"Operations Research '92", ed. Karmann, A., Mosler, K., Schader, M., Uebe, G. Physica-Verlag Heidelberg (1993), 352-369.

Semiparametric Approaches to Dimension Reduction

Wolfgang K. Härdle
FB Wirtschaftswissenschaften
Humboldt Universität zu Berlin
D-1086 Berlin, Germany

Berwin A. Turlach
C.O.R.E. and Institut de Statistique
Université Catholique de Louvain
B-1348 Louvain-la-Neuve, Belgium

Abstract

We give an overview on several semiparametric methods utilised to model high-dimensional data. Our approach is semiparametric in nature and is related to Generalised Linear Models. We focus on dynamic estimation techniques in this setting. In particular we discuss Generalized Additive Models (GAM), Alternating Conditional Expectations (ACE), Average Derivative Estimation (ADE), semiparametric weighted least squares (Single Index Models, SIM), Projection Pursuit Regression (PPR), and Sliced Inverse Regression (SIR). Their performance in practice and theory is compared.

1. Introduction and Motivation

Due to the increasing availability of computer power and graphical tools over the last decades non parametric estimation methods became more and more popular for the analysis of the relationship between a response variable $Y \in I\!R$ and its explanatory variable $X \in I\!R^d$. For the case $d = 1$ a rich basket of tools exists such as kernel estimators, nearest-neighbour estimators and spline estimators. A good access to these topics is given by the recent monographs of Eubank (1988), Härdle (1990), Müller (1988), Wahba (1990), and Hastie and Tibshirani (1990, Chapter 2 and 3). However, most of these techniques are very unappealing for $d > 1$ since they are based on the idea of local (weighted) averaging *(smoothing)*. Since in higher dimensions the observations X are sparsely distributed this process of local averaging results in a poor performance for reasonable sample sizes. This behaviour, also known as the *curse of dimensionality* (Huber, 1985), makes it necessary to search for methods and models which reduce the dimension of the smoothing problem — preferably to a one-dimensional problem which is the best-studied case. In this paper we describe some lines of thought and research to this goal.

This idea of dimension reduction is an old one and is build in parametric models such as

Generalized Linear Models (GLM). A GLM as defined by McCullagh and Nelder (1989) connects the mean μ of the response variable Y with the *linear predictor* $\eta = X^T\beta$ via a *link function* G_1,i.e., $G_1(\mu) = \eta$. The aim is to estimate β when the link function G_1 is fixed. From the point of view of dimension reduction this is equivalent to say that for given G_1 the projection β from $\mathbb{R}^d$ to $\mathbb{R}^1$ is searched such that η, the projected X variable, fits the (by G_1) transformed mean μ of the Y variable "best". Most of the ideas which we present here can be seen as generalizations of GLMs. Such a model can be generalized by nonparametric methods in two ways. Either the fixed form of the link function is abandoned, i.e., a flexible or parameter free form is allowed, or the linear form of the predictor is abandoned allowing for any unknown function of the explanatory variables.

Relaxing the form of the link function means to keep the linear predictor but to replace the link function by a non-parametric (preferably monotone) function. More generally several of these types of response models can be added, each using a different linear predictor and (non-parametric) link function. These models are known as Projection Pursuit Regression (PPR) models due to an algorithm developped by Friedman and Stuetzle (1981). If we take just one term, i.e., an unknown (inverse) link function operating on a linear combination of the explanatory variables, this is called a one term projection pursuit model, in econometrics also called a *Single Index Model* (SIM). This type of model will be discussed in Section 3.

Allowing for any functional form of influence for the predictor variables leads again into the dimensionality problems mentioned above. In order to avoid these problems Hastie and Tibshirani (1990) proposed to keep the link G_1 but to generalize the linear predictor by a sum of non-parametric univariate functions. This leads to so called *Generalized Additive Models* (GAM) which will be discussed in Section 4.

However, both of the above approaches make a priori assumptions on the structure of the model. Most generally, the hope that interesting features of high-dimensional data are retrievable from low-dimensional projections is expressed by saying that the conditional distribution of Y given X depends on X only through a p-dimensional variable $(X^T\beta_1, \ldots, X^T\beta_p)^T$. The hope is that p is much smaller than d. Here the model is $Y = g(X^T\beta_1, \ldots, X^T\beta_p, \varepsilon)$, where the β's are unknown vectors, ε is independent of X and g is an arbitrary unknown function on $\mathbb{R}^{p+1}$. In this model we of course have an identification problem. Without any assumptions on g it is impossible to identify the β's, we can only hope to identify the linear space which they span. Li (1991a) proposed a method called *Sliced Inverse Regression* (SIR) for estimating this linear space and some β's forming a base of this space. In his terminology which we will adopt here this space is called the effective dimension-reduction (e.d.r.) space and each vector of this space is called an e.d.r. direction. Once a set of e.d.r. directions is fixed Li (1991a) proposes to use standard smoothing techniques to smooth $(X^T\beta_1, \ldots, X^T\beta_p)^T$ against Y. We will discuss his approach in Section 5.

In Section 2 we will describe running examples which we will use to illustrate the different methods which we discuss. Section 6 finally shows how these methods can be implemented on a computer.

2. Nonparametric Approaches to Generalized Linear Models

We have argued that some dimension-reduction models can be seen as generalizations of GLMs were the assumptions on the link function are weakened or the linear form of the explanatory variables is abandoned. To fix ideas let $X \in I\!R^d$ denote the explanatory variable and $Y \in I\!R$ be the response variable. Here and in the following we use the term *link* where McCullagh and Nelder (1989) mean the inverse link,i.e., the mean μ of Y is connected with the predictor $\eta = X^T\beta$ via a link function G such that $\mu = G(\eta)$. Since in our examples this link is monotone there is no problem of confusion. As a running example we shall use the case of binary response models, i.e., $Y \in \{0,1\}$. The GLM then reads as $P[Y = 1|X = x] = G(x^T\beta)$.

Single Index Models

Single Index Models keep the linear component but generalize the link. In our running example this reads as

$$P[Y = 1|X = x] = g\left(x^T\beta\right) \tag{2.1}$$

with g an unknown univariate "smooth" function. Note that here some standardization of the parameter β is asked for, since as such, (2.1) does not identify β but rather the direction of β. The aim here is to estimate β and the unknown link. For illustration of statistical and numerical procedures to be described later we would like to introduce

example 1:

$$\begin{aligned} &X \sim \mathcal{N}_2(0, I_2), \quad \beta = (1,1)^T \\ &g(\eta) = L(\eta) + \rho\varphi'(\eta), \quad L(\eta) = \exp(\eta)/[1+\exp(\eta)], \quad \varphi(\eta) = \frac{1}{\sqrt{2\pi}}\exp\left(-\frac{\eta^2}{2}\right) \\ &P[Y = 1|X = x] = g(x^T\beta) \end{aligned} \tag{2.2}$$

This model is almost a Logit model, only the skew deviation term $\rho\varphi'(\eta)$ makes it different from a GLM. For $\rho = 0$ it falls into the class of GLMs. For later illustrations we have set $\rho = 0.6$ and have generated $n = 200$ datapoints (x_i, y_i) according to (2.2). A graphical inspection of the

data gives a taste of the nonparametric structure. Figure 1 shows a three dimensional scatterplot of the data. If we project the X variables in the 45^o line we obtain Figure 2. This picture shows the projected data $x_i^T\beta$ against y_i together with the link $g(\eta)$. All the graphics and computations were done in the computing environment XploRe (1992).

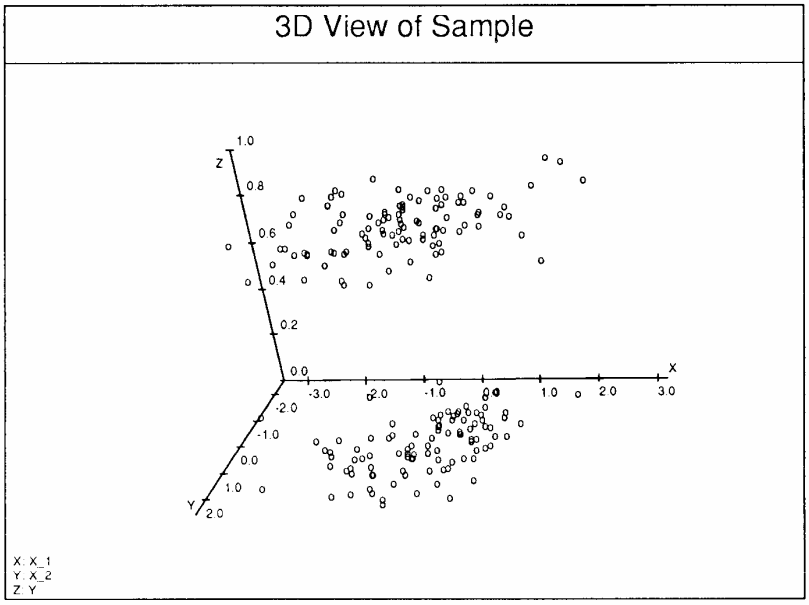


Figure 1: A three-dimensional scatterplot of the sample $\{(x_{1i}, x_{2i}, y_i)\}_{i=1}^{200}$ for example 1.

Generalized Additive Models

Generalized Additive Models keep the link but generalize the linear predictor to a sum of nonparametric functions. In our running example this reads as

$$P[Y = 1|X = x] = G\left(\alpha + \sum_{j=1}^{d} g_j(x_j)\right) \tag{2.3}$$

where x_j denotes the j^{th} component of the vector $x = (x_1, \ldots, x_d)^T$ and the g_j are unknown univariate "smooth" functions. Again some standardization is necessary since the model as such does not identify the unknown $\{g_j\}_{j=1}^d$. The aim here is to estimate the nonparametric functions g_j. For illustrations of later techniques let us introduce

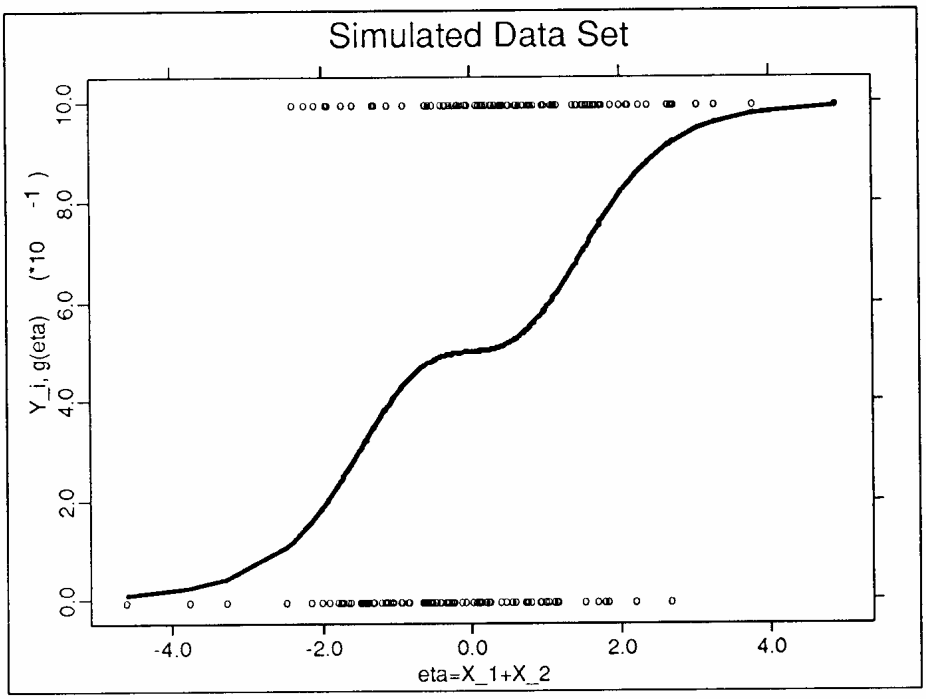


Figure 2: The observations y_i for example 1 plotted against $\eta = x^T\beta = x_{1i} + x_{2i}$. The link $g(\eta) = L(\eta) + 0.6\varphi'(\eta)$ is shown as the solid line.

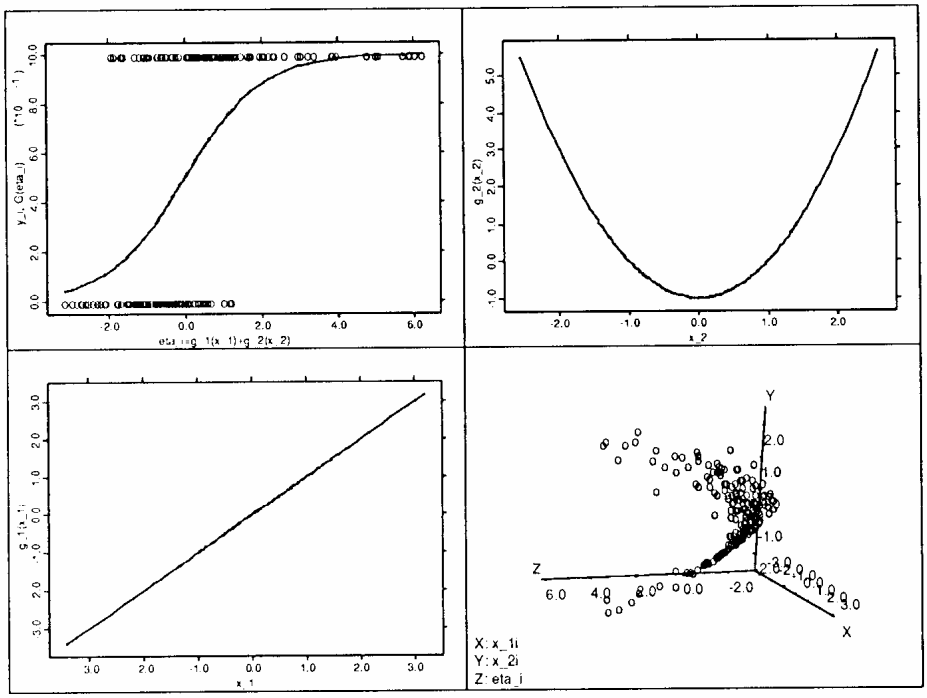


Figure 3: A four picture display with $\{(\eta_i, y_i)\}_{i=1}^{200}$ and $G(\eta_i)$ in the upper left. The nonparametric components are in the lower left and upper right. A rotated view of the surface $\{(x_{1i}, x_{2i}, \eta_i)\}_{i=1}^{200}$ is given in the lower right.

example 2:

$$\begin{aligned} &X \sim \mathcal{N}_2(0, I_2), \quad g_1(x_1) = x_1, \quad g_2(x_2) = (x_2)^2 - 1 \\ &G = L \\ &P[Y = 1|X = x] = G(g_1(x_1) + g_2(x_2)) \end{aligned} \tag{2.4}$$

This model is almost a Logit model, only the second predictor variable has a nonlinear influence on η. Figure 3 shows a four picture display with the data $\{(\eta_i, y_i)\}_{i=1}^{200}$ in the upper left corner together with the Logistic link. Note that the predictor is $\eta_i = g_1(x_{1i}) + g_2(x_{2i})$. The "nonparametric" components g_j are shown in the lower left und the upper right. An impression of the nonlinear components can be gained by rotating the three dimensional surface $\{(x_{1i}, x_{2i}, \eta_i)\}_{i=1}^{200}$. The rotated point cloud is shown in the lower right.

Sliced Inverse Regression

In our running example Sliced Inverse Regression estimates the effective dimension reduction directions $\beta_1, \ldots, \beta_p$ of the model

$$P[Y = 1|X = x] = g(x^T\beta_1, \ldots, x^T\beta_p). \tag{2.5}$$

where g is an arbitrary unknown function on $\mathbb{R}^p$. Note that SIR contains the above two models as special cases. To demonstrate this technique we will use a modified version of example 1 and example 3 defined below. Example 1 is modified by adding three additional components to X so that $X \sim \mathcal{N}_5(0, I_5)$. But these additional components will have no influence on Y.

example 3:

$$\begin{aligned} &X \sim \mathcal{N}_5(0, I_5), \quad g_1(x_1) = x_1, \quad g_2(x_2) = (x_2)^2 - 1 \\ &\varepsilon \sim \mathcal{N}_1(0, (0.6)^2) \\ &Y = g_1(x_1) + g_2(x_2) + \varepsilon \end{aligned} \tag{2.5}$$

3. Single Index Models

Model (2.1) is called a single index model or a one term projection pursuit model. This terminology is due to Friedman and Stuetzle (1981) who considered the more general model:

$$P[Y = 1|X = x] = \sum_{j=1}^{K} g_j(x^T\beta_j)$$

where the $\beta_j \in I\!R^d$ are unknown parameters and the g_j's are unknown smooth functions. In order to make the β_j's and the g_j's identifiable one has to impose restrictions on the scale, usually $\|\beta_j\| = 1$, or $\beta_{j1} = 1$.

Friedman and Stuetzle (1981) proposed to estimate K, β_j and g_j by the method of Projection Pursuit Regression (PPR) algorithm. This procedure estimates terms $g_j(x^T\beta_j)$ as long as the fraction of unexplained variance is below a user specified threshold. In each step that β_j is chosen which maximizes the fraction of unexplained variance given the previous terms (*projection pursuit*). The fitted model with $\hat{K}$ is

$$P[Y = 1|X = x] = \sum_{j=1}^{\hat{K}} \hat{g}_j(x^T\hat{\beta}_j).$$

A drawback of this method is that it is not evident which value of K is to be chosen. Research has therefore focused on one term projection pursuit models. In this line Hall (1989) constructs a root-n consistent estimator of β. A different method is that of Härdle and Stoker (1989) also called ADE for Average Derivative Estimation. It is based on the following idea. Define $m(x) = g(x^T\beta)$ and observe that for the average derivative δ, as defined below, we have

$$\delta = E_X[m'(X)] = E_X[\frac{dg}{d(x^T\beta)}(X^T\beta)]\beta. \tag{3.1}$$

Thus δ determines β up to scale. Let $f(x)$ denote the density of X and l its vector of the negative log-derivatives (partial), $l = -\frac{\partial \log f}{\partial x} = -\frac{f'}{f}$ (l is also called *score vector*). Under assumptions on f this enables us to write

$$\delta = E[m'(X)] = E[lY] \tag{3.2}$$

and to estimate δ by $\hat{\delta} = n^{-1}\sum_{i=1}^{n} \hat{l}_h(x_i)y_i$. Here $\hat{l}_h$ is an estimator of l based on a kernel density smoother with bandwidth h. For an easy access to kernel density smoothing see the book by Silverman (1986). With root-n estimates for δ precise estimates for the link can be obtained. The convergence rate for g is one dimensional. Stoker (1991) proposed alternative estimators for δ based on first estimating the partial derivatives $m'(x)$ and then to average over the observations. A Monte Carlo comparison of these methods is presented in Stoker and Villas-Boas (1992).

The estimation of the score vector l via a kernel density estimator involves a number of intensive calculations, especially when we optimize over h. Therefore discretization or WARPing ideas should be used (Turlach 1992). For our simulated example Figure 4 shows the result of this method. We calculated $\hat{\delta}$ and used the *Nadaraya-Watson* regression estimator to estimate $\hat{g}$. Note that the horizontal scale on this figure is different since (3.1) suggest that δ has different scale than β. In fact for ADE the scale of δ changes with g but it does not matter for the statistical interpretation of the link g that we are interested in.

The estimation of δ and its asymptotic covariance matrix $\hat{\Sigma}_\delta$ for example 1 was done with Program 1 in Section 6. Note that for this example we have $\delta = \binom{0.135}{0.135}$. The binning parameter d was chosen in such a way that maximal 20 bins were used in each coordinate, i.e., $d \approx \binom{0.33}{0.256}$. The estimate for the average derivative and the asymptotic covariance matrix was calculated using the three adjacent bins which equals a bandwidth $h \approx \binom{0.99}{0.78}$. As result we have

$$\hat{\delta} = \begin{pmatrix} 0.124 \\ 0.118 \end{pmatrix}, \qquad \hat{\Sigma}_\delta = \begin{pmatrix} 0.188 & 0.036 \\ 0.036 & 0.206 \end{pmatrix}.$$

These results allow us to test some hypothesis formally using a Wald statistic (see Stoker (1992), pp. 53–54). In particular, to test the restriction $R\delta = r_0$, the Wald statistic

$$W = n(R\hat{\delta} - r_0)^T (R\hat{\Sigma}_\delta R^T)^{-1} (R\hat{\delta} - r_0)$$

is compared to a $\chi^2(\text{rank } R)$ critical value. Table 3.1 gives some examples for this technique.

Restriction	Value W	d.f.	$P[\chi^2(\text{d.f.}) > W]$
$\delta^1 = \delta^2 = 0$	25.25	2	0
$\delta^1 = \delta^2 = 0.135$	0.365	2	0.83
$\delta^1 = \delta^2$	0.027	1	0.869

Table 3.1: Wald Statistics for some restrictions on δ.

The results in Turlach (1992) show that the results of such tests depend on the chosen bandwidth h, i.e., in practice it remains the problem of selecting h. The theoretical optimal bandwidth is calculated in Härdle, Hart, Marron, and Tsybakov (1992). They found that for an optimal estimation of δ the bandwidth h should be of the rate $n^{-\alpha}$ (α depending on the dimension d and the "smoothness" of f and g) where this rate is typically different from the optimal rate for estimating f or g for example. Thus one has to work with two different bandwidths for estimating δ and g.

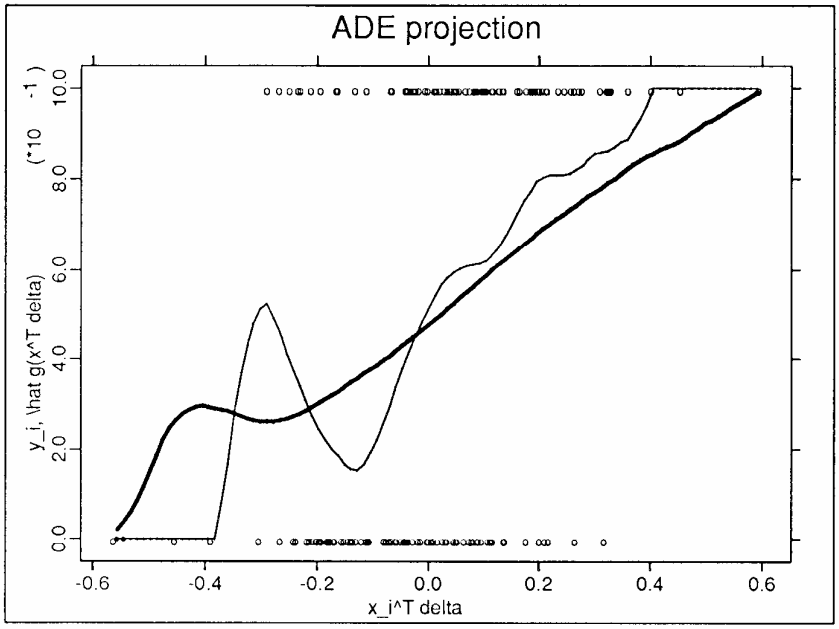


Figure 4: For the simulated data set of example 1 $x_i^T\hat{\delta}$ vs. y_i and two estimates of $\hat{g}(x_i^T\hat{\delta})$ are shown. The thick line shows the Nadaraya-Watson regression estimator for $\hat{g}$ with a bandwidth of $h = 0.3$, for the thin line $h = 0.1$ was chosen.

