDLE, W., SARDA, P., AND P. VIEU (1990): *Nonparametric Curve Estima-*

tion from Time Series. Lecture Notes in Statistics, 60. Heidelberg, New York: Springer-Verlag.

[35] HALL, P., (1992): *The Bootstrap and Edgeworth Expansion.* Springer-Verlag, New-York.

[36] HALL, P., (1993): "On Edgeworth Expansion and Bootstrap Confidence Bands in Nonparametric Curve Estimation," *Journal of the Royal Statistical Society Series B* 55, 291-304.

[37] HALL, P. AND I. JOHNSTONE (1992): "Empirical functional and efficient smoothing parameter selection," (with discussion). *Journal of the Royal Statistical Society Series B.* 54, 475-530.

[38] HÄRDLE, W. (1990). *Applied Nonparametric Regression.* New York: Cambridge University Press.

[39] HÄRDLE, W. (1991). *Smoothing Techniques with Implementation in S.* Heidelberg, New York, Berlin: Springer-Verlag.

[40] HÄRDLE, W. AND CARROLL, R. J. (1989): "Biased cross-validation for a kernel regression estimator and its derivatives," *Österreichische Zeitschrift für Statistik und Informatik.* 20, 53-64.

[41] HÄRDLE, W. AND MAMMEN, E. (1991) "Bootstrap Methods for Ninparametric Regression." *Nonparametric Functional estimation and Related Topics, Edited by G. roussas, Kluwer Publishing Company, Series C: Mathematical and Physical Sciences,* 335, 111–124.

[42] HÄRDLE, W., HALL, P. AND MARRON, J. S. (1988): "How far are automatically chosen regression smoothing parameters from their optimum?" (with discussion). *Journal of the American Statistical Association* 83, 86–101.

[43] HÄRDLE, W., HALL, P. AND MARRON, J. S. (1992): "Regression smoothing parameters that are not far from their optimum" *Journal of the American Statistical Association* 87, 227–233.

[44] HÄRDLE, W. AND M. JERISON, (1991): "Cross Section Engel Curves over Time," *Recherches Economiques de Louvain,* 57, 391–431.

[45] HÄRDLE, W., KERKYACHARIAN, G., PICARD, D. AND TSYBAKOV, A.B. (1995): "Wavelets and Econometric Applications." *book manuscript.*

[46] HÄRDLE, W., S. KLINKE, AND B. TURLACH (1995): *XploRe: An interactive statistical computing environment.* Springer Verlag, New York.

[47] HÄRDLE, W., AND O.B. LINTON (1994): "Applied Nonparametric Methods," *Chapter of the 4. Handbook of Econometrics, North Holland,* 38, 2295–2339.

[48] HÄRDLE, W., AND MARRON, J. S. (1985): "Optimal bandwidth selection in nonparametric regression function estimation," *Annals of Statistics* 13, 1465-81.

[49] HÄRDLE, W., AND M.MÜLLER (1993): "Nichtparametrische Glättungsmethoden in der alltäglichen statistichen Praxis," *Allg. Statistiches Archiv* 77, 9-31.

[50] HÄRDLE, W. AND D.W. SCOTT (1992): "Smoothing in Low and High Dimensions by Weighted Averaging Using Rounded Points." *Computational Statistics,* 1, 97–128.

[51] HÄRDLE, W., AND T. M. STOKER (1989): "Investigating Smooth Multiple Regression by the

Method of Average Derivatives," *Journal of the American Statistical Association* 84, 986-995.

[52] HÄRDLE, W., AND A.B. TSYBAKOV (1995): *Wavelets in econometrics.* Manuscript, Springer Verlag..

[53] HÄRDLE, W. AND P. VIEU (1991): "Kernel regression smoothing of time series," *Journal of Time Series Analysis* 13, 209-232.

[54] HART, J. AND P. VIEU (1990): "Data-driven bandwidth choice for density estimation based on dependent data," *Annals of Statistics* 18, 873–890.

[55] HART, D. AND T. E. WEHRLY (1986): "Kernel regression estimation using repeated measurements data," *Journal of the American Statistical Association* 81, 1080–8.

[56] HASTIE, T.J., AND R.J.TIBSHIRANI (1990): *Generalized Additive Models* Chapman and Hall.

[57] JONES, M.C., (1985): "Discussion of the paper by B.W.Silverman," *Journal of the Royal Statistical Society Series B* 47, 25-26.

[58] JONES, M.C., (1989): "Discretized and interpolated Kernel Density Estimates, " *Journal of the American Statistical Association* 84, 733-741.

[59] JONES, M.C., AND P.J. FOSTER (1993): "Generalized jacknifing and higher order kernels," Forthcoming in *Journal of Nonparametric Statistics.*

[60] JONES, M.C., J.S. MARRON, AND S.J. SHEATHER (1992): "Progress in data-based selection for Kernel Density estimation," *Australian Graduate School of Management* Working paper no 92-014.

[61] KOENKER, R., P.NG AND S. PORTNOY (1993): "Quantile Smoothing Splines," *Forthcoming in Biometrika* .

[62] LINTON, O.B. AND J.P. NIELSEN (1995): "A kernel method of estimating structured nonparametric regression using marginal integration," *Biometrika.* 82, 93-101.

[63] MACK, Y. P. (1981): "Local properties of k-NN regression estimates," *SIAM J. Alg. Disc. Meth.* 2, 311–23.

[64] MARRON, J.S. AND D. NOLAN (1989): "Canonical kernels for density estimation," *Statistics and Probability Letters* 7, 191-195.

[65] MARRON, J.S. AND M.P.WAND (1992): "Exact Mean Integrated Squared Error." *Annals of Statistics* 20, 712-736.

[66] MÜLLER, H. G. (1987): "On the asymptotic mean square error of L_1 kernel estimates of C_∞ functions," *Journal of Approximation Theory* 51, 193-201.

[67] MÜLLER, H. G. (1988): *Nonparametric Regression Analysis of Longitudinal Data.* Lecture Notes in Statistics, Vol. 46. Heidelberg/New York: Springer-Verlag.

[68] NADARAYA, E.A., (1964): "On estimating regression," *Theory of Probability and its Applications* 10, 186-190.

[69] PARK, B.U., AND B.A. TURLACH (1992): "Practical performance of several data-driven band-

width selectors (with discussion)," *Computational Statistics* 7, 251-271.

[70] RICE, J. A. (1984): "Bandwidth choice for nonparametric regression" *Annals of Statistics* 12, 1215–30.

[71] ROBINSON, P.M. (1983): "Nonparametric Estimators for Time Series." *Journal of Time Series Analysis* 185-208.

[72] ROSENBLATT, M., (1956): "Remarks on some nonparametric estimates of a density function," *Annals of Mathematical Statistics* 27, 642-669.

[73] RUPPERT, D., AND M.P.WAND (1995): "Multivariate Locally Weighted Least Squares Regression," *Annals of Statistics* 22, 1346-1370.

[74] SCHUSTER, E.F., (1972): "Joint asymptotic distribution of the estimated regression function at a finite number of distinct points," *Annals of Mathematical Statistics* 43, 84-8.

[75] SHIBATA, R.(1981): "An optimal selection of regression variables," *Biometrika*, 68, 45–54.

[76] SILVERMAN, B. W. (1984): "Spline smoothing: the equivalent variable kernel method." *Annals of Statistics* 12, 898–916.

[77] SILVERMAN, B. W. (1985): "Some aspects of the Spline Smoothing approach to Non-parametric Regression Curve Fitting," *Journal of the Royal Statistical Society Series B* 47, 1-52

[78] SILVERMAN, B. W. (1986). *Density estimation for statistics and data analysis.* London: Chapman and Hall.

[79] STONE, C.J., (1982): "Optimal global rates of convergence for nonparametric regression," *Annals of Statistics* 10, 1040-1053.

[80] STUTE, W. (1986): "Conditional Empirical Processes," *Annals of Statistics* 14, 638-647.

[81] TIBSHIRANI, R., (1984): "Local Likelihood estimation," PhD Thesis, Stanford University.

[82] TIKHONOV, A.N. (1963): "Regularization of incorrectly posed problems," *Soviet Math.*, 4, 1624–1627.

[83] WAHBA, G. (1990): *Spline Models for Observational Data.* CBMS-NSF Regional Conference Series in Applied Mathematics, no 59.

[84] WATSON, G.S. (1964): "Smooth regression analysis," *Sankhya Series A* 26, 359-372.

[85] WHITTAKER, E.T., (1923): "On a new method of graduation," *Proc. Edinburgh Math.Soc* 41, 63-75.

Testing Parametric Versus Semiparametric Modeling in Generalized Linear Models

Wolfgang HÄRDLE, Enno MAMMEN, and Marlene MÜLLER

We consider a generalized partially linear model $E(Y|\mathbf{X}, \mathbf{T}) = G\{\mathbf{X}^T\beta + m(\mathbf{T})\}$, where G is a known function, β is an unknown parameter vector, and m is an unknown function. We introduce a test statistic that allows one to decide between a parametric and a semiparametric model: (a) m is linear (i.e., $m(\mathbf{t}) = \mathbf{t}^T\gamma$ for a parameter vector γ), and (b) m is a smooth (nonlinear) function. Under linearity (a), we show that the test statistic is asymptotically normal. Moreover, we prove that the bootstrap works asymptotically. Simulations suggest that (in small samples) the bootstrap outperforms the calculation of critical values from the normal approximation. The practical performance of the test is demonstrated in applications to data on East–West German migration and credit scoring.

KEY WORDS: Binary choice models; Bootstrap test; Credit scoring; Generalized linear models; Migration; Smoothed quasi-likelihood.

1. INTRODUCTION

In the analysis of discrete response variables one often models the expected value of the response as a nonlinear monotone function of a linear combination of the explanatory variables. Examples are probit or logit models, where the nonlinear (link) function is the cumulative distribution function of a normal or logistic distribution (see McCullagh and Nelder 1989). Then the so-called *generalized linear model* has the form

$$E(Y|\mathbf{Z}) = G(\mathbf{Z}^T\theta) \tag{1}$$

with a known monotone function G and an unknown parameter θ. The model (1) combines computational feasibility (especially for discrete covariates) with good interpretability of the "index" $\mathbf{Z}^T\theta$ and thus has found wide application in all fields of applied statistics (see, e.g., Fahrmeir and Tutz 1994) and Maddala 1983). However, for some applications it may be argued that the assumption of linearity in (1) is too restrictive. In fact, it may not be even clear if the relationship between the influential variables and the response is monotone. A more complex relationship (allowing also for nonmonotone dependence) is given by the semiparametric *generalized partially linear model*

$$E(Y|\mathbf{Z}) = G\{\mathbf{X}^T\beta + m(\mathbf{T})\}, \tag{2}$$

where $\mathbf{Z} = (\mathbf{X}, \mathbf{T})$ is a split of $\mathbf{Z}$ into two components $\mathbf{X}$ and $\mathbf{T}$, β is an unknown parameter, and m is an unknown smooth function. For a discussion of model (2) and additional references see Severini and Staniswalis (1994).

As an example of a possible nonlinear dependence, consider a model on East–West German migration in 1991. This model uses data from the German Socio-Economic Panel for Mecklenburg-Vorpommern, a land of the Federal Republic of Germany (GSOEP 1991). The dependent variable is binary with $Y = 1$ (intention to migrate) or $Y = 0$ (intention to stay). The variable T = household income serves as an explanatory variable along with some socio-economic factors $\mathbf{X}$ = (age, sex, friends in west, city size, unemployment). Figure 1 shows a fit of the function m in the semiparametric model (2) using a logistic link function $G(u) = 1/\{1+\exp(-u)\}$. The estimated function is clearly nonlinear and shows a saturation in the intention to migrate for higher income households. The question is, of course, whether the observed nonlinearity is significant.

In this article we discuss a test of the parametric hypothesis (1); that is,

$$m(\mathbf{t}) = \mathbf{t}^T\gamma \quad \text{for a vector } \gamma, \tag{3}$$

versus the semiparametric alternative (2). Our test indicates whether a nonlinear shape observed in nonparametric fit of m is significant. Furthermore, the proposed test complements the work of Severini and Staniswalis (1994), who considered estimation under model (2). Optimal rates for the nonparametric component and efficient estimation of the parametric component have been discussed by Mammen and van de Geer (1997). With identity link, this model has also been analyzed by Green (1987), Robinson (1988), and Speckman (1983). A related model with semiparametric index has been given by Carroll, Fan, Gijbels, and Wand (1997). Most of the literature in this semiparametric context, though, was devoted to estimation rather than testing.

Our test is based on ideas of Hastie and Tibshirani (1990). For a more general setup, they proposed applying the likelihood ratio test and using chi-squared approximations for the calculation of critical values. Approximate degrees of freedom are derived by calculating the expectation of asymp-

Wolfgang Härdle is Professor of Econometrics, Institute for Statistics and Econometrics, Humboldt University, D-10178 Berlin, Germany. Enno Mammen is Associate Professor of Mathematical Statistics, Institute for Applied Mathematics, Ruprecht-Karls-University, Heidelberg, 69120 Heidelberg, Germany (E-mail: mammen@statlab.uni-heidelberg.de). Marlene Müller is Assistant Professor, Institute for Statistics and Econometrics, Humboldt-University, D-10178 Berlin, Germany. The research for this article was supported by Sonderforschungsbereich 373, "Quantifikation und Simulation Ökonomischer Prozesse" at Humboldt University. The work of M. Müller was supported in part by CentER, Tilburg University, The Netherlands. The authors thank Michael C. Burda for helpful discussions and comments on the economic applications. Furthermore, they are grateful to an associate editor and two referees for detailed comments on this article and its exposition.

Journal of the American Statistical Association, 93 , 1461-1474

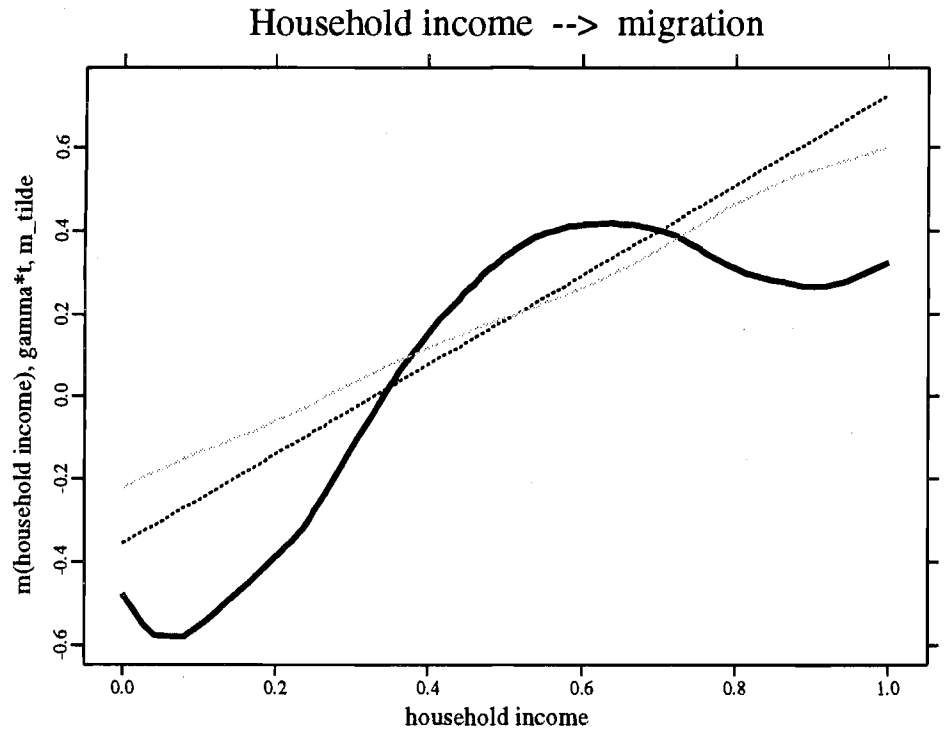


Figure 1. The Influence m(t) of Household Income (Transformed to [0, 1]) on Migration Intention. Nonparametric fit (thick black line), linear fit (thin black dashed line), and "biased" parametric estimate $\tilde{m}$ (thin gray dashed line), n = 402.

totic expansions of the test statistic under the null hypothesis. For this approach, only heuristic justification has been given. We propose the following modifications of this approach.

First, we correct for the bias of nonparametric estimates. Second, we modify the test statistic for the reason that different likelihoods (smoothed or unsmoothed likelihood) have been used in the calculation of the nonparametric or parametric component. For this modified test we can develop an asymptotic distribution theory. The test statistic does not have an asymptotic chi-squared distribution. We propose using the bootstrap for the calculation of critical values and show that bootstrap works.

The next section introduces estimators of m, γ, and β. These estimators will be used in the construction of the test statistics. The test and its asymptotic properties are discussed in Section 3. A small simulation study, the application to the migration example and another example on credit scoring, is discussed in Section 4. Remarks on the computation of the test statistics and proofs of our results are given in the Appendix.

2. ESTIMATION IN THE PARAMETRIC AND SEMIPARAMETRIC MODELS

For the estimation of the parametric component β and the nonparametric component m, we follow the approach of Severini and Staniswalis (1994). The method is based on quasi-likelihood estimation. The quasi-likelihood function is defined as

$$Q(\mu; y) = \int_{\mu}^{y} \frac{(s-y)}{V(s)}\, ds$$

where μ is the (conditional) expectation of Y; that is, $\mu = G\{\mathbf{X}^T\beta + m(\mathbf{T})\}$. It is assumed here that the conditional variance of Y is $\sigma^2 V(\mu)$, where σ is an unknown scale parameter and V is a known function. Quasi-likelihood functions are motivated by exponential families. Note that the maximum likelihood estimate $\hat{\theta}$, based on an iid sample $Y_1, \ldots, Y_n$ from an exponential family, is given by

$$\sum_{i=1}^{n} \frac{\partial}{\partial \theta}\, Q(\mu_i; Y_i) = 0.$$

In our model, the quasi-likelihood function is given as

$$\mathcal{L}(m, \beta) = \sum_{i=1}^{n} Q(\mu_i; Y_i), \tag{4}$$

where $(Y_1, X_1, T_1), \ldots, (Y_n, X_n, T_n)$ is a sample of independent observations and $\mu_i = G\{X_i^T\beta + m(T_i)\}$. The parameter β is supposed to lie in $B \subset \mathbb{R}^p$. The covariates $\mathbf{X}_i$ and $\mathbf{T}_i$ are $\mathbb{R}^p$ and $\mathbb{R}^q$ valued. We assume that the response variable Y_i is real valued. Multidimensional responses can be treated similarly.

The model assumption that the conditional variance of Y_i is equal to $\sigma^2 V(\mu_i)$ may be violated by the underlying data. For this reason, in our asymptotic analysis we do not suppose this condition. For the study of the bootstrap, we discuss this general case as well as the particular cases that the conditional variance is $\sigma^2 V(\mu_i)$ or that the conditional distribution of Y_i belongs to an exponential family; see Section 3.1.

(1998) Härdle, W., Mammen, E. and Müller, M.
Testing Parametric versus Semiparametric Modelling in Generalized Linear Models.

Journal of the American Statistical Association, 93 , 1461-1474

For the estimation of the nonparametric component m, we use the following smoothed quasi-likelihood:

$$\mathcal{L}^S(m(\cdot), \boldsymbol{\beta}) = \int \sum_{i=1}^{n} K_{\mathbf{h}}(\mathbf{t} - \mathbf{T}_i) Q[G\{\mathbf{X}_i^T \boldsymbol{\beta} + m(\mathbf{t})\}; Y_i]\, dt, \quad (5)$$

where $K_{\mathbf{h}}(\mathbf{u}) = (h_1 \cdot \ldots \cdot h_q)^{-1} K(h_1^{-1} u_1, \ldots, h_q^{-1} u_q)$ is a kernel (defined on $\mathbb{R}^q$) with bandwidth (vector) $\mathbf{h} = (h_1, \ldots, h_q)$. Following Severini and Staniswalis (1994) and Severini and Wong (1992), we put for $\boldsymbol{\beta} \in B$,

$$\hat{m}_{\boldsymbol{\beta}} = \operatorname*{argmax}_m \mathcal{L}^S(m, \boldsymbol{\beta}), \quad (6)$$

$$\hat{\boldsymbol{\beta}} = \operatorname*{argmax}_{\boldsymbol{\beta}} \mathcal{L}(\hat{m}_{\boldsymbol{\beta}}, \boldsymbol{\beta}), \quad (7)$$

and

$$\hat{m} = \hat{m}_{\hat{\boldsymbol{\beta}}}. \quad (8)$$

In (6), maximization runs over functions $m(\cdot)$. Because an integral is maximized by maximizing its integrand, the value $\eta = \hat{m}_{\boldsymbol{\beta}}(\mathbf{t})$ is defined as the maximizer of the "local likelihood" $\sum_{i=1}^{n} K_{\mathbf{h}}(\mathbf{t} - \mathbf{T}_i) Q[G\{\mathbf{X}_i^T \boldsymbol{\beta} + \eta\}; Y_i]$; see (5). Without loss of generality, we always assume that the constant vector is not contained in the design space. An intercept is automatically modeled by the nonparametric component. Under this assumption, the maximization in (6) and (7) is unique. (For a discussion of these estimates, see Severini and Staniswalis 1994.)

Our test will be based on a comparison of the semiparametric estimates with the estimators $(\tilde{\boldsymbol{\beta}}, \tilde{\boldsymbol{\gamma}})$ in the parametric model

$$(\tilde{\boldsymbol{\beta}}, \tilde{\boldsymbol{\gamma}}) = \operatorname*{argmax}_{\boldsymbol{\beta}, \boldsymbol{\gamma}} \mathcal{L}^P(\boldsymbol{\gamma}, \boldsymbol{\beta}). \quad (9)$$

Here $\mathcal{L}^P(\boldsymbol{\gamma}, \boldsymbol{\beta})$ is the quasi-likelihood function in model (1),

$$\mathcal{L}^P(\boldsymbol{\gamma}, \boldsymbol{\beta}) = \sum_{i=1}^{n} Q\{G(\mathbf{X}_i^T \boldsymbol{\beta} + \mathbf{T}_i^T \boldsymbol{\gamma}); Y_i\}. \quad (10)$$

The scale parameter σ can be estimated by

$$\hat{\sigma}^2 = \frac{1}{n} \sum_{i=1}^{n} (Y_i - \hat{\mu}_i)^2 / V(\hat{\mu}_i), \quad (11)$$

where $\hat{\mu}_i = G\{\mathbf{X}_i^T \hat{\boldsymbol{\beta}} + \hat{m}(\mathbf{T}_i)\}$.

A direct comparison of $\hat{m}(\mathbf{t})$ and $\mathbf{t}^T \tilde{\boldsymbol{\gamma}}$ may be misleading, because $\hat{m}$ has a smoothing bias which is typically nonnegligible. This also holds if the hypothesis of linearity is true. To avoid this effect, we add a bias to $\mathbf{t}^T \tilde{\boldsymbol{\gamma}}$ that will compensate for the bias of $\hat{m}(\mathbf{t})$. We do this by "smoothing" the function $\mathbf{t} \to \mathbf{t}^T \tilde{\boldsymbol{\gamma}}$. For this purpose, we consider the artificial dataset $\{\bar{Y}_i, \mathbf{X}_i, \mathbf{T}_i\}$: $i = 1, \ldots, n$, where $\bar{Y}_i = G(\mathbf{X}_i^T \tilde{\boldsymbol{\beta}} + \mathbf{T}_i^T \tilde{\boldsymbol{\gamma}})$ is the parametric fit of $E(Y_i | \mathbf{X}_i, \mathbf{T}_i)$. The function $\tilde{m}$ is defined by the following smoothing step:

$$\tilde{m} = \operatorname*{argmax}_m \int \sum_{i=1}^{n} K_{\mathbf{h}}(\mathbf{t} - \mathbf{T}_i) \times Q[G\{\mathbf{X}_i^T \tilde{\boldsymbol{\beta}} + m(\mathbf{t})\}; \bar{Y}_i]\, d\mathbf{t}. \quad (12)$$

In the Appendix we show that under the hypothesis $\tilde{m}(\mathbf{t})$ is asymptotically equivalent to $\mathbf{t}^T \tilde{\boldsymbol{\gamma}}$+ the bias of $\hat{m}(\mathbf{t})$. Therefore, in the difference $\hat{m}(\mathbf{t}) - \tilde{m}(\mathbf{t})$, the bias cancels asymptotically.

3. TESTING THE PARAMETRIC MODEL VERSUS THE SEMIPARAMETRIC MODEL

Our test procedures are based on a comparison of the parametric estimates $\tilde{\boldsymbol{\beta}}$ and $\tilde{m}$ with the semiparametric estimates $\hat{\boldsymbol{\beta}}$ and $\hat{m}$. A natural approach would be based on the likelihood ratio statistic $\mathcal{L}(\hat{m}, \hat{\boldsymbol{\beta}}) - \mathcal{L}(\tilde{m}, \tilde{\boldsymbol{\beta}})$. Unfortunately, this test statistic does not work, because in the construction of $\hat{m}$ and $\hat{\boldsymbol{\beta}}$, two different likelihood functions (smoothed and unsmoothed) have been used. [A Taylor expansion of the test statistic, in particular of the ith summand into $c_i \delta_i + d_i \delta_i^2$ with $\delta_i = \mathbf{X}_i^T(\hat{\boldsymbol{\beta}} - \tilde{\boldsymbol{\beta}}) + \hat{m}(\mathbf{T}_i) - \tilde{m}(\mathbf{T}_i)$, does not lead to a quadratic form.] This cannot be repaired by using the smoothed quasi-likelihood $\mathcal{L}^S$ instead of $\mathcal{L}$.

We propose the following test statistic:

$$R_1 = -2 \sum_{i=1}^{n} Q(\tilde{\mu}_i; \hat{\mu}_i), \quad (13)$$

with $\tilde{\mu}_i = G\{\mathbf{X}_i^T \tilde{\boldsymbol{\beta}} + \tilde{m}(\mathbf{T}_i)\}$ and $\hat{\mu}_i = G\{\mathbf{X}_i^T \hat{\boldsymbol{\beta}} + \hat{m}(\mathbf{T}_i)\}$ for $i = 1, \ldots, n$.

Note that for the case where the variance function V is constant, R_1 is equal to $\sum_{i=1}^{n} (\tilde{\mu}_i - \hat{\mu}_i)^2 / V$. In general, R_1 is equal to $\sum_{i=1}^{n} (\tilde{\mu}_i - \hat{\mu}_i)^2 / V(\bar{\mu}_i)$, where $\bar{\mu}_i$ is a point between $\tilde{\mu}_i$ and $\hat{\mu}_i$. Therefore R_1 can be interpreted as a weighted quadratic deviation.

If the distribution of Y does not belong to an exponential family, then calculation of R_1 involves evaluating n integrals. In these cases, the following two modifications of R_1, motivated by a Taylor expansion of R_1, are easier to compute:

$$R_2 = \sum_{i=1}^{n} \frac{[G'\{\mathbf{X}_i^T \hat{\boldsymbol{\beta}} + \hat{m}(\mathbf{T}_i)\}]^2}{V[G\{\mathbf{X}_i^T \tilde{\boldsymbol{\beta}} + \hat{m}(\mathbf{T}_i)\}]} \times \{\mathbf{X}_i^T(\hat{\boldsymbol{\beta}} - \tilde{\boldsymbol{\beta}}) + \hat{m}(\mathbf{T}_i) - \tilde{m}(\mathbf{T}_i)\}^2 \quad (14)$$

and

$$R_3 = \sum_{i=1}^{n} \frac{\{G'(\mathbf{X}_i^T \tilde{\boldsymbol{\beta}} + \mathbf{T}_i^T \tilde{\boldsymbol{\gamma}})\}^2}{V\{G(\mathbf{X}_i^T \tilde{\boldsymbol{\beta}} + \mathbf{T}_i^T \tilde{\boldsymbol{\gamma}})\}} \times \{\mathbf{X}_i^T(\hat{\boldsymbol{\beta}} - \tilde{\boldsymbol{\beta}}) + \hat{m}(\mathbf{T}_i) - \tilde{m}(\mathbf{T}_i)\}^2. \quad (15)$$

Theorem 1 discusses the asymptotics of these test statistics. The test statistics are asymptotically equivalent under the null hypothesis and have an asymptotic normal distribution.

Theorem 1. Suppose that the assumptions A1–A8 (in Sec. A.2 of the Appendix) apply. Then, under the hypothesis $m_0(\mathbf{t}) = \mathbf{t}^T \boldsymbol{\gamma}_0$, it holds that

(a) $R_1 = R_2 + o_p(v_n) = R_3 + o_p(v_n)$,

(1998) Härdle, W., Mammen, E. and Müller, M.
Testing Parametric versus Semiparametric Modelling in Generalized Linear Models.

Journal of the American Statistical Association, 93 , 1461-1474

and

$$\text{(b)}\ \ v_n^{-1}(R_1 - e_n) \xrightarrow{D} \mathrm{N}(0,1),$$

where e_n is a sequence with $e_n = h_{\mathrm{prod}}^{-1} \int K(\mathbf{u})^2\, d\mathbf{u}\lambda_1 + O(h_{\max}^2 h_{\mathrm{prod}}^{-1})$ and v_n^2 is defined by $v_n^2 = 2h_{\mathrm{prod}}^{-1} \int K^{(2)}(\mathbf{u})^2\, d\mathbf{u}\lambda_2$. Here we use the notation $h_{\max} = \max\{h_1, \cdots, h_q\}$ and $h_{\mathrm{prod}} = h_1 \cdot \ldots \cdot h_q$. The kernel $K^{(2)}$ is the convolution of K with itself. Furthermore,

$$\lambda_1 = E\, \frac{E\left[\frac{\sigma^2(\mathbf{X},\mathbf{T})G'(\eta)^2}{V^2(G(\eta))}\ |\mathbf{T}\right]}{E\left[\frac{G'(\eta)^2}{V\{G(\eta)\}}\ |\mathbf{T}\right]}\, p(\mathbf{T})^{-1},$$

and

$$\lambda_2 = E\, \frac{E\left[\frac{\sigma^2(\mathbf{X},\mathbf{T})G'(\eta)^2}{V^2\{G(\eta)\}}\ |\mathbf{T}\right]^2}{E\left[\frac{G'(\eta)^2}{V\{G(\eta)\}}\ |\mathbf{T}\right]^2}\, p(\mathbf{T})^{-1}.$$

where $\sigma^2(\mathbf{X},\mathbf{T})$ is the conditional variance of Y given $(\mathbf{X},\mathbf{T})$, and where $\eta = \mathbf{X}^T\boldsymbol{\beta}_0 + \mathbf{T}^T\boldsymbol{\gamma}_0$. If the conditional variance $\sigma^2(\mathbf{X},\mathbf{T})$ is correctly specified by $\sigma^2 V\{G(\eta)\}$, then λ_1 is equal to λ_2 and $\sigma^{-2}\lambda_1 = \sigma^{-2}\lambda_2$ is the Lebesgue measure of the support S_T of $\mathbf{T}$.

Note in particular that $\int K(\mathbf{u})^2\, d\mathbf{u} \neq \int\{K^{(2)}(\mathbf{u})\}^2\, d\mathbf{u}$. Therefore, for the case where $\lambda_1 = \lambda_2$, Theorem 1 implies that a chi-squared approximation is not appropriate for the distribution of R_1. This is because for kernel smoothing operators $\mathcal{K}$, it does not hold that $\mathcal{K}\mathcal{K} = \mathcal{K}$. This is in contrast to projection operators like B splines (see Buja, Hastie, and Tibshirani 1989). In particular, $\lambda_1 = \lambda_2$ holds if $Q(y;\mu)$ is the log-likelihood. Then R_1 is a modification of the (log)-likelihood ratio test statistic.

For the asymptotic mean e_n, an explicit formula can be given that contains conditional expectations of smoothed functions. Because this formula is rather lengthy, it is omitted here.

Theorem 1 states that the test statistics R_1, R_2, and R_3 are asymptotically equivalent under the null hypothesis. By standard arguments of asymptotic decision theory, the asymptotic equivalence remains valid for contiguous alternatives (i.e., $n^{-1/2}$ neighbored alternatives). In a parametric setting, this would imply that these three tests have asymptotically equivalent power. However, in our nonparametric setup the tests will have nontrivial power (power bounded away from the level and from 1) only for noncontiguous alternatives. Therefore, power functions may behave quite differently. A comparison of power functions based on simulations can be found in Section 4.

3.1 Bootstrap Tests

For two points $\mathbf{s}_n$ and $\mathbf{t}_n$ the nonparametric estimates $\hat{m}(\mathbf{s}_n)$ and $\hat{m}(\mathbf{t}_n)$ are asymptotically independent if the supports of the kernels $K_{\mathbf{h}}(\bullet - \mathbf{s}_n)$ and $K_{\mathbf{h}}(\bullet - \mathbf{t}_n)$ are disjoint. This may explain why, asymptotically, R_1 behaves approximately like a sum of $O(h_1^{-1} \cdot \ldots \cdot h_q^{-1})$ independent summands and has an asymptotic normal limit. Because, typically, $h_1^{-1} \cdot \ldots \cdot h_q^{-1}$ are not very large, it can be suspected that normal approximations do not work well for R_1 (see Härdle and Mammen 1993 for a related discussion). Therefore, we advise against using normal approximations for the calculation of quantiles. Instead, we propose using the bootstrap. We discuss here three versions of the bootstrap. The first version is the wild bootstrap, which is related to proposals of Wu (1986) (see also Beran 1986; Mammen 1992) and was first proposed by Härdle and Mammen (1993) in nonparametric setups. Note that in our model the conditional distribution of Y is not specified apart from A1 and A2.

The wild bootstrap procedure works as follows:

Step 1. Calculate residuals $\hat{\varepsilon}_i = Y_i - G(\mathbf{X}_i^T\hat{\boldsymbol{\beta}} + \hat{m}(\mathbf{T}_i))$.

Step 2. Generate n iid random variables $\varepsilon_1^*, \ldots, \varepsilon_n^*$ with mean 0 and variance 1 and that fulfill for a constant C that $|\varepsilon_i^*| \leq C$ (a.s.) for $i = 1, \ldots, n$.

Step 3. Put $Y_i^* = G(\mathbf{X}_i^T\tilde{\boldsymbol{\beta}} + \mathbf{T}_i^T\tilde{\boldsymbol{\gamma}}) + \hat{\varepsilon}_i\varepsilon_i^*$ for $i = 1, \ldots, n$.

Step 4. Calculate estimates $\hat{\boldsymbol{\beta}}^*, \hat{m}^*, \tilde{\boldsymbol{\beta}}^*, \tilde{\boldsymbol{\gamma}}^*$, and $\hat{m}^*$ based on the bootstrap samples $(\mathbf{X}_1, \mathbf{T}_1, Y_1^*), \ldots, (\mathbf{X}_n, \mathbf{T}_n, Y_n^*)$. Furthermore, calculate test statistics R_1^*, R_2^*, and R_3^*. The $(1-\alpha)$ quantiles of the distributions of R_1, R_2, and R_3 can be estimated by the $(1-\alpha)$ quantiles of the conditional distributions of R_1^*, R_2^*, or R_3^*.

Under the additional model assumption

$$\mathrm{var}(Y|\mathbf{X},\mathbf{T}) = \sigma^2 V\{G(\mathbf{X}^T\boldsymbol{\beta}_0 + \mathbf{T}^T\boldsymbol{\gamma}_0)\},$$

we propose the following modification of the resampling. In Step 3, put $Y_i^* = G(\mathbf{X}_i^T\tilde{\boldsymbol{\beta}} + \mathbf{T}_i^T\tilde{\boldsymbol{\gamma}}) + \hat{\sigma}V\{G[\mathbf{X}_i^T\hat{\boldsymbol{\beta}} + \hat{m}(\mathbf{T}_i)]\}^{1/2}\varepsilon_i^*$ for $i = 1, \ldots, n$ where $\hat{\sigma}^2$ is a consistent estimate of σ^2. In this case, the condition that $|\varepsilon_i^*|$ is bounded can be weakened to the assumption that ε_i^* has subexponential tails; that is, for a constant C, it holds that $E(e^{[|\varepsilon_i^*|/C]}) \leq C$ for $i = 1, \ldots, n$ (cf. A2).

In the special situation where $Q(\mu;y)$ is the log-likelihood (a semiparametric generalized linear model), the conditional distribution of Y_i is specified by $\mu_i = G(\mathbf{X}_i^T\boldsymbol{\beta} + \mathbf{T}_i^T\boldsymbol{\gamma})$. Then we recommend generating n independent $Y_1, \ldots, Y_n$ with distributions defined by $G(\mathbf{X}_1^T\tilde{\boldsymbol{\beta}} + \mathbf{T}_1^T\tilde{\boldsymbol{\gamma}}), \ldots, G(\mathbf{X}_n^T\tilde{\boldsymbol{\beta}} + \mathbf{T}_n^T\tilde{\boldsymbol{\gamma}})$. This is a version of the parametric bootstrap. In the binary response example that we considered earlier, Y_i is a Bernoulli variable with parameter $\mu_i = G(\mathbf{X}_i^T\boldsymbol{\beta} + \mathbf{T}_i^T\boldsymbol{\gamma})$. Hence here it is reasonable to resample from the Bernoulli distribution with parameter $\tilde{\mu}_i = G(\mathbf{X}_i^T\tilde{\boldsymbol{\beta}} + \mathbf{T}_i^T\tilde{\boldsymbol{\gamma}})$.

Theorem 2 shows that these three bootstrap procedures work (for their corresponding models).

Theorem 2. Suppose that the assumptions of Theorem 1 hold. In case of application of the second or third version of the bootstrap, assume that the aforementioned additional model assumptions hold. Then it holds for $j = 1, 2, 3$ that

$$d_K(\mathcal{L}^*(R_j^*), \mathcal{L}(R_j)) \xrightarrow{P} 0,$$

where $\mathcal{L}(R_j)$ is the distribution of R_j, $\mathcal{L}^*(R_j^*)$ is the conditional distribution of R_j^* (given the sample), and d_K denotes the Kolmogorov distance, which for two probability measures μ and ν (on the real line) is defined as

$$d_K(\mu,\nu) = \sup_{t\in\mathbb{R}} |\mu(X \leq t) - \nu(X \leq t)|.$$

(1998) Härdle, W., Mammen, E. and Müller, M.
Testing Parametric versus Semiparametric Modelling in Generalized Linear Models.

Application of these three versions of the bootstrap for β has been discussed by Mammen and van de Geer (1997), who estimated the nonparametric component by splines. The statement of the theorem also holds if the residuals are defined as $\hat{\varepsilon}_i = Y_i - G(\mathbf{X}_i^T\tilde{\beta} + \mathbf{T}_i^T\tilde{\gamma})$. We have seen in our simulations for binary responses that the normal approximation in Theorem 1 (b) is indeed inaccurate for small sample sizes (see Sec. 4), but that critical values are estimated quite well by the bootstrap.

Our test statistic depends on the choice of the bandwidth $\mathbf{h}$. Different values of $\mathbf{h}$ may lead to different observed significance levels (see Sec. 4). Small values of $\mathbf{h}$ have been motivated by asymptotic minimax theory (see Ingster 1993 and Lepski and Spokoiny 1995). In particular, the bandwidths proposed in these papers are of smaller order than optimal bandwidths for nonparametric estimation. However, it is difficult to adapt their abstract assumptions to practical settings.

We suggest applying the test for different choices of $\mathbf{h}$. Differences in observed critical values can be interpreted. Whereas test statistics with small choices of $\mathbf{h}$ will detect the appearance of wiggles of small length, large choices of $\mathbf{h}$ may detect better global deviations from linearity. So the inspection of the test statistic for different $\mathbf{h}$ gives an impression in which respect the function m differs significantly from linear functions.

Our approach can be generalized to tests of other parametric hypotheses on m; that is, $m \in \mathcal{M}$ for a parametric family $\mathcal{M} = \{m_\theta\colon\ \theta \in \Theta\}$. In particular, this includes tests of the hypothesis $m \equiv 0$. This test would check whether one of the coordinates of T has significant influence, and it can be used as tool in model choice.

Table 1. Relative Number of Rejections for the Test Statistics R_1, R_2, and R_3 Using the Bootstrap Method With $n^* = 200$ Compared to Relative Number of Rejections for Parametric LR Statistic (=LR(p)) and Semiparametric LR Statistic Using Approximate Degrees of Freedom (=LR(sp)); 500 Monte Carlo Replications.

α	.01	.05	.10	.15	.20
		n = 100, h = .6			
LR(p)	.010	.070	.138	.190	.248
LR(sp)	.014	.088	.220	.328	.428
R_1	.010	.052	.116	.178	.246
R_2	.010	.052	.116	.184	.250
R_3	.012	.052	.116	.178	.244
		n = 250, h = .5			
LR(p)	.012	.044	.098	.148	.194
LR(sp)	.020	.080	.158	.234	.322
R_1	.020	.052	.094	.138	.180
R_2	.020	.052	.096	.136	.184
R_3	.022	.052	.094	.142	.182
		n = 500, h = .4			
LR(p)	.008	.046	.094	.140	.198
LR(sp)	.020	.092	.164	.256	.338
R_1	.020	.056	.104	.160	.212
R_2	.020	.056	.104	.166	.214
R_3	.022	.054	.104	.158	.212

3.2 Testing Average Linearity

In case our test rejects the hypothesis of linearity, more insight into the reason for the rejection may be of interest. For the case of $q > 1$, we propose testing for average linearity in the direction of one covariate. For a given weight function $w(t_2, \ldots, t_q)$ with $\int w(t_2, \ldots, t_q)\, dt_2 \cdots dt_q = 1$, we consider the hypothesis that

$$\int m(t_1, \ldots, t_q) w(t_2, \ldots, t_q)\, dt_2 \cdots dt_q = a + bt_1 \quad \text{for all } t_1 \text{ and for fixed } a \text{ and } b. \tag{16}$$

Testing average linearity of m in t_1 is particularly appropriate in the following model, in which it is assumed that there is no interaction term of t_1 and $(t_2, \ldots, t_q)$:

$$m(t_1, \ldots, t_q) = m_1(t_1) + m_{2,\ldots,q}(t_2, \ldots, t_q) \quad \text{for some functions } m_1, m_{2,\ldots,q}. \tag{17}$$

(For a discussion of this additive model, see Buja et al. 1989; Hastie and Tibshirani 1990.) In this model, hypothesis (16) reduces to

$$m_1(t_1) = a + bt_1 \quad \forall\, t_1 \text{ and for fixed } a \text{ and } b. \tag{18}$$

Deviation from average linearity can be measured by the following test statistic:

$$R_4 = \min_{a,b} \sum_{i=1}^{n} \frac{[G'\{\mathbf{X}_i^T\hat{\beta} + \hat{m}(\mathbf{T}_i)\}]^2}{V[G\{\mathbf{X}_i^T\hat{\beta} + \hat{m}(\mathbf{T}_i)\}]} \times \{\hat{m}_1(\mathbf{T}_i) - a - b\mathbf{T}_i\}^2, \tag{19}$$

where $\hat{m}_1(t_1) = \int \hat{m}(t_1, \ldots, t_q) w(t_2, \ldots, t_q)\, dt_2 \cdots dt_q$. For the additive model (17), the nonparametric estimate $\hat{m}_1$ of the additive component m_1 has been considered by Härdle, Linton, and Severance-Lossin (1996), Fan, Härdle, and Mammen (1998), Linton and Nielsen (1995), and Tjøstheim and Auestad (1994). In a modified definition, the "marginal integration" in the calculation of $\hat{m}_1$ is replaced by a "marginal summation." For generalized additive models, asymptotics for the estimate $\hat{m}_1$ has been developed by Härdle, Huet, Mammen, and Sperlich (1998). These authors also provided a proof for asymptotic normality and consistency of bootstrap for a test statistic related to R_4.

4. SIMULATIONS AND APPLICATION

To verify the properties of our test procedure, we have conducted a small simulation study. We used simulated data from the following generalized (partially) linear model:

$$E(Y|\mathbf{X} = \mathbf{x}, T = t) = P(Y = 1|x_1, x_2, t) = F\{2x_1 + x_2 + m(t)\},$$

where F is the standard logistic distribution function $F(u) = 1/(1 + e^{-u})$; X_1, X_2, and T are independent; and X_1 and T have a uniform distribution on $[-1, 1]$. The variable X_2 is discrete and takes five values in $[-1, 1]$.

We performed simulations under the linearity hypothesis using $m(t) = t$. Sample sizes were $n = 100, 250$, and 500, and the number of replications in the bootstrap resampling used $n^* = 200$. The simulation results are based on

Table 2. Relative Number of Rejections Using Normal Approximations; 500 Monte Carlo Replications

α	.01	.05	.10	.15	.20
			$n = 100, h = .6$		
R_1	0	.002	.010	.012	.020
R_2	0	0	.006	.010	.012
R_3	.002	.001	.014	.024	.030
			$n = 250, h = .5$		
R_1	0	.008	.018	.026	.032
R_2	0	.006	.016	.022	.028
R_3	0	.014	.020	.030	.038
			$n = 500, h = .4$		
R_1	.004	.010	.016	.028	.034
R_2	.004	.010	.016	.026	.032
R_3	.006	.012	.020	.030	.036

500 replications. For smoothing in this section, the quartic kernel $K(u) = 15/16(1-u^2)^2 I(|u| \leq 1)$ was used.

Table 1 summarizes the results for $m(t) = t$. As can be seen, the bootstrap seems to be quite accurate for all three test statistics, at least for $\alpha \geq .05$.

As expected, the normal approximation of Theorem 1 can be quite inaccurate for small sample sizes and should not be used for the calculation of critical values of the test statistics R_1, R_2, and R_3. This can be seen from Table 2.

The values in Table 2 concern only the tail of the distributions of R_1, R_2, and R_3 and of the normal limit, given in Theorem 1. In the central region there are much larger differences between the distributions of R_1, R_2, and R_3 and the normal limit, given in Theorem 1, as can be seen in Figure 2. There density estimates for R_1, R_2, and R_3 (using the 500 Monte Carlo replications under the linear model $m(t) = t$) are plotted together with the limiting normal density. The normal limit and the distributions of the test statistics are nearly separated. (The density estimates for R_1, R_2, and R_3 are kernel estimates with bandwidth according to Silverman's rule of thumb; that is, $h = 1.06 \cdot 2.62 \cdot \hat{\sigma} \cdot n^{-1/5}$ for the quartic kernel. For better comparison, the normal density has been analogously convoluted with a quartic kernel.) Similar plots have been given by Härdle and Mammen (1993), who discussed a related test statistic for testing parametric versus nonparametric regression.

Finally, we ran our simulations with a function m consisting of a convex combination of the linear function $m(t) = t$ and the nonlinear function $m(t) = \cos(\pi t)$. Figure 3 shows the power functions of R_1 for these alternatives (black lines). The power has been plotted for four different significance levels. The power functions for R_2 and R_3 are almost the same and thus have been omitted. The dashed lines in Figure 3 show (simulated) power functions for a parametric likelihood ratio test $\mathrm{LR_p}$. The hypothesis "$m(x,t) = F\{c + x\beta + t\gamma\}$ for some β and γ" is tested against the alternative "$m(x,t) = F\{c + x\beta + t\gamma + \omega\cos(\pi t)\}$ for some c, β, γ and ω." In this setup R_1 achieves nearly the power of the parametric test $\mathrm{LR_p}$. We observed larger losses in other models.

For comparison, we have also included a likelihood ratio test of the parametric against semiparametric hypothesis, $\mathrm{LR_{sp}}$. Critical values have been calculated using chi-squared approximations and the definition of approximate degrees of freedom of Hastie and Tibshirani (1990). A more detailed description of this test has been given by Müller (1997). The gray curves in Figure 3 show the power of this test. It

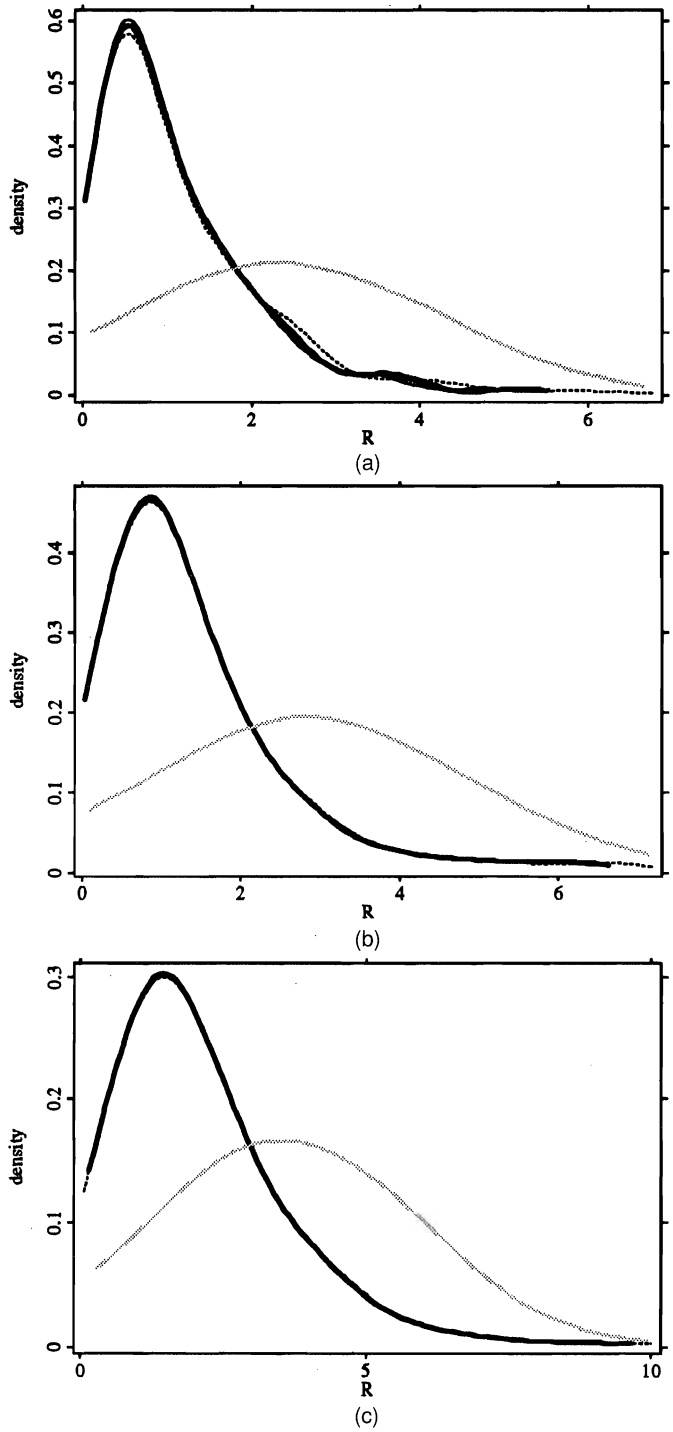


Figure 2. Density Estimates for R_1 (Thick Solid Line), R_2 (Thin Solid Line), R_3 (Thin Dashed Line), and Normal Density (Gray Line) (a) n = 100; (b) n = 250; (c) n = 500.

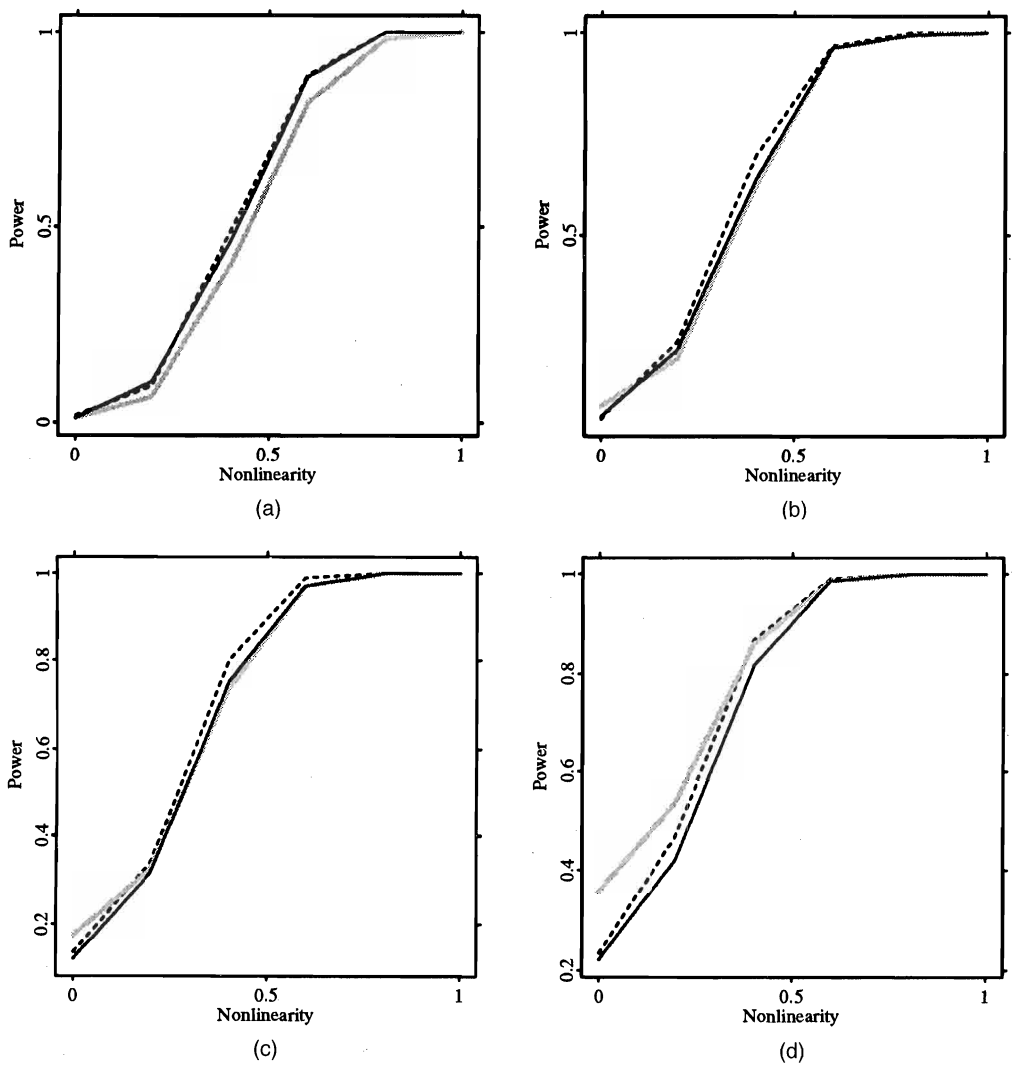


Figure 3. Power Functions of Test R_1 for α = .01 (a), .05 (b), .10 (c), .20 (d) (Black Solid Lines), x, t ∈ [-1, 1], and m(t) = (1 − ν)t + ν cos(πt), ν ∈ [0, 1], n = 500, h = .4, Compared to the Power of the Parametric LR Test LR_p (Dashed Lines) and the Power of the Semiparametric LR Test LR_{sp} Using Approximate Degrees of Freedom (Gray Lines).

achieves a power similar to that of R_1. However, it does not hold the nominal significance level under the hypothesis; see Table 1.

Let us now return to our introductory example on East–West German migration. Our interest in this subject has been inspired by an analysis of Burda (1993), who considered a sample of 3,710 East Germans surveyed in 1991 in the German Socio-Economic Panel (GSOEP 1991). Among other questions, the East German participants were asked if they could imagine moving to the Western part of Germany or West Berlin. As in Burda's study, we assign the value 1 to those who responded positively and 0 who did not. The economic model is based on the idea that a person will migrate if the utility (wage differential) exceeds the costs of migration. Of course, neither the wage differential nor the costs of migration are directly available, and hence proxy variables must be used. The original dataset of Burda (1993) contains 34 explanatory variables, with four of them continuous (age, income, rent, and job tenure) and the remainder dummy variables (sex, partner, homeowner, family/friends in West, and further variables on occupation, city size, region, and education).

It turns out that regional variables have an important impact on the responses. For instance, the estimation is particularly difficult for East Germans living in East Berlin, because other reasons besides the wage differential compared to costs may influence the intention to migrate. Also, the variables, which are most important, differ slightly between the five Eastern German states (plus East Berlin).

Table 3. Descriptive Statistics for Migration Data

	Yes	*No*		
Y, migration intention	39.9	60.1		
X_1, family/friends in West	88.8	11.2		
X_2, unemployed/job loss certain	21.1	78.9		
X_3, city size 10,000–100,000	35.8	64.2		
X_4, female	50.2	49.8		
	min	*max*	*Mean*	*SD*
X_5, age (yr)	18	65	39.93	12.89
T, household income (DM)	400	4,000	2,262.22	769.82

NOTE: Sample from Mecklenburg-Vorpommern, n = 402, results in percentages.

Table 4. Logit Coefficients and Coefficients in a Generalized Partially Linear Model for Migration Data

	Linear (logit)		Partial linear	
	Coefficient	(t value)	Coefficient	(t value)
Y, constant	−.358	(−.68)		
X_1, family/friends in West	.589	(1.54)	.599	(1.56)
X_2, unemployed/job loss certain	.780	(2.81)	.800	(2.87)
X_3, city size 10,000–100,000	.822	(3.39)	.842	(3.47)
X_4, female	−.388	(−1.68)	−.402	(−1.73)
X_5, age	−3.364	(−6.93)	−3.329	(−6.86)
T, household income	1.084	(1.90)		

NOTE: Sample from Mecklenburg-Vorpommern, n = 402, h = .3.

Unemployment, for example, plays a stronger role in the Northern, less industrialized part of East Germany. In the following we give the estimation results for Mecklenburg–Vorpommern (in the very north of Eastern Germany), which leads to a sample size of $n = 402$. We summarize some descriptive statistics in Table 3.

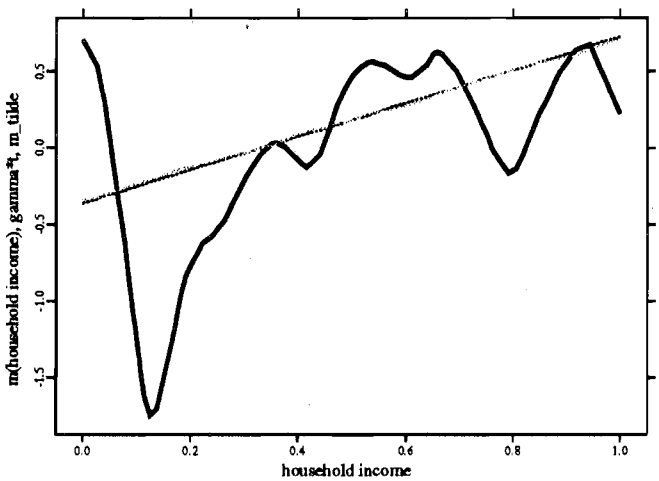


Figure 4. The Influence $m(t)$ of Household Income on Migration Intention. Nonparametric fit (thick black line), linear fit (thin black dashed line), and "biased" parametric estimate $\tilde{m}$ (thin gray dashed line); n = 402, bandwidths h = .1 (a) and h = .5 (b).

Table 5. Observed Significance Levels for the Linearity Test for Migration Data, n = 402, With n^* = 400 Bootstrap Replications

h	.1	.2	.3	.4	.5
R_1	.150	.065	.042	.045	.062
R_2	.122	.067	.042	.045	.062
R_3	.217	.075	.042	.042	.065
LR(sp)	.053	.066	.048	.035	

Table 4 shows the results of a logit fit, using a subset of covariates chosen previously by a model selection procedure based on logit models. For simplicity, both continuous variables (age and household income) have been linearly transformed to [0, 1]. The migration intention is definitely determined by age. However, unemployment, city size, and household income also are highly significant.

In a further analysis of this dataset, we fitted a generalized additive model with logit link. We used the same subset of covariates that were chosen by the parametric model selection procedure. This choice was motivated by our experience that typically values of parametric coefficients (and their t values) change only slightly if other covariates are modeled nonparametrically (see also Tables 4 and 7). So we conjecture that in semiparametric models, parametric model choice procedures will work well for the choice of the parametric components. Clearly, nonlinear influences will not be recognized for parametric and nonparametric components. (For nonparametric tests on the significance of covariates, see also the remark at the end of Sec. 3.1 and Härdle et al. 1998.)

In a first step, we modeled the influence of the age and income variables as nonparametric functions. Because age showed an almost perfectly linear influence, in a second step we modeled only the influence of household income as a nonparametric function. The coefficients for the parametric covariates are given in Table 4. The resulting fit, $\hat{m}$ (using bandwidth $h = .3$), for the function m is that shown in Figure 1, together with the linear fit (thin black dashed line) and the "biased" parametric fit $\tilde{m}$ (thin gray dashed line). Recall that the estimate $\tilde{m}$ is expected to be approximately equal to the sum of the parametric estimate and the bias of $\hat{m}$.

Figure 4 shows the functions $\hat{m}$ and $\tilde{m}$ (together with the linear fit) for bandwidths $h = .1$ and $h = .5$. The nonparametric estimate $\hat{m}$ in the migration example obviously seems to be a nonlinear function. However, it is difficult

Table 6. Descriptive Statistics for Credit Data; Sample for Credits for Cars, n = 284; results in percentages

	Yes	No		
Y, credit worthy	73.6	26.4		
X_1, previous credits okay	36.6	63.4		
X_2, employed	73.2	26.8		
	min	max	Mean	SD
X_3, duration (mo)	4	54	21.75	10.55
T_1, amount (DM)	428	14,179	3,902.31	2,621.95
T_2, age (yr)	19	75	34.16	10.81

Table 7. Logit Coefficients and Coefficient in Partially Linear Fit for Credit Scoring, n = 284

	Linear (logit)		Partial linear	
	Coefficient	t value	Coefficient	t value
Y, constant	1.480	2.78		
X_1, previous credits okay	.992	3.07	1.017	3.06
X_2, employed (%)	.526	1.67	.490	1.53
X_3, duration (mo)	−.035	−2.01	−.041	−2.43
T_1, amount (DM)	−1.080	−1.05		
T_2, age (yr)	.754	1.09		

to judge the significance of the nonlinearity. In general, it cannot be excluded that the difference between the nonparametric and the linear fit may be caused by boundary and bias problems of $\hat{m}$.

Table 5 shows the results of the application of our tests from Section 3. The number of bootstrap simulations is always chosen as $n^* = 400$. We observe that all three tests R_1, R_2, and R_3 show nearly the same behavior. The observed significance levels are given for different choices of the bandwidth h. Linearity is rejected (at the 5% level) only for bandwidths .3 and .4. The different behavior of the test for different h gives some indication on possible deviations of m from a linear function. The appearance of wiggles of small length is not significant, see Figure 4(a). However, it seems that the global shape of m cannot be well approximated by linear functions. This result is in accordance with the estimate in Figures 1 and 4(b), where a saturation of the intention to migrate appears for the upper third of the data.

At the end of this section we present the application of our test statistic in a binary choice regression with a two-dimensional nonparametric function m. The data are from a dataset on credit scoring (Fahrmeir and Hamerle 1984; Fahrmeir and Tutz 1994). The goal is to find factors related to credit worthiness. We used the subsample on car loans, which has a sample size of $n = 284$ out of 1,000. Table 6 presents some descriptive statistics for this subsample and a selection of covariates. The covariate "previous credit okay" indicates that previous loans were repaid without problems. The variable "employed" takes on the value 1 if the person taking the loan has been employed by the same employer for ≥ 1 year. In the following statistical analysis we took logarithms of amount and age and transformed these values linearly to the interval [0, 1].

A parametric logit model leads to the parameter estimates listed in Table 7. The coefficients of previous credits, employment, duration, and amount of credit have the expected

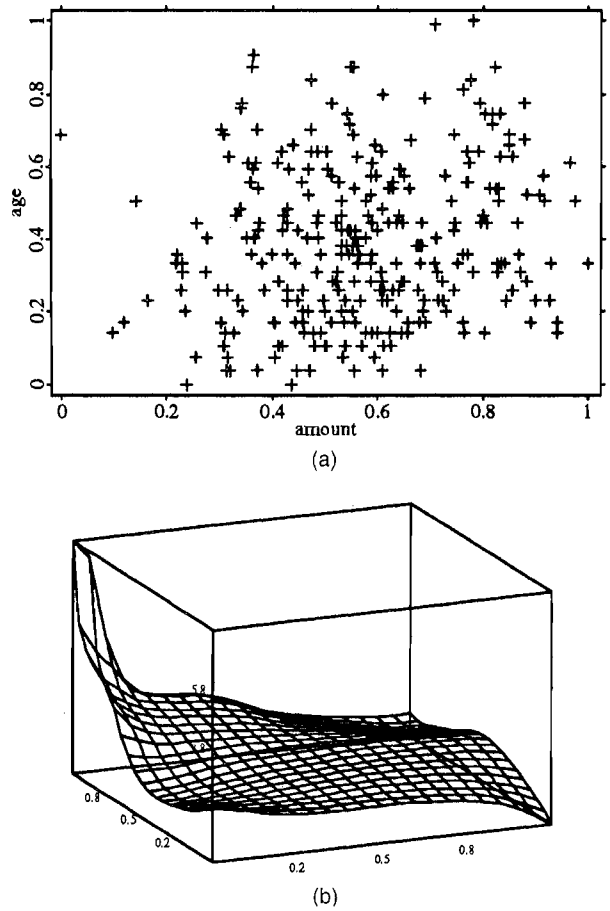


Figure 5. Scatterplot for Amount of Credit and Age (a); Influence $\hat{m}(t_1; t_2)$ of Amount and Age on Credit Worthiness (b), n = 284.

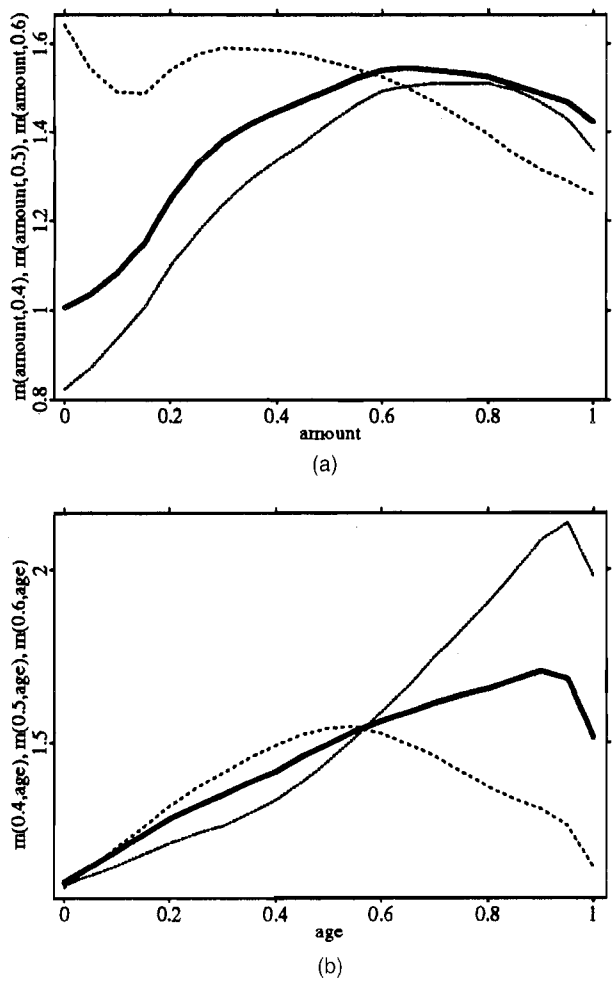


Figure 6. Influence of Amount on Credit Worthiness for Fixed Age (a); Influence of Age on Credit Worthiness for Fixed Amount (b); n = 284.

Table 8. Observed Significance Levels for the Linearity Test for Credit Scoring, n = 284, With 400 Bootstrap Replications

h	.2	.3	.4	.5	.6
R_1	.03	.09	.08	.13	.32
R_2	.01	.07	.07	.13	.32
R_3	.44	.38	.11	.12	.30

sign. The age variable shows a (globally) positive influence in the logit fit; this will change together with the amount variable in the semiparametric fit. Note also, that both coefficients for amount and age are not significant at the 10% level.

In a next step we fitted a generalized partially linear model to the data. Influence of amount and age has been fitted nonparametrically. The other variables have been modeled as linear covariates. For duration, this has been done because typically it is divisible by 6 months. Figure 5 shows a scatterplot of the two variables, amount and age and the two-variate estimate $\hat{m}$ (using a bandwidth $h = .4$ in both dimensions). It is difficult to check $\hat{m}$ graphically for significant deviations from linearity. The big peak of $\hat{m}$ is caused by only a few observations (as can be seen from the scatterplot). For a closer inspection of $\hat{m}$, Figure 6 shows the influence of amount and age separately. In the figure, one variable is held fixed at levels .4 (short dashes), .5 (thick line), and .6 (long dashes). For age, these levels correspond to 32.9, 37.75, and 43.30 years. For credit amounts, the corresponding original values are DM 1,735.90, 2,463.46, and 3,495.95. So, obviously, a higher amount of credit seems to get more risky in conjunction with higher age. Also, younger people seem to get less risky with increasing credit amount. Neither of these possible conclusions could be seen from the parametric logit fit.

Table 8 gives the observed significance levels of our test statistics for the credit data. For the test statistics R_1 and R_2, linearity is rejected at level .10 for $h < .5$. For $h = .2$, the rejection has even higher significance. This suggests that the deviations from linearity are more locally concentrated. Our inference in both applications was based on inspection of several tests. To get a resulting p value, one could consider a combination of the test statistics for several bandwidths and could calculate critical values for this combined statistic, again by bootstrap.

APPENDIX: COMPUTATIONAL AND MATHEMATICAL DETAILS

A.1 Computational Remarks

In this section we indicate how the estimates in (6) and (7) can be numerically computed. The following algorithm corresponds to that proposed by Severini and Staniswalis (1994), example 3, for the special case of a logistic link function where

$$\eta_j(\boldsymbol{\beta}) = \hat{m}_{\boldsymbol{\beta}}(T_j)$$

and

$$L_i(u) = Q\{G(u); Y_i\}. \tag{A.1}$$

Note that we have

$$L_i'(u) = \frac{Y_i - G(u)}{V(G(u))}\, G'(u) \tag{A.2}$$

and

$$L_i''(u) = \{Y_i - G(u)\}\left[\frac{G''(u)}{V(G(u))} - \frac{V'(G(u))G'(u)^2}{V(G(u))^2}\right] - \frac{G'(u)^2}{V(G(u))}. \tag{A.3}$$

Then maximizing the smoothed quasi-likelihood (5) requires solving

$$0 = \sum_{i=1}^{n} L_i'\{\mathbf{X}_i^T\boldsymbol{\beta} + \eta_j(\boldsymbol{\beta})\}K_{\mathbf{h}}(\mathbf{T}_i - \mathbf{T}_j). \tag{A.4}$$

Differentiation of (A.4) leads to

$$0 = \sum_{i=1}^{n} L_i''\{\mathbf{X}_i^T\boldsymbol{\beta} + \eta_j(\boldsymbol{\beta})\}K_{\mathbf{h}}(\mathbf{T}_i - \mathbf{T}_j)\{\mathbf{X}_i + \eta_j'(\boldsymbol{\beta})\}.$$

This gives

$$\eta_j'(\boldsymbol{\beta}) = \frac{-\sum_{i=1}^{n} L_i''\{\mathbf{X}_i^T\boldsymbol{\beta} + \eta_j(\boldsymbol{\beta})\}K_{\mathbf{h}}(\mathbf{T}_i - \mathbf{T}_j)X_i}{\sum_{i=1}^{n} L_i''\{\mathbf{X}_i^T\boldsymbol{\beta} + \eta_j(\boldsymbol{\beta})\}K_{\mathbf{h}}(\mathbf{T}_i - \mathbf{T}_j)}. \tag{A.5}$$

For $\boldsymbol{\beta} = \hat{\boldsymbol{\beta}}$, it holds that

$$0 = \sum_{i=1}^{n} L_i'\{\mathbf{X}_i^T\boldsymbol{\beta} + \eta_i(\boldsymbol{\beta})\}\{\mathbf{X}_i + \eta_i'(\boldsymbol{\beta})\}. \tag{A.6}$$

Equations (A.4), (A.5), and (A.6) suggest the following iterative Newton–Raphson-type algorithm to find $\hat{\boldsymbol{\beta}}$ and $\hat{m}(\mathbf{T}_j)$, where $j = 1, \ldots, n$.

- Start with $\hat{\boldsymbol{\beta}}^0 = \tilde{\boldsymbol{\beta}}, \hat{\eta}_j^0 = \mathbf{T}_j^T\tilde{\boldsymbol{\gamma}}$.
- Determine the iteration $k \to k+1$ by the stepwise application of the following two equations:

$$0 = \sum_{i=1}^{n} L_i'(\mathbf{X}_i^T\hat{\boldsymbol{\beta}}^k + \hat{\eta}_j^k)K_{\mathbf{h}}(\mathbf{T}_i - \mathbf{T}_j) + L_i''(\mathbf{X}_i^T\hat{\boldsymbol{\beta}}^k + \hat{\eta}_j^k)K_{\mathbf{h}}(\mathbf{T}_i - \mathbf{T}_j)(\hat{\eta}_j^{k+1} - \hat{\eta}_j^k)$$

and

$$0 = \sum_{i=1}^{n} L_i'(\mathbf{X}_i^T\hat{\boldsymbol{\beta}}^k + \hat{\eta}_i^{k+1})\tilde{\mathbf{X}}_i^k + L_i''(\mathbf{X}_i^T\hat{\boldsymbol{\beta}}^k + \hat{\eta}_i^{k+1})\tilde{\mathbf{X}}_i^k\tilde{\mathbf{X}}_i^{k^T}(\tilde{\boldsymbol{\beta}}^{k+1} - \tilde{\boldsymbol{\beta}}^k),$$

where

$$\tilde{\mathbf{X}}_j^k = \mathbf{X}_j - \frac{\sum_{i=1}^{n} L_i''(\mathbf{X}_i^T\hat{\boldsymbol{\beta}}^k + \hat{\eta}_j^{k+1})K_{\mathbf{h}}(\mathbf{T}_i - \mathbf{T}_j)\mathbf{X}_i}{\sum_{i=1}^{n} L_i''(\mathbf{X}_i^T\hat{\boldsymbol{\beta}}^k + \hat{\eta}_j^{k+1})K_{\mathbf{h}}(\mathbf{T}_i - \mathbf{T}_j)}.$$

Then $\hat{m}^k(\mathbf{T}_j) = \hat{\eta}_j^k$.

Alternatively, the functions $L_i''(u)$ can be replaced by their expectations, $-G'(u)^2/V\{G(u)\}$, to obtain a Fisher scoring–type procedure.

A.2 Assumptions

We now state the assumptions used in the results of Section 3. In the following, the underlying parameters are denoted by $\boldsymbol{\beta}_0, \boldsymbol{\gamma}_0$ and m_0. We use the notation

$$h_{\max} = \max\{h_1, \ldots, h_q\},$$

Journal of the American Statistical Association, 93 , 1461-1474

$$h_{\text{prod}} = h_1 \cdot \ldots \cdot h_q,$$
$$\rho = h_{\max}^2 + (nh_{\text{prod}})^{-1/2},$$

and

$$\tau = h_{\max} + (nh_{\text{prod}})^{-1/2}.$$

For the asymptotic expansions, we make the following assumptions:

A1. $(\mathbf{X}_1, \mathbf{T}_1, Y_1), \ldots, (\mathbf{X}_n, \mathbf{T}_n, Y_n)$ are iid tuples with values in $\mathbb{R}^q \times \mathbb{R}^p \times \mathbb{R}$.

A2. $E(Y_i|\mathbf{X}_i, \mathbf{T}_i) = G\{\mathbf{X}_i^T\boldsymbol{\beta}_0 + m_0(\mathbf{T}_i)\}$ with $\boldsymbol{\beta}_0 \in \mathbb{R}^p$. The conditional variance var $(Y_i|\mathbf{T}_i = \mathbf{t})$ has a bounded second derivative. Furthermore, the Laplace transform $E \exp t|Y_i|$ is finite for $t > 0$ small enough.

A3. $\mathbf{X}_i^T\boldsymbol{\beta}_0 + m_0(\mathbf{T}_i)$ has compact support S; $\mathbf{X}_i$ and $\mathbf{T}_i$ have compact convex support S_X and S_T; and $\mathbf{T}_i$ has a twice continuously differentiable density f_T with $\inf_{\mathbf{t}\in S_T} f_T(\mathbf{t}) > 0$.

A4. There exists an $\delta > 0$ such that $G^{(k)}(u)$, where $k = 1, \ldots, 3$, and $G'(u)^{-1}$ is bounded on $u \in S^\delta = \{v\colon \exists v' \in S$ with $|v' - v| \le \delta\}$. Furthermore, V^{-1}, V', and V'' are bounded on $G(S^\delta)$.

A5. The kernel K is a product kernel $K(\mathbf{u}) = K_1(u_1), \cdot \ldots \cdot, K_q(u_q)$. The kernels K_j are symmetric probability densities with compact support (e.g., $[-1, 1]$), where $j = 1, \ldots, q$.

A6. The estimate $\hat{\boldsymbol{\beta}}$ is defined as $\arg\max_{\boldsymbol{\beta}:\|\boldsymbol{\beta}-\boldsymbol{\beta}_0\|\le\rho} \mathcal{L}(\hat{m}_{\boldsymbol{\beta}}, \boldsymbol{\beta})$. For a δ_n with $\delta_n \to 0$, the estimate $\hat{m}_{\boldsymbol{\beta}}(\mathbf{t})$ is defined as $\arg\max_{\eta:|\eta-m_0(\mathbf{t})|\le\delta_n} \sum_{i=1}^n L_i(\mathbf{X}_i^T\boldsymbol{\beta} + \eta)K_{\mathbf{h}}(\mathbf{T}_i - \mathbf{t})$.

A7. $E[L_1'\{\mathbf{X}_1^T\boldsymbol{\beta}_0 + m_0(\mathbf{T}_1)\}|\mathbf{T}_1 = \mathbf{t}]$ and $E[L_1''\{\mathbf{X}_1^T\boldsymbol{\beta}_0 + m_0(\mathbf{T}_1)\}X_1|\mathbf{T}_1 = \mathbf{t}]$ are twice continuously differentiable functions for $\mathbf{t} \in S_T$.

A8. $h_{\text{prod}} n^{1/2} (\log n)^{-1} \to \infty$ and $h_{\max} = o(n^{-1/8}(\log n)^{-1/4})$.

A.3 Proofs

In this section we always assume that A1–A7 hold. The following lemmas give the stochastic expansions for $\hat{\boldsymbol{\beta}}$ and $\hat{m}$. Recall that the set S_T was the (compact) support of T_i. We denote $S_T^- = \{t \in S_T\colon \mathbf{t}+\boldsymbol{\eta} \in S_T$ for all $\boldsymbol{\eta}$ with $|\eta_j| \le h_j (j = 1, \ldots, q)\}$ and $S_T^{\mathbf{h}} = S_T \setminus S_T^-$. Furthermore, define

$$S_{i,1} = L_i'\{\mathbf{X}_i^T\boldsymbol{\beta}_0 + m_0(\mathbf{T}_i)\},$$
$$S_{i,2} = L_i''\{\mathbf{X}_i^T\boldsymbol{\beta}_0 + m_0(\mathbf{T}_i)\},$$
$$\tilde{\mathbf{X}}_i = \mathbf{X}_i - \{E[S_{i,2}|\mathbf{T}_i]\}^{-1}E[S_{i,2}\mathbf{X}_i|\mathbf{T}_i],$$

and

$$w_i(\mathbf{t}) = K_{\mathbf{h}}(\mathbf{t} - \mathbf{T}_i)\left\{n^{-1}\sum_{j=1}^n K_{\mathbf{h}}(\mathbf{t} - \mathbf{T}_j)\right\}^{-1}.$$

Lemma A.1

a. For all $C > 0$, it holds that

$$\sup_{\substack{\mathbf{t}\in S_T^- \\ \|\boldsymbol{\beta}-\boldsymbol{\beta}_0\|\le C\rho}} \left|\hat{m}_{\boldsymbol{\beta}}(\mathbf{t}) - \left(m(\mathbf{t}) - \{E(S_{1,2}|\mathbf{T}_1 = \mathbf{t})\}^{-1} \times \left[\frac{1}{n}\sum_{i=1}^n w_i(\mathbf{t})L_i'\{\mathbf{X}_i^T\boldsymbol{\beta}_0 + m_0(\mathbf{t})\} + E(S_{1,2}\mathbf{X}_1^T|\mathbf{T}_1 = \mathbf{t})(\boldsymbol{\beta} - \boldsymbol{\beta}_0)\right]\right)\right| = O_p(\rho^2 \log n).$$

b. The supremum in (a) taken over $\mathbf{t} \in S_T^{\mathbf{h}}, \|\boldsymbol{\beta} - \boldsymbol{\beta}_0\| \le C\rho$ is of stochastic order $O_p(\tau^2)$.

Proof. We prove only statement (a). Choose $C > 0$. We have, for $\mathbf{t} \in S_T^-, \|\boldsymbol{\beta} - \boldsymbol{\beta}_0\| \le C\rho$,

$$\sum_{i=1}^n L_i'\{\mathbf{X}_i^T\boldsymbol{\beta} + \hat{m}_{\boldsymbol{\beta}}(\mathbf{t})\}K_{\mathbf{h}}(\mathbf{t} - \mathbf{T}_i) = 0. \tag{A.7}$$

This follows from

$$\sup \sum_{i=1}^n L_i''(\mathbf{X}_i^T\boldsymbol{\beta} + \eta)K_{\mathbf{h}}(\mathbf{t} - \mathbf{T}_i) < 0 \tag{A.8}$$

with probability tending to one, where the supremum runs over $|\eta - m_0(\mathbf{t})| \le \delta_n, \mathbf{t} \in S_T^-$, and $\boldsymbol{\beta}$ with $\|\boldsymbol{\beta} - \boldsymbol{\beta}_0\| \le C\rho$.

Note that (A.8) implies that if we find an $\eta_{\boldsymbol{\beta}}(t)$ with $|\eta_{\boldsymbol{\beta}}(\mathbf{t}) - m_0(\mathbf{t})| \le \delta_n$ and

$$\sum_{i=1}^n L_i'\{\mathbf{X}_i^T\boldsymbol{\beta} + \eta_{\boldsymbol{\beta}}(\mathbf{t})\}K_{\mathbf{h}}(\mathbf{t} - \mathbf{T}_i) = 0,$$

then with probability tending (uniformly) to 1, we get $\hat{m}_{\boldsymbol{\beta}}(\mathbf{t}) = \eta_{\boldsymbol{\beta}}(\mathbf{t})$. Inequality (A.8) can be shown again by using that for $\delta > 0$ small enough,

$$\sup_{\eta\in I_n', \boldsymbol{\beta}\in I_n'', t\in I'''} \left|\frac{1}{n}\sum_{i=1}^n L_i''\{\mathbf{X}_i^T\boldsymbol{\beta} + \eta\}K_{\mathbf{h}}(\mathbf{t} - \mathbf{T}_i) - E[L_i''\{\mathbf{X}_i^T\boldsymbol{\beta} + \eta\}K_{\mathbf{h}}(\mathbf{t} - \mathbf{T}_i)]\right| = o_P(1), \tag{A.9}$$

$$\sup_{1\le i\le n} \sup_{u\in S^\delta, \mathbf{t}\in\mathbb{R}^q} |L_i'''(u)K_{\mathbf{h}}(\mathbf{t} - \mathbf{T}_i)| = O_P(h_{\text{prod}}^{-1}\log n), \tag{A.10}$$

and

$$\sup_{1\le i\le n} \sup_{u\in S^\delta, t\in\mathbb{R}^q} \|L_i'''(u)K_{\mathbf{h}}'(\mathbf{t} - \mathbf{T}_i)\| = O_P(h_{\text{prod}}^{-1}h_{\max}^{-1}\log n), \tag{A.11}$$

where the supremum in (A.9) runs over grids I', I'', and I''' with polynomially many elements. Equality (A.9) follows by application of the Markov inequality. Note that Y_i has bounded Laplace transform; see Assumption A2. Equalities (A.10)–(A.11) follow from $\max_{1\le i\le n}|Y_i| = O_P(\log n)$. This can be shown again by using that Y_i has bounded Laplace transform. For the proof of claim (A.9), one applies

$$E[L_i''\{\mathbf{X}_i^T\boldsymbol{\beta} + \eta\}K_{\mathbf{h}}(\mathbf{t} - \mathbf{T}_i)] = -E\frac{G'\{\mathbf{X}_i^T\boldsymbol{\beta} + \eta\}^2}{V[G\{\mathbf{X}_i^T\boldsymbol{\beta} + \eta\}]}.$$

Equation (A.7) implies that

$$\begin{aligned} 0 = {} & \frac{1}{n}\sum_{i=1}^n w_i(\mathbf{t})L_i'\{\mathbf{X}_i^T\boldsymbol{\beta}_0 + m_0(\mathbf{t})\} \\ & + \frac{1}{n}\sum_{i=1}^n w_i(\mathbf{t})L_i''\{\mathbf{X}_i^T\boldsymbol{\beta}_0 + m_0(\mathbf{t})\} \\ & \times \{\hat{m}_{\boldsymbol{\beta}}(\mathbf{t}) - m_0(\mathbf{t}) + \mathbf{X}_i^T(\boldsymbol{\beta} - \boldsymbol{\beta}_0)\} \\ & + R_1(\boldsymbol{\beta}, \mathbf{t})[\{\hat{m}_{\boldsymbol{\beta}}(\mathbf{t}) - m_0(\mathbf{t})\}^2 + \rho^2], \end{aligned} \tag{A.12}$$

with

$$\sup_{\substack{\mathbf{t}\in S_T^- \\ \|\boldsymbol{\beta}-\boldsymbol{\beta}_0\|\le C\rho}} |R_1(\boldsymbol{\beta}, \mathbf{t})| \le C_1 \quad \text{(a.s)},$$

(1998) Härdle, W., Mammen, E. and Müller, M.
Testing Parametric versus Semiparametric Modelling in Generalized Linear Models.

for a constant $C_1 > 0$ for n large enough. Furthermore, we have $|\hat{m}_\beta(\mathbf{t}) - m_0(\mathbf{t})| \le \delta_n \to 0$; see A6. This implies that

$$\begin{aligned}\hat{m}_\beta(\mathbf{t}) &= m_0(\mathbf{t}) - \left[\frac{1}{n}\sum_{i=1}^n w_i(\mathbf{t}) L_i''\{\mathbf{X}_i^T\beta_0 + m_0(\mathbf{t})\}\right]^{-1} \\ &\quad \times \left[\frac{1}{n}\sum_{i=1}^n w_i(\mathbf{t}) L_i'\{\mathbf{X}_i^T\beta_0 + m_0(\mathbf{t})\}\right. \\ &\quad \left. + \frac{1}{n}\sum_{i=1}^n w_i(\mathbf{t}) L_i''\{\mathbf{X}_i^T\beta_0 + m_0(\mathbf{t})\}\mathbf{X}_i^T(\beta - \beta_0)\right] \\ &\quad + R_2(\beta, \mathbf{t})\rho^2 \log n, \qquad \text{(A.13)}\end{aligned}$$

where

$$\sup_{\substack{t \in S_T^- \\ \|\beta - \beta_0\| \le C\rho}} |R_2(\beta, \mathbf{t})| = O_p(1).$$

For (A.13), it has been used that

$$\sup_{\mathbf{t} \in S_T^-} \left|\frac{1}{n}\sum_{i=1}^n w_i(\mathbf{t}) L_i'\{X_i^T\beta_0 + m_0(\mathbf{t})\}\right| = O_p(\rho\sqrt{\log n}).$$

This follows from

$$\sup_{\mathbf{t} \in S_T^-} \left|\frac{1}{n}\sum_{i=1}^n K_\mathbf{h}(\mathbf{t} - \mathbf{T}_i)[L_i'\{\mathbf{X}_i^T\beta_0 + m_0(\mathbf{t})\} - L_i'\{\mathbf{X}_i^T\beta_0 + m_0(\mathbf{T}_i)\}]\right| = O_p(\rho)$$

and

$$\sup_{\mathbf{t} \in S_T^-} \left|\frac{1}{n}\sum_{i=1}^n K_\mathbf{h}(\mathbf{t} - \mathbf{T}_i) L_i'\{\mathbf{X}_i^T\beta_0 + m_0(\mathbf{T}_i)\}\right| = O_p(\rho\sqrt{\log n}).$$

Recall that $E[L_i'\{\mathbf{X}_i^T\beta_0 + m_0(\mathbf{T}_i)\}|\mathbf{X}_i, \mathbf{T}_i] = 0$. For the statement of the lemma, it remains to show that

$$\sup_{\mathbf{t} \in S_T^-} \left|\frac{1}{n}\sum_{i=1}^n w_i(\mathbf{t}) L_i''\{\mathbf{X}_i^T\beta_0 + m_0(\mathbf{t})\} - E(S_{1,2}|\mathbf{T}_1 = \mathbf{t})\right| = O_p(\rho\sqrt{\log n}) \qquad \text{(A.14)}$$

and

$$\sup_{\mathbf{t} \in S_T^-} \left\|\frac{1}{n}\sum_{i=1}^n w_i(\mathbf{t}) L_i''\{\mathbf{X}_i^T\beta_0 + m_0(\mathbf{t})\}\mathbf{X}_i^T - E(S_{1,2}\mathbf{X}_1^T|\mathbf{T}_1 = \mathbf{t})\right\| = O_p(\rho\sqrt{\log n}). \qquad \text{(A.15)}$$

For the proof of (A.13), note first that

$$\sup_{\mathbf{t} \in S_T^-} \left|\frac{1}{n}\sum_{i=1}^n w_i(\mathbf{t})[L_i''\{\mathbf{X}_i^T\beta_0 + m_0(\mathbf{t})\} - L_i''\{\mathbf{X}_i^T\beta_0 + m_0(\mathbf{T}_i)\}]\right| = O_p(\rho);$$

see A4. With the help of A7,

$$\sup_{\mathbf{t} \in S_T^-} \left|\frac{1}{n}\sum_{i=1}^n w_i(\mathbf{t}) L_i''\{\mathbf{X}_i^T\beta_0 + m_0(\mathbf{T}_i)\} - E(S_{1,2}|\mathbf{T}_1 = \mathbf{t})\right| = O_p(\rho\sqrt{\log n}).$$

Equation (A.14) can be shown similarly.