This unappealing feature is avoided by an approach due to Ichimura (1993). Let ε denote the error term inherent to the response variable. Observing that (β_0 denotes the true parameter):

(1) The variation in Y results from both the variation in $X^T\beta_0$ and the variation in ε.

(2) On the contour line $X^T\beta_0 = c$, where c is a given constant, the variability in Y results only from the variation in ε.

(3) Observation (2) does not necessarily hold on a contour line defined by $X^T\beta = c$ for $\beta \neq \beta_0$. Along this contour line, the value of $X^T\beta_0$ changes and therefore the variability in Y again results from the variation in both $X^T\beta_0$ and ε.

To identify β_0 Ichimura (1993) thus proposes to estimate

$$S(\beta) = E[\{Y - g(X^T\beta)\}^2] \tag{3.3}$$

since β_0 is the minimizer of (3.3). Using

$$\hat{S}(\beta, h) = n^{-1}\sum_{i=1}^{n}[y_i - \hat{g}_{-i,h}(x_i^T\beta)]^2,$$

where $\hat{g}_{-i,h}$ denotes a leave-one-out kernel smoother of Y on $X^T\beta$, as estimator for $S(\beta)$ and minimizing $\hat{S}(\beta, h)$ with respect to β and h Härdle, Hall and Ichimura (1992) showed that this

yields a root-n consistent estimator of β_0 and an asymptotically optimal estimator of h_0, the bandwidth which should be used to calculate the kernel estimate of g.

A way of testing a GLM against this specific single index alternative has been given by Horowitz and Härdle (1992). They constructed a conditional moment test based on ideas of Bierens (1990) and Newey (1985). Another approach for such a test via Bootstrapping ideas was investigated by Rodríguez-Campos and Cao-Abad (1993) and Proença (1992).

4. Generalized Additive Models

For the Generalized Additive Model we have to estimate α and functions g_j in the model

$$P[Y = 1|X = x] = G\left(\alpha + \sum_{j=1}^{d} g_j(x_j)\right).$$

This estimation is a highly iterative procedure. Estimation of α and $g_1, \ldots, g_d$ in the above model is accomplished by an algorithm for fitting a weighted additive model (Hastie and Tibshirani, 1990). This iterative fitting of a weighted additive model is known as *local scoring* since it generalizes the Fisher scoring procedure. Each estimation of a weighted additive model is done in an iterative process known as *backfitting*. In the backfitting step non-parametric estimates for $g_1, \ldots, g_d$ are calculated. The explicit algorithm of finding the nonparametric components is given by (see Hastie and Tibshirani 1987):

Local Scoring Algorithm

Initialization $\hat{g}_j^{(0)} \equiv 0$ for $j = 1, \ldots, d$, $\hat{\alpha}^{(0)} = \text{logit}(\bar{y})$.

Loop over outer iteration counter m

$\hat{\eta}^{(m)}(x_i) = \hat{\alpha}^{(m)} + \sum_{j=1}^{d} \hat{g}_j^{(m)}(x_{ji})$

$\hat{p}_i = \text{logit}^{-1}(\hat{\eta}^{(m)}(x_i))$

$z_i = \hat{\eta}^{(m)}(x_i) + (y_i - \hat{p}_i)/[\hat{p}_i(1 - \hat{p}_i)]$

$w_i = \hat{p}_i(1 - \hat{p}_i), \qquad i = 1, \ldots, n.$

Obtain $\hat{\alpha}^{(m+1)}, \hat{g}_j^{(m+1)}, j = 1, \ldots, d$ by applying the backfitting algorithm to z_i with explanatory variables x_i and observation weights w_i.

until the deviance $D(y, \hat{p}) = -2\sum_i [y_i \log(\hat{p}_i) + (1 - y_i)\log(1 - \hat{p}_i)]$ converges.

Backfitting Algorithm

Initialization $\hat{g}_j^{(0)} \equiv 0$ for $j = 1, \ldots, d$, $\hat{\alpha}^{(0)} = \bar{y}$.

Repeat for $j = 1, \ldots, d$ repeat such cycles:

$$r_i = y_i - \hat{\alpha} - \sum_{\substack{k=1 \\ k \neq j}}^{d} \hat{g}_k(x_{ki}) \quad i = 1, \ldots, n$$

$$\hat{g}_j(x_{ji}) = S(r|w, x_{ji}) \quad i = 1, \ldots, n$$

Until $RSS = \sum_{i=1}^{n} \left(y_i - \hat{\alpha} - \sum_{j=1}^{d} \hat{g}_j(x_{ji}) \right)^2$ converges.

Here $S(r|w, x_{ji})$ denotes the value of the function obtained by smoothing the scatterplot (x_j, r) with weights w at the point x_{ji}.

Since non-parametric estimation methods are used in the backfitting step some typical questions arise. We have again the question of how to choose the smoothing parameter in this non-parametric fit regardless whether splines, kernel estimators or others are used, see Buja, Hastie and Tibshirani (1989). Another question is how to incorporate the weights in the non-parametric smoothing step (see Hastie and Tibshirani 1990, pp. 72–74). Especially in binary models, as we discuss them here, these weights can cause numeric problems. If the estimated probability $\hat{p}_i = P[Y_i = 1|X = x_i]$ is very close to 0 or 1 the weight w_i for this observation in the backfitting step will be very small. Thus the adjusted dependent variable z_i may be very big resulting in a large partial residual within the backfitting algorithm. This can result in a bad fit which leads in the next step of the local scoring to the same problem.

Program 3 in Section 6 demonstrates how the Generalized Additive Model can be estimated in XploRe (1992). The result of this fitting is visualized in Figure 5. The backfitting algorithm provides estimates of the function g_j in the multiple additive regression model $E[y|x] = \alpha + \sum_j g_j(x_j)$ with $E[g_j(X_j)] = 0$ for $j = 1, \ldots, d$. It is easily seen that in example 2 given by (2.4) we have $E[g_j(X_j)] = 0$, $j = 1, 2$. Thus for our example we would expect that α is estimated as 0. In fact the result is $\hat{\alpha} = 0.25$.

GAMs can be further generalized by relaxing the assumptions on the link function G. A general model would be in this case

$$\Psi(Y) = \sum_{i=1}^{d} g_i(X_i) + \varepsilon \tag{4.1}$$

where $\Psi, g_1, \ldots, g_d$ are arbitrary univariate functions and ε is independent from $X = (X_1, \ldots, X_d)^T$. For estimating model (4.1) Breiman and Friedman (1985) proposed the the method of Alternating Conditional Expectations (ACE). Here $\Psi, g_1, \ldots, g_p$ are estimated by the minimizers of the fraction of variance not explained by a regression of $\Psi(Y)$ on $\sum_{i=1}^{d} g_i(X_i)$, i.e. the minimizers of

$$\frac{E\left[\{\Psi(Y) - \sum_{i=1}^{d} g_i(X_i)\}^2\right]}{E\left[\Psi(Y)^2\right]}.$$

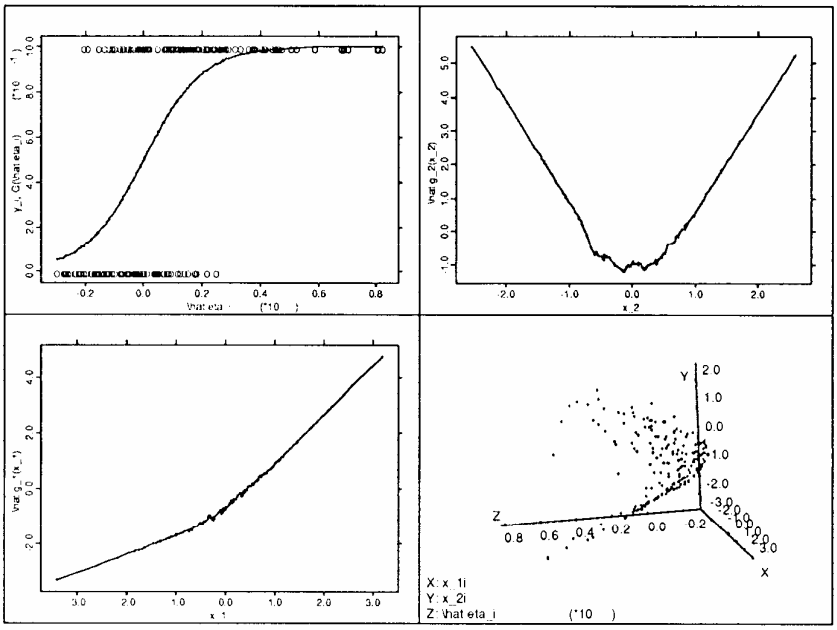


Figure 5: A four picture display with the results of the fitting procedure for the Generalized Additive Model. Legend is the same as for Figure 3 where η_i is replaced by $\hat{\eta}_i$.

Since

$$g_j(X_j) = E\left[\Psi(Y) - \sum_{i \neq j} g_i(X_i) \middle| X_j\right] \quad \text{and} \quad \Psi(Y) = E\left[\sum_{i=1}^{d} g_i(X_i) \middle| Y\right]$$

this can be done by iteratively estimating each of the above conditional expectation, using at each step the current estimates of the functions on the right hand side, until convergence is reached, i.e., estimating alternatively conditional expectations.

These estimates are chosen to optimize a correlation criterion. There are a number of properties of the ACE procedure that are somewhat misleading if one views ACE as a regression tool (see comments on Breiman and Friedman, 1985 and Hastie and Tibshirani, 1990, Chapter 7.2.6). To overcome these anomalies Tibshirani (1988) proposed a modification called AVAS (Additivity and Variance Stabilization). It differs from ACE by using an *asymptotic variance stabilizing transformation* instead of the estimate $\Psi(Y) = E\left[\sum_{i=1}^{d} g_i(X_i) \middle| Y\right]$. To our knowledge there is not much theoretical support for this technique until know and it is thus sill an open field of research. Especially global convergence of AVAS has not yet been established. There is also no consent which methods of inference should be applied for ACE and AVAS.

5. Sliced Inverse Regression

Sliced Inverse Regression (Li, 1991a) attempts to estimate the effective dimension-reduction (e.d.r.) directions $\beta_1, \ldots, \beta_p$ in the model

$$Y = g(X^T\beta_1, \ldots, X^T\beta_p, \varepsilon) \tag{5.1}$$

where the β's are unknown vectors, ε is independent of X and g is an arbitrary unknown function on $\mathbb{R}^{p+1}$. In our running example (5.1) reads as

$$P[Y = 1|X = x] = g(x^T\beta_1, \ldots, x^T\beta_p) \tag{5.2}$$

with g an arbitrary function on $\mathbb{R}^p$.

The idea to estimate the β's is to use inverse regression, i.e., to estimate $E[X|Y]$. Before we present the justification for this approach we have to introduce some further notation. Let Σ_X denote the covariance matrix of X and $Z = \Sigma_X^{-1/2}(X - E[X])$ the standardized version of X. With $\eta_i = \Sigma_X^{1/2}\beta_i$, $i = 1, \ldots, p$, Li (1991b) calls any vector in the linear space generated by the η's a standardized e.d.r. direction.

Observe now that the centered inverse regression curve $E[X|Y] - E[X]$, in general, describes a curve in $\mathbb{R}^d$. Li (1991a) showed that if for any $b \in \mathbb{R}^d$ we have that $E[b^T X|X^T\beta_1, \ldots, X^T\beta_p]$ is linear in $X^T\beta_1, \ldots, X^T\beta_p$ and (5.1) holds, the centered inverse regression curve is contained in the linear subspace spanned by $\Sigma_X\beta_i$, $i = 1, \ldots, p$. As a corrollary we have that the standardized inverse regression curve $E[Z|Y]$ is contained in the linear space generated by the standardized e.d.r. directions $\eta_1, \ldots, \eta_p$. Thus the covariance matrix $Cov[E[Z|Y]]$ is degenerate in any direction orthogonal to the η's. Therefore the eigenvectors η_i, $i = 1, \ldots, p$, associated with the largest p eigenvalues of $Cov[E[Z|Y]]$ are the standardized e.d.r. directions. So we have the following algorithm to estimate the e.d.r. directions

The Sliced Inverse Regression algorithm

I. Divide the range of Y into H slices

II. Within each slice compute the sample mean of X, denoted by x_h, $h = 1, \ldots, H$

III. Compute the sample covariance matrix for X, $\hat{\Sigma}_X = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})(x_i - \bar{x})^T$ and the weighted covariance matrix for the slice means, $\hat{\Sigma}_\eta = \sum_{i=1}^{H}\hat{p}_h(\bar{x}_h - \bar{x})(\bar{x}_h - \bar{x})^T$ where $\hat{p}_h$ is the proportion of cases that fall into the slice h, and $\bar{x}$ is the sample average of X.

IV. Conduct an eigenvalue decomposition of $\hat{\Sigma}_\eta$ with respect to $\hat{\Sigma}_X$. Order the eigenvectors $\beta_1, \ldots, \beta_d$ according to the descending order of the corresponding eigenvalues, $\hat{\lambda}_1 \geq \ldots \geq \hat{\lambda}_p$

Steps I and II produce a crude estimate of the inverse regression curve. Li (1991b) advocates the method of slicing due to its simplicity although more sophisticated nonparametric regression methods could be used. We want to study the performance of this algorithm on example 1 and example 3 introduced in Section 2. For example 1 we changed the explanatory variable X to have a $\mathcal{N}_5(0, I_5)$ distribution keeping the relationship between the first two components of X and Y as described in Section 2. Example 2 is not suitable for this method since it has two e.d.r. direction (the same as example 3). For estimating p e.d.f. directions the number H of slices has to be greater than p (see discussion of Li, 1991a). But in a binary response model one has in fact always $H = 2$. To our knowledge it is an open question whether this theory can be extended to binary (discrete) response models. The program for these calculations is listed in Section 6 and the results are given below for each example.

Example 1

In this example we have only one e.d.r. direction which is $\beta_1 = (1, 1, 0, 0, 0)^T$. Li (199a) gives (under additional assumptions) a result which allows to test for the significance of an e.d.r. direction β_i by the mean of the eigenvalues smaller than the eigenvalue associated with β_i. The eigenvalues estimated for this example are 0.152, 0, 0, 0, and 0. Thus we would interfere that there is only one e.d.r. direction. The estimate for this direction is

$$\hat{\beta}_1 = (0.7331, 0.6726, -0.0752, -0.0152, 0.0654)^T.$$

Note that as with the SIMs we can only estimate the e.d.r. direction, here $\hat{\beta}_1$ is standardized to have Euclidean norm 1. In this example the e.d.r. direction β_1 is recognized. If we change $\hat{\beta}_1$ to the same scale as δ, the Average Derivative Estimate, and do a regression of Y against $X^T\hat{\beta}_1$ we get a picture similar to Figure 4.

Example 3

In this example we have two e.d.r. directions, $\beta_1 = (1, 0, 0, 0, 0)^T$ and $\beta_2 = (0, 1, 0, 0, 0)^T$. However, the eigenvalues which are estimated in this examples are 0.4037, 0.0392, 0.0199, 0.0122, and 0.0076. Thus we would interfere again the existence of only one e.d.r. direction!. Even if we ask for estimates of two directions the result is

$$\hat{\beta}_1 = (0.9680, 0.1574, -0.0563, -0.1158, -0.1474)^T$$

and

$$\hat{\beta}_2 = (0.1810, 0.0903, 0.3044, 0.1951, 0.9101)^T.$$

Thus the first e.d.f. direction is fairly well identified whereas the second one is not at all identified. The problem is that in this example $E[X_2|Y] = 0$. Therefore the inverse regression curve is degenerated in the second e.d.r. direction and SIR fails to find it. A solution to this problem is to estimate higher orders too. If for example the conditional variance $Var[X_2|Y]$ varies with Y we could hope to identify this e.d.r. direction by exploring the variability of the conditional variance $Var[b^T X|Y]$. This approach is called SIR II by Li (1991b) and is a very promising development of research.

6. The implementation in XploRe

The above calculations have been performed in the language XploRe (1992). In this section we give some programs that are useful in solving the iterative procedure for Generalized Additive Models for example or for ADE. The Single Index Model for example 1 has been estimated using the ADE technique with the following program.

```
library(smoother)                              ;load the necessary libraries
library(addmod)
randomize(0)
x = normal(200 2)                              ;generate the explanatory variable
rho = 0.6
beta = #(1 1)
eta = x*beta                                   ;eta, notation as in (2.2)
g = 1./(1+exp(-eta)) - rho * eta.*pdfn(eta);calculate g(eta)
u = uniform (200)
y = u.<g                                       ;generate the response variable
d = (max(x)-min(x))/20                         ;choosing a binning parameter
(xb yb) = bindata(x d 0 y)                     ;binning the data
(del dvar) = adeind(xb yb d 3)
    ;estimate the average derivative and the asymptotic covariance matrix
est = (x*del)~y                                ;calculate the projection
gh1 = regest(est 0.1)                          ;find estimates for g
gh2 = regest(est 0.3)
show(est gh1 gh2 s2d)                          ;show results (Picture 4)
```

Program 1: This program generates and estimates example 1

The commands of XploRe (1992) are similar to GAUSS but more fine tuned for smoothing and nonparametric methods in high dimensions. The Generalized Additive Model (GAM) of example 2 was created using the following code:

```
randomize(0)
x = normal(200 2)
g1 = x[,1]
g2 = x[,2].*x[,2]-1
eta = g1+g2
px = 1./(1+exp(-px))
u = uniform(200)
y = u.<px
createdisplay(pic3, 2 2, s2d s2d s2d d3d)
show(eta~y eta~px s2d1, x[,1]~g1 s2d2, x[,2]~g2 s2d3, x~eta d3d1)
```

Program 2: This program generates Picture 3

The estimation of the GAM was done by

```
proc(fx alpha devs)=lscore(x y)
  dim = cols(x)
  gx = matrix(rows(x) dim 0)              ;initialize g_j
  xs = 1                                  ;used to store information
                                          ;to sort the covariates
  ybar = mean(y)
  alpha = ln(ybar/(1-ybar))               ;initialize alpha
  devs = 0
  loop = 1
  do
    eta = alpha + sumr(gx)
    p = 1./(1+exp(-eta))
    w = p.*(1-p)                          ;calculate the weights
    z = eta + (y-p)./w                    ;calculate the adjusted
                                          ;dependent variable
    (gx alpha xs)=backfit(x z xs w 0.4)   ;the backfitting step
    dev = -2*sum(y.*ln(p)+(1-y).*ln(1-p));calculate the deviance
    devs = defs|dev
    chg = abs(devs[loop,1]-dev)/dev
    loop = loop+1
  until( (chg < 0.001) || (loop == 6) )
  devs = devs[2:rows(devs),1]
endp
```

Program 3: This program implements the Local Scoring Algorithm

The following program calculated the SIR estimates for the two models used above and modified as described in Section 5.

```
library(xplore)                           ;load the necessary libraries
library(addmod)
randomize(0)
x = normal(200 5)                         ;generate the explanatory variable
```

```
rho = 0.6
beta = #(1 1 0 0 0)
eta = x*beta                                    ;eta, notation as in (2.2)
g = 1./(1+exp(-eta)) - rho * eta.*pdfn(eta);calculate g(eta)
y = (uniform(200).<g)                           ;generate the response variable
randomize(0)
x = normal(200 5)                               ;generate the explanatory variable
g1 = x[,1]
g2 = x[,2].*x[,2]-1
eps = 0.6*normal(200)
y = g1+g2+eps
b_gam = sir1(x y 2 10)
```

Program 4: This program generates the two modified models and applies SIR

REFERENCES

Bierens, H.J. (1990). "A consistent conditional moment test of functional form," *Econometrica*, **58**, 1443–1458.

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.

Buja, A., Hastie, T.J. and Tibshirani, R.J. (1989), " Linear smoothers and additive models (with discussion)," *The Annals of Statistics*, **17**, 453–555.

Eubank, R.L. (1988). *Smoothing Splines and Nonparametric Regression*, Marcel Dekker, New York and Basel.

Friedman, J.H. and Stuetzle, W. (1981), "Projection Pursuit Regression," *Journal of the American Statistical Association*, **76**, 817–823.

Härdle, W. (1990), *Applied Non-parametric Regression*, Econometric Society Monograps No. 19, Cambridge University Press.

Härdle, W., Hall, P. and Ichimura, H. (1992), "Optimal Smoothing in Single Index Models," *The Annals of Statistics*, to appear.

Härdle, W., Hart J., Marron J.S., and Tsybakov, A.B. (1992), "Bandwidth Choice for Average Derivative Estimation," *Journal of the American Statistical Association*, **87**, 218–226.

Härdle, W. and Stoker, T.M. (1989), "Investigating Smooth Multiple Regression by the Method of Average Derivatives," *Journal of the American Statistical Association*, **84**, 986–995.

Hall, P. (1989), "On Projection Pursuit Regression", *The Annals of Statistics*, **17**, 573–588

Hastie, T.J. and Tibshirani, R.J. (1987), "Non-parametric Logistic and Proportional Odds Regression," *Applied Statistics*, **36**, 260–276.

Hastie, T.J. and Tibshirani, R.J. (1990), *Generalized Additive Models*, Chapman and Hall, London.

Horowitz, J. and Härdle, W. (1992), " Testing a parametric model against a semiparametric alternative", *CentER Discussion Paper*.

Huber, P.J. (1985), "Projection Pursuit," *The Annals of Statistics*, **13**, 435–475.

Ichimura, H. (1993), "Semiparametric Least Squares (SLS) and Weighted SLS Estimation of Single-Index Models," *Journal of Econometrics*, special issue on "Nonparametric Approaches to Discrete Choice Models", ed. W. Härdle and C.F. Manski.

Li, K.C. (1991a) "Sliced Inverse Regression for Dimension Reduction (with discussion)," *Journal of the American Statistical Association*, **86**, 316–342.

Li, K.C. (1991b), Notes from the course MATH277–DATA ANALYSIS–Winter 1991, University of California, Los Angeles.

Müller, H.G. (1988), *Nonparametric Regression Analysis of Longitudinal Data*, Lecture Notes in Statistics 46, Springer, New York.

Newey, W.K. (1985), "Maximum likelihood specification testing and conditional moment test," *Econometrica*, **53**, 1047–1070.

McCullagh, P. and Nelder, J.A. (1989), *Generalized Linear Models. 2nd Edition*, Chapman and Hall, London.

Proença, I.M. (1992), *On the performance of a test against a semiparametric alternative*, manuscript in preparation.

Rodríguez-Campos, M.C. and Cao-Abad, R. (1993),"Nonparametric Bootstrap Confidence Intervals for Discrete Regression Functions," *Journal of Econometrics*, special issue on "Nonparametric Approaches to Discrete Choice Models", ed. W. Härdle and C.F. Manski.

Silverman, B.W. (1986), *Density Estimation for Statistical and Data Analysis*, Chapman and Hall, London.

Stoker, T.M. (1991), "Equivalence of Direct, Indirect and Slope Estimators of Average Derivatives," in *Nonparametric and Semiparametric Methods in Econometrics and Statistics*, Barnett, W.A., J.L. Powell and G.Tauchen, eds., Cambridge University Press.

Stoker, T.M. (1992), *Lectures on Semiparametrics Econometrics*, CORE Lecture Series, Louvain-la-Neuve.

Stoker, T.M. and Villas-Boas, J.M. (1992), "Monte Carlo Simulation of Average Derivative Estimators," *Discussion Paper*, Sloan School of Management, MIT.

Tibshirani, R.J. (1988), "Estimating optimal transformations for regression via additivity and variance stabilization," *Journal of the American Statistical Association*, **83**, 394–405.

Turlach, B. (1992), "On Discretization Methods for Average Derivative Estimation," *CORE Discussion Paper N° 9232.*

Wahba, G. (1990), *Spline Functions for Observational Data*, CBMS-NSF Regional Conference series, SIAM, Philadelphia.

XploRe (1992), XploRe 3.0 – – a computing environment for eXploratory Regression and data analysis. Available from XploRe Systems, C.O.R.E. Université Catholique de Louvain, Belgium.

The estimated mean trajectory of the coefficient (shown in Fig. 11) varies smoothly from -0.028 in early stages to -0.017 at about 1 year. After that, it stabilizes at that value but its uncertainty increases with time. Similar results are obtained after removing the two extremely large survival times confirming lack of information in later periods. On comparison with Figs 5(a) and 5(b), there may be an indication of a larger reaction of Hastie and Tibshirani's estimates in the presence of little information (confidence limits are not provided there).

Another modelling approach related to varying-coefficient models is hierarchical modelling (Lindley and Smith, 1972). The richness of the Bayesian structure, mentioned in Section 3.4.2, allows here the explanation of regression coefficients by additional covariates through stochastic relations. Dynamic hierarchical models (Gamerman and Migon, 1993) allow, in addition, coefficient variation with time. However, parametric relationships are an integral part of hierarchical (or dynamic) models and have a strong effect on the results even when the prior for the higher stage (or initial) parameter is vague. Smoothness in the appropriate direction is a consequence of the model.

Wolfgang Härdle and Marlene Müller (Humboldt University, Berlin): We would like to congratulate the authors for an excellent and interesting paper which gives a framework for a wide range of flexible regression models. The varying-coefficient model as presented in this paper is very powerful indeed. Its application in the examples in Sections 4 and 5 speak for the method proposed.

Our comment will address some aspects of inference for the estimation method described. Once the varying-coefficient regression model has been estimated it is natural to compare it with competing fits. Since the coefficients of the model are functions $\beta_j(\ )$ the comparison could be based on confidence bands for the coefficient functions. Another proposal would be a squared distance between competing coefficient functions. Suppose that the nonparametric $\hat{\beta}$ has to be tested against a parametric fit $\hat{g}$. Härdle and Mammen (1993) have derived the distribution of

$$nh^{1/2} \int (\hat{\beta} - \mathscr{K}\hat{g})^2$$

where h denotes the kernel bandwidth and $\mathscr{K}\hat{g}$ denotes the smoothed parametric model. Simulations suggest that this test (based on the quantile of the asymptotic normal distribution) is not very powerful. The correct bootstrap (the so-called 'wild bootstrap') yields much better results. Have the authors similar experiences for their test based on the 'approximate degrees of freedom'? The same comment applies to uniform confidence bands.

M. C. Jones (The Open University, Milton Keynes): My remarks concern only a rather technical point which may be of little practical consequence. Consider, for simplicity, model (5) with univariate X. Write $V = Y/X$ so that $V = \beta(X) + \epsilon/X$. One might, appropriately, fit a parametric β to (X, V) by weighted least squares using weights proportional to X^2. This *global* experience does not, it seems to me, necessarily carry over immediately to *local* nonparametric regression. In Jones (1993), I show how weighting affects Nadaraya–Watson estimators, in particular, and the answer (asymptotically) is only in terms of bias and not at all in terms of variance. Moreover, there is no argument for choosing weights inversely proportional to error variance. In fact, swift calculations involving (preferable) local linear fitting suggest no effect of weights whatsoever (asymptotically), and that the bias effect is one of Nadaraya–Watson's peculiarities. It seems, however, that there may be some sense in inverse variance weighting for splines (essentially as used by the authors), but only because of splines' effective local bandwidth choice. This appears to involve weight$(x) \times f(x)$ (Silverman, 1984); since variance of smoothers depends inversely on $\sigma^2(x)/f(x)$, inverse variance weighting is suggested.

All that I am trying to say is that the authors' weighting, which is applied to general versions of their methodology, is not quite that obviously appropriate, and that it is an issue that might repay further investigation; for example, perhaps it can be done without, although I would not expect great differences to result.

I do not mean to detract in the slightest from a most interesting and worthwhile further contribution to an important area of the subject, one to which the current authors continue to contribute enormously.

Charles Kooperberg (University of Washington, Seattle) **and Charles J. Stone** (University of California, Berkeley): It is implicit in the discussion in Section 5 of the application of varying-coefficient models to survival data that the penalized partial likelihood estimate for β_j is a natural cubic spline and hence linear in the right-hand tail. When there are scant data in this tail, and especially when there is a substantial

APPLIED NONPARAMETRIC SMOOTHING TECHNIQUES

Wolfgang HÄRDLE[1], Sigbert KLINKE[2] and Marlene MÜLLER[3]

[1,3]Humboldt-Universität zu Berlin

Wirtschaftswissenschaftliche Fakultät

Spandauer Str. 1, D-10178 Berlin, Germany

and

[2]Université Catholique de Louvain

C.O.R.E. & Institut de Statistique

Voie du Roman Pays 34, B-1348 Louvain-La-Neuve, Belgium

March 1993

Summary

Nonparametric smoothing methods are applied in statistics as a flexible tool in finding structure and connections within data. Well known means to do that are kernel, spline and nearest neighbor estimators. We present here kernel estimators, which are easy to handle in all dimensions, in various situations and applications. We first consider density estimators and show how these are used as an exploratory tool in univariate and multivariate situations. Some theory is provided for inferential issues. Next we give a short overview on smoothing techniques in univariate regression. The last chapter deals with multivariate regression, where we present two semiparametric applications. As we go along we present our computer implementations which are done entirely in *XploRe 3.1 - an interactive statistical computing environment.*

Härdle, W., Klinke, S. and Müller, M. (1994) Applied Nonparametric Smoothing Techniques.

Density estimation plays an important role in modern statistical research and practice. A variety of nonparametric methods has been proposed by various authors and there is a considerable amount of literature. This chapter concentrates on using and implementing kernel density estimation. To illustrate these ideas, we consider throughout this paper the following data sets.

Example 1
The Swiss Bank Notes data (see Flury and Riedwyl, 1988, page 5 ff.) consist of 200 measurements of old Swiss bank notes. It is known that the first 100 bank notes of the sample are genuine and the second 100 are forged. The data contain the measured values of length (X_1), left height (X_2), right height (X_3), distance of the inner frame to lower border (X_4) and to the upper border (X_5), and the length of the diagonal (X_6). One is interested in discriminating between the two groups.

Example 2
The nuclear sclerotic cataract data stem from the Beaver Dam Eye Study (see Mares-Perlman, Klein and Klein, 1992). It contains 1136 observations of age minus 60 (X_1), logarithm of zinc concentration in blood (X_2), and the level of the nuclear sclerotic cataract (Z) given in four categories labelled 1 to 4 (lowest to highest). To simplify the presentation we transform Z to a binary variable Y by pooling together levels 1, 2 ($Y = 0$) and 3, 4 ($Y = 1$). The interest lies in modelling the cataract as a function of the other two variables.

All computations and graphical presentations in this paper are done in XploRe 3.1. This system is an interactive, open statistical computing environment, which allows the user to implement complicated nonparametric algorithms in an easy way, to combine them into libraries and therefore to tailor the computing environment according to his own interests. The application of multivariate nonparametric methods is supported by predefined macros and libraries. The XploRe language is matrix oriented and offers various possibilities for interactive graphical representation. See XploRe (1993).

One important feature of kernel density estimates is that they provide easily an

impression of the data distribution. Consider *Example 2*. How can we immediately check for modes or skewness in the distribution of the variables X_1 or X_2? Binning the data and looking at histograms is often informative, but histograms change their shape with the origin of the bin sequence; see Scott (1992). Kernel estimates overcome this problem. Denote the observations of the variable X by $x_i, i = 1, \ldots, n$. The kernel density estimate $\hat{f}_h$ of the density f is constructed by averaging over scaled kernel functions centered in the points x_i, i.e.

$$\hat{f}_h(x) = \frac{1}{nh} \sum_{i=1}^{n} K\left(\frac{x - x_i}{h}\right). \tag{1.1}$$

Table 1.1 lists some commonly used kernel functions.

$K(\bullet)$	Kernel
$K(u) = \frac{1}{2}\,\mathrm{I}(\lvert u\rvert \le \frac{1}{2})$	Uniform
$K(u) = (1 - \lvert u\rvert)\,\mathrm{I}(\lvert u\rvert \le 1)$	Triangle
$K(u) = \frac{3}{4}(1 - u^2)\,\mathrm{I}(\lvert u\rvert \le 1)$	Epanechnikov
$K(u) = \frac{15}{16}(1 - u^2)^2\,\mathrm{I}(\lvert u\rvert \le 1)$	Quartic
$K(u) = \frac{1}{\sqrt{2\pi}} \exp(-\frac{u^2}{2}) = \varphi(u)$	Gaussian

Table 1.1.

The choice of the bandwidth h is the essential problem in kernel density estimation. One simple method is the "rule of thumb" for a Gaussian kernel proposed by Silverman (1986) which assumes the true underlying density to be Gaussian with variance σ^2. This yields

$$\hat{h} = 1.06 \hat{\sigma} n^{-1/5} \tag{1.2}$$

with $\hat{\sigma}$ the usual standard deviation estimator for σ. This procedure is realized in the XploRe macro `denauto` and gives us for the zinc variable X_2 of *Example 2* the curve in Figure 1.1.

The same method applied to the age variable X_1 of *Example 2* gives us the curve on the left of Figure 1.2. The calculated bandwidth is given below the figure. The estimated density shows a slight second mode and a pronounced skewness to the right. In the right part of Figure 1.2 we have calculated $\hat{f}_h$ for $h = \frac{1}{2}\hat{h}$. We see that more modes appear and that the distribution is most likely non-normal. The modes are of course functions of the bandwidth.

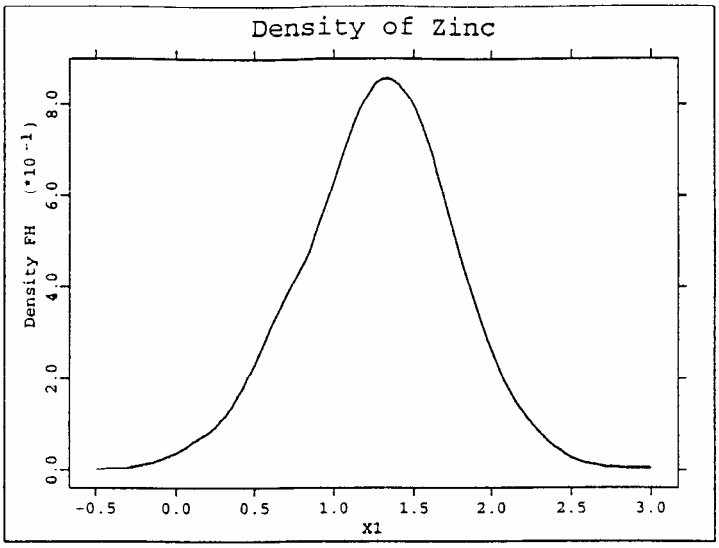


Figure 1.1. Kernel density estimate for the zinc distribution

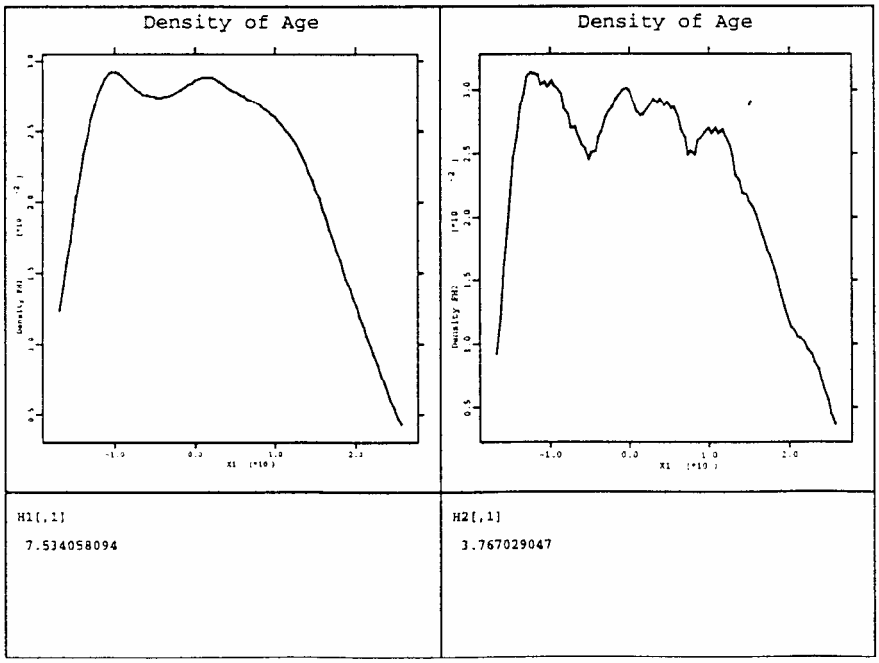


Figure 1.2. Kernel density estimates with bandwidths for the age distribution

Kernel density estimates for a given bandwidth can be calculated in XploRe

via the macro denest. The corresponding XploRe code for Figure 1.2 is given below.

```
proc()=main()
  x=read(nuc)                    ; read the data file "nuc.dat"
  library(smoother)              ; load the necessary library
  fh1=denauto(x[,1])             ; estimate fh1 by rule of thumb
  n=rows(x)
  h1=2.62*1.06*sigma*n^(-0.2)    ; bandwidth h1 used in "denauto"
  h2=h1./2                       ; bandwidth h2
  fh2=denest(x[,1] h2)           ; density estimate with h2
                                 ;
  createdisplay(d1, 2 (-2), s2d text s2d text)
                                 ; create 2 text & 2 graphics displays
                                 ;
  show(fh1 s2d1, fh2 s2d2, h1 text1, h2 text2)
                                 ; show fh1, fh2 and h1, h2
endp
```

Program 1.1. Macro for the density estimates of Figure 1.2

More information on data can be provided by two dimensional density estimates, calculated below with the XploRe macro denest2.

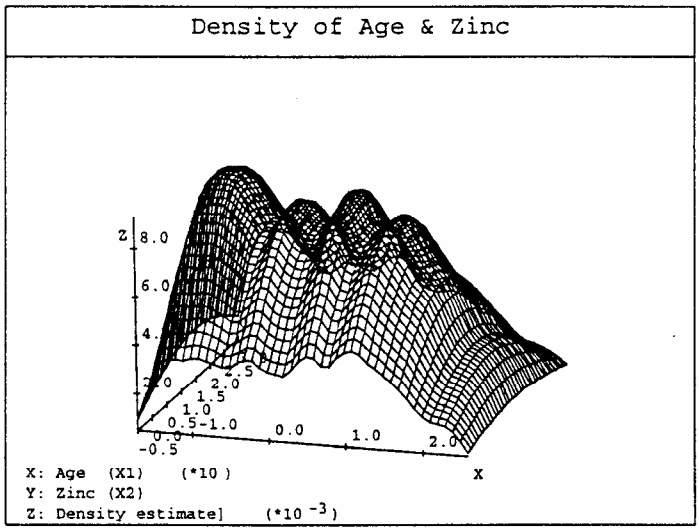


Figure 1.3. Kernel density estimate for age and zinc

One sees immediately the unimodal structure in the zinc direction (Y axis) as well as the different modes in the age direction (X axis). It is clear that the exploratory character of density estimates is restricted to low-dimensional data; the limit is for three dimensional data. Scott (1992) uses contour shells to reveal structure in three to four dimensional data. Let us here review a method for finding non-normal structures in high-dimensional data sets by combining projection and density estimation techniques.

Projection Pursuit techniques cover a wide field of interesting topics in data analysis (density estimation, regression, exploratory data analysis). The idea is to find an informative low-dimensional projection of a high-dimensional data set which help to describe the non-normal structure of the data. We recall that each projection of a Gaussian distribution has also a Gaussian distribution. Since there is an infinite number of possible projections there is a need for an automatic choice. Let $I(\alpha) : I\!R^n \to I\!R$ be a *projection index*, where α describes a projection vector with $|\alpha| = 1$. The aim is to maximize this index in order to detect non-normality. If we project the data points x_i, we obtain 1-dimensional data points $z_i = \alpha^T x_i$. The density estimation of z_i as a function of α should look non-normal for projections α revealing non-normality. The task is thus to find such projections.
Friedman (1987) defined an index by transforming the projected data such that normally distributed data will be uniformly distributed on $[-1, 1]$ and measured the non-uniformity by

$$I(\alpha) = \int_{-1}^{1} \left\{ p_R(R) - \frac{1}{2} \right\}^2 dR \tag{1.3}$$

where $R = 2\Phi(\alpha^T X) - 1$ and $p_R(R)$ is the probability density of R. After expanding $p_R(R)$ by Legendre-polynomials, we obtain the estimator

$$\hat{I}(\alpha) = \frac{1}{2} \sum_{j=1}^{J} \frac{2j+1}{n^2} \left[\sum_{i=1}^{n} P_j\{2\Phi(\alpha^T x_i)\} - 1 \right]^2 . \tag{1.4}$$

Here, J plays the role of a smoothing parameter. It is well known that this index is not very robust against outliers. We obtain usually skewed distributions in this case. Nevertheless we present it here to demonstrate the interplay between density estimation and projection pursuit techniques. To demonstrate the performance of this method we apply it to the Swiss Bank Notes data set (dimension 6) described in *Example 1*. The following picture is obtained by the `ppexpl` macro of XploRe.

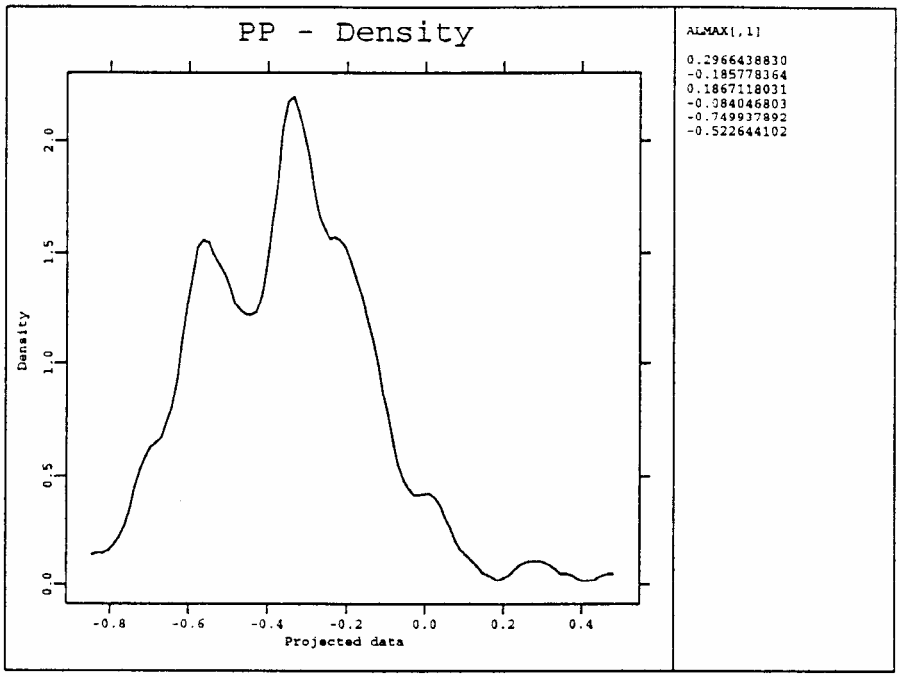


Figure 1.4. Estimated density of projected data in Exploratory Projection Pursuit

Figure 1.4 shows the projection of the data maximizing $\hat{I}(\alpha)$ for $J = 3$. The coordinates of the projection are displayed in the right window. We can clearly see that the data set separates into two clusters. For a projection pursuit discriminant analysis approach see Polzehl (1993).

In practical studies the choice of the kernel does not have a great influence on the resulting density estimate assuming the bandwidth h is optimal. One possible definition of optimal bandwidth is that h which minimizes a distance between the true underlying density and the density estimate. Usual distances are the *integrated squared error* $ISE(h)$ or its mean $MISE(h)$ which has the asymptotic representation $(n \to \infty, nh \to \infty, h \to 0)$

$$MISE(h) \approx C_1 n^{-1} h^{-1} + C_2 h^4, \tag{1.5}$$

with the constants $C_1 = \int K^2(u)du$, $C_2 = \frac{1}{4}\ \mu_2^2(K)\ \int \{f''(x)\}^2 dx$ and $\mu_2(K) = \int u^2 K(u)du$.

The concept of canonical kernels introduced by Marron and Nolan (1989) scales the kernels such that they are equivalent in view of $MISE$. More exactly, for

two kernels K_1, K_2,

$$MISE_{K_1}(h_1) \approx c_{K_1,K_2} MISE_{K_2}(h_2) \tag{1.6}$$

with a constant c_{K_1,K_2} independent of h_1, h_2, if the bandwidths h_1, h_2 fulfill

$$h_2 = h_1 \frac{\delta_2^*}{\delta_1^*}, \quad \delta_i^* = \left\{ \frac{\int K_i^2(u)du}{\mu_2^2(K_i)} \right\}^{1/5}. \tag{1.7}$$

Table 1.2 shows the transformation factors for the bandwidths if we change from one kernel of Table 1.1 to another. XploRe provides the macro `canker` which automatically calculates this transformation.