Lemma A.2.

a. For all $C > 0$, it holds that

$$\sup_{\substack{\mathbf{t} \in S_T^- \\ \|\beta - \beta_0\| \le C\rho}} \left\|\frac{\partial \hat{m}_\beta(\mathbf{t})}{\partial\beta} + \{E(S_{1,2}|\mathbf{T}_1 = \mathbf{t})\}^{-1} \times E(S_{1,2}\mathbf{X}_1|\mathbf{T}_1 = \mathbf{t})\right\| = O_p(\rho\sqrt{\log n}).$$

b. The supremum in a taken over $\mathbf{t} \in S_T^h, \|\beta - \beta_0\| \le C\rho$ is of stochastic order $O_p(\tau)$.

Proof. Lemma A.2 can be proved similarly to Lemma A.2; use that

$$\sum_{i=1}^n L_i''\{\mathbf{X}_i^T\beta + \hat{m}_\beta(\mathbf{t})\}K_\mathbf{h}(\mathbf{t} - \mathbf{T}_i)\frac{\partial}{\partial\beta}\hat{m}_\beta(\mathbf{t}) + \sum_{i=1}^n L_i''\{\mathbf{X}_i^T\beta + \hat{m}_\beta(\mathbf{t})\}X_i K_\mathbf{h}(\mathbf{t} - \mathbf{T}_i) = 0. \qquad \text{(A.15)}$$

Lemma A.3. For the estimate $\hat{\beta}$, the following stochastic expansion holds:

$$\hat{\beta} = \beta_0 + \{E(S_{1,2}\tilde{\mathbf{X}}_1\tilde{\mathbf{X}}_1^T)\}^{-1}\frac{1}{n}\sum_{i=1}^n S_{i,1}\tilde{\mathbf{X}}_i + O_p(\rho^2 \log n).$$

Proof. We show that with probability tending to 1 there exists a solution β with $\|\beta - \beta_0\| \le \rho$ of the following equation and that (with probability tending to 1) this solution is unique:

$$\frac{\partial}{\partial\beta}\sum_{i=1}^n L_i\{\mathbf{X}_i^T\beta + \hat{m}_\beta(\mathbf{T}_i)\} = 0. \qquad \text{(A.16)}$$

Expansion of the left side of (A.16) gives, with the help of Lemma A.2,

$$\begin{aligned}0 &= \frac{1}{n}\sum_{i=1}^n L_i'\{\mathbf{X}_i^T\beta + \hat{m}_\beta(\mathbf{T}_i)\}\left[\mathbf{X}_i + \frac{\partial}{\partial\beta}\hat{m}_\beta(\mathbf{T}_i)\right] \\ &= \frac{1}{n}\sum_{i=1}^n L_i'\{\mathbf{X}_i^T\beta_0 + m_0(\mathbf{T}_i)\}\left[\mathbf{X}_i + \frac{\partial}{\partial\beta}\hat{m}_\beta(\mathbf{T}_i)\right] \\ &\quad + \frac{1}{n}\sum_{i=1}^n L_i''\{\mathbf{X}_i^T\beta_0 + m_0(\mathbf{T}_i)\}\tilde{\mathbf{X}}_i\mathbf{X}_i^T(\beta - \beta_0) \\ &\quad + \frac{1}{n}\sum_{i=1}^n L_i''\{\mathbf{X}_i^T\beta_0 + m_0(\mathbf{T}_i)\}\tilde{\mathbf{X}}_i[\hat{m}_\beta(\mathbf{T}_i) - m_0(\mathbf{T}_i)] \\ &\quad + O_p(\rho^2 \log n). \qquad \text{(A.17)}\end{aligned}$$

Journal of the American Statistical Association, 93 , 1461-1474

This expansion holds uniformly for β with $\|\beta - \beta_o\| \le \rho$. For instance, it has been used that

$$\sup_{\substack{\mathbf{t}\in S_i^- \\ \|\beta-\beta_0\|\le\rho}} |\hat{m}_\beta(\mathbf{t}) - m(\mathbf{t})| = O_p(\rho\sqrt{\log n}).$$

This follows by standard techniques from Lemma A.1. By expansion of (A.15), it can be shown that

$$\frac{1}{n}\sum_{i=1}^n L_i'\{\mathbf{X}_i\beta_0 + m_0(\mathbf{T}_i)\}\left[\mathbf{X}_i + \frac{\partial}{\partial\boldsymbol{\beta}}\hat{m}_\beta(\mathbf{T}_i)\right] = \frac{1}{n}\sum_{i=1}^n L_i'\{\mathbf{X}_i^T\beta_0 + m_0(\mathbf{T}_i)\}\tilde{\mathbf{X}}_i + O_p(\rho^2).$$

Plugging this into the right side of (A.17) and replacing averages by their expectations gives that (with probability tending to 1) there exists a solution $\beta = \bar{\beta}$ of (A.16) with

$$\bar{\beta} = \beta_0 + \{E(S_{1,2}\tilde{\mathbf{X}}_1\tilde{\mathbf{X}}_1^T)\}^{-1}\frac{1}{n}\sum_{i=1}^n S_{i,1}\tilde{\mathbf{X}}_i + O_p(\rho^2\log n).$$

Because of $\bar{\beta} - \beta_0 = O_p(n^{-1/2})$, we have $\bar{\beta} = \hat{\beta}$ (with probability tending to 1). This shows Lemma A.3.

With the help of Lemmas A.1 and A.2, we get for the estimate $\hat{m}$ the following expansion.

Corollary A.4

a. For the estimate $\hat{m}$, the following stochastic expansion holds:

$$\sup_{\mathbf{t}\in S_T^-}\Big|\hat{m}(\mathbf{t}) - \{\bar{m}(\mathbf{t}) + \{E(S_{1,2}|\mathbf{T}_1 = \mathbf{t})\}^{-1} \times E(S_{1,2}\mathbf{X}_1^T|\mathbf{T}_1 = \mathbf{t})\,\{E(S_{1,2}\tilde{\mathbf{X}}_1\tilde{\mathbf{X}}_1)\}^{-1}\frac{1}{n}\sum_{i=1}^n S_{i,1}\tilde{\mathbf{X}}_i\}\Big| = O_p(\rho^2\sqrt{\log n}),$$

with $\bar{m}(\mathbf{t}) = m_0(\mathbf{t}) + E(S_{1,2}|\mathbf{T}_1 = \mathbf{t})^{-1}(1/n)\sum_{i=1}^n w_i(\mathbf{t})\, L_i'\{\mathbf{X}_i^T\beta_0 + m_0(\mathbf{t})\}$.

b. The supremum in (a) taken over $\mathbf{t} \in S_T^{\mathbf{h}}$ is of stochastic order $O_p(\tau^2)$.

In particular, we get $\sup_{\mathbf{t}\in S_T^-}|\hat{m}(\mathbf{t}) - \bar{m}(\mathbf{t})| = O_p(n^{-1/2})$ and $\sup_{\mathbf{t}\in S_T^{\mathbf{h}}}|\hat{m}(\mathbf{t}) - \bar{m}(\mathbf{t})| = O_p(\tau^2)$, and also $\sup_{\mathbf{t}\in S_T^-}|\hat{m}(\mathbf{t}) - m(\mathbf{t})| = O_p(\rho\sqrt{\log n})$ and $\sup_{\mathbf{t}\in S_T^{\mathbf{h}}}|\hat{m}(\mathbf{t}) - m(\mathbf{t})| = O_p(\tau)$. In Section 2 we introduced in (12) the modification $\tilde{m}(\mathbf{t})$ of the parametric estimate $\mathbf{t}^T\tilde{\boldsymbol{\gamma}}$. The purpose of this modification was to compensate for the bias of $\hat{m}(\mathbf{t})$ when comparing $\tilde{m}(\mathbf{t})$ and $\hat{m}(\mathbf{t})$. The next lemma shows that this modification works.

Lemma A.5. Suppose that the hypothesis (1) holds; that is, $m_0(\mathbf{t}) = \mathbf{t}^T\boldsymbol{\gamma}_0$. Then

$$\sup_{\mathbf{t}\in S_T^-}|\tilde{m}(\mathbf{t}) - \mathbf{t}^T(\tilde{\boldsymbol{\gamma}} - \boldsymbol{\gamma}_0) - E\{\bar{m}(\mathbf{t})|\mathbf{X}_1, \mathbf{T}_1, \ldots, \mathbf{X}_n, \mathbf{T}_n\}| = O_p(\rho^2\sqrt{\log n}).$$

Proof. The proof uses similar expansions as before. In particular, it uses the fact that with probability tending to 1,

$$\sum_{i=1}^n K_{\mathbf{h}}(\mathbf{t} - \mathbf{T}_i)\frac{G\{\tilde{\mu}_i(\mathbf{t})\} - G(\mathbf{X}_i^T\bar{\boldsymbol{\beta}} + \mathbf{T}_i^T\tilde{\boldsymbol{\gamma}})}{G\{\hat{\mu}_i(\mathbf{t})\}[1 - G\{\tilde{\mu}_i(\mathbf{t})\}]}\, G'\{\hat{\mu}_i(\mathbf{t})\} = 0,$$

where $\tilde{\mu}_i(\mathbf{t}) = \mathbf{X}_i^T\bar{\boldsymbol{\beta}} + \tilde{m}(\mathbf{t})$.

Proof of Theorem 1

Application of the foregoing expansions for the parametric and semiparametric estimates gives

$$\sup_{\mathbf{t}\in S_T^-}|[\hat{m}(\mathbf{t}) - \tilde{m}(\mathbf{t})] - [\bar{m}(\mathbf{t}) - E\{\bar{m}(\mathbf{t})|\mathbf{X}_1, \mathbf{T}_1, \ldots, \mathbf{X}_n, \mathbf{T}_n\}]| = O_p(\rho^2\sqrt{\log n}),$$

$$\sup_{\mathbf{t}\in S_T^-}|\bar{m}(\mathbf{t}) - E\{\bar{m}(\mathbf{t})|\mathbf{X}_1, \mathbf{T}_1, \ldots, \mathbf{X}_n, \mathbf{T}_n\}| = O_p((nh_{\text{prod}})^{-1/2}\sqrt{\log n}),$$

These equalities, together with the expansions for the suprema over S_T^-, imply that for $j = 1, 2, 3$

$$R_j = R + O_p(n\rho^2(nh_{\text{prod}})^{-1/2}\log n)$$

and

$$R = \sum_{i=1}^n \frac{G'(\eta_i)^2}{G(\eta_i)\{1 - G(\eta_i)\}} \times \{\bar{m}(\mathbf{T}_i) - E[\bar{m}(\mathbf{T}_i)|\mathbf{X}_1, \mathbf{T}_1, \ldots, \mathbf{X}_n, \mathbf{T}_n]\}^2,$$

where $\eta_i = \mathbf{X}_i^T\beta_0 + \mathbf{T}_i^T\boldsymbol{\gamma}_0$ for $i = 1, \ldots, n$. Under our assumptions, we have $n\rho^2(nh_{\text{prod}})^{-1/2}\log n = o(h_{\text{prod}}^{-1/2}) = o(v_n)$. This shows statement a. For statement b, note that, conditionally given $\mathbf{X}_1, \mathbf{T}_1, \ldots, \mathbf{X}_n, \mathbf{T}_n$, the statistic R is a U statistic. Proceeding following Härdle and Mammen (1993), one can verify de Jong's (1987) conditions for asymptotic normality of U statistics.

Proof of Theorem 2

As in the proof of Theorem 1, one shows for $j = 1, 2, 3$ that

$$d_K\{R_j^*,\ \mathrm{N}(e_n, v_n^2)\} \to 0 \quad \text{(in probability)}. \tag{A.18}$$

(Recall that e_n and v_n have been introduced in Theorem 1.) For this purpose, one notes first that for all three versions of the bootstrap, $|Y_i^*|$ has a bounded conditional Laplace transform (in a neighborhood of 0). This has been shown in the proof of theorem 5.1 of Mammen and van de Geer (1997). For the proof of (A.18), one proceeds now as in the proof of Theorem 1.

[Received May 1996. Revised June 1998.]

REFERENCES

Beran, R. (1986), "Comment on 'Jackknife, Bootstrap, and Other Resampling Methods in Regression Analysis' by C. F. J. Wu," *The Annals of Statistics*, 14, 1295–1298.

Buja, A., Hastie, T., and Tibshirani, R. (1989), "Linear Smoothers and Additive Models," (with discussion), *The Annals of Statistics*, 17, 453–555.

Burda, M. (1993), "The Determinants of East–West German Migration," *European Economic Review*, 37, 452–461.

Carroll, R., Fan, J., Gijbels, I., and Wand, M. (1997), "Generalized Partially Single-Index Models," *Journal of the American Statistical Association*, 90, 477–489.

Chen, R., Härdle, W., Linton, O., and Severance-Lossin, E. (1996), "Estimation and Variable Selection in Additive Nonparametric Regression Models," in *Proceedings of the COMPSTAT Satellite Meeting Semmering 1994*, eds. W. Härdle and M. Schimek, Heidelberg: Physica Verlag.

de Jong, P. (1987), "A Central Limit Theorem for Generalized Quadratic Forms," *Probability Theory and Related Fields*, 75, 261–277.

Fahrmeir, L., and Hamerle, A. (1984), *Multivariate Statistische Verfahren*, Berlin: De Gruyter.

Fahrmeir, L., and Tutz, G. (1994), *Multivariate Statistical Modelling Based on Generalized Linear Models*, Berlin: Springer-Verlag.

Fan, J., Härdle, W., and Mammen, E. (1998), "Direct Estimation of Low-Dimensional Components in Additive Models," unpublished manuscript submitted to *The Annals of Statistics*.

Green, P. J. (1987), "Penalized Likelihood for General Semi-Parametric Regression Models," *International Statistical Review*, 55, 245–259.

GSOEP (1991), "Das Sozio-Ökonomische Panel (SOEP) im Jahre 1990/91, Projektgruppe 'Das Sozio-ökonomische Panel,' Vierteljahreshefte zur Wirtschaftsforschung, pp. 146–155.

Härdle, W., Huet, S., Mammen, E., and Sperlich, S. (1998), "Semiparametric Additive Indices for Binary Response," technical report, Humboldt-Universität zu Berlin, Sonderforschungsbereich 373.

Härdle, W., and Mammen, E. (1993), "Testing Parametric Versus Nonparametric Regression," *The Annals of Statistics*, 21, 1926–1947.

Hastie, T. J., and Tibshirani, R. J. (1990), *Generalized Additive Models*, (Monographs on Statistics and Applied Probability, Vol. 43), London: Chapman and Hall.

Ingster, Y. I. (1993), "Asymptotically Minimax Hypothesis Testing for Nonparametric Alternatives, I–III," *Mathematical Methods of Statistics*, 2, 85–114; 171–189; 249–268.

Lepski, O. V., and Spokoiny, V. G. (1995), "Minimax Nonparametric Hypothesis Testing: The Case of an Inhomogeneous Alternative," unpublished manuscript submitted to Bernoulli.

Linton, O., and Nielsen, J. P. (1995), "A Kernel Method of Estimating Structured Nonparametric Regression Based on Marginal Integration," *Biometrika*, 82, 93–101.

Maddala, G. S. (1983), "Limited-Dependent and Qualitative Variables in Econometrics," in *Econometric Society Monographs* (No. 4), Cambridge, U.K.: Cambridge University Press.

Mammen, E. (1992), "When Does Bootstrap Work: Asymptotic Results and Simulations," *Lecture Notes in Statistics*, Vol. 77, Berlin: Springer-Verlag.

Mammen, E., and van de Geer, S. (1997), "Penalized Quasi-Likelihood Estimation in Partial Linear Models," *The Annals of Statistics*, 25, 1014–1035.

McCullagh, P., and Nelder, J. A. (1989), "Generalized Linear Models," in *Monographs on Statistics and Applied Probability*, (Vol. 37, 2nd ed.), London: Chapman and Hall.

Müller, M. (1997), "Computer-Assisted Generalized Partial Linear Models," unpublished manuscript submitted to *Proceedings of Interface '97*, Houston, TX.

Robinson, P. M. (1988), "Root n-Consistent Semiparametric Regression," *Econometrica*, 56, 931–954.

Severini, T. A., and Staniswalis, J. G. (1994), "Quasi-Likelihood Estimation in Semiparametric Models," *Journal of the American Statistical Association*, 89, 501–511.

Severini, T. A., and Wong, W. H. (1992), "Generalized Profile Likelihood and Conditionally Parametric Models," *The Annals of Statistics*, 20, 1768–1802.

Speckman, P. E. (1983), "Regression Analysis for Partially Linear Models," *Journal of the Royal Statistical Society*, Ser. B, 50, 413–436.

Tjøstheim, D., and Auestad, B. H. (1994), "Nonparametric Identification of Nonlinear Time Series: Projections," *Journal of the American Statistical Association*, 89, 1398–1409.

Wu, C. F. J. (1986), "Jackknife, Bootstrap, and Other Resampling Methods in Regression Analysis," *The Annals of Statistics*, 14, 1261–1350.

Scand. J. Stat. 26, 221-238
 Vol 26: 221–238, 1999

Testing a Regression Model When We Have Smooth Alternatives in Mind

WOLFGANG HÄRDLE

Humboldt-Universität zu Berlin

ALOIS KNEIP

Université catholique de Louvain

ABSTRACT. Goodness-of-fit tests based on residual sums of squares are standard procedures used when fitting regression models. Often we have a smooth alternative in mind, a qualitative feature that the χ^2-test does not take into account. We show that the power of detecting a smooth alternative increases when we smooth the current model as well. The proposed test is shown to be able to detect any continuous local alternative tending to zero slower than $n^{-1/2}$. Theoretical results also address minimax non-parametric hypothesis testing in Sobolev spaces. A simulation study is presented, and the procedure is applied to expenditure curve estimation.

Key words: goodness-of-fit tests, regression models, smooth alternatives

1. Introduction

Goodness-of-fit tests are designed to check whether a fitted model has captured all the systematic aspects of the data. There is no ideal test that has good power against all possible departures from a hypothesized model. Parametric goodness-of-fit tests may have poor power (or are even inconsistent) if one does not specify the correct type of model departure. Non-parametric tests like the Kolmogorov–Smirnov or the Cramér–von Mises are consistent against virtually all alternatives but have poor power in small samples (Durbin & Knott, 1972) unless the departure of the model is very smooth.

In this paper we propose an approach in between that is based on the idea of "smooth alternatives". This idea comes from the fact that a statistician who has set up a regression model and then wants to test its goodness-of-fit usually thinks of an alternative with a high degree of smoothness anyway. The traditional testing based on the fluctuation of the residual sum of squares (RSS) suffers in this situation form the fact that the random error induces a high variability of the residual sum of squares. Small alternatives can thus not be distinguished from noise. The approach proposed here is to smooth the data as well as the model in order to reduce the degrees of freedom. A test based on the resulting smoothed residual sum of squares then possesses a better performance for non-parametric smooth alternatives.

However, all popular smoothing procedures require the choice of a smoothing parameter h which crucially influences the effective power of such a test. To overcome this difficulty we present an enhanced testing procedure which is based on comparing the smoothed residual sum of squares over a large range of possible values h. The procedure is shown to be able to detect any continuous local alternative tending to zero slower than $n^{-1/2}$. Further theoretical results indicate that a high power is achieved uniformly over all alternatives of comparable degree of smoothness. This allows to draw conclusions on minimax non-parametric hypothesis testing in Sobolev spaces.

There have been a number of proposals using non-parametric smoothing techniques for goodness-of-fit tests. A detailed discussion of some important approaches is given in Hart (1997). For example, Cox *et al.* (1988) and Eubank & Spiegelmann (1990) smooth the residuals

Scand. J. Stat. 26, 221-238

from a hypothesized parametric model. Under the hypothesis that the model is correctly stated the estimated residual curve should fluctuate smoothly around zero. Under the assumption that the model is incorrectly stated this curve should tend to something different from zero. This deviation from zero could be measured by different means. Härdle & Mammen (1993), Raz (1990) use the integrated squared deviance. Hall & Hart (1990) employ the bootstrap to approach the limiting distribution of the smoothed residuals. Azzalini *et al.* (1988) also use the bootstrap (under the hypothesized parametric model) to construct a pseudo-likelihood ratio test for testing against a non-parametric alternative.

The setting we analyse is a regression model

$$Y_i = m(x_i) + \epsilon_i \qquad i = 1, \ldots, n, \tag{1.1}$$

where ε_i denotes the unknown error term, the design points $x_i \in J = [a, b] \subset \mathbb{R}$ are considered as given, and $m(\cdot)$ is the unknown smooth regression curve. We suppose that the errors are i.i.d. mean zero random variables with variance $\sigma^2 > 0$.

Suppose now that with known smooth basis functions $\{g_r\}_{r=1}^{L}$ our hypothesized model is

$$m = \sum_{r=1}^{L} \theta_r g_r \tag{1.2}$$

with unknown parameters $\underline{\theta} = (\theta_1, \ldots, \theta_L)^{\mathrm{T}}$. The problem is then to evaluate the goodness-of-fit of this model. If there is any deviation from (1.1), then there has to exist a parameter $\vartheta \neq 0$ and a function v: $J \to \mathbb{R}$ satisfying $\int_a^b v(x)^2\,dx = 1$, $\int_a^b v(x) g_r(x)\,dx = 0$, $r = 1, \ldots, L$, such that

$$m(x) = \sum_{r=1}^{L} \theta_r g_r(x) + \vartheta v(x).$$

If m and $g_1, \ldots, g_L$ are smooth, so is v. Technically we will only require that v is continuous, and it will be shown that asymptotically our test procedure is able to detect any continuous alternative if ϑ is of larger order than $n^{-1/2}$. However, it will become clear from the discussion of sections 3 and 5 that the effective power for moderate sample sizes depends crucially on the degree of smoothness of v.

The goodness-of-fit problem against a smooth alternative can now be stated as a test of

$$H_0\colon\ m = \sum_{r=1}^{L} \theta_r g_r \tag{1.3}$$

vs

$$H_1\colon\ m = \sum_{r=1}^{L} \theta_r g_r + \vartheta v \qquad \text{for some } \vartheta \neq 0 \text{ and some } v \in \mathscr{C}.$$

Here, $\mathscr{C}$ denotes the space of all continuous functions w: $[a, b] \to R$ with $\int_a^b w(x)^2\,dx = 1$, $\int_a^b w(x) g_r(x)\,dx = 0$, $r = 1, \ldots, L$.

In section 2 we recall the properties of power of a residual sum of squares based test. The application of smoothing procedures for detecting smooth alternatives is considered in section 3. We concentrate on the use of projection based smoothers (such as regression splines) which facilitate explicit power calculations. For simplicity, the arguments in sections 2 and 3 rely on the assumption that $\epsilon_i \sim N(0, \sigma^2)$ with known variance σ^2. Section 4 deals with some important generalizations which concern the use of local linear regression for smoothing as well as more general assumptions on the structure of the error term. A simulation study and an application to expenditure curve estimation are described in section 5.

(1999) Härdle, W. and Kneip, A.
Testing a Regression Model when we have smooth alternatives in mind.

2. When does the RSS indicate a model departure?

Assume that $\epsilon_i \sim N(0, \sigma^2)$ with known variance σ^2. The residual sum of squares (RSS) can then be used for checking the correctness of a regression model.

In the following, for any function w we use $\underline{w}$ to denote the vector $(w(x_1), \ldots, w(x_n))^T$. Let $\underline{Y} = (Y_1, \ldots, Y_n)^T$ denote the vector of observations and $\underline{g}_T = (g_r(x_1), \ldots, g_r(x_n))^T$ the rth model term. Define also $\underline{\epsilon} = (\epsilon_1, \ldots, \epsilon_n)^T$, $G = (g_1, \ldots, \underline{g}_L)$ and $\underline{\theta} = (\theta_1, \ldots, \theta_L)^T$. Throughout this paper we will assume that G is of full rank.

Denote by $\hat{\underline{\theta}} = (G^T G)^{-1} G^T \underline{Y}$ the least squares estimator of $\underline{\theta}$ in the model (1.2). The test based on

$$\text{RSS} := \frac{1}{\sigma^2} \|\underline{Y} - G\hat{\underline{\theta}}\|_2^2 = \frac{1}{\sigma^2} \sum_{i=1}^{n} \left\{ Y_i - \sum_{r=1}^{L} \hat{\theta}_r g_r(x_i) \right\}^2 = \frac{1}{\sigma^2} \|[I - G(G^T G)^{-1} G^T]\underline{Y}\|_2^2$$

is to compare this value with the critical value $C_{\alpha, n-L}$ of a χ^2_{n-L} distribution. This is of course motivated by the fact that $I - G(G^T G)^{-1} G$ is a projection matrix of rank $n - L$ and that under the assumption $\underline{\epsilon} \sim N(0, \sigma^2 I)$ and H_0

$$\text{RSS} = \frac{1}{\sigma^2} \|[I - G(G^T G)^{-1} G^T]\underline{Y}\|_2^2 = \frac{1}{\sigma^2} \|[I - G(G^T G)^{-1} G^T]\underline{\epsilon}\|_2^2 \sim \chi^2_{n-L}.$$

For what sizes of ϑ will we be able to detect the $(L+1)$st term under H_1? First note that under H_1

$$\text{RSS} = \frac{1}{\sigma^2} \|[I - G(G^T G)^{-1} G^T]\underline{Y}\|_2^2 = \frac{1}{\sigma^2} \|[I - G(G^T G)^{-1} G^T](\vartheta \underline{v} + \underline{\epsilon})\|_2^2$$

follows a non-central χ^2 distribution with $n - L$ degrees of freedom and non-centrality parameter $(1/\sigma^2)\vartheta^2 \|[I - G(G^T G)^{-1} G^T]\underline{v}\|_2^2$.

We must be somewhat careful when analysing the power of this test. By definition any alternative v satisfies $\int_a^b v(x)^2\, dx = 1$, $\int_a^b v(x) g_r(x)\, dx = 0$, $r = 1, \ldots, L$,. However, based on a finite number of design points we can at most approximate these integrals by finite sums of the type

$$\frac{1}{n} \sum_{i=1}^{n} v(x_i)^2 \quad \text{and} \quad \frac{1}{n} \sum_{i=1}^{n} v(x_i) g(x_i)$$

which are not necessarily equal to 1 and 0.

However, if for example the design points $x_1, \ldots, x_n$ are regularly spaced then these finite sums will converge to their integral values as $n \to \infty$. For any $v \in \mathscr{C}$ and every $\delta > 0$ there then exists an $n_v \in \mathbb{N}$ such that $v \in \mathscr{C}_{n,\delta}$ for all $n \geq n_v$, where $\mathscr{C}_{n,\delta}$ denotes the space of all functions $w \in \mathscr{C}$ with

$$1 + \delta \geq \frac{1}{n} \|[I - G(G^T G)^{-1} G^T]\underline{w}\|_2^2 = \frac{1}{n} \sum_{i=1}^{n} w(x_i)^2 - \frac{1}{n} \underline{w}^T G(G^T G)^{-1} G)^T \underline{w} \geq 1 - \delta.$$

If for given n we have $v \in \mathscr{C}_{n,\delta}$, the non-centrality parameter characterizing the distribution of RSS under H_1 can thus be bounded by

$$\frac{n\vartheta^2}{\sigma^2}(1 + \delta) \geq \frac{n\vartheta^2}{\sigma^2} \frac{1}{n} \|[I - G(G^T G)^{-1} G^T]\underline{v}\|_2^2 \geq \frac{n\vartheta^2}{\sigma^2}(1 - \delta).$$

The non-central χ^2 distribution is asymptotically normal, and for $v \in \mathscr{C}_{n,\delta}$ we obtain

$$n - L + \frac{n\vartheta^2}{\sigma^2}(1 + \delta) \geq E(\text{RSS}) \geq n - L + \frac{n\vartheta^2}{\sigma^2}(1 - \delta) \tag{2.1}$$

$$2\left(n - L + 2\frac{n\vartheta^2}{\sigma^2}(1+\delta)\right) \geq \text{var(RSS)} \geq 2\left(n - L + 2\frac{n\vartheta^2}{\sigma^2}(1-\delta)\right). \tag{2.2}$$

Since the critical value $C_{\alpha,n-L}$ of the χ^2_{n-L} distribution has the magnitude of $n - L + \text{const} \cdot (n-L)^{1/2}$, we see from a comparison with (2.1) and (2.2) that for fixed L the RSS based test will detect magnitudes of ϑ satisfying $(n-L)^{1/2} = O(n\vartheta^2)$. So we can expect a rejection of H_0 only if $\vartheta \gg n^{-1/4}$. In other words, terms under H_1 with order $\vartheta = o(n^{-1/4})$ will usually not be detected via an application of this χ^2-test. In this case the corresponding distribution of the RSS almost coincides with the Null distribution for large n. Summarizing we have seen that under H_1 we obtain for any $\delta > 0$ that as $n \to \infty$

$$\inf_{v \in \mathscr{C}_{n\delta}} \inf_{|\vartheta| \geq \beta_n \cdot n^{-1/4}} P(\text{RSS} > C_{\alpha,n-L}) \to 1 \tag{2.3}$$

$$\sup_{v \in \mathscr{C}_{n\delta}} \sup_{|\vartheta| \leq \beta_n^* n^{-1/4}} P(\text{RSS} > C_{\alpha,n-L}) \to \alpha$$

where β_n and β_n^* are sequences of constants with $\beta_n \to \infty$ and $\beta_n^* \to 0$ as $n \to \infty$. Note that our derivation of (2.3) did not involve assumptions about the structure of v, except $v \in \mathscr{C}_{n,\delta}$. The RSS-based test thus treats all possible alternatives in an identical way.

3. Thinking of a smooth alternative

The power of the RSS based test suffers from the fact that the random error induces a high variability of the residual sum of squares. Small alternatives can thus not be distinguished from noise.

When thinking of smooth alternatives, a natural way to reduce the influence of the random error consists in the application of smoothing procedures. If we concentrate on estimates at the design points, most popular smoothing procedures like kernel estimators, smoothing splines, etc., estimate $\underline{m}$ by multiplying an $n \times n$ smoother matrix W_h with the vector $\underline{Y}$ of observations. The structure of W_h depends on the method and on a smoothing parameter h. A discussion of the matrices W_h associated with different smoothing procedures is given in Hastie & Tibshirani (1990).

In this section we will concentrate on smoothing procedures with the property that W_h is a projection matrix (this condition will be relaxed in section 4). Among many possible methods, one might think of the following examples.

Example 1. Least squares approximation of polynomials of degree h. Then

$$W_h = B_h(B_h^{\mathrm{T}}B_h)^{-1}B_h^{\mathrm{T}}$$

where the elements of B_h are given by $(B_h)_{ij} = x_i^j$.

Example 2. Cubic regression spline smoothing (see, for example, de Boor, 1978). This is a projection method for fitting cubic splines. For a given sequence $a = t_1 < t_2 < \cdots < t_{k-1} < t_k = b$ of k knots we fit a spline function which is a cubic polynomial between two successive knots and is twice differentiable at each knot point. The space of all these functions possesses a basis of $h = k + 2$ so-called B-splines $b_1, \ldots, b_h$ (see de Boor, 1978). If knots are chosen in such a way that there are approximately the same number of data points in between two knots (equidistant knots in the case of regular spaced design), then h may be considered as smoothing parameter, and if $h < n$

$$W_h = B_h(B_h^{\mathrm{T}}B_h)^{-1}B_h^{\mathrm{T}}$$

where the elements of B_h are given by $b_i(x_j)$.

In the following we will thus assume that the procedure used to smooth the data possesses a corresponding smoother matrix W_h which is a symmetric projection matrix. The matrices W_h are determined by the choice of an integer-valued smoothing parameter h (degree of the polynomial in example 1 or number of B-splines in example 2), and $\text{rank}(W_h) = h$ if $h \leqslant n$. Note that here the amount of smoothing decreases as h increases.

3.1. The RSS_h-based test

Assume again that $\epsilon_i \sim N(0, \sigma^2)$ with known variance σ^2. A first approach would be to smooth the data and then apply the RSS based test to the smoothed data. However, with this approach we introduce a smoothing bias which complicates the analysis. More specifically, the RSS will contain the additional term $\|\underline{m} - W_h\underline{m}\|_2^2$. We then would have to control the magnitude of this term in order to derive properties of this test. This is complicated and tedious to calculate.

The method we shall use and that has been proposed by several authors, among them Hall & Hart (1990), is to smooth the model as well. Hence, introduce the pre-smoothed model

$$\tilde{\underline{m}} = W_h\underline{m} = \sum_{r=1}^{L} \theta_r W_h \underline{g}_r = \sum_{r=1}^{L} \theta_r \tilde{\underline{g}}_r. \tag{3.1}$$

If the null hypothesis H_0 defined by (1.2) holds, then also $\tilde{\underline{m}} = \sum_{r=1}^{L} \theta_r \tilde{\underline{g}}_r =: \tilde{G}\underline{\theta}$ holds. In contrast, H_1 implies that $\tilde{\underline{m}} = \tilde{G}\underline{\theta} + \vartheta W_h \underline{v}$.

We will require that to some extent the amount of smoothing done in (3.1) fits to the hypothesized model. The smoothed components $W_h\underline{g}_r$ should not be too far from the original $\underline{g}_r$. Note that necessarily for any $c \in \mathbb{R}^L$ one obtains $c^T G^T G c \geqslant c^T \tilde{G}^T \tilde{G} c = c^T G^T W_h G c$, but if $G \approx \tilde{G} = W_h G$ the difference will be small. This is in fact not very difficult to achieve since the g_r are known. We do not need a very accurate approximation, but in the following we will always assume some minimal condition: $h \in \{h_0, h_0 + 1, h_0 + 2, \ldots\} =: H$ for some $h_0 \in \mathbb{N}$ with $h_0 > L$ such that $\inf_{c \in R^L}\{2c^T G^T W_h G c - c^T G^T G c\} \geqslant 0$ holds for all $n \geqslant h \geqslant h_0$. In particular, $\tilde{G} = W_h G$ is then of full rank L.

For fixed $h \in H$, $h \leqslant n$, the proposed goodness-of-fit test now proceeds as follows.

1. Determine $\underline{\hat{m}}_h = (\hat{m}_h(x_1), \ldots, \hat{m}_h(x_n))' = W_h\underline{Y}$. Then calculate the least squares estimate $\underline{\tilde{\theta}}$ of $\underline{\theta}$ to be obtained by minimizing the sum of squared residuals $\|\underline{\hat{m}}_h - W_h G\vartheta\|_2^2$ for the smoothed model.
2. Determine

$$\text{RSS}_h = \frac{1}{\sigma^2}\|\underline{\hat{m}}_h - \tilde{G}\underline{\tilde{\theta}}\|_2^2 = \frac{1}{\sigma^2}\|(I - P_h)W_h\underline{Y}\|_2^2.$$

   Here, $P_h = \tilde{G}(\tilde{G}^T\tilde{G})^{-1}\tilde{G}^T$ denotes the projection matrix projecting into the linear space spanned by $W_h\underline{g}_1, \ldots, W_h\underline{g}_L$.
3. For a given level $\alpha > 0$ reject H_0 if $\text{RSS}_h > C_{\alpha,h-L}$, where $C_{\alpha,h-L}$ is the corresponding critical value obtained from a χ^2 distribution with $\text{tr}((I - P_h)W_h) = h - L$ degrees of freedom.

Step 3) follows from

$$\frac{1}{\sigma^2}\|(I - P_h)W_h\underline{Y}\|_2^2 = \frac{1}{\sigma^2}\|(I - P_h)W_h\underline{\epsilon}\|_2^2.$$

We have $P_h W_h = P_h$ and $(I - P_h)W_h$ is a projection matrix of rank $h - L$. Consequently

$$\frac{1}{\sigma^2}\|(I - P_h)W_h\underline{\epsilon}\|_2^2$$

follows a χ^2 distribution with $h - L$ degrees of freedom.

What can we say about the power of this test? For fixed h an answer is given by theorem 1 below which relies on some reasonable conditions on the asymptotic behaviour of the design and on the choice of the smoothing procedure. For all h, n let $\underline{v}_{h,n} = (v_{h,n}(x_1), \ldots, v_{h,n}(x_n))^T = W_h\underline{v}$.

Assumption 1

For $h \in H$, $n \in \mathbb{N}$ and $\delta > 0$ let $\mathscr{C}_{h,n,\delta}$ denote the space of all possible alternative $v \in \mathscr{C}$ such that

(a) $\frac{1}{n}\|(I - G(G^TG)^{-1}G^T)\underline{v}\|_2^2 \geq 1 - \delta$

(b) $\frac{1}{n}\|\underline{v} - W_h\underline{v}\|_2^2 = \frac{1}{n}\sum_{i=1}^{n}(v(x_i) - v_{h,n}(x_i))^2 \leq \delta.$

Then, for any $v \in \mathscr{C}$ and every $\delta > 0$ there exist an $h_v \in H$ and an $n_v \in \mathbb{N}$ such that $v \in \mathscr{C}_{h,n,\delta}$ holds for all $n \geq n_v$ and $h \in H$ with $n \geq h \geq h_v$.

Condition a refers to the asymptotic behaviour of the design. It resembles the requirements already discussed in section 2.

For regularly spaced design it follows from well-known results of approximation theory that for any continuous alternative $v \in \mathscr{C}$ we can achieve an arbitrarily good approximation of v by a polynomial of a sufficiently high degree or by cubic splines based on a sufficiently large number of equidistant knots. In this case, assumption 1 is thus satisfied for either one of the methods proposed in examples 1 and 2.

We can even say something more. An interesting aspect is the question whether we can achieve good power of a test uniformly over some interesting smoothness classes. We will not treat this question in full generality, but only concentrate on a particularly interesting class. For $p < \infty$ let $S_2(p) \subset \mathscr{C}$ denote the Sobolev space of all twice differentiable (in a distributional sense) alternatives $v \in \mathscr{C}$ satisfying $\int_a^b v''(x)^2dx \leq p$. We can infer from results of approximation theory (see, for example, Schumaker, 1981; Devore & Lorentz, 1991) that for fixed $p < \infty$ and regular spaced design we obtain

$$\sup_{v \in S_2(p)} \frac{1}{n}\sum_{i=1}^{n}(v(x_i) - v_{h,n}(x_i))^2 \to 0 \quad \text{as } h, n \to \infty$$

for either one of the methods proposed in examples 1 and 2. Note that $\int_a^b v(x)^2dx = 1$ as well as $\int_a^b v''(x)^2dx \leq p$ imply that $v(x)$ and $v'(x)$ are bounded uniformly for all $v \in S_2(p)$. In such situations the following assumption is satisfied.

Assumption 2

For any $\delta > 0$ there exist some $h_p \in H$ and some $n_p \in \mathbb{N}$ such that $S_2(p) \subset \mathscr{C}_{h,n,\delta}$ holds for all $n \geq n_p$ and $h \in H$ with $n \geq h \geq h_p$.

Write $\mathrm{RSS}_h(\vartheta v)$ to indicate the values of the RSS_h for a specific alternative under H_1. For fixed h the above test is justified by the following theorem.

Theorem 1

Let $h \in H$ be fixed and let $\frac{1}{5} \geq \delta > 0$. Under the above assumptions

$$\inf_{v \in \mathscr{C}_{h,n,\delta}} \inf_{|\vartheta| \geq \beta_n n^{-1/2}} P(\mathrm{RSS}_h(\vartheta v) > C_{\alpha,h-L}) \to 1,$$

then holds for any $\delta > 0$ as $n \to \infty$. Here β_n is an arbitrary sequence of constants with $\beta_n \to \infty$ as $n \to \infty$.

Proof. Let $v \in \mathscr{C}_{h,n,\delta}$ and for given $n \geq h$ set $\bar{v} = (I - G(G^{\mathrm{T}}G)^{-1}G^{\mathrm{T}})v$. Obviously $G\bar{v} = 0$. We then obtain

$$\begin{aligned}
\|(I - P_h)W_h v\|_2^2 = \|(I - P_h)W_h \bar{v}\|_2^2 &= \bar{v}^{\mathrm{T}} W_h \bar{v} - \bar{v}^{\mathrm{T}} W_h G(G^{\mathrm{T}} W_h G)^{-1} G^{\mathrm{T}} W_h \bar{v} \\
&= \bar{v}^{\mathrm{T}} W_h \bar{v} - \bar{v}(I - W_h)G(G^{\mathrm{T}} W_h G)^{-1} G^{\mathrm{T}}(I - W_h)\bar{v} \\
&\geq \bar{v}^{\mathrm{T}} W_h \bar{v} - 2\bar{v}(I - W_h)G(G^{\mathrm{T}} G)^{-1} G^{\mathrm{T}}(I - W_h)\bar{v} \\
&\geq \bar{v}^{\mathrm{T}} W_h \bar{v} - 2\|(I - W_h)\bar{v}\|_2^2 = \bar{v}^{\mathrm{T}}\bar{v} - 3\|(I - W_h)\bar{v}\|_2^2.
\end{aligned}$$

Recall that by assumption the matrix $2G^{\mathrm{T}}W_h G - G^{\mathrm{T}}G$ is positive definite. This implies that also the matrix $2(G^{\mathrm{T}}G)^{-1} - (G^{\mathrm{T}}W_h G)^{-1}$ is positive semi-definite which leads to the first inequality above.

Note that $v^{\mathrm{T}}(I - W_h)v \geq \bar{v}^{\mathrm{T}}(I - W_h)\bar{v}$. Definition of $\mathscr{C}_{h,n,\delta}$ now implies that

$$\delta \geq \frac{1}{n}\|v - W_h v\|_2^2 \geq \frac{1}{n}\|\bar{v} - W_h \bar{v}\|_2^2.$$

Since furthermore $\frac{1}{n}\bar{v}^{\mathrm{T}}\bar{v} \geq 1 - \delta$, it follows that

$$\|(I - P_h)W_h v\|_2^2 \geq n(1 - 4\delta). \tag{3.2}$$

This relation does not depend on the specific choice of $v \in \mathscr{C}_{h,n,\delta}$ and thus characterizes the whole class. Consequently,

$$\inf_{v \in \mathscr{C}_{h,n,\delta}} \vartheta^2 \|(I - P_h)W_h v\|_2^2 \geq \vartheta^2 n(1 - 4\delta) \geq \vartheta^2 n/5. \tag{3.3}$$

The theorem now is an immediate consequence of the fact that

$$\mathrm{RSS}_h(\vartheta v) = \frac{1}{\sigma^2}\|(I - P_h)W_h Y\|_2^2$$

follows a non-central χ^2 distribution with $h - L$ degress of freedom and non-centrality parameter

$$\frac{1}{\sigma^2}\vartheta^2\|(I - P_h)W_h v\|_2^2.$$

The theorem shows that for smooth alternatives an RSS_h-based test can be much more powerful than the RSS-based test of section 2. If there is some prior knowledge indicating a smooth alternative v, then we might reasonably choose a small h to perform the test. For example, if one can assume that $v \in S_2(p)$ for some known p, then it will be possible to choose a h such that $v \in \mathscr{C}_{h,n,1/10}$, say. By relation (3.3) the corresponding $\mathrm{RSS}_h(\vartheta v)$ adopts a non-central χ^2 distribution with $h - L$ degrees of freedom and non-centrality parameter

$$\frac{1}{\sigma^2}\vartheta^2\|(I - P_h)W_h v\|_2^2 \geq \vartheta^2 n\tfrac{6}{10}.$$

The test will then possess considerable power even for small values ϑ and moderate sample size. The point is that $h - L$ remains fixed as n increases.

3.2. The RSS_{max}-based test

In practice, the degree of smoothness of a possible alternative will rarely be known. An appropriate choice of h then poses a considerable problem. If h is too large, the test will possess a comparably poor power. Note that if $h = n$ we will usually have $RSS_h = RSS$. On the other hand, if h is too small then for wiggly alternatives $\|(I - P_h)W_h\underline{v}\|_2^2$ may be close to 0 which also results in a poor power of the test. One possibility to overcome this difficulty consists in trying to determine a power maximizing bandwidth. In a different context, some results in this direction are given by Hong (1993).

Our approach is based on a different type of reasoning. Recall that H_0 implies the validity of the smoothed model for *any* matrix W_h. Thus, a true model should pass the test for *any* h which motivates the idea of considering the values of RSS_h for a large range of possible $h \in H$. By definition the RSS_h test rejects H_0 if $RSS_h - C_{\alpha,h-L} > 0$. One might thus tend to look for the maximal difference $RSS_h - C_{\alpha,h-L}$ for different values of h and to reject H_0 if this maximal difference is larger than zero. However, by proceeding in this way we will automatically increase the actual level of significance. In order to be able to control the size of the test such a procedure only makes sense if we correct the RSS_h by a factor larger than $C_{\alpha,h-L}$. Since $C_{\alpha,h-L}$ is of the form $(h - L) + \text{const.}\ (h - L)^{1/2}$ we might think of using the factor $2(h - L)$.

The test we propose is based on this idea. We will consider the statistics

$$RSS_{max} = \sup_{h_0 \leq h \leq n} (RSS_h - 2(h - L)). \tag{3.4}$$

Remark 1. As is easily seen form the proof of theorem 2 below, it is possible to choose a correction term different form $2(h - L)$. In fact, all assertions of theorem 2 remain true if $2(h - L)$ is replaced by $h - L + \gamma(h - L)$ for some arbitrary $\gamma > 0$. An optimal choice of γ seems to be difficult. There is, however, a particular motivation for choosing the factor $2(h - L)$. Clearly, $(I - P_h)W_h\underline{Y}$ can be considered as an estimate of $\vartheta\underline{v}$ under H_1. The optimal value $h_{v,\text{opt}} \in H$ minimizing the mean squared error $E(\|\vartheta\underline{v} - (I - P_h)W_h\underline{Y}\|_2^2) = \vartheta^2\underline{v}^T\underline{v} - \vartheta^2\|(I - P_h)W_h\underline{v}\|_2^2 + \sigma^2(h - L)$ is obviously equivalent to the value of h which maximizes the difference between the non-centrality parameter

$$\frac{1}{\sigma^2}\vartheta^2\|(I - P_h)W_h\underline{v}\|_2^2$$

and $h - L$, the variance of the random error. The optimal smoothing parameter $h_{v,\text{opt}}$ can be estimated by Mallows C_L (Mallows, 1973), i.e. by minimizing $C_L = \|\underline{Y} - (I - P_h)W_h\underline{Y}\|_2^2 + 2\sigma^2 \operatorname{tr}((I - P_h)W_h) = \underline{Y}^T\underline{Y} - \|(I - P_h)W_h\underline{Y}\|_2^2 + 2\sigma^2(h - L)$. It is now immediately seen that minimizing C_L is equivalent to maximizing $RSS_h - 2(h - L)$. For theoretical results justifying the use of Mallows C_L for estimating $h_{v,\text{opt}}$ see Li (1987) or Kneip (1994).

The test now proceeds as follows

1. Under H_0 we have

$$RSS_{max} = \sup_{h_0 \leq h \leq n} \left(\frac{1}{\sigma^2}\|(I - P_h)W_h\underline{\epsilon}\|_2^2 - 2(h - L) \right).$$

For a given level α determine the corresponding critical value $C^{\alpha}_{\max,n}$ such that

$$P(RSS_{max} \geq C^{\alpha}_{\max,n} | H_0) = \alpha$$

2. Reject H_0 if the observed value of $\mathrm{RSS}_{\max}$ is larger than $C^{\alpha}_{\max,n}$

The distribution of $\mathrm{RSS}_{\max}$ under H_0 does not seem to posses an easily evaluable analytical structure under H_0. However, the critical values can be determined by Monte Carlo simulations. The test is justified by the following theorem.

Theorem 2

(i) *For any* $\alpha > 0$ *there exists a* $C < \infty$ *such that* $C^{\alpha}_{\max,n} \leqslant C$ *for all* n

(ii) *Write* $\mathrm{RSS}_{\max}(\vartheta v)$ *to indicate the values of* $\mathrm{RSS}_{\max}$ *for a specific alternative under* H_1. *Then, for* any $h \in H$ *and all* δ *with* $1/5 \geqslant \delta \geqslant 0$ *we obtain*

$$\inf_{v \in \mathcal{V}_{h,n,\delta}} \inf_{|\vartheta| \geqslant \beta_n \cdot n^{-1/2}} P(\mathrm{RSS}_{\max}(\vartheta v) > C^{\alpha}_{\max,n}) \to 1,$$

where β_n *is an arbitrary sequence of constants with* $\beta_n \to \infty$ *as* $n \to \infty$.

Proof. We first consider assertion (i). Without restriction let $\sigma^2 = 1$. Since the ϵ_i have finite fourth moment it follows from Whittle's inequality that under H_0 there exists a constant $\gamma_1 < \infty$ such that for $n \geqslant h$

$$E[(\|(I - P_h)W_h\underline{\epsilon}\|_2^2 - (h - L))^4] = E[(\underline{\epsilon}^{\mathrm{T}} W_h(I - P_h)W_h\underline{\epsilon} - E\{\underline{\epsilon}^{\mathrm{T}} W_h(I - P_h)W_h\underline{\epsilon}\})^4]$$

$$\leqslant \gamma_1[\mathrm{tr}((W_h(I - P_h)W_h)^2)]^2 = \gamma_1(h - L)^2.$$

Furthermore, there exists a constant $\gamma_2 \in \mathbb{R}$ such that $P(\underline{\epsilon}^{\mathrm{T}} W_{h_0}(I - P_{h_0})W_{h_0}\underline{\epsilon} - 2(h_0 - L)) \geqslant \gamma_2) = \alpha/2$. Let $\tilde{h}$ denote the smallest $h \in H$ such that $h - L + \gamma_2 \geqslant 0$. For all $n \geqslant h > \tilde{h}$ we then have

$$P(\|(I - P_h)W_h\underline{\epsilon}\|_2^2 - 2(h - L) \geqslant \gamma_2) \leqslant P(\|(I - P_h)W_h\underline{\epsilon}\|_2^2 - (h - L) \geqslant h - \tilde{h})$$

$$\leqslant \frac{\gamma_1(h - L)^2}{(h - \tilde{h})^4}.$$

There exists a constant $\gamma_3 < \infty$ such that

$$\sum_{i=\tilde{h}+1}^{\infty} \frac{\gamma_1(i - L)^2}{(i - \tilde{h})^4} = \gamma_3.$$

There thus exists a $\overline{h} \in H$ such that

$$\sum_{i=\overline{h}+1}^{\infty} \frac{\gamma_1(i - L)^2}{((i - \tilde{h})^4} \leqslant \alpha/2.$$

Consequently,

$$P\left(\sup_{n \geqslant h > \overline{h}} (\|(I - P_h)W_h\underline{\epsilon}\|_2^2 - 2(h - L)) \geqslant \gamma_2\right)$$

$$\leqslant \sum_{n \geqslant h > \overline{h}} P(\|(I - P_h)W_h\underline{\epsilon}\|_2^2 - 2(h - L) \geqslant \gamma_2) \leqslant \alpha/2.$$

Independent of the value of n, there exists a finite number of elements $h \in H$ such that $h < \overline{h}$, and it is immediately clear that there exists a constant $\gamma < \infty$ with $\gamma_4 \geqslant \gamma_2$ such that for all n sufficiently large $P(\sup_{h_0 \leqslant h \leqslant \overline{h}}(\|(I - P_h)W_h\underline{\epsilon}\|_2^2 - 2(h - L)) \geqslant \gamma_4) \leqslant \alpha/2$. We can thus infer that for all n sufficiently large

$$P(\mathrm{RSS}_{\max} \geq \gamma_4) \leq P\left(\sup_{h_0 \leq h \leq \bar{h}} (\|(I - P_h)W_h\epsilon\|_2^2 - 2(h - L)) \geq \gamma_4\right)$$
$$+ P\left(\sup_{n \geq h > \bar{h}} (\|(I - P_h)W_h\epsilon\|_2^2 - 2(h - L)) \geq \gamma_2\right)$$
$$\leq \alpha.$$

This proves assertion (i). For any $h \in H$ assertion (ii) is an immediate consequence of relation (3.3) and assertion (i).

By assumptions 1 and 2 the theorem implies the following corollary.

Corollary

(i) *For any fixed* $v \in \mathscr{C}$

$$\inf_{|\vartheta| \geq \beta_n \cdot n^{-1/2}} P(\mathrm{RSS}_{\max}(\vartheta v) > C^{\alpha}_{\max,n}) \to 1,$$

where β_n *is an arbitrary sequence of constants with* $\beta_n \to \infty$ *as* $n \to \infty$.

(ii) *For any* $p < \infty$ *we obtain*

$$\inf_{v \in S_2(p)} \inf_{|\vartheta| \geq \beta_n \cdot n^{-1/2}} P(\mathrm{RSS}_{\max}(\vartheta v) > C^{\alpha}_{\max,n}) \to 1$$

where β_n *is an arbitrary sequence of constants with* $\beta_n \to \infty$ *as* $n \to \infty$.

Remark 2. The theorem shows that the rate at which the magnitude of the local alternatives v is allowed to converge to 0 is $n^{-1/2}$. We can infer that the minimax rate over classes $S_2(p)$ of smooth alternatives is $n^{-1/2}$, and thus corresponds to the rate of convergence of parametric tests.

At a first glance this seems to be in contradiction with recent results in minimax nonparametric hypothesis testing. Ingster (1982) studies the situation where $w := \vartheta v$ belongs to a Sobolev class

$$w \in S_2^*(p) = \left\{ w \in L_2(J) \middle| \int w''(x)^2 \, dx \leq p \right\}$$

for some $p < \infty$. He shows that then the minimax rate of convergence of $\{\int_J w(x)^2 dx\}^{1/2} = \vartheta$ to zero is only $n^{-4/9}$. The point is, however, that our setup is different in that we make an explicit distinction between "magnitude" and "degree of smoothness" of an alternative by requiring that $\int v(x)^2 \, dx = 1$. Obviously $S_2(p) = S_2^*(p) \cap \mathscr{C}$. This is of no importance if $\vartheta = 1$, but if $\vartheta \to 0$ there will be more and more "wiggly" functions v with $v \notin S_2(p)$ but $\vartheta v \in S_2^*(p)$. For example, let $J = [0, 2]$ and $v(x) = \sin(2\pi kx)$. As $k \to \infty$, the function $v(x)$ becomes less and less smooth, and $v \notin S_2(p)$. Nevertheless, for *any* k there exists a corresponding value $\vartheta_k > 0$ such that $\int_J w''(x)^2 \, dx = \vartheta^2(2\pi k)^4 \int_J \sin(2\pi kx)^2 \, dx = \vartheta^2(2\pi k)^4 \leq p$ and, hence, $w = \vartheta v \in S_2^*(p)$ hold for all $\vartheta < \vartheta_k$. Our setup does not allow for this effect, and thus seems to be more "natural" if we have smooth alternatives in mind.

4. Generalizations

4.1. Local linear regression

Our test procedure generalizes to non-projection smoother matrices. This is an important aspect since the most widely used smoothing procedures like smoothing splines, kernel

Scand. J. Stat. 26, 221-238

estimators, etc., are not based on projections. We will exemplify this generalization for the case of local linear regression (see, for example, Fan & Gijbels, 1996). The basic arguments for other methods are similar.

The idea of local linear fitting is to estimate $m(x)$ by a locally weighted linear regression. The elements of the resulting smoother matrix W_h are then given by

$$(W_h)_{i,j} = \frac{w_j(x_i)}{\sum_{l=1}^{n} w_l(x_i)} \tag{4.1}$$

where

$$w_j(x_i) = K\left(\frac{x_j - x_i}{b}\right)\{S_{n,2} - (x_j - x_i)S_{n,1}\} \quad \text{and} \quad S_{n,r} = \sum_{l=1}^{n} K\left(\frac{x_l - x_i}{b}\right)(x_l - x_i)^r.$$

Here, K denotes a kernel function and $b = 1/h$ is a bandwidth. In order of W_h being well-defined for all $h \in (0, \infty)$, set $w_{ii} = 1$ and $w_{i,j} = 0$, $i \neq j$, if $\sum_{l=1}^{n} w_l(x_i) = 0$.

Remark 3. For the projection-based smoothing procedures discussed in section 3 a large value of h corresponds to a small amount of smoothing. In contrast, for local linear regression a large bandwidth induces a large amount of smoothing. To ensure comparability with the results discussed in the previous section we thus set $h = 1/b$.

For a fixed $h = 1/b$ the

$$\mathrm{RSS}_h = \frac{1}{\sigma^2}\|(I - P_h)W_h\underline{Y}\|_2^2$$

based test now takes the following form

The RSS_h*-based test.* For a given level $\alpha > 0$ reject H_0 if $\mathrm{RSS}_h > D_{h,n}^{\alpha}$, where $D_{h,n}^{\alpha}$ is the corresponding critical value obtained from the distribution $\mathscr{L}_{h,n}$ of the random variable

$$\frac{1}{\sigma^2}\|(I - P_h)W_h\underline{\epsilon}\|_2^2.$$

For $\epsilon_i \sim N(0, \sigma^2)$ the distribution $\mathscr{L}_{h,n}$ is well-defined, and critical values can always be evaluated by Monte Carlo simulations. The only difference to the situation discussed above consists in the fact that $\mathscr{L}_{h,n}$ does not possess an easily evaluable analytical structure and, in particular, it is not a χ^2 distribution.

In order to analyse this distribution more closely let $\hat{m}_h$ represent the resulting local linear estimator of m, and let $\tilde{m}_h(x) = \hat{m}_h(x) - \sum_{r=1}^{L}\tilde{\theta}_r\tilde{g}_r(x)$. We then obtain that under H_0

$$q_{\mathrm{mean},n}(h) := E(\mathrm{RSS}_h) = E\left(\frac{1}{\sigma^2}\|(I - P_h)W_h\underline{\epsilon}\|_2^2\right)$$

$$= \mathrm{tr}(W_h^{\mathrm{T}}(I - P_h)W_h) = \frac{1}{\sigma^2}\sum_{i=1}^{n}\mathrm{var}\,(\tilde{m}_h(x_i))$$

$$q_{\mathrm{var},n}(h) := \mathrm{var}(\mathrm{RSS}_h) = \frac{1}{\sigma^4}\mathrm{var}\left(\sum_{i=1}^{n}(\tilde{m}_h(x_i) - E(\tilde{m}_h(x_i))^2\right).$$

Note that for regularly spaced design we obtain that asymptotically

$$\frac{1}{\sigma^2}\sum_{i=1}^{n}\mathrm{var}(\hat{m}_h(x_i)) = \frac{d_K}{b} + o\left(\frac{1}{b}\right) = hd_K + \mathrm{o}(h)$$

as well as

$$\frac{1}{\sigma^4}\operatorname{var}\left(\sum_{i=1}^{n}(\hat{m}_h(x_i) - E(\hat{m}_h(x_i))^2)\right) = hd_K^* + o(h)$$

for some kernel dependent constants d_K, d_K^*. Similar as before, it is then easily verified that $q_{\text{mean},n}(h)$, $q_{\text{var},n}(h)$, and the corresponding critical level $D^{\alpha}_{h,n}$ increase in a way approximately proportional to h as h increases. Moreover, for large h ($\equiv$ small bandwidth) $\mathscr{L}_{h,n}$ is well approximated by a normal distribution.

A local linear version of the $\text{RSS}_{\max}$ based test can be defined by relying on reasonable sets of discretized bandwidths. For example, let b_0 denote a very large bandwidth and let $b_k = q^k b_0$ for some $1 < q < 0$ close to 1. With $h_k = 1/b_k$ we may then consider

$$\text{RSS}_{\max} = \sup_{k \geq 0, h_k \leq n} (\text{RSS}_{h_k} - 2q_{\text{mean},n}(h_k)). \tag{4.2}$$

Note that in section 3 the correction term $2(h - L)$ was due to $E(\text{RSS}_h | H_0) = h - L$. In the present case we have to replace $h - L$ by the true mean $q_{\text{mean},n}(h_k)$ of RSS_{h_k} under H_0. This leads to the following test

1. Under H_0 we have

$$\text{RSS}_{\max} = \sup_{k \geq 0, h_k \leq n} \left(\frac{1}{\sigma^2}\|(I - P_h)W_h\underline{\epsilon}\|_2^2 - 2q_{\text{mean},n}(h_k)\right).$$

   For a given level α determine the corresponding critical value $D^{\alpha}_{\max,n}$ such that $P(\text{RSS}_{\max} \geq D^{\alpha}_{\max,n} | H_0) = \alpha$.
2. Reject H_0 if the observed value of $\text{RSS}_{\max}$ is larger than $D^{\alpha}_{\max,n}$.

The critical level $D^{\alpha}_{\max,n}$ can again be determined by Monte Carlo simulations.

One might also use a more dense set of discretized bandwidths. The only important condition is that there exists some $\eta > 0$ such that for all n sufficiently large and all $k \in \mathbb{N}$ with $h_k < n$ we have $q_{\text{mean},n}(h_k) - q_{\text{mean},n}(h_{k-1}) \geq \eta$. By using arguments similar to those used in the proof of assertion (i) of theorem 2 it may then be proved that there exists a $D < \infty$ with $D^{\alpha}_{\max,n} \leq D$ for all $n \in \mathbb{N}$.

For regular spaced design it is well-known that

$$\frac{1}{n}\sum_{i=1}^{n}(v(x_i) - E(\tilde{v}_h(x_i))^2 = \frac{1}{n}\|\underline{v} - (I - P_h)W_h\underline{v}\|_2^2 \to 0 \quad \text{as } h = 1/b \to \infty$$

holds for any $v \in \mathscr{C}$, and this convergence is even uniform over $v \in S_2(p)$. Consequently, for each $\delta > 0$ one obtains $\|(I - P_h)W_h\vartheta\underline{v}\|_2^2 \geq n\vartheta^2(1 - \delta)$ for all h sufficiently large. Similar as above, one may then prove that the assertions of corollary 1 generalize to the present situation.

4.2. Unknown error variance

Up to now we have assumed that $\epsilon_i \sim N(0, \sigma^2)$ with known error variance σ^2. Under the more realistic assumption that σ^2 is unknown, the tests may be performed by replacing σ^2 by a consistent estimator. Such estimators of σ^2 can be obtained, for example, by the methods of Rice (1984), Gasser *et al.* (1986) or Hall *et al.* (1990). If m is continuous, then under weak technical conditions they all satisfy $|\hat{\sigma}^2 - \sigma^2| = o_P(1)$. As a consequence

Scand. J. Stat. 26, 221-238

$$\frac{1}{\hat{\sigma}^2}\|(I-P_h)W_h\vartheta\underline{v}\|_2^2 = \frac{1}{\sigma^2}\|(I-P_h)W_h\vartheta\underline{v}\|_2^2 + o_P(1)$$

and the above theoretical results remain asymptotically valid.

However, as long as σ^2 is known, we can derive the *exact* critical values for the RSS_h as well as for the $\mathrm{RSS}_{\max}$ based tests by a Monte Carlo approximation of the distribution of $1/\sigma^2\|(I-P_h)W_h\underline{\epsilon}\|_2^2$. This is no longer true if σ^2 is replaced by $\hat{\sigma}^2$. Then the resulting critical values are only asymptotically valid. Fortunately this effect can be eliminated by a slight modification of the simulation scheme.

Let us consider the method of Gasser *et al.* (1986). They propose an estimator $\hat{\sigma}^2$ of the form

$$\hat{\sigma}^2 \equiv \hat{\sigma}^2(\underline{Y}) = \frac{1}{n-2}\sum_{i=2}^{n-1}\frac{(Y_i - \alpha_i^{(1)}Y_{i-1} - \alpha_i^{(2)}Y_{i+1})^2}{(1-\alpha_i^{(1)2} - \alpha_i^{(2)2})} =: \underline{Y}^{\mathrm{T}}V\underline{Y},$$

where

$$\alpha_i^{(1)} := \frac{(x_{i+1}-x_i)}{(x_{i+1}-x_{i-1})} \quad \text{and} \quad \alpha_1^{(2)} := 1 - \alpha_i^{(1)}.$$

Obviously $E(\hat{\sigma}^2(\underline{\epsilon})) = \sigma^2$. A corrected version of this estimator providing unbiased estimates of σ^2 under H_0 is

$$\tilde{\sigma}^2 \equiv \tilde{\sigma}^2(\underline{Y}) = \frac{1}{1-\mathrm{tr}(G(G^{\mathrm{T}}G)^{-1}G^{\mathrm{T}}V)}\underline{Y}^{\mathrm{T}}(I-G(G^{\mathrm{T}}G)^{-1}G^{\mathrm{T}})V(I-G(G^{\mathrm{T}}G)^{-1}G^{\mathrm{T}})\underline{Y}. \tag{4.3}$$

Under H_0 we have $E(\tilde{\sigma}^2(\underline{Y})) = E(\tilde{\sigma}^2(\underline{\epsilon})) = \sigma^2$. Consistency of $\hat{\sigma}^2$ implies consistency of $\tilde{\sigma}^2$. Replacing σ^2 by $\tilde{\sigma}^2$ the estimated values of RSS_h are given by

$$\hat{\mathrm{RSS}}_h = \frac{1}{\tilde{\sigma}^2(\underline{Y})}\|(I-P_h)W_h\underline{Y}\|_2^2,$$

and under H_0 the distribution of

$$\hat{\mathrm{RSS}}_h = \frac{1}{\tilde{\sigma}^2(\underline{\epsilon})}\|(I-P_h)W_h\underline{\epsilon}\|_2^2 \tag{4.4}$$

does not depend on σ^2. Finitely exact critical values for the distribution of $\hat{\mathrm{RSS}}_h$ under H_0 can thus be obtained by Monte Carlo simulations from (4.4) by using standard normal errors.

4.3. Non-normal errors

Under some moment conditions on the error distribution, the above tests will remain asymptotically valid even for non-normal errors ϵ_i. For example when using either one of the methods proposed in examples 1 or 2 of section 3, under weak technical conditions we will obtain that the vectors $n^{-1/2}B_h^{\mathrm{T}}\underline{\epsilon}$ follow asymptotically a multivariate normal distribution for fixed h. One can then conclude that RSS_h is asymptotically χ^2 distributed. Since it has been shown in the proof of theorem 2 that under H_0 maximization of $\mathrm{RSS}_h - 2(h-L)$ over all h is asymptotically essentially equivalent to a maximization over only a finite number of elements $h \in H$, the asymptotic validity of the critical values $C^{\alpha}_{\max,n}$ for the $\mathrm{RSS}_{\max}$ based test is an immediate consequence. Similar results may be shown for the critical values of the tests relying on local linear smoothing.

Another approach to the treatment of non-normal errors may consist in the use of a bootstrap method: using a non-parametric estimator $\hat{m}_h$ of m relying on an approximately optimal choice

of h (determined, for example, by cross-validation), define residuals $\tilde{\epsilon}_i = Y_i - \hat{m}_h(x_i)$, for $1 \leq i \leq n$; calculate their mean, $\bar{\epsilon}$; and put $\hat{\epsilon}_i = \tilde{\epsilon}_i - \bar{\epsilon}$. Then determine a bootstrap approximation to the null distribution of RSS_h by resampling errors from the set $\tilde{\epsilon}_1, \ldots, \tilde{\epsilon}_n$.

5. Numerical results

5.1. Simulation study

A simulation study was carried out to investigate the power of the proposed tests. Since local linear estimators are more important in practice than the projection-based methods discussed in section 3, we concentrated on test statistics (4.1) and (4.2). The hypothesized model was $m(x) = \theta_1 + \theta_2 x$, i.e. m is a straight line. We considered two alternatives of different degree of smoothness.

$$v(x) = \sqrt{180}\{\tfrac{1}{12} - (x - \tfrac{1}{2})^2\}, \tag{5.1}$$

$$v(x) = \sqrt{2}\sin(4\pi x). \tag{5.2}$$

Design points $x_1, \ldots, x_2$ were regularly spaced on $J = [0, 1]$, and errors were taken to be Gaussian with variance $\sigma^2 = 1$. The Epanechnikov kernel $K(x) = \frac{4}{5}(1 - x^2)$ for $|x| \leq 1$ was applied in local linear fitting. Using the initial bandwidth $b_0 = 0.4$ a bandwidth sequence defined by $b_k = 0.8^k b_0$ was analysed. Sample sizes $n = 20$, 40 and 100 were considered, and the numerical work employed 10 000 simulations in each step.

Since we did not assume the error variance to be known, the modification (4.4) was used to determine critical values by Monte Carlo simulations. Recall that this procedure provides finitely exact critical values under H_0. The resulting values for $D^{\alpha}_{\max,n}$ were 4.14, 3.45, 2.21 for $n = 20$, 40, 120. In order to be able to judge the power of the test we compared it to the parametric F-Test which tests H_0 against H_1: $m(x) = \theta_1 + \theta_2 x + \vartheta v$ for a *prespecified* alternative v. This test is the best we could possibly do if we *knew* the true alternative. Since in practice the exact alternative is rarely known, this is not a very fair comparison, but it provides an upper bound to the maximal power a very good test could have in a given situation.

Based on alternative (5.1), Table 1 shows the number of rejections obtained for the RSS_h tests (bandwidths $b = 0.32$, 0.083) and for the $\mathrm{RSS}_{\max}$ test in comparison with the parametric F-test. We see that the results are surprisingly good, the $\mathrm{RSS}_{\max}$ based test being almost as powerful as an F-test. Tables 2 shows the corresponding results for alternative (5.2). Not surprisingly, the test is less powerful for this more "wiggly" alternative.

Table 1. *Rejections of H_0 under alternative* (5.1)

	Percentage of rejections			
	RSS_h $b = 0.32$	RSS_h $b = 0.08$	$\mathrm{RSS}_{\max}$	F-test
$n = 20, \vartheta = 1$	96	76	91	99
$n = 40, \vartheta = 1$	100	99	100	100
$n = 20, \vartheta = 0.5$	46	24	34	55
$n = 40, \vartheta = 0.5$	81	54	69	87
$n = 120, \vartheta = 0.5$	100	99	100	100
$n = 20, \vartheta = 0.25$	15	9	11	18
$n = 40, \vartheta = 0.25$	29	16	21	34
$n = 120, \vartheta = 0.25$	72	49	67	77

Scand. J. Stat. 26, 221-238

Table 2. *Rejections of H_0 under alternative (5.2)*

	Percentage of rejections			
	RSS_h $b = 0.32$	RSS_h $b = 0.08$	RSS_{max}	F-test
$n = 20, \vartheta = 1$	20	58	53	99
$n = 40, \vartheta = 1$	45	97	93	100
$n = 20, \vartheta = 0.5$	8	17	15	54
$n = 40, \vartheta = 0.5$	12	39	31	86
$n = 120, \vartheta = 0.5$	34	95	91	100
$n = 20, \vartheta = 0.25$	5	7	7	19
$n = 40, \vartheta = 0.25$	6	12	9	32
$n = 120, \vartheta = 0.25$	10	35	27	76

5.2. An application: household expenditures

In this section we consider an economic application. Starting with Engel (1857), a major issue in applied demand analysis has been the estimation of "cross sectional Engel curves". These are the conditional expectations of household expenditures on a commodity aggregate (like food, clothing, services, etc.) given total expenditure. Most work has been done in the context of parametric models of the form (1.1). The most important models are

$$m(x) = \theta_1 \cdot x + \theta_2 \cdot x \log(x), \tag{5.3}$$

$$m(x) = \theta_1 \cdot x + \theta_2 \cdot x \log(x) + \theta_3 \cdot x\{\log(x)\}^2, \tag{5.4}$$

$$m(x) = \theta_1 + \theta_2 \cdot x + \theta_3 \log(x), \tag{5.5}$$

where x denotes total expenditure. Each of these models has been frequently used in applications, see for the example Deaton (1986). Model (5.3) has been proposed by Working (1943), model (5.4) by Leser (1963), while model (5.5) stems from Deaton (1981).

We have tested these models by using the UK family expenditure survey (FES) data from 1968 to 1983. For each of these years the data reports the expenditures of approximately 7000 households. Each year households were selected at random from electorial registers. The data contains total expenditure and expenditures on nine commodity aggregates: housing, fuel, food, clothing, durables, transport, services, alcohol and tobacco, and "miscellaneous and other goods".

We normalized total expenditure by dividing through mean total expenditure (separately for each year). Let (Y_{jkt}, x_{jt}) denote the resulting data for each commodity $k = 1, \ldots, 9$. Here j indexes the household, and t denotes the respective year. Data for very rich and very poor households is sparse and not very reliable. We thus only considered the interval [0.25, 2.5] for the x_{jt}. Normalized total expenditures for approximately 95% of all households fall into this range.

In order to simplify computations we made a prebinning step. We chose a grid $0.25 =: x_0^* < x_1^* < \cdots < x_n^* < 2.5 =: x_{n+1}^*$ of $n = 231$ points and used the binned data (Y_{ikt}, x_i) for testing. Here, for given i, k, t Y_{ikt} denotes the average over all Y_{jkt} corresponding to some

$$x_{jt} \in \left[\frac{(x_{i-1}^* + x_i^*)}{2}, \frac{(x_i^* + x_{i+1}^*)}{2}\right], \quad \text{and} \quad x_i = x_i^*.$$

For fixed commodity k and year t we then applied a goodness-of-fit test to the models (5.3–5.5). Errors were heteroscedastic and a modification described in the appendix was applied. We have $k = 9$ and $t = 16$. Hence, in total 144 separate tests were done. There is theoretical reason to assume that a possible alternative will be very smooth. We thus decided to rely on the RSS_h based test using regression splines with a small number of knots. Table 3 reports the total

Table 3. *The number of rejections for the expenditure curve models*

Model	No. of rejections out of 144 tests $\alpha = 0.05$	$\alpha = 0.01$
(5.3)	124	100
(5.4)	53	29
(5.5)	56	34

number of rejections when smoothing was based on cubic regression splines with 5 knots at 0.2, 0.7, 1.25, 1.85, 2.7, ($h = 7$):

We see that the data quite drastically reject the hypothesis that either one of the models (5.3), (5.4), (5.5) is appropriate for modelling cross-sectional Engel curves. We also tried larger values of h. The rejection rates were less significant as is to be expected from the discussion of section 3. This is in line with observations made by Deaton (1986) who was unable to reject models (5.4) or (5.5) with established goodness-of-fit tests.

Acknowledgements

This research was supported by Deutsche Forschungsgemeinschaft, Sonderforschungsbereich 373 at the University of Berlin, and by the contract "Projet d'Actions de Recherche Concertées" (PARC No. 93/98-164) of the Belgian Government.

References

Azzalini, A., Bowman, A. W. & Härdle, W. (1989). On the use of nonparametric regression for model checking. *Biometrika* **76**, 1–11.

Cox, D., Koh, E., Wahba, G. & Yandell, B. S. (1988). Testing the (parametric) null hypothesis in (semiparametric) partial and generalized spline models. *Ann. Statist.* **16**, 113–119.

Deaton, A. S. (1981). Three essays on a Sri Lankan household survey. *Living Standards Measurement Study W. D. No. 11*, The World Bank, Washington.

Deaton, A. (1986). Demand analysis in *Handbook of econometrics*, (eds Z. Grilliches & M. D. Intrilligator), Vol. 3, 1777–1837. North Holland, New York.

de Boor, C. (1978). A *practical guide to splines*, Springer, New York.

Department of Employment (1969–1984). *FES, Family expenditure survey, reports for 1968–1983*. HMSO, London.

Devore, R. & Lorentz, G. (1991). *Constructive approximation*. Springer, New York.

Durbin, J. & Knott, M. (1972). Components of Cramér–von Mises statistics. *J. Roy. Statist. Soc. Ser. B* **34**, 290–307.

Engel, E. (1857). Die Produktions- und Consumptionsverhältnisse des Königreichs Sachsen. Reprinted 1895 in: *Bull. Inst. Internl. Statist.* **9**, 1–54.

Eubank, R. L. & Spiegelmann, C. H. (1990). Testing the goodness of fit of a linear model via nonparametric regression techniques. *J. Amer. Statist. Assoc.* **85**, 387–392.

Fan, J. & Gijbels, I. (1996). *Local polynomial modelling and its applications*. Chapman & Hall, London.

Gasser, Th. & Müller, H. G. (1984). Estimating regression functions and their derivatives by the kernel method. *Scand. J. Statist.* **11**, 171–185.

Gasser, Th., Sroka, L. & Jennen-Steinmetz, C. (1986). Residual variance and residual pattern in nonlinear regression. *Biometrika* **73**, 625–633.

Hall, P., Hart, J. D. & Titterington, D. M. (1990). Bootstrap test for difference between means in nonparametric regression. *J. Amer. Statist. Assoc.* **85**, 1039–1049.

Hall, P., Kay, J. M. & Titterington, D. M. (1990). Asymptotically optimal difference based estimation of variance in nonparametric regression. *Biometrika*, **77**, 521–529.

Härdle, W. & Mammen, E. (1993). Comparing nonparametric versus parametric fits. *Ann. Statist.* **21**, 1926–1947.

Hart, J. D. (1997). *Nonparametric smoothing and lack-of-fit tests*. Springer, New York.
Hastie, T. J. & Tibshirani, R. J. (1990). *Generalized additive models*. Chapman & Hall, London.
Hong (1993). Consistent specification testing using optimal nonparametric kernel estimation. Center for Analytical Economics Working Paper No. 93-13, Cornell University.
Ingster, Y. I. (1982). Minimax nonparametric detection of signals in white Gaussian noise. *Problems Inform. Transmission* **18**, 130–140.
Kneip, A. (1994). Ordered linear smoothers. *Ann. Statist.* **22**, 835–866.
Leser, C. E. V. (1963). The form of Engel functions. *Econometrica* **31**, 694–703.
Li. K. C. (1987). Asymptotic optimality for C_P, C_L, cross-validation and generalized cross-validation: discrete index set. *Ann. Statist.* **15**, 958–976.
Mallows, C. L. (1973). Some comments on C_P. *Technometrics* **15**, 661–675.
Raz, J. (1990). Testing for no effect when estimating a smooth function by nonparametric regression: a randomization approach. *J. Amer. Statist. Assoc.* **85**, 132–138.
Rice, J. A. (1984). Bandwidth choice for nonparametric regression. *Ann. Statist.* **12**, 1215–1231.
Schumaker, L. L. (1990). *Spline functions*. Wiley, New York.
Working, H. (1943). Statistical laws of family expenditure. *J. Amer. Statist. Assoc.* **38**, 43–56.

Received May 1996, in final form June 1998

Alois Kneip, Institut de Statistique, Université Catholique de Louvain, Voie du Roman Pays, 20, 1348 Louvain-la-Neuve, Belgium.

Appendix. Heteroscedastic errors

The data described in section 5.2 involves error terms which are heteroscedastic. An appropriate model consists in assuming that the errors ϵ_{ikt} satisfy $\mathrm{var}(\epsilon_{ikt}) = \sigma^2_{kt}(x_i^*)$, where the $\sigma^2_{kt}(\cdot)$ are smooth functions.

The approach used to deal with this situations can be described as follows: in a first step the variances $\sigma^2_{kt}(\cdot)$ are estimated. Following Gasser *et al.* (1986) we define squared pseudo-residuals

$$r^2_{ikt} := \frac{(Y_{ikt} - \alpha_i^{(1)} Y_{i-1,kt} - \alpha_i^{(2)} Y_{i-1,kt})^2}{(1 - \alpha_i^{(1)^2} - \alpha_i^{(2)^2})}$$

where

$$\alpha_i^{(1)} := \frac{(x^*_{i+1} - x^*_i)}{x^*_{i+1} - x^*_{i-1})}, \quad \alpha_1^{(2)} := 1 - \alpha_i^{(1)}$$

By using Gasser–Müller kernel estimators (Gasser & Müller, 1984) we then smooth these squared pseudo-residuals to obtain estimators $\hat\sigma^2_{kt}(\cdot)$ of $\sigma^2_{kt}(\cdot)$. For a second order kernel under some weak regularity conditions it can be shown that the $\hat\sigma^2_{kt}(\cdot)$ are consistent estimators of $\sigma^2_{kt}(\cdot)$ as $n \to \infty$.

In a second step data and model are transformed by multiplying with $1/\hat\sigma_{kt}(x_i^*)$, and the tests developed in sections 3 or 4 can then applied to the transformed values. More precisely, the tests are based on

$$Y^*_{ikt} := \frac{Y_{ikt}}{\hat\sigma_{kt}(x_i^*)}, \quad g^*_r(\cdot) := \frac{g_r(\cdot)}{\hat\sigma_{kt}(\cdot)}.$$

Note that the transformed error term

$$\epsilon^*_{ikt} := \frac{\epsilon_{ikt}}{\hat\sigma_{kt}(x_i^*)}$$

satisfies

Scand. J. Stat. 26, 221-238

$$\text{var}(\epsilon^*_{ikt}) \approx \text{var}\left(\frac{\epsilon_{ikt}}{\sigma_{kt}(x^*_i)}\right) = 1.$$

If $\hat{\sigma}^2_{it}(\cdot)$ is constructed in the manner described above, then it can be shown that the theoretical results of sections 3 and 4 generalize to the present situation, provided b is chosen in a reasonable way.

(1999) Härdle, W. and Kneip, A.
Testing a Regression Model when we have smooth alternatives in mind.

Journal Of Time Series Analysis (1999), Vol. 20, No 5, ISSN 0143-9782

Nonparametric Autoregression with Multiplicative Volatility and Additive Mean *

Lijian Yang
Department of Statistics and Probability
Michigan State University
East Lansing, Michigan 48824
U. S. A.

Wolfgang Härdle
Institut für Statistik und Ökonometrie
Humboldt Universität zu Berlin
Spandauer Str.1, D-10178 Berlin
Germany

Jens P. Nielsen
PFA Pension
Sundkrogsgade 4
2100 Copenhagen
Denmark

October 30, 1998

Abstract

For over a decade, nonparametric modelling has been successfully applied to study nonlinear structures in financial time series. It is well known that the usual nonparametric models often have less than satisfactory performance when dealing with more than one lag. When the mean has an additive structure, however, better estimation methods are available which fully exploit such a structure. Although in the past such nonparametric applications had been focused more on the estimation of the conditional mean, it is equally if not more important to measure the future risk of the series along with the mean. For the volatility function, i.e., the conditional variance given the past, a multiplicative structure is more appropriate than an additive one, as the volatility is a positive scale function and a multiplicative model provides a better interpretation of each lagged value's influence on such a function. In this paper we consider the joint estimation of both the additive mean and the multiplicative volatility. The technique used is marginally integrated local polynomial estimation. The procedure is applied to the DEM/USD (Deutsche Mark/US Dollar) daily exchange returns.