δ_j^*/δ_i^*	Uniform	Triangle	Epanechnikov	Quartic	Gaussian
Uniform	1.000	0.715	0.786	0.663	1.740
Triangle	1.398	1.000	1.099	0.927	2.432
Epanechnikov	1.272	0.910	1.000	0.844	2.214
Quartic	1.507	1.078	1.185	1.000	2.623
Gaussian	0.575	0.411	0.452	0.381	1.000

Table 1.2. Canonical kernel transformations from `canker`

Kernel density estimation is not an easy computational task. Some packages are very specialized in the sense that they can do only calculations related to kernels (e.g. N-kernel). On the other hand some are so general (e.g. GAUSS) that everything must be programmed (often with Do-loops). Others have not enough built-in flexibility (e.g. Splus supports only four kernels: Cosine, Gaussian, Uniform and Triangle). XploRe offers a great variety of kernels as well as predefined macros which realize the computation of kernel estimates. The user may add his own kernels of course. The fast computing algorithms are based on the WARPing technique, described e.g. in Härdle (1991). The basic idea is the "binning" of the data in bins of length δ starting at an origin x_0. Instead of evaluating the kernel for all differences $(x_i - x_j), i, j = 1, ..., n$ the kernel function needs now to be evaluated only at $i\delta h^{-1}, i = 1, ..., \ell$, where ℓ is number of bins which contains the support of the kernel function.
The calculation of the estimated density reduces from $\hat{f}_h(x_j)$ to

$$\bar{f}_h(\bar{x}_j) = \frac{n_j}{nh} \sum_{i=1}^{N_b} n_i K\left\{(i-j)\delta h^{-1}\right\} \tag{1.8}$$

computed on the grid $\bar{x}_j = x_0 + j\delta$ with N_b denoting the number of non-empty bins, n_i the number of observations in the i-th bin. The XploRe density estimation macros `denauto`, `denest` and `denest2` use the fact of the asymptotic

equivalence of the kernels described above and calculate therefore the density estimates based on the Quartic kernel, which has the advantage of a compact support. We give in Program 1.2 the XploRe code for the `denauto` macro, `denest` and `denest2` are programmed in an analogous way.

```
proc(fh)=denauto(x)
  d=(max(x)-min(x))./100                ; make 100 bins
  (xb yb)=bindata(x d)                  ; bin the data
  sigma=sqrt(cov(x))                    ; estimate the variance
  h=2.62*1.06*sigma*(rows(x))^(-0.2)    ; determine h by rule of thumb
                                        ; use transformation constant
                                        ; for change to quartic kernel
  wy=symweigh(0 d/h h/d &qua)           ; create weights for the
                                        ; symmetric quartic kernel
  wx=aseq(0 rows(wy))
  (xc yc or)=conv(xb yb wx wy)          ; calculate density func
  fh=(xc*d)^(yc/(n*d))
endp
```

Program 1.2. Automatic density estimation, XploRe macro `denauto`

Since we have an asymptotic Gaussian distribution for the kernel density estimate at fixed points x if $n \to \infty$, we can also construct confidence intervals for $\hat{f}_h(x)$. For a bandwidth $h = cn^{-1/5}$ holds the following formula (Silverman, 1986; Härdle, 1991)

$$n^{2/5}\{\hat{f}_h(x) - f(x)\} \xrightarrow{\mathcal{L}} N\left(\frac{c^2}{2} f''(x)\mu_2(K) \ , \ c^{-1} f(x) C_1 \right). \tag{1.9}$$

For small h (in relation to $n^{-1/5}$) the mean in (1.9) is negligible. This yields the asymptotic confidence interval

$$\left[\hat{f}_h(x) - u_{1-\alpha/2}\sqrt{\frac{\hat{f}_h(x)C_1}{nh}} \ , \ \hat{f}_h(x) + u_{1-\alpha/2}\sqrt{\frac{\hat{f}_h(x)C_1}{nh}} \right] \tag{1.10}$$

with $u_{1-\alpha/2}$ denoting the $\alpha/2$ quantile of the standard normal distribution.
The XploRe code below calculates the confidence intervals for the density estimation of the age data of *Example 2.* We show these confidence intervals in Figure 1.5. One sees that the modality structure remains inside the confidence intervals.
Formulas and computer algorithms for true confidence bands can be found in Bickel and Rosenblatt (1973), Härdle (1991). They are slightly more complicated but they have the same underlying idea to exploit the asymptotic limit distribution. For a bootstrap approach see Hall (1992, page 220 ff.).

```
proc()=main()
  x=read(nuc)                                ; read the data
  library(smoother)                          ; load the necessary library
  h=3.767                                    ; bandwidth
  fh=denest(x[,2] h)                         ; density estimate
  ci=2*sqrt((5/7).*fh[,2]./(rows(x)*h))
  cup=fh[,1]~(fh[,2].+ci)                    ; upper confidence bound
  clo=fh[,1]~(fh[,2].-ci)                    ; lower confidence bound
  show(fh s2d)                               ; display the result
endp
```

Program 1.3. Density estimation confidence intervals in XploRe

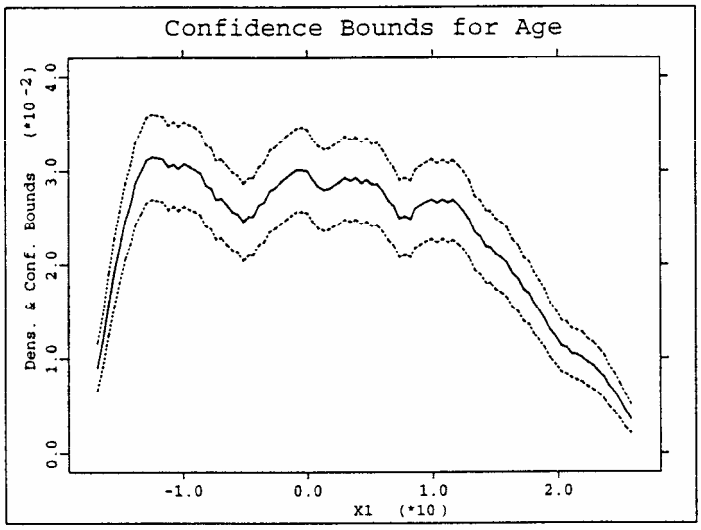


Figure 1.5. Density estimate and confidence intervals for the age data

2. Smooth regression in one dimension

The aim of this chapter is to give a very short overview and to recall the main ideas of univariate nonparametric regression methods, in particular kernel re-

gression smoothing. The univariate regression model assumes observations of two variables X and Y, i.e. data of the form $(x_i, y_i), i = 1, \ldots, n$ which are connected via an unknown regression function $m(\bullet)$ as follows:

$$y_i = m(x_i) + \epsilon_i \ , \ i = 1, \ldots, n, \tag{2.1}$$

ϵ_i denoting the error variables. The problem is now to estimate $m(\bullet)$. Nonparametric estimates suppose no prior knowledge of m. There is an obvious analogy with the nonparametric kernel density estimation. Plugging in kernel estimates for $f(x)$ and $f(x,y)$ in $m(x) = \mathbf{E}(Y|X = x) = f(x)^{-1} \int y f(x,y) dy$ leads to the popular Nadaraya–Watson estimate

$$\hat{m}_h(x) = \frac{n^{-1} \sum_{i=1}^{n} K_h(x - x_i) y_i}{n^{-1} \sum_{j=1}^{n} K_h(x - x_j)} \ . \tag{2.2}$$

This kernel estimator is essentially a local average of the y_i variables with corresponding x_i's close to x. This local averaging behavior is behind several other smoothing techniques, e.g. k-nearest-neighbor and spline smoothing. They are in an asymptotic sense equivalent to kernel smoothing with a bandwidth depending on x, see Härdle (1990), Chapter 3.

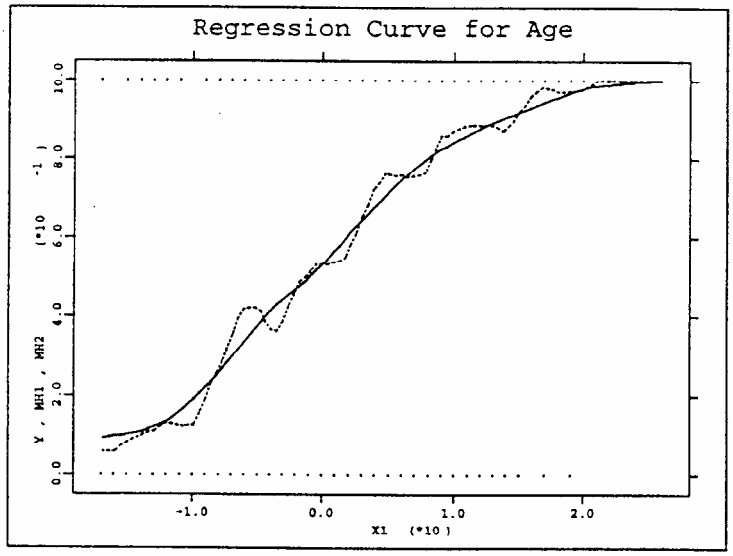


Figure 2.1. Regression curve in XploRe.

In XploRe the kernel smoothing is again performed via the WARPing technique in the macro `regest`. The bandwidth choice is as in density estimation crucial for the practical performance. The WARPing technique allows fast computation of the cross-validation bandwidth (macro `regcvl`). For a local plug-in choice we refer to Härdle and Marron (1991).

Figure 2.1 shows two kernel estimates according to (2.2) for the regression of Y on $X = X_1$ in *Example 2*. The bandwidths are chosen as $h = 2.5$ and $h = 7$. This is of course only a marginal relation between one variable and the response. A more refined analysis based on semiparametric models is presented below. More details on nonparametric regression, especially on bandwidth choice, are given in the monographs of Eubank (1988), Müller (1988), Hastie and Tibshirani (1990), and Härdle (1990).

3. Regression Smoothing in High Dimensions

The kernel methods described in Chapter 2 can be generalized to the multivariate case. In most practical applications though problems will arise due to the sparseness of data. Projection based methods or additive modelling avoid this data sparseness. We will present *Single Index Models* and *Generalized Additive Models*, both generalizations of *Generalized Linear Models* (GLM), see McCullagh and Nelder (1989). The GLM generalizes the linear regression model with systematic component $\eta = X^T\beta$ to $\mathbf{E}Y = G(\eta)$ with a known (inverse) link function G. In our *Example 2*, where we have binary responses, we model

$$P(Y = 1|X = x) = G(x^T\beta). \tag{3.1}$$

The Single Index Model idea is to generalize (3.1) to arbitrary smooth link functions g. This is what is called in the statistical literature a one term *Projection Pursuit Model*. Friedman and Stützle (1981) proposed an iterative method for estimating β. Härdle and Stoker (1989) derived a direct non-iterative method, the so-called *Average Derivative Estimation* (ADE). The ADE idea is as follows. For $m(x) = g(x^T\beta)$ the average derivative

$$\delta = \mathbf{E}m'(X) = \mathbf{E}\left[\frac{dg}{d(x^T\beta)}(X^T\beta)\right]\beta \tag{3.2}$$

determines β up to a scale factor. Since δ equals $\mathbf{E}\ell Y$, ℓ denoting the *score function* $-\partial \log f/\partial x = -f'/f$, it can be estimated by $\hat{\delta} = n^{-1}\sum_{i=1}^{n} \hat{\ell}_h(x_i)y_i$ with $\hat{\ell}$ based on a kernel density estimate with bandwidth h.

Figure 3.1 shows for *Example 2* the projected observations $x_i^T\hat{\delta}$ vs. the responses y_i as well as two link functions $\hat{g}$ (computed with `regest`) with bandwidths $h = 0.05$ and $h = 0.15$, respectively.

Since the two estimates of Figure 2.1 and Figure 3.1 looks a bit similar it would be interesting to test whether $\delta_2 = 0$, i.e. whether there is no influence of zinc. We have here $\hat{\delta} = \begin{pmatrix} 0.02 \\ -0.05 \end{pmatrix}$ and its estimated covariance $\hat{\Sigma}_\delta = \begin{pmatrix} 0.00044 & 0.00004 \\ 0.00004 & 0.12195 \end{pmatrix}$. Härdle and Turlach (1992) describe a test using a Wald statistic which goes back to Stoker. To test the hypothesis $R\delta = r$ we have to compare the test statistic $W = n(R\hat{\delta} - r)^T(R\hat{\Sigma}R^T)^{-1}(R\hat{\delta} - r)$ to the χ^2(rank R) value. For our running example this leads to $W = 29.177$, thus we reject the hypothesis.

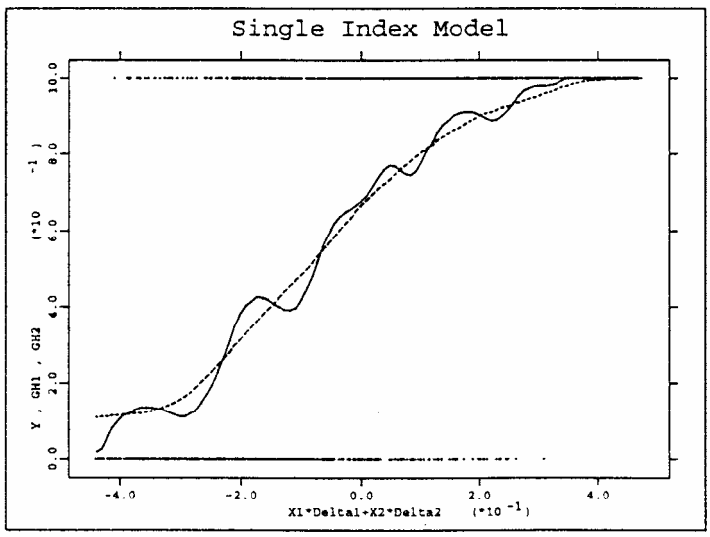


Figure 3.1. Single Index Model fit

One can see a clear asymmetry in the link functions, speaking against a symmetric, e.g. logistic, link. The XploRe code below uses the macro `adeind` which calculates the ADE.

```
proc()=main()
  x=read(nuc)                          ; read the data
  library(smoother)                    ; load the necessary libraries
  library(addmod)
  y=x[,3]                              ; the response variable
  x=x[,1:2]                            ; the age and zinc variables
  d=(max(x)-min(x))/20                 ; binning parameter
  (xb yb)=bindata(x d 0 y)             ; bin the data
  (del v) = adeind(xb yb d 3)          ; the ADE del and its asymptotic
                                       ; covariance matrix  v
  est=(x*del)~y                        ; the projected data
  gh1=regest(est 0.05)                 ; estimate g, h=0.05
  gh2=regest(est 0.15)                 ; estimate g, h=0.15
  show(est gh1 gh2 s2d)                ; display the results
endp
```

Program 3.1. The Single Index Model in XploRe

Generalized Additive Models (GAM) keep the link but generalize the projection $x^T\beta$ to a sum of nonparametric transformations. These fall into the class of *Additive Models*, i.e. one assumes

$$P(Y = 1|X = x) = G\left(\alpha + \sum_{j=1}^{d} g_j(x_j)\right). \tag{3.3}$$

For an introduction into this class of models we refer to the book of Hastie and Tibshirani (1990). The algorithm to estimate this model consists of *local scoring* and *backfitting* to determine the nonparametric transformations $g_1, \ldots, g_d$. Program 3.2 shows the realization of this iteration process in XploRe. The main part of the work is done in the local scoring macro `lscore`, which calls the backfitting macro `backfit`. The nonparametric estimates for $g_1, \ldots, g_d$ are obtained by a k-nearest-neighbor method with k the number of 30% of the data points. As link function we have taken the logistic link.

The output of Program 3.2 is displayed in Figure 3.2. The upper left picture shows the y_i vs. $\hat{\eta}_i = \hat{\alpha} + \sum_{j=1}^{d} \hat{g}_j(x_j)$ and the fit $G(\hat{\eta}_i)$. The lower left and upper right pictures show the estimated nonparametric components $\hat{g}_1(x_{1i})$ vs. x_{1i} and $\hat{g}_2(x_{2i})$ vs. x_{2i}. The lower right picture displays the 3-dimensional surface $(x_{1i}, x_{2i}, \hat{\eta}_i)$. We see that the nonparametric $\hat{g}_1$ for the age is almost linear. However $\hat{g}_2$ for zinc is nonlinear and has a negative slope. Recall that the ADE estimate was negative in the second component, too. But one should pay attention that $\hat{g}_1$ varies in $[-4, 6]$ whereas $\hat{g}_2$ takes values in $[-1, 1]$. So it turns out that the zinc variable has a smaller influence than the age variable.

```
proc()=main()
  x=read(data2\nuc)                    ; read the data
  library(smoother)                    ; load the necessary libraries
  library(addmod)
  y=x[,3]                              ; the response variable
  x=x[,1:2]                            ; the age and zinc variables
  (fx alpha dev)=lscore(x y 0.3)       ; run the local scoring algorithm
                                       ; which includes the backfitting
  eta=alpha+sumr(fx)
  mu=exp(eta)./(1+exp(eta))            ; calculate the link function
                                       ;
  createdisplay(h1, 2 2, s2d s2d s2d d3d)
                                       ; create the output display
  dat11 = eta~y
  dat12 = sort(eta~p)
  dat2  = sort(x[,1]~fx[,1])
  dat3  = sort(x[,2]~fx[,2])
  dat4  = x~eta
  show(dat11 dat12 s2d1, dat2 s2d2, dat3 s2d3, dat4 d3d1)
                                       ; show the results
  res=(y-mu)./(sqrt(mu.*(1-mu)))       ; calculate the residuals
  xres=x~res                           ; combine res with x
  write(xres xres)                     ; write xres to a file "xres.dat"
endp
```

Program 3.2. The Generalized Additive Model in XploRe

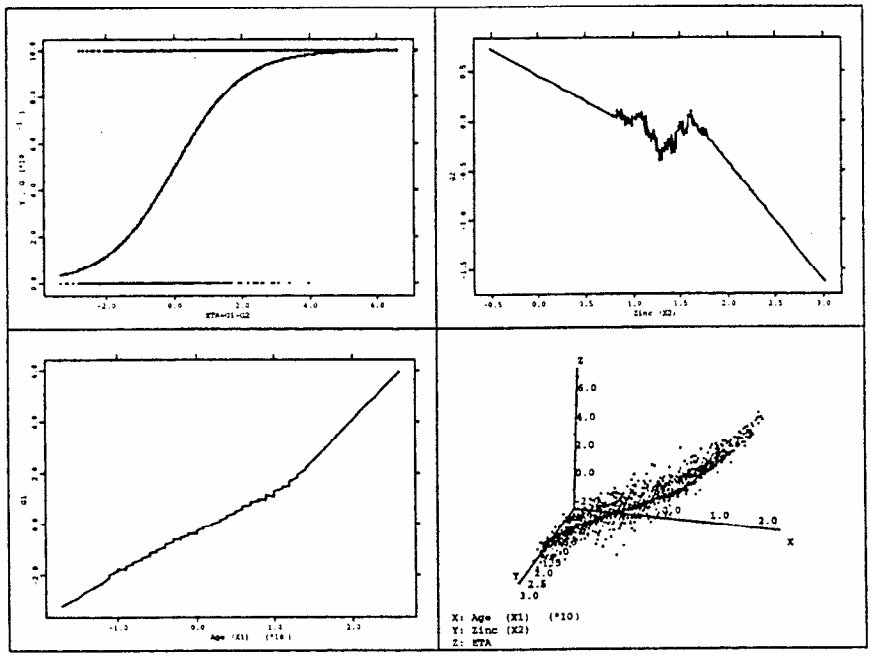


Figure 3.2. Generalized Additive Model fit

To assess the quality of this GAM estimate we provide a brushed residual plot. The following figure shows in the upper left a plot of the age and zinc data, masked by a point "·" if the corresponding residual lies in the lower quartile, by a "+" if the residual lies in the interquartile range and masked by a "O" if it is in the upper quartile. The three other plots show these residual groups separately.

The masking can be achieved directly by giving the corresponding commands in the program or brushing interactively on the screen since the four displays are linked.

It is easy to see that the residual plot underlines our conclusion from the GAM fit that the zinc has less influence on the results. Inferential issues are still open for GAMs as has been pointed out by Härdle and Müller (1993).

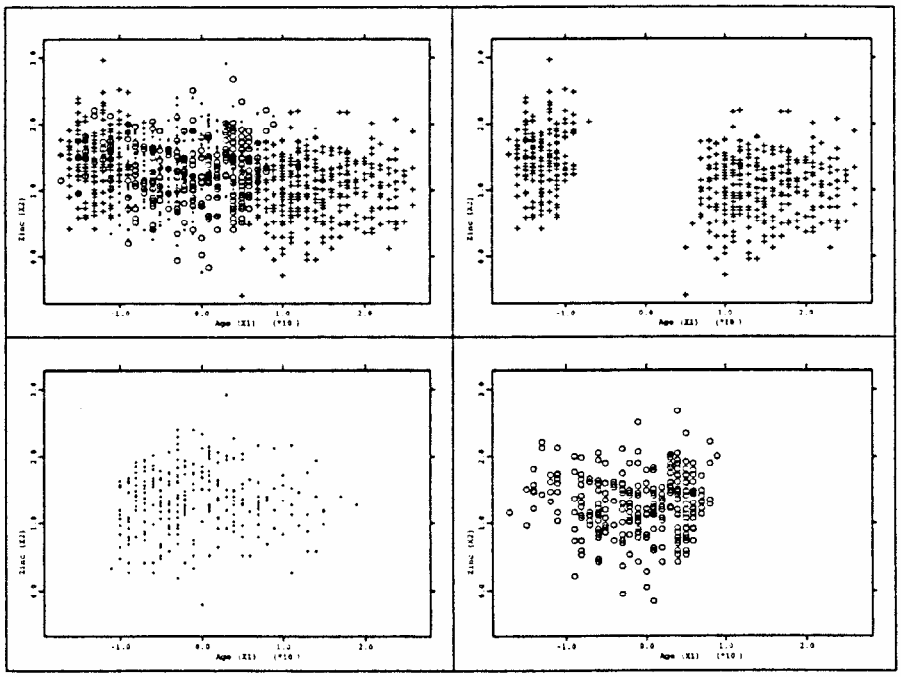


Figure 3.3. Brushed residual plot for the GAM fit

Program 3.3 gives the XploRe code for Figure 3.3.

```
proc()=main()
  xres=read(xres)                          ; read file "xres.dat"
  xres=sort(xres 3)                        ; sort data after residuals
  xr=xres[,1:2]                            ; sorted age and zinc values
                                           ;
  createdisplay(h2, 2 2, s2d s2d s2d s2d)
                                           ; create output display
  xr1=xr[1:284,]                           ; data with lower quartile res.
  xr2=xr[284:852,]                         ; data with interquartile res.
  xr1=xr[853:1136,]                        ; data with upper quartile res.
                                           ;
  show(xr1 xr2 xr3 s2d1, xr1 s2d2, xr2 s2d3, xr3 s2d4)
                                           ;
  link("s2d1\data_1" "s2d2\data_1")        ; link the 4 displays
  link("s2d1\data_2" "s2d3\data_1")
  link("s2d1\data_3" "s2d4\data_1")
  display(h2)                              ; show the linked displays
endp
```

Program 3.3. Residual plot in XploRe (Figure 3.4)

There are many more approaches in the analysis of high dimensional data. We would like to mention among others *Alternating Conditional Expectations* (ACE), Breiman and Friedman (1985), and *Sliced Inverse Regression* (SIR), a method introduced by Li (1991). Both methods are available in XploRe by the `acefit` and `sir1`, `sir2` macros.

Acknowlegdements

We are grateful to Julie Mares-Perlman, Barbara and Ron Klein for the permission to use their data on Nuclear Sclerotic Cataract (*Example 2*). We would also like to thank Christian Ritter (Louvain-la-Neuve) for attracting our interest to this data set.

References

BICKEL, P.J.; ROSENBLATT, M. (1973). On some global measures of the deviations of density function estimates. Ann. Statist. 1, 1071-1095.

BREIMAN, L.; FRIEDMAN, J.H. (1985). Estimating optimal transformations for multiple regression and correlation (with discussion). J. Amer. Statist. Assoc. 80, 580-619.

EUBANK, R.L. (1988). Spline smoothing and nonparametric regression. Marcel Dekker, New York.

FLURY, B.; RIEDWYL, H. (1988). Multivariate statistics. A practical approach. Chapman and Hall, London.

FRIEDMAN, J.H. (1987). Exploratory Projection Pursuit. J. Amer. Statist. Assoc. 82, 249-266.

HÄRDLE, W. (1990). Applied nonparametric regression. Econometric Society Monographs No. 19, Cambridge University Press, Cambridge.

HÄRDLE, W. (1991). Smoothing Techniques. With Applications in S. Springer, New York.

HÄRDLE, W.; MARRON, J.S. (1991). Fast and Simple Scatterplot Smoothing. CORE Discussion Paper No. 9143, Université Catholique de Louvain.

HÄRDLE, W.; MÜLLER, M. (1993). Discussion of: Hastie, T.J.; Tibshirani, R.J. (1993). Varying-Coefficient Models. J. Roy. Statist. Soc. B 55, in print.

HÄRDLE, W.; TURLACH, B. (1992). Nonparametric Approaches to Generalized Linear Models. In: Fahrmeir, L.; Francis, B.; Gilchrist, R.; Tutz, G. (Eds.): Advances in GLIM and Statistical Modelling. Lecture Notes in Statistics No. 78, Springer, New York, 213-225.

HASTIE, T.J.; TIBSHIRANI, R.J. (1990). Generalized Additive Models. Chapman and Hall, London.

HALL, P. (1992). The Bootstrap and the Edgeworth Expansion. Springer, New York.

LI, K.-C. (1991). Sliced Inverse Regression for Dimension Reduction (with discussion). J. Amer. Statist. Assoc. 86, 316-342.

MARES-PERLMAN, J.; KLEIN, B. and R. (1992). Beaver Dam Eye Study and Study on Nutritional Factors in Eye Disease. Funded by National Institutes of Health, National Eye Institute grants No. U10-EY06594 and R01-EY08012)

MARRON, S.; NOLAN, D. (1989). Canonical kernels for density estimation. Stat. Prob. Letters 7, 191-195.

McCULLAGH, P.; NELDER, J.A. (1989). Generalized Linear Models, 2nd. Ed., Chapman and Hall, London.

MÜLLER, H.-G. (1988). Nonparametric regression analysis of longitudinal data. Lecture Notes in Statistics No. 46, Springer, Berlin.

PARK, B.; TURLACH, B. (1992). Practical performance of several data driven bandwidth selectors (with discussion). Comp. Statist. 7, 251-271.

POLZEHL, J. (1993). Projection Pursuit Discriminant Analysis. CORE Discussion Paper, Université Catholique de Louvain.

SCOTT, D.W. (1992). Multivariate density estimation. Wiley, New York.

SILVERMAN, B.W. (1986). Density estimation for statistics and data analysis. Chapman and Hall, London.

XPLORE (1993). XploRe 3.1 - an interactive statistical computing environment. Available from XploRe Systems, Institut für Statistik und Ökonometrie, Wirtschaftswissenschaftliche Fakultät der Humboldt-Universität Berlin.
(XploRe 3.0 is available via anonymous ftp from the directory pub/xplore of amadeus.wiwi.hu-berlin.de = 141.20.100.2)

F. Faulbaum (Ed.) (1994). SoftStat '93
Advances in Statistical Software 4. Stuttgart - Jena - New York: Gustav Fischer, 261-266

XploRe - An Interactive Computing Environment

W. Härdle and T. Kötter

Summary

XploRe is an interactive statistical software for PCs. The design of XploRe is performed in a way that encourages immediate interaction with the data. Non-parametric smoothing methods in high dimensions are feasible through additive models and massive use of automatic smoothing methods. XploRe has highly interactive graphics and allows windows of different types.

1 Introduction

XploRe was designed as computational tool for statisticians. The aim was to offer facilities for easy access to and development of statistical algorithms (e.g. higher data objects, matrix computation, basic statistical routines, graphics) that also run on low cost computers. PCs were selected as the appropriate hardware base.

The first two releases of XploRe were coded in Pascal as menu driven software. The advantage was the easy handling of the software. However by the development of release 3 the wish for more flexibility caused the implementation of an interpreter and an integrated high level programming language. Many algorithms and all libraries of XploRe are coded in this incorporated language. Due to some perfomance advantages and better standardisation the development language of XploRe 3.0 was switched to C.

Nowadays additional features are necessary and important, especially in the domain of interaction and graphics. Today's statistical software should also provide integrated tools (e.g. an editor) and a help system.

The current version, XploRe 3.1, meets all above mentioned demands as they are

- a statistical interpreter including a high level statistical programming language
- matrices as basic data objects
- various graphical output (static, dynamic, 2- and 3-dimensional, multiple windows) including PostScript interface

SoftStat '93, Advances in Statistical Software 4, 261-266

- an editor for macros and data
- libraries for different statistical topics
  - smoothing facilities
  - highdimensional fitting techniques
  - teach ware
  - a complex matrix library
- a help system

In order to have a convenient handing special emphasis was given to an easy switching between the three components "editor", "interpreter" and "help system". The integration of all necessary tools avoids the need of using external help when working with XploRe . Nevertheless it is possible e.g. to write macros with an own editor. Data can be loaded and saved in two different formats. One is the ASCII-format which allows an easy exchange between other software or even editing by hand. The other one is a space and speed optimized internal format.

2 Basic Window

When starting XploRe the screen is divided into three parts:

1. the action screen
2. the command line (last line)
3. the icon list (last column)

The action screen which covers nearly the whole screen shows data, macros and graphics; essentially all output of XploRe is displayed here. The command line where the user types in his commands allows also to copy lines from the editor into it. Further on all previous commands can be retrieved through a 'last commands buffer'. On the icon list current available function keys are represented by small symbols. Since the function keys have different meanings in different parts of the XploRe software the icon list is meant to give the user a symbolic display of functions actually available.

3 Graphics in XploRe

One of the most important features of XploRe is the wide range of facilities to look at data.

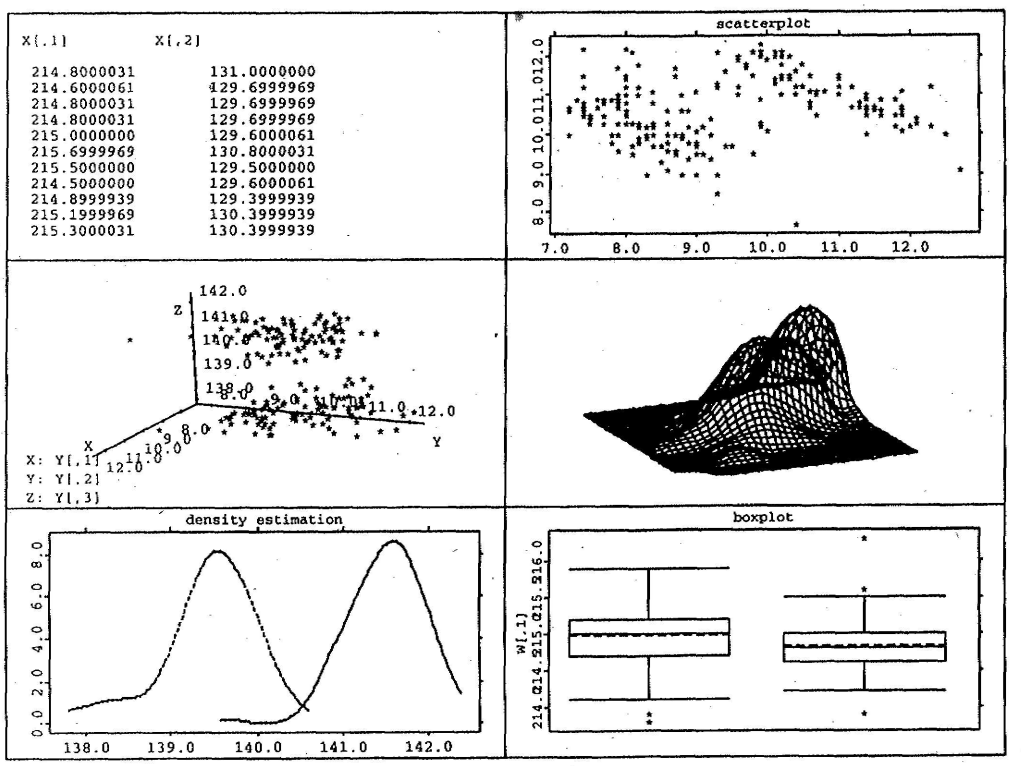
The action screen can be divided into different plot areas with various graphical styles. Apart from different possibilities to present the data the user may furthermore link them

Härdle, W. and Kötter, Th. (1994) XploRe - An Interactive Computing Environment.

SoftStat '93, Advances in Statistical Software 4, 261-266

or execute brushing operations like highlighting, masking, etc. which effect the single plots simultaneously.

The following picture shows the action screen divided into six different subscreens (text, scatterplot, projection of a 3-dimensional point cloud, 2-dimensional density estimation, line plot, box plot).



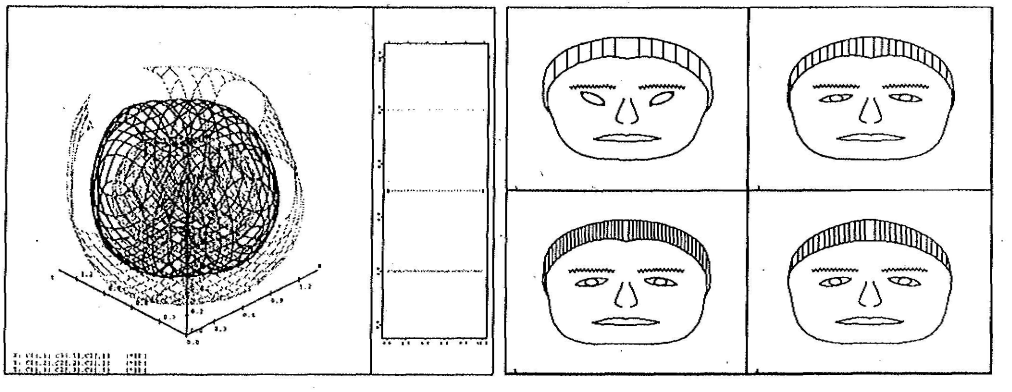
Further graphics types are sunflower plot, jitter plot, Chernoff–Flury faces, Draftman's plot, 2- and 3-dimensional contour plots (lines or surfaces, respectively) etc. .

The graphical options are chosen with the help of the function keys, which are represented here again by an icon list.

The choice of the graphical options are stored in the graphical display and automatically applied to graphics of the same kind (e.g. static 2-dimensional) so that they are displayed in an equal manner (e.g. useful by graphical output of an iteration process).

The following picture shows the contour surfaces for 20%, 50% and 80% of a simulated data set (six 3-dimensional normal distributions on the angles of a hexagon) and four Chernoff–Flury faces.

SoftStat '93, Advances in Statistical Software 4, 261-266



4 The Help System

Much effort was directed to implement not only a keyword sensitive help system but also to allow the user to look behind macros and data. Additional the user is able to extend the help system, e.g. for own written libraries.
The help system consists essentially of two keys:

F1 shows general help (e.g. in the interpreter an ordered list of all available commands or in a graphical display a list of operations which are invoked by function keys).

F10 works as 'content operator', an operator giving the "content" of a keyword known by XploRe , where the cursor is currently on. This means that:

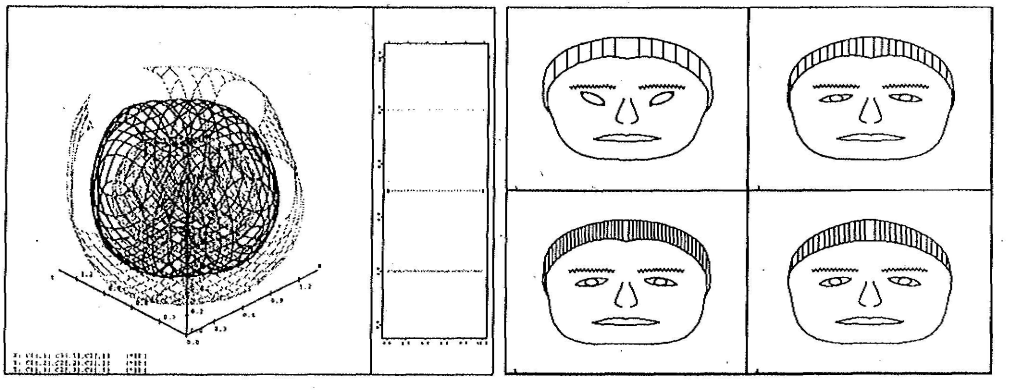
- pressing F10 on an XploRe command shows the appropriate help file. In this help file the user can call further help by applying the same procedure.
- pressing F10 on an XploRe macro shows the macro itself. Since all macros belonging to XploRe start with helpful comments they are self-explaining to the user. In addition, the help system can be easily extended to user written macros or libraries.
- pressing F10 on an XploRe variable invokes the data editor to show the actually data values held by this variable.

5 The Macro Language

XploRe provides a high level statistical programming language which allows the user to add further statistical algorithms, which then appear like regular XploRe commands. This open system strategy ensures that XploRe can be used in different research domains. The language contains full flow control and due to the management of local variables it allows even recursion.
Macros consist of

Härdle, W. and Kötter, Th. (1994) XploRe - An Interactive Computing Environment.

SoftStat '93, Advances in Statistical Software 4, 261-266



4 The Help System

Much effort was directed to implement not only a keyword sensitive help system but also to allow the user to look behind macros and data. Additional the user is able to extend the help system, e.g. for own written libraries.
The help system consists essentially of two keys:

F1 shows general help (e.g. in the interpreter an ordered list of all available commands or in a graphical display a list of operations which are invoked by function keys).

F10 works as 'content operator', an operator giving the "content" of a keyword known by XploRe , where the cursor is currently on. This means that:

- pressing F10 on an XploRe command shows the appropriate help file. In this help file the user can call further help by applying the same procedure.
- pressing F10 on an XploRe macro shows the macro itself. Since all macros belonging to XploRe start with helpful comments they are self-explaining to the user. In addition, the help system can be easily extended to user written macros or libraries.
- pressing F10 on an XploRe variable invokes the data editor to show the actually data values held by this variable.

5 The Macro Language

XploRe provides a high level statistical programming language which allows the user to add further statistical algorithms, which then appear like regular XploRe commands. This open system strategy ensures that XploRe can be used in different research domains. The language contains full flow control and due to the management of local variables it allows even recursion.
Macros consist of

Härdle, W. and Kötter, Th. (1994) XploRe - An Interactive Computing Environment.

SoftStat '93, Advances in Statistical Software 4, 261-266

`addmod`	additive modelling
`complex`	complex numbers and complex matrix routines
`csse`	constraint spline smoothing
`glm`	generalized linear modelling
`highdim`	highdimensional data analysis
`smoother`	density and regression estimation
`tware`	interactive teach ware
`xplore`	basic statistical tools

6 Final Remarks

The foregone version, XploRe 3.0, is now public domain which can be obtained via anonymous–ftp as well as the libraries `complex` and `xplore` and the documentation from the ftp–server `amadeus.wiwi.hu-berlin.de (141.20.100.2)`.

The main difference between the versions 3.0 and 3.1 lies only in the new memory management in XploRe 3.1 which is able to overcome the 640 KB limit of PCs and to manage virtual memory. This enables XploRe 3.1 to handle large data sets.

XploRe is a living software project. The design phase of XploRe 4.0 has recently been started. However an implementation will likely not be available before mid of 1994. The new features of XploRe 4.0 will be:

- "optimized" macro language (faster execution)
- integrated in graphical user interfaces (e.g. Windows 3.1)
- more debugging and editor tools
- more graphical interaction
- either running on PCs or SPARCs (using X/Windows)

7 Literature

XploRe 2.0 - a computing environment for eXploratory Regression and data analysis. *The Economics Journal*, Vol. 100, 1401–1403

Hilbe, J. Generalized Additive Models Software. *The American Statistician*, Vol. 47, No. 1, 59–64

Lee, D.K.C. N-Kernel and XploRe. *Journal of Economic Surveys*, Vol. 6, No. 1, 89–103

Ng, P.T., Sickles, R.C. 'XploRe'-ing the world of Nonparametric Analysis. *Journal of Applied Econometrics*, Vol. 5, 293–298

Härdle, W. and Kötter, Th. (1994) XploRe - An Interactive Computing Environment.

ISI meeting in Beijing, proceedings, IP10.2, 375-394

Nonparametric Time Series Analysis, a selective review with examples

Wolfgang Härdle
Institut für Statistik und Ökonometrie
Humboldt-Universität zu Berlin
10178 Berlin
Germany

Rong Chen
Department of Statistics
Texas A& M University
College Station, TX 77843
U.S.A.

Abstract

Nonlinear time series analysis has drawn much attention recently and has shown to be the appropriate tool in many fields, in particular in financial time series analysis. Following the principle of 'letting data speak for themselves,' researchers have developed nonparametric models for nonlinear time series. This article gives a survey on these nonparametric procedures in time series analysis. We also report on applications on the analysis of several real data including gold prices and foreign exchange rates.

Résumé

L'analyse des séries temporelles non linéaires a reçu beaucoup d'attention dans les dernières années. Les modèles non linéaires sont utilisés, notamment, dans l'analyse financiére. Cet article présente un survey des procédures non paramétriques en analyse des séries temporelles. Nous l'illustrons au moyen d'exemples portant sur l'analyse de séries du prix de l'or et de séries de taux de change.

Härdle, W. and Chen, R. (1995) Nonparametric Time Series Analysis,
a selective review with examples.

ISI meeting in Beijing, proceedings, IP10.2, 375-394

1 Introduction

Nonparametric smoothing techniques have been first studied in spectral density estimation. The major thrust in theoretical results came in the last ten to fifteen years, fueled by easier computing environments. Research in the nonparametric area has been concentrated on independent observations and researchers have long waited to extend these techniques to dependent observations and time series. Especially in financial markets, nonlinear and nonparametric time series analysis is useful in order to overcome limitations of the autoregressive moving-average models with constant volatility. Readers are referred to Tong (1990) and Priestley (1988) for details on nonlinear time series analysis. Tjøstheim (1994) gives an excellent review on recent developments in nonlinear time series analysis.

In this article, we review nonparametric model building procedures in time series analysis. Particularly we focus on nonlinear autoregressive models which assume the form of

$$X_t = f(X_{t-1}, \ldots, X_{t-p}) + \varepsilon_t, \quad t = 1, 2, \ldots, \tag{1}$$

where $\{\varepsilon_t\}$ is a sequence of i.i.d. random variables. Typically, the random shock ε_t is independent of X_s, for $s < t$. This model can be extended to nonlinear autoregressive conditional heteroscedastic models of the form

$$X_t = f(X_{t-1}, \ldots, X_{t-p}) + g(X_{t-1}, \cdots, X_{t-p})\varepsilon_t, \quad t = 1, 2, \ldots, \tag{2}$$

which is of particular interest in financial time series analysis, see Gourieroux and Monfort (1992) and Masry and Tjøstheim (1992). There are two different approaches to applying these models.

The first approach is to formulate parametric models for the mean functions $f(\cdot)$ and the volatility function $g(\cdot)$. Often this can be done based on the physical dynamic background and other substantive information of the data. Many models of this form have shown to be successful. The most common ones are the 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 bilinear models of Granger and Anderson (1978), Subba Rao (1981) and Subba Rao and Gabr (1980), the random coefficient models of Nicholls and Quinn (1982), the autoregressive conditional heteroscedastic (ARCH) models of Engle (1982) and the generalized ARCH models of Bollerslev (1986) and Bera and Higgins (1993). Many other related references can be found in Tong (1990) and Priestley (1988).

The second approach is to use nonparametric techniques to estimate the unknown functions f and g. Following the principle of 'letting the data speak for themselves,' this approach avoids the subjectivity of choosing a specific parametric model for a time series. Based on the estimated nonparametric functions, one can either make inference directly or formulate reasonable parametric functions, and hence, build a parameterized nonlinear model for the process. This approach only became practical in the recent years, attributable to powerful computers and easy-to-use interactive statistical and graphical softwares such as S (Becker, Chamber and Wilks, 1988) and XploRe (Härdle, Klinke and Turlach, 1995).

Since our emphasis is on model building related procedures, we present here a selective review of the literature on nonparametric time series analysis. We apologize for any omission of other relevant work in this area, especially on probabilistic aspects. Complimentary references can be found in Györfi, Härdle, Sarda and Vieu (1989), Tjøstheim (1994) and Hart (1994a).

Härdle, W. and Chen, R. (1995) Nonparametric Time Series Analysis,
a selective review with examples.

ISI meeting in Beijing, proceedings, IP10.2, 375-394

In the next section, we review some nonparametric approaches to nonlinear time series, mostly focusing on additive modeling. Practical implementation issues are discussed by analyzing real data. We give examples on river flow, chickenpox data, gold prices and foreign exchange rates. Section 3 is devoted to related nonlinear time series analysis methods which use nonparametric smoothing tools.

2 Nonparametric Approaches

2.1 The Whitening by Windowing Principle

Many nonparametric techniques have been developed under independent observations. For example, with independent random sample $X_1, \cdots, X_n$, a popular method of estimating the density function $f(x)$ is the kernel estimator

$$\hat{f}_h(x) = \frac{1}{nh} \sum_{i=1}^{n} K\left(\frac{x - X_i}{h}\right) \tag{3}$$

where $K(\cdot)$ is a kernel function, typically with finite support and $h > 0$, the so-called bandwidth. 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. That is, when the estimator is used on 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 a small window *almost independent*, then most of the techniques developed for independent observations apply in this situation. Hart (1994a) calls this feature *the whitening by windowing principle.* Various *mixing* conditions are commonly used for proving asymptotic properties of the smoothing techniques for dependent data. Basically these conditions try to control the dependency between X_i and X_j as the time distance $i - j$ increases. For example, a sequence is called 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 $X_i, \cdots, X_j$. A stronger condition is the ϕ-mixing (uniformly mixing) conditions (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. The rate at which α_k and ϕ_k go to zero plays an important role in showing asymptotic behavior of the nonparametric smoothing 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 imply strong mixing conditions. There are developed techniques in checking the geometric ergodicity, see Tweedie (1975), Tjøstheim (1990), Pham (1985) and Diebolt and Guegan (1990).

2.2 Nonparametric Model Building Procedures

In this section we list some common nonparametric approaches to inference the functions $f(\cdot)$ and $g(\cdot)$ in (2). In what follows, we mainly focus on the nonparametric time series techniques,

Härdle, W. and Chen, R. (1995) Nonparametric Time Series Analysis, a selective review with examples.

which are very general and straight forward. This approach, however, suffers from the 'curse of dimensionality.' To overcome this difficulty, researchers have proposed restrictions on the functions f and g. Common restrictions are additive of single index type and/or introduce functional-coefficients in a linear model. These approaches have better convergence rate and are easier to interpret, especially with graphics support from interactive statistical environments.

Local conditional mean (median) approach

Consider the general nonlinear AR(p) process $X_t = f(X_{t-1}, \ldots, X_{t-p}) + \varepsilon_t$. Let $Y_t = (X_{t-1}, \ldots, X_{t-p})$, and choose $\delta_n > 0$. For any $y \in I\!R^p$, let $I_n(y) = \{i : 1 < i < n \text{ and } ||Y_i - y|| < \delta_n\}$ and $N_n(y) = \#I_n(x)$. The conditional mean function estimator is given by $\hat{f}_n(y) = \{N_n(y)\}^{-1} \sum_{i \in I_n(y)} X_i$ and the local conditional median estimator is given by $\tilde{f}(y) = median\{X_i,\ i \in I_n(y)\}$. Under a strong mixing condition, Truong (1993) provides the strong consistency and asymptotic normality of the estimator, along with the optimal rate of convergence.

Nonparametric kernel estimation approach

Robinson (1983), Auestad and Tjøstheim (1990), Härdle and Vieu (1992), and others used a kernel estimator (or robustified versions of it) to estimate the conditional mean and variance under model (2). The function f is estimated by the Nadaraya–Watson estimator with product kernels:

$$\hat{f}(y_1, \ldots, y_p) = \frac{\sum_{t=p+1}^{n} \prod_{i=1}^{p} K\{(y_i - X_{t-i})/h_i\} X_t}{\sum_{t=p+1}^{n} \prod_{i=1}^{p} K\{(y_i - X_{t-i})/h_i\}}, \tag{4}$$

and the conditional variance g^2 is estimated by

$$\hat{g}^2(y_1, \ldots, y_p) = \frac{\sum_{t=p+1}^{n} \prod_{i=1}^{p} K\{(y_i - X_{t-i})/h_i\} X_t^2}{\sum_{t=p+1}^{n} \prod_{i=1}^{p} K\{(y_i - X_{t-i})/h_i\}} - \{\hat{f}(y_1, \ldots, y_p)\}^2, \tag{5}$$

where $K(\cdot)$ is a kernel function with bounded support and the h_i's are the bandwidths.