*Acknowledgements: This research was financially supported by Sonderforschungsbereich 373 "Quantifikation und Simulation Ökonomischer Prozesse" Deutsche Forschungsgemeinschaft, at Humboldt Universität zu Berlin. We appreciate the interests of our colleagues that motivated our work, especially Christian Hafner, Helmut Lütkepohl, and Rolf Tschernig. We have also benefited from the elegant formulation of the marginal integration technique by Eric Severance-Lossin and Stefan Sperlich. Finally, we want to thank the two anonymous referees who gave us many constructive comments.

Journal Of Time Series Analysis (1999), Vol. 20, No 5, ISSN 0143-9782

Keywords: Additive Mean, Geometric Ergodicity, Geometric Mixing, Local Polynomial Regression, Marginal Integration, Multiplicative Volatility, Stationary Probability Density.

1 Introduction

The prediction of financial time series based on daily data is, in general, difficult, since after differencing most of the structure in the mean disappears. This is why random-walk-based models have been used in this context. The situation is different, though, for high frequency time series such as foreign exchange rates. Autoregressive models have been applied for such data with specific assumptions on the error distribution, see Engle (1982), Engle and Ng (1993). Some of the most common nonlinear autoregressive models were proposed by Tong (1978, 1983), Haggan and Ozaki (1981), Chan and Tong (1986), and Granger and Teräsvirta (1993). In particular it is important not only to predict future values but also to evaluate the risk, or the volatility of the series. In the class of ARCH models the volatility or the scale of innovative random shocks is a function of past values. Over the past fifteen years, the strict parametric forms of these models have been questioned and more flexible nonparametric approaches have been studied as an alternative, see Robinson (1983, 1984), Meese and Rose (1991), Drost and Nijman (1993), Engle and Gonzalez-Rivera (1991). A more recent review is Härdle and Chen (1995).

One of the models studied for foreign exchange rates, for example, is the CHARN (conditional heteroskedastic autoregressive nonlinear) model with one lag (Bossaerts, Härdle, and Hafner, 1996)

$$Y_i = m(Y_{i-1}) + s(Y_{i-1})\xi_i \tag{1.1}$$

where $\{\xi_i\}_{i\geq 1}$ are i.i.d random variables $E(\xi_i) = E(\xi_i^3) = 0$, $E(\xi_i^2) = 1$ and $E(\xi_i^4) = m_4 < \infty$, and Y_0 is independent of the$\{\xi_i\}$'s. An analysis of the estimated residuals still revealed autocorrelation. Hence, more than one lagged variable in the modelling of the mean function $m(\bullet)$ and the scale function $s(\bullet)$ seems to be the necessary step in a further analysis.

We consider therefore in this paper the CHARN model of the form

$$Y_i = m(Y_{i-1}, Y_{i-2}, ..., Y_{i-d}) + s(Y_{i-1}, Y_{i-2}, ..., Y_{i-d})\xi_i \tag{1.2}$$

where $\{\xi_i\}_{i\geq 1}$ are as in (1.1) and $Y_0, Y_1, ..., Y_{d-1}$ are random variables independent of the $\{\xi_i\}$'s. The conditional volatility function is $v(Y_{i-1}, Y_{i-2}, ..., Y_{i-d}) = s^2(Y_{i-1}, Y_{i-2}, ..., Y_{i-d})$. This form of the CHARN model in financial time series has been studied by Gouriéroux and Monfort (1992) and Masry and Tjøstheim (1995a). The estimation problem for the functions $m(\bullet)$ and $v(\bullet)$ has been treated in Härdle and Tsybakov (1997) in the case of $d = 1$ with the local polynomial regression method. Härdle, Tsybakov and Yang (1998) studied vector autoregression with arbitrary number of lags and dimension. We define the CHARN model for general dimensions, however, from a practical point of view, the method can be expected to suffer from the statistical imprecision introduced by a large number of lags. In particular in the small sample size case. We illustrate the method with a foreign

Härdle, W., Nielsen, J.P. And Yang, L. (1999) Nonparametric Autoregression With Multiplicative Volatility And Additive Mean

exchange rate application. Through lag selection, see Tschernig and Yang (1997), we ended up using the first lag and the third lag of the time series.

Stone (1982) showed in the i.i.d. regression case that if the mean function $m(\bullet)$ is a sum of univariate functions, then the one dimensional convergence rate can be achieved for the estimation of $m(\bullet)$'s component functions. Tools for analysis of additive models in this context have been developed by Hastie and Tibshirani (1990), including the BRUTO algorithm for nonparametric modelling, which Chen and Tsay (1993a,b) applied to autoregressive time series. The "integration method" (but not the term marginal integration) was introduced by Auestad and Tjøstheim (1991) and further explored by Tjøstheim and Auestad (1994) for the precise analysis of additive model estimators which was previously unavailable. It provides closed form bias and variance expressions of the one dimensional function estimator. The term marginal integration was introduced in Linton and Nielsen (1995), who worked in the independent identically distributed regression setting. Marginal integration has recently been employed in the autoregression setting by Masry and Tjøstheim (1995a,b) and in the independent identically distributed regression setting by Linton and Härdle (1996) and Severance-Lossin and Sperlich (1995).

The idea of the integration method is quite straightforward: in the regression setting for instance, if the mean function $m(x_1, x_2, ..., x_d)$ is a sum of univariate functions, say

$$m(x_1, x_2, ..., x_d) = c + \sum_{\beta=1}^{d} m_\beta(x_\beta) \tag{1.3}$$

then

$$m_\beta(x_\beta) = \int m(x_1, x_2, ..., x_d) dF(x_1, ..., \widehat{x_\beta}, ..., x_d) - C$$

where $F(x_1, ..., \widehat{x_\beta}, ..., x_d)$ is the joint distribution function of all the variables $X_1, ..., X_d$ with the β-th X_β removed, and C is an additive constant. Hence each component function m_β is identified from $m(x_1, x_2, ..., x_d)$ through a simple integration procedure. Linton and Nielsen (1995) introduced the idea of applying integration estimation to multiplicative structures in dimension two, in this paper we extend the integration formula to multiplicative volatility functions of any dimension.

To estimate the parameters in the CHARN model, we have to estimate the conditional mean function $m(\bullet)$ and the conditional variance or volatility function $v(\bullet)$ at the same time. The flexibility of our CHARN model is important in a number of economic applications. For example prediction of financial time series, where the volatility function often plays an even more important role than the mean function. It is therefore beneficial to obtain the joint estimation of both $m(\bullet)$ and $v(\bullet)$ for model (1.2). The volatility function $v(\bullet)$ measures the scale and is always positive, therefore it seems more appropriate to model its changes multiplicatively rather than additively, as in the EGARCH model of Nelson (1991). In this paper we jointly estimate the additive (mean) and the multiplicative (volatility) functions with the integration method.

We therefore assume that the mean function $m(\bullet)$ is additive while the volatility function $v(Y_{i-1}, Y_{i-2}, ..., Y_{i-d}) = s(Y_{i-1}, Y_{i-2}, ..., Y_{i-d})^2$ is multiplicative

$$m(Y_{i-1}, Y_{i-2}, ..., Y_{i-d}) = c_m + \sum_{\beta=1}^{d} m_\beta(Y_{i-\beta}), \tag{1.4}$$

$$v(Y_{i-1}, Y_{i-2}, ..., Y_{i-d}) = c_v \prod_{\beta=1}^{d} v_\beta(Y_{i-\beta}) \tag{1.5}$$

where c_m and c_v are constants, $\{m_\beta(\bullet)\}_{\beta=1}^d$ and $\{v_\beta(\bullet)\}_{\beta=1}^d$ are sets of unknown functions. Besides the better rate of convergence for the estimation of $\{m_\beta(\bullet)\}_{\beta=1}^d$ and $\{v_\beta(\bullet)\}_{\beta=1}^d$ as discussed above, these univariate functions also allows one to quantify the impact of each lagged variable $Y_{i-\beta}$ on the mean and volatility more directly.

To formulate the identifiability conditions for the functions $\{m_\beta(\bullet)\}_{\beta=1}^d$ and $\{v_\beta(\bullet)\}_{\beta=1}^d$, the process Y_i has to converge to a stationary distribution. If we denote by $\mathbf{X}_i$ the vector $(Y_{i-1}, Y_{i-2}, ..., Y_{i-d})^T$, then $\{\mathbf{X}_i\}$ is a d-dimensional Markov process. Many authors, such as Tweedie (1975), Nummelin and Tuominen (1982), Mokkadem (1987), Tjøstheim (1990) and Diebolt and Guégan (1993) developed geometric ergodicity criteria for Markov processes. Here we state some general assumptions

A1: *The random variable ξ_i has a density function $p(\bullet)$. This density $p(\bullet)$ and the volatility function $v(\bullet)$ are strictly positive in a neighborhood of x;*

A2: *There exists an $r > 0$ such that for $\sum_{\beta=1}^d |y_{i-\beta}| > r$, the functions $m(\bullet)$ and $s(\bullet)$ satisfy:*

$$|m(y_{i-1}, y_{i-2}, ..., y_{i-d})| \leq C_1(1 + \sum_{\beta=1}^{d} |y_{i-\beta}|)$$

$$|s(y_{i-1}, y_{i-2}, ..., y_{i-d})| \leq C_2(1 + \sum_{\beta=1}^{d} |y_{i-\beta}|)$$

with $C_1 + C_2 E|\xi_1| < 1/d$.

These assumptions are standard in this context in order to prevent the process from either dying out or exploding. Ango Nze (1992) proved the following

Lemma 1.1 *Under assumptions A1 and A2, the process $\{\mathbf{X}_i\}$ is geometrically ergodic, i.e., it is ergodic, with stationary probability measure $\pi(\bullet)$ such that, for almost every $\mathbf{x}$,*

$$\|P^n(\bullet \mid \mathbf{x}) - \pi(\bullet)\|_{TV} = O(\rho^n)$$

for some $0 \leq \rho < 1$, where $P^n(\bullet \mid \mathbf{x})$ is the probability measure of $\mathbf{X}_n$ given $\mathbf{X}_d = \mathbf{x}$ and $\|\bullet\|_{TV}$ is the total variation distance.

This lemma ensures that the process $\{\mathbf{X}_i\}$ is asymptotically stationary. We denote by $F(\bullet)$ the stationary distribution function. For all $1 \leq \alpha \leq d$, we denote by $F_\alpha(\bullet)$ the stationary distribution function of the α-th variable, and $\overline{F}(\bullet)$ the stationary distribution function with the α-th variable deleted. We allow ourselves to use the short-hand notation Y_β for $Y_{i-\beta}$. Let x_β denote the deterministic version of $Y_{i-\beta}$. We can now state the identifiability conditions

A3: $Em_\beta(Y) = \int m_\beta(x_\beta) dF_\beta(x_\beta) = 0$, *for any Y that has distribution $F_\beta(\bullet)$, and for all $1 \leq \beta \leq d$;*

Härdle, W., Nielsen, J.P. And Yang, L. (1999) Nonparametric Autoregression With Multiplicative Volatility And Additive Mean

A4: $E\prod_{1\le\beta\le d,\beta\neq\alpha} v_\beta(Y_\beta) = \prod_{1\le\beta\le d,\beta\neq\alpha} v_\beta(x_\beta)d\overline{F}(\overline{x}) = 1$ *for any* $(Y_1, Y_2, ...Y_d)$ *that has distribution* $F(\bullet)$, *and for all* $1 \le \alpha \le d$.

Let $\mathbf{x} = (x_1, x_2, ..., x_d)^T \in \mathbb{R}^d$ be a point where we will estimate the mean and volatility functions. We define for every $1 \le \alpha \le d$, $M_\alpha(x_\alpha) = c_m + m_\alpha(x_\alpha)$, $V_\alpha(x_\alpha) = c_v v_\alpha(x_\alpha)$, then

$$m(\mathbf{x}) = \sum_{\beta=1}^{d} M_\beta(x_\beta) - (d-1)c_m, \; v(\mathbf{x}) = c_v^{-(d-1)} \prod_{\beta=1}^{d} V_\beta(x_\beta). \tag{1.6}$$

In what follows, we adopt the notation $\mathbf{X_i} = (Y_{i-\alpha}, \overline{Y}_i)$ to highlight a particular direction of interest $Y_{i-\alpha}$, for all $1 \le \alpha \le d$, while $\overline{Y}_i$ is the $d-1$ dimensional vector that consists of all the rest $Y_{i-\beta}$'s, $1 \le \beta \le d, \beta \neq \alpha$. Assumptions A3 and A4 yield the following marginal integration formulae for the unknown functions

$$\int m(x_\alpha, \overline{x})d\overline{F}(\overline{x}) = M_\alpha(x_\alpha) = c_m + m_\alpha(x_\alpha), \tag{1.7}$$

$$\int v(x_\alpha, \overline{x})d\overline{F}(\overline{x}) = V_\alpha(x_\alpha) = c_v v_\alpha(x_\alpha), \tag{1.8}$$

which show that the univariate functions $\{m_\beta(\bullet)\}_{\beta=1}^d$ and $\{v_\beta(\bullet)\}_{\beta=1}^d$ are identifiable from the functions $m(\bullet)$ and $v(\bullet)$ up to some constants. And similar formulae exist for these constants as well

$$c_m = \int m(x)dF(x) = E(Y), \; c_v = \left\{ \frac{1}{d} \sum_{\alpha=1}^{d} \int \prod_{1\le\beta\le d,\beta\neq\alpha} V_\beta(x_\beta)d\overline{F}(\overline{x}) \right\}^{\frac{1}{d-1}}. \tag{1.9}$$

These are the basic equations that will be used later in our estimation procedure.

In Section 2, we present the estimators of $\{m_\beta(\bullet)\}_{\beta=1}^d$ and $\{v_\beta(\bullet)\}_{\beta=1}^d$ and study their asymptotic properties. In Section 3, we discuss the application of the result to DM/USD daily return data. In Section 4, proofs of theorems are given. Inspection of the proofs in Section 4 shows that the result of the present paper also holds (with obvious reformulation) for the multivariate nonparametric regression model with heteroskedastic errors: $Y_i = m(X_{i1}, X_{i2}, ..., X_{id}) + s(X_{i1}, X_{i2}, ..., X_{id})\xi_i$, where ξ_i are as in (1.2), $(X_{i1}, X_{i2}, ..., X_{id}, Y_i)$ are i.i.d., and the design points $\{X_{i1}, X_{i2}, ..., X_{id}\}$ are independent of $\{\xi_i\}$.

2 The Estimators

The estimators given in this section are based on local polynomial regression, first studied by Stone (1977) and Katkovnik (1979). The idea, as will be seen below, is to estimate an unknown function locally by polynomials, whose coefficients are calculated through kernel weighted least squares, see also Tsybakov (1986), Ruppert and Wand (1994), Wand and Jones (1995) and Fan and Gijbels (1996).

Now we let $p > 0$ be any odd integer which will be the degree of polynomial used later. For any function $K(\bullet)$ we denote $\|K\|_2^2 = \int K^2(u)du$, while for a kernel function $K(\bullet)$ we define $K_h(u) = K(u/h)/h$, and $\mu_r(K) = \int u^r K(u)du$. We shall consider two kernel functions $K(\bullet)$ and $L(\bullet)$ that satisfy

Härdle, W., Nielsen, J.P. And Yang, L. (1999) Nonparametric Autoregression With Multiplicative Volatility And Additive Mean

A5: *Both kernels $K(\bullet)$ and $L(\bullet)$ are bounded, symmetric, compactly supported and Lipschitz continuous with $\int K(u)\,du = \int L(u)\,du = 1$; while $K(\bullet)$ is positive, the kernel $L(\bullet)$ is of order $q > (d-1)(p+1)/2$*

When estimating functions $m_\alpha(\bullet)$ and $v_\alpha(\bullet)$ for a particular α, a multiplicative kernel is used consisting of K for the α-th variable and L for all other variables.

We assume the following about the functions involved in the estimation

A6: *The functions $m_\alpha(\bullet)$'s and $v_\alpha(\bullet)$'s have bounded Lipschitz continuous $(p+1)$-th derivatives for all $1 \le \alpha \le d$.*

A7: *The stationary distribution function $F(\bullet)$ has a density $\varphi(\bullet)$. The function $\varphi(\bullet)$, together with the densities $\varphi_\alpha(\bullet)$ of $F_\alpha(\bullet)$ and $\overline{\varphi}(\bullet)$ of $\overline{F}(\bullet)$ are all uniformly bounded away from zero and infinity and have bounded Lipschitz continuous $(p+1)$-th derivatives, for all $1 \le \alpha \le d$.*

Lastly, we assume the following for two bandwidths, g for the kernel L, h for the kernel K

A8: *Bandwidths g and h satisfy $\frac{g^{d-1}}{h^2} \longrightarrow \infty$, $\frac{nhg^{2(d-1)}}{\ln^2(n)} \longrightarrow \infty$, $\frac{g^q}{h^{p+1}} \to 0$ and $h = h_0 n^{\frac{-1}{2p+3}}$.*

Note that this A8 requires that $L(\bullet)$ have the order as in A5. In particular, if one uses local linear regression, i.e., $p = 1$, then the order of $L(\bullet)$ is $q > d-1$.

One can define the integration estimator for $M_\alpha(x_\alpha)$ as

$$\widehat{M}_\alpha(x_\alpha) = \int \widehat{m}(x_\alpha, \overline{x}) d\widehat{\overline{F}}(\overline{x}) = (n-d+1)^{-1} \sum_{l=d}^{n} \widehat{m}(x_\alpha, \overline{Y}_l),$$

where $\widehat{m}(x_\alpha, \overline{x})$ is an estimate of $m(\bullet)$ at $(x_\alpha, \overline{x})$, and $\widehat{\overline{F}}(\overline{x})$ is the empirical cumulative distribution function (ecdf). The estimator $\widehat{M}_\alpha(x_\alpha)$ is hereby based on the sample version of equation (1.7). The estimator for c_m is simply the sample mean of Y_j's according to (1.9)

$$\hat{c}_m = \hat{E}(Y) = (n-d+1)^{-1} \sum_{j=d}^{n} Y_j$$

where $\widehat{E}$ is the empirical mean of Y. These estimators are then used to obtain estimators for $m_\alpha(x_\alpha)$ and $m(\mathbf{x})$

$$\widehat{m}_\alpha(x_\alpha) = \widehat{M}_\alpha(x_\alpha) - \hat{c}_m,$$

$$\widehat{m}(\mathbf{x}) = \hat{c}_m + \sum_{\beta=1}^{d} \widehat{m}_\beta(x_\beta) = \sum_{\beta=1}^{d} \widehat{M}_\beta(x_\beta) - (d-1)\hat{c}_m.$$

We now define $\widehat{m}(x_\alpha, \overline{Y}_l)$ as follows. For all $l = d, d+1, ..., n$, and $\lambda = 0, ..., p$ let

$$Z = \left\{ (Y_{i-\alpha} - x_\alpha)^\lambda \right\}_{(n-d+1)\times(p+1)},$$

$$W_l = \operatorname{diag}\left\{\frac{1}{(n-d+1)} K_h(Y_{i-\alpha} - x_\alpha) L_g(\overline{Y}_i - \overline{Y}_l)\right\}_{i=d}^{n},$$

where we denote

$$\operatorname{diag}(a) = \begin{bmatrix} a_1 & 0 & \cdots & 0 \\ 0 & a_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & a_k \end{bmatrix}$$

for any vector

$$a = \begin{pmatrix} a_1 \\ \vdots \\ a_k \end{pmatrix} \in I\!R^k.$$

Also write

$$Y = (Y_i)_{d \le i \le n},\ Y^2 = (Y_i^2)_{d \le i \le n},$$

and let e_λ be a $(p+1)$ vector of zeros whose $(\lambda+1)$-element is 1. Then

$$\widehat{m}(x_\alpha, \overline{Y}_l) = e_0^T \left(Z^T W_l Z\right)^{-1} Z^T W_l Y,$$

which is the usual local polynomial estimator of $m(\bullet)$ at $(x_\alpha, \overline{Y}_l)$ of order p in the α-th direction and order 0 in all the other directions. Our estimator $\widehat{M}_\alpha(x_\alpha)$ is therefore

$$\widehat{M}_\alpha(x_\alpha) = (n-d+1)^{-1} \sum_{l=d}^{n} e_0^T \left(Z^T W_l Z\right)^{-1} Z^T W_l Y.$$

.

Note that

$$E(Y_i^2 \mid X_i) = m^2(X_i) + v(X_i),$$

thus similar estimator for $V_\alpha(x_\alpha)$ based on equation (1.8) is defined as

$$\widehat{V}_\alpha(x_\alpha) = (n-d+1)^{-1} \sum_{l=d}^{n} \left\{ e_0^T \left(Z^T W_l Z\right)^{-1} Z^T W_l Y^2 - \widehat{m}(x_\alpha, \overline{Y}_l)^2 \right\},$$

and that of c_v is based on (1.9):

$$\widehat{c}_v = \left\{ \frac{1}{d} \sum_{\alpha=1}^{d} \frac{1}{(n-d+1)} \sum_{j=d}^{n} \prod_{1 \le \beta \le d, \beta \ne \alpha} \widehat{V}_\beta(Y_{j-\beta}) \right\}^{\frac{1}{d-1}}.$$

One then obtains estimators for $v_\alpha(x_\alpha)$ and $v(\mathbf{x})$ as the following:

$$\widehat{v}_\alpha(x_\alpha) = \widehat{V}_\alpha(x_\alpha) \widehat{c}_v^{-1},$$

$$\widehat{v}(\mathbf{x}) = \widehat{c}_v \prod_{\beta=1}^{d} \widehat{v}_\beta(x_\beta) = \widehat{c}_v^{-(d-1)} \prod_{\beta=1}^{d} \widehat{V}_\beta(x_\beta).$$

Our first theorem gives the estimation result of the mean functions

Theorem 1 *Under assumptions A1-A8, as* $n \to \infty$, *for any* α

$$\sqrt{nh}\left\{\widehat{M}_\alpha(x_\alpha) - M_\alpha(x_\alpha) - h^{p+1}b_{m\alpha}(x_\alpha)\right\} \xrightarrow{D} N\left\{0, \sigma^2_{m\alpha}(x_\alpha)\right\} \tag{2.1}$$

where

$$b_{m\alpha}(x_\alpha) = \frac{\mu_{p+1}(K_0^*)}{(p+1)!} m_\alpha^{(p+1)}(x_\alpha)$$

and

$$\sigma^2_{m\alpha}(x_\alpha) = \|K_0^*\|_2^2 \int \frac{v}{\varphi}(x_\alpha, w)\overline{\varphi}^2(w)dw.$$

While for any $\alpha \neq \beta$, *as* $n \to \infty$, *one has*

$$\operatorname{cov}\left[\sqrt{nh}\left\{\widehat{M}_\alpha(x_\alpha) - M_\alpha(x_\alpha)\right\}, \sqrt{nh}\left\{\widehat{M}_\beta(x_\beta) - M_\beta(x_\beta)\right\}\right] \to 0. \tag{2.2}$$

Furthermore, as $n \to \infty$

$$\sqrt{n}(\widehat{c}_m - c_m) \xrightarrow{D} N\left\{0, \sigma^2_{cm}(\mathbf{x})\right\}$$

for some implicitly-defined constant σ^2_{cm}. *The asymptotics of* $\sqrt{nh}\left\{\widehat{m}_\alpha(x_\alpha) - m_\alpha(x_\alpha)\right\}$ *are the same as those of the* $\sqrt{nh}\left\{\widehat{M}_\alpha(x_\alpha) - M_\alpha(x_\alpha)\right\}$, *while*

$$\sqrt{nh}\left\{\widehat{m}(\mathbf{x}) - m(\mathbf{x}) - h^{p+1}b_m(\mathbf{x})\right\} \xrightarrow{D} N\left\{0, \sigma^2_m(\mathbf{x})\right\} \tag{2.3}$$

where

$$b_m(\mathbf{x}) = \sum_{\alpha=1}^{d} b_{m\alpha}(x_\alpha)$$

and

$$\sigma^2_m(\mathbf{x}) = \sum_{\alpha=1}^{d} \sigma^2_{m\alpha}(x_\alpha).$$

The second theorem is about the estimation of the volatility functions

Theorem 2 *Under assumptions A1-A8, as* $n \to \infty$, *for any* α

$$\sqrt{nh}\left\{\widehat{V}_\alpha(x_\alpha) - V_\alpha(x_\alpha) - h^{p+1}b_{V\alpha}(x_\alpha)\right\} \xrightarrow{D} N\left\{0, \sigma^2_{V\alpha}(x_\alpha)\right\} \tag{2.4}$$

where

$$b_{V\alpha}(x_\alpha) = \frac{\mu_{p+1}(K_0^*)}{(p+1)!}\left\{V_\alpha^{(p+1)}(x_\alpha) + 2m_\alpha^{(p+1)}(x_\alpha)M(x_\alpha)\right\}$$

$$- \int 2b_m(x_\alpha, w)m(x_\alpha, w)\overline{\varphi}(w)dw$$

and

$$\sigma^2_{V\alpha}(x_\alpha) = \|K_0^*\|_2^2 \int \frac{v\left[m_4 v + 4m^2\right]}{\varphi}(x_\alpha, w)\overline{\varphi}^2(w)dw.$$

Also, as $n \to \infty$

$$\operatorname{cov}\left[\sqrt{nh}\left\{\widehat{V}_\alpha(x_\alpha) - V_\alpha(x_\alpha)\right\}, \sqrt{nh}\left\{\widehat{M}_\alpha(x_\alpha) - M_\alpha(x_\alpha)\right\}\right]$$

$$\rightarrow 2 \left\| K_0^* \right\|_2^2 \int \frac{vm}{\varphi}(x_\alpha, w)\overline{\varphi}^2(w)dw = c_{V\alpha}(x_\alpha) \tag{2.5}$$

while for any $\alpha \neq \beta$ *one has*

$$\begin{aligned} \operatorname{cov}\left[\sqrt{nh}\left\{\widehat{V}_\alpha(x_\alpha) - V_\alpha(x_\alpha)\right\}, \sqrt{nh}\left\{\widehat{V}_\beta(x_\beta) - V_\beta(x_\beta))\right\}\right] \rightarrow 0, \\ \operatorname{cov}\left[\sqrt{nh}\left\{\widehat{V}_\alpha(x_\alpha) - V_\alpha(x_\alpha)\right\}, \sqrt{nh}\left\{\widehat{M}_\beta(x_\beta) - M_\beta(x_\beta))\right\}\right] \rightarrow 0. \end{aligned} \tag{2.6}$$

Furthermore

$$\sqrt{n}(\widehat{c}_v - c_v - b_c h^{p+1}) \xrightarrow{D} N(0, \sigma_{cv}^2)$$

for some implicitly-defined constant σ_{cv}^2 *and*

$$b_c = \frac{1}{d(d-1)c_v^{d-2}} \sum_{\alpha=1}^{d} \int \sum_{1 \leq \beta \leq d, \beta \neq \alpha} \left\{ \prod_{1 \leq \gamma \leq d, \gamma \neq \alpha, \beta} V_\gamma(y_\gamma) \right\} b_{v\beta}(y_\beta)\varphi(y)dy.$$

For any α

$$\sqrt{nh}\left\{\widehat{v}_\alpha(x_\alpha) - v_\alpha(x_\alpha) - h^{p+1}b_{v\alpha}(x_\alpha)\right\} \xrightarrow{D} N\left\{0, \sigma_{v\alpha}^2(x_\alpha)\right\} \tag{2.7}$$

where

$$b_{v\alpha}(x_\alpha) = \frac{1}{c_v}\left\{b_{V\alpha}(x_\alpha) - b_c v_\alpha(x_\alpha)\right\}$$

and

$$\sigma_{v\alpha}^2(x_\alpha) = \frac{1}{c_v^2}\sigma_{V\alpha}^2(x_\alpha),$$

while

$$\sqrt{nh}\left\{\widehat{v}(\mathbf{x}) - v(\mathbf{x}) - h^{p+1}b_v(\mathbf{x})\right\} \xrightarrow{D} N\left\{0, \sigma_v^2(\mathbf{x})\right\}$$

where

$$b_v(\mathbf{x}) = v(\mathbf{x})\left\{\sum_{\beta=1}^{d} \frac{b_{V\beta}(x_\beta)}{V_\beta(x_\beta)} - (d-1)c_v^{-1}b_c\right\}$$

and

$$\sigma_v^2(\mathbf{x}) = v^2(\mathbf{x}) \sum_{\beta=1}^{d} \frac{\sigma_{V\beta}^2(x_\beta)}{V_\beta^2(x_\beta)}.$$

The next theorem summarizes all the previous results together in the form of joint asymptotic normality for all estimators

Theorem 3 *Under assumptions A1-A8, denote by* $\mathbf{B}(\mathbf{x})$ *the vector valued function*

$$\left\{b_{m1}(x_1), b_{m2}(x_2), ..., b_{md}(x_d), b_m(\mathbf{x}), b_{v1}(x_1), b_{v2}(x_2), ..., b_{vd}(x_d), b_v(\mathbf{x}), 0, \sqrt{n}b_c\right\}^T$$

and $\mathbf{\Sigma}(\mathbf{x})$ *the following matrix*

$$\begin{bmatrix} \mathbf{\Sigma_{11}} & \mathbf{\Sigma_{12}} & \mathbf{\Sigma_{13}} & \mathbf{\Sigma_{14}} & \mathbf{0}_{d\times 1} & \mathbf{0}_{d\times 1} \\ \mathbf{\Sigma_{21}} & \mathbf{\Sigma_{22}} & \mathbf{\Sigma_{23}} & \mathbf{\Sigma_{24}} & 0 & 0 \\ \mathbf{\Sigma_{31}} & \mathbf{\Sigma_{32}} & \mathbf{\Sigma_{33}} & \mathbf{\Sigma_{34}} & \mathbf{0}_{d\times 1} & \mathbf{0}_{d\times 1} \\ \mathbf{\Sigma_{41}} & \mathbf{\Sigma_{42}} & \mathbf{\Sigma_{43}} & \mathbf{\Sigma_{44}} & 0 & 0 \\ \mathbf{0}_{1\times d} & 0 & \mathbf{0}_{1\times d} & 0 & \sigma_{cm}^2 & 0 \\ \mathbf{0}_{1\times d} & 0 & \mathbf{0}_{1\times d} & 0 & 0 & \sigma_{cv}^2 \end{bmatrix}$$

where

$$\Sigma_{11} = \mathrm{diag}\left\{\sigma^2_{m\alpha}(x_\alpha)\right\}_{\alpha=1}^{d},\ \Sigma_{22} = \sigma^2_m(\mathbf{x}),\ \Sigma_{33} = \mathrm{diag}\left\{\sigma^2_{v\alpha}(x_\alpha)\right\}_{\alpha=1}^{d},\ \Sigma_{44} = \sigma^2_v(\mathbf{x}),$$

$$\Sigma_{12} = \Sigma_{21}^{\mathbf{T}} = \left\{\sigma^2_{m\alpha}(x_\alpha)\right\}_{1\le\alpha\le d},\ \Sigma_{13} = \Sigma_{31}^{\mathbf{T}} = \mathrm{diag}\left\{\frac{c_{V\alpha}(x_\alpha)}{c_v}\right\}_{\alpha=1}^{d},$$

$$\Sigma_{14} = \Sigma_{41}^{\mathbf{T}} = \left\{\frac{c_{V\alpha}(x_\alpha)}{c_v}\frac{v(\mathbf{x})}{V_\alpha(x_\alpha)}\right\}_{1\le\alpha\le d},\ \Sigma_{23} = \Sigma_{32}^{\mathbf{T}} = \left\{\frac{c_{V\alpha}(x_\alpha)}{c_v}\right\}_{1\le\alpha\le d}^{T},$$

$$\Sigma_{24} = \Sigma_{42}^{\mathbf{T}} = \sum_{\alpha=1}^{d}\frac{c_{V\alpha}(x_\alpha)}{c_v}\frac{v(\mathbf{x})}{V_\alpha(x_\alpha)},\ \Sigma_{34} = \Sigma_{43}^{\mathbf{T}} = \left\{\sigma^2_{v\alpha}(x_\alpha)\frac{v(\mathbf{x})}{V_\alpha(x_\alpha)}\right\}_{1\le\alpha\le d},$$

then, as $n \to \infty$

$$\sqrt{nh}\left\{\begin{array}{c}\widehat{m}_1(x_1) - m_1(x_1)\\ \widehat{m}_2(x_2) - m_2(x_2)\\ \vdots\\ \widehat{m}_d(x_d) - m_d(x_d)\\ \widehat{m}(\mathbf{x}) - m(\mathbf{x})\\ \widehat{v}_1(x_1) - v_1(x_1)\\ \widehat{v}_2(x_2) - v_2(x_2)\\ \vdots\\ \widehat{v}_d(x_d) - v_d(x_d)\\ \widehat{v}(\mathbf{x}) - v(\mathbf{x})\\ \frac{1}{\sqrt{h}}(\widehat{c}_m - c_m)\\ \frac{1}{\sqrt{h}}(\widehat{c}_v - c_v)\end{array}\right\} - \mathbf{B}(\mathbf{x})h^{p+1} \xrightarrow{D} N\left\{\mathbf{0}_{(2d+4)\times(2d+4)}, \mathbf{\Sigma}(\mathbf{x})\right\}$$

We comment here that although Theorem 3 is obtained for local polynomial of degree p, where p is an odd integer, the same result holds for p even, in particular, for $p = 0$, i.e., the Nadaraya-Watson estimator. We choose to have p odd here because it does not involve the derivatives of the design density in the bias and variance expressions, and thus "design-adaptive".

3 An Application

To illustrate our method with an example, we study the daily returns of the DEM/USD exchange rates from Jan.2 1980 to May 26 1986, a total of 1603 observations. The data is plotted in Figure 1.

We estimate the conditional mean and volatility functions of this series at lags 1 and 3. The choice of these two lags is based on the findings of Tschernig and Yang (1997), who have developed a nonparametric final prediction error criterion for determining significant lagged variables. For the estimation, we use subjectively selected bandwidths $h = 0.0062$, $g = 0.0074$, and the Nadaraya-Watson estimators. We found that except for some boundary

effects, the mean functions $m_\beta(\bullet)$'s are very close to zero. The estimated volatility function $\hat{v}_\beta(\bullet)$'s depicted in Figures 2 and 3, however, provide some fresh insights. Both the computation and graphics are done in XploRe, see Härdle, Klinke and Turlach (1995).

Figures 2 and 3 show that the lagged variables impact the volatility function asymmetrically as both $\hat{v}_1(\bullet)$ and $\hat{v}_3(\bullet)$ are quite skewed, especially $\hat{v}_3(\bullet)$; one can see this by comparing $\hat{v}_1(\bullet)$ and $\hat{v}_3(\bullet)$ with their ordinary least squares quadratic fits which are the thin lines in the pictures. Some kind of nonparametric testing would be needed in order to check the significance of these observed features.

Our observations about $\hat{v}_1(\bullet)$ and $\hat{v}_3(\bullet)$ have added weight to what some other studies had also suggested: that the basic GARCH model is perhaps inappropriate for the process we have here. Our analysis here has gone a step further in nonparametric estimation of times series as the significant lagged variables are first identified by a nonparametric criteria, see Tschernig and Yang (1997) for details. This example of identifying significant lags and measuring their impacts points to a new comprehensive nonparametric approach to time series analysis.

Figure 1: The daily returns

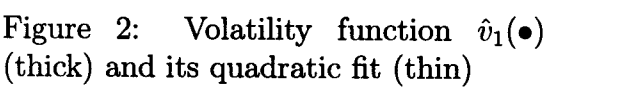


Figure 2: Volatility function $\hat{v}_1(\bullet)$ (thick) and its quadratic fit (thin)

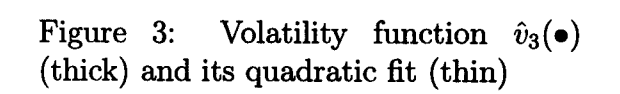


Figure 3: Volatility function $\hat{v}_3(\bullet)$ (thick) and its quadratic fit (thin)

4 Proofs

Theorems 1 through 3 are proved in this section by the marginal integration technique as in Severance-Lossin and Sperlich (1995). We make use of the following geometric mixing results

Lemma 4.1 (Davydov(1973)).
Under assumptions A1 and A2, if further, $\mathbf{X}_d$ *is distributed with the stationary distribution* $\pi(\bullet)$*, then the process* $\{\mathbf{X}_i\}$ *is geometrically strongly mixing with the mixing coefficients satisfying* $\alpha(n) \leq c_0\rho_0^n$ *for some* $c_0 > 0$ *and* $0 < \rho_0 < 1$.

By arguments which are very similar to those used in Härdle, Tsybakov, and Yang (1998), the above mixing lemma entails that the sample mean of any bounded continuous function of the observations Y_j converges in both probability and mean to the stationary population mean. The situation here is slightly more complicated than in that paper as one now has to average functions of two variables Y_j and $\overline{Y}_l$, one at a time. Nevertheless, the difference is more formal than substantial. We therefore neither state nor prove any such results here, but use them to derive the various formulae of asymptotic biases and variances as these are the new contributions of this paper.

The proof of the next lemma is standard and omitted. It employs the strong mixing condition of Lemma 1.1 and Lemma 4.1.

Lemma 4.2 *Let* $D_l = \left(Z^T W_l Z\right)^{-1} - \frac{1}{\varphi(x_\alpha, \overline{Y}_l)} H^{-1} S^{-1} H^{-1}$

$$Cov(D_l, D_k) = \rho^{|l-k|} \left\{ O_p(h + \ln n / \sqrt{nhg^{d-1}}) \right\}^2 \tag{4.1}$$

uniformly in x_α *and* $\overline{Y}_l$*, where* $H = \operatorname{diag}\left(h^\lambda\right)_{0 \leq \lambda \leq p}$.

Proofs of asymptotic normality in this section are based on the central limit theorem of Liptser and Shirjaev (1980). Conditions for applying this theorem will not be verified here as they are all standard. Set $S = (\int u^{s+t}K(u)du)_{0\le s,t\le p}$, which contains all the moments of S up to order $2p$. Denote $S^{-1} = (s_{st})_{0\le s,t\le p}$ and define

$$K_\lambda^*(u) = \sum_{t=0}^{p} s_{\lambda t}u^t K(u). \tag{4.2}$$

This $K_\lambda^*(\bullet)$ is called the λ-th equivalent kernel. It has the following moments

$$\int u^q K_\lambda^*(u)\,du = \left\{ \begin{array}{ll} 0 & q \le p, \quad q \ne \lambda \\ 1 & q = \lambda \\ \Lambda_\lambda & q = p+1 \end{array} \right\}. \tag{4.3}$$

and $K_0^*(\bullet)$ would yield the bias rates of $n^{-2p/(2p+1)}$ for local polynomial estimation, see Wand and Jones (1995).

To prove Theorem 1, we begin by observing the following simple equation

$$e_0^T \left(Z^T W_l Z\right)^{-1} Z^T W_l Z e_\lambda = \left\{ \begin{array}{ll} 0 & 0 \ne \lambda \\ 1 & 0 = \lambda \end{array} \right\} \tag{4.4}$$

thus

$$\widehat{M}_\alpha(x_\alpha) - M_\alpha(x_\alpha) = (n-d+1)^{-1}\sum_{l=d}^{n} e_0^T \left(Z^T W_l Z\right)^{-1} Z^T W_l Y$$

$$-(n-d+1)^{-1}\sum_{l=d}^{n} e_0^T \left(Z^T W_l Z\right)^{-1} Z^T W_l Z e_0 M_\alpha(x_\alpha)$$

$$-(n-d+1)^{-1}\sum_{l=d}^{n}\sum_{\nu=1}^{p} \frac{m_\alpha^{(\nu)}(x_\alpha)}{\nu!} e_0^T \left(Z^T W_l Z\right)^{-1} Z^T W_l Z e_\nu.$$

$$= (n-d+1)^{-1}\sum_{l=d}^{n} e_0^T \left(Z^T W_l Z\right)^{-1} Z^T W_l \{Y - M_\alpha(x_\alpha)\}$$

$$-(n-d+1)^{-1}\sum_{l=d}^{n}\sum_{\nu=1}^{p} \frac{m_\alpha^{(\nu)}(x_\alpha)}{\nu!} e_0^T \left(Z^T W_l Z\right)^{-1} Z^T W_l Z e_\nu.$$

Now Assumption A3 combined with the strong mixing properties of our process imply that for every $\beta = 1, 2, ..., d, \beta \ne \alpha$

$$(n-d+1)^{-1}\sum_{l=d}^{n} m_\beta(Y_{l-\beta}) = O_p(1/\sqrt{n}),$$

and thus by (4.4), one also has (using the mixing properties of the process, see Lemma 1.1, Lemma 4.1 and Lemma 4.2)

$$(n-d+1)^{-1}\sum_{l=d}^{n} e_0^T \left(Z^T W_l Z\right)^{-1} Z^T W_l Z e_0 m_\beta(Y_{l-\beta}) = O_p(1/\sqrt{n}).$$

So one has

$$\widehat{M}_\alpha(x_\alpha) - M_\alpha(x_\alpha) =$$

$$(n-d+1)^{-1}\sum_{l=d}^{n} e_0^T \left(Z^T W_l Z\right)^{-1} Z^T W_l \left\{ Y - \sum_{1\le\beta\le d,\beta\neq\alpha} m_\beta(Y_{l-\beta}) - M_\alpha(x_\alpha) \right\}$$

$$-(n-d+1)^{-1}\sum_{l=d}^{n}\sum_{\nu=1}^{p} \frac{m_\alpha^{(\nu)}(x_\alpha)}{\nu!} e_0^T \left(Z^T W_l Z\right)^{-1} Z^T W_l Z e_\nu + O_p(1/\sqrt{n})$$

$$= (n-d+1)^{-1}\sum_{l=d}^{n} e_0^T \left(Z^T W_l Z\right)^{-1} Z^T W_l$$

$$\left\{ Y - \sum_{\nu=1}^{p} \frac{m_\alpha^{(\nu)}(x_\alpha)}{\nu!} Z e_\nu - \sum_{1\le\beta\le d,\beta\neq\alpha} m_\beta(Y_{l-\beta}) - M_\alpha(x_\alpha) \right\},$$

or

$$\widehat{M}_\alpha(x_\alpha) - M_\alpha(x_\alpha) =$$

$$= (n-d+1)^{-1}\sum_{l=d}^{n} e_0^T \left(Z^T W_l Z\right)^{-1} Z^T W_l \left\{ Y - c_m - \sum_{\nu=0}^{p} \frac{m_\alpha^{(\nu)}(x_\alpha)}{\nu!} Z e_\nu - \sum_{1\le\beta\le d,\beta\neq\alpha} m_\beta(Y_{l-\beta}) \right\}. \tag{4.5}$$

Note that the λ-th element of $Z^T W_l \left\{ Y - c_m - \sum_{\nu=0}^{p} \frac{m_\alpha^{(\nu)}(x_\alpha)}{\nu!} Z e_\nu - \sum_{1\le\beta\le d,\beta\neq\alpha} m_\beta(Y_{l-\beta}) \right\}$

is

$$(n-d+1)^{-1}\sum_{j=d}^{n} (Y_{j-\alpha} - x_\alpha)^\lambda K_h(Y_{j-\alpha} - x_\alpha) L_g(\overline{Y}_j - \overline{Y}_l)$$

$$\left\{ Y_j - c_m - \sum_{\nu=0}^{p} \frac{m_\alpha^{(\nu)}(x_\alpha)}{\nu!} (Y_{j-\alpha} - x_\alpha)^\nu - \sum_{1\le\beta\le d,\beta\neq\alpha} m_\beta(Y_{l-\beta}) \right\}$$

$$= I_{\lambda l,1} + I_{\lambda l,2} + I_{\lambda l,3}$$

in which

$$I_{\lambda l,1} = (n-d+1)^{-1}\sum_{j=d}^{n} I_{\lambda l j,1}$$

where

$$I_{\lambda l j,1} = (Y_{j-\alpha} - x_\alpha)^\lambda K_h(Y_{j-\alpha} - x_\alpha) L_g(\overline{Y}_j - \overline{Y}_l) \left\{ m_\alpha(Y_{j-\alpha}) - \sum_{\nu=0}^{p} \frac{m_\alpha^{(\nu)}(x_\alpha)}{\nu!} (Y_{j-\alpha} - x_\alpha)^\nu \right\}, \tag{4.6}$$

$$I_{\lambda l,2} = \sum_{1\le\beta\le d,\beta\neq\alpha} (n-d+1)^{-1}\sum_{j=d}^{n} I_{\lambda l\beta j,2}$$

where

$$I_{\lambda l\beta j,2} = (Y_{j-\alpha} - x_\alpha)^\lambda K_h(Y_{j-\alpha} - x_\alpha) L_g(\overline{Y}_j - \overline{Y}_l) \left\{ m_\beta(Y_{j-\beta}) - m_\beta(Y_{l-\beta}) \right\}, \tag{4.7}$$

Härdle, W., Nielsen, J.P. And Yang, L. (1999) Nonparametric Autoregression With Multiplicative Volatility And Additive Mean

and

$$I_{\lambda l,3} = (n-d+1)^{-1} \sum_{j=d}^{n} I_{\lambda lj,3}, \tag{4.8}$$

where

$$I_{\lambda lj,3} = (Y_{j-\alpha} - x_\alpha)^\lambda K_h(Y_{j-\alpha} - x_\alpha) L_g(\overline{Y}_j - \overline{Y}_l) s(X_j) \xi_j. \tag{4.9}$$

Lemma 4.3 *As* $n \to \infty$,

$$E(I_{\lambda l_1 j_1,1} I_{\lambda l_2 j_2,1}) = \rho^{\min(|l_1-l_2|,|j_1-j_2|)} O(h^{2\lambda}/hg^{d-1})$$

uniformly, for $\lambda = 0,...,p$ *and* $l_1, l_2, j_1, j_2 = d,...,n$.

$$E(I_{\lambda l_1 \beta_1 j_1,1} I_{\lambda l_2 \beta_2 j_2,1}) = \rho^{\min(|l_1-l_2|,|j_1-j_2|)} O(h^{2\lambda}/hg^{d-1})$$

uniformly, for $\lambda = 0,...,p$ *and* $l_1, l_2, j_1, j_2 = d,...,n$ *and* $1 \le \beta \le d$.

$$E(I_{\lambda l_1 j_1,3} I_{\lambda l_2 j_2,3}) = \rho^{\min(|l_1-l_2|,|j_1-j_2|)} O(h^{2\lambda}/hg^{d-1})$$

uniformly, for $\lambda = 0,...,p$ *and* $l_1, l_2, j_1, j_2 = d,...,n$.

Proof. We only show this for the first case

$$E(I_{\lambda l_1 j_1,3} I_{\lambda l_2 j_2,3}) = \rho^{\min(|l_1-l_2|,|j_1-j_2|)} \int (w_\alpha - x_\alpha)^{2\lambda} K_h^2(w_\alpha - x_\alpha) L_g^2(\overline{w} - \overline{Y}_l) v(w) \varphi(w) dw \{1 + o(1)\},$$

where we have used Lemma 1.1. By a change of variable $w_\alpha = x_\alpha + hu_\alpha$, $\overline{w} = \overline{Y}_l + g\overline{u}$

$$E(I_{\lambda l_1 j_1,3} I_{\lambda l_2 j_2,3}) = \left\{hg^{d-1}\right\}^{-1} \{1 + o(1)\}$$

$\int (hu_\alpha)^{2\lambda} K^2(u_\alpha) L^2(\overline{u}) v(x_\alpha + hu_\alpha, \overline{Y}_l + g\overline{u}) \varphi(x_\alpha + hu_\alpha, \overline{Y}_l + g\overline{u}) du$. Q. E. D.

Now

$$O(h^{2\lambda}/nhg^{d-1}) \left\{O_p(h + \ln n/\sqrt{nhg^{d-1}})\right\}^2 = O_p\left\{h^{2\lambda+2}/nhg^{d-1} + h^{2\lambda} \ln^2 n/(n^2 h^2 g^{2(d-1)})\right\}$$

$$= h^{2\lambda}/nh O_p\left(h^2/g^{d-1} + \ln^2 n/nhg^{2(d-1)}\right) = o_p\left(h^{2\lambda}/nh\right)$$

by using assumption A8. Employing Lemma 4.2 and Lemma 4.3 now gives

$$\sum_{\lambda=0}^{p} (n-d+1)^{-1} \sum_{l=d}^{n} e_0^T \left\{\left(Z^T W_l Z\right)^{-1} - \frac{1}{\varphi(x_\alpha, \overline{Y}_l)} H^{-1} S^{-1} H^{-1}\right\} e_\lambda (I_{\lambda l,1} + I_{\lambda l,2} + I_{\lambda l,3})$$

$$= \sum_{\lambda=0}^{p} h^{-\lambda} o_p\left(h^\lambda/\sqrt{nh}\right) = o_p\left(1/\sqrt{nh}\right) = o_p\left\{h^{p+1}\right\} = o_p\left\{n^{-(p+1)/(2p+3)}\right\}.$$

If we only had to consider the diagonal terms, then this fact is easily recongnised (this is if we could ignore the correlation of the "I"-terms with the rest). The correlation can however

be taken care of by writing up the $I_{\lambda l,k}$'s as sums (se above), squaring the expression and conditioning on the "I-components". The exponential decay of the correlations in Lemma 4.2 and Lemma 4.3 ensures that the order of magnitude is the same as if only the diagonal terms were considered.

Proof of Theorem 1. Making the aforementioned substitution, one has in particular

$$\widehat{M}_\alpha(x_\alpha) - M_\alpha(x_\alpha) - o_p\left\{h^{p+1}\right\} =$$

$$= (n-d+1)^{-1}\sum_{l=d}^{n}\frac{1}{\varphi(x_\alpha,\overline{Y}_l)}e_0^T H^{-1}S^{-1}H^{-1}Z^T W_l$$

$$\left\{Y - c_m - \sum_{\nu=0}^{p}\frac{m_\alpha^{(\nu)}(x_\alpha)}{\nu!}Ze_\nu - \sum_{1\le\beta\le d,\beta\neq\alpha} m_\beta(Y_{l-\beta})\right\}$$

which, by using (4.6), (4.7), (4.7) and the definition (4.2), equals

$$= (n-d+1)^{-1}\sum_{l=d}^{n}\frac{1}{\varphi(x_\alpha,\overline{Y}_l)}(n-d+1)^{-1}\sum_{j=d}^{n}K_{0h}^*(Y_{j-\alpha}-x_\alpha)L_g(\overline{Y}_j-\overline{Y}_l)$$

$$\left[m_\alpha(Y_{j-\alpha}) - \sum_{\nu=0}^{p}\frac{m_\alpha^{(\nu)}(x_\alpha)}{\nu!}(Y_{j-\alpha}-x_\alpha)^\nu + \sum_{1\le\beta\le d,\beta\neq\alpha}\{m_\beta(Y_{j-\beta}) - m_\beta(Y_{l-\beta})\} + s(X_j)\xi_j\right]$$

$$= (n-d+1)^{-1}\sum_{j=d}^{n}\{1+o_p(1)\}\int dw\frac{K_{0h}^*(Y_{j-\alpha}-x_\alpha)}{\varphi(x_\alpha,\overline{Y}_j-gw)}\overline{\varphi}(\overline{Y}_j-gw)L(w)$$

$$\left[m_\alpha(Y_{j-\alpha}) - \sum_{\nu=0}^{p}\frac{m_\alpha^{(\nu)}(x_\alpha)}{\nu!}(Y_{j-\alpha}-x_\alpha)^\nu + \sum_{1\le\beta\le d,\beta\neq\alpha}\{m_\beta(Y_{j-\beta}) - m_\beta(Y_{j-\beta}-gw_\beta)\} + s(X_j)\xi_j\right].$$

And because L has order q, so the above equals

$$(n-d+1)^{-1}\sum_{j=d}^{n}\{1+o_p(1)\}\frac{K_{0h}^*(Y_{j-\alpha}-x_\alpha)}{\varphi(x_\alpha,\overline{Y}_j)}\overline{\varphi}(\overline{Y}_j)$$

$$\left\{m_\alpha(Y_{j-\alpha}) - \sum_{\nu=0}^{p}\frac{m_\alpha^{(\nu)}(x_\alpha)}{\nu!}(Y_{j-\alpha}-x_\alpha)^\nu + s(X_j)\xi_j\right\} + O_p(g^q). \tag{4.10}$$

Thus we have shown that

$$\widehat{M}_\alpha(x_\alpha) - M_\alpha(x_\alpha) = B + V + o_p(h^{p+1})$$

in which

$$B = (n-d+1)^{-1}\sum_{j=d}^{n}\frac{K_{0h}^*(Y_{j-\alpha}-x_\alpha)}{\varphi(x_\alpha,\overline{Y}_j)}\overline{\varphi}(\overline{Y}_j)\left\{m_\alpha(Y_{j-\alpha}) - \sum_{\nu=0}^{p}\frac{m_\alpha^{(\nu)}(x_\alpha)}{\nu!}(Y_{j-\alpha}-x_\alpha)^\nu\right\}$$

and

$$V = (n-d+1)^{-1}\sum_{j=d}^{n}\frac{K_{0h}^*(Y_{j-\alpha}-x_\alpha)}{\varphi(x_\alpha,\overline{Y}_j)}\overline{\varphi}(\overline{Y}_j)\{s(X_j)\xi_j\}.$$

Härdle, W., Nielsen, J.P. And Yang, L. (1999) Nonparametric Autoregression With Multiplicative Volatility And Additive Mean

Now (by using the mixing properties of our process)

$$B = \{1 + o_p(1)\} \int \frac{K_{0h}^*(z - x_\alpha)}{\varphi(x_\alpha, w)} \overline{\varphi}(w) \left\{ m_\alpha(Y_{j-\alpha}) - \sum_{\nu=0}^{p} \frac{m_\alpha^{(\nu)}(x_\alpha)}{\nu!} (Y_{j-\alpha} - x_\alpha)^\nu \right\} \varphi(z, w) dz dw.$$

After substituting $z = x_\alpha + hu$, B becomes

$$B = \{1 + o_p(1)\} \int \frac{K_0^*(u)}{\varphi(x_\alpha, w)} \overline{\varphi}(w) \left\{ m_\alpha(x_\alpha + hu) - \sum_{\nu=0}^{p} \frac{1}{\nu!} m_\alpha^{(\nu)}(x_\alpha)(hu)^\nu \right\} \varphi(x_\alpha + hu, w) du dw$$

which, by using the moment properties of the equivalent kernel as in (4.3), equals

$$\{1 + o_p(1)\} \frac{\mu_{p+1}(K_0^*)}{(p+1)!} m_\alpha^{(p+1)}(x_\alpha) b_{m\alpha}(x_\alpha) h^{p+1} = b_{m\alpha}(x_\alpha) h^{p+1} + o_p(h^{p+1}) \tag{4.11}$$

where $b_{m\alpha}(x_\alpha)$ is as given in Theorem 1. Meanwhile, V has mean zero and its variance is

$$(n - d + 1)^{-1} \int \left\{ \frac{K_{0h}^*(z - x_\alpha)}{\varphi(x_\alpha, w)} \overline{\varphi}(w) s(z, w) \right\}^2 \varphi(z, w) dz dw \{1 + o(1)\}$$

$$= n^{-1} h^{-1} \sigma_{m\alpha}^2(x_\alpha) \{1 + o(1)\}. \tag{4.12}$$

Equations (4.11) and (4.12) together establish (2.1). Equation (2.2) is derived by standard technique as in Linton and Härdle (1996). Equation (2.3) and all the other remaining formulas of Theorem 1, then follow directly from (2.1) and (2.2) as the various $\sqrt{nh}\left\{\widehat{M}_\alpha(x_\alpha) - M_\alpha(x_\alpha)\right\}$'s are all asymptotically uncorrelated, so the variance of $\sqrt{nh}\{\widehat{m}(\mathbf{x}) - m(\mathbf{x})\}$ is simply the sum of all their variances, the mean of $\sqrt{nh}\{\widehat{m}(\mathbf{x}) - m(\mathbf{x})\}$ is simply the sum of all their means. Q. E. D.

Proof of Theorem 2. We prove similar results for $\widehat{V}_\alpha(x_\alpha)$

$$\widehat{V}_\alpha(x_\alpha) - V_\alpha(x_\alpha) = (n - d + 1)^{-1} \sum_{l=d}^{n} \left\{ e_0^T \left(Z^T W_l Z \right)^{-1} Z^T W_l Y^2 - \widehat{m}(x_\alpha, \overline{Y}_l)^2 \right\} - V_\alpha(x_\alpha)$$

$$= (n - d + 1)^{-1} \sum_{l=d}^{n} e_0^T \left(Z^T W_l Z \right)^{-1} Z^T W_l \left\{ Y^2 - \widehat{m}(x_\alpha, \overline{Y}_l)^2 - V_\alpha(x_\alpha) \right\}$$

$$= (n-d+1)^{-1} \sum_{l=d}^{n} e_0^T \left(Z^T W_l Z \right)^{-1} Z^T W_l \left\{ Y^2 - m(x_\alpha, \overline{Y}_l)^2 + m(x_\alpha, \overline{Y}_l)^2 - \widehat{m}(x_\alpha, \overline{Y}_l)^2 - V_\alpha(x_\alpha) \right\}.$$

Now note that by Assumption A4

$$(n - d + 1)^{-1} \sum_{j=d}^{n} \prod_{\beta \neq \alpha} v_\beta(Y_{j-\beta}) = 1 + O_p(1/\sqrt{n})$$

and also that

$$Y_j^2 = m(X_j)^2 + 22 m(X_j) s(X_j) \xi_j + v(X_j)(\xi_j^2 - 1) + v(X_j)$$

So similar to (4.10), we have

$$\widehat{V}_\alpha(x_\alpha) - V_\alpha(x_\alpha) = T_1 + T_2 + T_3 + T_4 + T_5 + o_p(h^{p+1})$$

where

$$T_1 = (n-d+1)^{-1} \sum_{l=d}^{n} \left\{ m(x_\alpha, \overline{Y}_l)^2 - \widehat{m}(x_\alpha, \overline{Y}_l)^2 \right\},$$

$$T_2 = (n-d+1)^{-1} \sum_{j=d}^{n} \frac{K_{0h}^*(Y_{j-\alpha} - x_\alpha)}{\varphi(x_\alpha, \overline{Y}_j)} \overline{\varphi}(\overline{Y}_j) \left\{ m(X_j)^2 - m(x_\alpha, \overline{Y}_j)^2 \right\},$$

$$T_3 = (n-d+1)^{-1} \sum_{j=d}^{n} \frac{K_{0h}^*(Y_{j-\alpha} - x_\alpha)}{\varphi(x_\alpha, \overline{Y}_j)} \overline{\varphi}(\overline{Y}_j) \left\{ v(X_j) - V_\alpha(x_\alpha) \prod_{\beta \neq \alpha} v_\beta(Y_{j-\beta}) \right\},$$

$$T_4 = (n-d+1)^{-1} \sum_{j=d}^{n} \frac{K_{0h}^*(Y_{j-\alpha} - x_\alpha)}{\varphi(x_\alpha, \overline{Y}_j)} \overline{\varphi}(\overline{Y}_j) \left\{ 2m(X_j) s(X_j) \xi_j \right\},$$

$$T_5 = (n-d+1)^{-1} \sum_{j=d}^{n} \frac{K_{0h}^*(Y_{j-\alpha} - x_\alpha)}{\varphi(x_\alpha, \overline{Y}_j)} \overline{\varphi}(\overline{Y}_j) \left\{ v(X_j)(\xi_j^2 - 1) \right\}.$$

We derive the asymptotics of each of these terms. Recall that Theorem 1 provides the following

$$\sqrt{nh} \left\{ \widehat{m}(\mathbf{x}) - m(\mathbf{x}) - h^{p+1} b_m(\mathbf{x}) \right\} \xrightarrow{D} N\left\{ 0, \sigma_m^2(\mathbf{x}) \right\}$$

therefore

$$\begin{aligned} T_1 &= -(n-d+1)^{-1} \sum_{l=d}^{n} 2 \left\{ m(x_\alpha, \overline{Y}_l) - \widehat{m}(x_\alpha, \overline{Y}_l) \right\} m(x_\alpha, \overline{Y}_l) + o_p(h^{p+1}) \\ &= -2E \left\{ m(x_\alpha, \overline{Y}_n) - \widehat{m}(x_\alpha, \overline{Y}_n) \right\} m(x_\alpha, \overline{Y}_n) + o_p(h^{p+1}) \\ &= -h^{p+1} \int 2 b_m(x_\alpha, w) m(x_\alpha, w) \overline{\varphi}(w) dw + o_p(h^{p+1}). \end{aligned} \tag{4.13}$$

Next we see, by using substitution $z_1 = x_\alpha + hu$, that

$$\begin{aligned} T_2 &= \{1 + o_p(1)\} \int \frac{K_{0h}^*(z - x_\alpha)}{\varphi(x_\alpha, w)} \overline{\varphi}(w) \left\{ m(z,w)^2 - m(x_\alpha, w)^2 \right\} \varphi(z,w) dz dw \\ &= \frac{\mu_{p+1}(K_0^*)}{(p+1)!} \int 2 m_\alpha^{(p+1)}(x_\alpha) m(x_\alpha, w) \overline{\varphi}(w) dw + o_p(h^{p+1}) \\ &= \frac{2\mu_{p+1}(K_0^*)}{(p+1)!} m_\alpha^{(p+1)}(x_\alpha) M(x_\alpha) + o_p(h^{p+1}) \end{aligned} \tag{4.14}$$

$$\begin{aligned} T_3 &= \{1 + o_p(1)\} \int \frac{K_{0h}^*(z - x_\alpha)}{\varphi(x_\alpha, w)} \overline{\varphi}(w) \left\{ V_\alpha(z) \overline{V}_\alpha(w) - V_\alpha(x_\alpha) \overline{V}_\alpha(w) \right\} \varphi(z,w) dz dw \\ &= \frac{\mu_{p+1}(K_0^*)}{(p+1)!} \int V_\alpha^{(p+1)}(x_\alpha) \overline{V}_\alpha(w) \overline{\varphi}(w) dw + o_p(h^{p+1}) \end{aligned}$$

Härdle, W., Nielsen, J.P. And Yang, L. (1999) Nonparametric Autoregression With Multiplicative Volatility And Additive Mean

$$= \frac{\mu_{p+1}(K_0^*)}{(p+1)!} V_\alpha^{(p+1)}(x_\alpha) + o_p(h^{p+1}). \tag{4.15}$$

To calculate the terms T_4 and T_5, note first that they both have mean zero and are uncorrelated, so it is only necessary to calculate their variances and the sum.

$$\mathrm{var}(T_4) = (n-d+1)^{-1} E\left\{ \frac{K_{0h}^*(Y_{n-\alpha} - x_\alpha)\overline{\varphi}(\overline{Y}_n)}{\varphi(x_\alpha, \overline{Y}_n)} 2m(X_d)s(X_d) \right\}^2 \{1 + o(1)\}$$

$$= (n-d+1)^{-1} \int \left\{ \frac{K_{0h}^*(z - x_\alpha)}{\varphi(x_\alpha, w)} 2m(z,w)s(z,w)\overline{\varphi}(w) \right\}^2 \varphi(z,w)dzdw \{1 + o(1)\}$$

$$= \frac{1}{nh} \|K_0^*\|_2^2 \int \frac{4m^2 v}{\varphi}(x_\alpha, w)\overline{\varphi}^2(w)dw \{1 + o(1)\} \tag{4.16}$$

and similarly

$$\mathrm{var}(T_5) = \frac{1}{nh} \|K_0^*\|_2^2 \int \frac{m_4 v^2}{\varphi}(x_\alpha, w)\overline{\varphi}^2(w)dw \{1 + o(1)\}. \tag{4.17}$$

Putting together equations (4.13) through (4.17) gives the asymptotic expressions of $\widehat{V}_\alpha(x_\alpha)$ in Theorem 2. To get the formula for $c_{V\alpha}(x)$ in (2.5), note that the variance term V in the proof of Theorem 1 is uncorrelated to all the T_i's except T_4, and their asymptotic correlation is (plus some higher order term)

$$(n-d+1)^{-1} E\left\{ \frac{K_{0h}^*(Y_{d-\alpha} - x_\alpha)\overline{\varphi}(\overline{Y}_d)}{\varphi(x_\alpha, \overline{Y}_d)} 2m(X_d)s(X_d) \right\} \left\{ \frac{K_{0h}^*(Y_{d-\alpha} - x_\alpha)\overline{\varphi}(\overline{Y}_d)}{\varphi(x_\alpha, \overline{Y}_d)} s(X_d) \right\}$$

which can be verified to be exactly $\frac{1}{nh} c_{V\alpha}(x) \{1 + o(1)\}$ by the same technique used above. Equation (2.6) is easy to prove as (2.2) of Theorem 1.

To get the asymptotic properties of $\widehat{c}_v$, we use the above results on $\widehat{V}_\alpha(x)$ and the mixing properties of our process to get

$$\widehat{c}_v^{d-1} - c_v^{d-1} = \frac{1}{d} \sum_{\alpha=1}^{d} \frac{1}{(n-d+1)} \sum_{j=d}^{n} \prod_{1 \le \beta \le d, \beta \ne \alpha} \widehat{V}_\beta(Y_{j-\beta}) - c_v^{d-1}$$

$$= \frac{1}{d} \sum_{\alpha=1}^{d} \frac{1}{(n-d+1)} \sum_{j=d}^{n} \prod_{1 \le \beta \le d, \beta \ne \alpha} \left\{ V_\beta(Y_{j-\beta}) + \widehat{V}_\beta(Y_{j-\beta}) - V_\beta(Y_{j-\beta}) \right\} - c_v^{d-1}$$

$$= \frac{1}{d} \sum_{\alpha=1}^{d} \frac{1}{(n-d+1)} \sum_{j=d}^{n} \prod_{1 \le \beta \le d, \beta \ne \alpha} V_\beta(Y_{j-\beta}) - c_v^{d-1}$$

$$+ \frac{1}{d} \sum_{\alpha=1}^{d} \frac{1}{(n-d+1)} \sum_{j=d}^{n} \sum_{1 \le \beta \le d, \beta \ne \alpha} \left\{ \prod_{1 \le \gamma \le d, \gamma \ne \alpha, \beta} V_\gamma(Y_{j-\gamma}) \right\} \left\{ \widehat{V}_\beta(Y_{j-\beta}) - V_\beta(Y_{j-\beta}) \right\}$$

$$= \frac{1}{d} \sum_{\alpha=1}^{d} \frac{1}{(n-d+1)} \sum_{j=d}^{n} \sum_{1 \le \beta \le d, \beta \ne \alpha} \left\{ \prod_{1 \le \gamma \le d, \gamma \ne \alpha, \beta} V_\gamma(Y_{j-\gamma}) \right\} \left\{ \widehat{V}_\beta(Y_{j-\beta}) - V_\beta(Y_{j-\beta}) \right\} + O_p(\frac{1}{\sqrt{n}})$$

$$= S_1 + S_2 + S_3 + o_p(h^{p+1})$$

Härdle, W., Nielsen, J.P. And Yang, L. (1999) Nonparametric Autoregression With Multiplicative Volatility And Additive Mean

where

$$S_1 = \frac{1}{d}\sum_{\alpha=1}^{d}\frac{1}{(n-d+1)}\sum_{j=d}^{n}\sum_{1\le\beta\le d,\beta\neq\alpha}\left\{\prod_{1\le\gamma\le d,\gamma\neq\alpha,\beta}V_\gamma(Y_{j-\gamma})\right\}b_{v\beta}(Y_{j-\beta})h^{p+1}$$

$$S_2 = \frac{1}{d}\sum_{\alpha=1}^{d}\frac{1}{(n-d+1)^2}\sum_{j=d}^{n}\sum_{1\le\beta\le d,\beta\neq\alpha}\left\{\prod_{1\le\gamma\le d,\gamma\neq\alpha,\beta}V_\gamma(Y_{j-\gamma})\right\}$$
$$\times\left[\sum_{k=d}^{n}\frac{K^*_{0h}(Y_{k-\beta}-Y_{j-\beta})}{\varphi(Y_{j-\beta},\overline{Y}_k)}\overline{\varphi}(\overline{Y}_k)\left\{2m(X_k)s(X_k)\xi_k\right\}\right]$$

$$S_3 = \frac{1}{d}\sum_{\alpha=1}^{d}\frac{1}{(n-d+1)^2}\sum_{j=d}^{n}\sum_{1\le\beta\le d,\beta\neq\alpha}\left\{\prod_{1\le\gamma\le d,\gamma\neq\alpha,\beta}V_\gamma(Y_{j-\gamma})\right\}$$
$$\times\left[\sum_{k=d}^{n}\frac{K^*_{0h}(Y_{k-\beta}-Y_{j-\beta})}{\varphi(Y_{j-\beta},\overline{Y}_k)}\overline{\varphi}(\overline{Y}_k)\left\{v(X_k)(\xi_k^2-1)\right\}\right]$$

These three terms can be written as (again using the mixing properties)

$$S_1 = \frac{h^{p+1}}{d}\sum_{\alpha=1}^{d}\int\sum_{1\le\beta\le d,\beta\neq\alpha}\left\{\prod_{1\le\gamma\le d,\gamma\neq\alpha,\beta}V_\gamma(y_\gamma)\right\}b_{v\beta}(y_\beta)\varphi(y)dy + O_p(\frac{1}{\sqrt{n}})$$

and

$$S_2 = \sum_{k=d}^{n}\frac{2m(X_k)s(X_k)\xi_k}{(n-d+1)}\frac{1}{d}\sum_{\alpha=1}^{d}\sum_{1\le\beta\le d,\beta\neq\alpha}\int\left\{\prod_{1\le\gamma\le d,\gamma\neq\alpha,\beta}V_\gamma(y_\gamma)\right\}$$
$$\times\frac{K^*_{0h}(Y_{k-\beta}-y_\beta)}{\varphi(y_\beta,\overline{Y}_k)}\overline{\varphi}(\overline{Y}_k)\varphi(y)dy\left\{1+o_p(1)\right\}$$
$$= \sum_{k=d}^{n}\frac{2m(X_k)s(X_k)\xi_k}{(n-d+1)}\frac{1}{d}\sum_{\alpha=1}^{d}\sum_{1\le\beta\le d,\beta\neq\alpha}\int\left\{\prod_{1\le\gamma\le d,\gamma\neq\alpha,\beta}V_\gamma(y_\gamma)\right\}$$
$$\times\frac{K^*_0(u)\overline{\varphi}(\overline{Y}_k)\varphi(Y_{k-\beta}-hu,\overline{y})dud\overline{y}}{\varphi(Y_{k-\beta}-hu,\overline{Y}_k)}\left\{1+o_p(1)\right\}$$

$$= \sum_{k=d}^{n}\frac{2m(X_k)s(X_k)\xi_k}{(n-d+1)\varphi(Y_k)}\frac{1}{d}\sum_{\alpha=1}^{d}\sum_{1\le\beta\le d,\beta\neq\alpha}\int\left\{\prod_{1\le\gamma\le d,\gamma\neq\alpha,\beta}V_\gamma(y_\gamma)\right\}$$
$$\times\overline{\varphi}(\overline{Y}_k)\varphi(Y_{k-\beta},\overline{y})d\overline{y}\left\{1+o_p(1)\right\}$$

from which it is clear that S_2 satisfies a central limit theorem with $\sqrt{n}$ rate of convergence, which is also the case with S_3. Thus

$$\widehat{c}_v = \left[c_v^{d-1} + \frac{h^{p+1}}{d}\sum_{\alpha=1}^{d}\int\sum_{1\le\beta\le d,\beta\neq\alpha}\left\{\prod_{1\le\gamma\le d,\gamma\neq\alpha,\beta}V_\gamma(y_\gamma)\right\}b_{v\beta}(y_\beta)\varphi(y)dy + \frac{1}{\sqrt{n}}Z\right]^{\frac{1}{d-1}}$$

where $Z \xrightarrow{D} N(0, \sigma^2)$ for some σ^2, applying Taylor expansion gives the result on $\widehat{c}_v$. the rest of Theorem 2 follows directly. Q. E. D.

Proof of Theorem 3. Simply putting together the results of the previous two theorems. Note that the joint normality follows from the fact that the stochastic part of all the estimates are based on the ξ_j's and the $(\xi_j^2 - 1)$'s. Thus, any linear combinations of the estimates also have similar forms as the ones treated in Theorem 1. Q. E. D.

References

Ango Nze P. (1992) *Critères d'ergodicité de quelques modèles à représentation markovienne*, C.R. Acad. Sci. Paris, sér. I, **315**, 1301-1304.

Auestad, B.; Tjøstheim, D. (1991) Functional identification in nonlinear time series. In Nonparametric Functional Estimation and Related Topics, Ed. Roussas, G.G Amsterdam: Kluwer Academic Publishers, 493-507.

Bossaerts, P.; Härdle, W.; Hafner, C. (1996) *Foreign exchange-rates have surprising volatility*, In: Athens conference on applied probability and time series, **2**, ed. P. M. Robinson, Lecture Notes in Statistics, 115, 55-72, Springer Verlag.

Chan, K.S.; Tong, H. (1986) *On estimating thresholds in autoregressive models*, Journal of Time Series Analysis, **7**, 179-190.

Chen, R.; Tsay, R.S. (1993a) *Nonlinear additive ARX models*, Journal of the American Statistical Association, **88**, 955-967.

Chen, R.; Tsay, R.S. (1993b) *Functional-coefficient autoregressive models*, Journal of the American Statistical Association, **88**, 298-308.

Davydov, Yu.A. (1973) *Mixing conditions for Markov chains*, Theory of Probability and its Applications, **18**, 312-328.

Diebolt, J.; Guégan, D. (1993) *Tail behaviour of the stationary density of general nonlinear autoregressive processes of order one*, Journal of Applied Probability, **30**, 315-329.

Drost, F. C.; Nijman, T. E. (1993) *Temporal aggregation of GARCH processes*, Econometrica, **61**, 909–927.

Engle, R.F. (1982) *Autoregressive conditional heteroscedasticity with estimates of the variance of U.K. inflation*, Econometrica, **50**, 987-1008.

Engle, R.F.; Gonzalez-Rivera, G. (1991) *Semiparametric ARCH models*, Journal of Business and Economic Statistics, **9**, 345-360.

Engle, R.F.; Ng, V. (1993) *Measuring and testing the impact of news on volatility*, Journal of Finance, **48**, 1749-1778.

Fan, J.; Gijbels, I. (1996) *Local polynomial modelling and its applications*, Chapman and Hall.

Gouriéroux, Ch.; Monfort, A. (1992) *Qualitative threshold ARCH models*, Journal of Econometrics, **52**, 159-199.

Granger, C.; Teräsvirta, T. (1993) *Modelling nonlinear dynamic relationships*, Oxford University Press, Oxford.

Haggan, V.; Ozaki, T. (1981) *Modelling nonlinear vibrations using an amplitude-dependent autoregressive time series model*, Biometrika, **68**, 189–196.

Härdle, W.; Chen, R. (1995) *Nonparametric time series analysis, a selective review with examples*, Proceedings of the 50th session of the ISI, Peking.

Härdle, W.; Klinke, S.; Turlach, B. (1995) *XploRe - an interactive statistical computing environment*, Springer Verlag, Heidelberg.

Härdle, W.; Tsybakov, A.B. (1997) *Local polynomial estimators of the volatility function in nonparametric autoregression*, Journal of Econometrics, **81**, 223-242.

Härdle, W.; Tsybakov, A.B.; Yang, L. (1998) *Nonparametric vector autoregression*, Journal of Statistical Planning and Inference *to appear.*

Hastie, T. J.; Tibshirani, R. J. (1990) *Generalized additive models*, Monographs on Statistics and Applied Probability, 43, Chapman and Hall, London.

Katkovnik, V.Ya. (1979) *Linear and nonlinear methods of nonparametric regression analysis*, Automatika, 35-46.

Linton, O. B.; Härdle, W. (1996) *Estimation of additive regression models with known links*, Biometrika, **83**, 529-540.

Linton, O.; Nielsen, J.P. (1995) *A kernel method of estimating structured nonparametric regression based on marginal integration*, Biometrika, **82**, 93-100.

Liptser, R.Sh.; Shirjaev, A.N. (1980) *A functional central limit theorem for martingales*, Theory of Probability and its Applications, **25**, 667 - 688.

Masry, E.; Tjøstheim, D. (1995a) *Non-parametric estimation and identification of ARCH nonlinear time series: Strong convergence and asymptotic normality*, Econometric Theory, **11**, 258-289.

Masry, E. ; Tjøstheim, D. (1995b) *Additive nonlinear ARX time series and projection estimates*, to appear in Econometric Theory.

Meese, R.A.; Rose, A. (1991) *An empirical assessment of non-linearities in models of exchange rate determination*, Review of Economic Studies, **58**, 601-619.

Mokkadem, A. (1987) *Sur un modèle autorégressif nonlinéaire. Ergodicité et ergodicité géometrique*, Journal of Time Series Analysis, **8**, 195-204.

Nummelin, E.; Tuominen, P. (1982) *Geometric ergodicity of Harris–recurrent Markov chains with application to renewal theory*, Stochastic Processes and their Applications, **12**, 187-202.

Nelson, D.B. (1991) *Conditional heteroscedasticity in asset returns: A new approach*, Econometrica, **59**, 347-370.

Robinson, P.M. (1983) *Nonparametric estimators for time series*, Journal of Time Series Analysis, **4**, 185-207.

Robinson, P.M. (1984) *Robust nonparametric autoregression*, In: *Robust and nonlinear time series analysis*, eds. Franke, Härdle and Martin, Lecture Notes in Statistics, 26, Springer-Verlag, Heidelberg.

Ruppert, D.; Wand, M.P. (1994) *Multivariate locally weighted least squares regression*, Annals of Statistics, **22**, 1346-1370.

Severance-Lossin, E.; Sperlich, S. (1995) *Estimation of derivatives for additive separable models*, SFB 373 Discussion Paper 60, Humboldt Universität zu Berlin, available at http://www.wiwi.hu-berlin.de/pub/papers/sfb/dpsfb960060.ps.Z.

Stone, C.J. (1977) *Consistent nonparametric regression*, Annals of Statistics, **5**, 595 - 645.

Stone, C.J. (1982) *Optimal global rates of convergence for nonparametric regression*, Annals of Statistics, **10**, 1040-1053.

Tjøstheim, D. (1990) *Nonlinear time series and Markov chains*, Advances in Applied Probability, **22**, 587-611.

Tjøstheim, D.; Auestad, B. (1994) *Nonparametric identification of nonlinear time series: Projections*, Journal of the American Statistical Association, **89**, 1398-1409.

Tong, H. (1978) *On a threshold model*, in C. H. Chen (ed.), *Pattern recognition and signal processing*, Sijthoff and Noordholf, The Netherlands.

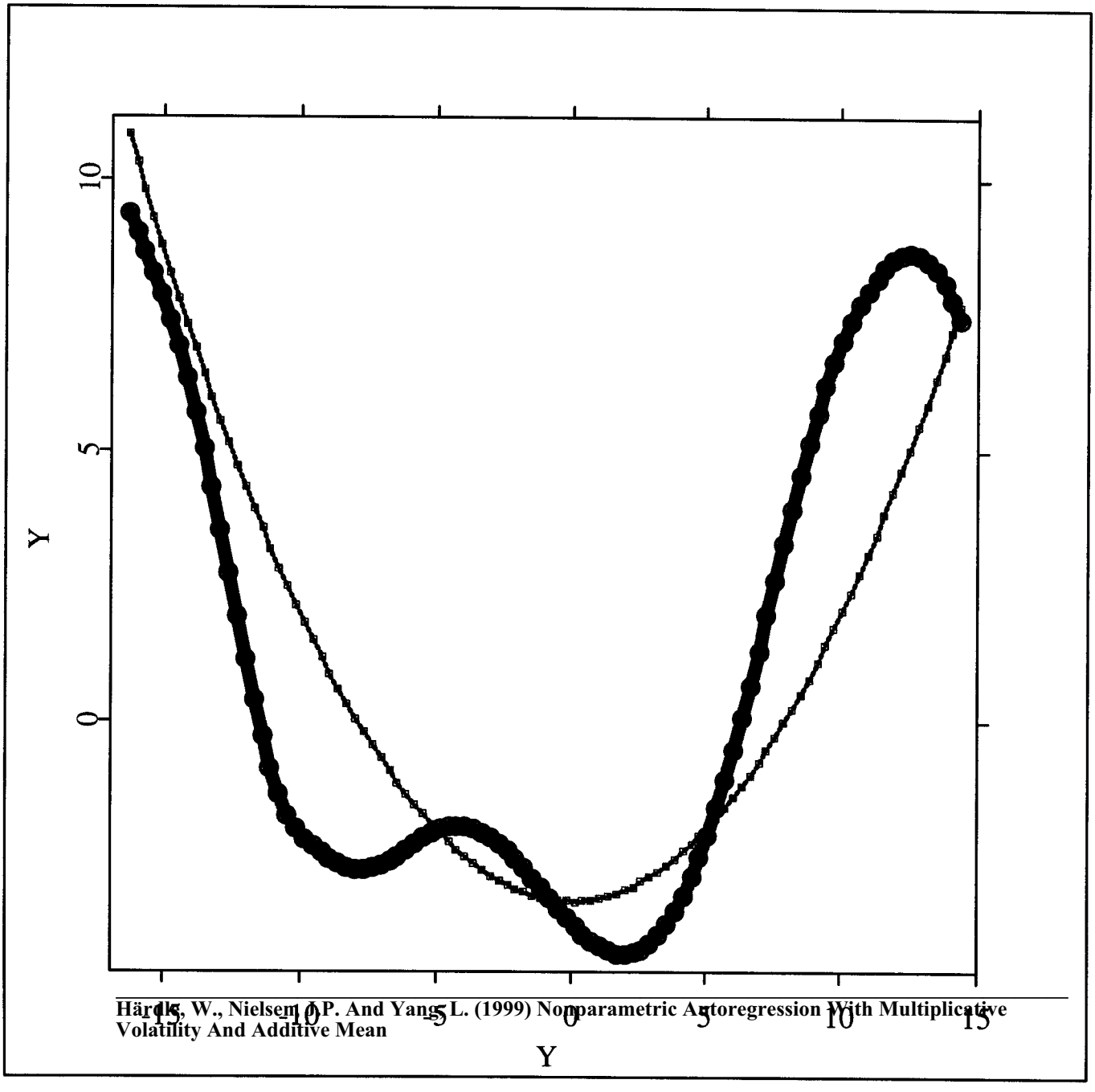
Tong, H. (1983) *Threshold models in nonlinear time series analysis*, Lecture Notes in Statistics, 21, Springer-Verlag, Heidelberg.

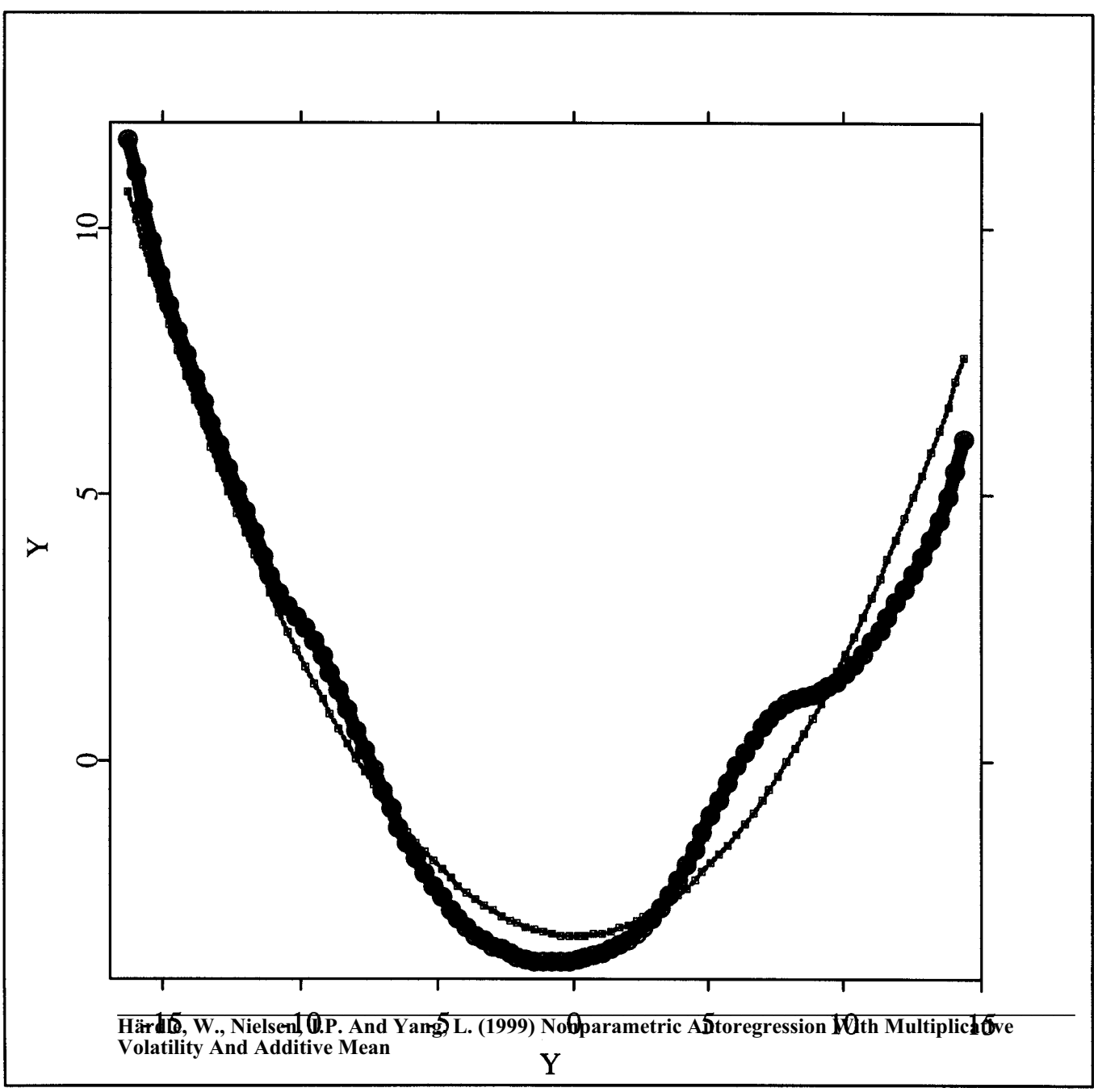
Tschernig, R.; Yang, L. (1997) *Nonparametric lag selection for time series*, SFB 373 Discussion Paper 59, Humboldt Universität zu Berlin.