Robinson (1983), Singh and Ullah (1985) and Masry and Tjøstheim (1992) show strong consistency and asymptotic normality for α-mixing observations. Bierens (1983, 1987) and Collomb and Härdle (1986) proved the uniform consistency of the estimator under the assumption of a ϕ-mixing process.

Härdle and Vieu (1992) applied the method to a gold price series, from 1978 to May 1986 ($n = 2041$). In figure 1, the returns $r_t = (x_t - x_{t-1})/x_{t-1}$ are plotted against the prices x_{t-1}. The model $r_t = f(x_{t-1}) + g(x_{t-1})\epsilon_t$ is estimated and the resulting plots for the conditional mean and variance are shown in figure 3 and 4, respectively. The bandwidths h were selected using the cross validation technique of Härdle and Vieu (1992). The cross validation function for estimating f in (4) is shown in figure 2. The cross validation function for the first term in (5) has its minimum at $h = 0.31$. (not shown). All computation are done in XploRe, using the WARPing technique (Härdle, Klinke, Turlach, 1995).

Local polynomial regression approach

Tsybakov (1986) and Härdle and Tsybakov (1994) used local polynomial nonparametric regression techniques to estimate the conditional mean and conditional variance in time

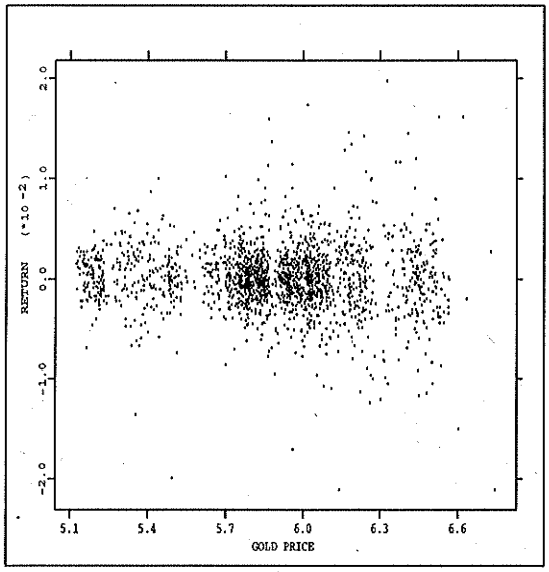


Figure 1: Gold price returns from 1978 to May 1986

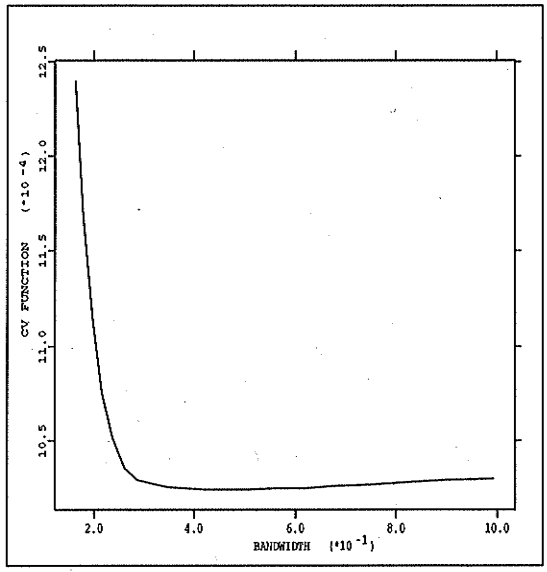


Figure 2: CV function for the conditional mean problem, $h_{min} = 0.45$

series. They considered the model $Y_t = f(Y_{t-1}) + g(Y_{t-1})\varepsilon_t$ where ε_t has mean 0 and variance 1. The functions f and g are estimated by minimization of

$$c_n(x) = arg\min_{c \in \mathbb{R}^l} \sum_{t=1}^{n} (Y_t - c^T U_{tn})^2 K\{(Y_{t-1} - x)/h_n\}$$

and

$$s_n(x) = arg\min_{s \in \mathbb{R}^l} \sum_{t=1}^{n} (Y_t^2 - s^T U_{tn})^2 K\{(Y_{t-1} - x)/h_n\}$$

where K is a kernel function, h_n is a positive bandwidth, and

$$U_{tn} = F(u_{tn}), \quad F(u) = (1\ u\ \cdots u^{l-1}/(l-1)!\,)^T, \quad u_{tn} = (Y_{t-1} - x)/h_n.$$

The estimators $\hat{f}(x)$ and $\hat{g}(x)$ are given by

$$\hat{f}(x) = c_n(x)^T F(0) \quad \text{and} \quad \hat{g}(x) = s_n(x)^T F(0) - \{c_n(x)^T F(0)\}^2$$

They proved asymptotic normality of these estimators under conditions satisfying the assumptions of Tweedie (1975) and Diebolt and Guegan (1990).

This procedure was applied to the YEN/DM exchange rate series from Oct. 1, 1992 to Sept. 30, 1993 with 23814 observations. The series is obtained by taking averages of the spot rate, defined as $(\ln A_t + \ln B_t)/2$ where A_t and B_t are ask- and bid- quotes, respectively, within non-overlap 20-minute window, adjusted for activity. The series is shown in figure 5. The conditional mean and variance functions f and g are estimated by the local polynomial method. We display the estimate of the function g in figure 6. This result is interesting since it shows that the volatility is increasing in high/low return situations. The increasing behavior at the extreme horizontal scale is due to boundary effects.

2.3 Nonlinear Additive AR Models

A nonlinear additive autoregressive (NAAR) model is defined as

$$X_t = f_1(X_{t-1}) + f_2(X_{t-2}) + \cdots + f_p(X_{t-p}) \tag{6}$$

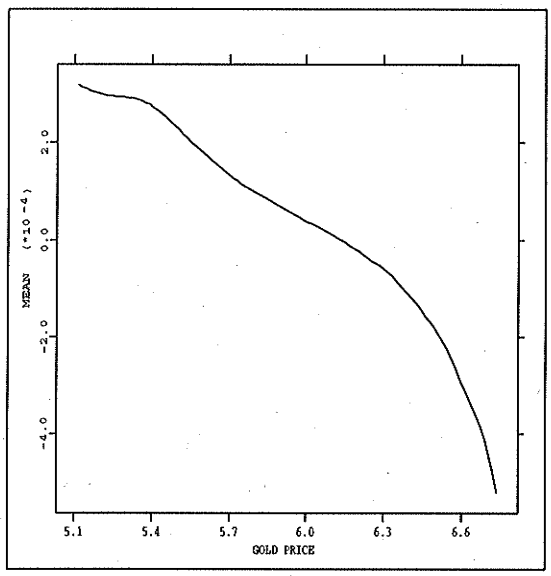


Figure 3: Conditional mean of gold prices returns

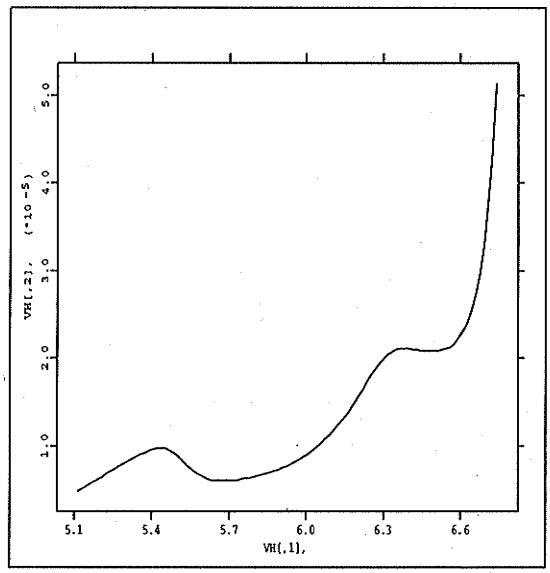


Figure 4: Conditional variance of gold prices returns

These models have been studied extensively in the regression context by Hastie and Tibshirani (1990). It is a generalization of the first-order nonlinear AR model of Jones (1978) and is very flexible as it encompasses linear AR models and many interesting nonlinear models as special cases. The models naturally generalize the linear regression models and allow interpretation of marginal changes i.e. the effect of one variable on the mean function. They are also interesting from a theoretical point of view since they combine flexible nonparametric modeling of many variables with statistical precision that is typical for just one explanatory variable. Note that the NAAR model can be easily extended to include exogenous variables.

Backfitting Algorithms

Chen and Tsay (1993a) used backfitting algorithms such as the Alternating Conditional Expectation (ACE) algorithm of Breiman and Friedman (1985) and the BRUTO algorithm of Hastie and Tibshirani (1990) to fit the additive model (6). 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 i we have $f_i(X_i) = E\{Y - \sum_{j \neq i} f_j(X_j) \mid X_i\}$. Consequently, we can treat $Y - \sum_{j \neq i} f(X_j)$ as the conditional response variable and use nonparametric smoothers to estimate f_i. In practice, all f_i's are unknown so that the estimates are iterated until they all converge. The effective hat matrix of this algorithm is computed in Härdle and Hall (1993), showing that the iteration results depend on the starting index.

The ACE algorithm has been applied to the riverflow data of the River Jokulsa Eystri in Iceland. This is a multiple time series data set, consisting of daily riverflow (Y_t), precipitation (Z_t), and temperature (X_t) from January 1, 1972, to December 31, 1974 ($n = 1096$). For further information see Tong (1990), who used threshold autoregressive models. The time series are plotted in figure 7. A procedure similar to the best subset regression is suggested by Chen and Tsay (1993a) to select the lag variables in the model. They found $\{Y_{t-1}, Y_{t-2}, Z_t, Z_{t-1}, X_{t-1}, X_{t-3}\}$ to be an appropriate explanatory set for the response variable Y_t. The transformations $f_1(Y_{t-1}), f_2(Y_{t-2}), f_3(Z_t), f_4(Z_{t-1}), f_5(X_{t-1}), f_6(X_{t-3})$ are shown in figure 8. Linear functions are suggested for the precipitation and piecewise linear functions for the

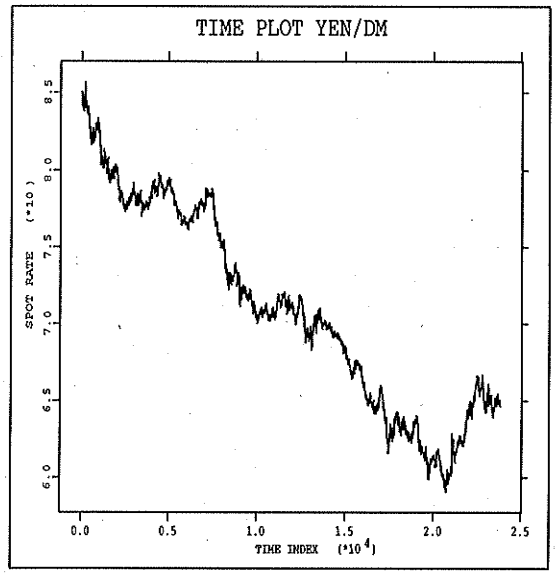


Figure 5: Time plot of YEN/DM exchange rate

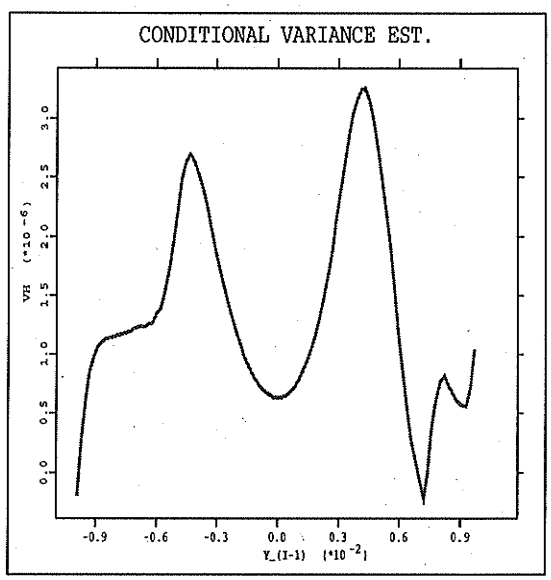


Figure 6: Conditional variance of YEN/DM exchange rate

lagged riverflow and temperature variables. In comparison to Tong's threshold model, the obtained model improves out-of-sample forecasts and is preferred by the AIC criterion.

Projection Estimator

One of the problems associated with the backfitting algorithms is that with highly correlated observations, the convergence can be slow, as noted in Chen and Tsay (1993a). Linton and Nielson (1994) and Chen and Härdle (1994) proposed a projection estimator for estimating the functions in additive regression models without using backfitting. At the same time, Tjøstheim and Auestad (1994a) and Masry and Tjøstheim (1994) proposed the same estimator for NAAR models. Specifically, the 'projection idea' is based on the following observation. If the model is of the additive form (6), and $m(x_1, \cdots, x_p) = c + \sum_{j=1}^{p} f_j(x_j)$ is the mean function, and $p_{-j}(\cdot)$ is the joint density of $X_1, \cdots, X_{j-1}, X_{j+1}, \cdots, X_d$, then for a fixed $x \in I\!R$,

$$f_j(x) + c = \int m(x_1, \cdots, x, \cdots, x_p) p_{-j}(x_1, \cdots, x_p) \prod_{s \neq j} dx_s,$$

provided $E_{X_s} f_s(X_s) = 0$, $s = 1, \cdots, p$. Using the Nadaraya-Watson estimator to estimate the mean function $m(\cdot)$, 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 with finite support. For $1 \leq j \leq p$ and any x in the domain of $f_j(\cdot)$, define, for $h_n > 0$, $h_n' > 0$,

$$\begin{aligned}
\hat{f}_j(x) &= \frac{1}{n} \sum_{i=1}^{n} \hat{m}(X_{i1}, \cdots, X_{i(j-1)}, x, X_{i(j+1)}, \cdots, X_{ip}) \\
&= \frac{1}{n} \sum_{i=1}^{n} \left[\frac{\sum_{l=1}^{n} [\prod_{s \neq j} K_{h_n'}(X_{ls} - X_{is})] K_{h_n}(X_{lj} - x) Y_l}{\sum_{t=1}^{n} [\prod_{s \neq j} K_{h_n'}(X_{ts} - X_{is})] K_{h_n}(X_{tj} - x)} \right] \\
&= \frac{1}{n} \sum_{l=1}^{n} K_{h_n}(X_{lj} - x) Y_l \left[\sum_{i=1}^{n} \frac{\prod_{s \neq j} K_{h_n'}(X_{ts} - X_{is})}{\sum_{t=1}^{n} [\prod_{s \neq j} K_{h_n'}(X_{ts} - X_{is})] K_{h_n}(X_{tj} - x)} \right]
\end{aligned}$$

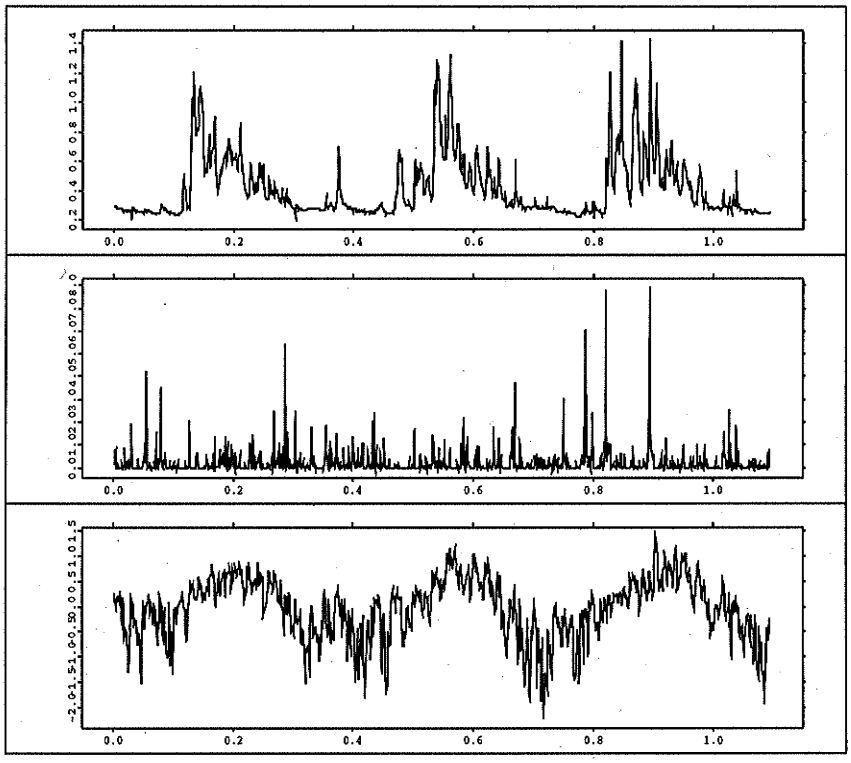


Figure 7: Time plots of riverflow data. (a) Daily riverflow $Y_t(m^3/s)$. (b) Daily precipitation Z_t (mm/day). (c) Daily temperature X_t (C^o).

$$\equiv \frac{1}{n}\sum_{l=1}^{n} K_{h_n}(X_{lj} - x) Y_l w(l, j, x), \tag{7}$$

where

$$w(l, j, x) = \sum_{i=1}^{n} \frac{\prod_{s \neq j} K_{h'_n}(X_{ls} - X_{is})}{\sum_{t=1}^{n}[\prod_{s \neq j} K_{h'_n}(X_{ts} - X_{is})] K_{h_n}(X_{tj} - x)}.$$

Note that, under proper conditions, $1/w(l, t, x)$ converges to $p_j(x \mid X_{l1}, \cdots, X_{ld})$, the conditional density of X_{lj} given X_{l1}, $\cdots$, $X_{l(j-1)}$, $X_{l(j+1)}$, $\cdots$, X_{ld} evaluated at x.

The asymptotic normality of the estimator was established by Chen and Härdle (1994) under independent observations and by Masry and Tjøstheim (1994) under a strong mixing condition. The rate of convergence for estimating $m(\cdot)$ is $n^{2/5}$ typical for regression smoothing with just one explanatory variable. Hence the estimator does not suffer from the 'curse of dimensionality.'

Spline Estimator

Wong and Kohn (1994) used spline nonparametric regression to estimate the components of an NAAR model. They adopted an equivalent Bayesian formation of the spline smoothing and used Gibbs sampler to estimate the components and the parameters of the model, through Monte Carlo simulation of the posterior distributions. The procedure essentially belongs to the backfitting family, but is shown to provide a truly $O(n)$ algorithm, where n is the sample size.

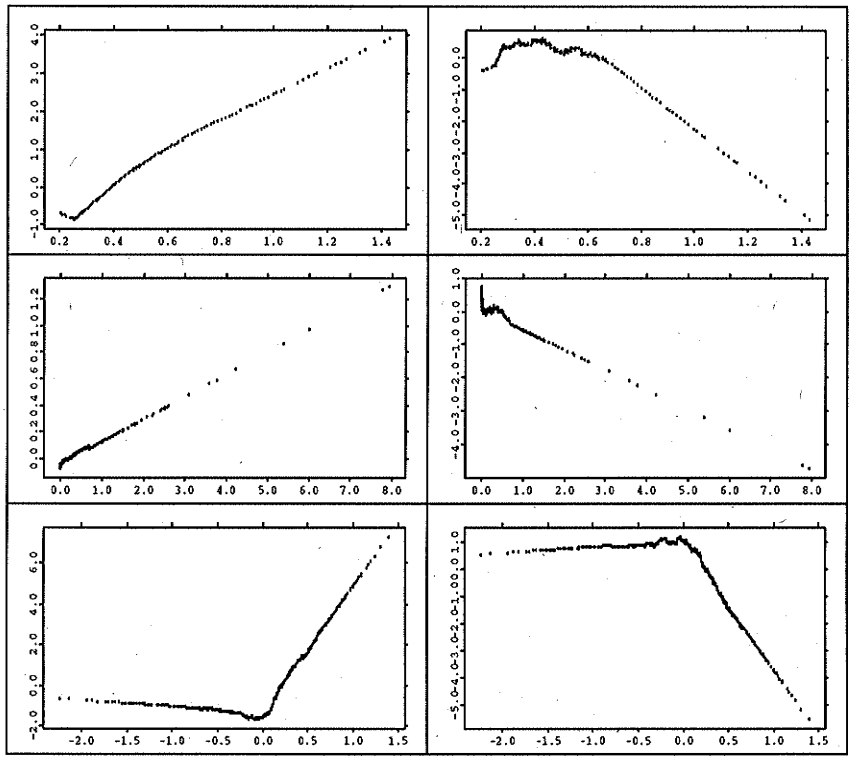


Figure 8: Results of the ACE algorithm for the riverflow data. The plots show the suggested transformations. First row: Y_{t-1} and Y_{t-2}; second row: Z_t and Z_{t-1}; third row: X_{t-1} and X_{t-3}.

Functional coefficient AR model approach

A functional coefficient autoregressive (FAR) model can be written as

$$X_t = f_1(X_{t-d})X_{t-1} + f_2(X_{t-d})X_{t-2} + \ldots + f_p(X_{t-d})X_{t-p} + \varepsilon_t.$$

The model generalizes the linear AR models by allowing the coefficients to change according to a threshold lag variable X_{t-d}. The model is general enough to include the TAR models (when the coefficient functions are step functions) and the EXPAR models (when the coefficient functions are exponential functions) along with many other models (e.g., the STAR models and Sine function models). Chen and Tsay (1993b) use an arranged local regression procedure to roughly identify the nonlinear function forms. For $y \in I\!R$ and $\delta_n > 0$, let $I_n(y) = \{t : 1 < t < n, |X_{t-d} - y| < \delta_n\}$. If we regress X_t on $X_{t-1}, \ldots, X_{t-p}$ using all the observations X_t such that $t \in I_n(y)$, then the estimated coefficients can be used as estimates of $f_i(y)$. One can then make inference directly or formulate parametric models based on the estimated nonlinear function forms. Chen and Tsay (1993b) proved the consistency of the estimator under geometric ergodicity conditions. Note that the locally weighted regression of Cleveland and Devlin (1988) can also be used here as well.