Tsybakov, A.B. (1986) *Robust reconstruction of functions by the local–approximation method*, Problems of Information Transmission, **22**, 133-146.

Tweedie, R.L. (1975) *Sufficient conditions for ergodicity and recurrence of Markov chains on a general state space*, Stochastic Processes and their Applications, **3**, 385-403.

Wand, M.P.; Jones, M.C. (1995) *Kernel smoothing*, Chapman and Hall, London.


10
5
0
Y
−15
−10
−5
0
5
10
15
Y


Y
10
5
0
−15
−10
−5
0
5
10
15
Y

Sociedad De Estadistica E Investigacion Operativa Test (1999)
Vol. 8, No 2, pp. 419-458

Sociedad de Estadística e Investigación Operativa
Test (1999) *Vol.* **8**, *No.* 2, pp. 419–458

Integration and backfitting methods in additive models – finite sample properties and comparison

Stefan Sperlich*
Departamento de Estadística y Econometría
Universidad Carlos III de Madrid, Spain.
Oliver B. Linton
Department of Economics
Yale University, USA.
Wolfgang Härdle
Institut für Statistik und Ökonometrie
Humboldt-Universität zu Berlin, Germany.

Abstract

We examine and compare the finite sample performance of the competing backfitting and integration methods for estimating additive nonparametric regression using simulated data. Although, the asymptotic properties of the integration estimator, and to some extent the backfitting method too, are well understood, its small sample properties are not well investigated. Apart from some small experiments in the above cited papers, there is little hard evidence concerning the exact distribution of the estimates. It is our purpose to provide an extensive finite sample comparison between the backfitting procedure and the integration procedure using simulated data.

Key Words: Additive models, curse of dimensionality, dimensionality reduction, model choice, nonparametric regression.
AMS subject classification: 62G07, 62G20, 62G35

1 Introduction

Additive models are widely used in both theoretical economics and in econometric data analysis. The standard text of Deaton and Muellbauer (1980) provides many examples in microeconomics for which the additive structure provides interpretability and allows solution of choice problems. Additive

*Correspondence to: Stefan Sperlich, Departamento de Estadística y Econometría, Universidad Carlos III de Madrid, 28903 Getafe-Madrid, Spain.

The research was supported by the National Science Foundation, NATO, and Deutsche Forschungsgemeinschaft, SFB 373.

Received: July 1997; Accepted: January 1999

Sociedad De Estadistica E Investigacion Operativa Test (1999)
Vol. 8, No 2, pp. 419-458

structure is desirable from a purely statistical point of view because it circumvents the curse of dimensionality. There has been much theoretical and applied work in econometrics on semiparametric and nonparametric methods, see Härdle and Linton (1994), Newey (1990), and Powell (1994) for bibliography and discussion. Some recent work has shown that additivity has important implications for the rate at which certain components can be estimated. In this paper we consider the finite sample performance of two popular estimators for additive models: the backfitting estimators of Hastie and Tibshirani (1990) and the integration estimators of Linton and Nielsen (1995).

Let (X, Y) be a random variable with X of dimension d and Y a scalar. Consider the estimation of the regression function $m(x) = E(Y \mid X = x)$ based on a random sample $\{(X_i, Y_i)\}_{i=1}^n$ from this population. Stone (1980, 1982) and Ibragimov and Hasminskii (1980) showed that the optimal rate for estimating m is $n^{-\ell/(2\ell+d)}$ with ℓ an index of smoothness of m. An additive structure for m is a regression function of the form

$$m(x) = c + \sum_{\alpha=1}^{d} m_\alpha(x_\alpha), \tag{1.1}$$

where $x = (x_1, \ldots, x_d)^T$ are the d-dimensional predictor variables and m_α are one-dimensional nonparametric functions operating on each element of the vector or predictor variables with $E\{m_\alpha(X_\alpha)\} = 0$. Stone (1985, 1986) showed that for such regression curves the optimal rate for estimating m is the one-dimensional rate of convergence with $n^{-\ell/(2\ell+1)}$. Thus one speaks of dimensionality reduction through additive modelling.

In practice, the backfitting procedures proposed in Breiman and Friedman (1985) and Buja, Hastie and Tibshirani (1989) are widely used to estimate the additive components. The latter (equation (18)) consider the problem of finding the projection of m onto the space of additive functions representing the right hand side of (1). Replacing population by sample, this leads to a system of normal equations with $nd \times nd$ dimensions. To solve this in practice, the backfitting or Gauss-Seidel algorithm, is usually used, see Venables and Ripley (1994). This technique is iterative and depends on the starting values and convergence criterion. It converges very fast but has, in comparison with the direct solution of the large linear system, the slight disadvantage of a more complicated "hat matrix", see Härdle and Hall (1993). These methods have been evaluated on numerous datasets

Sociedad De Estadistica E Investigacion Operativa Test (1999)
Vol. 8, No 2, pp. 419-458

and have been refined quite considerably since their introduction.

Recently, Linton and Nielsen (1995), Tjøstheim and Auestad (1994), and Newey (1994) have independently proposed an alternative procedure for estimating m_α based on integration of a standard kernel estimator. It exploits the following idea. Suppose that $m(x,z)$ is any bivariate function, and consider the quantities $\mu_1(x) = \int m(x,z)dQ_n(z)$ and $\mu_2(z) = \int m(x,z)dQ_n(x)$, where Q_n is a probability measure. If $m(x,z) = m_1(x) + m_2(z)$, then $\mu_1(\cdot)$ and $\mu_2(\cdot)$ are $m_1(\cdot)$ and $m_2(\cdot)$, respectively, up to a constant. In practice one replaces m by an estimate and integrates with respect to some known measure. The procedure is explicitly defined and its asymptotic distribution is easily derived: it converges at the one-dimensional rate and satisfies a central limit theorem. This estimation procedure has been extended to a number of other contexts like estimating the derivatives (Severance-Lossin and Sperlich, 1997), to the generalized additive model (Linton and Härdle, 1996), to dependent variable transformation models (Linton, Chen, Wang, and Härdle, 1997), to econometric time series models (Masry and Tjøstheim, 1995, 1997), to panel data models (Porter, 1996), and to hazard models with time varying covariates and right censoring (Nielsen, 1996). In this wide variety of sampling schemes and procedures the asymptotics have been derived because of the explicit form of the estimator. By contrast, backfitting or backfitting-like methods have until recently eluded theoretical analysis, until Opsomer and Ruppert (1997) provided conditional mean squared error expressions albeit under rather strong conditions on the smoothing matrices and design. More recently, Linton, Mammen, and Nielsen (1998) has established a central limit theorem for a modified form of backfitting which uses a bivariate integration step as well as the iterative updating of the other methods.

The purpose of this paper is to investigate the finite sample performance of the standard backfitting estimator and the integration estimator.

2 Methods and Theory

We suppose that

$$Y_i = m(X_i) + \varepsilon_i, \quad i = 1, \ldots, n,$$

where by definition $E(\varepsilon_i|X_i) = 0$; let also $\mathrm{Var}(\varepsilon_i|X_i) = \sigma^2(X_i)$ be the conditional variance function. We denote the marginal density of the

Sociedad De Estadistica E Investigacion Operativa Test (1999)
Vol. 8, No 2, pp. 419-458

d-dimensional explanatory variable by $p(x)$ with marginals $p_\alpha(x_\alpha)$, $\alpha = 1, \ldots, d$. We shall sometimes partition $X_i = (X_{\alpha i}, X_{\underline{\alpha} i})^T$ and $x = (x_\alpha, x_{\underline{\alpha}})^T$ into scalar and $d-1$-dimensional subvectors respectively calling x_α the direction of interest and $x_{\underline{\alpha}}$ the direction not of interest; denote by $p_{\underline{\alpha}}(x_{\underline{\alpha}})$ the marginal density of the vector $X_{\underline{\alpha} i}$. In the following we assume the following additive form for the regression function

$$m(x) = c + \sum_{\alpha=1}^{d} m_\alpha(x_\alpha), \quad x = (x_1, \ldots, x_d)^T, \quad c \text{ constant.}$$

A.1 Integration

A commonly used estimate of $m(x)$ is provided by the multidimensional local polynomial product kernel estimator which solves the following minimization problem

$$\begin{pmatrix} \widehat{\theta}_0 \\ \widehat{\theta}_1 \end{pmatrix} = \min_{\theta_0, \theta_1} \sum_{i=1}^{n} \{Y_i - P_q(\theta_0, \theta_1; X_i - x)\}^2 \prod_{\alpha=1}^{d} K_\alpha \left(\frac{x_\alpha - X_{\alpha i}}{h_\alpha} \right), \quad (2.1)$$

where K_α and h_α, $\alpha = 1, \ldots, d$, are scalar kernels and bandwidths respectively, while $P_q(\theta_0, \theta_1; t)$ is a $(q-1)^{th}$ order polynomial in the vector t with coefficients θ_0, θ_1 for which $P_q(\theta_0, \theta_1; 0) = \theta_0$ and e.g. $P_2(\theta_0, \theta_1; t) = \theta_0 + \theta_1 t$. Let $\widehat{m}(x) = \widehat{\theta}_0(x)$. Under regularity conditions, see Ruppert and Wand (1995) for example, the local polynomial estimator satisfies

$$\widehat{m}_h(x) - m(x) \xrightarrow{\mathcal{L}} N\left\{ h^q \mu_q(K) b(x), \frac{1}{nh^d} \nu_q(K) v(x) \right\}, \quad (2.2)$$

where $h = (\prod_{\alpha=1}^{d} h_\alpha)^{1/d}$ is the geometric average of the bandwidths, $\mu_q(K)$ and $\nu_q(K)$ are constants depending only on the kernels, while $v(x) = \sigma^2(x)/p(x)$ and $b(x)$ is the bias function depending on derivatives of m, and possibly p, up to and including order q. The (mean squared error) optimal bandwidth is of order $n^{-1/(2q+d)}$ for which the asymptotic mean squared error is of order $n^{-2q/(2q+d)}$, see Härdle and Linton (1994), which reflects the curse of dimensionality – as d increases, the rate of convergence decreases.

When $m(\cdot)$ satisfies the additive model structure, we can estimate $m(x)$

Sociedad De Estadistica E Investigacion Operativa Test (1999)
Vol. 8, No 2, pp. 419-458

with a better rate of convergence by imposing these restrictions. Let

$$\widetilde{m}_\alpha(x_\alpha) = \int \widehat{m}_{h_1}(x_\alpha, x_{\underline{\alpha}}) dQ_n(x_{\underline{\alpha}}) - \widehat{c}, \tag{2.3}$$

where $\widehat{c}$ is an estimate of c, while $Q_n(\cdot)$ is some easy to compute probability measure. The most convenient choice of $Q_n(\cdot)$ is the empirical measure of $\{X_{\underline{\alpha}i}\}_{i=1}^n$, which converges to the population distribution. It changes the integral in (2.3) to a sum over terms evaluated at $X_{\underline{\alpha}i}$ and implies for the constant $c = E(Y)$. The latter can be estimated root-n consistently by the sample mean $n^{-1}\sum_{i=1}^n Y_i$; an alternative estimate, which is not necessarily root- n consistent, is provided by $n^{-1}\sum_{i=1}^n \widetilde{m}_\alpha(X_{\alpha i})$. Whatever the estimates of c and $\widetilde{m}_\alpha(x_\alpha)$, we reestimate $m(x)$ by

$$\widetilde{m}_{h_1}(x) = \widehat{c} + \sum_{\alpha=1}^{d} \widetilde{m}_\alpha(x_\alpha). \tag{2.4}$$

Linton and Härdle (1996) derived the pointwise asymptotic properties of the empirical integration versions of $\widetilde{m}_\alpha(x_\alpha)$ and $\widetilde{m}_{h_1}(x)$. To simplify matters, we set $h_\alpha = h_1$ and $K_\alpha = K$, while $\prod_{\beta\neq\alpha} K_\beta = L$ and $h_\beta = h_2$ for all $\beta \neq \alpha$. Under their regularity conditions,

$$\widetilde{m}_{h_1}(x) - m(x) \xrightarrow{\mathcal{L}} N\left\{\frac{h_1^q}{q!}\overline{\mu}_q(K)b_0(x), \frac{1}{nh_1}\overline{\nu}_q(K)v_0(x)\right\}, \tag{2.5}$$

where $b_0(x) = \sum_\alpha b_{\alpha 0}(x_\alpha)$ and $v_0(x) = \sum_\alpha v_{\alpha 0}(x_\alpha)$ with $b_{\alpha 0}(x_\alpha) = \int b(x)p_{\underline{\alpha}}(x_{\underline{\alpha}})dx_{\underline{\alpha}}$ and $v_{\alpha 0}(x_\alpha) = \int v(x)p_{\underline{\alpha}}^2(x_{\underline{\alpha}})dx_{\underline{\alpha}}$. Here, $\overline{\mu}_q(K)$ and $\overline{\nu}_q(K)$ are constants depending only on the kernel K. By choosing $h_1 \propto n^{-1/(2q+1)}$ one can achieve the optimal rate of convergence i.e., mean squared error of order $n^{-2q/(2q+1)}$, which is independent of the dimensions d. See also Linton and Nielsen (1995) and Severance-Lossin and Sperlich (1997).

Remark 2.1. The bandwidths $h_1, \ldots, h_d$ should be chosen differently as we discuss further in the simulation section. To achieve the optimal rate of convergence, we must impose some restrictions on the bandwidth sequences. This condition, which corresponds to (A7) in Linton and Härdle (1996) is needed for bias reduction of the nuisance components. In Section 3 we will examine some bandwidth selection methods.

Sociedad De Estadistica E Investigacion Operativa Test (1999)
Vol. 8, No 2, pp. 419-458

A.2 Backfitting

Hastie and Tibshirani (1990) motivate the backfitting method as follows. First consider the analogous population problem:

$$\min_m E\{Y - m(X)\}^2 \quad \text{s.t.} \quad m(x) = \sum_{\alpha=1}^{d} m_\alpha(x_\alpha),$$

which can be formulated inside a Hilbert space framework: let $[\mathcal{H}_{YX}, \langle\cdot,\cdot\rangle]$ be the Hilbert space of random variables which are functions of Y and X with $\langle a, b\rangle = E(ab)$, let also $[\mathcal{H}_X, \langle\cdot,\cdot\rangle]$, and $[\mathcal{H}_{X_\alpha}, \langle\cdot,\cdot\rangle]$, $\alpha = 1,\ldots,d$ be corresponding subspaces, where for example $\mathcal{H}_{X_\alpha}$ contains only functions of X_α. The above problem is equivalent to finding the element of the subspace $\mathcal{H}_{X_1} \oplus \cdots \oplus \mathcal{H}_{X_d}$ closest to a point $Y \in \mathcal{H}_{YX}$ or equivalently the point $m \in \mathcal{H}_X$. By the projection theorem, there exists a unique solution which is characterized by the following first order conditions

$$E[\{Y - m(X)\}|X_\alpha] = 0 \Leftrightarrow m_\alpha(X_\alpha) = E\left[\left\{Y - \sum_{\gamma\neq\alpha} m_\gamma(X_\gamma)\right\}\middle| X_\alpha\right],$$

$\alpha = 1,\ldots,d$, which leads to the formal representation:

$$\begin{pmatrix} I & P_1 & \cdots & P_1 \\ P_2 & I & \cdots & P_2 \\ \vdots & & \ddots & \vdots \\ P_d & \cdots & P_d & I \end{pmatrix} \begin{pmatrix} m_1(X_1) \\ m_2(X_2) \\ \vdots \\ m_d(X_d) \end{pmatrix} = \begin{pmatrix} P_1 Y \\ P_2 Y \\ \vdots \\ P_d Y \end{pmatrix},$$

where $P_\alpha(\cdot) = E(\cdot|X_\alpha)$. By analogy, let S_α $(n \times n)$ be the smoother matrix which when applied to the $n \times 1$ vector $y = (Y_1,\ldots,Y_n)^T$ yields an $n \times 1$ vector estimate $S_\alpha y$ of the vector $\{E(Y_1|X_{\alpha 1}),\ldots,E(Y_n|X_{\alpha n})\}^T$. Substituting P_α by S_α we obtain the following

$$\begin{pmatrix} I & S_1 & \cdots & S_1 \\ S_2 & I & \cdots & S_2 \\ \vdots & & \ddots & \vdots \\ S_d & \cdots & S_d & I \end{pmatrix} \begin{pmatrix} \widehat{m}_1 \\ \widehat{m}_2 \\ \vdots \\ \widehat{m}_d \end{pmatrix} = \begin{pmatrix} S_1 y \\ S_2 y \\ \vdots \\ S_d y \end{pmatrix}.$$

This system can in principle be solved exactly for $\{\widehat{m}_\alpha(X_{\alpha 1}),\ldots,\widehat{m}_\alpha(X_{\alpha n})\}^T$ with $\alpha = 1,\ldots,d$. However, when nd is large the required matrix inversion

Sociedad De Estadistica E Investigacion Operativa Test (1999)
Vol. 8, No 2, pp. 419-458

is not feasible. Further, often the matrix on the left is not regular in practice and thus this equation cannot be solved directly. In practice, the backfitting (Gauss-Seidel) algorithm is used to solve these equations: given starting values $\check{m}_\alpha^{(0)}$, $\alpha = 1, \ldots, d$, update the $n \times 1$ vectors as follows

$$\check{m}_\alpha^{(r)} = S_\alpha \left\{ y - \sum_{\gamma \neq \alpha} \check{m}_\gamma^{(r-1)} \right\}, \quad r = 1, \ldots$$

until some prespecified tolerance is reached. The estimator is linear in y, but the algorithm only converges under strong restrictions on the smoother matrices. Recent work by Opsomer and Ruppert (1997) discuss some improvements to this algorithm which are guaranteed to provide a unique solution. They also derive the conditional mean squared error of the resulting estimator under strong conditions: this has a similar expression to (2.5) in large samples.

3 Simulation Results

A.1 Introduction

In a number of different additive models, we determined the bias, variance and mean squared error for both estimation procedures. We considered designs with distributions: the uniform $U[-3,3]^d$, the normal with mean 0, variance 1 and varying covariance $\rho = 0, 0.4, 0.8$, denoted as $N(\rho)$, for different numbers of observations and several dimensions. We drew all these designs once and kept them fixed for the investigation described in the following. The error term ε was always chosen as normal distributed with zero mean and variance $\sigma_\epsilon^2 = 0.5$. Since both estimators are linear, i.e.,

$$\widehat{m}_\alpha(x) = \sum_{i=1}^{n} w_{\alpha i}(x) Y_i$$

for some weights $\{w_{\alpha i}(x)\}$ we determined the conditional bias and variance as follows

$$\text{var}\{\widehat{m}_\alpha(x_\alpha)|X\} = \sigma_\epsilon^2 \sum_{i=1}^{n} w_{\alpha i}^2(x)$$

Sociedad De Estadistica E Investigacion Operativa Test (1999)
Vol. 8, No 2, pp. 419-458

$$\text{bias}\{\widehat{m}_\alpha(x_\alpha)|X\} = \sum_{i=1}^{n} w_{\alpha i}(x) m(X_i) - m_\alpha(x_\alpha)$$

for the additive function estimators and by analogy for the regression estimator. In the following notation the MSE denotes the mean squared error and the MASE the averaged MSE. We focused on the following questions:

a) What is a reasonable bandwidth choice for an optimal fit?

b) How sensitive are the estimators to the bandwidth?

c) What are the MASE, MSE, bias and variance, boundary effects?

d) We considered degrees of freedom, eigen analysis, singular values and eigen vectors.

e) We plotted the equivalent kernel weights of the estimates and

f) we investigated whether and when the asymptotics kick in.

We examined how well the estimation procedures performed in estimating one additive function. The parameters are $d = 2$ dimensions and $n = 100$ observations. We considered all combinations of the following additive functions for a two dimensional additive model:

$$m_1(x) = 2x; \qquad m_2(x) = x^2 - E(x^2)$$
$$m_3(x) = \exp(x) - E\{\exp(x)\}; \qquad m_4(x) = 0.5 \cdot \sin(-1.5x).$$

Our interest is mainly in the estimation of the marginal effect m_α. We first determined different optimal bandwidths for a given design distribution. In the second step we calculated for fixed designs bias, variance and mean average squared error (on the complete data set as well as on trimmed data) for both estimation procedures.

The advantages of using local polynomials are well known, especially with regard to the robustness against choice of bandwidth and the improvement in bias and consequently mean squared error if the requisite smoothness is present. In Severance-Lossin and Sperlich (1997) the consistency and asymptotic behavior of the integration estimator using local polynomial is shown. For these reasons we did the investigation for both, the Nadaraya Watson and the local linear estimator.

Sociedad De Estadistica E Investigacion Operativa Test (1999)
Vol. 8, No 2, pp. 419-458

A.2 Bandwidth Choice

The choice of an appropriate smoothing parameter is always a critical point in nonparametric and semiparametric estimation. For the integration estimator we need even two bandwidths, h_1 and h_2, see Section 2. There exist at least two rules for choosing them: the rule of thumb of Linton and Nielsen (1995) and the plug-in method suggested in Severance-Lossin and Sperlich (1997). Both methods give the MASE minimizing bandwidth, the first one approximately with the aid of parametric pre-estimators, the second one by using nonparametric pre-estimators. We give here the formulas for the case of local linear smoothers. The rule of thumb is

$$h_1 = \left\{ \frac{\tilde{\sigma}^2 \nu(K)(max - min)}{\mu_2(K)\left(\sum_{j=1}^{d} \hat{\beta}_j\right)^2} \right\}^{1/5} n^{-1/5},$$

where $\nu(K) = \|K\|_2^2$, $\mu_2(K) = \int t^2 K(t) dt$ and max and min are the sample maximum and minimum of the direction of interest. We obtained $\tilde{\beta}_j$ as the coefficients of $x_j^2/2$ from a least squares regression of Y on a constant, x_j, $x_j^2/2$ and $x_j x_k$ for all j, $k = 1, \ldots, d$, $j < k$, while $\tilde{\sigma}^2$ was obtained from the residuals of this regression by taking the average of the squares.

The formula for the nonparametric plug-in method we used for calculating the asymptotically optimal bandwidth is

$$h_1 = \left\{ \frac{\nu(K) \int \sigma^2 \frac{p_{\underline{\alpha}}^2(x_{\underline{\alpha}}) p_\alpha(x_\alpha)}{p(x_\alpha, x_{\underline{\alpha}})} dx_{\underline{\alpha}} dx_\alpha}{4\{\frac{1}{2}\mu_2(K)\}^2 \int \{m_\alpha^{(2)}(x_\alpha)\}^2 p_\alpha(x_\alpha) dx_\alpha} \right\}^{1/5} n^{-1/5}.$$

Note that this formula is not valid for h_2, the bandwidth for the direction not of interest. We took the bandwidth h_2 that minimized the MASE in the particular finite sample model.

For a fair comparison of the optimal bandwidth and the corresponding MASE of both estimators we applied several procedures. We started with considering the minimal MASE of the overall regression function and the minimizing bandwidths. Then we looked for the bandwidths minimizing the MASE in each direction separately. For taking into account the influence of boundary effects we looked also for the optimal bandwidths on trimmed data.

For small samples of 100 observations we could not discover any information by comparing the numerically MASE-minimizing bandwidths.

Sociedad De Estadistica E Investigacion Operativa Test (1999)
Vol. 8, No 2, pp. 419-458

They differed a lot depending on the particularly drawn design. Therefore we focused on once drawn, in that sense fixed, designs for the whole paper and considered only analytically determined bandwidths h_1. Thus we compared the results for bandwidths calculated with the rule of thumb proposed by Linton and Nielsen and the analytically optimal one.

Selected numerical Results, using both, the Nadaraya Watson and the local linear Smoother. Since the values of the MASE minimizing bandwidths that we found numerically for the particular designs in finite samples, were not particularly illuminating, we do not report them in the tables. In Table 1 the bandwidths of the rule of thumb by Linton and Nielsen and the asymptotically optimal bandwidths for each estimation procedure are shown. Here we concentrated on bandwidths that minimize the MASE in each direction separately. They are displayed for the additive components m_3, m_4 versus the particular model and design. The behavior for m_1, m_2 is the same, the results can be requested from the authors. One can see very well the strong influence of the distribution and the dependence of the additive function that has to be estimated. Furthermore, not only do the bandwidths determined by theory based rules differ a lot, we found them quite often far away from the MASE minimizing bandwidth value. This is also the case for the local linear smoothers. Mostly, the analytically chosen bandwidth was closer to the MASE minimizing one than the rule of thumb bandwidth, which, however, is much easier to calculate.

If the optimal value was infinity, we set it to 1 or in the case of a $N(0.8)$ distributed design to 2. In formulas where we had to integrate over a density from $-\infty$ to $+\infty$ we did this [for numerical reasons] over the interval $[-1.5, 1.5]$ for $N(0.8)$ and over $[-3, 3]$ else.

Table 2 gives the optimal bandwidths for different distributions, models, estimation routines and criteria when using local linear smoothing.

All findings from Table 1 are replicated here. Furthermore, note that for uncorrelated regressors the bandwidths are almost the same for backfitting and integration method, which is in accordance with the theoretically similar MASE. As mentioned above we will now consider the choice of bandwidth for the local linear estimation procedure in a more detailed way.

Sociedad De Estadistica E Investigacion Operativa Test (1999)
Vol. 8, No 2, pp. 419-458

Addit. Func.:	$\hat{m}_3$				$\hat{m}_4$			
Distribution:	U^2	$N(.0)$	$N(.4)$	$N(.8)$	U^2	$N(.0)$	$N(.4)$	$N(.8)$
Model		$m = m_1 + m_3$				$m = m_1 + m_4$		
rule of thumb	0.222	0.246	0.243	0.242	0.308	0.294	0.316	0.376
backfitting	0.191	0.175	0.175	0.260	0.426	0.310	0.310	0.282
integration h_1	0.191	0.175	0.194	0.307	0.426	0.309	0.352	0.387
Model		$m = m_2 + m_3$				$m = m_2 + m_4$		
rule of thumb	0.194	0.210	0.209	0.209	0.279	0.251	0.260	0.276
backfitting	0.191	0.175	0.175	0.260	0.426	0.310	0.310	0.282
integration h_1	0.191	0.175	0.194	0.307	0.426	0.309	0.352	0.387
Model		$m = m_3 + m_4$				$m = m_3 + m_4$		
rule of thumb	0.185	0.230	0.234	0.243	0.185	0.230	0.234	0.243
backfitting	0.191	0.175	0.175	0.260	0.426	0.310	0.310	0.282
integration h_1	0.191	0.175	0.194	0.307	0.426	0.309	0.352	0.387

Table 1: Asymptotically optimal bandwidths when using Nadaraya Watson smoother.

A.3 Robustness with respect to the Choice of Bandwidth

To find out how sensitive the estimators are with respect to the choice of bandwidth for the direction of interest h_1 we plotted MASE and $\text{MSE}_{x=0}$ against bandwidth for the two models $m = m_2 + m_3$ and $m = m_2 + m_4$. The parameters were kept unchanged or were mentioned in the caption of the respective figures. We present our results first for the uniform design on $[-3, 3]^2$, then for designs with distribution $N(0.0)$ and $N(0.4)$, see Figures 1 to 6.

The results for MASE have been trimmed in the pictures, since otherwise they would have been dominated by boundary effects (compare with tables in the next section). The results for the integration estimator are drawn throughout the paper as solid lines, those for the backfitting algorithm as dashed lines.

Obviously, the backfitting estimator is very sensitive to the choice of bandwidth. To get a small MASE it is crucially important for the backfitting method to choose a good smoothing parameter. For correlated designs oversmoothing seems slightly preferable, otherwise there is no particular advantage to either oversmoothing or undersmoothing. The behavior of the estimates for the highly correlated design is slightly strange and hard to interpret. This is true for both estimation procedures. Therefore we

Sociedad De Estadistica E Investigacion Operativa Test (1999)
Vol. 8, No 2, pp. 419-458

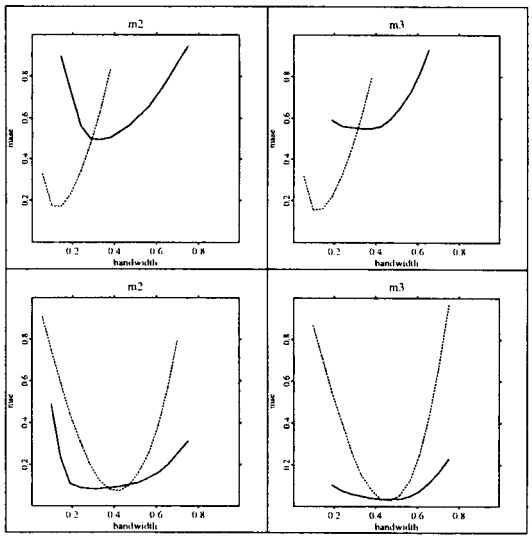


Figure 1: *Performance by bandwidth h_1 of MASE (top) and $MSE_{x=0}$ (bottom) in model $m = m_2 + m_3$, separately for m_2 (left), m_3 (right). Design is $X \sim U[-3,3]^2$.*

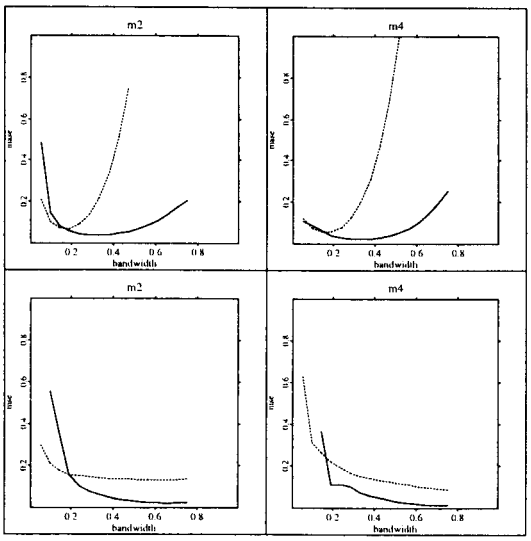


Figure 2: *Performance by bandwidth h_1 of MASE (top) and $MSE_{x=0}$ (bottom) in model $= m_2 + m_4$, separately for m_2 (left), m_4 (right). Design is $X \sim U[-3,3]^2$.*

Sociedad De Estadistica E Investigacion Operativa Test (1999)
Vol. 8, No 2, pp. 419-458

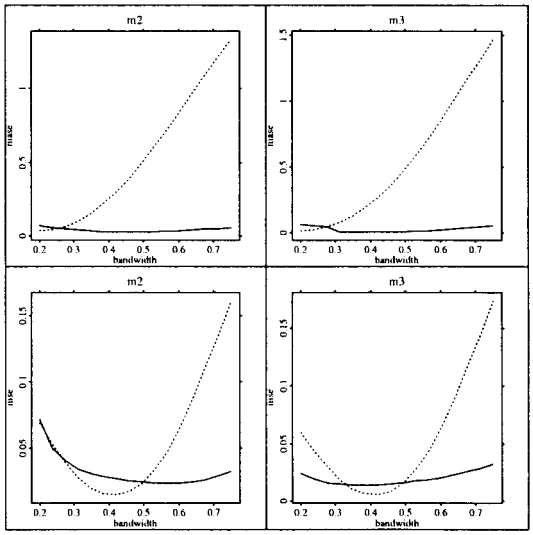


Figure 3: *Performance by bandwidth h_1 of MASE (top) and $MSE_{x=0}$ (bottom) in model $= m_2 + m_3$, separately for m_2 (left), m_3 (right). Design is $X \sim N(0.0)$.*

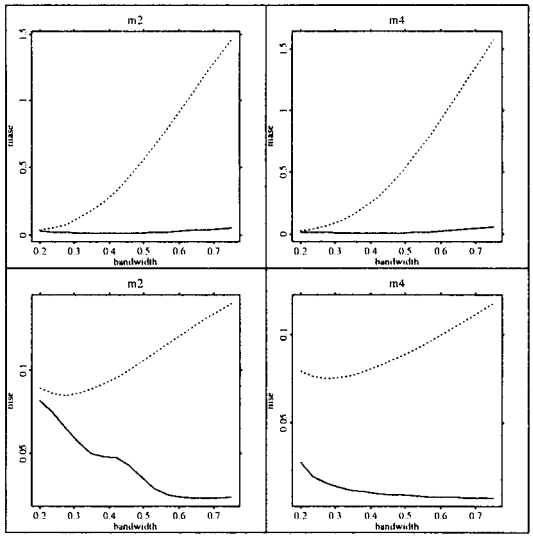


Figure 4: *Performance by bandwidth h_1 of MASE (top) and $MSE_{x=0}$ (bottom) in model $= m_2 + m_4$, separately for m_2 (left), m_4 (right). Design is $X \sim N(0.0)$.*

Sociedad De Estadistica E Investigacion Operativa Test (1999)
Vol. 8, No 2, pp. 419-458

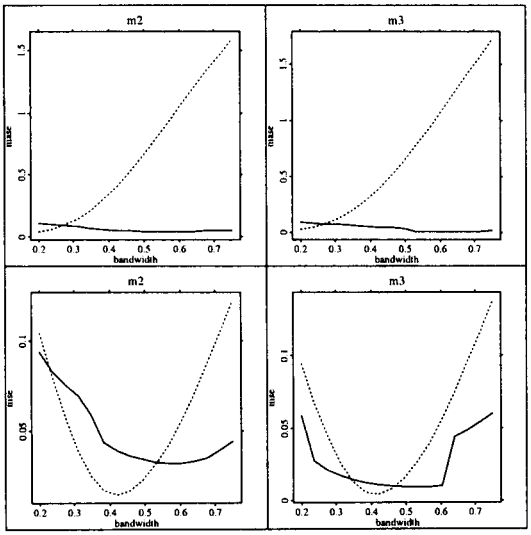


Figure 5: *Performance by bandwidth h_1 of MASE (top) and $MSE_{x=0}$ (bottom) in model $= m_2 + m_3$, separately for m_2 (left), m_3 (right). Design is $X \sim N(0.4)$.*

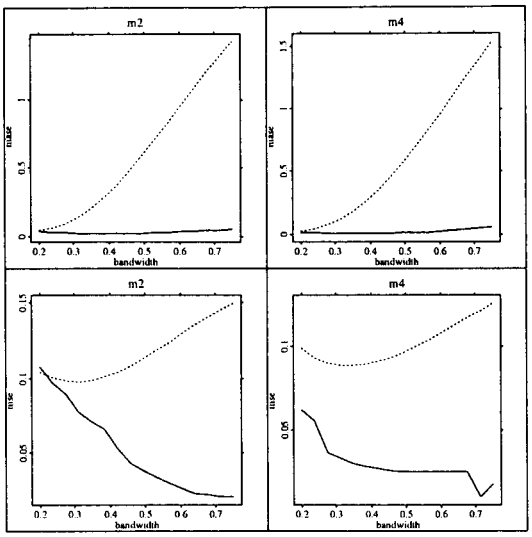


Figure 6: *Performance by bandwidth h_1 of MASE (top) and $MSE_{x=0}$ (bottom) in model $= m_2 + m_4$, separately for m_2 (left), m_4 (right). Design is $X \sim N(0.4)$.*

Sociedad De Estadistica E Investigacion Operativa Test (1999)
Vol. 8, No 2, pp. 419-458

Addit. Func.:	$\hat{m}_3$				$\hat{m}_4$			
Distribution:	U^2	$N(.0)$	$N(.4)$	$N(.8)$	U^2	$N(.0)$	$N(.4)$	$N(.8)$
Model	$m = m_1 + m_3$				$m = m_1 + m_4$			
rule of thumb	0.222	0.246	0.243	0.242	0.308	0.294	0.316	0.376
backfitting	0.191	0.267	0.267	0.284	0.426	0.423	0.423	0.374
integration h_1	0.191	0.267	0.324	0.571	0.426	0.423	0.513	0.752
Model	$m = m_2 + m_3$				$m = m_2 + m_4$			
rule of thumb	0.194	0.210	0.209	0.209	0.279	0.251	0.260	0.276
backfitting	0.191	0.267	0.267	0.284	0.426	0.423	0.423	0.374
integration h_1	0.191	0.267	0.324	0.571	0.426	0.423	0.513	0.752
Model	$m = m_3 + m_4$				$m = m_3 + m_4$			
rule of thumb	0.185	0.230	0.234	0.243	0.185	0.230	0.234	0.243
backfitting	0.191	0.267	0.267	0.284	0.426	0.423	0.423	0.374
integration h_1	0.191	0.267	0.324	0.571	0.426	0.423	0.513	0.752

Table 2: Asymptotically optimal bandwidths when using local linear smoother.

skipped the figures for the $N(0.8)$ distributed design.

For the integration estimator the results differ depending on the model. In general this method is by far not as sensitive to the choice of bandwidth as the backfitting procedure is. If we focus on the $\text{MSE}_{x=0}$ we have similar results as for the MASE but weakened concerning the sensitivity. Here the results differ more depending on the data generating model.

Since in a $[-3,3]^2$ rectangle $n = 100$ observations are fairly sparse and thus the behavior of the MASE or $\text{MSE}_{x=0}$ perhaps is not typical, we did the same investigation with $n = 100$ observations for the uniform design on $[-1.5,1.5]^2$. But, plotting the MASE and $\text{MSE}_{x=0}$ functions on the same scale as we did for the $U[-3,3]^2$ design, we detected that the general sensitivity is similar but certainly on a different range. Furthermore, the integration estimate improved a lot since it has been suffering more when data were sparse as e.g., in $[-3,3]^2$. All in all, our observations above are confirmed when data were not too sparse.

A.4 Simulation Results: Bias, Variance and MASE

Due to the excess of information we got out of doing our simulations, we concentrate on the results for the local linear case. Nevertheless we think it worthwhile to mention both, and, if there are differences in the results,

Sociedad De Estadistica E Investigacion Operativa Test (1999)
Vol. 8, No 2, pp. 419-458

to discuss them.

For the optimal bandwidths computed in Section 3.2 and given fixed designs the following tables present MASE, squared bias and variance on the complete data set and on trimmed data. In Table 3–5 the results are for the complete data set in the upper line, for the trimmed data in the lower line.

We found three main points:

1) It is not possible to declare one estimating procedure superior to the other one in general. The results are differing from model to model and for each additive component we want to estimate. We neither can say that one of the estimation procedures is in general outperforming the other one regarding the MASE nor that one of them is more biased or has less variance than the competing one.

2) The integration estimator is suffering more from boundary effects.

3) For increasing correlation both estimators get problems but much more the integration estimator. This is in line with the theory saying that the integration estimator is inefficient for correlated designs, see Linton (1997). He suggested an estimator for additive models constructed as a mixture of marginal integration and one-iteration-backfit and proved that for correlated designs this procedure dominates asymptotically the integration method in its variance part.

We want to emphasize our statements by looking closer to the behavior of squared bias, variance and MSE over the range.

The main difference from the results for using the Nadaraya Watson smoother is that the local linear smoother improves the integration estimator more than the backfitting estimator. The effects concerning the distribution of X and model structure are, as we expected, quite similar.

The following figures illustrate the behavior and the trade-off of variance and bias for both estimators in each additive direction. They are plotted on the range of the support. The boundaries of the data are cut off at a level of 5% each side, since otherwise their effects would dominate the pictures. The figures 7–11 reinforce clearly our observations and remarks concerning the Tables 3–5. They show the variance, squared bias and MASE over the whole range of X_1, respectively X_2, i.e. approximately over $(-3, 3)$ for

Sociedad De Estadistica E Investigacion Operativa Test (1999)
Vol. 8, No 2, pp. 419-458

Distr.		U^2	$N(.0)$	$N(.4)$	$N(.8)$	U^2	$N(.0)$	$N(.4)$	$N(.8)$	U^2	$N(.0)$	$N(.4)$	$N(.8)$
Model		$m = m_1 + m_2$				$m = m_1 + m_3$				$m = m_1 + m_4$			
$\hat{m}_\alpha$	ba	0.065	0.047	0.041	0.020	0.401	0.046	0.028	0.053	0.022	0.027	0.026	0.028
		0.060	0.038	0.031	0.014	0.393	0.037	0.018	0.033	0.018	0.018	0.016	0.019
	in	0.028	0.019	0.030	0.057	0.616	0.031	0.075	0.081	0.018	0.023	0.023	0.024
		0.023	0.013	0.017	0.047	0.555	0.024	0.059	0.071	0.014	0.016	0.015	0.017
$\hat{m}_\beta$	ba	0.116	0.083	0.079	0.047	0.479	0.073	0.053	0.058	0.033	0.033	0.036	0.048
		0.113	0.071	0.060	0.024	0.470	0.058	0.032	0.028	0.027	0.022	0.020	0.026
	in	0.436	0.090	0.116	0.530	0.402	0.137	0.234	0.528	0.191	0.043	0.074	0.137
		0.427	0.028	0.029	0.205	0.411	0.027	0.031	0.149	0.180	0.020	0.023	0.093
$\hat{m}$	ba	0.052	0.052	0.054	0.049	0.066	0.051	0.054	0.057	0.040	0.040	0.043	0.042
		0.045	0.032	0.031	0.028	0.057	0.030	0.029	0.035	0.333	0.024	0.024	0.024
	in	0.156	0.115	0.145	0.619	0.206	0.175	0.285	0.608	0.066	0.063	0.104	0.132
		0.143	0.041	0.041	0.252	0.159	0.043	0.053	0.194	0.051	0.031	0.041	0.089
Distr.		U^2	$N(.0)$	$N(.4)$	$N(.8)$	U^2	$N(.0)$	$N(.4)$	$N(.8)$	U^2	$N(.0)$	$N(.4)$	$N(.8)$
Model		$m = m_2 + m_3$				$m = m_2 + m_4$				$m = m_3 + m_4$			
$\hat{m}_\alpha$	ba	0.228	0.075	0.109	0.344	0.142	0.124	0.135	0.128	0.107	0.068	0.081	0.111
		0.206	0.057	0.079	0.119	0.141	0.107	0.116	0.099	0.101	0.046	0.055	0.081
	in	0.645	0.065	0.105	3.490	0.048	0.047	0.048	0.089	0.078	0.053	0.049	0.056
		0.572	0.031	0.064	0.782	0.041	0.022	0.026	0.078	0.070	0.026	0.022	0.041
$\hat{m}_\beta$	ba	0.271	0.060	0.086	0.238	0.132	0.112	0.121	0.110	0.077	0.051	0.062	0.096
		0.264	0.041	0.058	0.093	0.124	0.101	0.110	0.091	0.071	0.039	0.048	0.075
	in	0.310	0.070	0.191	1.499	0.061	0.048	0.480	1.322	0.177	0.057	0.603	2.413
		0.233	0.040	0.055	0.186	0.047	0.032	0.061	0.151	0.166	0.040	0.265	1.020
$\hat{m}$	ba	0.087	0.072	0.074	0.124	0.061	0.061	0.063	0.068	0.079	0.065	0.066	0.064
		0.076	0.042	0.041	0.067	0.052	0.035	0.035	0.037	0.069	0.038	0.037	0.038
	in	0.407	0.204	0.443	6.283	0.102	0.118	0.561	1.368	0.145	0.085	0.670	2.238
		0.308	0.136	0.236	0.785	0.084	0.076	0.083	0.189	0.123	0.044	0.257	0.681

Table 3: MASE, using the local linear smoother over all (upper) and over trimmed (lower) data. Here, ba stays for backfitting and in for marginal integration.

Sociedad De Estadistica E Investigacion Operativa Test (1999)
Vol. 8, No 2, pp. 419-458

Distr.		U^2	$N(.0)$	$N(.4)$	$N(.8)$	U^2	$N(.0)$	$N(.4)$	$N(.8)$	U^2	$N(.0)$	$N(.4)$	$N(.8)$
Model		$m = m_1 + m_2$				$m = m_1 + m_3$				$m = m_1 + m_4$			
$\hat{m}_\alpha$	ba	0.043	0.022	0.016	0.000	0.380	0.022	0.003	0.033	0.001	0.002	0.001	0.007
		0.039	0.020	0.014	0.000	0.338	0.019	0.003	0.017	0.000	0.002	0.001	0.004
	int	0.007	0.003	0.008	0.020	0.598	0.015	0.053	0.043	0.001	0.006	0.004	0.004
		0.005	0.002	0.003	0.017	0.481	0.010	0.036	0.036	0.000	0.005	0.003	0.002
$\hat{m}_\beta$	ba	0.078	0.046	0.037	0.002	0.425	0.033	0.008	0.015	0.004	0.004	0.003	0.010
		0.068	0.040	0.031	0.002	0.364	0.027	0.006	0.006	0.002	0.003	0.002	0.007
	int	0.358	0.054	0.065	0.415	0.329	0.097	0.175	0.431	0.142	0.013	0.034	0.055
		0.266	0.003	0.006	0.078	0.227	0.002	0.004	0.072	0.119	0.003	0.003	0.031
$\hat{m}$	ba	0.006	0.005	0.005	0.003	0.004	0.002	0.002	0.014	0.003	0.002	0.002	0.004
		0.005	0.004	0.004	0.003	0.001	0.001	0.001	0.007	0.002	0.001	0.001	0.002
	int	0.062	0.068	0.075	0.475	0.118	0.123	0.209	0.482	0.005	0.022	0.050	0.040
		0.032	0.008	0.010	0.103	0.062	0.009	0.015	0.099	0.003	0.006	0.012	0.024
Distr.		U^2	$N(.0)$	$N(.4)$	$N(.8)$	U^2	$N(.0)$	$N(.4)$	$N(.8)$	U^2	$N(.0)$	$N(.4)$	$N(.8)$
Model		$m = m_2 + m_3$				$m = m_2 + m_4$				$m = m_3 + m_4$			
$\hat{m}_\alpha$	ba	0.187	0.033	0.066	0.289	0.102	0.082	0.091	0.072	0.050	0.024	0.034	0.059
		0.156	0.028	0.049	0.058	0.087	0.072	0.080	0.058	0.039	0.017	0.026	0.047
	in	0.590	0.022	0.058	3.465	0.006	0.004	0.013	0.064	0.017	0.005	0.011	0.035
		0.387	0.007	0.033	0.574	0.004	0.000	0.004	0.047	0.010	0.002	0.002	0.019
$\hat{m}_\beta$	ba	0.219	0.023	0.051	0.198	0.105	0.086	0.095	0.074	0.050	0.025	0.036	0.060
		0.180	0.019	0.035	0.035	0.088	0.072	0.079	0.059	0.040	0.019	0.029	0.050
	in	0.238	0.030	0.132	1.473	0.018	0.018	0.455	1.300	0.132	0.027	0.577	2.391
		0.134	0.010	0.020	0.136	0.011	0.011	0.036	0.124	0.097	0.019	0.211	0.724
$\hat{m}$	ba	0.006	0.003	0.003	0.055	0.005	0.003	0.003	0.004	0.007	0.005	0.003	0.003
		0.003	0.002	0.001	0.013	0.003	0.002	0.002	0.002	0.004	0.002	0.002	0.001
	in	0.280	0.124	0.340	6.220	0.020	0.050	0.500	1.310	0.040	0.012	0.600	2.182
		0.098	0.077	0.153	0.624	0.010	0.033	0.041	0.138	0.023	0.005	0.170	0.539

Table 4: Averaged squared bias, using local linear smoother over all (upper) and over trimmed (lower) data. Here, ba stays for backfitting and in for marginal integration.

Sociedad De Estadistica E Investigacion Operativa Test (1999)
Vol. 8, No 2, pp. 419-458

Distr.		U^2	$N(.0)$	$N(.4)$	$N(.8)$	U^2	$N(.0)$	$N(.4)$	$N(.8)$	U^2	$N(.0)$	$N(.4)$	$N(.8)$
Model		$m = m_1 + m_2$				$m = m_1 + m_3$				$m = m_1 + m_4$			
$\hat{m}_\alpha$	ba	0.022	0.024	0.025	0.020	0.022	0.024	0.025	0.020	0.022	0.025	0.025	0.020
		0.015	0.014	0.014	0.012	0.015	0.014	0.014	0.012	0.015	0.014	0.014	0.013
	in	0.021	0.017	0.023	0.037	0.018	0.017	0.023	0.039	0.017	0.017	0.019	0.020
		0.015	0.009	0.011	0.025	0.013	0.010	0.011	0.026	0.012	0.009	0.010	0.013
$\hat{m}_\beta$	ba	0.038	0.037	0.042	0.045	0.054	0.040	0.045	0.043	0.029	0.029	0.034	0.038
		0.028	0.020	0.019	0.020	0.042	0.021	0.020	0.018	0.021	0.016	0.015	0.015
	in	0.078	0.036	0.051	0.115	0.073	0.040	0.059	0.097	0.049	0.029	0.040	0.083
		0.051	0.020	0.019	0.044	0.055	0.021	0.021	0.040	0.033	0.015	0.016	0.046
$\hat{m}$	ba	0.046	0.047	0.050	0.046	0.061	0.050	0.052	0.043	0.037	0.038	0.041	0.038
		0.036	0.025	0.026	0.024	0.049	0.027	0.027	0.022	0.028	0.021	0.021	0.019
	in	0.094	0.048	0.069	0.144	0.088	0.052	0.077	0.126	0.061	0.041	0.054	0.092
		0.066	0.027	0.030	0.067	0.067	0.029	0.033	0.064	0.044	0.023	0.025	0.053
Distr.		U^2	$N(.0)$	$N(.4)$	$N(.8)$	U^2	$N(.0)$	$N(.4)$	$N(.8)$	U^2	$N(.0)$	$N(.4)$	$N(.8)$
Model		$m = m_2 + m_3$				$m = m_2 + m_4$				$m = m_3 + m_4$			
$\hat{m}_\alpha$	ba	0.041	0.041	0.043	0.056	0.040	0.042	0.044	0.056	0.057	0.044	0.047	0.052
		0.030	0.020	0.020	0.025	0.030	0.021	0.020	0.025	0.043	0.022	0.021	0.023
	in	0.054	0.043	0.047	0.025	0.042	0.043	0.035	0.025	0.061	0.047	0.038	0.024
		0.039	0.020	0.022	0.012	0.031	0.019	0.016	0.012	0.046	0.021	0.017	0.012
$\hat{m}_\beta$	ba	0.053	0.037	0.035	0.041	0.027	0.026	0.026	0.035	0.027	0.026	0.026	0.036
		0.040	0.017	0.016	0.017	0.018	0.012	0.011	0.014	0.018	0.012	0.011	0.014
	in	0.073	0.040	0.059	0.026	0.043	0.030	0.026	0.022	0.045	0.030	0.026	0.022
		0.055	0.021	0.021	0.011	0.030	0.015	0.012	0.010	0.031	0.015	0.012	0.010
$\hat{m}$	ba	0.081	0.069	0.071	0.069	0.057	0.058	0.061	0.064	0.073	0.061	0.063	0.061
		0.065	0.038	0.039	0.036	0.044	0.032	0.032	0.033	0.058	0.034	0.033	0.031
	in	0.127	0.080	0.102	0.063	0.082	0.069	0.061	0.058	0.105	0.073	0.070	0.056
		0.098	0.043	0.048	0.026	0.064	0.035	0.029	0.024	0.082	0.037	0.030	0.023

Table 5: Averaged variance, using local linear over all (upper) and over trimmed (lower) data. Here, ba stays for backfitting and in for marginal integration.

Sociedad De Estadistica E Investigacion Operativa Test (1999)
Vol. 8, No 2, pp. 419-458

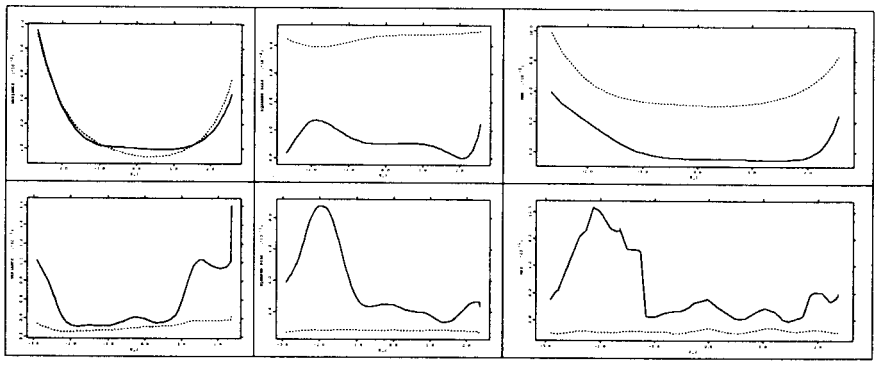


Figure 7: Variance (left), bias^2 (middle) and MASE (right) by bandwidth h_1 in model $m = m_1 + m_2$ for m_1 (top), m_2 (bottom) separately. Uniform design on $[-3,3]^2$, using local linear smoother.

$U^2[-3,3]$ and $(-1.7,1.7)$ for normal design with $cov = 0.8$. The absolute values in the vertical direction are not of interest here, otherwise, see Tables 3–5.

The *integration estimator* suffers more from sparseness of observations than the backfitting estimator does. In what follows the boundary effects are worse in the integration estimator and it does better for the normal distribution than for the uniform considering the MASE. At the mass of observations this estimator mostly has lower squared bias and variance for the estimators of the additive functions. Finally, we observe that an increasing ρ (covariance of the explanatory vector) affects strongly its MASE in a negative sense.

The *backfitting estimator* is less affected by boundary effects or correlation of the explanatory variables. For the regression estimator it fits the regression in general better than the integration estimator, at least the MASE is almost always smaller. But it pays for a low MSE (or MASE) for the regression with high MSE (MASE respectively) in the additive function estimation. Here we see the main difference of these estimators; the integration estimator is estimating the additive function by integrating out the directions not of interest, which means it is measuring the marginal influence of the considered input, whereas the backfitting estimator is looking in the space of additive models for the best fit of the response Y vs. X. For a more detailed discussion about their different interpretation, see Nielsen

Sociedad De Estadistica E Investigacion Operativa Test (1999)
Vol. 8, No 2, pp. 419-458

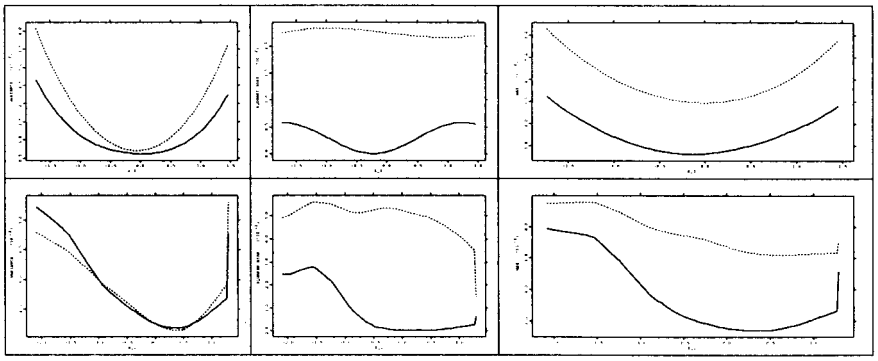


Figure 8: *Variance (left), bias^2 (middle) and MASE (right) by bandwidth h_1 in model $m = m_1 + m_2$, for m_1 (top), m_2 (bottom) separately. Standard normal design with cov= 0.0, using local linear smoother.*

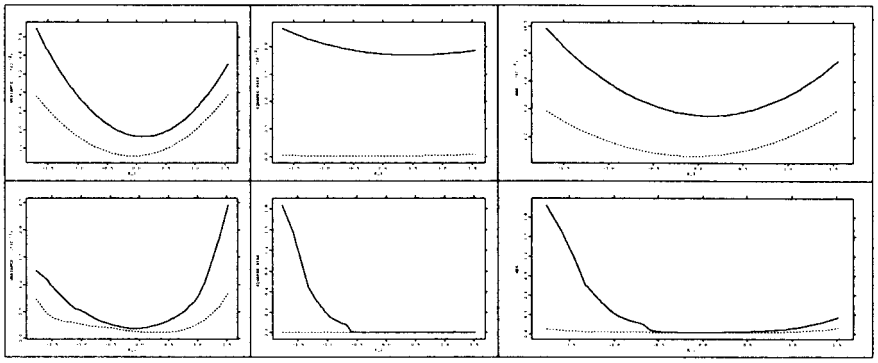


Figure 9: *Variance (left), bias^2 (middle) and MASE (right) by bandwidth h_1 in model $m = m_1 + m_2$, for m_1 (top), m_2 (bottom) separately. Standard normal design with cov= 0.8 , using local linear smoother.*

Sociedad De Estadistica E Investigacion Operativa Test (1999)
Vol. 8, No 2, pp. 419-458

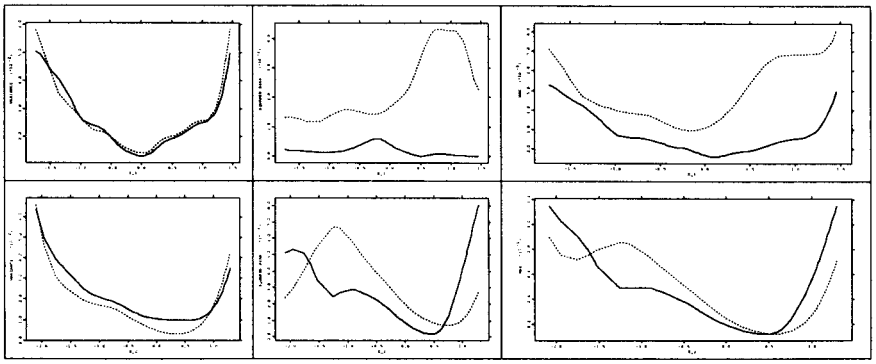


Figure 10: Variance (left), bias^2 (middle) and MASE (right) by bandwidth h_1 in model $m = m_3 + m_4$, for m_3 (top), m_4 (bottom) separately. Standard normal design with cov= 0.0 , using local linear smoother.

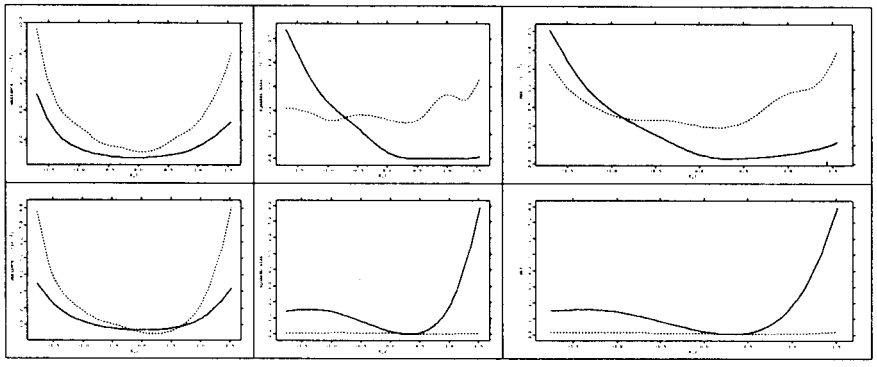


Figure 11: Variance (left), bias^2 (middle) and MASE (right) by bandwidth h_1 in model $m = m_3 + m_4$, for m_3 (top), m_4 (bottom) separately. Standard normal design with cov= 0.0 , using local linear smoother.

Sociedad De Estadistica E Investigacion Operativa Test (1999)
Vol. 8, No 2, pp. 419-458

and Linton (1997). An increase in the correlation ρ of the design again leads to a worse estimate.

Making Use of the Bandwidth Matrix For the purpose of correcting for correlation between the components of X we furthermore refined the estimation procedure by replacing the bandwidth vector h by its multivariate counterpart, a nonsingular bandwidth matrix H. This leads to the following multivariate kernel function:

$$K_H(u) = \frac{1}{\mid H \mid} K(H^{-1}u)$$

Motivated by the bandwidth matrix selection in the book of Wand and Jones (1995), the matrix H is constructed in the following way: Its diagonal elements are equivalent to the elements of the bandwidth vector h and its off-diagonal elements can be derived from the covariance matrix. In addition we included a factor δ which allows us to control the influence of the off-diagonal elements. Hence, for the two dimensional model one gets:

$$H = \begin{pmatrix} h_1 & \delta\rho \\ \delta\rho & h_2 \end{pmatrix}$$

Note that for $\delta = 0$ we would get the results of the previously applied estimation procedure. This can be checked from the tables of Section 3.4. Defining the bandwidth matrix in this way we were able to run the estimation on a grid for δ. Table 6 presents our results for the standard normal design for $cov(x_1, x_2) = 0.4$ and $cov(x_1, x_2) = 0.8$. The MASE is shown together with the value of δ, by which it is minimized, for trimmed data in brackets. To compare the results with the former one we present them in the following table together with the MASE which we got for diagonal bandwidth matrices in the integration procedure. Obviously the fit can be improved significantly if we use a proper off–diagonal element in the bandwidth matrices. Since interpretation using a nondiagonal bandwidth matrix is different for the backfitting, due to its iterative character and smoothing always in univariate subspaces, we skipped this investigation for the backfitting. However, such an investigation in theory and practice would be interesting for that method, too.

Using the local linear smoother, the optimal bandwidth matrix to estimate the linear function m_1 should be huge or even infinite on the first di-

Sociedad De Estadistica E Investigacion Operativa Test (1999)
Vol. 8, No 2, pp. 419-458

agonal, but we cannot ameliorate the results by changing the off–diagonals. For that reason we skipped models that include m_1.

cov	Model	m_j	δ_{min}		$MASE^{int.}_{with}$		$MASE^{int.}_{old}$	
	$m = m_2 + m_3$	m_2	0.9	(0.7)	0.090	(0.040)	0.105	(0.064)
		m_3	0.5	(0.6)	0.144	(0.030)	0.191	(0.055)
0.4	$m = m_2 + m_4$	m_2	1.5	(0.6)	0.045	(0.022)	0.048	(0.026)
		m_4	0.0	(0.1)	0.480	(0.061)	0.480	(0.061)
	$m = m_3 + m_4$	m_3	0.2	(0.2)	0.049	(0.022)	0.049	(0.022)
		m_4	1.5	(1.5)	0.590	(0.217)	0.603	(0.265)
	$m = m_2 + m_3$	m_2	0.6	(0.5)	0.244	(0.167)	3.490	(0.782)
		m_3	0.5	(1.5)	0.246	(0.038)	1.499	(0.186)
0.8	$m = m_2 + m_4$	m_2	0.4	(1.1)	0.079	(0.045)	0.089	(0.078)
		m_4	1.5	(0.8)	1.463	(0.194)	1.322	(0.151)
	$m = m_3 + m_4$	m_3	0.1	(0.4)	0.074	(0.052)	0.056	(0.041)
		m_4	1.5	(1.5)	2.385	(1.008)	2.413	(1.020)

Table 6: Performance (MASE) with vs without off–diagonals δ_{min} in the bandwidth matrix using loc. lin. smoother. Cov indicates the covariance between the regressors X_1 and X_2.

A.5 Singular and Eigenvalue Analysis

Eigendecomposition of the smoother matrix of an estimator can be used to describe the behavior of the smoother, especially when this matrix is symmetric and thus the eigenvalues are real. In that case this is much like the use of a *transfer function* to describe a linear filter for time series. This connection is made precise in Hastie and Tibshirani (1990). If the smoother matrix is not symmetric we have to turn to the singular value analysis since a eigendecomposition often would lead to complex eigenvalues.

In the following we present the first respectively the biggest singular values of the weight matrices. These smoothing matrices are symmetric for the backfitting, using local polynomial kernel smoothing but they are not symmetric for the integration estimator. Thus we did a singular value analysis for the integration procedure.

In Figure 12 to 15 we give the calculated values. Again in all figures the lines for the integration estimator are solid, for the backfitting estimator dotted. For each design distribution presented we give the results for two

Sociedad De Estadistica E Investigacion Operativa Test (1999)
Vol. 8, No 2, pp. 419-458

randomly drawn samples.

The slope of the eigen or singular values, see Figure 12 to 15, gives us an idea of the smoothness of the specific estimator. They almost always cross, often the backfitting eigenvalue is a little bit steeper, what depends on the bandwidth choice, but there seems to be no remarkable difference between the integration and the backfitting method regarding the eigenvalue analysis.

A.6 Degrees of Freedom

Another parameter we looked at is the degree of freedom of the smoothers. Hastie and Tibshirani (1990) give various interpretations for degrees of freedom in the context of nonparametric estimation as well as for testing nonparametrically. One of them is that they give us the amount of fitting. Further they can be used to approximate the distribution of test statistics. They also state that we can draw out of them some information about the smoothness of the estimator. So they propose for a fair comparison of different estimators to choose those smoothing parameters that give equal degrees of freedom for the different estimators. Our experience was that this leads to unreasonable bandwidths. So we have to doubt these interpretations at least for the integration estimator.

For all smoothing matrices we calculated the values for three different definitions of degrees of freedom, $\mathrm{tr}(W)$, $\mathrm{tr}(WW^T)$ and $n-\mathrm{tr}(2W-WW^T)$, but restrict ourselves in presenting only $\mathrm{tr}(W)$. The other results can be requested.

As already mentioned at the beginning of this paragraph the chosen asymptotically "optimal" bandwidth led us to totally different degrees of freedom as defined above.

Looking at Table 7, where the degrees are defined as the trace of W, we see that the degrees of freedom for the backfitting are almost always bigger than the degrees for the integration estimator. For both estimators the degrees are bigger in the case of normal distributed designs but it is hardly possible to detect a systematic difference in the degrees for the increasing correlation of the explanatory variables. What can be seen clearly is that the degrees of freedom are varying strongly with the choice of the model. This holds true for both estimators.

Sociedad De Estadistica E Investigacion Operativa Test (1999)
Vol. 8, No 2, pp. 419-458

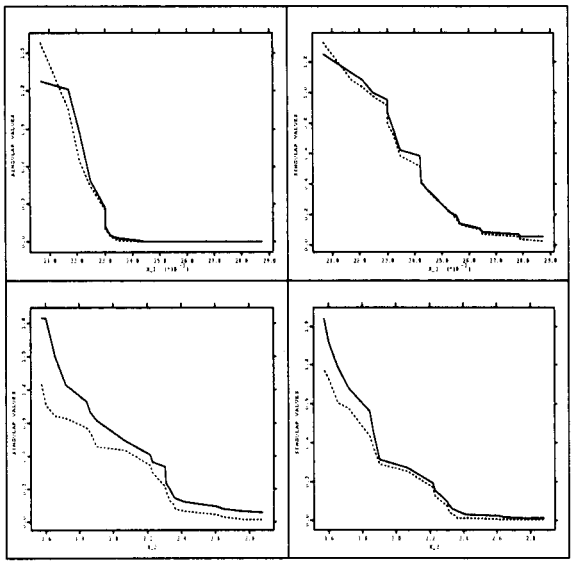


Figure 12: Eigen-/Singular value analysis using local linear smoother. Plotted are x_1 (top), x_2 (bottom) vs eigen/singular values for two uniformly distributed samples (left, right).

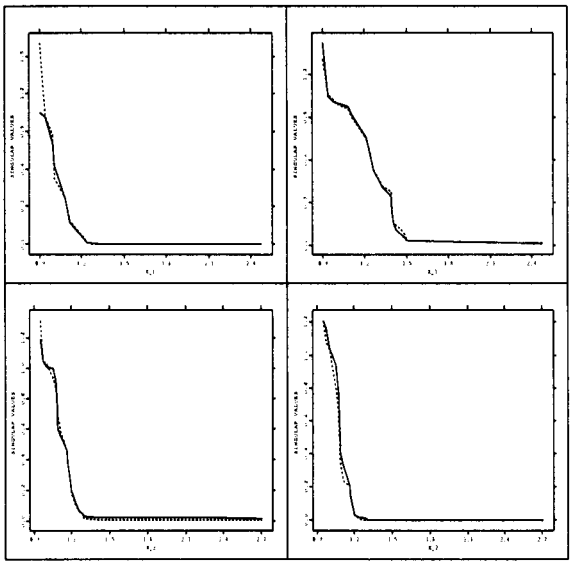


Figure 13: Eigen-/Singular value analysis using local linear smoother. Plotted are x_1 (top), x_2 (bottom) vs eigen/singular values for two normal (cov= 0.0) distributed samples (left, right).

Sociedad De Estadistica E Investigacion Operativa Test (1999)
Vol. 8, No 2, pp. 419-458

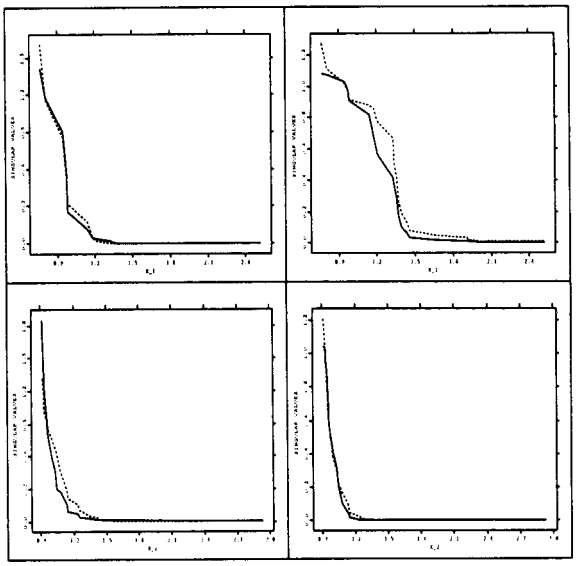


Figure 14: Eigen-/Singular value analysis using local linear smoother. Plotted are x_1 (top), x_2 (bottom) vs eigen/singular values for two normal (cov= 0.4) distributed samples (left, right).

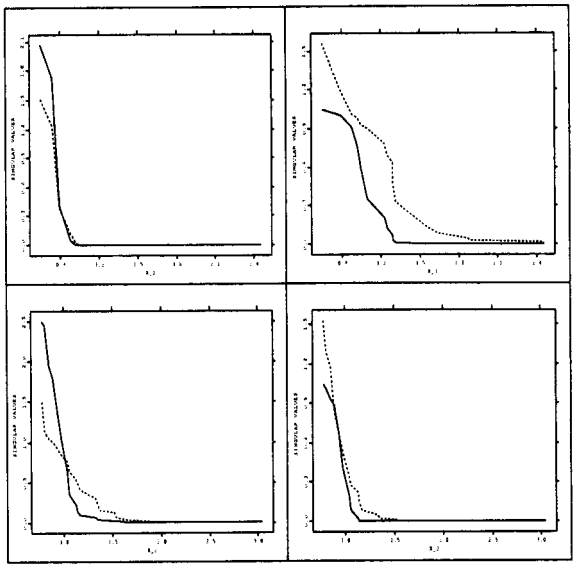


Figure 15: Eigen-/Singular value analysis using local linear smoother. Plotted are x_1 (top), x_2 (bottom) vs eigen/singular values for two normal (cov= 0.8) distributed samples (left, right).

Sociedad De Estadistica E Investigacion Operativa Test (1999)
Vol. 8, No 2, pp. 419-458

Distribution:		U^2	$N(.0)$	$N(.4)$	$N(.8)$	U^2	$N(.0)$	$N(.4)$	$N(.8)$
$m = m_\alpha + m_\beta$		$m = m_1 + m_2$				$m = m_1 + m_3$			
$\hat{m}_\alpha$	back.	3.63	4.19	4.01	2.43	3.60	4.19	4.00	2.43
	int.	3.52	3.96	3.31	2.11	3.64	3.61	3.29	2.11
$\hat{m}_\beta$	back.	8.42	7.54	8.30	8.86	12.70	8.15	8.87	8.19
	int.	8.53	8.00	6.66	3.84	12.99	8.65	7.39	3.52
$\hat{m}$	back.	13.05	12.73	13.30	12.29	17.30	13.33	13.87	11.62
	int.	11.05	10.96	8.97	4.95	15.64	11.26	9.68	4.63
$m = m_\alpha + m_\beta$		$m = m_1 + m_4$				$m = m_2 + m_3$			
$\hat{m}_\alpha$	back.	3.65	4.22	4.04	2.45	9.16	9.32	9.52	10.37
	int.	3.71	3.85	3.86	2.24	8.37	8.78	7.21	6.09
$\hat{m}_\beta$	back.	5.88	5.38	6.17	6.75	12.49	8.04	8.05	6.88
	int.	6.17	5.81	4.58	2.48	13.00	8.69	7.36	6.15
$\hat{m}$	back.	10.53	10.59	11.20	10.19	22.64	18.37	18.56	18.25
	int.	8.88	8.66	7.44	3.72	20.37	16.46	13.58	11.24
$m = m_\alpha + m_\beta$		$m = m_2 + m_4$				$m = m_3 + m_4$			
$\hat{m}_\alpha$	back.	9.33	9.37	9.61	10.41	13.73	10.02	10.26	9.51
	int.	9.34	8.78	8.34	6.09	13.80	9.23	9.24	5.69
$\hat{m}_\beta$	back.	5.78	5.31	5.49	5.49	5.70	5.31	5.39	5.67
	int.	6.33	5.73	6.33	5.22	6.25	5.79	6.33	5.22
$\hat{m}$	back.	16.10	15.69	16.10	16.89	20.43	16.32	16.65	16.18
	int.	14.67	13.52	13.66	10.31	19.05	14.02	14.56	9.916

Table 7: Degrees of Freedom measured by trace(W), using local linear smoother.

Note that the degrees of the function m in the integration method is the result of summing the degrees of its additive components minus one, as a result of eliminating in each estimation the sample mean. In the backfitting you take the sum of the degrees of the additive components and add one, see Opsomer and Ruppert (1997).

When we considered tr(WW^T), this is certainly different. Further, in the local linear case, considering tr(WW^T) led to different results at all. Here now the degrees were often much bigger for the integration method. However, since interpretation is hardly possible in that case, we skipped the presentation of these results.

A.7 The Equivalent Kernel Weights of the Estimators

What price do we pay to overcome the curse of dimensionality by choosing an additive model structure? To examine this we compared the two

Sociedad De Estadistica E Investigacion Operativa Test (1999)
Vol. 8, No 2, pp. 419-458

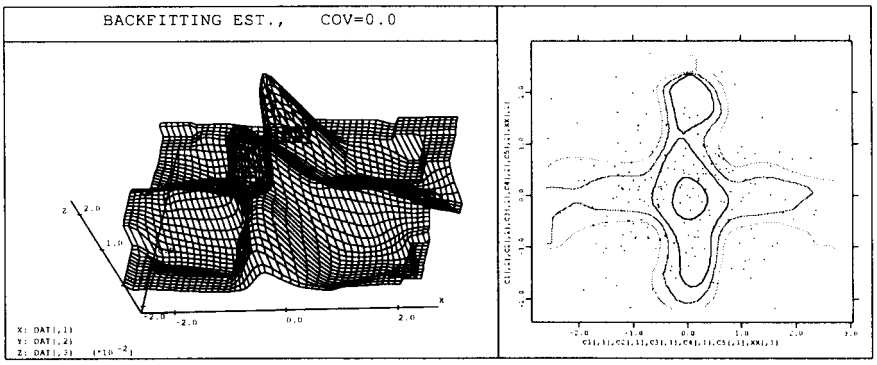


Figure 16: Equivalent kernels. 3-D and contour plot for the Backfitting estimator, using Nadaraya Watson. Regressors are standard normal with cov = 0.0.

additive model estimators, backfitting and integration procedure, with the bivariate Nadaraya Watson kernel smoother. Equivalent kernels are defined as the linear weights w of the estimates to fit the regression function at a particular point, in our case at $(0,0)$. For the integration estimator we used only a diagonal bandwidth matrix as in the beginning, even for the strongly correlated designs. We have considered $n = 100$, bivariate normal distributed designs with mean zero, variance 1 and increasing correlation $\rho = 0.0,\ 0.2,\ 0.4,\ 0.6$ and 0.8, but give only figures for $0.0,\ 0.4$ and 0.8. Please note that equivalent kernel weights depend only on the kernel function, the bandwidths and X but not on Y. So the results in Figure 16–24 presented hold for any underlying two dimensional model.

Since the local linear smoother is also taking into account the first derivative of the functions, we would get, depending on the data generating functions, positive and negative weights varying from point to point. Thus for the local linear smoother the pictures shown beneath would look like wild mountain scenery and so we skipped their presentation.

As we would have expected, both additive model estimators get their strength from the local panels orthogonal to the axes of X_1 and X_2 instead of uniformly in all directions like the bivariate Nadaraya Watson smoother. Since they are composed by components that behave like univariate smoothers, they can overcome the curse of dimensionality. For the backfitting this was already stated by Hastie and Tibshirani (1990). We can see clearly now that the integration estimator behaves very similar.

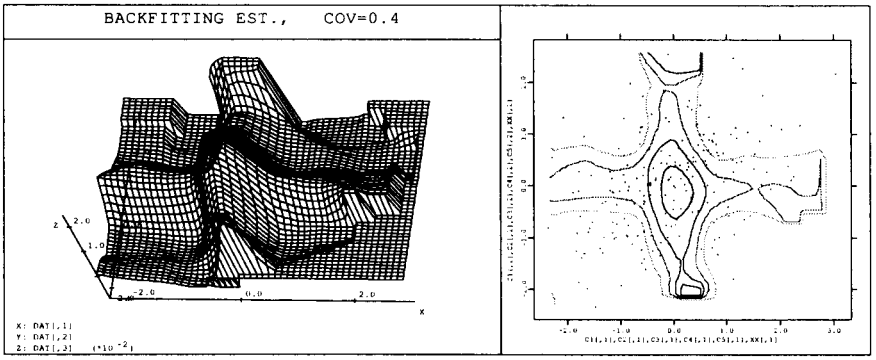


Figure 17: *Equivalent kernels. 3-D and contour plot for the Backfitting estimator, using Nadaraya Watson. Regressors are standard normal with cov* = 0.4.

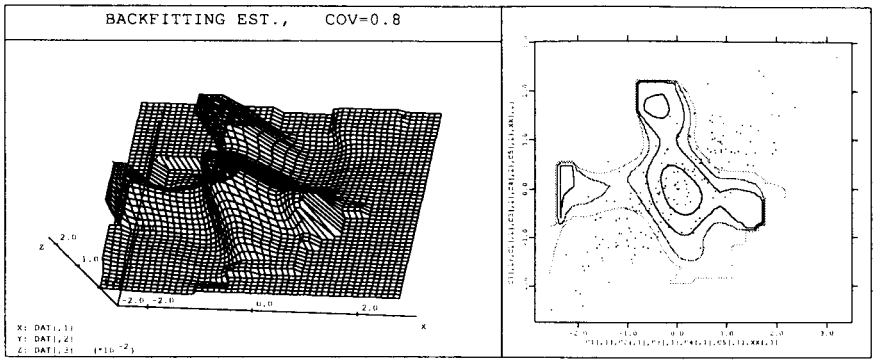


Figure 18: *Equivalent kernels. 3-D and contour plot for the Backfitting estimator, using Nadaraya Watson. Regressors are standard normal with cov* = 0.8.

Sociedad De Estadistica E Investigacion Operativa Test (1999)
Vol. 8, No 2, pp. 419-458

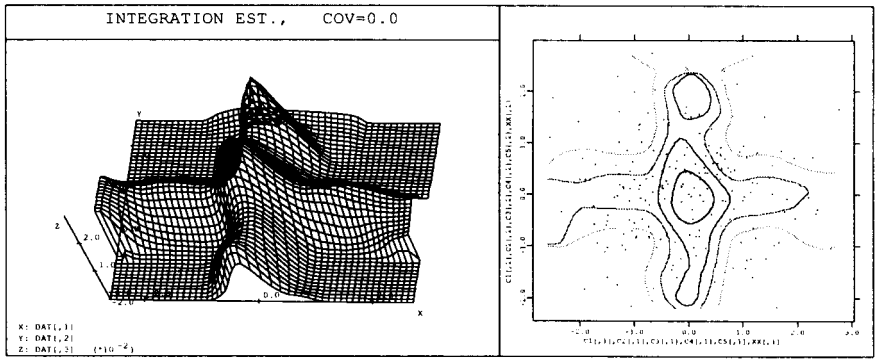


Figure 19: *Equivalent kernels. 3-D and contour plot for the Integration estimator, using Nadaraya Watson. Regressors are standard normal with cov* = 0.0.

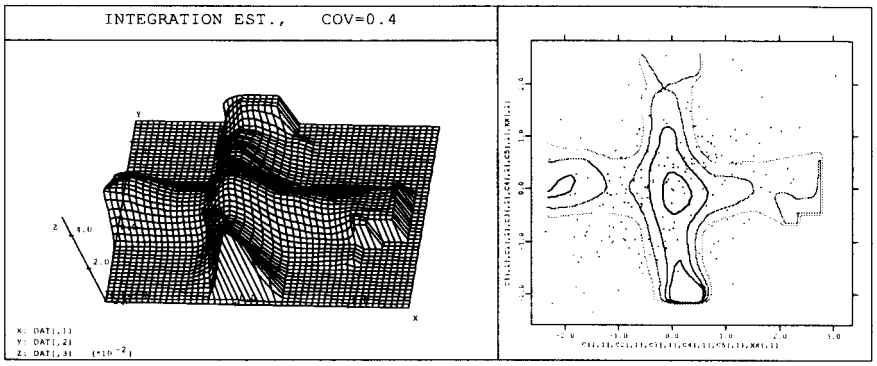


Figure 20: *Equivalent kernels. 3-D and contour plot for the Integration estimator, using Nadaraya Watson. Regressors are standard normal with cov* = 0.4.

Sociedad De Estadistica E Investigacion Operativa Test (1999)
Vol. 8, No 2, pp. 419-458

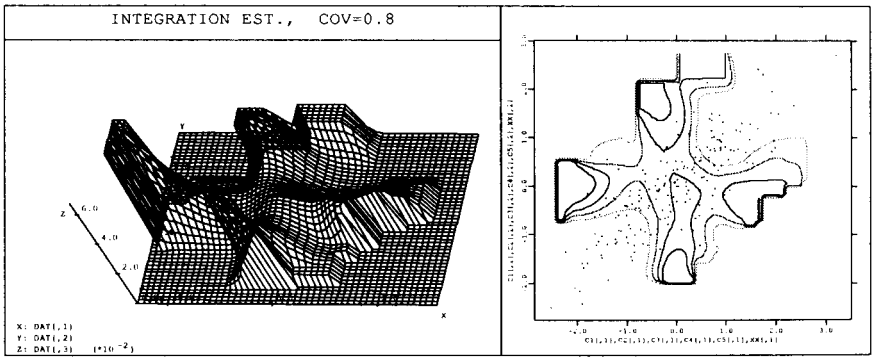


Figure 21: *Equivalent kernels. 3-D and contour plot for the Integration estimator, using Nadaraya Watson. Regressors are standard normal with cov* = 0.8.

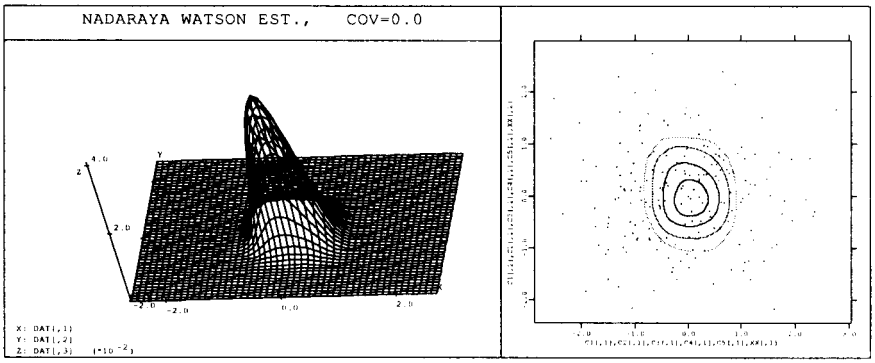


Figure 22: *Equivalent kernels. 3-D and contour plot for the Multidimensional Nadaraya Watson estimator. Regressors are standard normal with cov* = 0.0.

Sociedad De Estadistica E Investigacion Operativa Test (1999)
Vol. 8, No 2, pp. 419-458

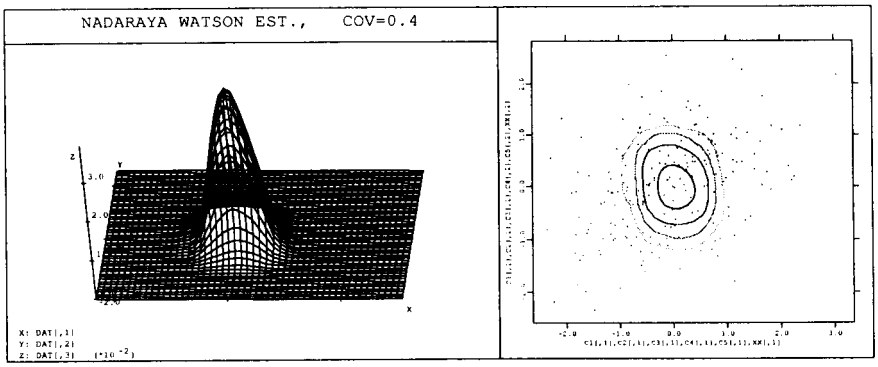


Figure 23: Equivalent kernels. 3-D and contour plot for the Multidimensional Nadaraya Watson estimator. Regressors are standard normal with cov = 0.4.

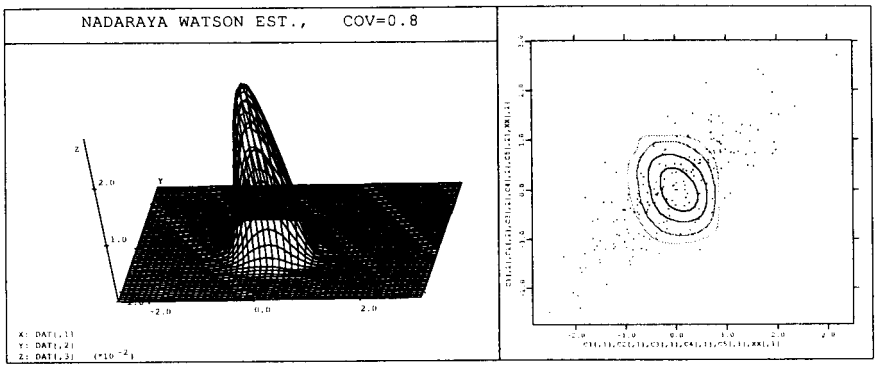


Figure 24: Equivalent kernels. 3-D and contour plot for the Multidimensional Nadaraya Watson estimator. Regressors are standard normal with cov = 0.8.

Sociedad De Estadistica E Investigacion Operativa Test (1999)
Vol. 8, No 2, pp. 419-458

The pictures for the additive smoothers look almost the same, except that the backfitting can also get some negative weights whereas the integration estimator cannot by its construction.

Both estimators run into deep problems to estimate properly in designs with increasing correlation. In contrast to the bivariate Nadaraya Watson smoother this can be seen in the figures for the backfitting as well as for the integration method. But we are not able to discover visually the reason why the integration estimator is doing worse for highly correlated explanatory variables than e.g. the backfitting.

A.8 Do the Asymptotics hold empirically?

For restricting our presentation on $n = 100$ observations we had mainly two reasons. First, in our simulations we had the same findings also for n different from 100, what is indicated also in this section, see below. Second, for $n > 100$ the difference between integration and backfitting method decreases in such an amount that it even would be hard to illustrate them at all. To answer the question about the asymptotics, we did a simulation study, using the local linear smoother, as follows.

We considered the model

$$Y = c + m_1(x_1) + m_2(x_2) + \varepsilon$$

with $m_1(x) = 2x$, $m_2(x) = x^2 - E(x^2)$ and $c = 0$. The error term ε has been normal distributed with mean zero and variance 0.5, the design X was uniform on $[-3,3]^2$ distributed. For $n = 250, 500, 1000$ and 2000 observations we calculated the estimates $\widehat{m}_1$, $\widehat{m}_2$ at $x = -1.5, -0.75, 0.0, 0.75$ and 1.5 and determined their biases B (which is always mentioned as $b \cdot h^2$ in theory) and variances V for each n. The bandwidths have been $h_1 := h_n = h_0 n^{-1/5}$ with $h_0 \approx 0.69$ and $h_2 := g_n = 2h_n/3$ for the nuisance direction.

Our first question was whether the rate of convergence mentioned by the theory holds also empirically. Therefore we considered the following regression

$$\begin{aligned} \ln(V) &= \beta_1 \ln(nh_n) \\ \ln(B) &= \beta_2 \ln(nh_n). \end{aligned}$$

Sociedad De Estadistica E Investigacion Operativa Test (1999)
Vol. 8, No 2, pp. 419-458

For the integration estimator we got for all five points $\beta_1 \approx \beta_2 \approx -1.003$, for the backfitting $\beta_1 \approx -1.02$ and $\beta_2 \approx -1.05$ which ensures the theory concerning the rate of convergence.

The second question we were interested in was the comparison of the empirical biases and variances calculated in our simulation study with the analytical ones. We present results only for the function m_2 in the above mentioned setting, but have to remark that the biases certainly depend on the particular data generating model as well as on the chosen design, at least in practice. To consider the function m_1 that is linear in this model is useless since we know that a local linear estimator is fitting such a function almost always exactly by definition and thus this would not be typical in practice. For the comparison see Table 8 for the analytical values and Table 9, 10 for the empirical values.

n	250	500	1000	2000
variance (equal for all points)	0.0147	0.0085	0.0048	0.0028
bias (equal for all points)	0.0529	0.0400	0.0306	0.0225

Table 8: Analytical bias and variance.

As we can see the estimator is doing very well for an increasing number of observations and at least for a low dimensional model the integration estimator obviously reaches his asymptotics pretty fast.

Since we could not calculate (in GAUSS) with weight matrices for the backfitting procedure when n was ≥ 1000, we had to determine the empirical bias and variance by doing 400 replications for huge n and did the regression described above separately for 250 and 500, respectively for 1000 and 2000. We can conclude from β_1 and β_2 that bias and variance also diminish almost in the theoretical one dimensional rate. Obviously the constant h_0 of the bandwidth is chosen too big here, as can be seen in Table 10. The variance calculated with the aid of the weight matrices is smaller than expected whereas the bias is much bigger.

Since in this subsection we were not interested in the direct comparison of the MSE or something similar for backfitting and integration method, we did not look for an optimal bandwidth in each direction neither for each method. So one should only look on the tables respectively the asymptotic behavior of the estimates, but not for a comparison of the absolute values.

Sociedad De Estadistica E Investigacion Operativa Test (1999)
Vol. 8, No 2, pp. 419-458

n		250	500	1000	2000
	-1.5	0.01897	0.00986	0.00535	0.00305
	-0.75	0.01813	0.00999	0.00536	0.00303
variance at	$+0.0$	0.01825	0.01019	0.00548	0.00306
	$+0.75$	0.01807	0.00996	0.00535	0.00301
	$+1.5$	0.01765	0.00978	0.00546	0.00309
	-1.5	0.06237	0.04866	0.03206	0.02552
	-0.75	0.06681	0.04932	0.03196	0.02509
bias at	$+0.0$	0.05892	0.04705	0.03303	0.02567
	$+0.75$	0.06410	0.04853	0.03139	0.02453
	$+1.5$	0.07067	0.05071	0.03313	0.02471

Table 9: Small sample bias and variance for Integration estimator.

A.9 In higher dimensions

Due to the excess of information in this paper we only present results for $d = 4$, $n = 500$. Other simulations we did result in the same statements made for this special case. Here we did 100 replications and calculated bias and variance empirically by doing 400 replications. We took the analytically optimal bandwidth for the estimation of the additive functions, compare our discussion at the very beginning of our simulation study. The additive functions in our model have been

$$\begin{aligned} m_1(x) &= 2x, \qquad m_2(x) = x^2 - E(x^2), \\ m_3(x) &= \exp(x) - E\{\exp(x)\} \quad \text{and} \quad m_4(x) = 0.5 \cdot \sin(-1.5x). \end{aligned}$$

n		250	500	1000	2000
	-1.5	0.01431	0.00793	0.01503	0.00619
	-0.75	0.01411	0.00798	0.01231	0.00684
variance at	$+0.0$	0.01404	0.00801	0.01509	0.00755
	$+0.75$	0.01409	0.00796	0.01073	0.00528
	$+1.5$	0.01391	0.00791	0.01417	0.00684
	-1.5	0.31041	0.22552	0.16990	0.11953
	-0.75	0.30895	0.22489	0.16621	0.12171
bias at	$+0.0$	0.31176	0.22576	0.17181	0.13314
	$+0.75$	0.31097	0.22529	0.17411	0.13399
	$+1.5$	0.31137	0.22625	0.18929	0.12787

Table 10: Small sample bias and variance for Backfitting estimator.

Sociedad De Estadistica E Investigacion Operativa Test (1999)
Vol. 8, No 2, pp. 419-458

	$\hat{m}_1$		$\hat{m}_2$		$\hat{m}_3$		$\hat{m}_4$		$\hat{m}$	
Distr.	U^2	N(0.0)	U^2	N(0.0)	U^2	N(0.0)	U^2	N(0.0)	U^2	N(0.0)
h_1	20	20	0.212	0.211	0.138	0.194	0.309	0.307		
back.	0.051	0.018	0.180	0.159	0.078	0.036	0.074	0.075	0.028	0.024
int.	0.041	0.056	0.100	0.037	0.156	0.057	0.135	0.106	0.250	0.540

Table 11: MASE in higher dimensions ($d = 4$) for additive components and regression function. First row gives the estimated function.

Some final results are presented in Table 11 together with the bandwidth we used. The bandwidth for the directions not of interest in the integration estimator has been chosen as 0.45. The trends already discovered in the simpler cases were enforced in that study. The regression function itself is estimated well by backfitting whereas the marginal influences of the explanatory variables sometimes are better estimated by the integration estimator. Since the integration estimator suffers much more from boundary effects and data sparseness, what is especially the case in higher dimensions, the average mean squared error looks quite often worse. This concerns mainly the simulation example where the design is normal distributed.

4 Conclusion

A common misunderstanding of the integration method is that it must inherit the poor properties of the high dimensional regression estimator. Of course, this is absurd. It amounts to saying that the sample mean must behave poorly because the individual observations from which it is constructed are inconsistent estimates of the mean themselves. In any event, we have not found this to be the case. In fact, we have found many similarities between the integration and backfitting methodologies in terms of what they do to the data (for example the eigenanalysis) and indeed their statistical performance. In particular, both integration and backfitting suffer some small sample cost. The backfitting method seems to work better at boundary points and when there is high correlation among the covariates, while the integration method works better in most of the other cases and especially in estimating the components as opposed to the function itself.

Sociedad De Estadistica E Investigacion Operativa Test (1999)
Vol. 8, No 2, pp. 419-458

Acknowledgements

We would like to thank R.J. Carroll, J. Horowitz, J.P. Nielsen, M. Neumann, R. Tschernig, and two anonymous referees for helpful comments.

References

Breiman, L. and J.H. Friedman (1985). Estimating optimal transformations for multiple regression and correlation (with discussion). *Journal of the American Statistical Association*, **80**, 580-619.

Buja, A., T. Hastie and R. Tibshirani (1989). Linear smoothers and additive models (with discussion). *The Annals of Statistics*, **17** , 453-555.

Deaton, A. and J. Muellbauer (1980). *Economics and Consumer Behavior.* Cambridge University Press, Cambridge.

Härdle, W., and P. Hall (1993). On the backfitting algorithm for additive regression models. *Statistica Neederlandica*, **47**, 43-57.

Härdle, W., and O.B. Linton (1994). Applied nonparametric methods, *The Handbook of Econometrics*, vol. IV, ch. 38 (R.F. Engle and D.F. McFadden, eds.) Elsevier, Amsterdam.

Hastie, T. and R. Tibshirani (1990). *Generalized Additive Models.* Chapman and Hall, London.

Ibragimov, I.A. and R.Z. Hasminskii (1980). On nonparametric estimation of regression, *Soviet Math. Dokl.*, **21**, 810-814.

Linton, O.B. (1997). Efficient estimation of additive nonparamteric regression models. *Biometrika*, **84**, 469-473.

Linton, O.B., R. Chen, N. Wang, and W. Härdle (1995). An analysis of transformation for additive nonparametric regression. *Journal of the American Statistical Association*, **92**, 1512-1521.

Linton, O.B. and W. Härdle (1996). Estimation of additive regression models with known links. *Biometrika*, **83**, 529-540.

Linton, O.B., E. Mammen and J. Nielsen (1998). The Existence and Asymptotic Properties of a Backfitting Projection Algorithm under weak conditions. Manuscript, Yale University.

Linton, O.B. and J.P. Nielsen (1995). A kernel method of estimating structured nonparametric regression based on marginal integration. *Biometrika*, **82**, 93-100.

Sociedad De Estadistica E Investigacion Operativa Test (1999)
Vol. 8, No 2, pp. 419-458

Masry, E. and D. Tjøstheim (1995). Nonparametric estimation and identification of nonlinear ARCH time series: strong convergence and asymptotic normality. *Econometric Theory*, **11**, 258-289.

Masry, E. and D. Tjøstheim (1997). Additive nonlinear ARX time series and projection estimates. *Econometric Theory*, **13**, 214-252.

Newey, W.K. (1990). Semiparametric efficiency bounds. *Journal of Applied Econometrics*, **5**, 99-135.

Newey, W.K. (1994). Kernel estimation of partial means. *Econometric Theory*, **10**, 233-253.

Nielsen, J.P. (1996). Multiplicative and additive marker dependent hazard estimation based on marginal integration. Manuscript, PFA Pension.

Nielsen, J.P. and O.B. Linton (1997). An optimization interpretation of integration and backfitting estimators for separable nonparametric models. *Journal of the Royal Statistical Society, Series B*, **60**, 217-222.

Opsomer, J.D. and D. Ruppert (1997). Fitting a bivariate additive model by local polynomial regression. *The Annals of Statistics*, **25**, 212-243.

Porter, J. (1996). Essays in Semiparametric Econometrics. *PhD Thesis, MIT.*

Powell, J.L. (1994). Estimation in semiparametric models. *The Handbook of Econometrics*, vol. IV, ch. 41 (R.F. Engle and D.F. McFadden, eds.) Elsevier, Amsterdam.

Ruppert, D. and M.P. Wand (1995). Multivariate Locally Weighted Least Squares. *The Annals of Statistics*, **22**, 1346-1370.

Severance-Lossin, E. and S. Sperlich (1997). Estimation of Derivatives for Additive Separable Models. *Discussion Paper, SFB 373, Humboldt-University Berlin, Germany.*

Stone, C.J. (1980). Optimal rates of convergence for nonparametric estimators. *The Annals of Statistics*, **8**, 1348-1360.

Stone, C.J. (1982). Optimal global rates of convergence for nonparametric regression. *The Annals of Statistics*, **8**, 1040-1053.

Stone, C.J. (1985). Additive regression and other nonparametric models. *The Annals of Statistics*, **13**, 685-705.

Stone, C.J. (1986). The dimensionality reduction principle for generalized additive models. *The Annals of Statistics*, **14**, 592-606.

Tjøstheim, D. and B. Auestad (1994). Nonparametric identification of nonlinear time series: projections. *Journal of the American Statistical Association*, **89**, 1398-1409.

Sociedad De Estadistica E Investigacion Operativa Test (1999)
Vol. 8, No 2, pp. 419-458

Venables, W.N. and B. Ripley (1994). *Modern applied statistics with S-Plus.* Springer Verlag, New York.

Wand, M.P. and M.C. Jones (1995). *Kernel Smoothing.* Monographs on Statistics and Applied Probability, vol. 60. Chapman and Hall, London.

Communications In Statistics Theory & Methods (1999), 28, pp. 2025-2036

LARGE SAMPLE THEORY OF ESTIMATION OF ERROR DISTRIBUTION FOR A SEMIPARAMETRIC MODEL *

Hua Liang[a] Wolfgang Härdle[b]
Institut für Statistik und Ökonometrie
Humboldt-Universität zu Berlin
D-10178 Berlin, Germany [a,b]
Institute of Systems Science, Beijing 100080, China [a]

Abstract

The paper studies large sample theory of estimators of the error the distribution for the semiparametric model $Y = X^\tau\beta + g(T) + \varepsilon$. Under appropriate conditions, we prove that the estimators converge in probability, almost surely converge and uniformly almost surely converge. Asymptotic normality and the rates of convergence of the estimators are also investigated. Finally we establish the law of the iterated logarithm for the estimators.

Key Words and Phrases: *Weak, strong consistency; uniformly strong consistency; rates of convergence; asymptotic normality; law of the iterated logarithm; semiparametric model.*

1 INTRODUCTION

Consider the model given by

$$Y_i = X_i^\tau\beta + g(T_i) + \varepsilon_i, i = 1, \ldots, \tag{1}$$

where $X_i = (x_{i1}, \ldots, x_{ip})^\tau (p \geq 1)$ and $T_i (T_i \in [0,1])$ are known fixed design points, $\beta = (\beta_1, \ldots, \beta_p)^\tau$ is an unknown parameter vector and g is an unknown function, and $\varepsilon_1, \ldots, \varepsilon_n$ are i.i.d. random variables with a common unknown density function $f(u)$, and mean 0 and finite variance σ^2. The model was introduced by Engle, et al. (1986) to study the effect of weather or electricity demand. More recent work dealt with the estimation of β at a parametric rate. Chen (1988), Chen and Shiau (1991), Heckman (1986, 1988), Robinson (1988), Schick (1996) and Speckman (1988) constructed $\sqrt{n}-$consistent estimates of β under the nonsingularity of the matrix $E[\{X_1 - E(X_1|T_1)\}\{X_1 - E(X_1|T_1)\}^\tau]$ which guarantees the identifiability of

*The research for both authors was supported by Sonderforschungsbereich 373 "Quantifikation und Simulation Ökonomischer Prozesse". The first author was also supported by Alexander von Humboldt Foundation .

Liang, H. and Härdle, W. (1999) Large Sample Theory of Estimation Of Error Distribution For A Semiparametric Model

the parameter β, under various conditions on $E(X_1|T_1)$ and $g(\bullet)$ when (X_i, T_i) are random design points. Cuzick (1992a) constructed efficient estimates of β when the error density is known and has finite Fisher information. The same problem was solved by Cuzick (1992b) and Schick (1993) when the error distribution is unknown.

In this paper, the estimators of $f(u)$, $\widehat{f}_n(u)$, are obtained by using nonparametric regression to approximate $g(t)$. Under appropriate conditions, we prove that $\widehat{f}_n(u)$ converge in probability, almost surely converge and uniformly almost surely converge. Then we consider asymptotic normality and the the convergence rates of $\widehat{f}_n(u)$. Finally we establish the law of the iterated logarithm for $\widehat{f}_n(u)$.

The paper is organized as follows. In the following we give the assumptions on the X_i and T_i. Section 2 lists some lemmas. Section 3 proves that $\widehat{f}_n(u)$ converge in probability, almost surely converge and uniformly almost surely converge. Section 4 gives the convergence rates of $\widehat{f}_n(u)$. Section 5 obtains asymptotic normality and the law of the iterated logarithm. Simulation results are presented in Section 6. For the convenience and simplicity, we shall employ $C(0 < C < \infty)$ to denote some constant not depending on n but may assume different values at each appearance.

Assume $\{X_i = (x_{i1}, \ldots, x_{ip})^\tau, T_i, Y_i, i = 1, \ldots, n.\}$ satisfy the model (1). Let $W_{ni}(t) = W_{ni}(t; T_1, \ldots, T_n)$ be probability weight functions depending only on the design points $T_1, \ldots, T_n$. Denote $\widetilde{\mathbf{X}}^\tau = (\widetilde{X}_1, \ldots, \widetilde{X}_n)$, $\widetilde{X}_i = X_i - \sum_{j=1}^n W_{nj}(T_i)X_j$, and $\widetilde{\mathbf{Y}} = (\widetilde{Y}_1, \ldots, \widetilde{Y}_n)^\tau$, $\widetilde{Y}_i = Y_i - \sum_{j=1}^n W_{nj}(T_i)Y_j$.

If β were known, we could take $g_n(t) = \sum_{j=1}^n W_{nj}(t)(Y_j - X_j^\tau\beta)$ as the estimator of $g(t)$. Generally we can take $W_{nj}(t)$ as Nadaraya-Watson kernel. Surveys of nonparametric methods can be found in Härdle (1990), which gives extensive discussions of various statistical estimation. Base on the modified model $\{Y_i = X_i^\tau\beta + g_n(T_i) + \varepsilon_i, \quad i = 1, \ldots, n\}$ with $g_n(t)$, we get the estimator $\widehat{\beta}_n$ of β

$$\widehat{\beta}_n = (\widetilde{\mathbf{X}}^\tau\widetilde{\mathbf{X}})^{-1}\widetilde{\mathbf{X}}^\tau\widetilde{\mathbf{Y}}$$

and "true" estimator $\widehat{g}_n(t)$ of $g(t)$

$$\widehat{g}_n(t) = \sum_{i=1}^n W_{nj}(t)(Y_i - X_i^\tau\widehat{\beta}_n)^2$$

Set $\widehat{\varepsilon}_i = Y_i - X_i^\tau\widehat{\beta}_n - \widehat{g}_n(T_i)$ for $i = 1, \ldots, n$. Define the estimators of $f(u)$ as follows,

$$\widehat{f}_n(u) = \frac{1}{2na_n}\sum_{i=1}^n I_{(u-a_n \le \widehat{\varepsilon}_i \le u+a_n)}, \quad u \in R^1 \tag{2}$$

where $a_n(> 0)$ is a bandwidth, and I_A denotes the indicator function of the set A.

In the following we list the sufficient conditions for our main result.

Condition 1. *There exist functions $h_j(\cdot)$ defined on $[0, 1]$ such that*

$$x_{ij} = h_j(T_i) + u_{ij} \quad 1 \le i \le n, \quad 1 \le j \le p$$

where u_{ij} is a sequence of real numbers such that

$$B = \lim_{n\to\infty} \frac{1}{n}\sum_{i=1}^{n} u_i u_i^\tau$$

is a positive matrix, and

$$\lim_{n\to\infty} n^{-1/2}\log^{-1} n \left|\sum_{i=1}^{n} u_{ij}\right| < \infty \quad \textit{holds for all } 1 \le j \le p,$$

for $u_i = (u_{i1}, \ldots, u_{ip})^\tau$.

Condition 2. *$g(\cdot)$ and $h_j(\cdot)$ are Lipschitz continuous of order 1.*

Condition 3. *Weight functions $W_{ni}(\cdot)$ satisfy:*

$$(i) \quad \max_{1\le i\le n} \sum_{j=1}^{n} W_{ni}(T_j) = O(1),$$

$$(ii) \quad \sup_t \max_{1\le i\le n} W_{ni}(t) = O(h_n), \quad h_n = n^{-2/3},$$

$$(iii \quad \sup_t \sum_{j=1}^{n} W_{nj}(t) I_{(|T_j - t| > c_n)} = O(c_n), \quad c_n = n^{-1/2},$$

$$(iv) \quad \max_{1\le i\le n} |W_{ni}(t) - W_{ni}(s)| \le M_0|s-t| \quad \textit{for } s,t \in [0,1]$$

and some positive number $M_0 > 0$.

Remark. Here we give two weight functions satisfied Condition 3 to explain the reasonability of Condition 3.

$$W_{ni}^{(1)}(t) = \frac{1}{h_n}\int_{s_{i-1}}^{s_i} K\Big(\frac{t-s}{h_n}\Big)ds, \text{ or } \quad W_{ni}^{(2)}(t) = K\Big(\frac{t-T_i}{h_n}\Big)\Big/\sum_{j=1}^{n} K\Big(\frac{t-T_j}{h_n}\Big),$$

where $s_i = \frac{1}{2}(T_i + T_{i-1}), i = 1, \ldots, n-1, s_0 = 0, s_n = 1$. $K(\cdot)$ is Parzen-Rosenblatt kernel function, h_n is a bandwidth parameter.

2 SOME LEMMAS

In this section, we list some lemmas proved which are used in the following proving main results. First we give an exponential inequality for bounded independent random variables.

Lemma 1 (Bernstein inequality). *Let $Z_1, \ldots, Z_n$ be independent r.v's satisfying $P\{|Z_i| \le m\} = 1$, each i, where $m < \infty$. Then, for $\eta > 0$,*

$$P\Big\{\Big|\sum_{i=1}^{n} Z_i\Big| \ge n\eta\Big\} \le 2\exp\Big\{-n^2\eta^2/[2\sum var(Z_i) + \frac{2}{3}mn\eta]\Big\}$$

for all $n = 1, \ldots$.

Lemma 2 (Gao, et al. (1995)). *Suppose Conditions 1-3 hold. If $E|\varepsilon_1|^3 < \infty$ and $\max_i \sum_{j=1}^{p} u_{ij}^2 \le C_0 < \infty$. Then*

$$\sup_t |\widehat{g}_n(t) - g(t)| = O_p(n^{-1/3}\log n),$$

and

$$\limsup_{n\to\infty}\Big(\frac{n}{\log\log n}\Big)^{1/2}\|\widehat{\beta}_n-\beta\|<\infty \quad a.s.$$

Where and below we denote Euclidean norm by $\|\cdot\|$.

Lemma 3 (See Devroye, et al. (1980)) *Let* μ_n *and* μ *be 1-dimensional empirical distribution and theoretical distribution, respectively,* $a>0$ *and* Ia *be an interval with length* a. *Then for any* $\zeta>0$, $0<b\le 1/4$ *and* $n\ge\max\{1/b, 8b/\zeta^2\}$,

$$P\Big(\sup\{|\mu_n(Ia)-\mu(Ia)|: 0<\mu(Ia)\le b\}\ge\zeta\Big) \le 16n^2\exp\{-n\zeta^2/(64b+4\zeta)\} +8n\exp\{-nb/10\}.$$

3 CONSISTENCY

In this section we shall prove that $\widehat{f}_n(u)$ converge in probability, almost surely converges and uniformly almost surely converge. Below we always denote

$$f_n(u)=\frac{1}{2na_n}\sum_{i=1}^{n}I_{(u-a\le\varepsilon_i\le u+a_n)},$$

for fixed $u\in C(f)$, where $C(f)$ in the set of continuous points of f.

Theorem 3.1. *There exists a* $M>0$ *such that* $\|X_i\|\le M$ *for* $i=1\sim n$. *Under the assumptions of lemma 2. If*

$$0<a_n\to 0,\quad n^{1/3}a_n\log^{-1}n\to\infty.$$

Then $\widehat{f}_n(u)\to f(u)$ *in probability as* $n\to\infty$.

Proof. Simply calculation shows that the mean of $f_n(u)$ converge to $f(u)$, and its variance does to 0. This implies that $f_n(u)\to f(u)$ in probability as $n\to\infty$.

Now, we prove $\widehat{f}_n(u)-f_n(u)\to 0$ in probability.

If $\varepsilon_i<u-a_n$, then $\widehat{\varepsilon}_i\in(u-a_n,u+a_n)$ implies that $u-a_n+X_i^T(\widehat{\beta}_n-\beta)+(\widehat{g}_n(T_i)-g(T_i))<\varepsilon_i<u-a_n$. If $\varepsilon_i>u+a_n$, then $\widehat{\varepsilon}_i\in(u-a_n,u+a_n)$ implies that $u+a_n<\varepsilon_i<u+a_n+X_i^T(\widehat{\beta}_n-\beta)+(\widehat{g}_n(T_i)-g(T_i))$. Write

$$C_{ni}=X_i^T(\widehat{\beta}_n-\beta)+(\widehat{g}_n(T_i)-g(T_i)) \quad \text{for } i=1,\dots,n.$$

It follows from lemma 2 that, for any $\zeta>0$, there exists a $\eta_0>0$ such that

$$P\{n^{1/3}\log^{-1}n\sup_i|C_{ni}|>\eta_0\}\le\zeta$$

The above arguments yield that

$$\begin{aligned} |\widehat{f}_n(u) - f_n(u)| &\leq \frac{1}{2na_n} I_{(u\pm a_n - |C_{ni}| \leq \varepsilon_i \leq u \pm a_n)} + \frac{1}{2na_n} I_{(u\pm a_n \leq \varepsilon_i \leq u \pm a_n + |C_{ni}|)} \\ &\stackrel{\text{def}}{=} I_{1n} + I_{2n}, \end{aligned}$$

where

$$I_{(u\pm a_n - |C_{ni}| \leq \varepsilon_i \leq u \pm a_n)} = I_{(u+a_n - |C_{ni}| \leq \varepsilon_i \leq u+a_n) \cup (u-a_n - |C_{ni}| \leq \varepsilon_i \leq u-a_n)}.$$

We complete the proof of the theorem by dealing with I_{1n} and I_{2n}. For any $\zeta' > 0$ and large enough n,

$$\begin{aligned} P\{I_{1n} > \zeta'\} &\leq \zeta + P\{I_{1n} > \zeta', \sup_i |C_{ni}| \leq \eta_0\} \\ &\leq \zeta + P\Big(\sum_{i=1}^{n} I_{(u\pm a_n - C\eta_0 n^{-1/3}\log n \leq \varepsilon_i \leq u \pm a_n)} \geq 2na_n\zeta'\Big) \end{aligned}$$

According to the continuity of f on u, using Chebyshev's inequality we know the second term above is less than

$$\frac{1}{2a_n\zeta'} P\Big(u \pm a_n - C\eta_0 n^{-1/3}\log n \leq \varepsilon_i \leq u \pm a_n\Big) = C\eta_0 \log n \Big/ (2\zeta' n^{1/3} a_n)(f(u) + o(1)).$$

It follows from $a_n n^{1/3} \log^{-1} n \to \infty$ that

$$\limsup_{n\to\infty} P\{I_{1n} > \zeta'\} \leq \zeta.$$

Since ζ is arbitrary, we obtain $I_{1n} \to 0$ in probability as $n \to \infty$. We can similarly prove that I_{2n} tends to zero in probability as $n \to \infty$. Thus, we complete the proof of Theorem 3.1.

Theorem 3.2 *Under the assumptions of Theorem 3.1. If*

$$0 < a_n \to 0, \quad n^{1/3} a_n \log^{-2} n \to \infty. \tag{3}$$

Then $\widehat{f}_n(u) \to f(u)$ *for* $u \in C(f)$ *a.s. as* $n \to \infty$.

Proof. Set $f_n^E(u) = Ef_n(u)$, for $u \in C(f)$. Using the continuity of f on u and $a_n \to 0$ we can show that

$$f_n^E(u) \to f(u) \quad \text{as } n \to \infty \tag{4}$$

Now let us consider $f_n(u) - f_n^E(u)$.

$$\begin{aligned} f_n(u) - f_n^E(u) &= \frac{1}{2na_n} \sum_{i=1}^{n} \Big\{ I_{(u-a_n \leq \varepsilon_i \leq u+a_n)} - EI_{(u-a_n \leq \varepsilon_i \leq u+a_n)} \Big\} \\ &\stackrel{\text{def}}{=} \frac{1}{2na_n} \sum_{i=1}^{n} U_{ni}. \end{aligned}$$

Then $U_{n1}, \ldots, U_{nn}$ are independent with $EU_{ni} = 0$, and $|U_{ni}| \leq 1$, moreover

$$\text{var}(U_{ni}) \leq P(u - a_n \leq \varepsilon_i \leq u + a_n) = 2a_n f(u)(1 + o(1)) \leq 4a_n f(u),$$

for large enough n. It follows from lemma 1 that, for any $\zeta > 0$,

$$\begin{aligned} P\{|f_n(u) - f_n^E(u)| \geq \zeta\} &= P\{|\sum_{i=1}^{n} U_{ni}| \geq 2na_n\zeta\} \\ &\leq 2\exp\{-4n^2a_n^2\zeta^2 \Big/ [8na_nf(u) + 4/3na_n\zeta]\} \\ &= 2\exp\{-3na_n\zeta^2 \Big/ [6f(u) + \zeta]\}. \end{aligned} \tag{5}$$

Condition (3) and Borel-Cantelli lemma imply

$$f_n(u) - f_n^E(u) \to 0 \quad a.s. \tag{6}$$

In the following, we shall prove

$$\widehat{f}_n(u) - f_n(u) \to 0 \quad a.s. \tag{7}$$

According to lemma 2, we have with probability one that

$$\begin{aligned} |\widehat{f}_n(u) - f_n(u)| &\leq \frac{1}{2na_n} I_{(u\pm a_n - Cn^{-1/3}\log n \leq \varepsilon_i \leq u \pm a_n)} + \frac{1}{2na_n} I_{(u\pm a_n \leq \varepsilon_i \leq u \pm a_n + Cn^{-1/3}\log n)} \\ &\stackrel{\text{def}}{=} J_{1n} + J_{2n}. \end{aligned} \tag{8}$$

Denote

$$f_{n1}(u) = \frac{1}{2a_n} P(u \pm a_n - Cn^{-1/3}\log n \leq \varepsilon_i \leq u \pm a_n). \tag{9}$$

Then $f_{n1}(u) \leq Cf(u)(n^{1/3}a_n)^{-1}\log n$, for large enough n. By the condition (3), we obtain

$$f_{n1}(u) \to 0, \quad \text{as } n \to \infty. \tag{10}$$

Now let us deal with $J_{n1} - f_{n1}(u)$. Set

$$Q_{ni} = I_{(u\pm a_n - Cn^{-1/3}\log n \leq \varepsilon_i \leq u \pm a_n)} - P(u \pm a_n - Cn^{-1/3}\log n \leq \varepsilon_i \leq u \pm a_n),$$

for $i = 1, \ldots, n$. Then $Q_{n1}, \ldots, Q_{nn}$ are independent, and $|Q_{ni}| \leq 1$, $EQ_{ni} = 0$, and

$$\mathrm{Var}(Q_{ni}) \leq 2Cn^{-1/3}(\log n)f(u)$$

By lemma 1, we have

$$\begin{aligned} P\{|J_{n1} - f_{n1}(u)| > \zeta\} &= P\{|\sum_{i=1}^{n} Q_{ni}| > \zeta\} \\ &\leq 2\exp\{-Cna_n\zeta^2 \Big/ (n^{-1/3}a_n^{-1}f(u)\log^{-1} n + \zeta)\} \\ &\leq 2\exp\{-Cna_n\zeta\}. \end{aligned} \tag{11}$$

Employing Borel-Cantelli lemma we conclude that

$$J_{n1} - f_{n1}(u) \to 0 \quad a.s.$$

Combining (10) with the above conclusion, we obtain $J_{n1} \to 0$ a.s. Similar argument yields $J_{n2} \to 0$ a.s. Moreover, (8) implies (7). From (4), (6) and (7), we complete the proof of Theorem 3.2.

Theorem 3.3. *Under the assumptions of Theorem 3.2. If f is uniformly continuous on R^1 and*

$$0 < a_n \to 0, \quad n^{1/3} a_n \log^{-2} n \to \infty. \tag{12}$$

Then $\sup_u |\widehat{f}_n(u) - f(u)| \to 0$ *a.s.*

Proof. We still use the notations in the proof of Theorem 3.2 denote the empirical distribution of $\varepsilon_1, \ldots, \varepsilon_n$ by μ_n and the distribution of ε_1 by μ. Since f is uniformly continuous, thus $\sup_u f(u) = f_0 < \infty$. It is easy to show

$$\sup_u |f(u) - f_n^E(u)| \to 0 \quad \text{as } n \to \infty \tag{13}$$

Write

$$f_n(u) - f_n^E(u) = \frac{1}{2a_n}\{\mu_n([u - a_n, u + a_n]) - \mu([u - a_n, u + a_n])\}$$

and denote $b_n^* = 2f_0 a_n$, $\zeta_n = 2a_n\zeta$ for any $\zeta > 0$. Then for large enough n, $0 < b_n^* < 1/4$ and $\sup_u \mu([u - a_n, u + a_n]) \le b_n^*$ for all n. From lemma 3, we have, for large enough n,

$$\begin{aligned} P\{\sup_u |f_n(u) - f_n^E(u)| \ge \zeta\} &= P\{\sup_u |\mu_n([u - a_n, u + a_n]) - \mu([u - a_n, u + a_n])| \ge 2a_n\zeta\} \\ &\le 16n^2 \exp\{-na_n^2\zeta^2 \Big/ (32f_0 a_n + 2a_n\zeta)\} + 8n \exp\{-na_n^2 f_0/5\}. \end{aligned}$$

From (12) and Borel-Cantelli lemma, it follows that

$$\sup_u |f_n(u) - f_n^E(u)| \to 0 \quad \text{a.s.} \tag{14}$$

Combining (14) with (13) we obtain

$$\sup_u |f_n(u) - f(u)| \to 0 \quad \text{a.s.} \tag{15}$$

In the following we shall prove that

$$\sup_u |\widehat{f}_n(u) - f_n(u)| \to 0 \quad \text{a.s.} \tag{16}$$

It is obvious that $\sup_u |f_{n1}(u)| \to 0$, as $n \to \infty$. Set $d_n = f_0 n^{-1/3} \log n$. For large enough n, we have $0 < d_n < 1/4$ and

$$\sup_u \mu\{(u \pm a_n - Cn^{-1/3} \log n, u \pm a_n)\} \le Cd_n \text{ for all } n.$$

It follows from lemma 3 that

$$\begin{aligned} P(\sup_u |J_{n1} - f_{n1}(u)| > \zeta) &\leq P[|\mu_n\{(u \pm a_n - Cn^{-1/3}\log n, u \pm a_n)\} \\ &\quad -\mu\{(u \pm a_n - Cn^{-1/3}\log n, u \pm a_n)\}| \geq 2a_n\zeta] \\ &\leq 16n^2 \exp\Big(-\frac{4na_n^2\zeta^2}{64f_0 n^{-1/3}\log n + 8a_n\zeta}\Big) + 8n\exp(-n^{2/3}\log n/10). \end{aligned}$$

By (12) and the above arguments, it follows that $\sup_u |J_{n1} - f_{n1}(u)| \to 0$ a.s., and hence $\sup_u |J_{n1}| \to 0$ a.s. We have $\sup_u |J_{n2}| \to 0$ similarly. In the proof of Theorem 3.2, it can be shown that, with probability one and for large enough n,

$$\sup_u |\widehat{f}_n(u) - f_n(u)| \leq \sup_u |J_{n1}| + \sup_u |J_{n2}|.$$

This implies (16), and so does the conclusion of Theorem 3.3.

4 CONVERGENCE RATE

Theorem 4.1. *Under the assumptions of Theorem 3.2. If f is locally Lipschitz continuous of order 1 on u. Then for $a_n = n^{-1/6}\log^{1/2} n$,*

$$\widehat{f}_n(u) - f(u) = O(n^{-1/6}\log^{1/2} n). \quad \text{a.s} \tag{17}$$

Proof. The proof is analogous to that for Theorem 3.2 completely. By the assumption of Theorem 4.1, there exist $c_0 > 0$ and $\delta_1 = \delta_1(u) > 0$ such that $u' \in (u - \delta_1, u + \delta_1)$ implying $|f(u') - f(u)| \leq c_0|u' - u|$. Hence for large enough n,

$$\begin{aligned} |f_n^E(u) - f(u)| &\leq \frac{1}{2a_n}\int_{u-a_n}^{u+a_n} |f(u) - f(u')|du' \\ &\leq c_0 a_n/2 = O(n^{-1/6}\log^{1/2} n). \end{aligned} \tag{18}$$

Since f is bounded on $(u - \delta_1, u + \delta_1)$, we have for large enough n,

$$\begin{aligned} f_{n1}(u) &= \frac{1}{2a_n}P(u \pm a_n - Cn^{-1/3}\log n \leq \varepsilon_i \leq u \pm a_n) \\ &\leq Cn^{-1/3}a_n^{-1}\log n \sup_{u' \in (u-\delta_1, u+\delta_1)} f(u') \\ &= O(n^{-1/6}\log^{1/2} n). \end{aligned}$$

Replacing ζ by $\zeta_n = \zeta n^{-1/6}\log^{1/2} n$ in (5), then for large enough n,

$$P(|f_n(u) - f_n^E(u)| \geq 2\zeta n^{-1/6}\log^{1/2} n) \leq 2\exp\{-3n^{1/2}\log^{3/2} n\zeta \Big/ (6f_0 + \zeta)\}.$$

Where $f_0 = \sup_{u' \in (u-\delta_1, u+\delta_1)} f(u')$. Instead of (15), we have

$$f_n(u) - f_n^E(u) = O(n^{-1/6}\log^{1/2} n). \quad \text{a.s.} \tag{19}$$

The similar argument as (11) yields

$$P\{|J_{n1} - f_{n1}(u)| > \zeta n^{-1/6}\log^{1/2} n\} \quad \leq \quad 2\exp(-Cn^{2/3}\log^{1/2} n).$$

Hence, $J_{n1} - f_{n1}(u) = O(n^{-1/6}\log^{1/2} n)$ a.s. (18) and (19) imply that we have proved

$$f_n(u) - f(u) = O(n^{-1/6}\log^{1/2} n). \quad \text{a.s.}$$

Using the arguments below (8) in the proof of Theorem 3.2, the proof is completed.

5 ASYMPTOTIC NORMALITY AND LAW OF THE ITERATED LOGARITHM

Theorem 5.1. *Under the assumptions of Theorem 3.2. If f is locally Lipschitz continuous of order 1 on u and*

$$0 < na_n^3 \to 0, \quad n^{5/12}a_n\log^{-1} n \to \infty.$$

Then

$$\sqrt{2na_n/f(u)}\{\widehat{f}_n(u) - f(u)\} \to N(0,1) \quad \textit{in distribution as } n \to \infty.$$

Theorem 5.2. *Under the assumptions of Theorem 3.2. If f is locally Lipschitz continuous of order 1 on u and*

$$\lim_{n\to\infty}(na_n^3/\log\log n) = 0, \quad \lim_{n\to\infty}(n^{1/2}a_n\log\log n\log^{-2} n) = \infty.$$

Then

$$\limsup_{n\to\infty} \pm\Big\{\frac{na_n}{f(u)\log\log n}\Big\}^{1/2}\{\widehat{f}_n(u) - f(u)\} = 1, \quad a.s.$$

The proofs of the above two theorems can be completed by slightly modifying the proofs of theorems 2 and 3 of Chai and Li(1993), we omitted the details.

6 SIMULATION

In order to evaluate the practical performance of the estimate given in (2), in this section we present the results of a small scale simulation study of the estimators $\widehat{f}_n(u)$. Although it is not possible to completely characterize the sampling behavior of the estimator under general situations, the result presented below is suggestive.

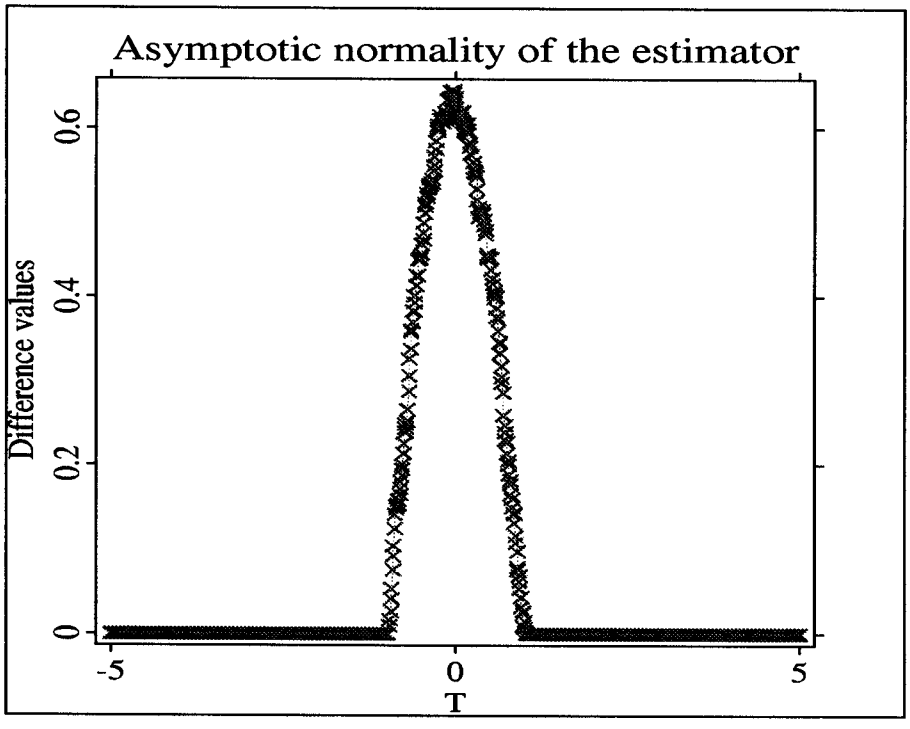


Figure 1: *The behavior of asymptotic normality of* $\widehat{f}_n(u)$.

The dependent variable Y_i was generated from an underlying partially linear model

$$Y_i = X_i^T\beta + \sin(\pi T_i) + \epsilon_i, \qquad i = 1, \ldots, n = 300$$

with true value $\beta = (1, 0.75)^T$. Both of X_i and T_i were generated as uniform on $[0, 1]$, while ϵ_i was standard normal distributed. The simulation number is 500.

The main point of the simulation is to investigate the difference of the estimator $\widehat{f}_n(u)$ with real density $\phi(\cdot)$, the standard normal distribution function. We only consider their behavior on $[-5, 5]$. In procedure of simulation, we select bandwidth by using Cross-validation method.

Figure 1 presents the results for $\widehat{f}_n(u) - f(u)$; while Figure 2 presents the result for $\sqrt{2na_n/\{f(u)\}}\{\widehat{f}_n(u) - f(u)\}$, in which we take $a_n = n^{-1/3}\log^{-1} n$. The simulation results indicate that the estimator is very close to the true function, and $\sqrt{2na_n/\{f(u)\}}\{\widehat{f}_n(u) - f(u)\}$ just like normal distribution.

REFERENCES

Bickel, P.J. (1982). On Adaptive Estimation. *Annals of Statistics*, **10** 647-671.

Chai, G. X. and Li, Z.Y. (1993). Asymptotic Theory for Estimation of Error Distributions in Linear Model. *Science in China, Ser. A* **4** 408-419.

Chen, H. (1988). Convergence Rates for Parametric Components in a Partly Linear Model. *Annals of Statistics*, **16** 136-146.

Chen, H. and Shiau, J.G. (1991). A Two-Stage Spline Smoothing Method for Partially Linear Models. *Journal of Statistical Planning & Inference*, **25** 187-201.

Cuzick, J. (1992a). Semiparametric Additive Regression. *Journal of the Royal Statistical Society, Series B*, **54** 831-843.

Cuzick, J. (1992b). Efficient Estimates in Semiparametric Additive Regression Models with Unknown Error Distribution. *Annals of Statistics*, **20** 1129-1136.

Devroye, L. P. and Wagneer, T.J. (1980). The Strong Uniform Consistency of Kernel Estimates. *Journal of Multivariate Analysis*, **5** 59-77.

Engle, R. F., Granger, C.W.J., Rice,J. and Weiss, A. (1986). Semiparametric Estimates of the Relation Between Weather and Electricity Sales. *Journal of the American Statistical Association*, **81** 310-320.

Gao, J. T., Hong, S.Y. and Liang, H. (1995). Convergence rates of a class of estimates in partly linear models. *Acta Math. Sinica*, **38**, 658-669.

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

Heckman, N.E. (1986). Spline Smoothing in Partly Linear Models. *Journal of the Royal Statistical Society, Series B*, **48** 244-248.

Heckman, N.E.(1988). Minimax Estimates in a Semiparametric Model. *Journal of the American Statistical Association*, **83** 1090-1096.

Robinson, P.M. (1988). Root-N-Consistent Semiparametric Regression. *Econometrica*, **56** 931-954.

Schick, A. (1993). On Efficient Estimation in Regression Models. *Annals of Statistics*, **21** 1486-1521.

Schick, A. (1996). Root-N Consistent Estimation in Partly Linear Regression Models. *Statistics & Probability Letters*, **28** 353-358.

Serfling, R.J. (1980).*Approximation Theorems of Mathematical Statistics.* John Wiley & Sons 95-96.

Speckman, P. (1988). Kernel Smoothing in Partial Linear Models. *Journal of the Royal Statistical Society, Series B*, **50** 413-436.

The Annals Of Statistics (1999), Vol. 27, No 5, pp. 1519-1535

ESTIMATION IN A SEMIPARAMETRIC PARTIALLY LINEAR ERRORS-IN-VARIABLES MODEL

Hua Liang[1], Wolfgang Härdle [2] and Raymond J. Carroll[3]

Chinese Academy of Sciences, Humboldt-Universität zu Berlin and Texas A&M University

We consider the partially linear model relating a response Y to predictors (X, T) with mean function $X^{\mathrm{T}}\beta+g(T)$ when the X's are measured with additive error. The semiparametric likelihood estimate of Severini and Staniswalis (1994) leads to biased estimates of both the parameter β and the function $g(\cdot)$ when measurement error is ignored. We derive a simple modification of their estimator which is a semiparametric version of the usual parametric correction for attenuation. The resulting estimator of β is shown to be consistent and its asymptotic distribution theory is derived. Consistent standard error estimates using sandwich–type ideas are also developed.

Key Words and Phrases: Errors-in-Variables; Measurement Error; Nonparametric Likelihood; Orthogonal Regression; Partially Linear Model; Semiparametric Models; Structural Relations.

Short title: Partially Linear Models and Measurement Error.

AMS 1991 subject classification: Primary: 62J99, 62H12, 62E25, 62F10 Secondary: 62H25, 62F10, 62F12, 60F05.

[1]Supported by an award from the Alexander von Humboldt Foundation, and by Sonderforschungsbereich 373 "Quantifikation und Simulation Ökonomischer Prozesse".

[2]Supported by Sonderforschungsbereich 373 "Quantifikation und Simulation Ökonomischer Prozesse".

[3]Supported by a grant from the National Cancer Institute (CA–57030), by the Texas A&M Center for Environmental and Rural Health via a grant from the National Institute of Environmental Health Sciences (P30-ESO9106), and by Sonderforschungsbereich 373 "Quantifikation und Simulation Ökonomischer Prozesse".

The Annals Of Statistics (1999), Vol. 27, No 5, pp. 1519-1535

1 INTRODUCTION AND BACKGROUND

Consider the semiparametric partially linear model based on a sample of size n,

$$Y_i = X_i^{\mathrm{T}}\beta + g(T_i) + \epsilon_i, \tag{1}$$

where X_i is a possibly vector–values covariate, T_i is a scalar covariate, the function $g(\cdot)$ is unknown, and the model errors ϵ_i are independent with conditional mean zero given the covariates. The partially linear model was introduced by Engle, et al. (1986) to study the effect of weather on electricity demand, and further studied by Heckman (1986), Chen (1988), Speckman (1988), Cuzick (1992a,b), Liang & Härdle (1997) and Severini & Staniswalis (1994).

We are interested in the estimation of the unknown parameter β and the unknown function $g(\cdot)$ in model (1) when the covariates X_i are measured with error. Instead of observing X_i, we observe

$$W_i = X_i + U_i, \tag{2}$$

where the measurement errors U_i are independent and identically distributed, independent of (Y_i, X_i, T_i), with mean zero and covariance matrix Σ_{uu}. We will assume that Σ_{uu} is known, taking up the case that it is estimated in section 5. The measurement error literature has been surveyed by Fuller (1987) and Carroll, et al. (1995).

If the X's are observable, estimation of β at ordinary rates of convergence can be obtained by a local–likelihood algorithm, as follows. For every fixed β, let $\widehat{g}(T, \beta)$ be an estimator of $g(T)$. For example, in the Severini and Staniswalis implementation, $\widehat{g}(T, \beta)$ maximizes a weighted likelihood assuming that the model errors ϵ_i are homoscedastic and normally distributed, with the weights being kernel weights with symmetric kernel density function $K(\cdot)$ and bandwidth h. Having obtained $\widehat{g}(T, \beta)$, β is estimated by a least squares operation:

$$\text{minimize} \sum_{i=1}^{n} \left\{ Y_i - X_i^{\mathrm{T}}\beta - \widehat{g}(T_i, \beta) \right\}^2.$$

In this particular case, the estimate for β can be determined explicitly. Let $\widehat{g}_{y,h}(\cdot)$ and $\widehat{g}_{x,h}(\cdot)$ be the kernel regressions with bandwidth h of Y and X on T, respectively. Then

$$\widehat{\beta}_x = \left[\sum_{i=1}^{n} \{X_i - \widehat{g}_{x,h}(T_i)\} \{X_i - \widehat{g}_{x,h}(T_i)\}^{\mathrm{T}} \right]^{-1} \sum_{i=1}^{n} \{X_i - \widehat{g}_{x,h}(T_i)\} \{Y_i - \widehat{g}_{y,h}(T_i)\}. \tag{3}$$

One of the important features of the estimator (3) is that it does not require undersmoothing, so that bandwidths of the usual order $h \sim n^{-1/5}$ lead to the result

$$n^{1/2}(\widehat{\beta}_n - \beta) \Rightarrow \text{Normal}(0, B^{-1}CB^{-1}), \tag{4}$$

Carroll, R., Härdle, W. and Liang, H. (1999), Estimation In A Semiparametric Partially Linear Errors In Variables Model

where B is the covariance matrix of $X - E(X|T)$ and C is the covariance matrix of $\epsilon\{X - E(X|T)\}$.

The least squares form of (3) can be used to show that if one ignores the measurement error and replaces X by W, the resulting estimate is inconsistent for β. The form though suggests even more. It is well–known that in linear regression, inconsistency caused by the measurement error can be overcome by applying the so–called "correction for attenuation". In the context of semiparametric models, this suggests that we use the estimator

$$\widehat{\beta}_n = \left[\sum_{i=1}^{n} \{W_i - \widehat{g}_{w,h}(T_i)\} \{W_i - \widehat{g}_{w,h}(T_i)\}^{\mathrm{T}} - n\Sigma_{uu}\right]^{-1} \sum_{i=1}^{n} \{W_i - \widehat{g}_{w,h}(T_i)\} \{Y_i - \widehat{g}_{y,h}(T_i)\}. \quad (5)$$

The estimator (5) can be derived in much the same way as the Severini–Staniswalis estimator. For every β, let $\widehat{g}(T, \beta)$ maximize the weighted likelihood ignoring the measurement error. Then form the estimators of β via a negatively penalized operation:

$$\text{minimize} \sum_{i=1}^{n} \left\{Y_i - W_i^{\mathrm{T}}\beta - \widehat{g}(T_i, \beta)\right\}^2 - \beta^{\mathrm{T}}\Sigma_{uu}\beta. \quad (6)$$

The negative sign in the second term in (6) looks odd until one remembers that the effect of the measurement error is attenuation, i.e., to underestimate β in absolute value when it is scalar, and thus one must correct for attenuation by making β larger, not by shrinking it further towards zero.

In this paper, we analyze the estimate (5), and show that it is consistent, asymptotically normally distributed with a variance different from (4). Just as in the Severini–Staniswalis algorithm, the kernel weight with ordinary bandwidths of order $h \sim n^{-1/5}$ may be used.

The outline of the paper is as follows. In Section 2, we define the weighting scheme to be used and hence the estimators of β and $g(\cdot)$. Section 3 is the statement of the main results for β, while the results for $g(\cdot)$ are stated in Section 4. Section 5 states the corresponding results when the measurement error variance Σ_{uu} is estimated. Section 6 gives a numerical illustration. Final remarks are given in Section 7. All proofs are delayed until the appendix.

2 DEFINITION OF THE ESTIMATORS

For technical convenience we will assume that the T_i are confined to the interval $[0, 1]$. Throughout, we shall employ $C(0 < C < \infty)$ to denote some constant not depending on n, but which may assume different values at each appearance. In our proofs and statement of results, we will let the X's be independent random variables.

Let $\omega_{ni}(t) = \omega_{ni}(t; T_1, \ldots, T_n)$ be weight functions depending only on the design points $T_1, \ldots, T_n$.

Carroll, R., Härdle, W. and Liang, H. (1999), Estimation In A Semiparametric Partially Linear Errors In Variables Model

For example

$$\omega_{ni}(t) = \frac{1}{h_n}\int_{s_{i-1}}^{s_i} K\left(\frac{t-s}{h_n}\right)ds \quad 1 \le i \le n \tag{7}$$

where $s_0 = 0$, $s_n = 1$ and $s_i = (1/2)(T_{(i)} + T_{(i+1)})$, $1 \le i \le n-1$, $T_{(i)}$ are the order statistics of T_i, h_n is a sequence of bandwidth parameters which tends to zero as $n \to \infty$ and $K(\cdot)$ is a nonnegative kernel function, which is supported to have compact support and to satisfy

$$\text{supp}(K) = [-1,1], \sup|K(x)| \le C < \infty, \int K(u)du = 1 \text{ and } K(u) = K(-u).$$

In the paper, for any a sequence of variables or functions $(S_1,\ldots,S_n)$, we always denote $\mathbf{S}^{\mathrm{T}} = (S_1,\ldots,S_n)$, $\widetilde{S}_i = S_i - \sum_{j=1}^n \omega_{nj}(T_i)S_j$, $\widetilde{\mathbf{S}}^{\mathrm{T}} = (\widetilde{S}_1,\ldots,\widetilde{S}_n)$. For example, $\widetilde{\mathbf{W}}^{\mathrm{T}} = (\widetilde{W}_1,\ldots,\widetilde{W}_n)$, $\widetilde{W}_i = W_i - \sum_{j=1}^n \omega_{nj}(T_i)W_j$; $\widetilde{g}_i = g(T_i) - \sum_{k=1}^n \omega_{nk}(T_i)g(T_k)$, $\widetilde{\mathbf{G}} = (\widetilde{g}_1,\ldots,\widetilde{g}_n)^{\mathrm{T}}$.

The fact that $g(t) = E(Y_i - X_i^{\mathrm{T}}\beta|T=t) = E(Y_i - W_i^{\mathrm{T}}\beta|T=t)$ suggests

$$\widehat{g}_n(t) = \sum_{j=1}^n \omega_{nj}(t)(Y_j - W_j^{\mathrm{T}}\widehat{\beta}_n) \tag{8}$$

as the estimator of $g(t)$.

In some cases, it may be reasonable to assume that the model errors ϵ_i are homoscedastic with common variance σ^2. In this event, since $E\{Y_i - X_i^{\mathrm{T}}\beta - g(T_i)\}^2 = \sigma^2$ and $E\{Y_i - W_i^{\mathrm{T}}\beta - g(T_i)\}^2 = E\{Y_i - X_i^{\mathrm{T}}\beta - g(T_i)\}^2 + \beta^{\mathrm{T}}\Sigma_{uu}\beta$, we define

$$\widehat{\sigma}_n^2 = n^{-1}\sum_{i=1}^n (\widetilde{Y}_i - \widetilde{W}_i^{\mathrm{T}}\widehat{\beta}_n)^2 - \widehat{\beta}_n^{\mathrm{T}}\Sigma_{uu}\widehat{\beta}_n. \tag{9}$$

as the estimator of σ^2.

3 MAIN RESULTS

Let the components of X_i be $X_i = (X_{ij})$. Denote $h_j(T_i) = E(X_{ij}|T_i)$, $V_i = X_i - E(X_i|T_i)$, $1 \le i \le n$, $1 \le j \le p$. We make the following assumptions.

Assumption 1.1. $\sup_{0\le t\le 1} E(\|X_1\|^4|T=t) < \infty$ and $B = E(V_1V_1^{\mathrm{T}})$ is a positive definite matrix.

Assumption 1.2. $g(\cdot)$ and $h_j(\cdot)$ are Lipschitz continuous of order 1.

Assumption 1.3. The weight functions $\omega_{ni}(\cdot)$ satisfy:

$$(i) \qquad \max_{1\le i\le n}\sum_{j=1}^n \omega_{nj}(T_i) = O_P(1),$$

$$(ii) \qquad \max_{1\le i,j\le n} \omega_{ni}(T_j) = O_P(b_n),$$

$$(iii) \qquad \max_{1\le i\le n}\sum_{j=1}^n \omega_{nj}(T_i)I(|T_j - T_i| > c_n) = O_P(c_n),$$

Carroll, R., Härdle, W. and Liang, H. (1999), Estimation In A Semiparametric Partially Linear Errors In Variables Model

where $b_n = n^{-4/5}$, $c_n = n^{-1/5}\log n$.

Assumption 1.4. $E(\epsilon_i) = E(U_i) = 0$ and $\sup_i E(\epsilon_i^4 + \|U_i\|^4) < \infty$.

Our two main results concern the limit distributions of the estimates of β and σ^2.

Theorem 3.1 *Suppose that Assumptions 1.1-1.4 hold. Then $\widehat{\beta}_n$ is an asymptotically normal estimator, i.e.*

$$n^{1/2}(\widehat{\beta}_n - \beta) \Rightarrow N(0, B^{-1}\Gamma B^{-1}),$$

with $\Gamma = E\left[(\epsilon - U^{\mathrm{T}}\beta)\{X - E(X|T)\}\right]^{\otimes 2} + E\{(UU^{\mathrm{T}} - \Sigma_{uu})\beta\}^{\otimes 2} + E(UU^{\mathrm{T}}\epsilon^2)$, *where* $A^{\otimes 2} = AA^{\mathrm{T}}$. *Note that* $\Gamma = E(\epsilon - U^{\mathrm{T}}\beta)^2 B + E\{(UU^{\mathrm{T}} - \Sigma_{uu})\beta\}^{\otimes 2} + \Sigma_{uu}\sigma^2$ *if ϵ is homoscedastic and independent of (X, T).*

Theorem 3.2 *Suppose that the conditions of Theorem 3.1 hold, and that the ϵ's are homoscedastic with variance σ^2, and independent of (X_i, T_i). Then*

$$n^{1/2}(\widehat{\sigma}_n^2 - \sigma^2) \Rightarrow N(0, \sigma_*^2),$$

where $\sigma_*^2 = E\{(\epsilon - U^{\mathrm{T}}\beta)^2 - (\beta^{\mathrm{T}}\Sigma_{uu}\beta + \sigma^2)\}^2$.

Remarks

- It is relatively easy to estimate the covariance matrix of $\widehat{\beta}_n$. Let $\dim(X)$ be the number of the components of X. A consistent estimate of B is just

$$\{n - \dim(X)\}^{-1}\sum_{i=1}^{n}\{W_i - \widehat{g}_{w,h}(T_i)\}^{\otimes 2} - \Sigma_{uu} \stackrel{\text{def}}{=} B_n.$$

  In the general case, one can use (25) below to construct a consistent sandwich–type estimate of Γ, namely

$$n^{-1}\sum_{i=1}^{n}\left\{\widetilde{W}_i(\widetilde{Y}_i - \widetilde{W}_i^{\mathrm{T}}\widehat{\beta}_n) + \Sigma_{uu}\widehat{\beta}_n\right\}^{\otimes 2}.$$

  In the homoscedastic case, namely that ϵ_i is independent of (X_i, T_i, U_i) with variance σ^2, and with U being normally distributed, a different formula can be used. Let $C(\beta) = E\{(UU^{\mathrm{T}} - \Sigma_{uu})\beta\}^{\otimes 2}$. Then a consistent estimate of Γ is

$$(\widehat{\sigma}_n^2 + \widehat{\beta}_n^{\mathrm{T}}\Sigma_{uu}\widehat{\beta}_n)\widehat{B}_n + \widehat{\sigma}_n^2\Sigma_{uu} + C(\widehat{\beta}_n).$$

- In the classical functional model (Kendall and Stuart, 1992), instead of obtaining an estimate of Σ_{uu} through replication, it is instead assumed that the ratio of Σ_{uu} to σ^2 is known. Without loss of generality, we set this ratio equal to the identity matrix. The resulting analogue of the parametric estimators to the partially linear model is to solve the following minimization problem:

$$\sum_{i=1}^{n} \left| \frac{\widetilde{Y}_i - \widetilde{W}_i^{\mathrm{T}}\beta}{\sqrt{1+\|\beta\|^2}} \right|^2 = \min!,$$

where here and in the sequel $\|\cdot\|$ denotes the Euclidean norm. One can use the techniques of this paper to show that this estimator is consistent and asymptotically normally distributed. The asymptotic variance of the estimate of β for the case where ϵ_i is independent of (X_i, T_i) can be shown to be

$$B^{-1}\left[(1+\|\beta\|^2)^2\sigma^2 B + \frac{E\{(\epsilon - U^{\mathrm{T}}\beta)^2\Gamma_1\Gamma_1^{\mathrm{T}}\}}{1+\|\beta\|^2}\right]B^{-1},$$

where $\Gamma_1 = (1+\|\beta\|^2)U + (\epsilon - U^{\mathrm{T}}\beta)\beta$.

4 ASYMPTOTIC RESULTS FOR THE NONPARAMETRIC PART

Theorem 4.1 *Suppose that Assumptions 1.1-1.4 hold and that $\omega_{ni}(t)$ are Lipschitz continuous of order 1 for all $i = 1, \ldots, n$. Then for fixed T_i, the asymptotic bias and asymptotic variance of $\widehat{g}_n(t)$ are respectively, $\sum_{i=1}^n \omega_{ni}(t)g(T_i) - g(t)$ and $\sum_{i=1}^n \omega_{ni}^2(t)(\beta^{\mathrm{T}}\Sigma_{uu}\beta + \sigma^2)$. These are all of order $O(n^{-2/5})$ for the kernel estimators.*

5 ESTIMATED ERROR VARIANCE

Although in some cases the measurement error covariance matrix Σ_{uu} has been established by independent experiments, in others it is unknown and must be estimated. The usual method of doing so (Carroll, et al., 1995, Chapter 3) is by partial replication, so that we observe $W_{ij} = X_i + U_{ij}, \quad j = 1, \ldots m_i$.

For notational convenience, we consider here only the case that $m_i \leq 2$, and assume that a fraction δ of the data has such replicates. Let $\overline{W}_i$ be the sample mean of the replicates. Then a consistent, unbiased method of moments estimate for Σ_{uu} is

$$\widehat{\Sigma}_{uu} = \frac{\sum_{i=1}^n \sum_{j=1}^{m_i}(W_{ij} - \overline{W}_i)(W_{ij} - \overline{W}_i)^{\mathrm{T}}}{\sum_{i=1}^n (m_i - 1)}.$$

The estimator changes only slightly to accommodate the replicates, becoming

$$\widehat{\beta}_n = \left[\sum_{i=1}^{n}\left\{\overline{W}_i - \widehat{g}_{w,h}(T_i)\right\}^{\otimes 2} - n(1-\delta/2)\widehat{\Sigma}_{uu}\right]^{-1}\sum_{i=1}^{n}\left\{\overline{W}_i - \widehat{g}_{w,h}(T_i)\right\}\{Y_i - \widehat{g}_{y,h}(T_i)\}, \quad (10)$$

where $\widehat{g}_{w,h}(\cdot)$ is the kernel regression of the $\overline{W}_i$'s on T_i.

Using the techniques in the appendix, one can show that the limit distribution of (10) is Normal$(0, B^{-1}\Gamma_2 B^{-1})$, with

$$\begin{aligned}\Gamma_2 &= (1-\delta)E\left[(\epsilon - U^{\mathrm{T}}\beta)\{X - E(X|T)\}\right]^{\otimes 2} + \delta E\left[(\epsilon - \overline{U}^{\mathrm{T}}\beta)\{X - E(X|T)\}\right]^{\otimes 2} \\ &\quad +(1-\delta)E\left([\{UU^{\mathrm{T}} - (1-\delta/2)\Sigma_{uu}\}\beta]^{\otimes 2} + UU^{\mathrm{T}}\epsilon^2\right) \\ &\quad +\delta E\left([\{\overline{UU}^{\mathrm{T}} - (1-\delta/2)\Sigma_{uu}\}\beta]^{\otimes 2} + \overline{UU}^{\mathrm{T}}\epsilon^2\right). \end{aligned} \quad (11)$$

In (11), $\overline{U}$ refers to the mean of two U's. In the case that ϵ is independent of (X,T), the sum of the first two terms simplifies to $\{\sigma^2 + \beta^{\mathrm{T}}(1-\delta/2)\Sigma_{uu}\beta\}B$.

Standard error estimates can also be derived. A consistent estimate of B is

$$\widehat{B}_n = \{n - \dim(X)\}^{-1}\sum_{i=1}^{n}\left\{\overline{W}_i - \widehat{g}_{w,h}(T_i)\right\}^{\otimes 2} - (1-\delta/2)\widehat{\Sigma}_{uu}.$$

Estimates of Γ_2 can be also easily developed. In the homoscedastic case with normal errors, the sum of the first two terms can be estimated by $(\widehat{\sigma}_n^2 + (1-\delta/2)\widehat{\beta}_n^{\mathrm{T}}\widehat{\Sigma}_{uu}\widehat{\beta}_n)\widehat{B}_n$. The sum of the last two terms is a deterministic function of $(\beta, \sigma^2, \Sigma_{uu})$, and these estimates are simply substituted into the formula.

A general sandwich–type estimator is developed as follows. Define $\kappa = n^{-1}\sum_{i=1}^{n} m_i^{-1}$, and define

$$R_i = \widetilde{\overline{W}}_i(\widetilde{Y}_i - \widetilde{\overline{W}}_i^{\mathrm{T}}\widehat{\beta}_n) + \widehat{\Sigma}_{uu}\widehat{\beta}_n/m_i + \frac{\kappa}{\delta}(m_i - 1)\left\{\frac{1}{2}(W_{i1} - W_{i2})(W_{i1} - W_{i2})^{\mathrm{T}} - \widehat{\Sigma}_{uu}\right\}.$$

Then a consistent estimate of Γ_2 is the sample covariance matrix of the R_i's.

6 NUMERICAL EXAMPLE

To illustrate our method, we consider data from the Framingham Heart Study. We consider $n = 1615$ males with Y being their average blood pressure in a fixed 2–year period, T being their age and W being the logarithm of the observed cholesterol level, for which there are two replicates.

We do two analyses. In the first, we use both cholesterol measurements, so that in the notation of Section 5, $\delta = 1$. In this analysis, there is not a great deal of measurement error. Thus, in our second analysis, which is given for illustrative purposes, we use only the first cholesterol

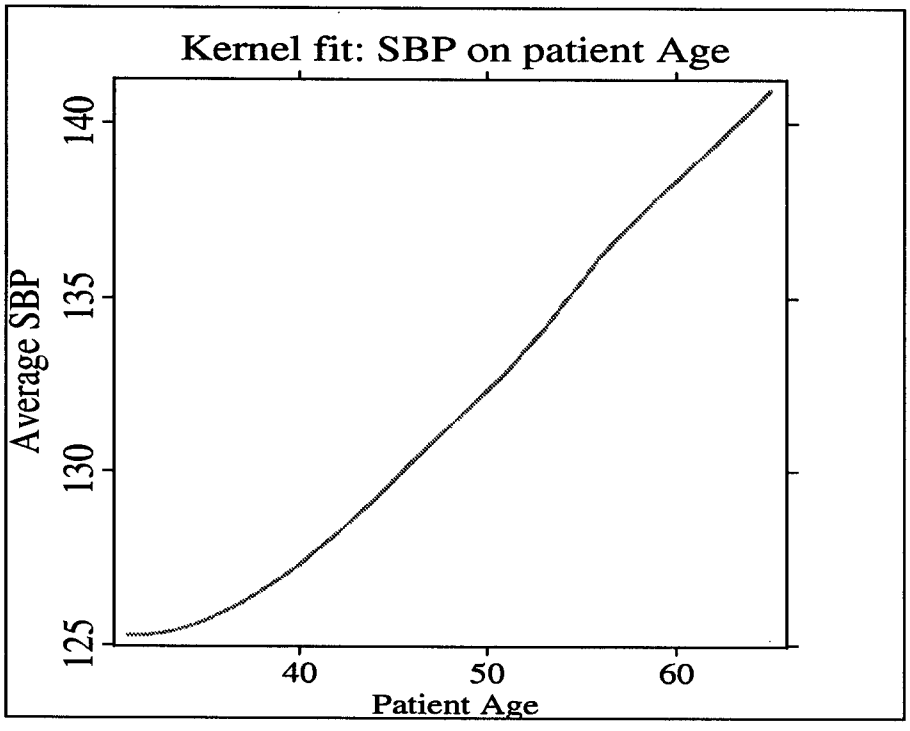


Figure 1: *Estimate of the function $g(T)$ in the Framingham data ignoring measurement error.*

measurement, but fix the measurement error variance at the value obtained in the first analysis, in which case $\delta = 0$. For nonparametric fitting, we chose the bandwidth using crossvalidation to predict the response. In precise terms, we compute the squared error using a geometric sequence of 191 bandwidths ranging in $[1, 20]$. The optimal bandwidth is selected to minimize the squared error among these 191 candidates. An analysis ignoring the measurement error found some curvature in T, see Figure 1 for the estimate of $g(T)$. All calculations were performed in XploRe (Härdle, et al., 1995).

Our results are as follows. First consider the case that the measurement error is estimated, and both cholesterol values are used to estimate Σ_{uu}. The estimator of β ignoring the measurement error is 9.438, with estimated standard error 0.187. When we account for the measurement error, the estimate increased to $\widehat{\beta} = 12.540$, and the standard error increased to 0.195.

In the second analysis, we fix the measurement error variance and use only the first cholesterol value. The estimator of β ignoring the measurement error was 10.744, with estimated standard error 0.492. When we account for the measurement error, the estimate increased to $\widehat{\beta} = 13.690$, and the standard error increased to 0.495.

7 DISCUSSION

The nonparametric regression estimator (8) is based on locally weighted averages. Clearly, results such as Theorem 3.1 should apply if (8) is replaced by a locally linear kernel regression estimator, or by a spline estimator, although our proofs do not apply to these estimators.

We have treated the case that the parametric part X of the model has measurement error and the nonparametric part T is measured exactly. An interesting problem is to interchange the roles of X and T, so that the parametric part is measured exactly and the nonparametric part is measured with error, i.e., $E(Y|X,T) = \theta T + g(X)$. Fan and Truong (1993) have shown in this case that with normally distributed measurement error, the nonparametric function $g(\cdot)$ can be estimated only at logarithmic rates, and not with rate $n^{-2/5}$. We conjecture even so that θ can be estimated at parametric rates, but this remains an open problem.

Acknowledgments. *The authors are extremely grateful to the Editor, the Associate Editor and an anonymous referee for their many valuable suggestions and comments which greatly improved the presentation of the paper.*

REFERENCES

Carroll, R.J., Ruppert, D. and Stefanski, L.A. (1995). *Nonlinear Measurement Error Models.* Chapman and Hall, New York.

Chen, H. (1988). Convergence rates for parametric components in a partly linear model. *Annals of Statistics*, **16**, 136-146.

Cuzick, J. (1992a). Semiparametric additive regression. *Journal of the Royal Statistical Society, Series B*, **54**, 831-843.

Cuzick, J. (1992b). Efficient estimates in semiparametric additive regression models with unknown error distribution. *Annals of Statistics*, **20**, 1129-1136.

Engle, R. F., Granger, C.W.J., Rice,J. and Weiss, A. (1986). Semiparametric estimates of the relation between weather and electricity sales. *Journal of the American Statistical Association*, **81**, 310-320.

Fan, J. & Truong, Y. K. (1993). Nonparametric regression with errors in variables. *Annals of Statistics*, **21**, 1900–1925.

Fuller, W. A. (1987). *Measurement Error Models.* Wiley, New York.

Härdle,W., Klinke, S. and Turlach, B.A. (1995). *XploRe: An Interactive Statistical Computing Environment.* Springer-Verlag,

Heckman, N. E. (1986). Spline smoothing in partly linear models. *Journal of the Royal Statistical Society, Series B*, **48**, 244-248.

Kendall, M. and Stuart, A. (1992). *The Advanced Theory of Statistics 2*, 4th ed, Charles Griffin, London.

Liang, H. and Härdle, W. (1997). *Asymptotic normality of parametric part in partially linear heteroscedastic regression models.* DP 33, SFB 373, Humboldt University of Berlin.

Speckman, P. (1988). Kernel smoothing in partial linear models. *Journal of the Royal Statistical Society, Series B*, **50**, 413-436.

Severini, T.A. and Staniswalis, J.G. (1994). Quasilikelihood estimation in semiparametric models. *Journal of the American Statistical Association*, **89**, 501-511.

A APPENDIX

In this appendix, we prove several lemmas required. Lemma A.1 provides bounds for $h_j(T_i) - \sum_{k=1}^n \omega_{nk}(T_i)h_j(T_k)$ and $g(T_i) - \sum_{k=1}^n \omega_{nk}(T_i)g(T_k)$. The proof is immediate.

Lemma A.1 *Suppose that Assumptions 1.1–1.4 hold. Then*

$$\max_{1\le i\le n} |G_j(T_i) - \sum_{k=1}^n \omega_{nk}(T_i)G_j(T_k)| = O_p(c_n) \quad \text{for } j = 0,\ldots,p,$$

where $G_0(\cdot) = g(\cdot)$ *and* $G_l(\cdot) = h_l(\cdot)$ *for* $l = 1,\ldots,p$.

Lemma A.2 *If Assumptions 1.1-1.4 hold, then* $n^{-1}\tilde{\mathbf{X}}^{\mathrm{T}}\tilde{\mathbf{X}} = B + o_P(1)$.

Proof. Denote $\overline{h}_{ns}(T_i) = h_s(T_i) - \sum_{k=1}^n \omega_{nk}(T_i)X_{ks}$. It follows from $X_{js} = h_s(T_j) + V_{js}$ that the $(s,m)-$th element of $\tilde{\mathbf{X}}^{\mathrm{T}}\tilde{\mathbf{X}}$ $(s,m = 1,\ldots,p)$ is

$$\begin{aligned}\sum_{j=1}^n \tilde{X}_{js}\tilde{X}_{jm} &= \sum_{j=1}^n V_{js}V_{jm} + \sum_{j=1}^n \overline{h}_{ns}(T_j)V_{jm} + \sum_{j=1}^n \overline{h}_{nm}(T_j)V_{js} + \sum_{j=1}^n \overline{h}_{ns}(T_j)\overline{h}_{nm}(T_j) \\ &\stackrel{\text{def}}{=} \sum_{j=1}^n V_{js}V_{jm} + \sum_{q=1}^3 R_{nsm}^{(q)}.\end{aligned}$$

The strong law of large numbers implies that $n^{-1}\sum_{i=1}^n V_iV_i^{\mathrm{T}} = B + o_P(1)$, and Lemma A.1 means $R_{nsm}^{(3)} = o_P(n)$, which together with the Cauchy-Schwarz inequality shows that $R_{nsm}^{(1)} = o_P(n)$ and $R_{nsm}^{(2)} = o_P(n)$. This completes the proof of the lemma.

Lemma A.3 *(Bernstein's inequality) Let* $\Gamma_1,\ldots,\Gamma_n$ *be independent random variables with zero means and bounded ranges:* $|\Gamma_i| \le M$. *Then for each* $\eta > 0$,

$$P\Big\{\Big|\sum_{i=1}^n \Gamma_i\Big| > \eta\Big\} \le 2\exp\left\{-\eta^2\Big/\left[2\left\{\sum_{i=1}^n var(\Gamma_i) + M\eta\right\}\right]\right\}.$$

Denote $\epsilon_j' = \epsilon_j I(|\epsilon_j| \le n^{1/4})$ and $\epsilon_j'' = \epsilon_j - \epsilon_j' = \epsilon_j I(|\epsilon_j| > n^{1/4}), j = 1,\ldots,n$. We next establish several results for nonparametric regression.

Lemma A.4 *Assume that Assumptions 1.3–1.4 hold. Then*

$$\max_{1\le i\le n}\left|\sum_{k=1}^{n}\omega_{nk}(T_i)\epsilon_k\right| = o_P\{n^{-2/5}\log(n)\}.$$

Proof. Fix $L > 0$ but arbitrarily large. Let

$$B_{nL} = \left\{\max_{1\le i\le n}\sum_{j=1}^{n} w_{nj}(T_i) \le L, \max_{1\le i,j\le n} w_{nj}(T_i) \le Lb_n\right\}.$$

Then

$$P\left\{\max_{1\le i\le n}|\sum_{j=1}^{n} w_{nj}(T_i)\epsilon_j| > n^{-2/5}\log(n)\right\} \le P\{I(B_{nL}) = 0\}$$
$$+P\left\{\max_{1\le i\le n}|\sum_{j=1}^{n} w_{nj}(T_i)\epsilon_j| > n^{-2/5}\log(n), I(B_{nL}) = 1\right\}. \quad (12)$$

Since by Assumption 1.3 $P\{I(B_{nL}) = 1\}$ can be made arbitrarily small by choosing L sufficiently large, it suffices to show that the second term in (12) converges to zero for any L.

Application of Bernstein's inequality to (12) is complicated by the fact that the terms $w_{nj}(T_i)$ and $I(B_{nL}) = 1$ are random. We first condition on these terms and will later uncondition. For sufficiently large C, first note that

$$P\left\{\max_{1\le i\le n}|\sum_{j=1}^{n} w_{nj}(T_i)\{\epsilon'_j - E(\epsilon'_j)\}| > Cn^{-2/5}\log(n)|\{w_{nj}(T_i)\}, I(B_{nL}) = 1\right\}$$
$$\le \sum_{i=1}^{n} P\left\{|\sum_{j=1}^{n} w_{nj}(T_i)\{\epsilon'_j - E(\epsilon'_j)\}| > Cn^{-2/5}\log(n)|\{w_{nj}(T_i)\}, I(B_{nL}) = 1\right\}.$$

Now apply Bernstein's inequality with $\eta = Cn^{-2/5}\log(n)$ and $M = 2Lb_n n^{1/4}$. Then the right hand side of the last expression is bounded by

$$2I(B_{nL})\sum_{i=1}^{n}\exp\left\{-\frac{C^2 n^{-4/5}\log^2(n)}{4LCb_n n^{1/4-2/5}\log(n) + 2\sum_{j=1}^{n} w_{nj}^2(T_i)\mathrm{var}(\epsilon'_j)}\right\}. \quad (13)$$

First note that $b_n = n^{-4/5}$ and $\mathrm{var}(\epsilon'_j) < \infty$. On the set that $I(B_{nL}) = 1$, we have thus that

$$\sum_{j=1}^{n} w_{nj}^2(T_i) \le \sum_{j=1}^{n} w_{nj}(T_i)\max_{1\le i,j\le n} w_{nj}(T_i) \le L^2 b_n.$$

This means that (13) is bounded by $2nI(B_{nL})\exp\{-(C/L)\log(n)\} \le n^{-3/2}$ for sufficiently large C. Since this last expression is independent of the $\{w_{nj}(T_i)\}$ except through $I(B_{nL})$, we have that

$$P\left\{\max_{1\le i\le n}|\sum_{j=1}^{n} w_{nj}(T_i)\{\epsilon'_j - E(\epsilon'_j)\}| > Cn^{-2/5}\log(n)|I(B_{nL}) = 1\right\} \le n^{-3/2}.$$

This shows that

$$\max_{1\le i\le n} |\sum_{j=1}^{n} w_{nj}(T_i)\{\epsilon_j' - E(\epsilon_j')\}| = o_p\{n^{-2/5}\log(n)\}. \tag{14}$$

Now consider $V_n = \max_{1\le i\le n}\sum_{j=1}^{n} w_{nj}(T_i)\{\epsilon_j'' - E(\epsilon_j'')\}$. Let p and q be such that $1 \le p < 2$, $1/p + 1/q = 1$ and $1/q < 2/5 - 1/4$. By Hölder's inequality,

$$|V_n| \le \max_{1\le i\le n}\left\{\sum_{j=1}^{n} w_{nj}^q(T_i)\right\}^{1/q}\left\{\sum_{j=1}^{n}|\epsilon_j'' - E(\epsilon_j'')|^p\right\}^{1/p}.$$

By assumption 1.3(ii), $w_{nj}^q(T_i) = O_P(b_n^q)$ so that $\sum_j w_{nj}^q(T_i) = O_P(nb_n^q) = O_P(n^{1-4q/5})$, and thus

$$|V_n| \le O_P\{n^{(1-4q/5)/q}\}\left\{\sum_{j=1}^{n}|\epsilon_j'' - E(\epsilon_j'')|^p\right\}^{1/p}.$$

Clearly,

$$n^{-1}\sum_{j=1}^{n}\left[|\epsilon_j'' - E(\epsilon_j'')|^p - E\left\{|\epsilon_j'' - E(\epsilon_j'')|^p\right\}\right] = o_P(1). \tag{15}$$

Also, again using Hölder's inequality,

$$E|\epsilon_j''|^p = E\left\{|\epsilon_j|^p I(\epsilon_j > n^{1/4})\right\} \le (E|\epsilon_j|^4)^{p/4}\{P(|\epsilon_j| > n^{1/4})\}^{1-p/4},$$

which by Chebychev's inequality is bounded by $\le n^{-1+p/4}(E|\epsilon_j|^4)^{p/4}$. It thus follows that

$$\sum_{j=1}^{n} E|\epsilon_j'' - E(\epsilon_j'')|^p = O_P(n^{p/4}). \tag{16}$$

Replacing (16) into (15), we get

$$\sum_{j=1}^{n}|\epsilon_j'' - E(\epsilon_j'')|^p = O_P(n^{p/4}),$$

which along with the fact that $1/q < 2/5 - 1/4$, we find that

$$\max_{1\le i\le n}\sum_{j=1}^{n} w_{nj}(T_i)\{\epsilon_j'' - E(\epsilon_j'')\} = O_P(n^{(1-4q/5)/q+1/4}) = o_P(n^{-2/5}).$$

This completes the proof of Lemma A.4.

Lemma A.5 *Suppose that Assumptions 1.1–1.4 hold. Then*

$$\begin{aligned}\sum_{i=1}^{n} U_i\tilde{g}_i &= o_p(n^{1/2});\\ \sum_{i=1}^{n}\epsilon_i\tilde{g}_i &= o_p(n^{1/2}).\end{aligned}$$

The same holds if $g(T_i)$ is replaced by $h_j(T_i)$.

Carroll, R., Härdle, W. and Liang, H. (1999), Estimation In A Semiparametric Partially Linear Errors In Variables Model

Proof. We prove only the first step, as the other steps follow in a similar fashion. Let $\xi_n = n^{1/2}/\log(n)$.

$$P(|\sum_{i=1}^{n} U_i\widetilde{g}_i| > \xi_n) \leq P(|\sum_{i=1}^{n} U_i\widetilde{g}_i| > \xi_n, \max_i |\widetilde{g}_i| \leq c_n \log n) + P(\max_i |\widetilde{g}_i| > c_n \log n).$$

The second term is $o_P(1)$ by Lemma A.1. For the first term, let r_i be the event that $|\widetilde{g}_i| \leq c_n \log(n)$. Then,

$$\begin{aligned} P[|\sum_{i=1}^{n} U_i\widetilde{g}_i| > \xi_n, \{I(r_i) = 1\ \forall i)\}] \quad &\leq \quad \xi_n^{-2} \sum_{i=1}^{n} E[U_i\widetilde{g}_i\{I(r_i) = 1\}]^2 \\ &\quad + \xi_n^{-2} \sum_{i \neq k}^{n} E[U_iU_k\widetilde{g}_i\widetilde{g}_k I(r_k) = 1\ \forall k\}]. \end{aligned} \tag{17}$$

Since $\widetilde{g}_i\{I(r_i) = 1\} \leq c_n \log(n)$ is independent of U_i, the first term in (17) is of order $O\{n\xi_n^{-2}c_n^2 \log^2(n)\} = o(1)$. The second term is easily seen to equal zero.