For illustration of the ALR procedure, we consider the chickenpox data used by Chen and Tsay (1993b) and described by Sugihara and May (1990) with 533 observations. Natural logarithms are taken for variance stabilization. In the implementation in XploRe, we require the sample size within each window to be at least K $(> p)$ to ensure the accuracy of the coefficient estimates. Lacking an optimal selection criterion, we select the structure parameters (d, p, K) by $d = 1, \ldots, d_{max}$, $p = 1, \ldots, p_{max}$, $K = 3p, \ldots, n_{min}$. Several

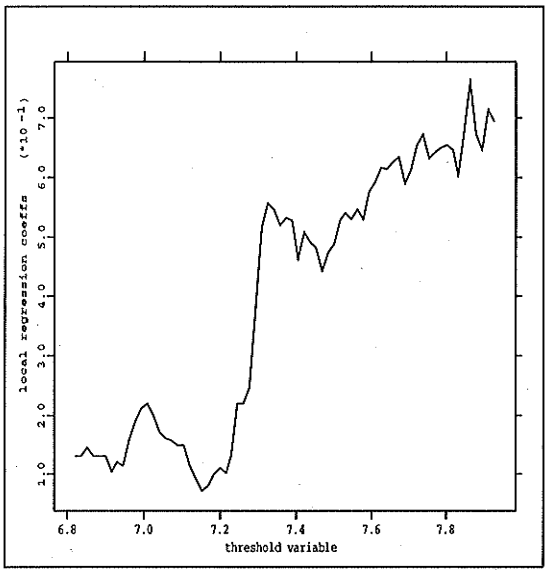


Figure 9: Local estimates of $f_1(x)$

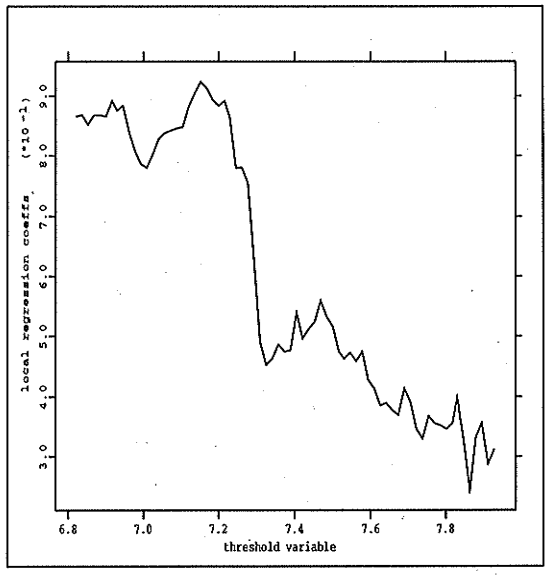


Figure 10: Local estimates of $f_2(x)$

nonlinearity tests indicate strong nonlinearity for the threshold lag $d = 12$, which is plausible because we have monthly data. The most significant lags are 1 and 24. The resulting model is

$$X_t = f_1(X_{t-12})X_{t-1} + f_2(X_{t-12})X_{t-24} + \varepsilon_t.$$

The scatterplots of the estimated functions are shown in figures 9 and 10, respectively. To formulate a parametric model based on the estimated functions, we note a level shift around the value $X_{t-12} = 7.2$. Hence a TAR model is suggested, for details see Chen and Tsay (1993b).

Adaptive spline threshold AR model approach

Lewis and Stevens (1991) proposed the adaptive spline threshold autoregressive (ASTAR) model with the form $X_t = \sum_{j=1}^{s} c_j K_j(X) + \varepsilon_t$, 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 $(X_{t-1}, \cdots, X_{t-p})$. For example,

$$X_t = c_1 + c_2 X_{t-1} + c_3(a_1 - X_{t-5})_+ + c_4 X_{t-1}(X_{t-3} - a_2)_+(a_3 - X_{t-4})_+ + \varepsilon_t,$$

where $u_+ = u$ if $u > 0$ and $u_+ = 0$ if $u \leq 0$, is an ASTAR model.

The modeling and estimation procedures follow the Multivariate Adaptive Regression Splines (MARS) algorithm by Friedman (1988). It is basically a regression tree procedure using truncated regression splines.

Hermite expansion approach

Gallant and Tauchen (1990) used Hermite expansion to approximate the nonlinear one-step-ahead conditional density of the process given its past. Letting z denote an M-vector, the particular Hermite expansion employed has the form $h(z) \propto [P(z)]^2\phi(z)$, where $P(z)$ denotes a multivariate polynomial of degree K_z and $\phi(z)$ denotes the density function of the (multivariate) Gaussian distribution with mean zero and the identity matrix as its covariance

matrix. The model is fitted using maximum likelihood procedures on a truncated expansion together with a model selection strategy that determines the truncation point K_z. Note that we can view the truncation point as the smoothing parameter.

2.4 Implementation Issues

One of the important implementation issues of the nonparametric smoothing tools is the bandwidth selection method in finite sample. There have been 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, 1986), Park and Marron (1990) and Park and Turlach (1992).

One of the usual criteria for selecting the bandwidth is the averaged squared error

$$d_A(h) = \frac{1}{n}\sum_{i=1}^{n}\{m(X_i) - \hat{m}_h(X_i)\}^2 w(X_i).$$

which is an approximation of the integrated squared error

$$d_I(h) = \int \{\hat{m}(x) - m(x)\}^2 f(x) w(x) dx$$

Note that the criteria still involve the unknown function $m(\cdot)$. Hence it has to be estimated.

For the nonparametric kernel estimator, Härdle and Vieu (1992), Härdle (1990) proposed to use the leave-out cross-validation function

$$CV(h) = \frac{1}{n}\sum_{i=1}^{n}\{Y_i - \hat{m}_{h,i}(X_i)\}^2 w(X_i),$$

where

$$\hat{m}_{h,i}(x) = \frac{n^{-1}\sum_{j\neq i} K\{(x - X_j)/h\}Y_j}{n^{-1}\sum_{j\neq i} K\{(x - X_j)/h\}},$$

to select the bandwidth. Let $\hat{h}$ be the bandwidth that minimizes $CV(h)$. They proved that, under a 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 and Vieu (1990). These bandwidth selection methods are efficiently implemented in the XploRe smoother library, see Härdle, Klinke and Turlach (1995).

3 Tests and Prediction with Nonparametric Techniques

3.1 A Nonparametric Nonlinearity Test

Hjellvik and Tjøstheim (1994) proposed a nonlinearity test 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 index 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.

3.2 Additivity Tests

Additivity is commonly used in the statistical literature to simplify data analysis, especially in analysis of variance and in multivariate smoothing. Chen, Liu and Tsay (1994) proposed three nonparametric procedures for testing additivity in nonlinear time series analysis.

The first procedure combines some smoothing techniques with analysis of variance. First, a shrunken range, $\delta(y_{max} - y_{min})$ is partitioned into m equal intervals, (a_i, a_{i+1}) for $i = 0, \cdots, (m-1)$ where $a_i = y_{min} + (1-\delta)(y_{max} - y_{min})/2 + i\delta(y_{max} - y_{min})/m$ and $\delta \in (0,1)$ is a shrinking factor. This avoids the complication of the "boundary effect" often encountered in nonparametric smoothing procedures. For $t = 3, \cdots, n$, an observation Y_t is classified into the (i,j)th cell if $Y_{t-1} \in (a_{i-1}, a_i)$ and $Y_{t-2} \in (a_{j-1}, a_j)$ and is denoted by X_{ijk} where k is used to distinguish different observations in the same cell. If Y_{t-1} or Y_{t-2} is outside the shrunken range, Y_t is dropped from further consideration. Finally, an unbalanced two-way analysis of variance procedure is carried out to obtain an F statistic for testing the null hypothesis $H_0 : f_{ij} = 0$ for all i and j in the model $X_{ijk} = \mu + \alpha_i + \beta_j + f_{ij} + \epsilon_{ijk}$, where f_{ij} denotes a non-additive function.

The second is a Lagrange multiplier test using nonparametric estimation. It consists of three steps.

1. An additive model $Y_t = f_1(Y_{t-k_1}) + \cdots + f_p(Y_{t-k_p}) + \varepsilon_t$ is estimated using the ACE algorithm with the restriction that the response variable can only be linearly transformed. Denote the estimates of $f_i(\cdot)$ by $\hat{f}_i(\cdot)$ and the residuals by $\hat{\varepsilon}_t = Y_t - \sum_{i=1}^{p} \hat{f}_i(Y_{t-k_i})$.

2. Regress the cross-product terms $Y_{t-k_{j_1}} Y_{t-k_{j_2}}$ on $Y_{t-k_1}, \cdots, Y_{t-k_p}$ for $1 \le j_1 < j_2 \le p$, and the third-order cross-product terms $Y_{t-k_{j_1}} Y_{t-k_{j_2}} Y_{t-k_{j_3}}$ on $Y_{t-k_1}, \cdots, Y_{t-k_p}$ for $1 \le j_1 \le j_2 \le j_3 \le p$ except for $j_1 = j_2 = j_3$ using the ACE algorithm. This procedure results in $K = p(p-1)/2 + \{p(p+1)(p+2)/6 - p\} = (p-1)p(p+7)/6$ residual series, say, $e_1(t), \cdots, e_K(t)$. Here the transformations of the response variables are also restricted to be linear.

3. Linearly regress the residual series $\hat{\varepsilon}_t$ obtained from Step 1 on $e_1(t), \cdots, e_K(t)$ obtained from Step 2. Compute the test statistic nR^2 where n is the sample size and R^2 is the conventional coefficient of determination in linear regression analysis.

The third is a permutation test which uses smoothing techniques to obtain the test statistic and its reference distribution.

1. As for Step 1 of the Lagrange multiplier test.

2. Regress the estimated residuals $\hat{\varepsilon}_t$ from Step 1 on the cross-product terms $Y_{t-k_i} Y_{t-k_j}$ for $1 \le i < j \le p$ using the ACE algorithm and obtain the sum of squares of residuals of this regression.

3. Form a new series of residuals $e(t)$ by permuting the $\hat{\varepsilon}_t$. Regress it on the same cross-product terms as those of Step 2 using the ACE algorithm and obtain the sum of squares of residuals. Repeat this step N times.

4. The p-value of the permutation test is determined by the proportion of the sum of squares of residuals in Step 3 that is smaller than sum of squares of residuals in Step 2.

Theoretical justification and simulation evidence of the tests are given in Chen, Liu and Tsay (1994).

3.3 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 lag variables and exogenous variables is too big and we wish to select those most significant components.

For linear time series models, lag selection and order determination are usually done using information criterion such as FPE, AIC and BIC (Akaike 1970, 1974, 1979), along with other model checking procedures such as residual analysis.

In fully nonparametric approach to time series analysis, Auestad and Tjøstheim (1990) and Tjøstheim and Auestad (1994b) proposed to use the FPE criterion and Cheng and Tong (1992) proposed to use the cross validation criterion.

More specifically, Tjøstheim and 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 (1). Let $Y_t, t = 1, \cdots, N$ be a stationary strong mixing nonlinear AR process. Let $\boldsymbol{i} = (i_1, \cdots, i_p)$ and $\boldsymbol{X}_t(\boldsymbol{i}) = (Y_{t-i_1}, \cdots, Y_{t-i_p})'$. Define

$$\hat{FPE}(\boldsymbol{i}) = \frac{1}{n}\sum_t [Y_t - \hat{f}\{\boldsymbol{X}_t(\boldsymbol{i})\}]^2 w\{\boldsymbol{X}_t(\boldsymbol{i})\} \frac{1 + (nh^p)^{-1} J^p B_p}{1 - (nh^p)^{-1}\{2K^p(0) - J^p\} B_p} \tag{8}$$

where

$$J = \int K^2(x)dx, \quad B_p = n^{-1}\sum_t \frac{w^2\{\boldsymbol{X}_t(\boldsymbol{i})\}}{\hat{p}\{\boldsymbol{X}_t(\boldsymbol{i})\}}$$

and $\hat{f}(\boldsymbol{X}_t(\boldsymbol{i}))$ is the kernel conditional mean estimator in (4). Note that the $\hat{FPE}$ is essentially a penalized sum of squares of residuals, where the last term in (8) penalizes small bandwidth h and large order p.

Cheng and Tong (1992) used leave-one-out cross validation procedure to select the order of a general nonlinear AR model. Let $\boldsymbol{X}_t(d) = (Y_{t-1}, \cdots, Y_{t-d})$ and

$$CV(d) = \frac{1}{N - r + 1}\sum_t \{Y_t - \hat{f}_{-t}(\boldsymbol{X}_t(d)\}^2 W\{\boldsymbol{X}_t(d)\}$$

where $\hat{f}_{-t}$ is the kernel conditional mean estimator with Y_t deleted. They proved that, under some regularity conditions,

$$CV(d) = RSS(d)\{1 + 2K(0)\gamma h^{-d}/N + o_p(1/h^d N)\}$$

where $\gamma = \int W(x)dx / \int W(x)f(x)dx$ and h is the bandwidth. Again, we can view this as a penalized sum of squares of residuals.

In additive model approach, using the ACE algorithm, a procedure similar to the best subset regression is suggested by Chen and Tsay (1993a) to select the lag variables in the NAAR model. Chen and Härdle (1994) proposed a procedure for selecting the most significant lags in an additive model. This variable selection procedure is based on the size of $S_j = E_{X_j}[f_j^2(X_j)]$ for model (6). Let J be a subset of $\{1, \cdots, p\}$ such that $J = \{j : S_j > s\}$ for some $s > 0$ and assume for $j \notin J$, $S_j = 0$. Define $\hat{S}_j = n^{-1}\sum_{i=1}^n \{\hat{f}_j(X_{ij}) - \bar{f}_j\}^2 = n^{-1}\sum_{i=1}^n \hat{f}_j^2(X_{ij}) - \bar{f}_j^2$, where $\bar{f}_j = n^{-1}\sum_{i=1}^n \hat{f}_j(X_{ij})$ and $\hat{f}_j$'s are estimated using the projection estimator (7). Note that $\hat{S}_j$ estimates the quantity S_j. Hence, a large $\hat{S}_j$ implies that the variable X_j should be included in the model. The variable selection procedure selects the indices j such that $\hat{S}_j > b_n$ where b_n is some prescribed level. Denote $\hat{J} = \{j : \hat{S}_j \geq b_n\}$. Chen and Härdle (1994) proved, with i.i.d. observations, under suitable conditions and as b_n satisfy certain constraints, $P(\hat{J} = J) \to 1$.

3.4 Prediction

Consider the nonlinear AR(1) model $X_t = \phi(X_{t-1}) + \varepsilon_t$. Since the conditional mean $E(X_{t+k} \mid X_t = x)$ is the least squares predictor for k-step ahead prediction, Auestad and Tjøstheim (1990) and Härdle and Vieu (1992) and Härdle (1990) proposed using the ordinary Nadaraya–Watson estimator

$$\hat{m}_{h,k}(x) = \frac{\sum_{t=1}^{n-k} K\{(x - X_t)/h\} X_{t+k}}{\sum_{t=1}^{n-k} K\{(x - X_t)/h\}} \tag{9}$$

to estimate $E(X_{t+k} \mid X_t = x)$ directly.

Note that the variables $X_{t+1}, \ldots, X_{t+k-1}$ consist of substantial information about the conditional mean function $E(X_{t+k} \mid X_t)$. Chen and Hafner (1994) proposed a multistage kernel smoother which utilizes these information. For example, consider two-step ahead forecasting. Due to the Markov property, we have

$$m_{h,2}(x) = E[X_{t+2} \mid X_t = x] = E[E(X_{t+2} \mid X_{t+1}, X_t) \mid X_t = x] = E[E(X_{t+2} \mid X_{t+1}) \mid X_t = x].$$

Define $f(y) = E(X_{t+2} \mid X_{t+1} = y)$. Ideally, if we knew $f(\cdot)$, we would use the pairs $(f(X_{i+1}), X_i), i = 1, \ldots, (n-1)$ to estimate $E(X_{t+2} \mid X_t)$, instead of using the pairs (X_{i+2}, X_i) as the estimator in (9). Note that the error between X_{t+2} and $f(X_{t+1})$ is $O(1)$. Hence, if we can estimate the function $f(\cdot)$ with an estimator $\hat{f}(\cdot)$ that has a smaller error rate and use the pairs $(\hat{f}(X_{i+1}), X_i)$ to estimate $E(X_{t+2} \mid X_t)$, we should achieve a smaller error. This observation motivated the following estimator, which is called 'multistage smoother'. It is defined as

$$\hat{m}_{h_1,h_2}(x) = \frac{\sum_{t=1}^{n-1} K\{(x - X_t)/h_2\} \hat{f}_{h_1}(X_{t+1})}{\sum_{t=1}^{n-1} K\{(x - X_t)/h_2\}} \tag{10}$$

where

$$\hat{f}_{h_1}(y) = \frac{\sum_{j=1}^{n-1} K\{(y - X_j)/h_1\} X_{j+1}}{\sum_{j=1}^{n-1} K\{(y - X_j)/h_1\}}.$$

The new smoother is proved to have a smaller mean squared error.

The estimators in (9) and (10) are applied to the gold price example. Because the returns are very small in absolute value, the figure for a two–step prediction does not look very different from figure 3. For this reason and to show that multi–step prediction is easily adaptable, we computed a ten–step prediction with both estimators. For the multistage smoother we need a recursive algorithm, which computes at the kth step a smoother of the $(k-1)$th smoother, beginning with the simple one–step predictor. At each step the optimal bandwidth according to the cross validation criterion is obtained. The estimates are shown in figures 11 and 12.

Collomb, Härdle and Hassani (1987) proposed to predict future observations based on the mode function $m(x) = \arg\max_y f(y \mid x)$ where $f(y \mid x)$ denotes the conditional density function of Y given X. They estimated the conditional density function from a sequence of ϕ-mixing observations using kernel estimation and showed uniform convergence of the estimator.

3.5 Trend Estimation

Suppose $\{X_1, \cdots, X_n\}$ is a possibly nonstationary time series with trend $\mu(j) = E(X_j)$. Under the assumption that the trend is smooth, a traditional way of estimating the trend function Hannan and Gasser (1971), Nadaraya (1974), Priestley and Chao (1972) proposed

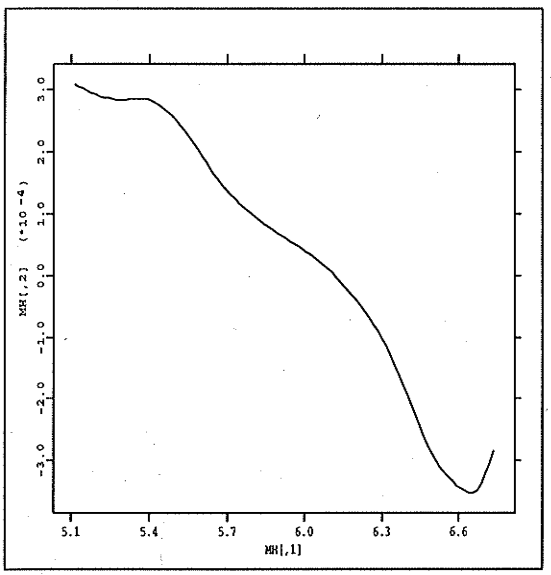


Figure 11: 10–step prediction using the direct Nadaraya–Watson estimator

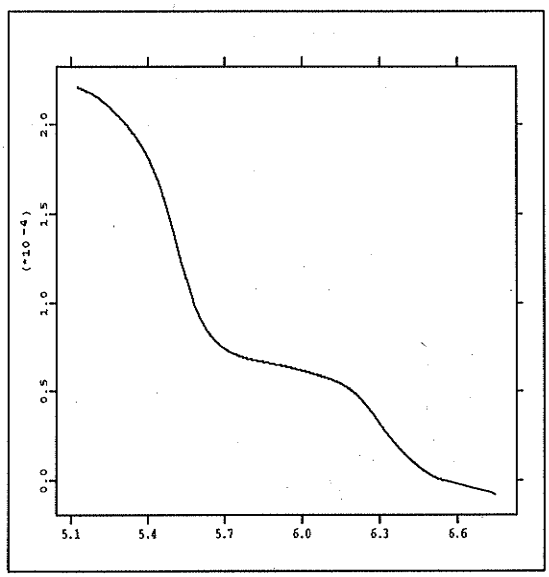


Figure 12: 10–step prediction using the multistage smoother

to use a robust M-estimator to estimate the trend and they showed the consistency of the estimator. Hart (1991) uses the kernel smoother of Gasser-Müller (1979) form

$$\hat{\mu}_{jh} = \frac{1}{h}\sum_{i=1}^{n} X_i \int_{(i-1)/n}^{i/n} K\left(\frac{(j-0.5)/n - \mu}{h}\right) du.$$

Hart (1994b) proposed a method called time series cross-validation for selecting the bandwidth for trend estimation. He noted that the ordinary leave-one cross-validation tends to select a bandwidth many orders of magnitude too small, if the data are highly positively correlated.

3.6 Serial Dependency Test

Skaug and Tjøstheim (1993) proposed a nonparametric test for independency between two variables. This test can be used in checking the residual behavior of an estimated nonlinear time series model. 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 is a weight function with compact support. Using kernel density estimators, we obtain an estimator

$$\hat{I} = \frac{1}{n}\sum\{\hat{p}(X_i,Y_i) - \hat{p}_i(X_i)\hat{p}_i(Y_i)\}^2 w(X_i,Y_i).$$

which, under the null hypothesis that X and Y are independent, should be small. For detailed implementation, see Skaug and Tjøstheim (1993)

3.7 Density Estimation with Correlated Observations

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) where the kernel function $K(\cdot)$ is typically a probability density function. The key in density estimation is the bandwidth selection. There are a number of different methods proposed, including the cross-validation (Rudemo 1982, Bowman 1984) and the plug-in rules of Sheather (1983, 1986), Park and Marron (1990) and Park and Turlach (1992).

The earliest work on density estimation for stationary process is that of Roussas (1969) and Rosenblatt (1970). The properties of the kernel estimator under dependent observations were investigated by Robinson (1983) and Hall and Hart (1990). 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 are studied by Hart and Vieu (1990), under certain regularity conditions. Detailed information and references can be found in Györfi, Härdle, Sarda and Vieu (1989) and Hart (1994a).

Acknowledgment: Research of the first author is supported in part by INRA and INSEE, France. Research of the second author is supported in part by the National Science Foundation under grants DMS-9301193 and by Sonderforschungsbereich 373, "Quantifikation und Simulation Ökonomischer Prozesse".

References

[1] Akaike, H. (1970), "Statistical predictor identification." *Ann. Inst. Statist. Math.*, **22**, 203-217

[2] Akaike, H. (1974), "A New Look at Statistical Model Identification," *IEEE transactions on Automatic Control*, **AC-19**, 716–722.

[3] Akaike, H. (1979), "A Bayesian extension of the minimum AIC procedure of autoregressive model fitting." *Biometrika*, **66**, 237-242

[4] Auestad, B. and Tjøstheim, D. (1990). Identification of nonlinear time series: First order characterization and order estimation, *Biometrika* **77**: 669–687.

[5] Beck, R.A., Chambers, J. M. and Wilks, A. R. (1988), *The New S Language*, Chapman and Hall, New York.

[6] Bierens, H.J. (1983), Uniform consistency of kernel estimators of a regression function under generalized conditions. *J. Am. Statist. Assoc.* , **78**, 699-707

[7] Bierens, H.J. (1987), *Kernel estimators of regression functions*, Cambridge University Press: Advances in Econometrics.

[8] Billingsley, P. (1968) *Convergence of Probability Measures*, New York: Wiley.

[9] Bollerslev, T. (1986). Generalized autoregressive conditional heteroscedasticity, *Journal of Econometrics* **31**: 307–327.

[10] Bowman, A.W. (1994), An alternative method of cross-validation for the smoothing of density estimates. *Biometrika* , **71**, 353-360

[11] Box, G. E. P. and Jenkins, G. M. (1976). *Time Series Analysis: Forecasting and Control*, Holden-Day, San Francisco.

[12] Breiman, L. and Friedman, J. H. (1985). Estimating optimal transformations for multiple regression and correlations (with discussion), *Journal of the American Statistical Association*, **80**, 580–619.

[13] Chan, K.S. and Tong, H (1986), On estimating thresholds in autoregressive models. *J. Time Series Analysis*, **7**, 179-190

[14] Chatfield, C. (1984), *The Analysis of Time Series: An Introduction*, 3rd ed., Chapman and Hall, London

[15] Chen, R. and Hafner, C. (1994). A nonparametric predictor for nonlinear linear time series, *Technical report*, Department of Statistics, Texas A&M University.

[16] Chen, R. and Härdle, W. (1995), "Estimation and Variable Selection in Nonparametric Additive Models," Discussion paper, Institut für Statistik und Ökonometric, Humboldt-Universität zu Berlin, available via ftp 141.20.100.2

[17] Chen, R., Liu, J.S. and Tsay, R.S. (1994), "Additivity Tests for Nonlinear Autoregressive Models," *Biometrika*, in press.