Lemma A.6 *Suppose that Assumptions 1.1–1.4 hold. Then*

$$\begin{aligned} n^{-1/2} \sum_{i=1}^{n} \sum_{j=1}^{n} \omega_{nj}(T_i)\epsilon_j U_i \quad &= \quad o_P(1); \\ n^{-1/2} \sum_{i=1}^{n} \sum_{j=1}^{n} \omega_{nj}(T_i)\epsilon_j \epsilon_i \quad &= \quad o_P(1); \\ n^{-1/2} \sum_{i=1}^{n} \sum_{j=1}^{n} \omega_{nj}(T_i)U_j U_i \quad &= \quad o_P(1). \end{aligned}$$

Proof. We prove only the first step, as the other steps follow in a similar fashion. Let r_{ij} be the event that $|w_{nj}(T_i)| \leq Cb_n \log n$.

$$\begin{aligned} P\{n^{-1/2}|\sum_{i=1}^{n} \sum_{j=1}^{n} \omega_{nj}(T_i)\epsilon_j U_i| > \xi\} \quad &\leq \quad P\{n^{-1/2}|\sum_{i=1}^{n} \sum_{j=1}^{n} \omega_{nj}(T_i)\epsilon_j U_i| > \xi, I(r_{ij} = 1\ \forall i, j)\} \\ &+ \quad P\{\max_{i,j} |w_{nj}(T_i)| > Cb_n \log n\}. \end{aligned}$$

The second term tends to zero by assumption 1.3(ii). For the first term, note that

$$\begin{aligned} &P\{n^{-1/2}|\sum_{i=1}^{n} \sum_{j=1}^{n} \omega_{nj}(T_i)\epsilon_j U_i| > \xi, I(r_{ij} = 1\ \forall i, j)\} \\ &\leq n^{-1}\xi^{-2} E\{\sum_{i=1}^{n} \sum_{j=1}^{n} \omega_{nj}(T_i)\epsilon_j U_i I(r_{ij} = 1\ \forall i, j)\}^2 \\ &= n^{-1}\xi^{-2} \sum_{i=1}^{n} E\left\{\sum_{j=1}^{n} \omega_{nj}(T_i)\epsilon_j I(r_{ij} = 1\ \forall i, j)\right\}^2 EU_i^2. \end{aligned}$$

Carroll, R., Härdle, W. and Liang, H. (1999), Estimation In A Semiparametric Partially Linear Errors In Variables Model

The last equation holds because U_i and $\sum_{j=1}^n \omega_{nj}(T_i)\epsilon_j I(r_{ij} = 1\ \forall i,j)$ are independent for each i, and U_i are iid with mean zero. It suffices to prove

$$\max_i E\left\{\sum_{j=1}^n \omega_{nj}(T_i)\epsilon_j I(r_{ij} = 1\ \forall i,j)\right\}^2 \to 0.$$

In fact,

$$E\left\{\sum_{j=1}^n \omega_{nj}(T_i)\epsilon_j I(r_{ij} = 1\ \forall i,j)\right\}^2 = \sum_{j=1}^n E\left\{\omega_{nj}(T_i)\epsilon_j I(r_{ij} = 1\ \forall i,j)\right\}^2$$
$$+ \sum_{j\neq k}^n E\left\{\omega_{nj}(T_i)\epsilon_j \omega_{nk}(T_i)\epsilon_k I(r_{ij} = 1\ \forall i,j)\right\}.$$

The second term equals zero. The first term equals

$$\sum_{j=1}^n E\left[\{\omega_{nj}(T_i)\epsilon_j\}^2 \{I(r_{ij}) = 1\ \forall i,j\}\right],$$

and this is of order $O\{nb_n^2 \log^2(n)\} = o(1)$, as required.

Lemma A.7 *Assume that Assumptions 1.1-1.4 hold. Then*

$$p\lim_{n\to\infty} n^{-1}\widetilde{\mathbf{W}}^{\mathrm{T}}\widetilde{\mathbf{W}} = B + \Sigma_{uu}, \tag{18}$$
$$p\lim_{n\to\infty} n^{-1}\widetilde{\mathbf{W}}^{\mathrm{T}}\widetilde{\mathbf{Y}} = B\beta, \tag{19}$$
$$p\lim_{n\to\infty} n^{-1}\widetilde{\mathbf{Y}}^{\mathrm{T}}\widetilde{\mathbf{Y}} = \beta^{\mathrm{T}}B\beta + \sigma^2. \tag{20}$$

Proof. Since $W_i = X_i + U_i$ and $\widetilde{W}_i = \widetilde{X}_i + \widetilde{U}_i$. For the (s,m) matrix element we obtain

$$n^{-1}(\widetilde{\mathbf{W}}^{\mathrm{T}}\widetilde{\mathbf{W}})_{sm} = n^{-1}(\widetilde{\mathbf{X}}^{\mathrm{T}}\widetilde{\mathbf{X}})_{sm} + n^{-1}(\widetilde{\mathbf{U}}^{\mathrm{T}}\widetilde{\mathbf{X}})_{sm} + n^{-1}(\widetilde{\mathbf{X}}^{\mathrm{T}}\widetilde{\mathbf{U}})_{sm} + n^{-1}(\widetilde{\mathbf{U}}^{\mathrm{T}}\widetilde{\mathbf{U}})_{sm}. \tag{21}$$

At first, we prove that the 2nd and 3rd terms converge to zero. It follows from the strong law of large numbers and Lemma A.2 that

$$n^{-1}\sum_{j=1}^n X_{js}U_{jm} \to 0 \quad a.s. \tag{22}$$

Observe that

$$n^{-1}(\widetilde{\mathbf{X}}^{\mathrm{T}}\widetilde{\mathbf{U}})_{sm} = n^{-1}\Big[\sum_{j=1}^n X_{js}U_{jm} - \sum_{j=1}^n \left\{\sum_{k=1}^n \omega_{nk}(T_j)X_{ks}\right\}U_{jm}$$
$$- \sum_{j=1}^n \left\{\sum_{k=1}^n \omega_{nk}(T_j)U_{km}\right\}X_{js} + \sum_{j=1}^n \left\{\sum_{k=1}^n \omega_{nk}(T_j)X_{ks}\right\}\left\{\sum_{k=1}^n \omega_{nk}(T_j)U_{km}\right\}\Big].$$

Similar to the proof of Lemma A.4, we can prove that $\sup_{1\leq j\leq n} |\sum_{k=1}^n \omega_{nk}(T_j)U_{km}| = o_P(1)$, which together with (22) and Assumption 1.3 (ii) deduce that the above each term tends to zero. The same reason implies that $n^{-1}(\widetilde{\mathbf{U}}^{\mathrm{T}}\widetilde{\mathbf{X}})_{sm}$ also tends to zero.

Carroll, R., Härdle, W. and Liang, H. (1999), Estimation In A Semiparametric Partially Linear Errors In Variables Model

Secondly, we prove

$$n^{-1}(\widetilde{\mathbf{U}}^{\mathrm{T}}\widetilde{\mathbf{U}})_{sm} \to \sigma_{sm}^2 \tag{23}$$

here σ_{sm}^2 is the $(s,m)-$th element of Σ_{uu}.

$$\begin{aligned} n^{-1}(\widetilde{\mathbf{U}}^{\mathrm{T}}\widetilde{\mathbf{U}})_{sm} &= n^{-1}\Big[\sum_{j=1}^{n} U_{js}U_{jm} - \sum_{j=1}^{n}\Big\{\sum_{k=1}^{n}\omega_{nk}(T_j)U_{ks}\Big\}U_{jm} \\ &\quad - \sum_{j=1}^{n}\Big\{\sum_{k=1}^{n}\omega_{nk}(T_j)U_{km}\Big\}U_{js} + \sum_{j=1}^{n}\Big\{\sum_{k=1}^{n}\omega_{nk}(T_j)U_{ks}\Big\}\Big\{\sum_{k=1}^{n}\omega_{nk}(T_j)U_{km}\Big\}\Big]. \end{aligned}$$

Obviously $n^{-1}\sum_{j=1}^{n} U_{js}U_{jm} \to \sigma_{sm}^2$. It follows from Lemmas A.4 and A.6 that (23) holds. Using (21), (23) and the arguments for $n^{-1}(\widetilde{\mathbf{U}}^{\mathrm{T}}\widetilde{\mathbf{X}})_{sm} \to 0$ and $n^{-1}(\widetilde{\mathbf{X}}^{\mathrm{T}}\widetilde{\mathbf{U}})_{sm} \to 0$, we complete the proof of (18).

We now prove (19). Note that $\widetilde{\mathbf{W}}^{\mathrm{T}}\widetilde{\mathbf{Y}} = \widetilde{\mathbf{W}}^{\mathrm{T}}(\widetilde{\mathbf{X}}\beta + \widetilde{\mathbf{G}} + \widetilde{\epsilon})$. From Lemma 1, $\sum_{j=1}^{n}\widetilde{g}_j^2 = O_P(c_n^2 n)$, so that Since

$$|\sum_{j=1}^{n} X_{js}\widetilde{g}_j| \le \left(\sum_{j=1}^{n} X_{js}^2 \sum_{j=1}^{n}\widetilde{g}_j^2\right)^{1/2} \le O_P(c_n n^{1/2})\left(\sum_{j=1}^{n} X_{js}^2\right)^{1/2} = O_P(Cnc_n),$$

and

$$\begin{aligned} (\widetilde{\mathbf{W}}^{\mathrm{T}}\widetilde{\mathbf{G}})_s &= \sum_{j=1}^{n}\widetilde{X}_{js}\widetilde{g}_j + \sum_{j=1}^{n}\widetilde{U}_{js}\widetilde{g}_j \\ &= \sum_{j=1}^{n}\Big\{X_{js} - \sum_{k=1}^{n}\omega_{nk}(T_j)X_{ks}\Big\}\widetilde{g}_j + \sum_{j=1}^{n}\widetilde{U}_{js}\widetilde{g}_j. \end{aligned}$$

Obviously $n^{-1}\sum_{j=1}^{n}\widetilde{U}_{js}\widetilde{g}_j$ tends to zero. Therefore $n^{-1}(\widetilde{\mathbf{W}}^{\mathrm{T}}\widetilde{\mathbf{G}})_s$ tends to zero.

The proof for that $n^{-1}(\widetilde{\mathbf{W}}^{\mathrm{T}}\widetilde{\epsilon})_s$ tends to zero is similar to that of $n^{-1}(\widetilde{\mathbf{W}}^{\mathrm{T}}\widetilde{\mathbf{U}})_s \to 0$. Combining the above arguments and (18), we complete the proof of (19). The proof of (20) can be completed by the similar arguments. The details are omitted.

Lemma A.8 *Assume that Assumptions 1.1-1.4 hold. Then*

$$\begin{aligned} n^{-1/2}\sum_{i=1}^{n}\widetilde{\epsilon}_i\widetilde{X}_i &= n^{-1/2}\sum_{i=1}^{n}\epsilon_i V_i + o_P(1); \\ n^{-1/2}\sum_{i=1}^{n}\widetilde{X}_i\widetilde{U}_i^T &= n^{-1/2}\sum_{i=1}^{n}V_i U_i^T + o_P(1). \end{aligned}$$

Proof. We show only the first step, as the second step follows in a similar fashion. Let $h(T) = E(X|T)$ and $h_i = h(T_i)$. By direct calculation,

$$n^{-1/2}\sum_{i=1}^{n}\epsilon_i(V_i - \widetilde{X}_i) = n^{-1/2}\sum_{i=1}^{n}\epsilon_i\widetilde{h}_i - n^{-1/2}\sum_{i=1}^{n}\epsilon_i\sum_{j=1}^{n}w_{nj}(T_i)\left\{X_j - h(T_j)\right\}.$$

The first term is $o_P(1)$ by Lemma A.4. The second terms follows using Assumption 1.1 by using the same method of proof as in Lemma A.6, upon remembering that for $j \neq k$,

$$E\left[\{X_j - h(T_j)\}\{X_k - h(T_k)\}\,|T_1, ..., T_n)\right] = 0.$$

Proof of Theorem 3.1. Denote $\Delta_n = (\widetilde{\mathbf{W}}^{\mathrm{T}}\widetilde{\mathbf{W}} - n\Sigma_{uu})/n$. By Lemma A.7 and a direct calculation,

$$\begin{aligned} n^{1/2}(\widehat{\beta}_n - \beta) &= n^{-1/2}\Delta_n^{-1}(\widetilde{\mathbf{W}}^{\mathrm{T}}\widetilde{\mathbf{Y}} - \widetilde{\mathbf{W}}^{\mathrm{T}}\widetilde{\mathbf{W}}\beta + n\Sigma_{uu}\beta) \\ &= n^{-1/2}\Delta_n^{-1}(\widetilde{\mathbf{X}}^{\mathrm{T}}\widetilde{\mathbf{G}} + \widetilde{\mathbf{X}}^{\mathrm{T}}\widetilde{\epsilon} + \widetilde{\mathbf{U}}^{\mathrm{T}}\widetilde{\mathbf{G}} + \widetilde{\mathbf{U}}^{\mathrm{T}}\widetilde{\epsilon} - \widetilde{\mathbf{X}}^{\mathrm{T}}\widetilde{\mathbf{U}}\beta - \widetilde{\mathbf{U}}^{\mathrm{T}}\widetilde{\mathbf{U}}\beta + n\Sigma_{uu}\beta). \end{aligned}$$

By Lemmas A.1–A.2, A.4–A.6 and A.8 it is an easy calculation to show that

$$n^{1/2}(\widehat{\beta}_n - \beta) = n^{-1/2}\Delta_n^{-1}\sum_{i=1}^{n}\left(V_i\epsilon_i - V_iU_i^{\mathrm{T}}\beta + U_i\epsilon_i - U_iU_i^{\mathrm{T}}\beta + \Sigma_{uu}\beta\right) + o_P(1) \tag{24}$$

$$\stackrel{\mathrm{def}}{=} n^{-1/2}\sum_{i=1}^{n}\zeta_{in} + o_P(1). \tag{25}$$

Since $\lim_{n\to\infty} n^{-1}\sum_{i=1}^{n} V_i = 0$ and $\lim_{n\to\infty} n^{-1}\sum_{i=1}^{n} V_iV_i^{\mathrm{T}} = B$, and $\sup_i E(\epsilon_i^4 + \|U\|^4) < \infty$, it follows that the sequence of $k-$th elements $\{\zeta_{in}^{(k)}\}$ of $\{\zeta_{in}\}$ $(k = 1, \ldots, p)$ satisfy, for any given $\zeta > 0$, $n^{-1}\sum_{i=1}^{n} E\{{\zeta_{in}^{(k)}}^2 I(|\zeta_{in}^{(k)}| > \zeta n^{1/2})\} \to 0$ as $n \to \infty$. This means that the Lindeberg's condition for the central limit theorem holds. Moreover, note that

$$\begin{aligned} \mathrm{cov}(\zeta_{ni}) = E\left\{V_i(\epsilon_i - U_i^{\mathrm{T}}\beta)\right\}^{\otimes 2} + E\left\{(U_iU_i^{\mathrm{T}} - \Sigma_{uu})\beta\right\}^{\otimes 2} + E(U_iU_i^{\mathrm{T}}\epsilon_i^2) \\ + E(V_iU_i^{\mathrm{T}}\beta\beta^{\mathrm{T}}U_iU_i^{\mathrm{T}}) + E(U_iU_i^{\mathrm{T}}\beta\beta^{\mathrm{T}}U_i)V_i. \end{aligned}$$

These arguments ensure that

$$\lim_{n\to\infty} n^{-1}\sum_{i=1}^{n}\mathrm{cov}(\zeta_{ni}) = E[(\epsilon - U^{\mathrm{T}}\beta)\{X - E(X|T)\}]^{\otimes 2} + E\{(UU^{\mathrm{T}} - \Sigma_{uu})\beta\}^{\otimes 2} + E(UU^{\mathrm{T}}\epsilon^2).$$

Theorem 3.1 now follows.

Proof of Theorem 3.2. Denote

$$A_n = n^{-1}\begin{bmatrix} \widetilde{\mathbf{Y}}^{\mathrm{T}}\widetilde{\mathbf{Y}} & \widetilde{\mathbf{Y}}^{\mathrm{T}}\widetilde{\mathbf{W}} \\ \widetilde{\mathbf{W}}^{\mathrm{T}}\widetilde{\mathbf{Y}} & \widetilde{\mathbf{W}}^{\mathrm{T}}\widetilde{\mathbf{W}} \end{bmatrix}; \qquad A = \begin{bmatrix} \beta^{\mathrm{T}}B\beta + \sigma^2 & \beta^{\mathrm{T}}B \\ B\beta & B + \Sigma_{uu} \end{bmatrix};$$

$$\widetilde{A}_n = n^{-1}\begin{bmatrix} (\epsilon + \mathbf{V}\beta)^{\mathrm{T}}(\epsilon + \mathbf{V}\beta) & (\epsilon + \mathbf{V}\beta)^{\mathrm{T}}(\mathbf{U} + \mathbf{V}) \\ (\mathbf{U} + \mathbf{V})^{\mathrm{T}}(\epsilon + \mathbf{V}\beta) & (\mathbf{U} + \mathbf{V})^{\mathrm{T}}(\mathbf{U} + \mathbf{V}) \end{bmatrix}.$$

Note that $\widehat{\sigma}_n^2 = (1, -\widehat{\beta}_n^{\mathrm{T}})A_n(1, -\widehat{\beta}_n^{\mathrm{T}})^{\mathrm{T}} - \widehat{\beta}_n^{\mathrm{T}}\Sigma_{uu}\widehat{\beta}_n^{\mathrm{T}}$. A direct calculation using Lemma A.6 yields that $n^{1/2}(\widehat{\sigma}_n^2 - \sigma^2) = n^{1/2}\sum_{j=1}^{4} S_{jn} + n^{-1/2}(\epsilon - \mathbf{U}\beta)^{\mathrm{T}}(\epsilon - \mathbf{U}\beta) - n^{1/2}(\beta^{\mathrm{T}}\Sigma_{uu}\beta + \sigma^2) + o_P(1)$, where $S_{1n} = (1, -\widehat{\beta}_n^{\mathrm{T}})(A_n - \widetilde{A}_n)(1, -\widehat{\beta}_n^{\mathrm{T}})^{\mathrm{T}}$, $S_{2n} = (1, -\widehat{\beta}_n^{\mathrm{T}})(\widetilde{A}_n - A)(0, \beta^{\mathrm{T}} - \widehat{\beta}_n^{\mathrm{T}})^{\mathrm{T}}$, $S_{3n} = (0, \beta^{\mathrm{T}} - \widehat{\beta}_n^{\mathrm{T}})(\widetilde{A}_n - A)(1, -\beta^{\mathrm{T}})^{\mathrm{T}}$, $S_{4n} = -(\beta - \widehat{\beta}_n)^{\mathrm{T}}B(\beta - \widehat{\beta}_n)$. It follows from Theorem 3.1 and Lemma A.7

Carroll, R., Härdle, W. and Liang, H. (1999), Estimation In A Semiparametric Partially Linear Errors In Variables Model

that $n^{1/2}S_{jn}$ converges to zero in probability for $j = 2, 3, 4$. To show that $n^{1/2}S_{1n} = o_P(1)$ is more detailed, but follows from Lemmas A.1, A.4–A.6. This means that

$$n^{1/2}(\widehat{\sigma}_n^2 - \sigma^2) = n^{-1/2}\sum_{i=1}^{n}\left\{(\epsilon_i - U_i^{\mathrm{T}}\beta)^2 - (\beta^{\mathrm{T}}\Sigma_{uu}\beta + \sigma^2)\right\} + o_P(1).$$

Theorem 3.2 now follows immediately.

Proof of Theorem 4.1. Since $\widehat{\beta}_n$ is a consistent estimator of β, its asymptotic bias and variance equal the relative ones of $\sum_{j=1}^{n}\omega_{nj}(t)(Y_j - W_j^{\mathrm{T}}\beta)$, which is denoted by $\widehat{g}_n^*(t)$. By a simple calculation,

$$\begin{aligned} E\widehat{g}_n^*(t) - g(t) &= \sum_{i=1}^{n}\omega_{ni}(t)g(T_i) - g(t), \\ \widehat{g}_n^*(t) - E\widehat{g}_n^*(t) &= \sum_{i=1}^{n}\omega_{ni}^2(t)(\beta^{\mathrm{T}}\Sigma_{uu}\beta + \sigma^2). \end{aligned}$$

Both terms are order $O(n^{-2/5})$ by Lemma A.1 and Assumption 1.3 (iii). Theorem 4.1 then follows.

Institute of Systems Science
Chinese Academy of Sciences
Beijing 100080
China

Institut für Statistik und Ökonometrie
Humboldt-Universität zu Berlin
D-10178 Berlin
Germany
E-Mail: haerdle@wiwi.hu-berlin.de

Department of Statistics
Texas A&M University
College Station, TX 77843–3143
USA
E-Mail: carroll@stat.tamu.edu

ELSEVIER

Journal of Statistical Planning and Inference 91 (2000) 413–426

journal of statistical planning and inference

www.elsevier.com/locate/jspi

Bootstrap approximation in a partially linear regression model ☆

Hua Liang[a,b,*], Wolfgang Härdle[a], Volker Sommerfeld[a]

[a] *Institut für Statistik und Ökonometrie, Humboldt-Universität zu Berlin, D-10178 Berlin, Germany*

[b] *Institute and Systems Science, Chinese Academy of Sciences, Beijing 10080, China*

Abstract

Consider the semiparametric regression model $Y_i = X_i^{\mathrm{T}}\beta + g(T_i) + \varepsilon_i$ $(i = 1, \ldots, n)$, where (X_i, T_i) are known and fixed design points, β is a p-dimensional unknown parameter, $g(\cdot)$ is an unknown function on $[0, 1]$, and ε_i are i.i.d. random errors with mean 0 and variance σ^2. In this paper, we first construct bootstrap statistics β_n^* and σ_n^{2*} by resampling. Then we prove that for the estimators β_n and σ_n^2 of the parameters β and σ^2, $\sqrt{n}(\beta_n^* - \beta_n)$ and $\sqrt{n}(\beta_n - \beta)$, $\sqrt{n}(\sigma_n^{2*} - \sigma_n^2)$ and $\sqrt{n}(\sigma_n^2 - \sigma^2)$ have the same limit distributions, respectively. The advantage of the bootstrap approximation is explained. The feasibility of this approach is also shown in a simulation study.

MSC: primary: 62G05; secondary: 60F15

Keywords: Semiparametric regression model; Bootstrap approximation; Asymptotic normality

1. Introduction

Consider the model given by

$$Y_i = X_i^{\mathrm{T}}\beta + g(T_i) + \varepsilon_i, \quad i = 1, \ldots, \tag{1.1}$$

where $X_i = (x_{i1}, \ldots, x_{ip})^{\mathrm{T}}$ $(p \geqslant 1)$ and T_i $(T_i \in [0, 1])$ are known design points, $\beta = (\beta_1, \ldots, \beta_p)^{\mathrm{T}}$ is an unknown parameter vector, g is an unknown function, and $\varepsilon_1, \ldots, \varepsilon_n$ are i.i.d. random variables with mean 0 and unknown variance σ^2.

This model is important because it can be used in applications where one can assume that the responses Y_i and predictors X_i are linearly dependent, but Y_i's are nonlinearly related to the independent variables T_i. Engle et al. (1986) studied the effect of weather

☆ This research was supported by Sonderforschungsbereich 373 "Quantifikation und Simulation Ökonomischer Prozesse". The work of Hua Liang was partially supported by Alexander von Humboldt Foundation.

* Correspondence address. Department of Statistics, Texas A& M University, College Station, TX 77843-3143, USA.

E-mail address: hliang@stat.tamu.edu (H. Liang).

0378-3758/00/$ - see front matter
PII: S0378-3758(00)00191-9

on electricity demand. Liang et al. (1997) used the model to investigate the relationship between income and age from German data. Heckman (1986), Speckman (1988) and Chen (1988) considered the asymptotic normality of estimators of β and σ^2. Later Cuzick (1992a,b) and Schick (1993) discussed asymptotic properties and asymptotic efficiency for these estimators. Liang and Cheng (1993) proposed the second-order asymptotic efficiency of least-squares estimator and maximum likelihood estimator of β. Hong and Cheng (1993) considered bootstrap approximation of the estimators for the parameters in the model (1.1) in the case where $\{X_i, T_i,\ i = 1,\ldots,n\}$ are i.i.d. random variables and $g(\cdot)$ is estimated by a kernel smoother.

The bootstrap technique is a useful tool for the approximation of an unknown probability distribution and therefore for its characteristics like moments or confidence regions. This approximation can be performed by different estimators of the true underlying distribution that should be well adapted to the special situation. In this paper, we use the empirical distribution function which puts mass $1/n$ at each residual in order to approximate the underlying error distribution (for more details see Section 2). This classical bootstrap technique was introduced by Efron (for a review see e.g. Efron and Tibshirani, 1993). Note that for a heteroscedastic error structure, a wild bootstrap procedure (see e.g. Wu, 1986 or Härdle and Mammen, 1993) would be more appropriate.

Hong and Cheng (1993) proved that their bootstrap approximation is the same as the classic methods, but failed to explain the advantage of the bootstrap method, which will be discussed in this paper. We will construct bootstrap statistics for parameters β and σ^2, and study their asymptotic normality when (X_i, T_i) are known design points and $g(\cdot)$ is estimated by general nonparametric fitting. We will show, analytically as well as numerically, that the bootstrap techniques provide a reliable method to approximate the asymptotic distributions of the estimates.

The effect of the smoothing parameter is studied in a simulation study. Thereby it turns out that the estimators of the parametric part are quite robust against the choice of the smoothing parameter. More details can be found in Section 4.

The paper is organised as follows. In the rest of this section we explain the basic idea for estimating the parameters. Section 2 constructs bootstrap statistics of β and σ^2. Section 3 lists basic assumptions and states the main results. In Section 4, we present a simulation study in order to support the asymptotic results. Section 5 presents the proof of the main result. All technical lemmas are in the appendix. For convenience and simplicity, we shall employ C $(0 < C < \infty)$ to denote some constant not depending on n, but that may assume different values at each appearance.

Generally, we are used to estimate the linear parameter β by backfitting and local likelihood methods, based on which the asymptotic variances of the two estimates are the same. Here we adopt the local likelihood method. Specifically, fixing β one estimates $g(\cdot)$ as a function of β and obtain $\hat{g}(\beta)$, which is a nonparametric estimation problem. Then letting $g = \hat{g}(\beta)$, one estimates the parametric component, and this is a parametric problem. Detailed discussions can also be found in Severini and Staniswalis (1994).

Härdle, W., Liang, H. and Sommerfeld, V. (1999) Bootstrap Approximation Of The Estimates For Parameters In A Semiparametric Regression Model

To estimate g for fixed β, let $\omega_{ni}(t) = \omega_{ni}(t; T_1, \ldots, T_n)$ be positive weight functions depending only on the design points $T_1, \ldots, T_n$. Assume $\{X_i = (x_{i1}, \ldots, x_{ip})^{\mathrm{T}}, T_i, Y_i, i = 1, \ldots, n\}$ satisfy the model (1.1). $\hat{g}_\beta(t) = \sum_{j=1}^n \omega_{nj}(t)(Y_j - X_j^{\mathrm{T}}\beta)$ is just the nonparametric estimate of $g(t)$ for fixed β. Giving the estimator $\hat{g}_\beta(t)$, an estimate of β, say β_n, is obtained based on $Y_i = X_i^{\mathrm{T}}\beta + \hat{g}_\beta(T_i) + \varepsilon_i$ for $i = 1, \ldots, n$.

Denote $\tilde{\boldsymbol{X}} = (\tilde{X}_1, \ldots, \tilde{X}_n)^{\mathrm{T}}$, $\tilde{X}_i = X_i - \sum_{j=1}^n \omega_{nj}(T_i)X_j$, $\tilde{\boldsymbol{Y}} = (\tilde{Y}_1, \ldots, \tilde{Y}_n)^{\mathrm{T}}$, $\tilde{Y}_i = Y_i - \sum_{j=1}^n \omega_{nj}(T_i)Y_j$. It follows from the arguments in the last two paragraphs that the least-squares estimate β_n can be expressed as

$$\beta_n = (\tilde{\boldsymbol{X}}^{\mathrm{T}}\tilde{\boldsymbol{X}})^{-1}\tilde{\boldsymbol{X}}^{\mathrm{T}}\tilde{\boldsymbol{Y}}.$$

In addition, the estimate of σ^2, say σ_n^2, is naturally defined as

$$\sigma_n^2 = \frac{1}{n}\sum_{i=1}^n \{Y_i - X_i^{\mathrm{T}}\beta_n - g_n(T_i)\}^2,$$

which is equal to $1/n\sum_{i=1}^n(\tilde{Y}_i - \tilde{X}_i^{\mathrm{T}}\beta_n)^2$, where $g_n(t) = \sum_{j=1}^n \omega_{nj}(t)(Y_j - X_j^{\mathrm{T}}\beta_n)$ is the estimate of $g(t)$.

2. Bootstrap approximations

The statistics β_n and σ_n^2 have asymptotic normal distributions under mild assumptions. In this section, we propose a bootstrap method as an alternative to the normal asymptotic method. For its simplicity resulted from that when we use bootstrap the estimation of nuisance parameters is done automatically.

In the partially linear regression model the observable column n-vector $\hat{\varepsilon}$ of residuals is given by

$$\hat{\varepsilon} = \boldsymbol{Y} - \boldsymbol{G}_n - \boldsymbol{X}^{\mathrm{T}}\beta_n,$$

where $\boldsymbol{G}_n = \{g_n(T_1), \ldots, g_n(T_n)\}^{\mathrm{T}}$. Denote $\mu_n = 1/n\sum_{i=1}^n \hat{\varepsilon}_i$. Let $\hat{F}_n$ be the empirical distribution of $\hat{\varepsilon}$, centered at the mean, so $\hat{F}_n$ puts mass $1/n$ at $\hat{\varepsilon}_i - \mu_n$ and $\int x\,\mathrm{d}\hat{F}_n(x) = 0$. Given $\boldsymbol{Y}$, let $\varepsilon_1^*, \ldots, \varepsilon_n^*$ be conditionally independent with common distribution $\hat{F}_n$. Let ε^* be the n-vector whose ith component is ε_i^*, and let

$$\boldsymbol{Y}^* = \boldsymbol{X}^{\mathrm{T}}\beta_n + \boldsymbol{G}_n + \varepsilon^*.$$

Informally, ε^* is obtained by resampling the centered residuals, and $\boldsymbol{Y}^*$ is generated from the data using the regression model with β_n as the vector of parameters and $\hat{F}_n$ as the distribution of the disturbance terms ε^*.

We define the bootstrap estimates of β and σ^2 as follows:

$$\beta_n^* = (\tilde{\boldsymbol{X}}^{\mathrm{T}}\tilde{\boldsymbol{X}})^{-1}\tilde{\boldsymbol{X}}^{\mathrm{T}}\tilde{\boldsymbol{Y}}^* \quad \text{and} \quad \sigma_n^{2*} = \frac{1}{n}\sum_{i=1}^n(\tilde{Y}_i^* - \tilde{X}_i^{\mathrm{T}}\beta_n^*)^2,$$

where $\tilde{Y}_i^* = Y_i^* - \sum_{j=1}^n \omega_{nj}(T_i)Y_j^*$ and $\tilde{\boldsymbol{Y}}^* = (\tilde{Y}_1^*, \ldots, \tilde{Y}_n^*)^{\mathrm{T}}$.

The bootstrap principle ensures that, as we demonstrate below, the distributions of $\sqrt{n}(\beta_n^* - \beta_n)$ and $\sqrt{n}(\sigma_n^{2*} - \sigma_n^2)$, which can be computed directly from the data, approximate the distributions of $\sqrt{n}(\beta_n - \beta)$ and $\sqrt{n}(\sigma_n^2 - \sigma^2)$, respectively. As will be

shown later, this approximation is likely to be very good, provided n is large enough. This fact is stated later in Theorem 3.1.

3. Main results

We first state the sufficient conditions for our main results.

Condition 1. *There exist functions $h_j(\cdot)$ defined on $[0,1]$ such that*

$$x_{ij} = h_j(T_i) + u_{ij}, \quad 1 \leqslant i \leqslant n, \quad 1 \leqslant j \leqslant p, \tag{3.1}$$

where u_{ij} is a sequence of real numbers which satisfy $\lim_{n\to\infty}(1/n)\sum_{i=1}^{n} u_i = 0$ and

$$\lim_{n\to\infty} \frac{1}{n} \sum_{i=1}^{n} u_i u_i^{\mathrm{T}} = B \tag{3.2}$$

is a positive-definite matrix, and

$$\limsup_{n\to\infty} \frac{1}{a_n} \max_{i \leqslant k \leqslant n} \left| \sum_{i=1}^{k} u_{j_i m} \right| < \infty, \quad m = 1, \ldots, p \tag{3.3}$$

holds for all permutation $(j_1, \ldots, j_n)$ of $(1, \ldots, n)$, where $u_i = (u_{i1}, \ldots, u_{ip})^{\mathrm{T}}$ and $a_n = n^{5/6} \log^{-1} n$.

Condition 2. *$g(\cdot)$ and $h_j(\cdot)$ are Lipschitz continuous of order 1.*

Condition 3. *The weight functions $\omega_{ni}(\cdot)$ satisfy the following*:

(i) $\max_{1 \leqslant i \leqslant n} \sum_{j=1}^{n} \omega_{ni}(T_j) = \mathrm{O}(1)$, $\max_{1 \leqslant j \leqslant n} \sum_{i=1}^{n} \omega_{ni}(T_j) = \mathrm{O}(1)$;
(ii) $\max_{1 \leqslant i,j \leqslant n} \omega_{ni}(T_j) = \mathrm{O}(b_n)$;
(iii) $\max_{1 \leqslant i \leqslant n} \sum_{j=1}^{n} \omega_{nj}(T_i) I_{(|T_j - T_i| > c_n)} = \mathrm{O}(c_n)$,

where $b_n = n^{-2/3}, c_n = n^{-1/3} \log n$.

These conditions are not more complicated than that given in the related literature. They are usually needed for establishing asymptotic normality for the estimators of the parameters. Specifically, imposing Condition 1 in that we can conclude that $(1/n)\tilde{\boldsymbol{X}}^{\mathrm{T}}\tilde{\boldsymbol{X}}$ converges to B. In fact, (3.1) of Condition 1 is analogous to the case

$$h_j(T_i) = E(x_{ij}|T_i) \quad \text{and} \quad u_{ij} = x_{ij} - E(x_{ij}|T_i)$$

when (X_i, T_i) are random variables. Eq. (3.2) is similar to the result of the strong law of large numbers for random errors. Eq. (3.3) is similar to the law of the iterated logarithm. More detailed discussions may be found in Speckman (1988) and Gao et al. (1995).

The weight functions satisfying Condition 3 are presented in Liang et al. (1997) and interested readers can find them there.

Theorem 3.1. *Suppose Conditions* 1–3 *hold. If* $E\varepsilon_1^4 < \infty$ *and* $\max_{1\leqslant i\leqslant n}||u_i|| \leqslant C_0 < \infty$. *Then*

$$\sup_x |P^*\{\sqrt{n}(\beta_n^* - \beta_n) < x\} - P\{\sqrt{n}(\beta_n - \beta) < x\}| \to 0 \tag{3.4}$$

and

$$\sup_x |P^*\{\sqrt{n}(\sigma_n^{2^*} - \sigma_n^2) < x\} - P\{\sqrt{n}(\sigma_n^2 - \sigma^2) < x\}| \to 0 \tag{3.5}$$

hold in probability, where and below P^* *and* E^* *denote the conditional probability and conditional expection given* $\boldsymbol{Y}$.

Up to now, we have showed that the bootstrap method performs at least as well as the normal approximation with the error rate of $o_p(1)$ and $o(1)$, respectively. It is natural to expect that the bootstrap method should perform better than this, however. Our numerical experience means that it is case. In fact, this is true analytically, as shown in the following theorem.

Theorem 3.2. *Let* $M_{jn}(\beta)$ $[(\sigma^2)]$ *and* $M_{jn}^*(\beta)$ $[(\sigma^2)]$ *be the jth moments of* $\sqrt{n}(\beta_n - \beta)$ $[\sqrt{n}(\sigma_n^2 - \sigma^2)]$ *and* $\sqrt{n}(\beta_n^* - \beta_n)[\sqrt{n}(\sigma_n^{2^*} - \sigma_n^2)]$, *respectively. Then under Conditions* 1–3 *and* $E\varepsilon_1^6 < \infty$ *and* $\max_{1\leqslant i\leqslant n}||u_i|| \leqslant C_0 < \infty$,

$$M_{jn}^*(\beta) - M_{jn}(\beta) = O_P(n^{-1/3}\log n) \quad \textit{and} \quad M_{jn}^*(\sigma^2) - M_{jn}(\sigma^2) = O_P(n^{-1/3}\log n)$$

for $j = 1, 2, 3, 4$.

The proof of Theorem 3.2 can be completed by the arguments of Liang (1994) and the similar procedures. We omit the details.

Theorem 3.2 indicates that the bootstrap distributions yield a much better approximation for the first four moments of β_n^* and $\sigma_n^{2^*}$, which are very important quantities in characterizing distributions. Indeed, by Theorem 3.1 and Lemma 3.1 given later, one can only obtain that

$$M_{jn}^*(\beta) - M_{jn}(\beta) = o_P(1) \quad \text{and} \quad M_{jn}^*(\sigma^2) - M_{jn}(\sigma^2) = o_P(1)$$

for $j = 1, 2, 3, 4$, in contrast to Theorem 3.2.

4. Numerical results

In this section, we present a small simulation study in order to illustrate the finite sample behavior of the estimators. We investigate the model

$$Y_i = X_i^{\mathrm{T}}\beta + g(T_i) + \varepsilon_i, \tag{4.1}$$

where $g(T_i) = \sin(T_i)$, $\beta = (1, 5)^{\mathrm{T}}$, and $\varepsilon_i \sim \text{Uniform}(-0.3, 0.3)$. The independent variables $X_i = (X_{i1}, X_{i2})^{\mathrm{T}}$ and T_i are realizations of a Uniform$(0, 1)$ distributed random variable. We analyze the cases of sample sizes 30, 50, 100 and 300. For nonparametric

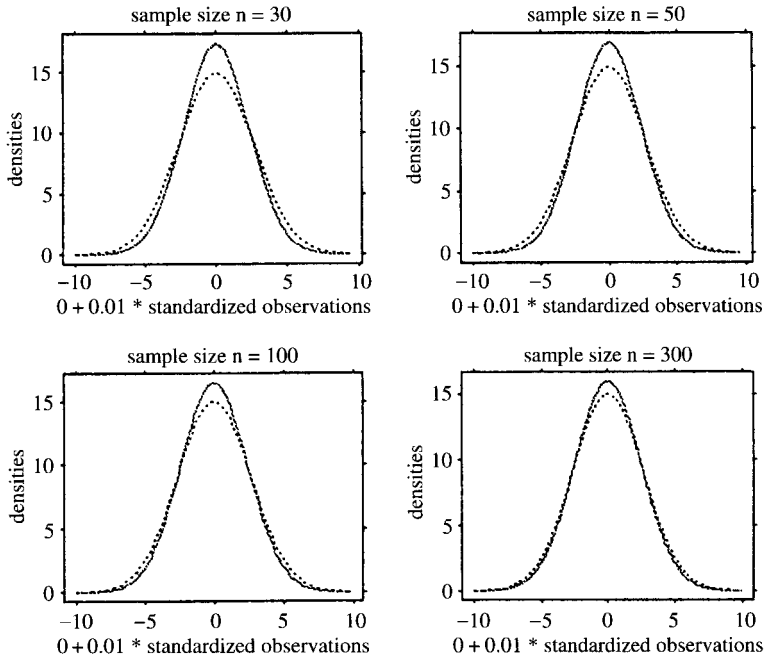


Fig. 1. Plots of the smoothed bootstrap density (solid) and the normal approximation (dashed).

fitting, we use a Nadaraya–Watson kernel weight function with an Epanechnikov kernel. We perform the smoothing with different bandwidths using some grid search. It turns out that the results for the parametric part are quite robust against the bandwidth chosen in the nonparametric part. In the following, we present only the simulation results for the parameter β_2. The discussions for β_1 are similar.

We implement for the cases of sample sizes $30, 50, 100, 300$ and use XploRe (see Härdle et al., 1995) to calculate each case. The asymptotic variance $\sigma^2 B^{-1}$ is estimated by $\hat{\sigma}^2 \hat{B}^{-1}$ with $\hat{B} = 1/n \sum_{i=1}^{n} \tilde{X}_i \tilde{X}_i^{\mathrm{T}}$ and $\hat{\sigma}^2 = 1/n \sum_{i=1}^{n} (\tilde{Y}_i - \tilde{X}_i^{\mathrm{T}} \beta_n)^2$. In Fig. 1, we plot the asymptotic normal distributions and the corresponding bootstrap distributions of the smoothed densities of the estimated true distribution of $\sqrt{n}(\hat{\beta}_2 - \beta_2)/\hat{\sigma}$. It turns out that the bootstrap distribution well approximates the asymptotic normal distribution even for moderate sample sizes of $n = 30$.

5. Proof of Theorem 3.1

We outline the proof. First we decompose $\sqrt{n}(\beta_n - \beta)$ and $\sqrt{n}(\beta_n^* - \beta_n)$ into three terms, and σ_n^2 and $\sigma_n^{2^*}$ into five terms, respectively. Then we calculate the tail probability value of each term. Some notation is introduced. $\varepsilon = (\varepsilon_1, \ldots, \varepsilon_n)^{\mathrm{T}}$, $\tilde{\varepsilon} = (\tilde{\varepsilon}_1, \ldots, \tilde{\varepsilon}_n)^{\mathrm{T}}$, $\tilde{\varepsilon}_i = \varepsilon_i - \sum_{j=1}^{n} \omega_{nj}(T_i)\varepsilon_j$, $\tilde{g}_i = g(T_i) - \sum_{k=1}^{n} \omega_{nk}(T_i) g(T_k)$, $\tilde{\boldsymbol{G}} = (\tilde{g}_1, \ldots, \tilde{g}_n)^{\mathrm{T}}$.

We have from the definitions of β_n and β_n^*, and σ_n^2 and $\sigma_n^{2^*}$,

$$\begin{aligned}\sqrt{n}(\beta_n-\beta) &= \sqrt{n}(\tilde{X}^{\mathrm{T}}\tilde{X})^{-1}(\tilde{X}^{\mathrm{T}}\tilde{G}+\tilde{X}^{\mathrm{T}}\tilde{\varepsilon})\\ &= \sqrt{n}(\tilde{X}^{\mathrm{T}}\tilde{X})^{-1}\left[\sum_{i=1}^{n}\tilde{X}_i\tilde{g}_i-\sum_{i=1}^{n}\tilde{X}_i\left\{\sum_{j=1}^{n}\omega_{nj}(T_i)\varepsilon_j\right\}+\sum_{i=1}^{n}\tilde{X}_i\varepsilon_i\right]\\ &\stackrel{\mathrm{def}}{=} n(\tilde{X}^{\mathrm{T}}\tilde{X})^{-1}(H_1-H_2+H_3),\end{aligned}$$

$$\begin{aligned}\sqrt{n}(\beta_n^*-\beta_n) &= \sqrt{n}(\tilde{X}^{\mathrm{T}}\tilde{X})^{-1}(\tilde{X}^{\mathrm{T}}\tilde{G}_n^*+\tilde{X}^{\mathrm{T}}\tilde{\varepsilon}^*)\\ &= \sqrt{n}(\tilde{X}^{\mathrm{T}}\tilde{X})^{-1}\left[\sum_{i=1}^{n}\tilde{X}_i\tilde{g}_{ni}^*-\sum_{i=1}^{n}\tilde{X}_i\left\{\sum_{j=1}^{n}\omega_{nj}(T_i)\varepsilon_j^*\right\}+\sum_{i=1}^{n}\tilde{X}_i\varepsilon_i^*\right]\\ &\stackrel{\mathrm{def}}{=} n(\tilde{X}^{\mathrm{T}}\tilde{X})^{-1}(H_1^*-H_2^*+H_3^*),\end{aligned}$$

where $\tilde{G}_n^*=(\tilde{g}_{n1}^*,\ldots,\tilde{g}_{nn}^*)^{\mathrm{T}}$ with $\tilde{g}_{ni}^*=g_n(T_i)-\sum_{k=1}^{n}\omega_{nk}(T_i)g_n(T_k)$ for $i=1,\ldots,n$.

$$\begin{aligned}\sigma_n^2 &= \frac{1}{n}\tilde{Y}^{\mathrm{T}}\left\{\mathscr{F}-\tilde{X}(\tilde{X}^{\mathrm{T}}\tilde{X})^{-1}\tilde{X}^{\mathrm{T}}\right\}\tilde{Y}\\ &= \frac{1}{n}\varepsilon^{\mathrm{T}}\varepsilon-\frac{1}{n}\varepsilon^{\mathrm{T}}\tilde{X}(\tilde{X}^{\mathrm{T}}\tilde{X})^{-1}\tilde{X}^{\mathrm{T}}\varepsilon+\frac{1}{n}\tilde{G}^{\mathrm{T}}\{\mathscr{F}-\tilde{X}(\tilde{X}^{\mathrm{T}}\tilde{X})^{-1}\tilde{X}^{\mathrm{T}}\}\tilde{G}\\ &\quad-\frac{2}{n}\tilde{G}^{\mathrm{T}}\tilde{X}(\tilde{X}^{\mathrm{T}}\tilde{X})^{-1}\tilde{X}^{\mathrm{T}}\varepsilon+\frac{2}{n}\tilde{G}^{\mathrm{T}}\varepsilon\\ &\stackrel{\mathrm{def}}{=} I_1-I_2+I_3-2I_4+2I_5,\end{aligned}$$

$$\begin{aligned}\sigma_n^{2^*} &= \frac{1}{n}\tilde{Y}^{*\mathrm{T}}\left\{\mathscr{F}-\tilde{X}(\tilde{X}^{\mathrm{T}}\tilde{X})^{-1}\tilde{X}^{\mathrm{T}}\right\}\tilde{Y}^*\\ &= \frac{1}{n}\varepsilon^{*\mathrm{T}}\varepsilon^*-\frac{1}{n}\varepsilon^{*\mathrm{T}}\tilde{X}(\tilde{X}^{\mathrm{T}}\tilde{X})^{-1}\tilde{X}^{\mathrm{T}}\varepsilon^*+\frac{1}{n}\tilde{G}_n^{*\mathrm{T}}(\mathscr{F}-\tilde{X}(\tilde{X}^{\mathrm{T}}\tilde{X})^{-1}\tilde{X}^{\mathrm{T}})\tilde{G}_n^*\\ &\quad-\frac{2}{n}\tilde{G}_n^{*\mathrm{T}}\tilde{X}(\tilde{X}^{\mathrm{T}}\tilde{X})^{-1}\tilde{X}^{\mathrm{T}}\varepsilon^*+\frac{2}{n}\tilde{G}_n^{*\mathrm{T}}\varepsilon^*\\ &\stackrel{\mathrm{def}}{=} I_1^*-I_2^*+I_3^*-2I_4^*+2I_5^*.\end{aligned}$$

Here $\mathscr{F}$ is the identity matrix of order p. In the appendix we shall prove that $H_{1j},H_{2j}=o_P(1)$ and $H_{1j}^*,H_{2j}^*=o_{P^*}(1)$ for $j=1,\ldots,p$, and $I_i=o_P(n^{-1/2})$ and $I_i^*=o_{P^*}(n^{-1/2})$ for $i=2,3,4,5$.

From (A.3) and Lemma A.6, we only need to prove $\sqrt{n}(\tilde{X}^{\mathrm{T}}\tilde{X})^{-1}\tilde{X}^{\mathrm{T}}\varepsilon^*$ converges in distribution to a p-variate normal random variate with mean 0 and covariance matrix σ^2B^{-1}.

Let q_{ii} be the ith diagonal element of the matrix $\tilde{X}(\tilde{X}^{\mathrm{T}}\tilde{X})^{-1}\tilde{X}^{\mathrm{T}}$. According to Proposition 2.2 of Huber (1973), if we know $\max_i q_{ii}\to 0$ as $n\to\infty$, then $\sqrt{n}(\tilde{X}^{\mathrm{T}}\tilde{X})^{-1}\tilde{X}^{\mathrm{T}}\varepsilon^*$ is asymptotically normal. Since the covariance matrix of $\sqrt{n}(\tilde{X}^{\mathrm{T}}\tilde{X})^{-1}\tilde{X}^{\mathrm{T}}\varepsilon^*$ is given by $n(\tilde{X}^{\mathrm{T}}\tilde{X})^{-1}\int u^2\,\mathrm{d}\hat{F}_n(u)$, recalling the definition of $\hat{F}_n(u)$ and the result given in

Härdle, W., Liang, H. and Sommerfeld, V. (1999) Bootstrap Approximation Of The Estimates For Parameters In A Semiparametric Regression Model

Lemma A.2, the asymptotic variance of $\sqrt{n}(\beta_n^* - \beta_n)$ is $\sigma^2 B^{-1}$. Notice the definition of q_{ii}. Since $n^{-1}(\tilde{X}^{\mathrm{T}}\tilde{X}) \to B$ by Lemma A.2, it follows from Lemma 3 of Wu (1981) that $\max_i q_{ii} \to 0$. This completes the proof of (3.4).

We now prove (3.5). First we give the following preliminary results. In Lemma A.5, letting V_i be ε_i^*, E and P be E^* and P^*, we have

$$\max_{1\leqslant i\leqslant n}\left|\sum_{k=1}^{n}\omega_{nk}(T_i)\varepsilon_k^*\right| = O_{P^*}(n^{-1/4}\log^{-1/2} n).$$

This equation, Lemma A.3 and the fact

$$|\sqrt{n}I_3^*| \leqslant C\sqrt{n}\max_{1\leqslant i\leqslant n}\left\{\left|g_n(T_i) - \sum_{k=1}^{n}\omega_{nk}(T_i)g_n(T_k)\right|^2 + \left|\sum_{k=1}^{n}\omega_{nk}(T_i)\varepsilon_k^*\right|^2\right\}$$

yield that $|\sqrt{n}I_3^*| = o_{P^*}(1)$.

Using similar arguments as for proving $\sqrt{n}(\tilde{X}^{\mathrm{T}}\tilde{X})^{-1}\tilde{X}^{\mathrm{T}}\varepsilon^* \to N(0, \sigma^2 B^{-1})$, one concludes that

$$\sqrt{n}I_2^* = o_{P^*}(1), \quad \sqrt{n}I_4^* = o_{P^*}(1).$$

We decompose I_5^* into three terms, and prove that each term tends to zero. More precisely,

$$\begin{aligned} I_5^* &= \frac{1}{n}\sum_{i=1}^{n}\tilde{g}_{ni}^*\varepsilon_i^* - \frac{1}{n}\sum_{k=1}^{n}\omega_{ni}(T_k)\varepsilon_k^{*2} - \frac{1}{n}\sum_{i=1}^{n}\sum_{k\neq i}^{n}\omega_{ni}(T_k)\varepsilon_i^*\varepsilon_k^* \\ &\stackrel{\mathrm{def}}{=} I_{51}^* + I_{52}^* + I_{53}^*. \end{aligned} \tag{5.1}$$

It follows from Lemma A.3 and Chebychev's inequality that

$$\sqrt{n}I_{51}^* = o_{P^*}(1), \tag{5.2}$$

and

$$\sqrt{n}I_{52}^* \leqslant b_n\sqrt{n}\sum_{i=1}^{n}\varepsilon_i^{*2} = O_{P^*}(\log^{-2} n) = o_{P^*}(1). \tag{5.3}$$

Denote $\varepsilon_j^{*'} = \varepsilon_j^* I(|\varepsilon_j^*| \leqslant n^{1/4})$ and $\varepsilon_j^{*''} = \varepsilon_j^* - \varepsilon_j^{*'}$ for $j = 1, \ldots, n$. Let $I_n^* = \sum_{i=1}^{n}\sum_{j\neq i}\omega_{nj}(T_i)(\varepsilon_j^{*'} - E\varepsilon_j^{*'})(\varepsilon_i^{*'} - E\varepsilon_i^{*'})$. Observe that

$$\sqrt{n}|I_{53}^*| \leqslant \frac{1}{\sqrt{n}}\left|\sum_{i=1}^{n}\sum_{k\neq i}\omega_{nk}(T_i)\varepsilon_i^*\varepsilon_k^* - I_n^*\right| + I_n^* \stackrel{\mathrm{def}}{=} \frac{1}{\sqrt{n}}(J_{1n}^* + I_n^*).$$

Simple calculations show that

$$\begin{aligned} J_{1n}^* &\leqslant \max_{1\leqslant j\leqslant n}\left|\sum_{i=1}^{n}\omega_{nj}(T_i)\varepsilon_i^*\right|\left(\sum_{i=1}^{n}|\varepsilon_i^{*''}| + E|\varepsilon_i^{*''}|\right) \\ &\quad + \max_{1\leqslant j\leqslant n}\left|\sum_{i=1}^{n}\omega_{nj}(T_i)(\varepsilon_i^{*'} - E\varepsilon_i^{*'})\right|\left(\sum_{i=1}^{n}|\varepsilon_i^{*''}| + E|\varepsilon_i^{*''}|\right) \\ &= o_{P^*}(1). \end{aligned}$$

Letting ε_i^* be V_i, E^* and P^* be E and P in Lemma A.7, respectively, we have $I_n^* = o_{P^*}(\sqrt{n})$. It follows that

$$\sqrt{n} I_{53}^* = o_{P^*}(1). \tag{5.4}$$

A combination of (5.2)–(5.4) implies that $\sqrt{n} I_5^* = o_{P^*}(1)$.

From the above arguments and the third result of Lemma A.1, the proof of (3.5) is equivalent to showing

$$\frac{1}{\sqrt{n}} \left\{ \varepsilon^{*\mathrm{T}} \varepsilon^* - \int u^2 \, \mathrm{d}\hat{F}_n(u) \right\} \to N(0, \operatorname{Var} \varepsilon_1^2),$$

which can be verified by a central limit theorem. We therefore complete the proof of Theorem 3.1. □

Acknowledgements

The authors thank Dr. Michael Neumann for his valuable suggestions and comments. The authors are grateful to two referees for their valuable comments which greatly improve the presentation of this paper.

Appendix. Some lemmas

Under the conditions of Theorem 3.1, Gao et al. (1995) obtained asymptotic normality of β_n and σ_n^2 and the convergence rate of g_n, which are given in Lemma A.1. Lemma A.2 presents the limit of $1/n\tilde{\boldsymbol{X}}^{\mathrm{T}}\tilde{\boldsymbol{X}}$. The proof is found in Chen (1988) and Speckman (1988). Lemma A.3 provides the boundedness for $g(T_i) - \sum_{k=1}^n \omega_{nk}(T_i) g(T_k)$ and $g_n(T_i) - \sum_{k=1}^n \omega_{nk}(T_i) g_n(T_k)$, whose proofs are immediate. Lemma A.4 shows that $\sqrt{n} H_{1j}$ and $\sqrt{n} H_{1j}^*$ are $O(n^{1/3} \log n)$ in different probability senses. Lemma A.5 gives a general result for nonparametric regression, whose proof depends on an exponential inequality, Bernstein's inequality, for bounded independent random variables.

Lemma A.1. *Suppose that the conditions of Theorem* 3.1 *hold. Then*

$$\sqrt{n}(\beta_n - \beta) \to \mathrm{N}(0, \sigma^2 B^{-1}), \quad \sup_{t \in [0,1]} |g_n(t) - g(t)| = \mathrm{O_p}(n^{-1/3} \log n), \tag{A.1}$$

and

$$\sqrt{n}(\sigma_n^2 - \sigma^2) \to \mathrm{N}(0, \operatorname{Var}(\varepsilon_1^2)). \tag{A.2}$$

Lemma A.2. *If Conditions* 1–3 *hold, then*

$$\lim_{n \to \infty} \frac{1}{n} \tilde{\boldsymbol{X}}^{\mathrm{T}} \tilde{\boldsymbol{X}} = B.$$

Lemma A.3. *Suppose that Conditions* 2 *and* 3(iii) *hold. Then*

$$\max_{1\leqslant i\leqslant n}\left|g(T_i)-\sum_{k=1}^{n}\omega_{nk}(T_i)g(T_k)\right|=\mathrm{O}(n^{-1/3}\log n),$$

$$\max_{1\leqslant i\leqslant n}\left|g_n(T_i)-\sum_{k=1}^{n}\omega_{nk}(T_i)g_n(T_k)\right|=\mathrm{O_P}(n^{-1/3}\log n).$$

The same conclusion as the first part holds for $h_j(T_i)-\sum_{k=1}^{n}\omega_{nk}(T_i)h_j(T_k)$ *for* $j=1,\ldots,p$.

Lemma A.4. *Suppose that Conditions* 1–3 *hold and* $E|\varepsilon_1|^3<\infty$. *Then*

$$\sqrt{n}H_{1j}=\mathrm{O}(n^{1/2}\log^{-1/2}n)\quad\textit{and}\quad\sqrt{n}H_{1j}^{*}=\mathrm{O}(n^{1/2}\,log^{-1/2}n)\quad\textit{for } j=1,\ldots,p. \tag{A.3}$$

Proof. Their proofs can be completed by the same methods as Lemmas 2.4 and 2.5 of Liang (1999). We omit the details. □

Bernstein's inequality: Let $V_1,\ldots,V_n$ be independent random variables with zero means and bounded ranges: $|V_i|\leqslant M$. Then for each $\eta>0$,

$$P\left\{\left|\sum_{i=1}^{n}V_i\right|>\eta\right\}\leqslant 2\exp\left\{-\eta^2\Big/\left[2\left(\sum_{i=1}^{n}\operatorname{var}V_i+M\eta\right)\right]\right\}.$$

Lemma A.5. *Suppose that Condition* 3 *holds. Let* V_i *be independent with mean zero and* $\sup_i EV_i^4<\infty$, *then*

$$\max_{1\leqslant i\leqslant n}\left|\sum_{k=1}^{n}\omega_{nk}(T_i)V_k\right|=\mathrm{O_P}(n^{-1/4}\log^{-1/2}n).$$

Proof. Denote $V_j'=V_jI_{(|V_j|\leqslant n^{1/4})}$ and $V_j''=V_j-V_j'$ for $j=1,\ldots,n$. Let $M=Cb_nn^{1/4}$ for b_n given in Condition 3. From Bernstein's inequality

$$\begin{aligned}
&P\left\{\max_{1\leqslant i\leqslant n}\left|\sum_{j=1}^{n}\omega_{nj}(T_i)(V_j'-EV_j')\right|>C_1n^{-1/4}\log^{-1/2}n\right\}\\
&\quad\leqslant\sum_{i=1}^{n}P\left\{\left|\sum_{j=1}^{n}\omega_{nj}(T_i)(V_j'-EV_j')\right|>C_1n^{-1/4}\log^{-1/2}n\right\}\\
&\quad\leqslant 2n\exp\left\{-\frac{C_1n^{-1/2}\log^{-1}n}{\sum_{j=1}^{n}\omega_{nj}^2(T_i)EV_j^2+2CC_1b_n\log^{-1/2}n}\right\}\\
&\quad\leqslant 2n\exp\{-C_1^2C\log n\}\leqslant C_n^{-1/2}\quad\text{for some large } C_1>0.
\end{aligned}$$

This implies that

$$\max_{1\leqslant i\leqslant n}\left|\sum_{j=1}^{n}\omega_{nj}(T_i)(V_j'-EV_j')\right|=\mathrm{O_P}(n^{-1/4}\log^{-1/2}). \tag{A.4}$$

On the other hand, we know

$$\max_{1\leqslant i\leqslant n}\left|\sum_{j=1}^{n}\omega_{nj}(T_i)EV_j''\right| \leqslant \max_{1\leqslant k\leqslant n}\max_{1\leqslant i\leqslant n}|\omega_{nk}(T_i)|\sum_{j=1}^{n}n^{-3/4}E|V_j|^4$$

$$\leqslant Cn^{-5/12}\log n \max_{1\leqslant i\leqslant n}E|V_i|^4$$

$$= o(n^{-1/4}\log^{-1/2}n) \tag{A.5}$$

and

$$\sum_{j=1}^{n}E|\varepsilon_j''|\leqslant n^{-3/4}\sum_{j=1}^{n}E\varepsilon_j^4\leqslant n^{1/4}\sup_i E|\varepsilon_i|^4. \tag{A.6}$$

Moreover, the Hartman–Winter theorem yields that

$$\sum_{j=1}^{n}(|\varepsilon_j''|-E|\varepsilon_j''|)=O\left[\left\{\sum_{j=1}^{n}E|\varepsilon_j''|^2\log\log\left(\sum_{j=1}^{n}E|\varepsilon_j''|^2\right)\right\}^{1/2}\right]$$

$$= O_P(n^{1/4}(\log\log n)^{1/2}). \tag{A.7}$$

It follows from (A.6) and (A.7) that

$$\sum_{j=1}^{n}|\varepsilon_j''| = O_P(n^{1/4}(\log\log n)^{1/2}), \tag{A.8}$$

and

$$\max_{1\leqslant i\leqslant n}\left|\sum_{j=1}^{n}\omega_{nj}(T_i)\varepsilon_j''\right| \leqslant \max_{1\leqslant k,i\leqslant n}|\omega_{nk}(T_i)|\sum_{j=1}^{n}|\varepsilon_j''|$$

$$= O(n^{-5/12}(\log\log n)^{1/2}) = O_P(n^{-1/4}(\log n)^{-1/2}). \tag{A.9}$$

Combining the results of (A.4), (A.5), (A.7), and (A.9) we obtain

$$\max_{1\leqslant i\leqslant n}\left|\sum_{k=1}^{n}\omega_{nk}(T_i)V_k\right| = O_P(n^{-1/4}\log^{-1/2}n). \tag{A.10}$$

This completes the proof of Lemma A.5. □

Lemma A.6. *Suppose that Conditions* 1–3 *hold and* $E|\varepsilon_1|^4<\infty$. *Then*

$$\sqrt{n}H_{2j} = o(n^{1/2}) \quad \textit{and} \quad \sqrt{n}H_{2j}^* = o_{P^*}(n^{1/2}) \quad \textit{for } j=1,\ldots,p.$$

Proof. Denote $h_{nij}=h_j(T_i)-\sum_{k=1}^{n}\omega_{nk}(T_i)h_j(T_k)$. Observe the fact

$$\sqrt{n}H_{2j}=\sum_{i=1}^{n}\left\{\sum_{k=1}^{n}\tilde{x}_{kj}\omega_{ni}(T_k)\right\}\varepsilon_i$$

$$= \sum_{i=1}^{n} \left\{ \sum_{k=1}^{n} u_{kj} \omega_{ni}(T_k) \right\} \varepsilon_i + \sum_{i=1}^{n} \left\{ \sum_{k=1}^{n} h_{nkj} \omega_{ni}(T_k) \right\} \varepsilon_i$$

$$- \sum_{i=1}^{n} \left[\sum_{k=1}^{n} \left\{ \sum_{q=1}^{n} u_{qj} \omega_{nq}(T_k) \right\} \omega_{ni}(T_k) \right] \varepsilon_i.$$

Using Conditions 3(i) and (ii) and the remark in Lemma A.3, we handle each term as (A.10) by letting $V_i = \varepsilon_i$ in Lemma A.5. Each term can be proved to be $o_P(n^{1/2})$ by using Lemma A.5 and the arguments for proving Lemma A.5. The same technique is also applied to $\sqrt{n} H_{2j}^*$. We omit the details. □

Lemma A.7. *Under the conditions of Lemma* A.5. $I_n = o_P(n^{1/2})$, *where*

$$I_n = \sum_{i=1}^{n} \sum_{j \neq i} \omega_{nj}(T_i)(V_j' - EV_j')(V_i' - EV_i').$$

Proof. Let $j_n = [n^{1/2} \log^2 n]$, ($[a]$ denotes the integer portion of a). $A_j = \{[(j-1)n/j_n] + 1, \ldots, [jn/j_n]\}$, $A_j^c = \{1, 2, \ldots, n\} - A_j$ and $A_{ji} = A_j - \{i\}$. Then I_n can be decomposed as follows:

$$I_n = \sum_{j=1}^{j_n} \sum_{i \in A_j} \sum_{k \in A_{ji}} \omega_{nk}(T_i)(V_k' - EV_k')(V_i' - EV_i')$$

$$+ \sum_{j=1}^{j_n} \sum_{i \in A_j} \sum_{k \in A_j^c} \omega_{nk}(T_i)(V_k' - EV_k')(V_i' - EV_i')$$

$$\stackrel{\text{def}}{=} \sum_{j=1}^{j_n} U_{nj} + \sum_{j=1}^{j_n} V_{nj}$$

$$\stackrel{\text{def}}{=} I_{1n} + I_{2n}, \tag{A.11}$$

where

$$U_{nj} = \sum_{i \in A_j} p_{nij}(V_i' - EV_i') \stackrel{\text{def}}{=} \sum_{i \in A_j} u_{nij},$$

$$V_{nj} = \sum_{i \in A_j} q_{nij}(V_i' - EV_i') \stackrel{\text{def}}{=} \sum_{i \in A_j} v_{nij}$$

and

$$p_{nij} = \sum_{k \in A_{ji}} \omega_{nk}(T_i)(V_k' - EV_k'),$$

$$q_{nij} = \sum_{k \in A_j^c} \omega_{nk}(T_i)(V_i' - EV_i').$$

Notice that $\{v_{nij}, i \in A_j\}$ are conditionally independent random variables given $E_{nj} = \{V_k, k \in A_j^c\}$ with $E(v_{nij}|E_{nj}) = 0$ and $E(v_{nij}^2|E_{nj}) \leqslant \sigma^2(\max_{1 \leqslant i \leqslant n}|q_{nij}|^2) \stackrel{\text{def}}{=} \sigma^2 q_{nj}^2$ for $i \in A_j$, and satisfy $\max_{1 \leqslant i \leqslant n}|v_{nij}| \leqslant 2n^{1/4} q_{nj}$ for $q_{nj} = \max_{1 \leqslant i \leqslant n}|q_{nij}|$.

On the other hand, by the same reason as that for Lemma A.5,

$$q_n = \max_{1\leqslant j\leqslant j_n} |q_{nj}| = \max_{1\leqslant j\leqslant j_n} \max_{1\leqslant i\leqslant n} \left| \sum_{k\in A_j^c} \omega_{nk}(T_i)(V_k' - EV_k') \right|$$
$$= O_P(n^{-1/4}\log^{-1/2} n).$$

Denote the numbers of the elements in A_j by $\#A_j$. Take $M = 2n^{1/4}q_n$ and $\eta = Cn^{1/2}j_n^{-1} \times (\log n)^{-1/2}$.

By applying Bernstein's inequality and the fact that $\#A_j \leqslant n/j_n$, we have, for $j = 1,\ldots,j_n$,

$$P\left\{|V_{nj}| > \left|\frac{C\sqrt{n}}{\sqrt{\log n}\, j_n}\right| E_{nj}\right\} \leqslant 2\exp\left\{-\frac{C^2 n\,(\log^{-1} n) j_n^{-2}}{\sigma^2 q_n^2 \#A_j + \eta n^{1/4} q_n}\right\}$$
$$\leqslant 2\exp(-C^2 n^{1/2} j_n^{-1}) \leqslant Cn^{-1/2}.$$

It follows from the bounded dominant convergence theorem that

$$P\left\{|V_{nj}| > \frac{C\sqrt{n}}{\sqrt{\log n}\, j_n}\right\} \leqslant Cn^{-1/2} \quad \text{for } j = 1,\ldots,j_n.$$

Then

$$I_{2n} = o_P(\sqrt{n}). \tag{A.12}$$

Noting that $\{V_k, 1\leqslant k\leqslant n\}$ are i.i.d. random variables, and the definition of U_{nj}, we know that

$$P\{|I_{1n}| > C\sqrt{n}(\log^{-1/2} n)\} \leqslant Cn^{-1}(\log n) E\left\{\sum_{j=1}^{j_n} U_{nj}\right\}^2$$
$$= Cn(\log^{-1} n)\left(\sum_{j=1}^{j_n} EU_{nj}^2 + \sum_{j_1\neq j_2}^{j_n} |EU_{nj_1}EU_{nj_2}|\right)$$
$$\leqslant Cn^{-1}(\log n)(j_n + j_n^2)(\#A_j)^2 b_n^2 [E(V_1' - EV_1')^4 + \{E(V_1' - EV_1')^2\}^2]$$
$$\leqslant Cn^{-1/2}. \tag{A.13}$$

Hence $I_{1n} = o_P(\sqrt{n})$. Combining (A.11), (A.12) with (A.13), we complete the proof of Lemma A.7. □

References

Chen, H., 1988. Convergence rates for parametric components in a partly linear model. Ann. Statist. 16, 136–146.

Cuzick, J., 1992a. Semiparametric additive regression. J. Roy. Statist. Soc. Ser. B 54, 831–843.

Cuzick, J., 1992b. Efficient estimates in semiparametric additive regression models with unknown error distribution. Ann. Statist. 20, 1129–1136.

Efron, B., Tibshirani, R.J., 1993. An Introduction to the Bootstrap. Chapman & Hall, New York.

Engle, R.F., Granger, C.W.J., Rice, J., Weiss, A., 1986. Semiparametric estimates of the relation between weather and electricity sales. J. Amer. Statist. Assoc. 81, 310–320.

Härdle, W., Liang, H. and Sommerfeld, V. (1999) Bootstrap Approximation Of The Estimates For Parameters In A Semiparametric Regression Model

Gao, J.T., Hong, S.Y., Liang, H., 1995. Convergence rates of a class of estimates in partly linear models. Acta Math. Sinica Ser. New 38, 658–669.

Härdle, W., Klinke, S., Turlach, B.A., 1995. XploRe: An Interactive Statistical Computing Environment. Springer, Berlin.

Härdle, W., Mammen, E., 1993. Comparing nonparametric versus parametric regression fits. Ann. Statist. 21, 1926–1947.

Heckman, N.E., 1986. Spline smoothing in partly linear models. J. Roy. Statist. Soc. Ser. B 48, 244–248.

Hong, S.Y., Cheng, P., 1993. Bootstrap approximation of estimation for parameter in a semiparametric regression model. Sci. China Ser. A 23, 239–251.

Huber, P.J., 1973. Robust regression: asymptotic, conjectures and Monte-Carlo. Ann. Statist. 1, 799–821.

Liang, H., 1994. The Berry–Esseen bounds of error variance estimation in a semiparametric regression model. Commun. Statist. Theory Methods 23, 3439–3452.

Liang, H., 1999. Law of the iterated logarithm for parameters in a partly linear regression model. Taiwanese J. Math. 3, 517–528.

Liang, H., Cheng, P., 1993. Second order asymptotic efficiency in a partial linear model. Statist. Probab. Lett. 18, 73–84.

Liang, H., Härdle, W., Werwatz, A., 1997. Asymptotic properties of nonparametric regression estimation in partly linear models. Discussion paper no. 55, Institut für Statistik und Ökonometrie, Humboldt-Universität zu Berlin.

Schick, A., 1993. On efficient estimation in regression models. Ann. Statist. 21, 1486–1521. (Correction and Addendum 23, 1862–1863.)

Severini, T.A., Staniswalis, J.G., 1994. Quasilikelihood estimation in semiparametric models. J. Amer. Statist. Assoc. 89, 501–511.

Speckman, P., 1988. Kernel smoothing in partial linear models. J. Roy. Statist. Soc. Ser. B 50, 413–436.

Wu, C.F.J., 1981. Asymptotic theory of nonlinear least squares estimation. Ann. Statist. 9, 501–513.

Wu, C.F.J., 1986. Jackknife, Bootstrap and other resampling methods in regression analysis (with discussion). Ann. Statist. 14, 1261–1295.

Härdle, W., Liang, H. and Sommerfeld, V. (1999) Bootstrap Approximation Of The Estimates For Parameters In A Semiparametric Regression Model

ELSEVIER

Journal of Econometrics 95 (2000) 333–345

JOURNAL OF Econometrics

www.elsevier.nl/locate/econbase

Internet-based econometric computing

W. Härdle[a,*], J. Horowitz[b]

[a]*Institut für Statistik und Ökonometrie, Wirtschaftswissenschaftliche Fakultät, Spandauer Straße 1, Humboldt-Universität zu Berlin, 10178 Berlin, Germany*

[b]*Department of Economics, University of Iowa, Iowa City, IA 52242, USA*

Abstract

Modern econometrics requires implementation of highly specialized software. In contrast to mathematical arguments used in implementing new econometric techniques the corresponding software algorithms require specific platforms. The specialization of hardware and software, in fact, seriously impedes the adoption of new methods in applied research. It complicates the proliferation of new techniques and makes it difficult to motivate students to use the methods and to help students to develop an intuitive understanding of the methods in applications. We discuss the potential for reducing these problems through Internet-based econometric computing and instruction. We refer to existing examples of net-based teaching and present concrete examples for interactive teaching of elementary econometrics and statistics. © 2000 Published by Elsevier Science S.A. All rights reserved.

JEL classification: C19; C63; C87

Keywords: Computational techniques; Internet-based methods; Java; XploRe; Econometrics software

1. Introduction

An important characteristic of modern econometrics is the rapid development of new, mathematically complex methods whose implementation requires highly

* Corresponding author.

E-mail address: haerdle@wiwi.hu-berlin.de (W. Härdle)

0304-4076/00/$ - see front matter
PII: S0304-4076(99)00041-X

Journal Of Econometrics 95 (2000), pp. 333-345

specific software and, sometimes, hardware. Software is typically written in the developer's favorite proprietary language and is available to others, if at all, only if they own the proprietary system. For example a program written in GAUSS can be run only by a user who has the GAUSS system on his computer and preferably has the same type of computer too. We believe that this specialization of hardware and software seriously impedes the adoption of new methods in applied research. Moreover, it greatly complicates the tasks of teaching students to use the methods and helping students to develop an intuitive understanding of the performance of the methods in applications. This paper discusses the potential for reducing these problems through Internet-based econometric computing and instruction. We describe an existing network architecture and give concrete examples for interactive teaching of an elementary statistics course and for the dissemination of semiparametric methods.

Section 2 provides a more detailed discussion of the heterogeneity of hardware and software that creates barriers between researchers, potential users of new methods and students. We point out, however, that specialization of hardware and software can have benefits because it permits the developer to take advantage of unique features of hardware or software systems that are especially suited to the problem being solved. The methods for Internet-based computing that we discuss preserve these benefits. Section 3 explains what Internet-based computing is and how it differs from the existing technology of downloading software from worldwide web (WWW) sites. Section 4 describes an existing architecture that implements these ideas and is in use in undergraduate teaching at the Humboldt-Universität zu Berlin and several other universities. Section 5 presents concluding comments.

2. The problem of heterogeneity

Three hardware platforms are in widespread use for statistical computing and graphical data interaction: Macintosh, UNIX, Windows. The first has a simple graphically oriented user interface and allows highly interactive dialogues with data. UNIX is used for high-speed and distributed computing but is often less satisfactory in graphical interaction. Windows aims at facilitating both high-speed computing and graphics but is weaker at present than UNIX for Internet access. Distributed computing is simply not possible under Windows unless one uses certain add-ons.

Many software platforms for econometric computing exist but are not easily interchangeable. The reasons for this include the history of software development, the targeted user groups, and the optimization of certain software for specific hardware configurations. The original version of GAUSS (http://www.aptech.com), for example, was optimized for INTEL chips and, therefore, could

Journal Of Econometrics 95 (2000), pp. 333-345

not be transferred to Macs or UNIX platforms. Now GAUSS is available on UNIX, but the UNIX version does not have a graphical device that would allow for, e.g. interactive changes in the layout of graphs. SPLUS (http://www.mathsoft.com/Splus) was developed for UNIX systems and was only later transferred to PCs. Consequently, the PC version is different from the UNIX version. EVIEWS was developed for DOS and is now available for Windows but not for UNIX or Macs. TSP is a DOS program and is not easily transferred to a Windows/NT platform. SPSS exists for Windows but has still a batch structure that makes many mouse clicks necessary in order to generate implicitly the batch commands. STATA (http://www.stata.com), SAS (http://www.sas.com/) and SHAZAM (http://shazam.econ.ubc.ca) are unusual in that mutually compatible versions exist for all platforms. Besides the software that we mentioned here as examples, there are many other platforms, but they are similar in being different.

Heterogeneity of software platforms creates no problems if there is no need to exchange programs. Exchange of graphs, document files, and ASCII-based data sets can be carried out by FTP, provided that the user has the appropriate graphics plug-in and document reader (e.g., Ghostscript or Acrobat). However, there is also a need for exchangeable computer programs for implementing advanced econometric methods, as these are becoming increasingly complex mathematically, and writing the necessary programs can be a difficult and time-consuming task.

Graduate-level instruction in econometrics provides one example of the usefulness of exchangeability. It is not unusual for a faculty member at one university to give a short course at another. In some cases, a faculty member at one university may use electronic communication to present a course at several geographically dispersed locations. An econometric estimator may require heavy computing that is available on the researcher's home machine. During the course, modifications of this estimator and different applications may be discussed, and these may require access to the software at multiple locations. Exchangeability of software is necessary to enable students at all locations to carry out computational and empirical exercises that the instructor has prepared at his own university.

Collaboration among researchers at different locations provides another example of the desirability of exchangeability. In this case, the goal is to enable each collaborator to carry out computations using the same software. Ideally such cooperation should be based on a pool of easily accesible software and computing power for all parties. For effective progress on a project that involves heterogeneous hardware and software, it is desirable for partners to have the ability to contribute methods despite being at different locations and working with different computing environments. In addition, it may simply be a problem for a researcher who is a visitor in another establishment to be able to continue using his own programs.

Journal Of Econometrics 95 (2000), pp. 333-345

On the other hand, heterogeneity of software does have the important advantage of enabling a developer of new methods to choose the software system that is best suited to the problem under consideration. Therefore the problem of exchangeability should not be solved by standardizing econometric software but by making software from different sources accessible to diversely equipped users.

3. Internet-based econometric methods

The phrase 'Internet-based econometric methods' can refer to several different concepts. One concept is to maintain applications programs on a server for users who have access to the software systems required to run the programs. The programs are downloaded by FTP and executed on the user's computer. An example is the STATLIB server of SPLUS at Carnegie-Mellon University (http://lib.stat.cmu.edu/S/). Other examples include the ELSA archive at the University of California, Berkeley (http://elsa.berkeley.edu/) and the CodEc software archive (http://netec.mcc.ac.uk/CodEc.html). These sites and others like them permit a user to retrieve programs quickly. The user must, however, have access to the software system that executes the programs (e.g. SPLUS for programs written in that language, GAUSS, FORTRAN, etc.).

STATLIB and ELSA also illustrate two different archival policies that are worth noting. STATLIB accepts externally written applications programs with minimal requirements for formatting, documentation, and testing, whereas ELSA has relatively stringent requirements. The developer of a program in SPLUS can easily submit it to STATLIB, but the cost to a user of learning to use a STATLIB program may be high. Indeed, there is no guarantee that the program even works. In contrast, submitting a program to ELSA is relatively costly because of the documentation and testing that are required, but it is relatively easy for a user to implement a program that has been downloaded from ELSA.

One outlet for net-based proliferation of econometric methods is the Common Gateway Interface (CGI). This user interface allows outside Internet users to enter text into a CGI window that may then be interpreted as data or program lines. An archive of GAUSS programs for econometricians with a CGI interface may be found at (http://eclab.econ.pdx.edu/gpe/). The CGI communication technique allows the distant user to send certain commmands and thereby to try methods developed by others. It is not possible, though, to contruct commands to read one's own data from the user's disk or display results in an interactive graphic. Own data may be entered only by hand or by cut/paste. Afterwards the data strings are sent to a serving computer. There exist numerous CGI interfaced net calculators An XLISP-STAT

Journal Of Econometrics 95 (2000), pp. 333-345

(http://www.cern.ch/WebMaker/examples/xlisp/www/cldoc 1.html) based calculator may be found at (http://www.stat.ucla.edu/calculators).

An alternative way to do net-based econometric computing is via Java. The Java technology enables the programmer to produce code that is independent of the user's operating system and applications software. Typically, Java is used to support interactiveness in browsers. A java applet that has been loaded into a browser can perform operations such as interactive graphics, least-squares mean-regression, and nonparametric density estimation. Essentially, a Java applet functions as an application module that operates on all machines without requiring the user to have a specific operating system or specific applications software. Several user interface tools are available for this technology. For example, sliders, point clicks, spreadsheets, and interactive drawing are available. The current technical implementation of Java is based on SUN's Java Development Kit (JDK) 1.1 (http://www.javasoft.com/). For Java applets to run in common browsers it is necessary to have the corresponding Java Runtime Environment (JRE) 1.1 enabled.

An example of a Java interaction based on point clicking, is the support vector machine (SVM) running in the Royal Holloway College in London (http://svm.dcs.rhbnc.ac.uk/pagesnew/1D-Reg.shtml). One uses the mouse to enter points on a set of coordinate axes that are displayed on the screen. The SVM computes a nonparametric mean-regression using orthogonal polynomials. This regression application can be used to illustrate the effects of outliers and other pathologies of data. It cannot, however, be used for real-data applications because there is no capability for loading a user's data into the applet. The reason is that the browser-supported JRE is not able to perform local file input and output (I/O).

Another example for educational interaction is the effect of bin width on the appearance of a histogram. This example has been developed by Phillip Stark at the University of California, Berkeley, and is available with many others at SticiGui© Java Tools (http://www.stat.berkeley.edu/users/stark/Java/index.htm). The user points with the mouse to a slider and moves a bar that is linked to the histogram bin width. A graph displays the histogram that is produced with the chosen bin width. As with the SVM examples, these are useful for illustrating the properties of statistical procedures in an instructional setting but cannot be used for real-data applications, since the user may not use his own data files.

A limited capability for data entry is provided by the procedures that are available at the Webstat Project at the University of South Carolina (http://www.stat.sc.edu/ ~ west/webstat/). The user may enter data by typing them into a spreadsheet or by copying and pasting them from a file or downloading them by FTP from a server. Data entry by reading a local file is not available because applets loaded through a browser do not allow local I/O operations. Basic statistical operations may be performed, including computing means and variances, linear regression, plotting data. Graphs may be displayed

Journal Of Econometrics 95 (2000), pp. 333-345

via a menu-controlled user interface. One drawback of this approach is that the data-entry procedure makes the use of large data sets infeasible. Another is that no source code is available to the user. Therefore, the user cannot modify the procedures to support special needs or extend them to classes of problems that they currently do not accommodate.

The foregoing implementations of statistical methods are most useful for teaching introductory econometrics and statistics students. An example is given by the virtual statistics laboratory of Rice University (http://www.ruf.rice.edu/ ~ lane/stat_sim/index.html). In this example, the student generates a scatterplot and then draws a regression line by eye. The display shows the residual mean-squared error and other goodness-of-fit statistics for the line that has been drawn. The student can then use the mouse to draw a new line and, by repeating this process, find the line that minimizes the mean-squared error. Other Java-supported teaching methods are available at this virtual laboratory for correlation, sampling distributions, and approximations to the normal distribution, among others.

The Java technology used by Webstat and the laboratory at Rice University are not ideally suited to carrying out platform-independent complex computations with large data sets. Technically, pure applets are loaded into the browser like plug-ins. They run independently of the Internet once they are loaded using the browser-provided Java runtime environment (JRE) Certain operations (e.g. I/O file transfer) are not supported by browsers although they are permitted in the Java language. The Java applet runs in a browser via a JRE that uses certain 'classes' provided by the browser. A class is like a keyhole that accepts only certain keys, in this case Java commands. For security reasons free browsers do not come with an I/O class. One may, however, write Java applets that use local file I/O operations. In consequence, these must run independently of a browser. For realistic Internet-based econometric analysis it is, therefore, useful to start applets independently of the Java classes provided by a browser. Simple econometric computations can be done entirely by such an applet. For example, computation of means and variances, t-tests, and smoothing can be entirely written in Java and run on the user's machine with the user's data. One may imagine a whole econometric package written entirely in Java.

The disadvantage of this approach is that handling large data sets and carrying out intensive computing operations slows the user's machine. The price for platform-independent computing is that calculations written in Java are slower than those written in generic languages such as C or FORTRAN. In general, intensive computing is best performed on the provider's server, whereas graphics and editing are best performed on the user's machine. In other words, the Java applet technology is most useful when there is a good balance between the tasks carried out on the provider's server and those carried out on the user's machine. This raises the question of scaling computing loads between a user and a server, the latter usually being the econometric method provider with a fast computer.

Journal Of Econometrics 95 (2000), pp. 333-345

4. Client-server econometric computing

The foregoing discussion shows that a well designed client-server architecture should combine accessibiliy of methods on the Internet with computing loads that are distributed in a way that assigns tasks to client and server machines so as to take advantage of the strengths of each. In this section, we describe an existing computing environment that does this. This environment is based on the software environment XploRe (http://www.XploRe-stat.de) but could also be implemented using any other system that provides the needed Java interface. We consider three uses of the client-server technology:

1. Instruction in econometrics via a Java interface and tutorials and interactive course texts.
2. Methods for enabling outside researchers to supply programs that other outside researchers can access over the Internet and use in applications.
3. The creation of Method and Data Technology centers for a group of suppliers

4.1. Instruction in econometrics

In teaching econometrics, it is important to give students opportunities to experience the numerical properties of the methods they are using, develop intuition about how methods perform, and apply these methods to data. One way of doing this is to carry out instruction in a setting that has the required computers, software, and data – for example, a computerized classroom in which each student sits at a microcomputer that has access to the required application programs. This approach, while effective, can have the disadvantage of requiring much input of faculty time and effort to develop applications software and prepare data. Downloading software and ready-to-use data does not reduce these costs unless the software and data are compatible with the statistical software system being used for instruction.

The Java interface of XploRe offers the possibility of implementing econometric methods immediately and digesting formulas more easily by applying them to data directly. An example is given in a tutorial about generalized linear models (GLM) (http://www.xplore-stat.de/tutorials/glmstart.html). (See Fig. 1).

It is assumed that the student has some background knowledge of GLM. The tutorial instructs the student on how to apply what kind of model to what kind of data. The question of natural link functions is presented. The software may control a possible mismatch of data type and user chosen link by rejecting, e.g. a logistic link for normal response data. The tutorial is written in HTML language. It is possible to inspect parts of the GLM module and, thereby, to introduce the student into the operational phase of applying the GLM technique to data. The Java interface may be opened via a simple mouse click on the

Journal Of Econometrics 95 (2000), pp. 333-345

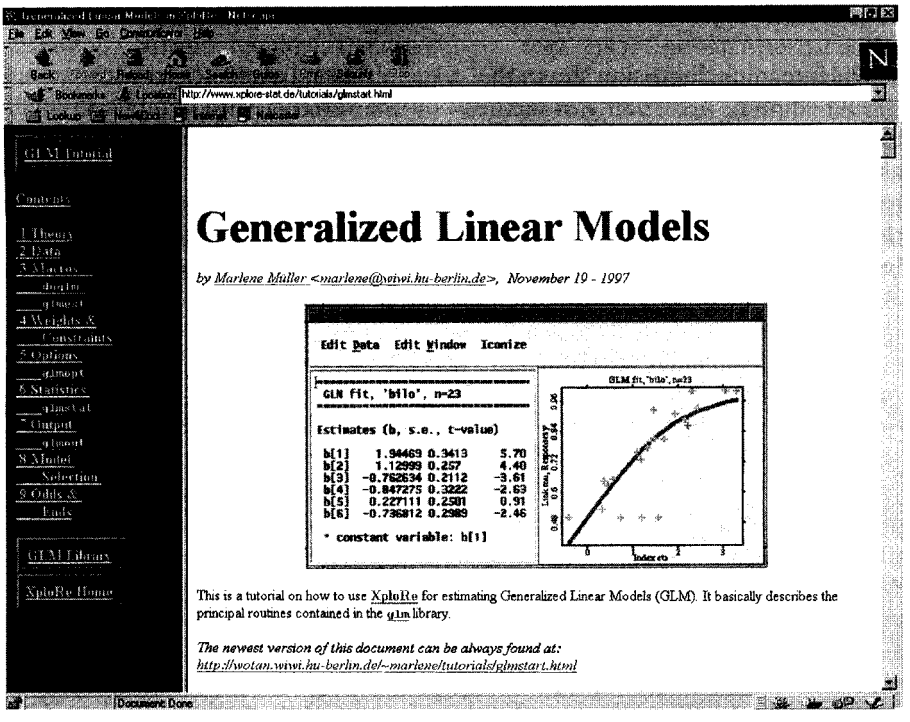


Fig. 1. The GLM tutorial.

user's desktop. This gives the student an independent computing window that allows him to apply GLM techniques to his own data. The computations are performed on a server. This may be the instructor's PC or workstation or a high-speed remote server. The platform-independent client window may be opened on any machine and any platform in the world. This enables the student to do his practical GLM course homework at home or even in an Internet cafe.

The same technique is applicable to complete textbooks. A LATEX2HTML converter (http://cbl.leeds.ac.uk/nikos/tex2html/doc/latex2html/latex2html.html) may create an HTML document from a Latex text with links to the Java interface. Similar linking techniques to Java interfaces apply for PDF documents. Examples exist for books on Applied Multivariate Statistical Analysis (http://www.quantlet.de/ ~ scripts/scripts/sma/ma.html) and a course on non- and semiparametric modelling (http://www.quantlet.de/ ~ scripts/scripts/spm/spm.html). A PDF document of a book on wavelets and their statistical applications is provided by (http://www.quantlet.de/ ~ scripts/scripts/wav/wavpdf.pdf). The course text is typically available as a PDF, Postscript or HTML document. The basic idea is that the student may carry out homework exercises without

Journal Of Econometrics 95 (2000), pp. 333-345

necessary physical access to the software environment used in class. Exercises and examinations may be carried out via email and HTML document transfer. Classes can be taught to a distant audience and no overhead emerges at these places since the Java interface is embedded into the course text. A student may write his own application in XploRe and make it available via the Internet to other students or professors.

4.2. Supply of technology

In writing his own application the user introduces in a sense a new technology that might be useful to add to the system of programs used for a specific course. This proliferation of technology among students and teachers is possible only if there is an easy and standardized way to 'publish' user-written programs and macros. XploRe provides us with this technology. The student may easily produce a web-visible HTML page of the results of his homework. Moreover he may add new algorithms and techniques to an existing faculty system of methods. (See Fig. 2).

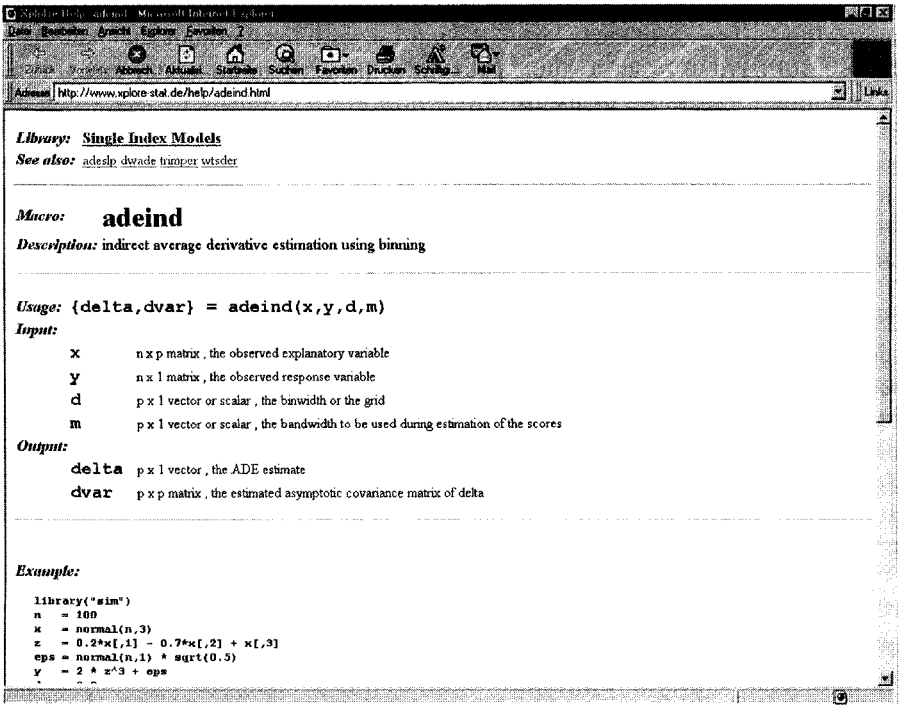


Fig. 2. The help file for ADE made by a method supplier.

Journal Of Econometrics 95 (2000), pp. 333-345

The semiparametric technique for Average Derivative Estimation (ADE), for example, is hard to find in standard econometric packages, although much theoretical research has been carried out on its asymptotic properties. A researcher who has written a macro for ADE may put this technique on the web so that other researchers can try it with their own data. See the help page (http://www.XploRe-stat.de/help/adeind.html).

The Java interface, which may be started independently, serves as a client which makes contact with the server that is provided by the researcher, inventor and supplier of the help page. Many other examples of this kind, for example on ADE with discrete covariates, may be found on the Help System Pages (http://www.xplore-stat.de/help/Xpl Start.html). This technique leads in fact directly to a virtual computing and methodology laboratory.

*4.3. MD*Tech Centers*

The possibility of offering techniques to other researchers creates centers of technology which may be called Method and Data Technologies (MD*Tech) centers. They may provide outside users (professional clients, students, companies, etc.) with methodological techniques and data descriptions. An example would be a place that offers, say, high-frequency finance data and fast forecasting methods. The clients may check the market with the methods that suit them but do not have to order a whole package of possibly unneccesary techniques. An example for methods of image processing is given by (http://lcavwww.epfl.ch/javaproject/index.html). Any user on the net may process several sample images via a Java interface. Buttons control the supplier written filtering and smoothing or false color imaging. A user's own data set may not be introduced via the browser as explained above but a user who learns about a method from the web site may contact the supplier for further assistance.

Fig. 3 displays an example of an outside view of an MD*Tech center (http://www.mdtech.de). The client wants to apply a density estimation technique on his own data set. In the above-mentioned course text on non- and semiparametric modelling (http://www.quantlet.de/ ~ scripts/scripts/spm/spm.html) the method of automatic smoothing with Silverman's rule of thumb is described (right-half of Fig. 3). The student/user/client may now take his own data (upper left-half of Fig. 3) and apply the MD*Tech provided technique of automatic density estimation to his data. The resulting density estimate is provided as a graph in the lower left half of Fig. 3.

Any individual or a research group may form such a MD*Tech center. The basic idea is that research-oriented individuals or groups offer their knowledge and expertise via the Internet together with a client/server based computing service. The service and methodology provider and the outside users of such a center may profit from this arrangement in two ways. The MD*Tech center that develops new methods (faculty, research group, etc.) is able to provide

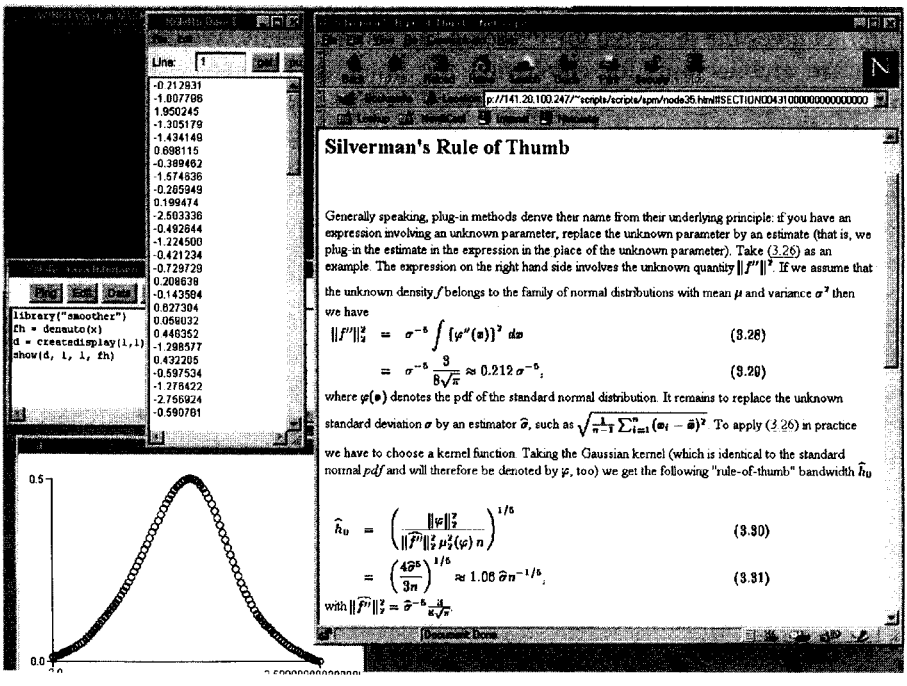


Fig. 3. An instruction text on density estimation with a Java window.

newest technology in shorter cycles than usual software updates. Second the outside user need not buy a whole software package on its own since the methods may be tailored for his needs. In addition, the Java interface reduces costs for both sides. The service provider may offer problem solutions for all platforms and, thereby, reduces programming work. The user may apply the provided methods to his data sets regardless of the user's platform. In summary: this kind of Internet-based econometric computing can be of real value for clients and service providers. In the long run, a MD*Tech center may even provide computing services for other institutions and thereby create a profitable marketing of university hosted technology.

5. Conclusions

Modern econometrics requires transparent use of highly specialized methods that are usually implemented on specific hard- and software platforms. The supply and proliferation of new econometric technology is complicated through this heterogeneity. Typically, software is written and optimized for a specific

Journal Of Econometrics 95 (2000), pp. 333-345

platform and thus not available for applied research at other places. The Internet gives the potential for reducing this heterogeneity. We discuss possible software architectures for employing the potential of Internet-based econometric computing. With existing browser technology program pieces may be used in the world wide web with the exception of local file I/O.

A client/server concept encompasses this problem and offers the possibility of intra- and Internet-based teaching. We refer to existing examples of net-based teaching and present concrete examples for interactive teaching of elementary econometrics and statistics. For web-based courses it is vital to have interaction between the clients data and the server provided methodology. The Java language is the appropriate tool for the architecture of this interaction. It allows platform-independent computing via standard browsers. The net-based teaching may be seen as an export of technology to outside clients or students. This point of view leads to the potential for a group of teachers and researchers to form Method and Data Technology (MD*Tech) centers. In such MD*Tech centers methods for special econometric problems may be collected and via a Java interface applied researchers may use the provided methods with their own data.

Acknowledgements

We would like to thank Knut Bartels, Alan Kirman, Christian Müller, Marlene Müller, and Erich Neuwirth for helpful suggestions and corrections. The paper was financially supported by the Sonderforschungsbereich 373 'Quantifikation und Simulation Ökonomischer Prozesse', Deutsche Forschungsgemeinschaft. The research of Joel L. Horowitz was supported in part by NSF grant SBR-9617925.

References

Applied Multivariate Statistical Analysis book (passwd available from authors) (http://www.quantlet.de/~scripts/scripts/sma/ma.html)
Average Derivative Estimation helpfile (http://www.XploRe-stat.de/help/adeind.html).
CGI calculators (http://www.stat.ucla.edu/calculators).
CodEc software archive (http://netec.mcc.ac.uk/CodEc.html)
ELSA archive (http://elsa.berkeley.edu/)
GLM tutorial (http://www.xplore-stat.de/tutorials/glmstart.html)
GAUSS software (http://www.aptech.com)
GAUSS programming for Econometricians (http://eclab.econ.pdx.edu/gpe/)
Help System Pages (http://www.xplore-stat.de/help/Xpl Start.html).
Image processing with Java (http://lcavwww.epfl.ch/javaproject/index.html)
LATEX2HTML converter (http://cbl.leeds.ac.uk/nikos/tex2html/doc/latex2html/latex2html.html)
MD*Tech-Method and Data Technologies (http://www.mdtech.de)

Journal Of Econometrics 95 (2000), pp. 333-345

Non- and Semiparametric Modelling course text (passwd avble from authors) (http://www.quantlet.de/~ scripts/scripts/spm/spm.html)
SAS software (http://www.sas.com/)
SHAZAM software (http://shazam.econ.ubc.ca)
Splus software (http://www.mathsoft.com/Splus)
Stata software (http://www.stata.com)
STATLIB server of SPLUS (http://lib.stat.cmu.edu/S/)
SticiGui© Java Tools (http://www.stat.berkeley.edu/users/stark/Java/index.htm)
SUN's Java Development Kit (JDK) (http://www.javasoft.com/).
Support Vector Machine http://svm.dcs.rhbnc.ac.uk/pagesnew/1D-Reg.shtml)
Virtual Stat Lab (http://www.ruf.rice.edu/~ lane/stat sim/index.html)
wavelet book in PDF format (passwd avable from authors) (http://www.quantlet.de/~ scripts/scripts/wav/wavpdf.pdf)
Webstat Project (http://www.stat.sc.edu/ ~ west/webstat/)
XLISP-STAT (http://www.cern.ch/WebMaker/examples/xlisp/www/cldoc1.html)
ploRe – the internet interactive statistical computing environment (http://www.XploRe-stat.de)

Finance and Stochastics 4 (2000), pp. 189-207

Discrete Time Option Pricing with Flexible Volatility Estimation *

Wolfgang Härdle
Institut für Statistik und Ökonometrie
Humboldt–Universität zu Berlin
Spandauer Str.1
D-10178 Berlin
Germany
e-mail: haerdle@wiwi.hu-berlin.de

Christian M. Hafner
CORE, Université catholique de Louvain
Louvain–la–Neuve, Belgium
and
Sonderforschungsbereich 373
Humboldt–Universität zu Berlin, Germany
e-mail hafner@wiwi.hu-berlin.de

Abstract

By extending the GARCH option pricing model of Duan (1995) to more flexible volatility estimation it is shown that the prices of out-of-the-money options strongly depend on volatility features such as asymmetry. Results are provided for the properties of the stationary pricing distribution in the case of a threshold GARCH model. For a stock index series with a pronounced leverage effect, simulated threshold GARCH option prices are substantially closer to observed market prices than the Black/Scholes and simulated GARCH prices.

Keywords: option pricing, volatility, GARCH, threshold GARCH, leverage effect
JEL classification: C15, C22, G13
Mathematics Subject Classification (1991): 90A09, 60H10

*This research was partially financed by contributions from the Deutsche Forschungsgemeinschaft, Sonderforschungsbereich 373 "Quantification and Simulation of Economic Processes" and by the Belgian Program on Interuniversity Poles of Attraction initiated by the Belgian State, Prime Minister's Office, Science Policy Programming. The authors would like to acknowledge helpful comments from Frédéric Jouneau, Eckhard Platen, Martin Schweizer and an anonymous referee. Manuscript received: ...; final version received: ...

Finance and Stochastics 4 (2000), pp. 189-207

1 Introduction

It has long been recognized in the option pricing literature that the Black/Scholes prices reveal certain empirical anomalies, *e.g.* the well-known 'smile' effect. In recent years, the most prominent explanation for these anomalies has been stochastic volatility of the underlying asset. Empirically less significant are the effects of trading in discrete time (Bossaerts and Hillion, 1997) and feedback effects of hedging on the stock price process (Platen and Schweizer, 1998).

Since their introduction by Engle (1982), autoregressive conditional heteroskedasticity models (ARCH) have been successfully applied to financial time series. It is thus natural to consider pricing models for options on assets whose prices follow ARCH–type processes. To this end, Duan (1995) established a discrete–time option pricing model for the case of a GARCH volatility process. The aim of our paper is to show that for a given preference structure the results of Duan may be very sensitive to alternative specifications of the volatility process. This concerns the statistical properties of the asset price process under the equivalent martingale measure as well as the simulated prices.

The shape of the news impact curve, defined by Engle and Ng (1993) as today's volatility as a function of yesterday's return, is one of the dominating pricing factors. For instance, it is relevant to find out whether the news impact curve is symmetric or asymmetric, how fast it increases and whether it saturates for large returns. In general, far in– and far out–of–the–money options are underpriced and at–the–money options overpriced by Black/Scholes in the case of stochastic volatility. However, as the simulations of Hull and White (1987) already show, the degree of mispricing strongly depends on the volatility parameters and even more strongly on the correlation between volatility and the stock price.

In order to alleviate mispricing due to volatility misspecifation, flexible volatility models are required. If there is a correlation between stock price and volatility, one could use the EGARCH model of Nelson (1991). This model, however, has the drawback that stationarity conditions and the asymptotics of quasi maximum likelihood estimation (QMLE) are not completely solved. An alternative way is to introduce thresholds for the news impact curve as in the threshold GARCH (TGARCH) model by Zakoian (1994, for the conditional standard deviation) and Glosten, Jagannathan and Runkle (1993, for the conditional variance). If the number of thresholds can be determined from the data this approach has the appealing property that it is the first step towards a nonparametric model without any parametric restriction. In fact, recent papers on nonparametric volatility estimation show that these models are able to reveal volatility features that would be difficult to capture with parametric models. Bossaerts, Härdle and Hafner (1996) obtain asymmetry of nonparametric news impact curves for major foreign exchange rates. Also, they show that the conditional kurtosis may not be constant, which is not consistent with the standard conditional normality assumption.

However, an exhaustive analysis of the complex structure of high frequency financial time series and its impact on option pricing has to be left to future research. Here, we focus solely

Härdle, W. and Hafner, C. M. (2000) Discrete Time Option Pricing with Flexible Volatility Estimation

on the volatility specification, knowing that the effects of *e.g.* skewness and kurtosis may not be negligible. We extend the results of Duan (1995) to the case of a TGARCH process and provide extensive Monte Carlo simulation results for three typical parameter constellations. In particular, we compare the simulated GARCH option prices with corresponding TGARCH and Black/Scholes prices. In an empirical analysis, we show that the observed call option market prices indeed reflect the asymmetry found for the news impact curve of a DAX series.

Section 2 gives a review of recent developments of volatility models in discrete time, Section 3 extends the GARCH option pricing model to TGARCH, Section 4 provides a simulation study for GARCH and TGARCH option prices and in Section 5 price predictions are obtained for calls on the DAX and compared with market prices.

2 A succinct review of flexible ARCH models

It is well–known that returns of financial time series exhibit nonconstant volatility patterns. A general time series model for financial returns would be

$$y_t = \mu_t + \varepsilon_t, \tag{1}$$

with $\varepsilon_t = \sigma_t \xi_t, \quad \xi_t \sim i.i.d.(0,1)$, and μ_t and σ_t being respectively the mean and standard deviation conditional on the past. σ_t can be either stochastic itself or determined by the past history of the time series. If μ_t is interpreted as the risk premium, it can be linked to σ_t, as in the ARCH–in–mean (ARCH–M) model of Engle, Lilien, and Robins (1987). For $\mu_t = r + f(\sigma_t^2)$, r would typically be the riskfree interest rate and f the logarithm or square root.

The ARCH(q) model (Engle, 1982) assumes a linear dependence of the conditional variance on squared past residuals,

$$\sigma_t^2 = \omega + \sum_{i=1}^{q} \alpha_i \varepsilon_{t-i}^2. \tag{2}$$

Bollerslev (1986) generalized the ARCH(q) model to an analogue of ARMA processes for σ_t^2. The GARCH(p, q) model takes the form

$$\sigma_t^2 = \omega + \sum_{i=1}^{q} \alpha_i \varepsilon_{t-i}^2 + \sum_{j=1}^{p} \beta_j \sigma_{t-j}^2. \tag{3}$$

The original idea of these ARCH type models of volatility was that the value of today's σ_t is computable from recent innovation scale factors and/or past residuals of the process. The standard ARCH models have a defect though in the sense that they do not model possible asymmetric volatility shocks. "Good news" do not necessarily have the same impact on volatility as "bad news". Engle and Ng (1993) provide a survey of many parametric models proposed to overcome the symmetry problem. Important representatives in this context are the EGARCH model and the threshold ARCH models.

Nelson (1991) introduced the exponential GARCH (EGARCH) model,

$$\log \sigma_t^2 = \omega_t + \sum_{k=1}^{\infty} \beta_k g(\xi_{t-k}) \tag{4}$$

with deterministic coefficients ω_t, β_k, and $g(\xi_t) = \gamma(|\xi_t| - E[|\xi_t|]) + \theta\xi_t$. The EGARCH model has several important advantages over the classic ARCH formulation of conditional heteroskedasticity. It models volatility more naturally in a multiplicative way, and the piecewise linear function g may model the observable asymmetry of σ_t^2 as a function of past innovations. A disadvantage though is that for some common fat–tailed distributions of ξ_t the unconditional variance is not finite. Also, it implies an exponential increase of the news impact curve, which has not been found favorable in many empirical investigations.

The idea of threshold ARCH (TARCH) models is to keep the functional form of the standard GARCH model, but to let the coefficients α depend on past innovations. Glosten, Jagannathan and Runkle (1993) consider the simple case where α depends only on the sign of the past innovation, *i.e.*

$$\sigma_t^2 = \omega + \alpha_1 \varepsilon_{t-1}^2 I(\varepsilon_{t-1} < 0) + \alpha_2 \varepsilon_{t-1}^2 I(\varepsilon_{t-1} \geq 0) + \beta \sigma_{t-1}^2. \tag{5}$$

Here, $I(A)$ denotes an indicator function that takes the value one if the logical expression A is true, and zero otherwise. Obviously, this model coincides with the standard GARCH(1,1) model if $\alpha_1 = \alpha_2$. The case $\alpha_1 > \alpha_2$ describes the leverage effect, which is usually observed for stock returns.

In the same spirit, Zakoian (1994) modeled the conditional standard deviation. For one threshold at zero, a threshold ARCH model of order q can be written as

$$\sigma_t = \omega + \sum_{i=1}^{q} \alpha_i^+ \varepsilon_{t-i}^+ + \sum_{i=1}^{q} \alpha_i^- \varepsilon_{t-i}^-, \tag{6}$$

with $\varepsilon_t^+ = \max(\varepsilon_t, 0)$ and $\varepsilon_t^- = \min(\varepsilon_t, 0)$. Rabemananjara and Zakoian (1993) applied this model in a generalized form to the French stock market. Recently, this model was generalized by El Babsiri and Zakoian (1996) by specifying $\varepsilon_t = \sigma_{t,+}\xi_t^+ + \sigma_{t,-}\xi_t^-$, where $\sigma_{t,+}$ and $\sigma_{t,-}$ are TGARCH processes. Thus, depending on the sign of the innovations one possibly obtains different volatility processes.

A first step towards a flexible nonparametric modelling of volatility was made by the paper of Gouriéroux and Monfort (1992). Their Qualitative Threshold ARCH (QTARCH) model had σ_t^2 as a step function of the past returns y_t. For instance, a QTARCH model of order one takes the form

$$\sigma_t = \sum_{j=1}^{J} s_j I(y_{t-1} \in A_j), \tag{7}$$

where $\{A_j\}_{j=1}^J$ is a partition of the real line, s_j are the step heights and J is the number of steps.

A direct advantage of model (7) is that the functional form is no longer bound to a specific one, since step functions are dense in the L_2 function space. A disadvantage though is that the choice of J is not flexible. Gouriéroux and Monfort (1992) assumed a known and fixed number of steps J.

A more flexible model is described in Härdle and Tsybakov (1997) where the volatility is modelled as an unknown function of the past return,

$$\sigma_t^2 = g(y_{t-1}). \tag{8}$$

An extension to the multivariate case $\sigma_t^2 = g(y_{t-1}, \ldots, y_{t-q})$ is given by Härdle, Tsybakov and Yang (1998). In that paper a multivariate time series volatility matrix is modelled as an unknown function of the past values of the process. From the smoothing literature it is well known that the flexibility of free functional form estimation has to be paid with reduced statistical precision, especially in higher dimension. In the case considered here the consequence for the practical use of smoothing techniques for time series must be a limit on the number of lags or an introduction of a lower dimensional structure.

The newer literature pursues the second way by considering additive models or multiplicative structures of volatility, see Härdle, Lütkepohl and Chen (1997), Yang, Härdle and Nielsen (1999) and Hafner (1998). Also, a nonparametric analogue of the heterogenous ARCH (HARCH) model of Müller *et al.* (1997) can be established as

$$\sigma_t^2 = \omega + \sum_{j=1}^{q} g_j \left(\sum_{i=1}^{j} y_{t-i} \right), \tag{9}$$

where g_j are nonparametric additive factor functions. This model is economically appealing, since it regards volatility as the accumulation of different market components. These components are described by the trader's frequency of acting in the market, each component having a different impact on volatility.

3 Option pricing with alternative ARCH models

We consider a discrete–time economy where interest rates and returns are paid after each time interval of fixed equispaced length. This contrasts the usual formulation in terms of continuously compounded interest rates and returns, but we keep the notation consistent with the notation traditionally used in the ARCH literature.

Let $S_t, t = 0, 1, 2, \ldots$ be the price of a stock at time t and $y_t = (S_t - S_{t-1})/S_{t-1}$ be its one–period return excluding dividend payments. Suppose that there is a price for risk, measured in terms of a risk premium that is added to the risk–free interest rate r to build the expected next–period return. It is sensible to allow dependence of risk premia on the conditional variance. As Duan (1995), we adopt the ARCH–M model of Engle, Lilien, and Robins (1987) with the risk

premium being a linear function of the conditional standard deviation,

$$y_t = r + \lambda\sigma_t + \varepsilon_t \tag{10}$$

$$\varepsilon_t \mid \mathcal{F}_{t-1} \sim N(0, \sigma_t^2) \tag{11}$$

$$\sigma_t^2 = \omega + \alpha\varepsilon_{t-1}^2 + \beta\sigma_{t-1}^2. \tag{12}$$

In (12), ω, α, and β are constant parameters satisfying stationarity and positivity conditions. The constant parameter λ may be interpreted as the unit price for risk. $\mathcal{F}_t$ denotes the set of all information prior to and including time t. For notational convenience we restrict our discussion to the GARCH(1,1) case.

The above model is estimated under the empirical measure P. If one wants to apply the risk–neutral pricing methodology (see Cox and Ross, 1976), the measure has to be transformed such that the resulting discounted price process is a martingale. This guarantees that there are no arbitrage opportunities (Harrison and Kreps, 1979). Due to the incompleteness of markets, however, there is a multitude of such transformations (Harrison and Pliska, 1981). Unlike the complete market situation, incompleteness leaves the trader unable to construct a self–financing portfolio that exactly replicates the option's payoff. Thus, hedging involves a risk, and option prices generally depend on risk preferences. It should be emphasized that the non-availability of a perfect hedging strategy is of eminent importance for the hedging practice.

To apply no arbitrage pricing by choosing a particular pricing measure, one has to impose assumptions about the pricing of volatility. Many papers investigated option prices under stochastic volatility for the case that volatility has zero systematic risk (*i.e.* the volatility risk premium is zero, see e.g. Hull and White, 1987, and Renault and Touzi, 1996). Melino and Turnbull (1990) allowed for nonzero, constant and exogenous volatility risk premia. As the empirical results of Wiggins (1987) show, the non–pricing of changes in volatility may not be justified.

Duan identified an equivalent martingale measure Q by requiring that the conditional return distribution remains normal, and

$$\mathrm{Var}^P(y_t \mid \mathcal{F}_{t-1}) = \mathrm{Var}^Q(y_t \mid \mathcal{F}_{t-1}) \tag{13}$$

almost surely with respect to P. This is what he terms the 'locally risk-neutral valuation relationship' (LRNVR). He shows that a representative agent with, for example, constant relative risk aversion and normally distributed relative changes in aggregate consumption maximizes his expected utility using the LRNVR. The LRNVR incorporates a constant volatility risk premium that is directly linked to the risk premium in the mean. The alternative concept of minimizing the quadratic loss of a hedge portfolio, as pioneered by Föllmer and Sondermann (1986) and Föllmer and Schweizer (1991), will in general lead to a different choice of the pricing measure.

To obtain a martingale process under the new measure, one has to introduce a new error term, η_t, that incorporates the time–varying risk premium effect. Hence, by defining $\eta_t =$

$\varepsilon_t + \lambda\sigma_t$, the LRNVR leads to the following model under the pricing measure Q:

$$
\begin{aligned}
y_t &= r + \eta_t && (14)\\
\eta_t \mid \mathcal{F}_{t-1} &\sim N(0, \sigma_t^2) && (15)\\
\sigma_t^2 &= \omega + \alpha(\eta_{t-1} - \lambda\sigma_{t-1})^2 + \beta\sigma_{t-1}^2. && (16)
\end{aligned}
$$

For the GARCH(1,1) model, the variance of the stationary distribution under the empirical measure P is $\mathrm{Var}^P(\varepsilon_t) = \omega/(1-\alpha-\beta)$, see Bollerslev (1986). For the LRNVR–measure the variance of the stationary distribution increases to $\mathrm{Var}^Q(\eta_t) = \omega/(1-\alpha(1+\lambda^2)-\beta)$ due to the fact that the volatility process under Q is driven by noncentral rather than central chi–square distributed innovations. We will see below that the change of the unconditional variance crucially depends on the specification of the news impact curve.

As noted above, the restriction of having a quadratic and symmetric news impact function may not always be reasonable, as many empirical studies of stock returns showed. For the above model, this assumption can be relaxed to some nonlinear news impact function $g(\cdot)$. The following model is a nonparametric (or semiparametric) analogue to the GARCH model. Under the empirical measure P we have

$$
\begin{aligned}
y_t &= r + \lambda\sigma_t + \varepsilon_t\\
\varepsilon_t \mid \mathcal{F}_{t-1} &\sim_P N(0, \sigma_t^2)\\
\sigma_t^2 &= g(\varepsilon_{t-1}) + \beta\sigma_{t-1}^2.
\end{aligned}
$$

For this general framework with no prior information on $g(\cdot)$, estimation is a delicate issue, because iterative estimators are required. However, if β is sufficiently small one can truncate at some lag and estimate a conventional semiparametric additive model.

Under the LRNVR equivalent martingale measure Q, the model becomes

$$
\begin{aligned}
y_t &= r + \eta_t\\
\eta_t \mid \mathcal{F}_{t-1} &\sim_Q N(0, \sigma_t^2)\\
\sigma_t^2 &= g(\eta_{t-1} - \lambda\sigma_{t-1}) + \beta\sigma_{t-1}^2.
\end{aligned}
$$

Note that once an estimate of $g(\cdot)$ is obtained under P, it can readily be used for the pricing under Q.

However, we decided not to use this general semiparametric model because a thorough analysis of the properties of the estimators is still in progress. Instead, we consider a flexible parametric model that will be investigated below in a simulation study, *i.e.* the threshold GARCH model of Glosten, Jagannathan and Runkle (1993), where the news impact function can be written as $g(x) = \omega + \alpha_1 x^2 I(x < 0) + \alpha_2 x^2 I(x \geq 0)$. To give some motivation for this model, we estimated a very simple nonparametric model, $y_t = \sigma(y_{t-1})\xi_t$, with ξ_t *i.i.d.*$(0,1)$, for the returns on the German stock index DAX, which will be further analyzed in Section 5. The estimate of the news impact curve $\sigma^2(\cdot)$ is shown in Figure 2. To have an idea about

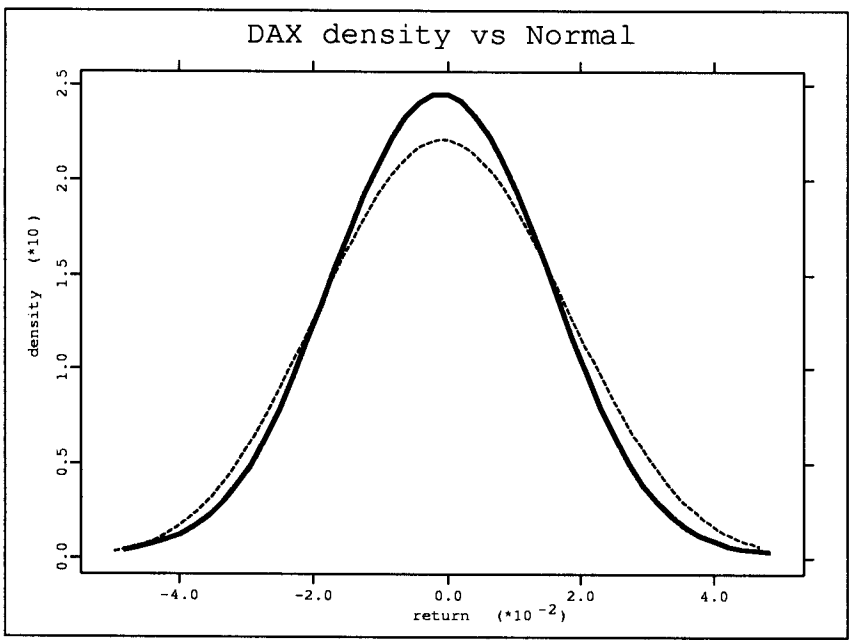


Figure 1: *Kernel estimate of the DAX return distribution (solid line) versus a Kernel estimate of a normal distribution (dashed line) with the same mean and variance. We used a bandwidth of 0.03 and the quartic kernel* $K(u) = 15/16(1-u^2)^2 I(|u| < 1)$. *The boundary regions are skipped in the figure.*

the distribution of the returns, a nonparametric density estimate *vis à vis* a smoothed normal density is provided in Figure 1.

It is obvious that $g(\cdot)$ is not symmetric around zero. The TGARCH model captures this effect by having $\alpha_1 > \alpha_2$. We are aware of the fact that other parametric models may as well describe this feature, but the TGARCH model has proven to be a sufficiently flexible and tractable model for stock returns (see, *e.g.*, Rabemananjara and Zakoian, 1993), whereas the EGARCH model, as noted above, suffers from several theoretical and practical drawbacks.

Recall that the innovation distribution is normal. Thus, it follows for the TGARCH model similarly as in Bollerslev (1986) that the unconditional variance under P is $\mathrm{Var}^P(\varepsilon_t) = \omega/(1-\bar{\alpha}-\beta)$, with $\bar{\alpha} = (\alpha_1+\alpha_2)/2$. The following proposition provides the unconditional variance for $\eta_t = \varepsilon_t + \lambda\sigma_t$ under Q.

Proposition 1 *The unconditional variance of the TGARCH(1,1) model under the LRNVR equivalent martingale measure* Q *is*

$$Var^Q(\eta_t) = \frac{\omega}{1-\psi(\lambda)(\alpha_1-\alpha_2)-\alpha_2(1+\lambda^2)-\beta} \tag{17}$$

with

$$\psi(u) = \frac{u}{\sqrt{2\pi}}\exp(-\frac{1}{2}u^2) + (1+u^2)\Phi(u)$$

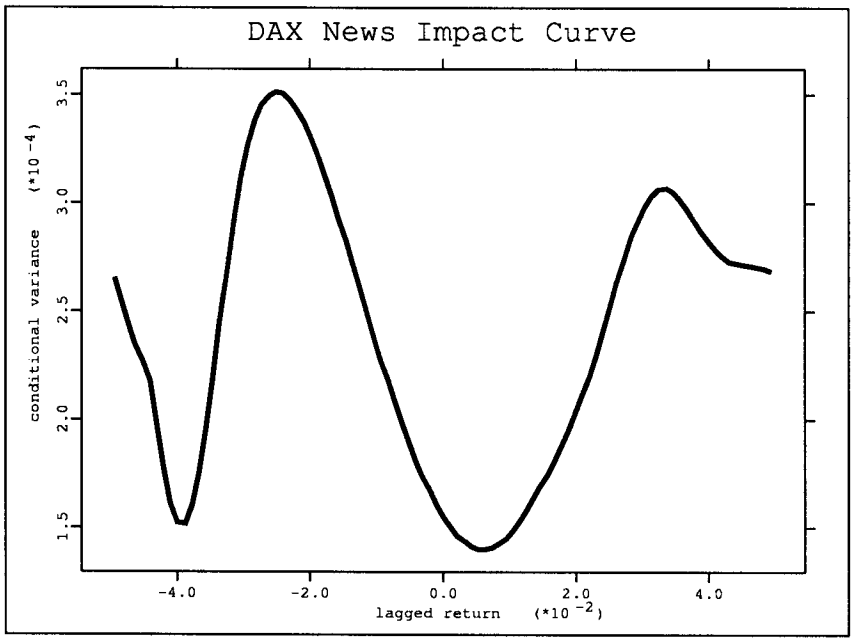


Figure 2: *Local linear estimate of the news impact curve for the DAX. The model is* $y_t = \sigma(y_{t-1})\xi_t$. *Shown is the estimate of the function* $\sigma^2(y_{t-1})$ *for a bandwidth choice of 0.03. The boundary regions are skipped in the figure.*

and $\Phi(u)$ *denoting the cumulative standard normal distribution function.*

Proof: see Appendix.

The function ψ is positive and $\psi(\lambda) > 1/2$ for the realistic case $\lambda > 0$. We can make the following statements about the change of the unconditional variance: For $\alpha_1 = \alpha_2$, (17) coincides with the GARCH(1,1) result. For $\alpha_1 > \alpha_2$ (the leverage effect case), the unconditional variance increases even stronger than in the symmetric GARCH case. For $\alpha_1 < \alpha_2$, the unconditional variance will be smaller than for the leverage effect case, and we can distinguish two cases: If the inequality

$$\alpha_1 < \alpha_2 \frac{2\psi(\lambda) - 1 - 2\lambda^2}{2\psi(\lambda) - 1} \tag{18}$$

holds, then the unconditional variance under Q will be even smaller than the unconditional variance under P. If (18) does not hold, then we have as above $\mathrm{Var}^P(\varepsilon_t) \leq \mathrm{Var}^Q(\eta_t)$. However, the quotient on the right hand side of (18) takes negative values for realistic values of the unit risk premium (*i.e.* small positive values), such that for most empirical studies (18) will not hold.

Of course the stationary variance affects the option price: the larger (smaller) the variance, the higher (lower) the option price. This is especially relevant for long maturity options, where the long run mean of volatility is one of the important determinants of the option price. Thus, options may be 'underpriced' when employing the GARCH model if in fact there is a leverage

effect.

A second pecularity of the LRNVR approach is that under Q and for positive risk premia, today's innovation is negatively correlated with tomorrow's GARCH conditional variance, contrary to the zero correlation under P. More precisely, we have $\mathrm{Cov}^Q(\eta_t/\sigma_t, \sigma_{t+1}^2) = -2\lambda\alpha\mathrm{Var}^Q(\eta_t)$ with GARCH parameter α. This suggests that short run predictions of volatility under Q (which affects the option price) depend not only on squared past innovations, but also on their signs. In particular, for $\lambda > 0$ a negative (positive) past innovation tends to increase (decrease) volatility and thus the option price. The following proposition states that the covariance depends on the asymmetry of the news impact function when we use a TGARCH instead of a GARCH model.

Proposition 2 *For the TGARCH(1,1) model, the covariance under the LRNVR equivalent martingale measure Q of the innovation at time t and the conditional variance at time $t+1$ can be expressed as*

$$Cov^Q(\frac{\eta_t}{\sigma_t}, \sigma_{t+1}^2) = -2\,Var^Q(\eta_t)\left\{\lambda\alpha_2 + \left[\frac{1}{\sqrt{2\pi}}\exp(-\frac{1}{2}\lambda^2) + \lambda\Phi(\lambda)\right](\alpha_1 - \alpha_2)\right\}, \qquad (19)$$

where $Var^Q(\eta_t)$ is given in Proposition 1.

Proof: see Appendix.

Assume in the following that we have a positive unit risk premium λ. Again, we can distinguish three cases: For $\alpha_1 = \alpha_2$ (the symmetry case), we obtain $\mathrm{Cov}^Q(\eta_t/\sigma_t, \sigma_{t+1}^2) = -2\lambda\alpha_2\mathrm{Var}^Q(\eta_t)$, *i.e.* the GARCH(1,1) result. For $\alpha_1 < \alpha_2$ (the reverted leverage case) the covariance increases and if

$$\lambda\alpha_2 + \left[\frac{1}{\sqrt{2\pi}}\exp(-\frac{1}{2}\lambda^2) + \lambda\Phi(\lambda)\right](\alpha_1 - \alpha_2) < 0, \qquad (20)$$

then the correlation becomes positive. Finally, for $\alpha_1 > \alpha_2$ (the leverage case) the covariance is negative and increases in absolute value.

This shows that also the reverting behavior of volatility to the stationary variance under Q may be quite different from the symmetric GARCH case. The negative covariance is even larger for the leverage effect case. This indicates that options may be 'overpriced' ('underpriced') if the past innovation is positive (negative), the asset price follows a TGARCH process with $\alpha_1 > \alpha_2$, and the GARCH model ($\alpha_1 = \alpha_2$) is used for volatility predictions.

4 A Monte Carlo study

Because the discounted price process is a martingale under the equivalent martingale measure Q, we can apply the risk–neutral valuation methodology of Cox and Ross (1976). The Q–price of a call option at time t, C_t, is given by

$$C_t = (1+r)^{-n}\mathrm{E}^Q[\max(S_T - K, 0) \mid \mathcal{F}_t] \qquad (21)$$

with T denoting the maturity date, $n \equiv T-t$ the time to maturity and K the exercise price. For European options, the no–arbitrage price of a put option, P_t, is determined by put–call parity, *i.e.*, $P_t = C_t - S_t + (1+r)^{-n}K$. Because there is no analytic expression for the expectation in (21), we have to use numerical techniques to simulate the option price. That is, the distribution of the payoff function $\max(S_T - K, 0)$ at the terminal date is simulated by generating m stock price processes

$$S_{T,i} = S_t \prod_{s=t+1}^{T} (1 + y_{s,i}), \quad i = 1, \ldots, m, \tag{22}$$

where $y_{s,i}$ is the return of the i^{th} replication at time s, and then discount the mean of the payoff–function with the risk–free rate, *i.e.*

$$C_t = (1+r)^{-n} \frac{1}{m} \sum_{i=1}^{m} \max(S_{T,i} - K, 0). \tag{23}$$

Throughout the simulation study we used the following parameters: $r = 0$, $S_0 = 100$, $n = 30$ days, $m = 400,000$, $\lambda = 0.01$. The moneyness S_0/K was varied from 0.85 to 1.15, which is the typical range of traded options. We do not compare the effects of different times to maturity n, because it is known from previous work that many of the features caused by stochastic volatility such as smiles disappear when the time to maturity is increased. In general, these effects qualitatively stay the same, but quantitatively become more and more insignificant. This was confirmed by our experiments, so we focused on only one short maturity.

To reduce the variance of the payoffs, the antithetic variable technique of Hammersley and Handscomb (1964) was used. This turned out to be sufficient, since the standard errors of the obtained option prices were small due to our large number of replications m.

In order to study the effects of an asymmetric news impact function on option prices, we consider three situations, characterized by the degree of short–run autocorrelation of squared returns and by the degree of persistence. For a GARCH(1,1) process it can be shown that the first order autocorrelation of squared returns, ρ_1, is given by

$$\rho_1 = \alpha(1 - \alpha\beta - \beta^2)/(1 - 2\alpha\beta - \beta^2), \tag{24}$$

and $\rho_j = (\alpha + \beta)\rho_{j-1}$, $j \geq 2$. Table 1 reports the parameter constellations and characteristics of the three types.

Type 1 is described by high persistence and small first order autocorrelation, Type 2 by high persistence and large first order autocorrelation, and Type 3 by low persistence and small first order autocorrelation. Type 1 is typical for financial time series (daily or intra–daily), because it is usually observed that the autocorrelation function of squared returns drops quickly for the first lags but then declines very slowly. Type 2 describes a situation where there are very strong ARCH effects, and Type 3 resembles the case of highly aggregated data, *e.g.* monthly or quarterly series. In all cases the parameter ω is chosen such that $\sigma^2 = 0.0002$, *i.e.* the unconditional variance remains the same.

Type	α	β	$\alpha+\beta$	ρ_1
1	0.1	0.85	0.95	0.1791
2	0.5	0.45	0.95	0.8237
3	0.1	0.5	0.6	0.1077

Table 1: *Characterization of types for the GARCH(1,1) model*

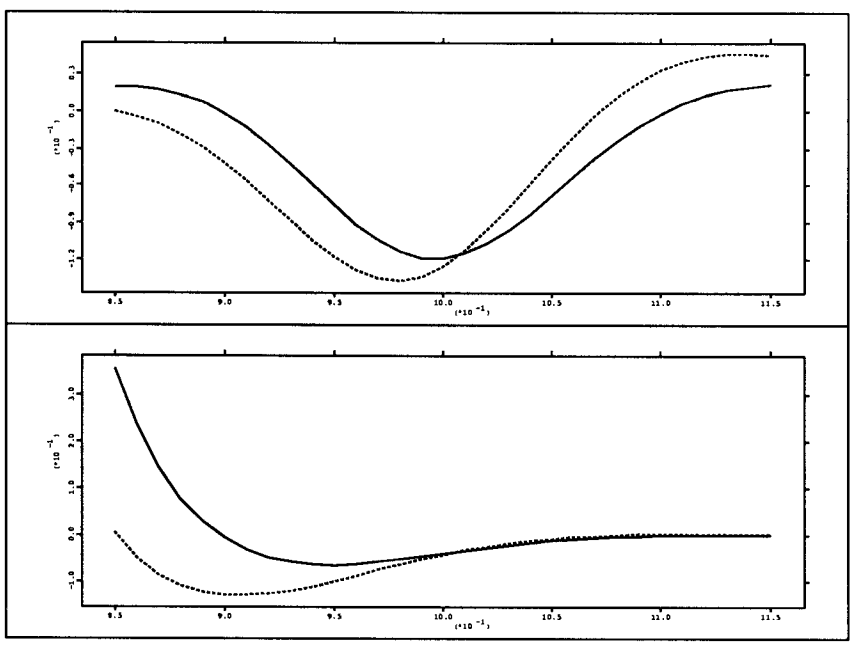


Figure 3: *Difference of simulated GARCH (solid) and TGARCH (dashed) option prices to BS prices as a function of moneyness for Type 1 and the leverage effect case. The upper plot shows the absolute difference, the lower plot the absolute difference divided by the BS price.*

Concerning the nonlinear news impact function $g(\cdot)$, we have chosen the Threshold ARCH Model of Glosten, Jagannathan and Runkle (1993) and Zakoian (1994) with two asymmetry cases: The first case, which we may call 'leverage-effect' case, is

$$g_1(x) = \omega + 1.2\alpha x^2 I(x < 0) + 0.8\alpha x^2 I(x \geq 0)$$

and the second, 'reverted leverage-effect' case

$$g_2(x) = \omega + 0.8\alpha x^2 I(x < 0) + 1.2\alpha x^2 I(x \geq 0).$$

For Type 1 and the leverage effect case the simulation results are depicted in Figure 3. We show a plot of the absolute and relative difference of GARCH and TGARCH prices to the corresponding Black and Scholes price. The relative difference is defined as the absolute

difference divided by the Black and Scholes price. Due to the small grid (we used steps of 0.01 for the moneyness), the functions appear very smooth. For the GARCH case we obtain the well–known result that the price difference to Black/Scholes has a U–shape with respect to the moneyness. As a consequence of the monotonously increasing call price in the moneyness, the relative difference is largest in absolute value for out-of-the-money options, whereas the relative difference becomes more and more negligible the higher the moneyness. This may also explain the often observed skewness of the smile effect. For the TGARCH option prices we basically observe a similar deviation to Black/Scholes but with one major difference: For the leverage effect, out-of-the-money options are priced lower and in-the-money-options higher than under a GARCH model. This is intuitively plausible: If an option is far out-of-the-money and time to maturity is short, the only possibility to be of positive value at the expiration date is that the underlying stock appreciates several times in a row with large returns. This, however, is less probable for the leverage case, because positive returns have in this case a smaller impact on volatility than in the symmetric case, provided that the above parameter constellation holds.

To economize on space, we present the results for Types 2 and 3 and for the reverted leverage effect case in Table 2 for selected values of the moneyness. For the leverage effect case, the described deviation of TGARCH prices from GARCH prices is also visible for Types 2 and 3. For the reverted leverage effect case the arguments are reversed. Now it is more probable that an out-of-the-money option will end up in the money, and therefore the TGARCH prices of far out-of-the-money options are higher than the GARCH prices. As one might expect, the deviations of the simulated prices to Black/Scholes and between the GARCH and TGARCH prices are highest for Type 2, *i.e.* for very strong short–run ARCH effects, and smallest for the low persistence Type 3. The latter case is expected, because the differences should disappear the more the homoskedastic case is approached.

5 Application to the pricing of DAX Calls

The GARCH pricing methodology was applied to German stock index and option data. As stock index we used the daily closing notation of the DAX, January 1st, 1988 to March 31, 1992. The closing notation of this index is usually fixed at about 13:30 local time (Frankfurt). For call options on this index we used the transaction price records of the German futures exchange (DTB) for January to March 1992. In order to synchronize stock and option observation times, we linearly interpolated between the last option price before 13:30 and the first one after, unless there was more than two hours difference.

There was no evidence for autocorrelation in the DAX returns, but squared and absolute returns were highly autocorrelated. We estimated the GARCH(1,1)–M model

$$y_t = \lambda\sigma_t + \varepsilon_t \tag{25}$$

$$\varepsilon_t \mid \mathcal{F}_{t-1} \sim N(0, \sigma_t^2) \tag{26}$$

		GARCH		TGARCH			
				leverage effect		reverted lev.eff.	
Type	Moneyness	% diff	SE	% diff	SE	% diff	SE
Type 1	0.85	35.947	1.697	0.746	1.359	75.769	2.069
	0.90	-0.550	0.563	-12.779	0.498	11.606	0.631
	0.95	-6.302	0.261	-9.786	0.245	-3.153	0.278
	1.00	-3.850	0.132	-4.061	0.125	-3.806	0.139
	1.05	-1.138	0.057	-0.651	0.052	-1.692	0.061
	1.10	-0.020	0.025	0.347	0.022	-0.400	0.028
	1.15	0.162	0.012	0.347	0.010	-0.013	0.014
Type 2	0.85	199.068	5.847	104.619	4.433	293.704	7.884
	0.90	0.489	1.136	-23.964	0.891	22.140	1.469
	0.95	-30.759	0.370	-39.316	0.305	-24.518	0.454
	1.00	-20.975	0.167	-22.362	0.141	-20.804	0.198
	1.05	-6.038	0.077	-5.427	0.063	-7.148	0.095
	1.10	-0.302	0.042	0.202	0.033	-0.966	0.054
	1.15	0.695	0.027	0.991	0.021	0.351	0.037
Type 3	0.85	-2.899	1.209	-11.898	1.125	6.687	1.297
	0.90	-5.439	0.496	-8.886	0.479	-1.982	0.513
	0.95	-4.027	0.249	-4.970	0.245	-3.114	0.254
	1.00	-2.042	0.128	-2.077	0.126	-2.025	0.130
	1.05	-0.710	0.055	-0.559	0.053	-0.867	0.056
	1.10	-0.157	0.023	-0.047	0.022	-0.267	0.023
	1.15	-0.009	0.010	0.042	0.010	-0.059	0.011

Table 2: *Simulation results for selected values of the moneyness. The percentage differences of GARCH and TGARCH option prices to the Black and Scholes prices are given and the corresponding standard errors (SE) of the simulation.*

	ω	α	α_1	α_2	β	λ	$-2\log L$
GARCH	1.66E-05	0.1438			0.7756	0.0691	-7,697.66
	(1.04E-06)	(0.0061)			(0.012)	(0.0178)	
TGARCH	1.91E-05		0.2005	0.0454	0.7736	0.0385	-7,719.24
	(1.359E-06)		(0.0084)	(0.0113)	(0.0157)	(0.0175)	

Table 3: *GARCH and TGARCH estimation results for DAX returns, 88/01/01–91/12/30 (QMLE–standard errors in parentheses)*

$$\sigma_t^2 = \omega + \alpha\varepsilon_{t-1}^2 + \beta\sigma_{t-1}^2 \tag{27}$$

by QMLE for the DAX series until 1991/12/30. A constant in (25) was not significant and excluded from the estimation. The estimation results are reported in Table 3. All parameters are significantly different from zero. The degree of persistence, $\alpha + \beta = 0.9194$, is significantly smaller than one. Thus, the unconditional variance is finite. The risk premium parameter λ is positive as expected from economic theory.

The QMLE results for the TGARCH model

$$\sigma_t^2 = \omega + \alpha_1\varepsilon_{t-1}^2 I(\varepsilon_{t-1} < 0) + \alpha_2\varepsilon_{t-1}^2 I(\varepsilon_{t-1} \geq 0) + \beta\sigma_{t-1}^2 \tag{28}$$

are also given in Table 3. Considering the log–likelihood value, the TGARCH fit improves the GARCH fit. A likelihood–ratio test rejects the GARCH model at all conventional levels. α_1 and α_2 are significantly different, and thus the asymmetry of the news impact function is significant. Because $\alpha_1 > \alpha_2$, we have the usual leverage effect for stock market series.

Having estimated the model for the period 1988 to 1991, the next step is to simulate option prices for the evaluation period January to March 1992 and to compare them with the observed market prices. In this paper we concentrate on call options. Since the DAX options traded at the DTB are of the European type, results for put options can be obtained by put–call parity. We selected nine call options with maturities January 17, March 20 and June 19, 1992. To resemble approximately the in-the-money, at-the-money and out-of-the-money cases, we have chosen the exercise prices 1550, 1600, and 1650 for the January option and 1600, 1650 and 1700 for the March and June options. We simulated prices for the January options from January 3 to January 16 (ten days), for the March options from January 3 to March 19 (57 days) and for the June options from January 3 to March 31 (64 days). The June option with exercise price 1700 was introduced on January 16, so that for the first ten trading days of January we do not have observations for this particular option. Also, the non-availibility of market prices due to thin trading reduced slightly the number of observations, given as k in Table 4.

One question is how to specify the starting value for the volatility process. We experimented with two different strategies: First, to set the starting value equal to the current estimate of

volatility (GARCH or TGARCH) by extrapolating the volatility process, keeping the parameters fixed. Second, to use the volatility as implied by observed market prices using the formula of Black and Scholes (1973). In the following, however, we refrain from pursuing the second approach and report only the results for the strategy using GARCH or TGARCH estimates. The argument, as emphasized by a referee, is that one should rely on a consistent, self-contained valuation method without referring to parameters used in different models. One may imagine that Black/Scholes implied volatilities will perform poorly as soon as the valuation methods under stochastic volatility will be standard for practitioners.

For calculation of the Black/Scholes prices at time t, the implied volatility at time $t-1$ was used. A similar procedure was used in Bossaerts and Hillion (1993), where 15 minute old implied volatilities were plugged into the Black/Scholes formula which then performed well.

In order to have a goodness-of-fit criterion, we define relative residuals as

$$u_t \equiv \frac{C_t - C_{Market,t}}{C_{Market,t}}$$

where C_t is either the Black/Scholes, the GARCH or the TGARCH price. Residuals should be looked at in relative terms, because a trader will always prefer a cheap option which is 'underpriced' by the same amount as an expensive option, simply by multiplying his position in the cheap option. A similar argument applies for the case of selling 'overpriced' options. Due to the symmetry we can consider a quadratic loss-criterion, *i.e.*

$$Q = \sum_{t=1}^{k} u_t^2.$$

The results for the three models are given in Table 4.

Overall, both the GARCH and TGARCH option pricing models perform substantially better than the Black and Scholes model. For in–the–money and at–the–money options, the improvement of the TGARCH prediction over GARCH is small. For out–of–the–money options, however, there is a large reduction in the loss criterion. Recall from the simulation study that options reacting most sensitive to stochastic volatility and leverage effects are out–of–the–money options. In our real data situation, this is most distinct for the January 1650 option, where Black/Scholes performs very poorly, and TGARCH is performing much better than GARCH. For the March and June options, the improvements are slightly less distinct. This is explained by the fact that the index increased to a level of 1736 on March 20 and 1717 on March 31, turning the options with exercise price 1700 from out–of–the–money to in–the–money. This is also the explanation for the fact that Q is highest for the January 1650 options. There are only ten trading days, but this option is for some of these days far out-of-the-money. For example, the DAX index closed at 1578 on January 8.

Because in all cases TGARCH outperformed GARCH, we conclude that the market is aware of the asymmetry of the volatility. Thus, the correct specification of the volatility model strongly matters for the prediction of option prices.

T	K	k	BS	GARCH	TGARCH
Jan	1550	10	0.017	0.014	0.014
	1600	10	0.099	0.029	0.028
	1650	10	4.231	1.626	1.314
Mar	1600	47	1.112	0.961	0.954
	1650	53	1.347	1.283	1.173
	1700	56	1.827	1.701	1.649
Jun	1600	53	1.385	1.381	1.373
	1650	56	2.023	1.678	1.562
	1700	51	2.460	2.053	1.913
total		346	14.500	10.725	9.980

Table 4: *The loss criterion Q for DAX calls with maturity T and exercise price K using BS, GARCH and TGARCH option prices. The number of observations is given by k.*

6 Conclusions

In this paper, we show that out-of-the-money options strongly depend on the volatility specification. In particular, if there is a leverage effect, out-of-the-money options may be severely overpriced by assuming a symmetric news impact function, as in the GARCH model. For this to show, a simulation study was performed which used as volatility generating processes constant (Black and Scholes), GARCH and Threshold GARCH processes. The TGARCH option prices of about more than five percent out–of–the–money options significantly deviated from the GARCH prices. In a real data example it was shown for calls on the German stock index DAX that the simulated TGARCH prices were closer to market prices than both Black/Scholes and GARCH prices. In fact, under time–varying volatility and short maturity Black/Scholes seems to perform quite poorly, whereas GARCH and TGARCH both do reasonably well. The difference between GARCH and TGARCH becomes obvious when looking at the prices for options with high exercise price. Concluding, it can be stated that for the examined period January to March 1992 traders at the German futures exchange were aware of both the underlying stochastic volatility and the underlying leverage effect. The observed market prices reflect both of these features.

Future research will have to investigate the performance of standard hedge portfolios under different choices of the martingale measure as well as under misspecification of the time series model for the underlying stock.

Appendix

Proof of Proposition 1. Let $z_t = \eta_t/\sigma_t - \lambda$. Under Q, $z_t \mid \mathcal{F}_{t-1} \sim N(-\lambda, 1)$. The conditional variance σ_t^2 can be written as

$$\sigma_t^2 = \omega + \alpha_1 \sigma_{t-1}^2 z_{t-1}^2 I(z_{t-1} < 0) + \alpha_2 \sigma_{t-1}^2 z_{t-1}^2 I(z_{t-1} \geq 0) + \beta \sigma_{t-1}^2.$$

Taking expectations, the integral expression for the negative support can be verified to be

$$\begin{aligned} \mathrm{E}^Q[z_t^2 I(z_t < 0) \mid \mathcal{F}_{t-1}] &= \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{0} z^2 e^{-\frac{1}{2}(z+\lambda)^2} dz \\ &= \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\lambda} (u-\lambda)^2 e^{-\frac{1}{2}u^2} du \\ &= \frac{\lambda}{\sqrt{2\pi}} e^{-\frac{1}{2}\lambda^2} + (1+\lambda^2)\Phi(\lambda) \qquad (29) \\ &=: \psi(\lambda). \qquad (30) \end{aligned}$$

Since

$$\mathrm{E}^Q[z_t^2 \mid \mathcal{F}_{t-1}] = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} z^2 e^{-\frac{1}{2}(z+\lambda)^2} dz = 1 + \lambda^2,$$

we have for the positive support

$$\mathrm{E}^Q[z_t^2 I(z_t \geq 0) \mid \mathcal{F}_{t-1}] = 1 + \lambda^2 - \psi(\lambda). \qquad (31)$$

Thus, we arrive at

$$\mathrm{E}^Q[\sigma_t^2] = \omega + \alpha_1 \psi(\lambda) \mathrm{E}^Q[\sigma_{t-1}^2] + \alpha_2 [1 + \lambda^2 - \psi(\lambda)] \mathrm{E}^Q[\sigma_{t-1}^2] + \beta \mathrm{E}^Q[\sigma_{t-1}^2]. \qquad (32)$$

Noting that the unconditional variance is independent of t, the result is obtained. Q.E.D.

Proof of Proposition 2.

At first, the conditional covariance is determined:

$$\begin{aligned} \mathrm{Cov}_{t-1}^Q(\frac{\eta_t}{\sigma_t}, \sigma_{t+1}^2) &= \mathrm{E}_{t-1}^Q \left[\frac{\eta_t}{\sigma_t} \sigma_{t+1}^2\right] = \omega \mathrm{E}_{t-1}^Q \left[\frac{\eta_t}{\sigma_t}\right] \\ &+ \alpha_1 \mathrm{E}_{t-1}^Q \left[\frac{\eta_t}{\sigma_t} (\eta_t - \lambda\sigma_t)^2 I(\eta_t - \lambda\sigma_t < 0)\right] \\ &+ \alpha_2 \mathrm{E}_{t-1}^Q \left[\frac{\eta_t}{\sigma_t} (\eta_t - \lambda\sigma_t)^2 I(\eta_t - \lambda\sigma_t \geq 0)\right] \\ &+ \beta\sigma_t \mathrm{E}_{t-1}^Q [\eta_t] \qquad (33) \end{aligned}$$

where $\mathrm{E}_t(\cdot)$ and $\mathrm{Cov}_t(\cdot)$ abbreviates $\mathrm{E}(\cdot \mid \mathcal{F}_t)$ and $\mathrm{Cov}(\cdot \mid \mathcal{F}_t)$, respectively. Because of (15), the first and fourth conditional expectation on the right hand side of (33) are zero. The second conditional expectation can be shown to be

$$\mathrm{E}_{t-1}^Q \left[\frac{\eta_t}{\sigma_t} (\eta_t - \lambda\sigma_t)^2 I(\eta_t - \lambda\sigma_t < 0)\right] = -2\sigma_t^2 \left[\frac{1}{\sqrt{2\pi}} \exp(-\frac{1}{2}\lambda^2) + \lambda\Phi(\lambda)\right]. \qquad (34)$$

Because of $\mathrm{E}^Q_{t-1}\left[\frac{\eta_t}{\sigma_t}(\eta_t-\lambda\sigma_t)^2\right]=-2\lambda\sigma_t^2$, we can write for the third conditional expectation in (33)

$$\mathrm{E}^Q_{t-1}\left[\frac{\eta_t}{\sigma_t}(\eta_t-\lambda\sigma_t)^2 I(\eta_t-\lambda\sigma_t\geq 0)\right]=-2\sigma_t^2\left[\lambda-\frac{1}{\sqrt{2\pi}}\exp(-\frac{1}{2}\lambda^2)-\lambda\Phi(\lambda)\right]. \tag{35}$$

Plugging (34) and (35) into (33), we obtain

$$\mathrm{Cov}^Q_{t-1}(\frac{\eta_t}{\sigma_t},\sigma^2_{t+1})=-2\sigma_t^2\left\{\lambda\alpha_2+\left[\frac{1}{\sqrt{2\pi}}\exp(-\frac{1}{2}\lambda^2)+\lambda\Phi(\lambda)\right](\alpha_1-\alpha_2)\right\}. \tag{36}$$

Noting that $\mathrm{Cov}^Q(\eta_t/\sigma_t,\sigma^2_{t+1})=\mathrm{E}^Q[\mathrm{Cov}^Q_{t-1}(\eta_t/\sigma_t,\sigma^2_{t+1})]$, the result is obtained. Q.E.D.

References

Black, F., Scholes, M. (1973), The Pricing of Options and Corporate Liabilities, *Journal of Political Economy* 81: 637–659.

Bollerslev, T. (1986), Generalized Autoregressive Conditional Heteroskedasticity, *Journal of Econometrics* 31: 307–327.

Bossaerts, P., Härdle, W., Hafner, C. M. (1996), Foreign Exchange Rates Have Surprising Volatility, in: P.M.Robinson (ed.), Athens Conference on Applied Probability and Time Series, vol.2, Lecture Notes in Statistics 115, 55–72, Springer Verlag.

Bossaerts, P., Hillion, P. (1993), A Test of a General Equilibrium Stock Option Pricing Model, *Mathematical Finance* 3:311–347.

Bossaerts, P., Hillion, P. (1997), Local Parametric Analysis of Hedging in Discrete Time, *Journal of Econometrics* 81:243–272.

Cox, J.C., Ross, S.A. (1976), The Valuation of Options for Alternative Stochastic Processes, *Journal of Financial Economics* 3: 145–166.

Duan, J.–C. (1995), The GARCH option pricing model, *Mathematical Finance* 5: 13–32.

El Babsiri, M., Zakoian, J.-M. (1996), Contemporaneous Asymmetry in Weak GARCH Processes, CORE DP 9604, Louvain-la-Neuve, Belgium.

Engle, R. (1982), Autoregressive Conditional Heteroskedasticity with Estimates of the Variance of U.K. Inflation, *Econometrica* 50: 987–1008.

Engle, R., Lilien, D., Robins, R. (1987), Estimating Time Varying Risk Premia in the Term Structure: The ARCH–M Model, *Econometrica* 55: 391–407.

Engle, R., Ng, V. (1993), Measuring and Testing the Impact of News on Volatility, *Journal of Finance* 48: 1749–1778.

Föllmer, H., Schweizer, M. (1991), Hedging of Contingent Claims under Incomplete Information, in: *Applied Stochastic Analysis*, ed. by M.H.A.Davis and R.J.Elliot, Gordon and Breach, London, 389–414.

Föllmer, H., Sondermann, D. (1986), Hedging of Non-redundant Contingent Claims, in: Hildenbrand, W., Mas-Colell, A. (eds.), *Contributions to Mathematical Economics*, Amsterdam, North Holland, 205–223.

Glosten, L.R., Jagannathan, R., Runkle, D.E. (1993), On the Relation Between the Expected Value and the Volatility of the Nominal Excess Return on Stocks, *Journal of Finance* 48: 1779–1801.

Gouriéroux, C., Monfort, A. (1992), Qualitative Threshold ARCH Models, *Journal of Econometrics* 52: 159–199.

Härdle, W., Lütkepohl, H., Chen, R. (1997), A Review of Nonparametric Time Series Analysis, *International Statistical Review* 65: 49–72.

Härdle, W., Tsybakov, A. (1997), Local Polynomial Estimation of the Volatility Function, *Journal of Econometrics* 81:223–242.

Härdle, W., Tsybakov, A., Yang, L. (1998), Nonparametric Vector Autoregression, *Journal of Statistical Planning and Inference* 68: 221–245.

Hafner, C. M. (1998), Estimating High Frequency Foreign Exchange Rate Volatility with Nonparametric ARCH Models, *Journal of Statistical Planning and Inference* 68: 247–269.

Hammersley, J.M., Handscomb, D.C. (1964), *Monte Carlo Methods*, Methuen, London.

Harrison, M., Kreps, D. (1979), Martingales and Arbitrage in Multiperiod Securities Markets, *Journal of Economic Theory* 20: 381–408.

Harrison, M., Pliska, S. (1981), Martingales and Stochastic Integrals in the Theory of Continuous Trading, *Stochastic Processes Applications* 11: 215–260.

Hull, J., White, A. (1987), The Pricing of Options on Assets with Stochastic Volatilities, *Journal of Finance* 42: 281–302.

Melino, A., Turnbull, S.M. (1990), Pricing Foreign Currency Options with Stochastic Volatility, *Journal of Econometrics* 45: 239–265.

Müller, U.A., Dacorogna, M.M., Davé, R., Olsen, R.B., Pictet, O.V., von Weizsäcker, J.E. (1997), Volatilities of Different Time Resolutions – Analyzing the Dynamics of Market Components, *Journal of Empirical Finance* 4: 213–240.

Nelson, D. (1991), Conditional Heteroskedasticity in Asset Returns: A New Approach, *Econometrica* 59: 347–370.

Platen, E., Schweizer, M. (1998), On Feedback Effects From Hedging Derivatives, *Mathematical Finance* 8: 67-84.

Rabemananjara, R., Zakoian, J.M. (1993), Threshold ARCH Models and Asymmetries in Volatility, *Journal of Applied Econometrics* 8: 31–49.

Renault, E., Touzi, N. (1996), Option Hedging and Implied Volatilities in a Stochastic Volatility Model, *Mathematical Finance* 6: 277–302.

Wiggins, J. (1987), Option Values under Stochastic Volatility: Theory and Empirical Estimates, *Journal of Financial Economics* 19: 351–372.

Yang, L., Härdle, W., Nielsen, J.P. (1999), Nonparametric Autoregression with Multiplicative Volatility and Additive Mean, *Journal of Time Series Analysis*, in press.

Zakoian, J.M. (1994), Threshold Heteroskedastic Functions, *Journal of Economic Dynamics and Control* 18: 931–955.

Volume 9, No. 2 (2000), pp. 160–175 Allerton Press, Inc.

MATHEMATICAL METHODS OF STATISTICS

SECOND ORDER MINIMAX ESTIMATION IN PARTIAL LINEAR MODELS

G. GOLUBEV[1] AND W. HÄRDLE[2]*

[1]CMI, Université de Provence
39 rue F. Joliot-Curie, 13453 Marseille Cedex 13, France
golubev@gyptis.univ-mrs.fr

[2]Humboldt Universität zu Berlin
Spandauer Straße 1, 10178 Berlin, BRD
haerdle@wiwi.hu-berlin.de

The problem of estimation of the finite-dimensional parameter in a partial linear model is considered. We derive upper and lower bounds for the second order minimax risk and show that the second order minimax estimator is a penalized maximum likelihood estimator.

Key words: partial linear model, nonparametric estimation, second order minimax risk.

AMS 1991 Subject Classification: Primary 62G05, 62G20; secondary 62C20.

1. Introduction

In a partial linear model we estimate an unknown column-vector $\theta \in \mathbb{R}^d$ based on the observations

$$Y_i = \theta^T Z_i + m(X_i) + \xi_i, \qquad i = 1, \dots, n, \tag{1}$$

where $(\cdot)^T$ denotes transposition and ξ_i are i.i.d. random variables with zero mean and a finite variance $\sigma^2 = \mathbf{E}\xi_i^2$. The regressors $X_i \in [0,1]$ are i.i.d. random variables with a known and strictly positive density $q(x)$ on the interval $[0,1]$. It is assumed that they do not depend on ξ_i. The function $m(x)$, $x \in [0,1]$, here is an unknown nuisance function such that the random variables $m(X_i)$ have zero mean. We

*The work was supported in part by the Deutsche Forschungsgemeinschaft, Sonderforschungsbereich 373 "Quantifikation und Simulation Ökonomischer Prozesse", Humboldt-Universität zu Berlin.

Math. Methods of Statistics, 9, 2, 160 - 175

assume in this paper that the vectors $Z_i \in \mathbb{R}^d$ are non-random and such that the matrix ZZ^T, where $Z = (Z_1, \dots, Z_n)$, is non-singular.

In the parametric part of this partial linear model the "noise" $m(X_i) + \xi_i$ has zero mean. One could therefore be tempted to use, for instance, the "naive" mean-square estimator

$$\widehat{\theta}_n = \arg\min_{\theta \in \mathbb{R}^d} \sum_{i=1}^{n} (Y_i - \theta^T Z_i)^2.$$

It is easy to see that its risk is given by

$$\mathbf{E}\,(\widehat{\theta}_n - \theta)(\widehat{\theta}_n - \theta)^T = (ZZ^T)^{-1}\{\sigma^2 + \mathbf{E}\,m^2(X)\}.$$

The risk is blown up by $\mathbf{E}\,m^2(X)$. The reason is that the estimator $\widehat{\theta}_n$ does not use the prior information about smoothness of the function $m(x)$. If it is sufficiently smooth, then the performance of $\widehat{\theta}_n$ can be substantially improved. Using the method by Robinson [11] it can be shown that there exists an estimator $\widehat{\theta}_{\mathrm{ef}}$ such that

$$\mathbf{E}\,(\widehat{\theta}_{\mathrm{ef}} - \theta)(\widehat{\theta}_{\mathrm{ef}} - \theta)^T = \{1 + o(1)\}(ZZ^T)^{-1}\sigma^2, \quad n \to \infty. \tag{2}$$

If the noise ξ_i is Gaussian such estimators are often called asymptotically efficient or adaptive, see Bickel *et al.* [2]. Asymptotically efficient estimates are traditionally constructed in partial linear models in two ways: by using kernel estimators as in Speckman [16] or by penalization of the log-likelihood. The penalized mean-square spline estimator is defined by

$$\widehat{\theta} = \arg\min_{\theta \in \mathbb{R}^d} \min_{m} \left\{ \frac{1}{n} \sum_{i=1}^{n} \left[Y_i - \theta^T Z_i - m(X_i) \right]^2 + \mu_n \int_0^1 [m^{(\beta)}(t)]^2 \, dt \right\},$$

where $m^{(\beta)}(\cdot)$ denotes the derivative of order β and the minimum is taken over all functions m such that $\int_0^1 q(x)m(x)\,dx = 0$. If the smoothing parameter μ_n tends to zero as $n \to \infty$ then $\widehat{\theta}$ is the asymptotically efficient estimator when the noise is Gaussian.

The goal of the present paper is to provide a second order minimax theory for partial linear models. The available results by Rice [12], Heckman [8], Speckman [16], Carroll and Härdle [1], Mammen and van de Geer [9], Bhattacharia and Zhao [4], Chen [5] specify only the order of the second order term. Typically the estimators proposed in these papers have the second order term of order $n^{-2\beta/(2\beta+1)}$, where β is the "smoothness" of the nuisance function. Our objective is to compute this term up to a constant. So we can discriminate between different estimators and thereby propose the best one. The importance of the second order theory becomes more transparent when we deal with a data driven choice of smoothing parameters in partial linear models. The motivation of such choice is essentially based on the second order arguments.

To shed some light on the second order minimax risk in partial linear models we first present here simple heuristic arguments. Suppose that there is an "oracle" which provides us with the additional data

$$Y_i' = m(X_i) + \xi_i, \quad i = 1, \dots, n. \tag{3}$$

(2000) Golubev, Y. and Härdle, W.
On the second order minimax estimation in partial linear models.

Math. Methods of Statistics, 9, 2, 160 - 175

Based on these data one can estimate the unknown function $m(\cdot)$. Local polynomials, splines or orthogonal series estimators are usually used to construct an estimator $\widehat{m}(X_i)$. All these estimators are linear and therefore can be represented as

$$\widehat{m}(X_k) = \sum_{i=1}^{n} K(X_k, X_i) Y_i',$$

where $K(\cdot,\cdot)$ is a kernel such that $\mathbf{E}\, K(X_i, y) = 0$ for all $y \in [0,1]$. Subtracting $\widehat{m}(X_i)$ from the observations Y_i one arrives at new artificial data

$$Y_i'' = \theta^T Z_i + m(X_i) - \widehat{m}(X_i) + \xi_i, \quad i = 1, \dots, n. \tag{4}$$

The random variables $m(X_i) - \widehat{m}(X_i) + \xi_i$ have almost zero mean, so that the unknown parameter θ can be estimated based on Y_i'' by the least-squares method $\widehat{\theta} = (ZZ^T)^{-1} Z^T Y''$. A simple algebra easily reveals that

$$\begin{aligned}\mathbf{E}\,(\widehat{\theta} - \theta)(\widehat{\theta} - \theta)^T &\approx \mathbf{E}\,(ZZ^T)^{-1}\left\{\sigma^2 E + (ZZ^T)^{-1} \sum_{i=1}^{n} [m(X_i) - \widehat{m}(X_i)]^2 Z_i Z_i^T\right\} \\ &\approx (ZZ^T)^{-1}\left\{\sigma^2 + \mathbf{E}\int_0^1 [m(x) - \widehat{m}(x)]^2 q(x)\, dx\right\}.\end{aligned}$$

Here and later in the text E stands for the $d \times d$ identity matrix. Thus we see that the second order term in the risk expansion is defined by the mean-square error of recovering $m(x)$ in the model (3). This fact plays a very important role and its proof will be given later (see Theorem 2 below).

To simplify some technical details we assume that the function $m(x)$ belongs to the set

$$\mathbf{W}_0 = \left\{m : \int_0^1 [m^{(\beta)}(x)]^2 dx \le L, \ \int_0^1 q(x) m(x)\, dx = 0\right\}.$$

It is assumed that β is positive integer.

The results presented in the paper can be extended in different directions. The errors may be heteroscedastic, i.e., the variance $\mathbf{var}\,\xi_i$ may be a function of (X_i, Z_i), in particular, of $\theta^T Z_i + m(X_i)$. This case is important in generalized partially linear models, where the variance is a function of the mean. Generalized linear models have been investigated by Severini and Staniswalis [14]. Various generalizations and applications of partially linear models can be found in the recent book by Härdle, Liang, and Gao [7]. But we intentionally choose the simplest partial linear model to demonstrate how the second order theory works in semiparametric estimation. We will comment on some possible extensions of our theory later in the text.

The outline of the paper is as follows. We first derive a lower bound for the minimax risk. Here we follow the method proposed by Pinsker [10] and developed for distribution function estimation in Golubev and Levit [6]. We then study in Section 3 penalized least-squares estimators and show that under a proper choice of penalization these estimators are the second order minimax estimators.

(2000) Golubev, Y. and Härdle, W.
On the second order minimax estimation in partial linear models.

Math. Methods of Statistics, 9, 2, 160 - 175

2. The Functional Class $\mathbf{W}_0$

Our consideration will be based on the so-called orthogonal series approach. The cornerstone idea of this approach is to parametrize the functional class $\mathbf{W}_0$. We do this by constructing an orthonormal system in the Hilbert space $\mathbf{L}_q^2[0,1]$ equipped with the norm $\|\cdot\|_q$ and with the inner product $\langle\cdot,\cdot\rangle_q$,

$$\|f\|_q^2 = \int_0^1 q(x)f^2(x)\,dx, \quad \langle f,g\rangle_q = \int_0^1 q(x)f(x)g(x)\,dx.$$

We write $f \perp g$ when $\langle f,g\rangle_q = 0$. Recall that the Kolmogorov diameter d_m of the set $\mathbf{W}_0$ is defined by

$$d_s^2 = \inf_{\varphi_k} \sup_{m\in\mathbf{W}_0} \left\| m - \sum_{k=1}^{s} \langle m,\varphi_k\rangle_q \varphi_k \right\|_q^2,$$

where inf is taken over all orthonormal systems in $\mathbf{L}_q^2[0,1]$. Define the orthonormal system $\{\psi_k\}_1^\infty$ by

$$\sup_{m\in\mathbf{W}_0} \left\| m - \sum_{k=1}^{s} \langle m,\psi_k\rangle \psi_k \right\|_q^2 = d_s^2.$$

In other words, the linear space spanned by the functions $\psi_1,\dots,\psi_s$ provides the best approximation of $\mathbf{W}_0$ in $\mathbf{L}_q^2[0,1]$. Since $\mathbf{W}_0$ is an ellipsoid in $\mathbf{L}_q^2[0,1]$, it is easy to see that the $\{\psi_k\}_1^\infty$ coincide with the main axes of $\mathbf{W}_0$ (for more details see Tikhomirov [17]). Let $\psi_0(x) = 1$. Then ψ_k, $k = 1,\dots,\beta-1$, are the orthonormal polynomials in $\mathbf{L}_q^2[0,1]$, whereas the remaining functions are obtained by

$$\psi_{\beta+l} = \frac{\arg\max_{\varphi\in\mathbf{W}_0,\ \varphi\perp(\psi_0,\dots,\psi_{\beta+l-1})} \|\varphi\|_q}{\max_{\varphi\in\mathbf{W}_0,\ \varphi\perp(\psi_0,\dots,\psi_{\beta+l-1})} \|\varphi\|_q}.$$

It is also not very difficult to show (see Tikhomirov [17]) that $d_s^2 = \|\psi_{s+1}^{(\beta)}\|_q^{-2}$. The Lagrange multipliers method together with integration by parts reveal that $\psi_s(t)$ are the solutions of the following boundary value problem:

$$(-1)^\beta \frac{d^{2\beta}}{dx^{2\beta}}\psi_s(x) = \lambda_s q(x)\psi_s(x),$$
$$\frac{d^k}{dx^k}\psi_s(x)\Big|_{x=0} = \frac{d^k}{dx^k}\psi_s(x)\Big|_{x=1} = 0, \quad k = \beta,\dots,2\beta-1.$$

In particular, for $\beta = 1$ and $q(x) = 1$ we get the well-known cosine-basis $\psi_k(t) = \sqrt{2}\cos(\pi k t)$ with the corresponding eigenvalues $\lambda_k = (\pi k)^2$. The asymptotic behavior of λ_k plays a very important role in approximation theory. It is known that as $s \to \infty$

$$\lambda_s = [1+o(1)](\pi s)^{2\beta}\left[\int_0^1 q^{1/(2\beta)}(x)\,dx\right]^{2\beta}. \tag{5}$$

(2000) Golubev, Y. and Härdle, W.
On the second order minimax estimation in partial linear models.

Math. Methods of Statistics, 9, 2, 160 - 175

For more detail we refer to Utreras [18] and Speckman [15]. Once the basis ψ_k is obtained we can represent any function $m(t) \in \mathbf{W}_0$ as the Fourier series

$$m(t) = \sum_{k=1}^{\infty} \nu_k \psi_k(t), \qquad \text{with} \quad \sum_{k=1}^{\infty} \nu_k^2 \lambda_k \le L, \tag{6}$$

where $\nu_k = \langle m, \psi_k \rangle$.

3. A Lower Bound

The next theorem provides a lower bound for the second order term of the minimax risk in the partial linear model (1). It is assumed only that the random variables ξ_i have an absolutely continuous density $p_\xi(x)$, $x \in \mathbb{R}^1$, with finite Fisher information

$$I_\xi = \int_{-\infty}^{\infty} \frac{p_\xi'^2(x)}{p_\xi(x)}\, dx < \infty.$$

Theorem 1. *As* $n \to \infty$,

$$\inf_{\tilde\theta} \sup_{m \in \mathbf{W}_0} \sup_{\theta \in \mathbb{R}^d} \mathbf{E}\,(\tilde\theta - \theta)(\tilde\theta - \theta)^T \ge (ZZ^T)^{-1} I_\xi^{-1} \left[1 + \frac{1 + o(1)}{n} \sum_{s=1}^{\infty} h_s\right], \tag{7}$$

where inf *is taken over all estimators,*

$$h_s = [1 - \omega\sqrt{\lambda_s}]_+, \qquad [x]_+ = \max(0, x), \tag{8}$$

and ω *is a root of the equation*

$$\frac{1}{nI_\xi} \sum_{s=1}^{\infty} \lambda_s \left[\frac{1}{\omega\sqrt{\lambda_s}} - 1\right]_+ = L. \tag{9}$$

Thus we see that the second order term in the lower bound is controlled by the quantity $\Delta_n = I_\xi^{-1} \sum_{s=1}^{\infty} h_s / n$. The statistical interpretation of this value is well-known. The theorem due to Pinsker [10] states that Δ_n is the asymptotically minimax risk in the following smoothing problem. Suppose that we wish to estimate the infinite-dimensional vector $(\nu_1, \nu_2, \dots)^T$ based on the data

$$s_i = \nu_i + n^{-1/2} \varepsilon_i, \quad i = 1, 2, \dots,$$

where ε_i are i.i.d. $\mathcal{N}(0, I_\xi^{-1})$ and the parameters of interest ν_i obey the condition (6). Then as $n \to \infty$

$$\inf_{\hat\nu} \sup_{\nu} \sum_{k=1}^{\infty} \mathbf{E}\,(\hat\nu_k - \nu_k)^2 = [1 + o(1)]\Delta_n,$$

where inf is taken over all possible estimators, whereas sup is taken over ν_k such that $\sum_{k=1}^{\infty} \lambda_k \nu_k^2 \le L$. The value of Δ_n can be computed as follows. From (5) one concludes that

$$\Delta_n = [1 + o(1)] n^{-1} C(\beta) (L n I_\xi)^{1/(2\beta+1)} \left(\int_0^1 q^{1/(2\beta)}(x)\, dx\right)^{-2\beta/(2\beta+1)},$$

Math. Methods of Statistics, 9, 2, 160 - 175

where $C(\beta) = \pi^{-2\beta/(2\beta+1)}(2\beta+1)^{1/(2\beta+1)}[\beta/(\beta+1)]^{2\beta/(2\beta+1)}$ is the Pinsker constant.

4. Penalized Least-Squares Estimators

In order to show that the lower bound (7) is precise we study now penalized least-squares estimators. We will suppose that ξ_i are such that $\mathbf{E}\,|\xi_i|^{2+\delta} < \infty$ for some $\delta > 0$. Recall the main idea of the penalized likelihood. Assume for a moment that ξ_i are i.i.d. Gaussian and the Fourier coefficients ν_k are also i.i.d. Gaussian $\mathcal{N}(0, \Sigma^2)$, where Σ is the diagonal matrix having arbitrary entries $\Sigma_{kk} = \sigma_k$. Let $\Psi_{ki} = \psi_k(X_i)$. Then it is well known that the estimator

$$\widehat{\theta} = \arg\min_{\theta\in\mathbb{R}^d} \min_{\nu} \{\sigma^{-2}\|Y - Z^T\theta - \Psi^T\nu\|^2 + \|\Sigma^{-1}\nu\|^2\} \tag{10}$$

is Bayesian. Although the above assumptions about ν_k are not fulfilled in the minimax setting, nevertheless we use $\widehat{\theta}$ in this situation. The following theorem shows how to compute the risk of $\widehat{\theta}$ for matrices Σ satisfying the condition (13) below. For brevity we use the following notation:

$$h_k = \frac{n\sigma_k^2}{\sigma^2 + n\sigma_k^2}. \tag{11}$$

Theorem 2. *Assume that*

$$\lim_{n\to\infty} \frac{\max_{j,k} Z_{kj}^2}{\sum_{i=1}^n Z_{ki}^2} \log^{1/2} n = 0, \qquad \max_n \|(ZZ^T)^{-1}\| \max_k \sum_{i=1}^n Z_{ki}^2 < \infty, \tag{12}$$

$$\lim_{n\to\infty} \frac{\log^{1/2} n}{n} \left(\sum_{k=1}^{\infty} h_k\right)^2 = 0. \tag{13}$$

Then for the estimator $\widehat{\theta}$ defined by (10) *we have uniformly in $m \in \mathbf{W}_0$ as $n \to \infty$*

$$\mathbf{E}\,(\widehat{\theta} - \theta)(\widehat{\theta} - \theta)^T = (ZZ^T)^{-1}\left\{\sigma^2 + [1 + o(1)]\sum_{k=1}^{\infty}\left[\nu_k^2(1 - h_k)^2 + \frac{\sigma^2}{n}h_k^2\right]\right\}, \tag{14}$$

where $\nu_k = \langle m, \psi_k\rangle_q$.

In the next theorem we find the optimal in the minimax sense penalization matrix Σ and show that the lower bound from Theorem 1 cannot be improved when ξ_k are Gaussian. Thus we construct the second order minimax estimator.

Theorem 3. *Let $\widehat{\theta}$ be the estimator from* (10) *with $\Sigma = H^{1/2}(E - H)^{-1/2}\sigma/\sqrt{n}$, where H is diagonal, $H_{ss} = h_s$, and h_s are defined by* (8) *and* (9). *If* (12) *is fulfilled then as $n \to \infty$*

$$\sup_{m\in\mathbf{W}_0} \sup_{\theta\in\mathbb{R}^d} \mathbf{E}\,(\widehat{\theta} - \theta)(\widehat{\theta} - \theta)^T = (ZZ^T)^{-1}\sigma^2\left[1 + \frac{1 + o(1)}{n}\sum_{s=1}^{\infty} h_s\right]. \tag{15}$$

(2000) Golubev, Y. and Härdle, W.
On the second order minimax estimation in partial linear models.

Math. Methods of Statistics, 9, 2, 160 - 175

If ξ_i are Gaussian then

$$\inf_{\tilde\theta}\sup_{m\in\mathbf{W}_0}\sup_{\theta\in\mathbb{R}^d}\mathbf{E}(\tilde\theta-\theta)(\tilde\theta-\theta)^T=(ZZ^T)^{-1}\sigma^2\left[1+\frac{1+o(1)}{n}\sum_{s=1}^{\infty}h_s\right].$$

Proof. Noting that $h_s=0$ when $s>Cn^{1/(2\beta+1)}$ we see that the condition (13) is fulfilled. Thus we are in a position to apply Theorem 2. The second order term in the right-hand side of (14) is evaluated as follows. Since $m\in\mathbf{W}_0$ we get by (6)

$$\sup_{m\in\mathbf{W}_0}\sum_{k=1}^{n}\nu_k^2(1-h_k)^2=L\max_k\lambda_k^{-1}(1-h_k)^2=L\omega^2.$$

Therefore one obtains by (9)

$$\begin{aligned}L\omega^2+\frac{\sigma^2}{n}\sum_{k=1}^{n}h_k^2&=\frac{\omega\sigma^2}{n}\sum_{s=1}^{\infty}\lambda_s^{1/2}\left[1-\omega\sqrt{\lambda_s}\right]_+ +\frac{\sigma^2}{n}\sum_{s=1}^{\infty}\left[1-\omega\sqrt{\lambda_s}\right]_+^2\\&=\frac{\sigma^2}{n}\sum_{s=1}^{\infty}h_s,\end{aligned}$$

thus proving (15). The rest of the proof follows from Theorem 1. □

Remark 1. When the distribution of the noise is known but non-Gaussian the penalized maximum likelihood estimator

$$\widehat\theta_p=\arg\max_{\theta\in\mathbb{R}^d}\max_{\nu_k}\left\{\sum_{i=1}^{n}\log p_\xi\left[Y_i-\theta^TZ_i-\sum_k\nu_k\psi_k(X_i)\right]-\frac{1}{2}\|\Sigma^{-1}\nu\|^2\right\}$$

may be used. We are sure that under some additional smoothness assumptions on the density $p_\xi(\cdot)$ one can prove that the risk of $\widehat\theta_p$ admits expansion (14) with $\sigma^2=I_\xi^{-1}$.

Remark 2. The assumption $\int_0^1 q(t)m(t)\,dt=0$ may seem very restrictive from a practical point of view. Let us indicate how to extend our approach to the ordinary Sobolev class

$$\mathbf{W}=\left\{m:\int_0^1[m^{(\beta)}(x)]^2\,dx\le L\right\}.$$

Consider the new regressors $Z_i'=(Z_i^T,1)^T$ and the new parameter $\theta'\in\mathbb{R}^{d+1}$, $\theta'^T=(\theta_1,\dots,\theta_d,\nu_0)$, where $\nu_0\in\mathbb{R}^1$ is yet another nuisance parameter. Since

$$Z'Z'^T=\begin{pmatrix}ZZ^T & \bar Z\\ \bar Z^T & 1\end{pmatrix},\qquad \bar Z_i=\sum_{k=1}^{d}Z_{ik},\tag{16}$$

we easily obtain

$$(Z'Z'^T)^{-1}=\begin{pmatrix}(ZZ^T-\bar Z\bar Z^T)^{-1} & -(ZZ^T-\bar Z\bar Z^T)^{-1}\bar Z\\ -\bar Z^T(ZZ^T-\bar Z\bar Z^T)^{-1} & 1+\bar Z^T(ZZ^T-\bar Z\bar Z^T)^{-1}\bar Z\end{pmatrix},$$

Math. Methods of Statistics, 9, 2, 160 - 175

so that to find the minimax risk over $\theta \in \mathbb{R}^d$ we have just to replace $(ZZ^T)^{-1}$ by $(ZZ^T - \bar{Z}\bar{Z}^T)^{-1}$ in Theorems 1–3. In particular, (15) can be rewritten as

$$\sup_{m \in \mathbf{W}} \sup_{\theta \in \mathbb{R}^d} \mathbf{E}\,(\widehat{\theta} - \theta)(\widehat{\theta} - \theta)^T = (ZZ^T - \bar{Z}\bar{Z}^T)^{-1}\sigma^2\left[1 + \frac{1 + o(1)}{n}\sum_{s=1}^{\infty} h_s\right].$$

Remark 3. The popular spline estimator

$$\widehat{\theta}_{\mathrm{spl}} = \arg\min_{\theta \in \mathbb{R}^d} \min_{m} \left\{ \frac{1}{n}\sum_{i=1}^{n} \left[Y_i - \theta^T Z_i - m(X_i)\right]^2 + \mu_n \int_0^1 [m^{(\beta)}(t)]^2\, dt \right\}, \tag{17}$$

can be also evaluated by Theorem 2. It suffices to note that

$$\int_0^1 [m^{(\beta)}(x)]^2\, dx = \sum_{k=1}^{\infty} \nu_k^2 \lambda_k.$$

Therefore the estimator (17) is equivalent to (10) with the new predictors matrix Z' defined by (16) and the regularization matrix $\Sigma_{kk} = \sigma(\mu_n n \lambda_k)^{-1/2}$. Thus one obtains that $h_k = (1 + \mu_n \lambda_k)^{-1}$ and in view of (14), (5)

$$\sup_{m \in \mathbf{W}} \mathbf{E}\,(\widehat{\theta}_{\mathrm{spl}} - \theta)(\widehat{\theta}_{\mathrm{spl}} - \theta)^T = (ZZ^T - \bar{Z}\bar{Z}^T)^{-1}$$
$$\times\, \sigma^2\left\{1 + [1 + o(1)]\left[\frac{L\mu_n}{4} + \frac{\sigma^2}{n\mu_n^{1/(2\beta)}\pi}\left(\int_0^1 q^{1/(2\beta)}(x)\, dx\right)^{-1}\int_0^{\infty} \frac{1}{(1 + x^{2\beta})^2}\, dx\right]\right\}.$$

Interesting simulation results about second order performance of the spline estimators can be found in Schimek [13].

5. Appendix

5.1. Proof of Theorem 1. We begin with a lower bound for the Bayesian risk in a slightly more general situation. Assume that the nuisance function $m(\cdot)$ has the form $m(x) = \sum_{k=1}^{\infty} \nu_k \varphi_k(x)$, where $\varphi_k(x)$ is a certain orthonormal system in $\mathbf{L}_q^2[0,1]$ such that $\int_0^1 q(x)\varphi_k(x)\, dx = 0$.

Let the nuisance parameters ν_k be i.i.d. $\mathcal{N}(0, \sigma_k^2)$ and let the parameters of interest θ_k be also Gaussian random variables with zero mean and $\mathbf{E}\,\theta\theta^T = \varepsilon^{-2}(ZZ^T)^{-1}$. Denote by $R_\epsilon^n(\tilde{\theta}) = \mathbf{E}\,(\tilde{\theta} - \theta)(\tilde{\theta} - \theta)^T$ the Bayesian risk.

Lemma 1. *Let* $\sup_{k \geq 1} \int_0^1 \varphi_k^4(x)dx < A$. *Then*

$$\inf_{\tilde{\theta}} R_\epsilon^n(\tilde{\theta}) \geq I_\xi^{-1}(ZZ^T)^{-1}\frac{I_\xi}{\varepsilon^2 + I_\xi}\left[1 + \frac{I_\xi}{n(\varepsilon^2 + I_\xi)}\sum_{k=1}^{n} h_k\left(1 - \frac{A}{n}\sum_{s=1}^{n} h_s\right)\right],$$

where

$$h_k = \frac{n\sigma_k^2 I_\xi}{1 + n\sigma_k^2 I_\xi}. \tag{18}$$

(2000) Golubev, Y. and Härdle, W.
On the second order minimax estimation in partial linear models.

Math. Methods of Statistics, 9, 2, 160 - 175

Proof. Denote $\mu = (\theta_1, \dots, \theta_d, \nu_1, \dots)^T$. It follows from the Van Trees [19] inequality that the mean-square risk of any estimator $\widehat{\mu}$ based on the data $(Y_1, \dots, Y_n)$ is bounded from below by

$$\mathbf{E}\{(\widehat{\mu}-\mu)(\widehat{\mu}-\mu)^T \mid X_1, \dots, X_n\} \geq (I + I_\mu)^{-1}, \tag{19}$$

where I_μ is the Fisher information matrix of the prior distribution. This matrix is diagonal with the entries

$$[I_\mu]_{kk} = \begin{cases} \varepsilon^2 Z Z^T, & k \leq d, \\ \sigma_k^{-2}, & k > d. \end{cases}$$

The matrix I in (19) is the ordinary information matrix with

$$I_{kl} = \mathbf{E}\left\{ \frac{\partial}{\partial \mu_k} \sum_{j=1}^{n} \log p_\xi \left[Y_j - \theta^T Z_j - \sum_{k=1}^{n} \nu_k \phi_k(X_j) \right] \right. \tag{20}$$
$$\left. \times \frac{\partial}{\partial \mu_l} \sum_{j=1}^{n} \log p_\xi \left[Y_j - \theta^T Z_j - \sum_{k=1}^{n} \nu_k \phi_k(X_j) \right] \;\middle|\; X_1, \dots X_n, \right\}.$$

Therefore it is clear that I admits the representation

$$I = I_\xi \begin{pmatrix} Z Z^T & Z\Phi^T \\ \Phi Z^T & \Phi\Phi^T \end{pmatrix}, \tag{21}$$

where Φ is the matrix with entries $\Phi_{kl} = \varphi_k(X_l)$. Denote for brevity by Σ the diagonal matrix with entries $\Sigma_{kk} = \sigma_k$ and let

$$A = \begin{pmatrix} Z Z^T + \varepsilon^2 I_\xi^{-1} Z Z^T & 0 \\ 0 & nE + I_\xi^{-1} \Sigma^{-2} \end{pmatrix}, \quad B = \begin{pmatrix} 0 & Z\Phi^T \\ \Phi Z^T & \Phi\Phi - nE \end{pmatrix}.$$

Then we have by (20), (21)

$$(I + I_\mu)^{-1} = I_\xi^{-1}(A+B)^{-1} = I_\xi^{-1}(E + A^{-1}B)^{-1}A^{-1} \tag{22}$$
$$\geq I_\xi^{-1}[E - A^{-1}B + (A^{-1}B)^2 - (A^{-1}B)^3]A^{-1}.$$

It is obvious that A can be inverted as

$$A^{-1} = \begin{pmatrix} V_\varepsilon^{-1} & 0 \\ 0 & H/n \end{pmatrix},$$

where $V_\varepsilon = Z Z^T (1 + \varepsilon^2 I_\xi^{-1})$. The matrix H is diagonal with entries $H_{kk} = h_k$, where h_k are defined by (18). A simple algebra reveals that

$$A^{-1}B = \begin{pmatrix} 0 & V_\varepsilon^{-1} Z\Phi^T \\ H\Phi Z^T/n & H(\Phi\Phi^T/n - E) \end{pmatrix},$$

$$(A^{-1}B)^2 = \begin{pmatrix} V_\varepsilon^{-1} Z\Phi^T H\Phi Z^T/n & * \\ H(\Phi\Phi^T/n - E)H\Phi Z^T/n & * \end{pmatrix},$$

$$(A^{-1}B)^3 = \begin{pmatrix} V_\varepsilon^{-1} Z\Phi^T(\Phi\Phi^T/n - E)H\Phi Z^T/n & * \\ * & * \end{pmatrix}.$$

(2000) Golubev, Y. and Härdle, W.
On the second order minimax estimation in partial linear models.

Math. Methods of Statistics, 9, 2, 160 - 175

Here and later in the text $*$ denotes a matrix that is not needed in further calculations. Therefore by the above equations and (19), (20) we get that for any estimator $\widehat{\theta}$

$$R^n(\widehat{\theta}) \geq I_\xi^{-1} V_\epsilon^{-1} \Big[E + \frac{1}{n} V_\epsilon^{-1} \mathbf{E}\, Z\Phi^T H \Phi Z^T - \frac{1}{n} V_\epsilon^{-1} \mathbf{E}\, Z\Phi^T H(\Phi\Phi/n - E) H \Phi Z^T \Big]. \tag{23}$$

We have evidently $\mathbf{E}\,\Phi^T H\Phi = \mathbf{E}\,\mathrm{tr}\, H$. For the last term in the right-hand side of (23) one obtains by the Cauchy–Schwarz inequality

$$\begin{aligned} &\mathbf{E}\,\Phi^T H(\Phi\Phi^T/n - E) H\Phi_{ls} \\ &\quad = \mathbf{E}\,\frac{1}{n} \sum_{k,m=1}^{n} h_k h_m \sum_{p=1}^{n} \varphi_k(X_l)\big(\varphi_k(X_p)\varphi_m(X_p) - \delta_{km}\big)\varphi_m(X_s) \\ &\quad \leq E\frac{1}{n}\Big(\sum_{m=1}^{n} h_m\Big)^2 \sup_{k\geq 1} \int_0^1 \varphi_k^4(x)\, dx. \end{aligned}$$

These relations together with (23) prove the lemma. □

Proof of Theorem 1. We follow the idea proposed by Pinsker [10] and developed for second order minimax estimation in Belitser and Levit [3] and Golubev and Levit [6]. Choose (see (8), (9))

$$\sigma_k^2 = (1-\delta)\sigma^2 \frac{h_k}{n(1-h_k)} = (1-\delta)\sigma^2 \frac{[1-\omega\sqrt{\lambda_k}]_+}{n\omega\sqrt{\lambda_k}},$$

where $\delta > 0$. Let ζ_k be i.i.d. $\mathcal{N}(0,1)$. Suppose that the nuisance parameters ν_k have the form $\nu_k = \sigma_k\zeta_k$. Our first step is to show that the nuisance function $m(x) = \sum_{k=1}^{\infty} \nu_k \psi_k(k)$ belongs to the Sobolev class $\mathbf{W}_0$ with a high probability. We have by (9)

$$\mathbf{P}\{m \notin \mathbf{W}_0\} = \mathbf{P}\Big\{\sum_{k=1}^{\infty} \sigma_k^2(\zeta_k^2 - 1)\lambda_k > \delta L\Big\}. \tag{24}$$

Denote for brevity

$$v_k = \sigma_k^2\lambda_k, \quad \eta = \frac{1}{\sqrt{2}\|v\|}\sum_{i=1}^{\infty} v_i(\xi_i^2 - 1), \quad \text{and} \quad m = \frac{\max_k |v_k|}{\|v\|}.$$

By the Markov inequality using the formula

$$-\log(1-x) = \sum_{k=1}^{\infty} \frac{x^k}{k}$$

Math. Methods of Statistics, 9, 2, 160 - 175

one obtains, for any $0 < t < (\sqrt{2}m)^{-1}$,

$$\begin{aligned}
\mathbf{P}\{\eta > x\} &\le \exp(-tx)\mathbf{E}\exp(t\eta) \\
&= \exp(-tx)\prod_{i=1}^{\infty}\exp\left[-\frac{tv_i}{\sqrt{2}\|v\|} - \frac{1}{2}\log\left(1 - \frac{\sqrt{2}tv_i}{\|v\|}\right)\right] \\
&= \exp(-tx)\exp\left[\sum_{k=2}^{\infty}\sum_{i=1}^{\infty}\frac{1}{2k}\left(\frac{\sqrt{2}tv_i}{\|v\|}\right)^k\right] \\
&\le \exp(-tx)\exp\left[\frac{1}{m^2}\sum_{k=2}^{\infty}\frac{1}{2k}(\sqrt{2}tm)^k\right] \\
&\le \exp(-tx)\exp\left[-\frac{1}{2m^2}\log(1-\sqrt{2}tm) - \frac{t}{\sqrt{2}m}\right].
\end{aligned}$$

Minimization of the last expression with respect to t yields

$$\mathbf{P}\{\eta > x\} \le \exp\left\{\frac{1}{2m^2}\log[1+\sqrt{2}xm] - \frac{x}{\sqrt{2}m}\right\}. \tag{25}$$

It is easy to see using (5) that the following relations hold

$$\max_k |v_k| \asymp C(1-\delta)\frac{1}{n}\omega^{-2}, \qquad \|v\| \asymp C(1-\delta)\frac{1}{n}\omega^{-2-1/(2\beta)},$$

$$\sum_{k=1}^{\infty} v_k \asymp C(1-\delta)\frac{1}{n}\omega^{-2-1/(\beta)}.$$

Here and later in the text C is a generic constant depending on β, different in different occasions. With this in mind and with $x = 2^{-1/2}\delta L\|v\|^{-1}$ we have in view of (24), (25)

$$\mathbf{P}\{m \notin \mathbf{W}_0\} \le \exp\{-\tfrac{1}{2}[1+O(\delta)]x^2\} \le \exp\{-C\delta^2 n^{-1/(2\beta+1)}\}. \tag{26}$$

Now we are ready to complete the proof. Let

$$\pi(x) = \frac{1}{(2\pi/n)^{d/2}\det{}^{1/2} ZZ^T}\exp\left\{-\frac{x^T Z^T Z x}{2n}\right\}$$

be the Gaussian probability density in $\mathbb{R}^d$ of the prior distribution of the vector θ. By the triangle inequality one obtains

$$\begin{aligned}
\inf_{\tilde\theta}\sup_{m\in\mathbf{W}_0}\sup_{\theta\in\mathbb{R}^d}\mathbf{E}(\tilde\theta-\theta)(\tilde\theta-\theta)^T &\ge \inf_{\tilde\theta}\sup_{m\in\mathbf{W}_0}\sup_{ZZ^T\theta\theta^T<n^2E}\mathbf{E}(\tilde\theta-\theta)(\tilde\theta-\theta)^T \\
&\ge \inf_{ZZ^T\tilde\theta\tilde\theta^T<n^2E}\sup_{m\in\mathbf{W}_0}\sup_{ZZ^T\theta\theta^T<n^2E}\mathbf{E}(\tilde\theta-\theta)(\tilde\theta-\theta)^T \\
&\ge \inf_{ZZ^T\tilde\theta\tilde\theta^T<n^2E}\sup_{m\in\mathbf{W}_0}\mathbf{E}\int_{ZZ^T\theta\theta^T<n^2E}(\tilde\theta-\theta)(\tilde\theta-\theta)^T\pi(\theta)\,d\theta \\
&\ge \inf_{ZZ^T\tilde\theta\tilde\theta^T<n^2E}\mathbf{E}\,\mathbf{1}\{m\in\mathbf{W}_0\}\int_{ZZ^T\theta\theta^T<n^2E}(\tilde\theta-\theta)(\tilde\theta-\theta)^T\pi(\theta)\,d\theta \\
&\ge \inf_{\tilde\theta}\mathbf{E}(\tilde\theta-\theta)(\tilde\theta-\theta)^T - n^2(ZZ^T)^{-1}\{\mathbf{P}\{m\notin\mathbf{W}_0\} + \exp(-Cn)\}.
\end{aligned}$$

Thus using (26) with $\delta = \log^{-1} n$ and Lemma 1 we complete the proof. □

(2000) Golubev, Y. and Härdle, W.
On the second order minimax estimation in partial linear models.

Math. Methods of Statistics, 9, 2, 160 - 175

5.2. PROOF OF THEOREM 2. Let $\mu = (\theta_1, \dots, \theta_d, \nu_1, \dots)^T$. Differentiating (10) with respect to μ one obtains that $\widehat{\mu}$ is a root of the equation

$$\begin{pmatrix} ZZ^T & Z\Psi^T \\ \Psi Z^T & \Psi\Psi^T + \sigma^2\Sigma^{-2} \end{pmatrix} (\mu - \widehat{\mu}) = -\begin{pmatrix} Z^T\xi \\ \Psi^T\xi \end{pmatrix} + \begin{pmatrix} 0 & 0 \\ 0 & \sigma^2\Sigma^{-2} \end{pmatrix} \mu.$$

Therefore

$$\mathbf{E}\{(\widehat{\mu} - \mu)(\widehat{\mu} - \mu)^T \mid X_1, \dots, X_n\} = S^{-1} + S^{-1} \begin{pmatrix} 0 & 0 \\ 0 & \sigma^4\Sigma^{-2}\nu\nu^T\Sigma^{-2} - \sigma^2\Sigma^{-2} \end{pmatrix} S^{-1}, \tag{27}$$

where $S = A + B$, with

$$A = \begin{pmatrix} ZZ^T & 0 \\ 0 & nE + \sigma^2\Sigma^{-2} \end{pmatrix}, \qquad B = \begin{pmatrix} 0 & Z\Psi^T \\ \Psi Z^T & \Psi\Psi^T - nE \end{pmatrix}. \tag{28}$$

In order to compute S^{-1} we use the Taylor formula with respect to $A^{-1/2}BA^{-1/2}$ in the right-hand side of the equation

$$S^{-1} = A^{-1/2}(E + A^{-1/2}BA^{-1/2})^{-1}A^{-1/2}.$$

Let $H = (E + \sigma^2\Sigma^{-2}/n)^{-1}$. Thus we have to check that the operator norm of the matrix

$$A^{-1/2}BA^{-1/2} = \begin{pmatrix} 0 & (ZZ^T)^{-1/2}Z\Psi^T H^{1/2}/\sqrt{n} \\ H^{1/2}\Psi Z^T(ZZ^T)^{-1/2}/\sqrt{n} & H^{1/2}(\Psi\Psi^T/n - E)H^{1/2} \end{pmatrix} \tag{29}$$

is sufficiently small. This is proved in the following lemma.

Lemma 2. *Let* $\|(ZZ^T)^{-1}\| \sum_{i=1}^n Z_{ki}^2 \le C_0$ *for some* $C_0 < \infty$. *Then for*

$$x \le \min\left\{n, \min_k \left[\sum_{i=1}^n \frac{Z_{ki}^2}{\max_i Z_{ki}^2}\right]\right\}$$

and for sufficiently large $C > 0$ *we have*

$$\mathbf{P}\left\{\|A^{-1/2}BA^{-1/2}\|^2 > C(1 + x)\frac{1}{n}\operatorname{tr}^2 H\right\} \le \exp(-x^2/C).$$

Proof. Denoting $\zeta_{kli} = \psi_k(X_i)\psi_l(X_i) - \delta_{kl}$ we have

$$\left\|H^{1/2}(\Psi\Psi^T/n - E)H^{1/2}\right\|^2 \le \frac{1}{n}\sum_{k,l=1}^{\infty} h_k h_l \left[\frac{1}{\sqrt{n}}\sum_{i=1}^n \zeta_{kli}\right]^2.$$

(2000) Golubev, Y. and Härdle, W.
On the second order minimax estimation in partial linear models.

Math. Methods of Statistics, 9, 2, 160 - 175

It follows that

$$\mathbf{P}\left\{\left\|H^{1/2}(\Psi\Psi^T - nE)H^{1/2}\right\| > \frac{1+x}{n}\sum_{k,l=1}^{\infty} h_k h_l \mathbf{E}\,\zeta_{kl1}^2\right\} \tag{30}$$

$$\le \mathbf{P}\left\{\sum_{k,l=1}^{\infty} h_k h_l\left[\left(\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\zeta_{kli}\right)^2 - \mathbf{E}\,\zeta_{kl1}^2\right] > Cx\mathrm{tr}^2 H\right\}$$

$$\le \mathbf{P}\left\{\frac{1}{n}\sum_{k,l=1}^{\infty} h_k h_l \sum_{i=1}^{n}(\zeta_{kli}^2 - \mathbf{E}\,\zeta_{kli}^2) > Cx\mathrm{tr}^2 H\right\}$$

$$+ \mathbf{P}\left\{\frac{1}{n}\sum_{k,l=1}^{\infty} h_k h_l \sum_{i\ne j}^{n}\zeta_{kli}\zeta_{klj} > Cx\mathrm{tr}^{\,2} H\right\}.$$

Since the random variables $\zeta_{kli}^2 - \mathbf{E}\,\zeta_{kli}^2$ are independent and bounded we obtain by the Markov inequality and by the Taylor formula

$$\mathbf{P}\left\{\frac{1}{n}\sum_{k,l=1}^{\infty} h_k h_l \sum_{i=1}^{n}(\zeta_{kli}^2 - \mathbf{E}\,\zeta_{kli}^2) > Cx\,\mathrm{tr}^2 H\right\} \tag{31}$$

$$\le \exp(-C\lambda x\,\mathrm{tr}^2 H)\mathbf{E}\exp\left\{\frac{\lambda}{n}\sum_{k,l=1}^{\infty} h_k h_l \sum_{i=1}^{n}(\zeta_{kli}^2 - \mathbf{E}\,\zeta_{kli}^2)\right\}$$

$$\le \exp(-C\lambda x\,\mathrm{tr}^2 H)\exp(C\lambda^2\mathrm{tr}^4 H/n)$$

provided that

$$\lambda\mathrm{tr}^2 H/n \le 1. \tag{32}$$

By the same arguments for the last term in (30) we have

$$\mathbf{P}\left\{\frac{1}{n}\sum_{k,l=1}^{\infty} h_k h_l \sum_{i\ne j}^{n}\zeta_{kli}\zeta_{klj} > Cx\,\mathrm{tr}^2 H\right\} \tag{33}$$

$$\le \exp\{-C\lambda x\,\mathrm{tr}^2 H\}\mathbf{E}\,\exp\left\{\frac{\lambda}{n}\sum_{k,l=1}^{\infty} h_k h_l \sum_{i\ne j}^{n}\zeta_{kli}\zeta_{klj}\right\}$$

$$\le \exp\{-C\lambda x\,\mathrm{tr}^2 H\}\exp\{\lambda^2\,\mathrm{tr}^4 H\}.$$

Choosing $\lambda = x\,\mathrm{tr}^{-2} H$ we get in view of (30)–(33) that for $x < n$

$$\mathbf{P}\left\{\left\|H^{1/2}(\Psi\Psi^T - nE)H^{1/2}\right\| > C(1+x)\frac{1}{n}\mathrm{tr}^2 H\right\} \le \exp(-x^2/C). \tag{34}$$

We use almost the same arguments to evaluate the operator norm of $Z\Psi^T H^{1/2}$. Notice that

$$\|Z\Psi^T H^{1/2}\|^2 \le \frac{1}{n}\sum_{s=1}^{\infty}\sum_{k=1}^{d} h_s\left(\sum_{i=1}^{n} Z_{ki}\psi_s(X_i)\right)^2. \tag{35}$$

(2000) Golubev, Y. and Härdle, W.
On the second order minimax estimation in partial linear models.

Math. Methods of Statistics, 9, 2, 160 - 175

Denoting for brevity $\|Z_k\|^2 = \sum_{i=1}^n Z_{ki}^2$ we obtain

$$
\begin{aligned}
(36)\qquad & \mathbf{P}\left\{\sum_{s=1}^{\infty} h_s \left(\sum_{i=1}^{n} Z_{ki}\psi_s(X_i)\right)^2 > (1+x)\,\mathrm{tr}\,H\|Z_k\|^2\right\} \\
&= \mathbf{P}\left\{\sum_{s=1}^{\infty} h_s \left[\left(\sum_{i=1}^{n} Z_{ki}\psi_s(X_i)\right)^2 - \|Z_k\|^2\right] \geq x\,\mathrm{tr}\,H\|Z_k\|^2\right\} \\
&\leq \mathbf{P}\left\{\sum_{s=1}^{\infty} h_s \sum_{i=1}^{n} Z_{ki}^2[\psi_s^2(X_i) - 1] \geq \frac{x}{2}\,\mathrm{tr}\,H\|Z_k\|^2\right\} \\
&\quad + \mathbf{P}\left\{\sum_{s=1}^{\infty} h_s \sum_{i\neq j}^{n} Z_{ki}Z_{kj}\psi_s(X_i)\psi_s(X_j) \geq \frac{x}{2}\,\mathrm{tr}\,H\|Z_k\|^2\right\}.
\end{aligned}
$$

Next by the Markov inequality and the Taylor formula we get

$$
\begin{aligned}
(37)\qquad & \mathbf{P}\left\{\sum_{s=1}^{\infty} h_s \sum_{i=1}^{n} Z_{ki}^2[\psi_s^2(X_i) - 1] \geq \frac{x}{2}\,\mathrm{tr}\,H\|Z_k\|^2\right\} \\
&\leq \exp\left(-\frac{1}{2}\lambda x\,\mathrm{tr}\,H\|Z_k\|^2\right)\mathbf{E}\,\exp\left\{\lambda\sum_{s=1}^{\infty} h_s \sum_{i=1}^{n} Z_{ki}^2[\psi_s^2(X_i) - 1]\right\} \\
&\leq \exp\left(-\frac{1}{2}\lambda x\,\mathrm{tr}\,H\|Z_k\|^2\right)\exp\left\{C\lambda^2\,\mathrm{tr}^2 H\sum_{i=1}^{n} Z_{ki}^4\right\}
\end{aligned}
$$

provided that

$$
(38)\qquad \lambda \max_i Z_{ki}^2\,\mathrm{tr}H < 1.
$$

The last term in (36) is evaluated in the same way

$$
\begin{aligned}
&\mathbf{P}\left\{\sum_{s=1}^{\infty} h_s \sum_{i\neq j}^{n} Z_{ki}Z_{kj}\psi_s(X_i)\psi_s(X_j) \geq \frac{x}{2}\,\mathrm{tr}H\|Z_k\|^2\right\} \\
&\quad \leq \exp\left(-\tfrac{1}{2}\lambda x\,\mathrm{tr}H\|Z_k\|^2\right)\exp\left(C\lambda^2\,\mathrm{tr}\,H^2\|Z_k\|^4\right).
\end{aligned}
$$

Hence with $\lambda = Cx\,\mathrm{tr}^{-1}H\,\|Z_k\|^{-2}$ we conclude by (35)–(38) that

$$
\mathbf{P}\left\{\|Z\Psi^T H^{1/2}\| > C(1+x)\frac{1}{n}\,\mathrm{tr}\;H\|Z_k\|^2\right\} \leq \exp\left(-\frac{x^2}{C}\right).
$$

The proof of the lemma follows now from the above inequality, (34) and (29). □

Proof of Theorem 2. We have evidently by (10) $\|Z^T(\widehat{\theta} - \theta)\| \leq \|\xi + \Psi^T\nu\|$. Therefore

$$
\mathbf{E}\,\|\widehat{\theta} - \theta\|^{2+2\delta} \leq C(\delta)n^{1+\delta}\|(ZZ^T)^{-1/2}\|^{-1-\delta},
$$

(2000) Golubev, Y. and Härdle, W.
On the second order minimax estimation in partial linear models.

Math. Methods of Statistics, 9, 2, 160 - 175

and one obtains by the Hölder inequality and Lemma 2 with $x = (C \log n)^{1/2}$

$$
\begin{aligned}
(39)\quad \mathbf{E}(\widehat{\theta}-\theta)(\widehat{\theta}-\theta)^T &= \mathbf{E}(\widehat{\theta}-\theta)(\widehat{\theta}-\theta)^T \mathbf{1}\{\|A^{-1/2}BA^{-1/2}\| \le \varepsilon\} \\
&\quad + \mathbf{E}(\widehat{\theta}-\theta)(\widehat{\theta}-\theta)^T \mathbf{1}\{\|A^{-1/2}BA^{-1/2}\| > \varepsilon\} \\
&\le \mathbf{E}(\widehat{\theta}-\theta)(\widehat{\theta}-\theta)^T \mathbf{1}\{\|A^{-1/2}BA^{-1/2}\| \le \varepsilon\} \\
&\quad + n\|(ZZ^T)^{-1/2}\|^{-2}\mathbf{P}^{\delta/(1+\delta)}\{\|A^{-1/2}BA^{-1/2}\| > \varepsilon\}O(E) \\
&\le \mathbf{E}(\widehat{\theta}-\theta)(\widehat{\theta}-\theta)^T \mathbf{1}\{\|A^{-1/2}BA^{-1/2}\| \le \varepsilon\} + n^{-1}\|(ZZ^T)^{-1}\|^{-1}O(E),
\end{aligned}
$$

where $O(E)$ is a bounded $d \times d$-matrix. Noting that (cf. (29))

$$
(A^{-1/2}BA^{-1/2})^2 = \begin{pmatrix} (ZZ^T)^{-1/2}Z\Psi^T H \Psi Z^T (ZZ^T)^{-1/2}/n & * \\ H^{1/2}(\Psi\Psi^T/n - E)H\Psi Z^T (ZZ^T)^{-1/2}/\sqrt{n} & * \end{pmatrix}
$$

we have by the Taylor formula

$$
S^{-1} = \begin{pmatrix} (ZZ^T)^{-1} & * \\ * & * \end{pmatrix} + \begin{pmatrix} (ZZ^T)^{-1}Z\Psi^T H\Psi Z^T (ZZ^T)^{-1}[E + \varepsilon O(E)]/n & * \\ * & * \end{pmatrix}.
$$

Therefore

$$
\begin{aligned}
(40)\quad &\mathbf{E}\, S^{-1}\mathbf{1}\{\|A^{-1/2}BA^{-1/2}\| \le \varepsilon\} \\
&\qquad \le \begin{pmatrix} (ZZ^T)^{-1} & * \\ * & * \end{pmatrix} + \begin{pmatrix} (ZZ^T)^{-1}[E + \varepsilon O(E)]n^{-1}\sum_{k=1}^{\infty} H_{kk} & * \\ * & * \end{pmatrix}.
\end{aligned}
$$

The last term in the right-hand side of (27) is bounded from above by the same arguments, so that we have

$$
\begin{aligned}
(41)\quad & S^{-1}\begin{pmatrix} 0 & 0 \\ 0 & \sigma^4\Sigma^{-2}\nu\nu^T\Sigma^{-2} - \sigma^2\Sigma^{-2} \end{pmatrix} S^{-1} \\
&= \begin{pmatrix} (ZZ^T)^{-1/2}Z\Psi^T H(\sigma^4\Sigma^{-2}\nu\nu^T\Sigma^{-2} - \sigma^2\Sigma^{-2})H\Psi Z^T (ZZ^T)^{-1/2}/n^2 & * \\ * & * \end{pmatrix}.
\end{aligned}
$$

Noticing that $\Sigma^{-2} = n\sigma^{-2}(H^{-1} - E)$ one obtains

$$
\begin{aligned}
(42)\qquad & \mathbf{E}\,\Psi^T H(\sigma^4\Sigma^{-2}\nu\nu^T\Sigma^{-2} - \sigma^2\Sigma^{-2})H\Psi \\
&\qquad = E\,\mathrm{tr}\,\{H(\sigma^4\Sigma^{-2}\nu\nu^T\Sigma^{-2} - \sigma^2\Sigma^{-2})H\} \\
&\qquad = E\left[n^2\sum_{k=1}^{\infty}(1-h_k)^2\nu_k^2 - n\sum_{k=1}^{\infty}(1-h_k)h_k\right].
\end{aligned}
$$

On the other hand we have by Lemma 2 and by the Cauchy–Schwarz inequality

$$
\begin{aligned}
&\mathbf{E}\,\Psi^T H(\sigma^4\Sigma^{-2}\nu\nu^T\Sigma^{-2} - \sigma^2\Sigma^{-2})H\Psi\mathbf{1}\{\|A^{-1/2}BA^{-1/2}\| \le \varepsilon\} \\
&\qquad = \mathbf{E}\,\mathrm{tr}\,\{H(\sigma^4\Sigma^{-2}\nu\nu^T\Sigma^{-2} - \sigma^2\Sigma^{-2})H\} + O(E).
\end{aligned}
$$

Thus the proof of the theorem follows from (27), (39)–(42). □

(2000) Golubev, Y. and Härdle, W.
On the second order minimax estimation in partial linear models.

Math. Methods of Statistics, 9, 2, 160 - 175

Acknowledgments

We would like to thank the anonymous referee for numerous comments on an earlier version of the paper that led to a much improved exposition of the material.

References

[1] R. J. Carroll and W. Härdle, *A note on second-order effects in a semiparametric context*, Statistics, 20 (1989), 179–186.

[2] P. J. Bickel, C. A. J. Klaassen, Y. Ritov, and J. A. Wellner, *Efficient and Adaptive Estimation for Semiparametric Models*, Springer, New York, 1998.

[3] E. N. Belitser and B. Y. Levit, *On minimax filtering on ellipsoids*, Math. Methods Statist., 4 (1995), 259–273.

[4] P. K. Bhattacharia and P.-L. Zhao, *Semiparametric inference in partial linear models*, Ann. Statist., 25 (1997), 244–262.

[5] H. Chen, *Convergence rates for parametric components in a partly linear model*, Ann. Statist., 16 (1998), 136–146.

[6] G. K. Golubev and B. Y.Levit, *On the second order minimax estimation of distribution functions*, Math. Methods Statist., 5 (1996), 1–31.

[7] W. Härdle, H. Liang, and J. Gao, *Partially linear models*, Electronic version: http:// www.xplore-stat.de/ebooks.html (1999).

[8] N. E. Heckman, *Spline smoothing in partly linear models*, J. Roy. Statist. Soc., Ser. B, 48 (1986), 244–248.

[9] E. Mammen and S. Van de Geer, *Penalized quasi-likelihood estimation in partial linear models*, Ann. Statist., 25 (1997), 1014–1035.

[10] M. S. Pinsker, *Optimal filtering of square integrable signals in Gaussian white noise*, Problems Inform. Transmission, 16 (1980), 120–133.

[11] P. M. Robinson, *Asymptotically efficient estimation in the presence of heteroscedasticity of unknown form*, Econometrica, 55 (1987), 875–891.

[12] J. A. Rice, *Convergence rates for partially splined models*, Statist. Probab. Lett., 4 (1986), 203–208.

[13] M. G. Schimek, *Estimation and inference in partially linear models with splines*, J. Statist. Plann. Infer. (2000), to appear.

[14] T. A. Severini and J. G. Staniswalis, *Quasilikelihood estimation in semiparametric models*, J. Amer. Statist. Assoc., 89 (1994), 501–511.

[15] P. Speckman, *Spline smoothing and optimal rates of convergence in nonparametric regression models*, Ann. Stat., 13 (1985), 970–983.

[16] P. Speckman, *Kernel smoothing in partial linear models*, J. Roy. Statist. Soc., Ser. B, 50 (1988), 413–416.

[17] V. Tikhomirov, *Fundamental Principles of the Theory of Extremal Problems*, Wiley, New York, 1986.

[18] F. Utreras, *Sur le choix des parametres d'adjustement dans le lissage par fonction spline*, Numer. Math., 34 (1980), 15-28.

[19] H. L. Van Trees, *Detection, Estimation and Modulation Theory*, Vol. 1, Wiley, New York, 1968.

[*Received March* 1999; *revised May* 2000]

Statistical Inference for Stochastic Processes **3:** 263–276, 2000.

Semiparametric Bootstrap Approach to Hypothesis Tests and Confidence Intervals for the Hurst Coefficient

PETER HALL[1], WOLFGANG HÄRDLE[2], TORSTEN KLEINOW[2] and PETER SCHMIDT[3]

[1] *Centre for Mathematics and its Applications, Australian National University, Canberra, ACT 0200, Australia*

[2] *Institut für Statistik und Ökonometrie, Humboldt-Universität zu Berlin, Spandauer Str. 1, D–10178 Berlin, Germany*

[3] *Quantitative Research, Bankgesellschaft Berlin AG, Germany*

Abstract. A major application of rescaled adjusted range analysis (R–S analysis) is to the study of price fluctuations in financial markets. There, the value of the Hurst constant, H, in a time series may be interpreted as an indicator of the irregularity of the price of a commodity, currency or similar quantity. Interval estimation and hypothesis testing for H are central to comparative quantitative analysis. In this paper we propose a new bootstrap, or Monte Carlo, approach to such problems. Traditional bootstrap methods in this context are based on fitting a process chosen from a wide but relatively conventional range of discrete time series models, including autoregressions, moving averages, autoregressive moving averages and many more. By way of contrast we suggest simulation using a single type of continuous-time process, with its fractal dimension. We provide theoretical justification for this method, and explore its numerical properties and statistical performance by application to real data on commodity prices and exchange rates.

Key words: Box-counting method, commodity price, financial market, fractal dimension, fractional Brownian motion, Gaussian process, long-range dependence, Monte Carlo, R–S analysis, self affineness, self similarity

1. Introduction

R–S analysis has its roots in early work of the British hydrologist H.E. Hurst, who investigated dependence properties of phenomena such as levels of the River Nile. The Hurst constant H, as the index of dependence is often called, always lies between 0 and 1, and equals 1/2 for processes that have independent increments. Particular interest focuses on the hypothesis that $H > 1/2$, indicating relatively long-range dependence. For example, Hurst observed that $H = 0.91$ in the case of Nile data, indicating a strength of dependence that was well beyond what could be adequately explained assuming independent increments.

Today, a principal application of R–S analysis is to the study of fluctuations in financial markets, where the value of H is variously interpreted as an indicator of range of dependence, of irregularity and of nervousness. (Adler, 1981, coined the

Statistics of Stochastic Processes, 3, 263-276

word 'erraticism' to denote a quantitative measure of 'nervousness'.) To elucidate this point we note that the fractal dimension D of sample paths of a locally self-similar or self-affine random process increases monotonically with the irregularity of those paths; and that $D = 2 - H$ (see e.g. Berry and Hannay, 1978; Sayles and Thomas, 1978; Adler, 1981, (Chapter 8); Mandelbrot et al., 1984; Hall et al., 1996). Therefore, a process with higher Hurst constant is more regular, or less erratic, or less 'nervous' than one with a lower value. For example, a time series of commodity prices that is characterised by a larger Hurst constant enjoys greater stability, over at least short periods of time; and trade in that commodity might be said to be subject to less nervousness. See for example Peters (1994).

Thus, point and interval estimation of the Hurst constant can be basic to quantitative descriptions of market fluctuations. And testing for significant differences between two Hurst constants, or between one constant and the value 1/2, is fundamental to comparative quantitative analysis of market 'nervousness'. In this paper we suggest bootstrap, or Monte Carlo, methods for constructing confidence intervals and hypothesis test for Hurst indices.

Our methods are based on the estimator $\widehat{H}$ of H derived from R–S analysis, and involve simulating the sampled process using a time-adjusted version of fractional Brownian motion. We argue that, since the 'S' part of R–S analysis corrects for inhomogeneities in the data, it is unnecessary to reproduce them in the bootstrap algorithm.

This approach differs fundamentally from more traditional methods currently used for simulation, where the model is taken to be a relatively conventional discrete time series such as an autoregression, or moving average, or autoregressive moving average. See for example Peters (1994, Chapter 5). Instead, we suggest simulating a single type of continuous stochastic process, where the degree of irregularity is determined empirically through an estimator of H. We justify this approach through theoretical analysis, and assess its numerical and statistical properties using applications to real data on stock prices.

The idea of basing the bootstrap method on a continuous rather than a discrete stochastic process has been suggested before, but in the very different context of bootstrap methods for spatial samples of data on surface roughness (Davies and Hall, 1998). There, the 'S' part of R–S analysis is usually omitted, since the observed process is generally scale-homogeneous. Such bootstrap methods are nonstandard, not least because they conform to neither the parametric nor nonparametric bootstrap approaches. They fall midway between the two, and might fairly be said to be semiparametric bootstrap methods.

2. Methodology and Theory

2.1. R–S ANALYSIS

We observe a stochastic process X_t at time points $t \in \mathcal{I} = \{0, \ldots, N\}$. Let n be an integer that is small relative to N (asymptotically, as $n/N \to \infty$), and let A denote

(2000) Hall, P., Härdle, W., Kleinow, T. and P. Schmidt
On Semiparametric Bootstrap Approach to Hypothesis Tests a. Confidence Intervals for Hurst Coefficients.

Statistics of Stochastic Processes, 3, 263-276

the integer part of N/n. Divide the 'interval' $\mathcal{I}$ into A consecutive 'subintervals', each of length n and with overlapping endpoints. In every subinterval correct the original datum X_t for location, using the mean slope of the process in the subinterval, obtaining $X_t - (t/n)(X_{an} - X_{(a-1)n})$ for all t with $(a-1)n \leqslant t \leqslant an$ and for all $a = 1, \ldots, A$. Over the a'th subinterval $\mathcal{I}_a = \{(a-1)n, (a-1)n+1, \ldots, an\}$, for $1 \leqslant a \leqslant A$, construct the smallest box (with sides parallel to the coordinate axes) such that the box contains all the fluctuations of $X_t - (t/n)(X_{an} - X_{(a-1)n})$ that occur within $\mathcal{I}_a$. Then, the height of the box equals

$$R_a = \max_{(a-1)n \leqslant t \leqslant an} \left\{ X_t - \frac{t}{n}(X_{an} - X_{(a-1)n}) \right\} - \\ - \min_{(a-1)n \leqslant t \leqslant an} \left\{ X_t - \frac{t}{n}(X_{an} - X_{(a-1)n}) \right\}.$$

Let S_a denote the empirical standard error of the n variables $X_t - X_{t-1}$, for $(a-1)n + 1 \leqslant t \leqslant an$. If the process X is stationary then S_a varies little with a. In other cases, dividing R_a by S_a corrects for the main effects of scale inhomogeneity in both spatial and temporal domains.

The total area of the boxes, corrected for scale, is proportional in n to

$$\left(\frac{R}{S}\right)_n := A^{-1} \sum_{a=1}^{A} \frac{R_a}{S_a}. \tag{2.1}$$

The slope $\widehat{H}$ of the regression of $\log (R/S)_n$ on $\log n$, for k values of n, may be taken as an estimator of the Hurst constant H describing long-range dependence of the process X. See the example Beran (1994, Chapter 1) and Peters (1994, Chapters 4–6).

This R–S analysis, or 'rescaled adjusted range' analysis, dates from Hurst (1951). If the process X is stationary then correction for scale is not strictly necessary, and we may take each S_a to be the constant 1. In that case the R–S statistic $\widehat{H}$ is a version of the box-counting estimator that is widely used in physical science applications; see for example Carter et al. (1988), Sullivan and Hunt (1988) and Hunt (1990). The box-counting estimator is related to the capacity definition of fractal dimension (Barnsley, 1988, p. 172ff), and the R–S estimator may be interpreted in the same way. Statistical properties of the box-counting estimator have been discussed by Hall and Wood (1993).

A more detailed analysis, exploiting dependence among the errors in the regression of $\log (R/S)_n$ on $\log n$, may be undertaken in place of R–S analysis. See Kent and Wood (1997) for a version of this approach in the case where scale correction is unnecessary. However, as Kent and Wood show, the advantages of the approach tend to be asymptotic in character, and sample sizes may need to be extremely large before real improvements are obtained.

(2000) Hall, P., Härdle, W., Kleinow, T. and P. Schmidt
On Semiparametric Bootstrap Approach to Hypothesis Tests a. Confidence Intervals for Hurst Coefficients.

Statistics of Stochastic Processes, 3, 263-276

2.2. APPROXIMATING THE DISTRIBUTION OF $\hat{H}$

Depending on the value of H, and on the nature of the stochastic process X, the asymptotic distribution of $\widehat{H}$ (as $N \to \infty$, for fixed k) can be Normal or Rosenblatt; the latter was introduced by Taqqu (1975), following work of Rosenblatt (1961). (More concisely, in the Rosenblatt case the asymptotic distribution of $\widehat{H}$ is that of a finite linear form in correlated Rosenblatt-distributed random variables, but for simplicity we shall refer to this as a Rosenblatt distribution). Indeed, the asymptotic distribution of $\widehat{H}$ can be Rosenblatt for $3/4 < H < 1$ and Normal for $0 < H \leqslant 3/4$; see Section 2.4. The Rosenblatt distribution that is relevant here is particularly complex, and its shape depends intimately on the unknown value of H. The distribution has not been tabulated.

If the value of k is large, i.e. the number of values of n for the linear regression is larger than the Rosenblatt approximation becomes, by virtue of the central limit theorem, similar to the Normal approximation. However, the asymptotic variance is difficult to calculate. Moreover, it is known from the work of Hall and Wood (1993) and Constantine and Hall (1994) that, due to long-range dependence, statistical performance of the estimator $\widehat{H}$ generally deteriorates for large k, and in fact optimal mean squared error properties are often achieved by keeping k fixed as N increases.

These considerations motivate Monte Carlo analysis, rather than more conventional asymptotic methods, in the range $3/4 < H < 1$. Even when H lies outside this interval there is much to be said for taking a Monte Carlo approach, however. Monte Carlo simulation can be expected to capture many of the penultimate, second-order effects that describe departure of the distribution of $\widehat{H}$ from its asymptotic limit, so that even if the limiting distribution were known, the Monte Carlo approach would be expected to provide somewhat greater accuracy than the conventional asymptotic approximation. The second-order effects arise from finiteness of N, and from the fact that stochastic fluctuations of the scale correction in R–S analysis influence the true distribution of $\widehat{H}$ even though they do not affect the limit distribution.

A more familiar example of the same phenomenon is the use of Student's t distribution to approximate the distribution of a Studentised ratio, even when the sampled distribution is not exactly Normally distributed. The Student's t approximation represents a 'penultimate' form of the Normal 'ultimate' limiting distribution. Even for data from a skew distribution of Student's t approach generally captures finite-sample properties better than the Normal approximation, despite the fact that it does not capture all second-order departures from Normality.

We shall show in Section 2.4 that in many cases the limiting distribution of $\widehat{H}$ depends only on H and a temporal scale factor. The spatial scale of the process X, and the process's potential heteroscedasticity and non-Gaussianity, do not feature in first-order asymptotic results. In large part this is a result of the 'S' component of R–S analysis. Therefore, the limiting distribution of $\widehat{H}$ is the same as it would be if X_t were ζ_t, were ζ is an elementary self-similar Gaussian process.

(2000) Hall, P., Härdle, W., Kleinow, T. and P. Schmidt
On Semiparametric Bootstrap Approach to Hypothesis Tests a. Confidence Intervals for Hurst Coefficients.

Statistics of Stochastic Processes, 3, 263-276

The Gaussian process that we have in mind is fractional Brownian motion, defined by $P(\zeta_0 = 0) = 1$, $E(\zeta_t) = 0$ and $E(\zeta_{s+t} - \zeta_s)^2 = |t|^\alpha$ for all s and t, where $\alpha = 2H \in (0, 2)$. Equivalently, ζ_t is defined to be that Gaussian process with zero mean and covariance

$$\gamma(s, t) \equiv \text{cov}(\zeta_s, \zeta_t) = \tfrac{1}{2}(|s|^\alpha + |t|^\alpha - |s - t|^\alpha).$$

See for example Beran (1994, p. 51ff) and Peters (1994, p. 183ff).

We may simulate from a discrete approximation to ζ_t, say on the points $t_j = j/\nu$ for a large integer ν, by forming the $(2p\nu+1) \times (2p\nu+1)$ covariance matrix, M, of which the (i, j)th element is $\gamma(t_i, t_j)$ for $-p\nu \leqslant i, j \leqslant p\nu$ (p an integer); and then using the spectral decomposition of M to generate Gaussian random $(2p\nu + 1)$-vectors with this covariance. Alternatively, methods of Davies and Harte (1987), or those of Wood and Chan (1994) or of the many authors whose work is surveyed by Wood and Chan, may be employed.

Denote the original data set $\{X_1, \ldots, X_N\}$ by $\mathcal{X}$. Our bootstrap algorithm is as follows. Compute the estimator $\widehat{H}$, and in the steps below, take $\alpha = 2\widehat{H}$ when constructing the fractional Brownian motion ζ, conditional on $\mathcal{X}$. Let X_t^*, for $0 \leqslant t \leqslant N$, denote a realisation of the process ζ. Compute the corresponding value $\widehat{H}^*$ of $\widehat{H}$. Take the conditional distribution of $\widehat{H}^*$, given the data $\mathcal{X}$, to be a Monte Carlo approximation to the unconditional distribution of $\widehat{H}$; or alternatively, take the conditional distribution of $\widehat{H}^* - \widehat{H}$ to approximate the unconditional distribution of $\widehat{H} - H$. These approaches give rise respectively to the two percentile methods discussed in Section 2.3.

Some of the second-order properties that this approach does not capture may be addressed by fitting a smooth estimate of scale to the process ζ. For example, we might model the variance function $\sigma(t)^2 = \text{var}(X_t)$, and thereby compute an estimator $\hat{\sigma}(\cdot)$ of $\sigma(\cdot)$; and simulate from the process $\hat{\sigma}(t)|t|^{-\alpha/2}\zeta_t$ rather than from ζ_t. In this case we should translate the time interval so as to avoid the origin.

2.3. CONFIDENCE REGIONS AND HYPOTHESIS TESTING

Confidence intervals and hypothesis tests for H may be constructed using either of the two standard bootstrap percentile methods. For example, a nominal 95% confidence interval for H is given by $(\widehat{H}^{(1)}, \widehat{H}^{(2)})$, where $\widehat{H}^{(1)}$ and $\widehat{H}^{(2)}$ are defined by either $P(\widehat{H}^* \leqslant \widehat{H}^{(1)}|\mathcal{X}) = P(\widehat{H}^* \geqslant \widehat{H}^{(2)}|\mathcal{X}) = 0.025$ or $P(\widehat{H}^* - \widehat{H} \leqslant \widehat{H} - \widehat{H}^{(2)}|\mathcal{X}) = P(\widehat{H}^* - \widehat{H} \geqslant \widehat{H} - \widehat{H}^{(1)}|\mathcal{X}) = 0.025$. A test at the 5% level of the null hypothesis that $H = 1/2$, corresponding to X being a random walk, is to reject the null if $(\widehat{H}^{(1)}, \widehat{H}^{(2)})$ does not contain the point 1/2.

Given two independent samples from long-range dependent processes, leading to respective estimators $\widehat{H}_1$ and $\widehat{H}_2$ of Hurst constants, we may generate independent realisations from respective stochastic processes $\zeta^{(1)}$ and $\zeta^{(2)}$, and thereby compute a bootstrap approximation to the distribution of $\widehat{H}_1 - \widehat{H}_2$ or of $\widehat{H}_1 -$

Statistics of Stochastic Processes, 3, 263-276

$\widehat{H}_2 - (H_1 - H_2)$. As before, this may be used as the basis of percentile-bootstrap confidence intervals and hypothesis tests for $H_1 - H_2$.

These techniques, being based on the percentile bootstrap, lack the pivotalness that bootstrap methods for confidence procedures should ideally enjoy. However, they have asymptotically correct levels, as N increases. Moreover, even when the statistic $\widehat{H}$ admits a Normal asymptotic distribution we lack a simple, computable variance estimator with which to correct for scale. And when the limiting distribution is Rosenblatt, rather than Normal, scale corrections are not sufficient to produce pivotalness, since the shape of the Rosenblatt distribution depends on the unknown Hurst constant through more than simply scale. For these reasons we argue that the percentile-t bootstrap, often suggested in simpler problems as a pivotal method for constructing confidence intervals and hypothesis tests with relatively accurate levels (Hall, 1992, p. 14f; Efron and Tibshirani, 1993, p. 158f; Shao and Tu, 1995, p. 94f; Davison and Hinkley, 1997, p. 29f), is not appropriate in the present setting.

Instead, level accuracy may be enhanced by using the double bootstrap (Hall, 1992, p. 20ff; Efron and Tibshirani, 1993, p. 263ff; Shao and Tu, 1995, p. 155ff; Davison and Hinkley, 1997, p. 103ff). However, the relatively high orders of accuracy associated with double-bootstrap confidence procedures in simpler problems cannot be expected to be generally available here, since our Gaussian model based on fractional Brownian motion does not necessarily reflect all second-order features of the distribution of the sampled stochastic process X. It seems difficult to improve on this situation without introducing relatively complex high-order models for X.

2.4. THEORETICAL PROPERTIES

Suppose the data $X_t, t \in \mathcal{I}$, are generated as $X_t = g(Y_{\epsilon t}, t)$, where

(a) g is a smooth bivariate function,
(b) Y is a Gaussian process whose sample paths have fractal dimension $D = 2-H$, and
(c) ϵ denotes a small positive constant.

The function g represents a possibly nonlinear transformation of Y, implying in particular that the observed process X is not necessarily Gaussian. Importantly, it allows a wide range of different types of inhomogeneity. By taking ϵ small we ensure that even if t_1 is moderately distant from t_2, X_{t_1} can be strongly correlated with X_{t_2}. This confers long-range dependence on the observed process. There is no difficulty in extending our results to the case where X is a function of a vector of Gaussian processes, say $X_t = g(Y_{\epsilon t}^{(1)}, \ldots, Y_{\epsilon t}^{(k)}, t)$. Here the Hurst index that prevails equals 2 minus the fractal dimension of sample paths of the process $Y^{(j)}$ that has the roughest sample paths. It is also possible to incorporate a smooth, monotone, nonlinear transformation of the time variable t. However, the simpler

Statistics of Stochastic Processes, 3, 263-276

setting prescribed by condition (a) conveys the important characteristics of these more complex models.

We claim that, under models of the type characterised by (a)–(c), $\widehat{H}$ is consistent for H and has an asymptotic distribution that is either Normal or of the type introduced by Rosenblatt (1961). To formulate this assertion as a mathematical theorem we first elaborate on (a)–(c), as follows. Assume that

(A) the derivatives

$$g_{j_1 j_2}(y, t) = (\partial/\partial y)^{j_1} (\partial/\partial t)^{j_2} g(y, t)$$

are bounded for each $j_1, j_2 \geqslant 0$, and g_{10} does not vanish;

(B) the Gaussian process Y satisfies $E(Y_t) \equiv 0$, and for constants $c > 0, \alpha = 2H \in (1/2, 2)$ and $\beta > \min(1/2, 2 - \alpha)$, $E(Y_{s+t} - Y_s)^2 = c|t|^\alpha + O(|t|^{\alpha+\beta})$, uniformly in $s \in \mathcal{J} = [0, 1]$, as $\to 0$; and

(C) $\epsilon = 1/N \to 0$,

and we define $\widehat{H}$ by regression of $\log(R/S)_n$ on $\log n$ for a fixed number, k, of values $\ell_1 m, \ldots, \ell_k m$ of n, where $\ell_1, \ldots, \ell_k$ are fixed and $m = m(\epsilon) \to \infty$ as $\epsilon \to 0$, in such a manner that $m^{-1} + m\epsilon = O(\epsilon^a)$ for some $a > 0$. Define $\zeta = m\epsilon$ and

$$t_\xi = \begin{cases} \xi^{2(1-H)} & \text{if } 3/4 < H < 1 \\ (\xi \log \xi^{-1})^{1/2} & \text{if } H = 3/4 \\ \xi^{1/2} & \text{if } 0 < H < 3/4, \end{cases}$$

which converges to 0 as $\epsilon \to 0$. Then, we claim that $\widehat{H} - H$ may be expressed as $t_\xi Z_\xi$, where Z_ξ has a proper limiting distribution as $\epsilon \to 0$. (The regularity conditions may be relaxed in many circumstances. For example, the restriction in (B) that $\alpha > (1/2)$ may be dropped if $g(y, t) \equiv y$, and also in some other cases. The boundedness condition on derivatives of g may also be relaxed).

Crucially, the limiting distribution of $\widehat{H}$ depends only on H and $\ell_1, \ldots, \ell_k$; it does not depend on g or on the scale constant, c, appearing in the first-order approximation to covariance. The main effects of scale and heteroscedasticity, entering through g and c, have cancelled due to rescaling by the terms S_a in (2.1). The limiting distribution is Normal when $0 < H \leqslant 3/4$, and a finite linear combination of correlated Rosenblatt distributions when $3/4 < H < 1$. Outline proofs of all these assertions are given in the appendix.

The results are foreshadowed by those of Hall and Wood (1993) for box-counting estimators, of which $\widehat{H}$ may be regarded as a scale-corrected version. We do not give the form of the limits, since it is complex (particularly in the Rosenblatt case), but it is of the type discussed by Hall and Wood (1993, p. 252). The relationships between statistical properties of a Gaussian process (e.g. Y), and of a smooth function of that process (e.g. X), have been addressed by Hall and Roy (1994).

The fact that the limiting distribution depends only on H and $l_1, \ldots, l_k$ justifies the bootstrap methods suggested in Section 2.2. Specifically, since the bootstrap

(2000) Hall, P., Härdle, W., Kleinow, T. and P. Schmidt
On Semiparametric Bootstrap Approach to Hypothesis Tests a. Confidence Intervals for Hurst Coefficients.

Statistics of Stochastic Processes, 3, 263-276

algorithm preserves the way in which H and $l_1, \ldots, l_k$ contribute to the limiting distribution, and since $\widehat{H} \to H$ at a rate that is polynomial in ζ (indeed, at rate t_ζ), then the bootstrap produces confidence intervals and hypothesis tests that have asymptotically correct coverage. The fractional Brownian motion ζ, used as the basis for our simulations, is just one of many that could have been employed, satisfying condition (B) above.

Note particularly that we keep k fixed as ϵ decreases. If our regularity conditions were to allow $k = k(\epsilon)$ to diverge then the Rosenblatt limit would change to Normal, but as discussed by Constantine and Wood (1994), this would generally be at the expense of increased mean squared error of $\widehat{H}$.

3. Application to Data

The aim of this section is to obtain an estimator $\widehat{H}$ of the Hurst coefficient H and to construct hypothesis tests and confidence intervals for H for the logarithm of the price process of certain German stocks.

Denote the logarithm of the price process of a stock (or index) by $\{X_t : 0 \leqslant t \leqslant T\}$. To estimate the Hurst coefficient H we applied R–S analysis, as described in Section 2.1, to N discrete observations $\{X_n : n = 1, \ldots, N\}$ of $\{X_t\}$ at times $t_1 \leqslant t_2 \leqslant \cdots \leqslant t_N$.

For the empirical study we used 6 900 observations ($N = 6\,900$) of 24 German blue chip stocks obtained form the Datastream/Primark's database from 8 January 1973 to 18 June 1999. The blue chips are included in the DAX, an index comprising 30 German blue chip stocks. We analysed Datastream performance indices of prices in order to avoid jumps in the respective time series due to dividend payments or rights issues. The obtained Hurst coefficients are shown in Table I. Figure 1 shows the R–S plot for the price process of the stock of Volkswagen. The R–S plot also includes a line with slope 0.5, which correspond to Brownian motion. As one can see, the R–S line has a different slope than it would have if the underlying process corresponded to a Brownian motion.

In the first step of our empirical analysis we tested whether the Hurst coefficient of an asset was significantly different from 0.5 or not. A significant difference from 0.5 would indicate that X_t did not follow a Brownian motion. In order to test the null hypothesis that $H = 0.5$, against the alternative $H \neq 0.5$, i.e.

$$h_0 \colon H = 0.5, \qquad h_1 \colon H \neq 0.5,$$

we approximated the distribution of $\widehat{H} - H$ conditional on the null hypothesis, and calculated the p-values, $P\{|\widehat{H} - E\widehat{H}| > |H_{\text{observed}} - E\widehat{H}| | h_0\}$, of the estimated $\widehat{H}$. For this approximation the bootstrap algorithm described in Section 2 was used. For $H = 0.5$ the fractional Brownian motion coincided with usual Brownian motion, which we simulated as a random walk. An estimate of the conditional density of $\widehat{H}^* - \widehat{H}$, computed from 400 simulated random walks of length 6,900, is shown

Statistics of Stochastic Processes, 3, 263-276

Table 1. Estimated Hurst coefficient of German stocks

Asset	$\widehat{H}$	p-value
Allianz	0.5642	0.6
BASF	0.5390	0.24
Bayer	0.5288	0.073
BMW	**0.5851**	**0.05**
Commerzbank	0.5536	0.88
Dt. Bank	0.5743	0.22
Daimler	**0.5859**	**0.05**
Degussa Hüls	0.5629	0.68
Dresdner Bank	0.5625	0.7
Hoechst	0.5420	0.37
Hypo Vereinsbank	0.5533	0.86
Karstadt	0.5552	0.95
Lufthansa	0.5584	0.89
Linde	0.5583	0.90
MAN	0.5605	0.79
Mannesmann	**0.5856**	**0.05**
Münchner Rück NA	0.5589	0.88
Preussag	**0.5884**	**0.035**
RWE	0.5398	0.29
Schering	0.5772	0.17
Siemens	**0.6007**	**0**
ThyssenKrupp	0.5794	0.13
Veba	0.5426	0.38
Volkswagen	**0.6049**	**0**

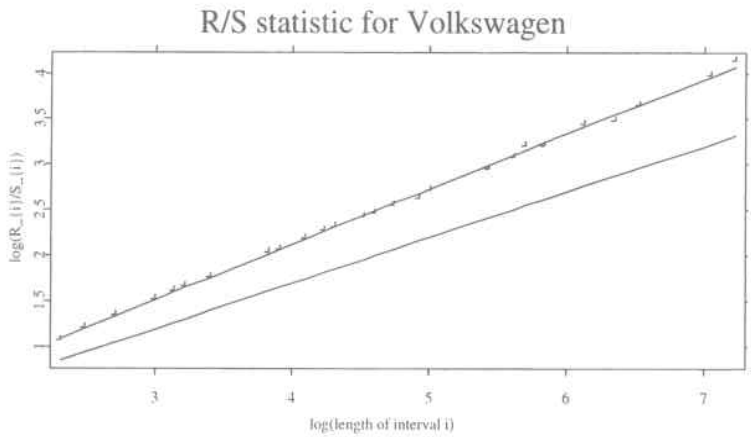


Figure 1. R–S plot for VW, $\widehat{H} = 0.606$.

(2000) Hall, P., Härdle, W., Kleinow, T. and P. Schmidt
On Semiparametric Bootstrap Approach to Hypothesis Tests a. Confidence Intervals for Hurst Coefficients.

Statistics of Stochastic Processes, 3, 263-276

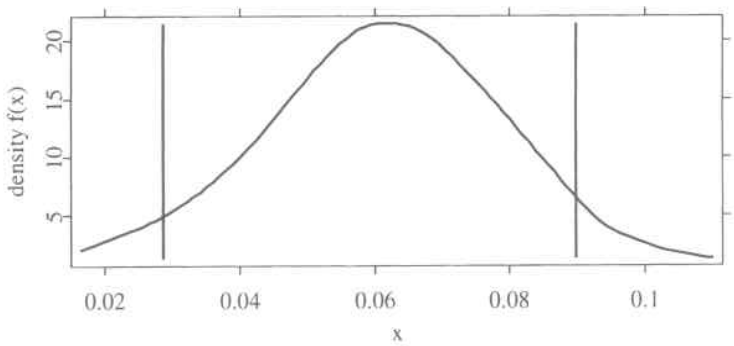


Figure 2. Estimated density of $\widehat{H} - H$ for 400 simulated Brownian motions with length 6 900. The vertical lines determine the 0.05, 0.95 quantiles.

in Figure 2. Table 1 shows the p-values for the estimated Hurst coefficient of the stocks.

Our analysis suggests that the difference between the estimated Hurst index of the prices of BMW, Daimler, Mannesmann, Preussag, Siemens and Volkswagen, and the value the Hurst index would take if the stochastic process describing prices were Brownian motion, is so great that it cannot be adequately explained by stochastic fluctuations.

We studied the assets for which the estimated Hurst coefficient H was significantly different from 0.5. For our further analysis we assumed that the logarithm of the price processes are self similar with stationary increments, i.e.

$$c^{-H}(X_{ct})_{t\in\mathbb{R}} =_d (X_t)_{t\in\mathbb{R}} \quad \text{for all} \quad c > 0 \tag{3.1}$$

and for any $k \geqslant 1$ and any time points $t_1, \ldots, t_k$,

$$(X(t_1), \ldots, X(t_k)) =_d (X(t_1 + c), \ldots, X(t_k + c)) \quad \text{for all} \quad c \in \mathbb{R}. \tag{3.2}$$

Here, $Y =_d Z$ means that Y and Z have the same distribution. These assumptions are often made in literature on financial market analysis. A well-known model is the multifractal model of asset returns (MMAR) introduced by Calvet et al. (1997). In this model the logarithms of prices are assumed to follow a fractional Brownian motion, i.e.

$$X(t) - X(0) = B_H(\theta(t)),$$

where $\theta(t)$ is a multifractal process with continuous, non-decreasing paths and stationary increments.

Under Assumptions 3.1 and 3.2 the autocorrelation function $\rho(k) = E[\{X(t) - EX(t)\}\{X(t+k) - EX(t+k)\}]$ of $X(t)$ is approximately of the form ck^{2H-2}. More precisely, the following holds (see Beran, 1994):

$$\frac{\rho(k)}{H(2H-1)k^{2H-2}} \to 1, \qquad 0 < H < 1, H \neq \frac{1}{2}, \quad k \to \infty.$$

(2000) Hall, P., Härdle, W., Kleinow, T. and P. Schmidt
On Semiparametric Bootstrap Approach to Hypothesis Tests a. Confidence Intervals for Hurst Coefficients.

Statistics of Stochastic Processes, 3, 263-276

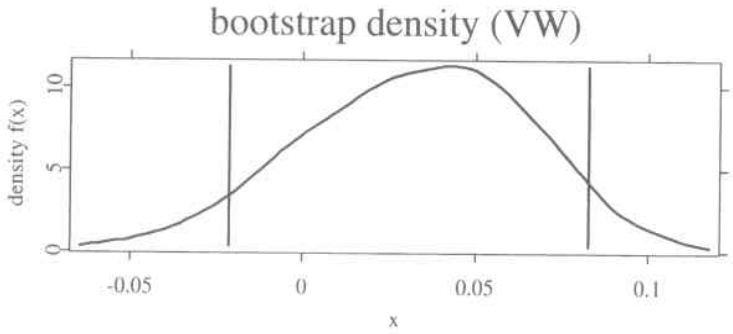


Figure 3. Bootstrap density of $\widehat{H} - H$ for the Volkswagen stock. The vertical lines determine the 0.05, 0.95 quantiles

This means that for $\widehat{H} > 0.5$, X_t has long memory. Stocks where long memory was detected are displayed in bold face in Table I.

The second step of our analysis was construction of confidence intervals. For this purpose we approximated the distribution of $\widehat{H} - H$ by that of $\widehat{H}^* - \widehat{H}$, where $\widehat{H}^*$ denotes the estimated value of the Hurst coefficient of simulated fractional Brownian motions with coefficient $\hat{\alpha} = 2\widehat{H}$. That is, we computed the conditional (on $X(t)$) distribution of the bootstrap form of $\widehat{H}^* - \widehat{H}$, as an approximation to the unconditional distribution of $\widehat{H} - H$. We applied the bootstrap method described in Section 2. To simulate fractional Brownian motion we used methods from Section 2.2 with $p = 1$ as well as the algorithm described in Beran (1994, p. 216). The latter is based on the finite Fourier transform of the autocovariance function of fractional Gaussian noise. Both methods lead to similar results.

The bootstrap densities for the different Hurst values of the assets which have significantly larger Hurst coefficient than a Brownian motion were approximately the same except for the mean value. For this reason we calculated only the density of $\widehat{H}^* - \widehat{H}$ of the Volkswagen stock. It is shown in Figure 3. The confidence intervals for the other assets were obtained by correcting this density for the different estimated Hurst coefficient. Table II shows the resulting confidence regions.

Table II. Confidence regions for Hurst coefficients

Asset	0.9 Confidence region	0.95 Confidence region
BMW	[0.475, 0.579]	[0.466, 0.594]
Daimler	[0.476, 0.581]	[0.467, 0.596]
Mannesmann	[0.476, 0.580]	[0.467, 0.596]
Preussag	[0.481, 0.585]	[0.472, 0.601]
Siemens	[0.506, 0.610]	[0.497, 0.626]
Volkswagen	[0.514, 0.619]	[0.505, 0.634]

(2000) Hall, P., Härdle, W., Kleinow, T. and P. Schmidt
On Semiparametric Bootstrap Approach to Hypothesis Tests a. Confidence Intervals for Hurst Coefficients.

Statistics of Stochastic Processes, 3, 263-276

Acknowledgements

The research of this paper was carried out within the Sonderforschungsbereich 373 at Humboldt University Berlin and was printed using funds made available by the Deutsche Forschungsgemeinschaft

Appendix: Technical Arguments

Put $Z_t = g(Y_t, t)$ and let $\mathcal{J} = [0, 1]$. Given $B > 0$, choose an integer B' so large that $B'\alpha > 2B$. Then, uniformly in $t_1, t_2 \in \mathcal{J}$,

$$Z_{t_1} - Z_{t_2} = \sum_{j_1=1}^{B'} \sum_{j_2=1}^{B'} \frac{1}{j_1! j_2!} (Y_{t_1} - Y_{t_2})^{j_1} (t_1 - t_2)^{j_2} g_{j_1 j_2}(Y_{t_2}, t_2) + O_p(|t_1 - t_2|^B). \tag{A.1}$$

This formula provides the opportunity to develop Taylor expansions of quantities such as R_a/S_a. It turns out that only the first term in such expansions contribute to asymptotic results. Nevertheless, higher-order Taylor-expansion terms should be included since, prior to correction for their means and analysis of their size, they are potential first-order contributors to limit theory for $(R/S)_n$. In our work the contributions of these high-order terms will be denoted by $Q_1, Q_2, \ldots$. For the sake of simplicity we ignore the mean correction in the definition of S_a.

Let $\mathcal{T} \subseteq \mathcal{J}$ denote a set of $n + 1$ equally-spaced points $t_0 < \cdots < t_n$ within an interval of width $\delta = n\epsilon$, and write $S_\mathcal{T}$ and $U_\mathcal{T}$ for the empirical standard errors of the 'samples' $\{Z_{t_i} - Z_{t_{i-1}}, 1 \leqslant i \leqslant n\}$ and $\{Y_{t_i} - Y_{t_{i-1}}, 1 \leqslant i \leqslant n\}$, respectively. Then by (A.1), for all $\eta > 0$,

$$S_\mathcal{T}^2 = g_{10}(Y_{t_2}, t_2)^2 U_\mathcal{T}^2 + Q_1 + O_p(\epsilon^{(\alpha/2)+B-\eta}), \tag{A.2}$$

$$\begin{aligned} R_\mathcal{T} &\equiv \max_{t \in \mathcal{T}} Z_t - \min_{t \in \mathcal{T}} Z_t \\ &= s|g_{10}(Y_{t_2}, t_2)|(Y_{T_\mathcal{T}} - Y_{T'_\mathcal{T}}) + Q_2 + O_p(\delta^B), \end{aligned} \tag{A.3}$$

where $T_\mathcal{T} = \operatorname{argmax}_{t \in \mathcal{T}} Z_t$, $T'_\mathcal{T} = \operatorname{argmin}_{t \in \mathcal{T}} Z_t$, and s denotes the sign of g_{10}. Hence, for all $\eta > 0$,

$$\frac{R_\mathcal{T}}{S_\mathcal{T}} = \frac{s}{U_\mathcal{T}} (Y_{T_\mathcal{T}} - Y_{T'_\mathcal{T}}) + Q_3 + O_p(\delta^{\alpha/2} \epsilon^{B-(\alpha/2)-\eta} + \delta^B \epsilon^{-(\alpha/2)-\eta}), \tag{A.4}$$

where Q_3 represents a series of ratios of terms, of the form $V/U_\mathcal{T}$, in Taylor expansions (in this sense, each summand is like the first term on the right-hand side of (A.4)), and the $O_p(\cdot)$ remainder is of the stated order uniformly in $\mathcal{T}$. Note particularly that in forming the leading ratio in (A.4) the contribution $g_{10}(Y_{t_2}, t_2)$ has cancelled from the leading terms in (A.2) and (A.3), and likewise the effect of the constant c (see condition (B) in Section 2.4) may be seen to cancel. This results from the scaling aspect of R–S analysis, and explains why the process ζ from

Statistics of Stochastic Processes, 3, 263-276

which we simulate when applying the bootstrap does not need to reflect either the properties of g or the value of c.

We deal with each ratio, V/W where $W = U_{\mathcal{T}}$, by expressing it as $w^{-1}(v + \Delta_V)(1 + \frac{1}{2}\Delta_W + \frac{3}{8}\Delta_W^2 + \cdots)$, where $\Delta_V = V - v$, $\Delta_W = -(W^2 - w^2)/w^2$, $v = E(V)$ and $w^2 = E(W^2)$. For purposes of exposition we shall confine attention to the three main terms in such an expansion, i.e. to $(v/w) + (\Delta_V/w) + \frac{1}{2}v(\Delta_W/w)$, in the case $V = Y_{T_{\mathcal{T}}} - Y_{T'_{\mathcal{T}}}$ and $W = U_{\mathcal{T}}$. (Without loss of generality, $s = 1$). Other terms may be treated similarly, although the argument is lengthy.

Let Δ_{Va}, Δ_{Wa}, v_a and w_a denote versions of Δ_V, Δ_W, v and w when $\mathcal{T} = \mathcal{I}_a$, the latter defined in Section 2.1. Note that, by condition (B), $w_a = w^0\{1 + O(\xi^\beta)\}$ uniformly in a, where w^0 does not depend on a or n. Since $\beta > \min(\frac{1}{2}, 2 - \alpha)$ (see condition (B)) then $\xi^\beta = o(t_\xi)$. Arguing thus it may be proved that A^{-1} times the sum over $1 \leqslant a \leqslant A$ of v_a/w_a equals $C\delta^{\alpha/2}(w^0)^{-1}\{1 + o(t_\xi)\}$, where $C > 0$ is a constant not depending on n.

Put $u = A^{-1}\delta^{-\alpha/2}w^0$, and let $S_\xi(n)$ equal u times the sum over $1 \leqslant a \leqslant A$ of the term Δ_{Va}/w_a. Methods of Hall and Wood (1993) may be used to show that $S_\xi(n)$ has variance asymtotic to a constant multiple of t_ξ^2, and that for the k values of n being considered, the variables $S_\xi(n)/t_\xi$ have a joint asymptotic distribution which is k-variate Normal when $0 < H \leqslant 3/4$, and k-variate Rosenblatt (Rosenblatt, 1961; Taqqu, 1975) when $3/4 < H < 1$.

By considering properties of the variogram estimator of fractal dimension, methods of Constantine and Hall (1994) may be employed to prove that u times the sum over a of $v_a\Delta_{Wa}/w_a$ equals $o_p(t_\xi)$. (Here it is critical that m diverge to infinity). If B is sufficiently larger then u times the sum over a of the $O_p(\cdot)$ remainder at (A.4) also equals $O_p(t_\xi)$, and similar methods may be applied to terms represented by Q_3 in the Taylor expansion. (The high-order contributions to bias of $\widehat{H}$ include terms of order ξ^α, but since we assumed $\alpha > (1/2)$ then this equals $o(t_\xi)$.) Arguing thus we may ultimately show that $(R/S)_n = C\delta^{\alpha/2}(w^0)^{-1}\{1 + S_\xi(n) + o_p(t_\xi)\}$. Hence, $\log(R/S)_n$ equals a quantity which does not depend on n and which goes into the intercept term in the regression, plus $(\alpha/2)\log n + S_\xi(n) + o_p(t_\xi)$. The result asserted in Section 2.3 follows from this property.

References

Adler, R. J.: 1981, *The Geometry of Random Fields*, Wiley, New York.

Barnsley, M.: 1988, *Fractals Everywhere*, Academic Press, New York.

Beran, J.: 1994, *Statistics for Long-Memory Processes*, Chapman and Hall, London.

Berry, M. V. and Hannay, J. H.: 1978, Topography of random surfaces, *Nature* **273**, 573.

Calvet, L., Fisher, A. and Mandelbrot, B. B.: 1997, A Multifractal Model of Asset Returns, Cowles Foundation Discussion Paper #1164

Carter, P. H., Cawley, R. and Mauldin, R. D.: 1988, Mathematics of dimension measurements of graphs of functions. in *Proc. Symp. Fractal Aspects of Materials, Disordered Systems*, D. A. Weitz, L. M. Sander and B. B. Mandelbrot (eds) Materials Research Society, Pittsburgh, PA, pp. 183–186.

(2000) Hall, P., Härdle, W., Kleinow, T. and P. Schmidt
On Semiparametric Bootstrap Approach to Hypothesis Tests a. Confidence Intervals for Hurst Coefficients.

Statistics of Stochastic Processes, 3, 263-276

Constantine, A. G. and Hall, P.: 1994, Characterizing surface smoothness via estimation of effective fractal dimension, *J. Roy. Statist. Soc. Ser. B* **56**, 97–113.

Davies, S. and Hall, P.: 1998, Fractal analysis of surface roughness using spatial data, *J. Roy. Statist. Soc. Ser. B*, to appear.

Davies, R. B. and Harte, D. S.: 1987, *Tests for Hurst effect*, *Biometrika* **74**, 95–101.

Davison, A. C. and Hinkley, D. V.: 1997, *Bootstrap Methods and their Application*, Cambridge: Cambridge University Press.

Efron, B. and Tibshirani, R.: 1993, *An Introduction to the Bootstrap*, Chapman and Hall, London.

Hall, P.: 1992, *The Bootstrap and Edgeworth Expansion*, Springer, New York.

Hall, P., Matthews, D. and Platen, E.: 1996, Algorithms for analyzing nonstationary time series with fractal noise, *J. Computat. Graph. Statist.* **5**, 351–364.

Hall, P. and Roy, R.: 1994, On the relationship between fractal dimension and fractal index for stationary stochastic processes, *Ann. Appl. Probab.* **4**, 241–253.

Hall, P. and Wood, A. T. A.: 1993, On the performance of box-counting estimators of fractal dimension, *Biometrika* **80**, 246–252.

Hunt, F.: 1990, Error analysis and convergence of capacity dimension algorithms, *SIAM J. Appl. Math.* **50**, 307–321.

Hurst, H. E. 1951, Long-term storage capacity of reservoirs, *Trans. Amer. Soc. Civil Engineers* **116**, 770–799.

Kent, J. T. and Wood, A. T. A.: 1993, Estimating the fractal dimension of a locally self-similar Gaussian process by using increments, *J. Roy. Statist. Soc. Ser. B* **59**, 679–699.

Mandelbrot, B. B., Passoja, D. E. and Paullay, A. J.: 1984, Fractal character of surfaces of metals, *Nature* **308**, 721–722.

Peters, E. E.: 1994, *Fractal Market Analysis: Applying Chaos Theory to Investment and Economics*, Wiley, New York.

Rosenblatt, M.: 1961, Independence and dependence, in *Proc. 4th Berkeley Symp. Math. Statist. Probab.*, J. Neyman (ed) University of California Press, Berkeley, pp. 411–433.

Sayles, R. S. and Thomas, T. R.: 1978, Surface topography as a nonstationary random process. *Nature* **271**, 431–434.

Shao, J. and Tu, D.: 1995, *The Jackknife and Bootstrap*, Springer, New York.

Sullivan, F. and Hunt, F.: 1988, How to estimate capacity dimension, *Nuclear Phys. B. (Proc. Suppl.)* **5A**, 125–128.

Taqqu, M. S.: 1975, Weak convergence to fractional Brownian motion and to the Rosenblatt process, *Z. Wahrsch. Verw. Gebiete* **31**, 287–302.

(2000) Hall, P., Härdle, W., Kleinow, T. and P. Schmidt
On Semiparametric Bootstrap Approach to Hypothesis Tests a. Confidence Intervals for Hurst Coefficients.