[18] Chen, R. and Tsay, R. S. (1993a). Nonlinear additive ARX models, *Journal of the American Statistical Association* **88**: 955–967.

[19] Chen, R. and Tsay, R. S. (1993b). Functional-coefficient autoregressive models, *Journal of the American Statistical Association* **88**: 298–308.

[20] Cheng, B. and Tong, H. (1992), "On consist non-parametric order determination and chaos (with discussion)," *J. R. Statist. Soc* B **54**, 427-474

[21] Cleveland, W. S. and Devlin, S. J. (1988). Locally weighted regression: An approach to regression analysis by local fitting, *Journal of the American Statistical Association* **83**: 596–610.

[22] Collomb, G and Härdle, W. (1986), Strong uniform convergence rates in robust nonparametric time series analysis and prediction: kernel regression estimation from dependent observations. *Stochastic Process and their Applications*, **23**, 77-89

[23] Collomb, G, Härdle, W. and Hassani, S. (1987), A note on prediction via estimation of the conditional mode function. *J. Statistical Planning and Inference*, **15**, 227-236.

[24] Diebolt, J. and Guegan, D. (1990), Probabilistic properties of the general nonlinear autoregressive process of order one. *Technical report*, N^o 128, L.S.T.A., Universit Paris VI.

[25] Engle, R. F. (1982). Autoregressive conditional heteroscedasticity with estimates of the variance of u.k. inflation, *Econometrica* **50**: 987–1008.

[26] Friedman, J. H. (1988). Multivariate adaptive regression splines (with discussion), *Ann. Statist.*, **19**, 1-141

[27] Gallant, A.R. and Tauchen, G. (1990), A nonparametric approach to nonlinear time series analysis, estimation and simulation. in *IMA volumes on mathematics and its applications*, Ed. Brillinger et al. Springer Verlag, New York

[28] Gasser, T. and Müller, H.G. (1979). Kernel estimation of regression functions. In *Smoothing Techniques for Curve Estimation*, eds. T. Gasser and M. Rosenblatt), 23-68

[29] Gourieroux, C. and Monfort, A. (1992), Qualitative threshold ARCH models, *J. Econometrics*, **52**, 159-199

[30] Granger, C. W. J. and Anderson, A. P. (1978). *An Introduction to Bilinear Time Series Models*, Vandenhoeck & Ruprecht, Göttingen und Zürich.

[31] Granger, C. and Teräsvirta, T. (1992) *Modeling Nonlinear Dynamic Relationships*, Oxford University Press, Oxford

[32] Györfi, L. Härdle, W. Sarda, P. and Vieu, P. (1989) *Nonparametric Curve Estimation from Time Series*. Lecture Notes in Statistics 60. Springer-Verlag, Heidelberg.

[33] Haggan, V. and Ozaki, T. (1981). Modeling nonlinear vibrations using an amplitude-dependent autoregressive time series model, *Biometrika* **68**: 189–196.

[34] Hall, P. and Hart, J.D. (1994), Convergence rates in density estimation for data from infinite-order moving average process. , *Probability Theory and Related Fields*, **87**, 253-274

[35] Härdle, W. (1990), *Applied Nonparametric Regression*, Cambridge University Press: Cambridge

[36] Härdle, W. and Hall, P (1993), On the backfitting algorithm for additive regression models. *Statistica Neerlandica*, **47**, 43-57

[37] Härdle, W., Klinke, S. and Turlach, B. (1995) *XploRe, an interactive statistical computing environment*, Springer Verlag, Heidelberg.

[38] Härdle, W. and Tuan, P.D. (1986), Some theory on M-smoothing of time series, *J. Time Series Analysis*, **7**, 191-204

[39] Härdle, W. and Tsybakov, A.B. (1995), Locally polynomial estimators of volatility function. Discussion paper, Institut für Statistik und Ökonometric, Humboldt-Universität zu Berlin, available via ftp 141.20.100.2

[40] Härdle, W. and Vieu, P. (1992). Kernel regression smoothing of time series, *Journal of Time Series Analysis* **13**: 209–232.

[41] Hart, J.D. (1991), Kernel regression estimation with time series errors. *J. R. Statist. Soc* B , **53**, 173-187

[42] Hart, J.D. (1994a), Smoothing time-dependent data: a survey of data driven methods. *J. Nonparametric Statistics*, in press.

[43] Hart, J.D. (1994b), Automated kernel smoothing of dependent data by using time series cross-validation. *J. R. Statist. Soc* B , **56**, 529-542

[44] Hart, J.D. and Vieu, P. (1990). Data-driven bandwidth choice for density estimation based on dependent data. *Ann. Statist.*, **18**, 1080-1088

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

[46] Hjellvik, V. and Tjøstheim, D. (1994). Nonparametric tests of linearity for time series, *Biometrika* p. to appear.

[47] Jones, D. A. (1978). Non-linear autoregressive processes, *Journal of the Royal Statistical Society, Series A* **360**: 71–95.

[48] Lewis, P. A. W. and Stevens, G. (1991). Nonlinear modeling of time series using multivariate adaptive regression splines (mars), *Journal of the American Statistical Association* **87**: 864–877.

[49] Linton, O. and Nielsen, J.P. (1994), A kernel method of estimating structured nonparametric regression based on marginal integration, *Biometrika*, in press.

[50] Masry, E and Tjøstheim, D. (1992), Non-parametric estimation and identification of ARCH and ARX nonlinear time series: convergence properties and rates. Preprint, Dept. of Mathematics, Univ. of Bergen.

[51] Masry, E and Tjøstheim, D. (1994), "Additive nonlinear ARX time series and projection estimates," preprint, Dept. of Math., Univ. of Bergen.

[52] Nicholls, D.F. and Quinn, B.G. (1982), *Random coefficient autoregressive models: An introduction.* Lecture Notes in Statistics, **Vol. No. 11**, New York: Springer-Verlag.

[53] Park, B.U. and Marron, J.S. (1990), Comparison of data-driven bandwidth selectors. *J. Am. Statist. Assoc.* , **85**, 66-72

[54] Park, B.U. and Turlach, B. (1992), Practical performance of several data driven bandwidth selectors (with discussion), *Computational Statistics*, **7**, 251-270

[55] Pham, D. T. (1985), Bilinear Markovian representations and bilinear models, *Stochastic Process. Appl.*, **20**, 295-306

[56] Priestley, M. B. (1988). *Non-linear and Non-stationary Time Series Analysis*, Academic Press, New York.

[57] Robinson, P. M. (1983). Non-parametric estimation for time series models, *Journal of Time Series Analysis* **4**: 185–208.

[58] Rosenblatt, M. (1970), Density estimation and Markov sequences. In *Nonparametric Techniques in Statistical Inference*, (M.L. Puri, ed.) 199-213. Cambridge University Press.

[59] Roussas, G.G. (1969), Nonparametric estimation in Markov process. *Annals of the Institute of Statistical Mathematics*, **21**. 73-87

[60] Rudemo, M. (1982), Empirical choice of histograms and kernel density estimators. *Scandinavian J. of Statist.*, **9**, 65-78

[61] 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

[62] Sheather, S.J. (1986), An improved data-based algorithm for choosing the window width when estimating the density at a point. *Computational Statistics and Data Analysis*, **4**, 61-65

[63] Skaug, H.J. and Tjøstheim, D. (1993) Non-parametric tests of serial independence. *The M. Priestley Birthday Volume* (ed. T. Subba Rao), pp. 207-229

[64] Singh, R.S. and Ullah, A. (1985) Nonparametric time series estimation of joint DGP, conditional DGP and vector autoregression. *Econometric Theory*, **1**.

[65] Subba Rao, T. (1981). On the theory of bilinear time series models, *Journal of the Royal Statistical Society, Series B* **43**: 244–255.

[66] Subba Rao, T. and Gabr, M. M. (1980). *An introduction to bispectral analysis and bilinear time series models*, Vol. 24 of *Lecture Notes in Statistics*, Springer, New York.

[67] Sugihara, G and May, R. M. (1990). Nonlinear forecasting as a way of distinguishing chaos from measurement error in time series, *Nature*, **344**: 734-741

[68] Tibshirani, R. (1988). Estimating transformations for regression via additivity and variance stabilization, *Journal of the American Statistical Association* **83**: 194–405.

[69] Tjøstheim, D. (1990), "Nonlinear Time Series and Markov Chains," *Advanced Applied Probability*, **22**, 587–611

[70] Tjøstheim, D. (1994), "Nonlinear time series, a selective review," *Scand. J. Statist*, **21**, 97-130.

[71] Tjøstheim, D. and Auestad, B (1994a), Non-parametric identification of non-linear time series: projection. *J. Am. Statist. Assoc.* , in press

[72] Tjøstheim, D. and Auestad, B (1994b), Non-parametric identification of non-linear time series: selecting significant lags. *J. Am. Statist. Assoc.* , in press

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

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

[75] Tong, H. (1990). *Nonlinear Time Series Analysis: A Dynamic Approach*, Oxford University Press, Oxford.

[76] Truong, Y. K. (1993). A nonparametric framework for time series analysis, *New Directions in Time Series Analysis*, Springer, New York.

[77] Tsybakov, A.B. (1986) Robust reconstruction of functions by the local approximation method, *Problems of Information Transmission*, **22**, 133-146

[78] 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.

[79] Wong, C-M and Kohn, R. (1994), A Bayesian approach to estimating and forecasting additive nonparametric autoregressive models, *J. Time Series Analysis*, in press

5. Karlsruher Ökonometrie-Workshop, Editor Bol., Physika Verlag

A New Method for Volatility Estimation with Applications in Foreign Exchange Rate Series

Peter Bossaerts Christian Hafner Wolfgang Härdle *

April 1995

Abstract

The statistical properties of three foreign exchange rate series are analyzed using a redefinition of the time scale to cope with the inherent seasonal heteroskedasticity. A conditional heteroskedastic autoregressive nonlinear (CHARN) model is estimated by local linear regression techniques. The results show significant nonlinearities for the mean function as well as for the variance function.

1 Introduction

The behaviour of foreign exchange (FX) rates has been subject of many recent investigations. This is, of course, partly due to the fact that the market for foreign currencies is by far the largest financial market. A correct understanding of the foreign exchange rate dynamics has important implications for international asset pricing theories, the pricing of contingent claims and policy–oriented questions.

The most important exchange rates to analyze are, of course, the US Dollar, the Japanese Yen and the Deutsche Mark. European cross rates are of limited comparability to the "big" rates because of restrictions in the European Monetary System (EMS), at least before October 1992, when the variability bands were quite narrow.

High frequency financial data analysis is a booming research field. This is due to improved real–time information systems, relatively cheap data supply by institutions such as Olsen & Associates and improved storing facilities. Also, after having found that GARCH(1,1) processes fit daily and weekly FX rates well in most cases, the topic of temporal aggregation (Drost, Nijman (1993)) arose and the question if ARCH-type models still fit high–frequency data. The literature is still very

*First author's affiliation: California Institute of Technology and Tilburg University; Mailing address: CentER, Tilburg University, PO Box 90153, NL-5000 LE Tilburg, The Netherlands; e-mail: pbs@rioja.caltech.edu; Second and third author's affiliation: Humboldt Universität zu Berlin; Mailing address: Institut für Statistik und Ökonometrie, Humboldt Universität zu Berlin, Spandauer Straße 1, D-10178 Berlin, Germany; e-mail: hafner@wiwi.hu-berlin.de. Comments on a preliminary presentation at the fifth $(EC)^2$ conference are gratefully acknowledged. The XploRe macros for local polynomial estimation and the data were generously provided by Marlene Müller and Olsen Associates, respectively.

short. Recently, some papers by people associated with Olsen appeared, of which a review is given by Guillaume et al. (1994).

A GARCH(1,1) model has at least two drawbacks: it imposes a symmetrical influence of lagged residuals on the volatility (this plays a minor role in FX markets), and leptokurtocity not only in the unconditional but also in the conditional density. Engle, González-Rivera (1991) compute relative efficiencies (as variance ratios of MLE and QMLE) for a variety of distributional assumptions. For example, if the true conditional density is a Student's t with 5 *df*, the relative efficiency is as low as 41%. This situation becomes worse when dealing with intra-daily data, because it is known that the deviation of the unconditional return density from normality increases when the sampling interval is decreased.

In this paper a nonparametric approach is chosen. After a short explanation of the data and a necessary deseasonalization, both conditional mean and conditional variance are estimated local linearly.

2 The Foreign Exchange Market and the Data Set

The foreign exchange market is by far the largest financial market. According to the Wall Street Journal of March 1 1990, the average daily FX trading volume is $ 650 billion. Compared to this, the NYSE's largest volume day, Oct. 19 1987, only had $ 21 billion of volume.

The market is decentralized with the main trading locations being New York, London and Tokyo. It is an electronic market, active 24 hours a day. Banks act as market makers and place bid– and ask–quotes on the screen. Central information collectors such as Reuters provide the quotes for the market makers. Actual trade takes place over the phone. This is the reason why there is no information about actual prices and trading volume. By far the largest part of trading occurs in US Dollars, which assumes in a way the role of the numéraire for the minor rates. Although there is some important central–bank intervention money, by far the largest part of the FX market is pure speculation by the market makers.

The data set was acquired from Olsen & Associates, Zürich. It contains the following numbers of quotes during the time Oct 1 1992, 0:00:00 and Sept 30 1993, 23:59:59 GMT:

- DEM/USD : 1,472,241 records
- JPY/USD : 570,840 records
- JPY/DEM : 158,979 records.

For each pair of bid– and ask–quotes, the time in GMT, the quoting bank and the location of the bank are notated.

3 Seasonal Heteroskedasticity and the Time Scale

First it is necessary to deal with the seasonal volatility. We use a deformed time scale, which seems to be more flexible than the dummy–variable method by Baillie, Bollerslev (1990). For the

Bossaerts, P., Hafner, Ch. and Härdle, W. (1996)
A New Method for Volatility Estimation with Applications in Foreign Exchange Rate Series.

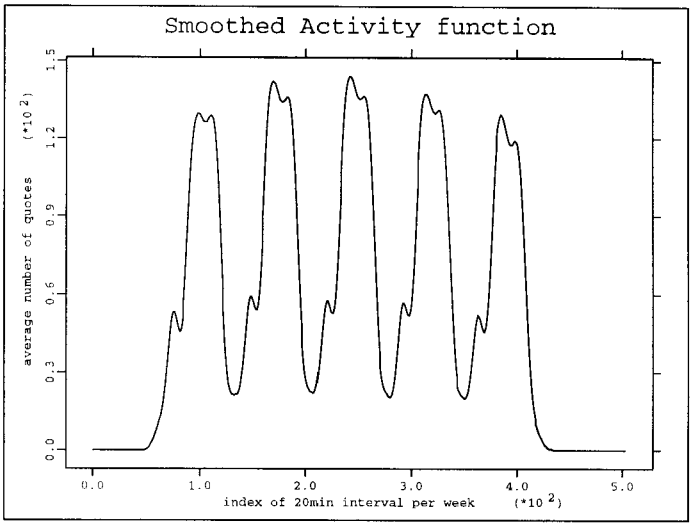


Figure 1: Smoothed activity, defined as the number of quotes, as a function of 20–minute intervals during a week for the DEM/USD rate. A Quartic Kernel with bandwidth $h = 10$ was used.

statistical properties of a time series under deformed time see Stock (1988), who analyzes US-GNP and interest rates, and Ghysels, Gouriéroux and Jasiak (1994).

Usual time series analysis is based not on a physical time scale, but on a business one. For intra–daily data, we can analogously define the time intervals to be longer in low business periods and shorter in busy ones. This idea is not new: Mandelbrot, Taylor (1967) defined the transaction–based "clock" referring to the transaction volume in stock markets, using the fact that volume and volatility are highly correlated. Without information about volume in FX markets, we redefined time based on activity, which is also highly correlated with volatility.

For each 20–minute interval, activity is measured by the number of quotes. Activity is averaged over the weeks and smoothed by a Kernel smoother. The obtained activity function is shown in Figure 1. It is seen that the five major peaks correspond to the working days Monday to Friday, whereas within one day there is a trimodal pattern, corresponding to the openings of the main market centers Tokyo, London and New York.

Denote the activity function in Figure 1 by $a(\cdot)$. The new time scale $t^*(t)$ is defined as

$$t^*(t) = c \int_0^t a(\tau) d\tau, \tag{1}$$

where t denotes physical time, and the constant c is chosen such that one week in deformed time corresponds to one week in physical time, i.e.

$$c = \frac{504}{\int_0^{504} a(\tau) d\tau}.$$

In some cases, there is no quote in the new time interval. This happens because an averaging method is used. The numbers of records are thus reduced from 26280 20–minute intervals per year to 25434 for the DEM/USD rate.

Bossaerts, P., Hafner, Ch. and Härdle, W. (1996)
A New Method for Volatility Estimation with Applications in Foreign Exchange Rate Series.

FX rate	n	mean	std.dev.	skewness	kurtosis
DEM/USD	25434	$5.73E-06$	$7.96E-04$	0.17	12.25
JPY/USD	25247	$-4.89E-06$	$7.98E-04$	0.09	15.71
JPY/DEM	23814	$-1.13E-05$	$8.26E-04$	−0.14	10.66

Table 1: distributional characteristics of the three exchange rate returns

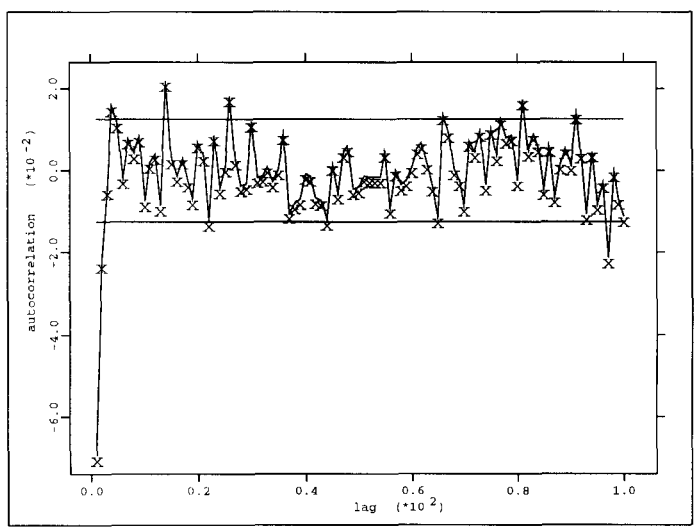


Figure 2: Correlogram for the DEM/USD returns and the first hundred lags. The horizontal lines correspond to the 95% confidence band of a Gaussian white noise.

4 Properties of foreign exchange rates under redefined time

In table 1, four characteristics about the distributions of the returns are given. The skewness is not significantly different from zero for all three rates and the sign changes. The kurtosis, however, reveals substantial differences to a normal density, which has a kurtosis of 3. The return distribution is leptokurtic, i.e. it has fatter tails and a higher peak than a normal distribution.

In Figure 2 the correlogram for the return series is given for the first hundred lags. The first two autocorrelations are significantly negative. However, this does not imply that the market is inefficient. To claim this, one would have to assume a certain equilibrium model for the foreign exchange market. One might interpret this result as a *mean reversion* effect, which was reported in various papers for asset markets. For foreign exchange markets, Goodhart and Figliuoli (1991) and Guillaume et al. (1994) report negative autocorrelation for ultra–high frequencies. Two economic explanations are possible:

1. traders have at the same time different information sets (this would imply market inefficency) or interpret the same news differently, and

2. banks have to perform inventory rebalancing if they hold open positions longer than just a few minutes. This is confirmed by the fact that negative autocorrelation disappears when the

Bossaerts, P., Hafner, Ch. and Härdle, W. (1996)
A New Method for Volatility Estimation with Applications in Foreign Exchange Rate Series.

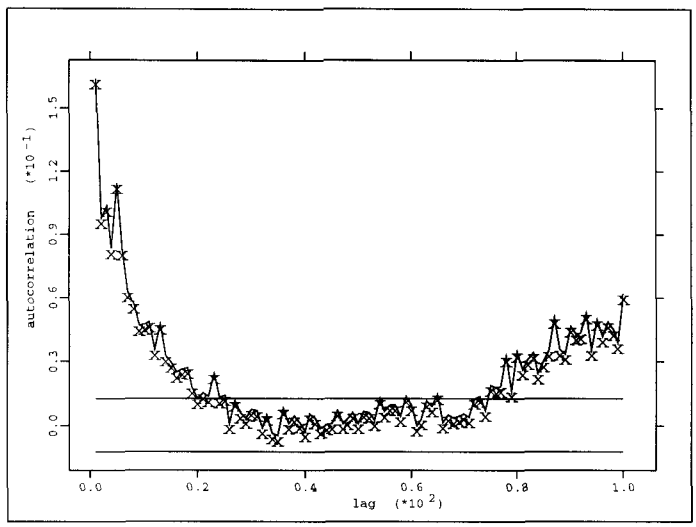


Figure 3: Correlogram for the DEM/USD rate and the first hundred lags of the squared returns. The horizontal lines correspond to a 95% confidence band of Gaussian white noise.

data are aggregated.

The modified Box–Ljung statistic

$$Q_1(k) = n(n+2)\sum_{\tau=1}^{k}(n-\tau)^{-1}r^2(\tau;\Delta s_t)$$

is rejecting the null hypothesis of Gaussian white noise for $k = 20$: $Q_1(k) \sim \chi^2_k$ as. for Gaussian white noise, $Q_1(20) = 167.6$ for DEM/USD, significant at 1%.

In order to get an impression of the immanent conditional heteroskedastic effects, regard Figure 3. This gives the correlogram of the *squared* returns.

The ACF shows a typical declining structure of an autoregressive process. But now the autoregression is in the squared return, which has a close relationship to the variance. Whether the autoregression in the variance is linear or nonlinear cannot be answered yet, but at least we know that there is some kind of conditional heteroskedasticity in the return series.

The Box–Ljung statistic for the squared returns rejects the Null hypothesis of Gaussian white noise:

$$Q_2(k) = n(n+2)\sum_{\tau=1}^{k}(n-\tau)^{-1}r^2(\tau;(\Delta s_t)^2),$$

$Q_2(k) \sim \chi^2_k$ for Gaussian white noise, $Q_2(20) = 2445.7$ for DEM/USD, significant at 1%.

5 Local linear estimation of a CHARN model

This section deals with local linear estimation of the conditional mean ("mean function") and the conditional variance ("variance function") of the three return series. Local linear estimation is a special case of local polynomial estimation (LPE). For details about LPE see Fan and Müller

Bossaerts, P., Hafner, Ch. and Härdle, W. (1996)
A New Method for Volatility Estimation with Applications in Foreign Exchange Rate Series.

(1995) and the monograph by Fan and Gijbels (1995). The Nadaraya–Watson estimate, also a special case, is equivalent to local constant estimation.

A parametric extension of ARCH is the QTARCH model by Gouriéroux and Monfort (1992). Consider the simplest case of a univariate QTARCH(1) model. Also, let $\{y_t\}$ denote a onedimensional process, $\{A_j, j \in J\}$ a partition of $I\!R$, and $\{\xi_t\}$ an IID sequence with mean zero and variance one. Then a QTARCH(1) can be written as

$$y_t = \sum_{j=1}^{J} \alpha_j I(y_{t-1} \in A_j) + \sum_{j=1}^{J} \beta_j I(y_{t-1} \in A_j)\xi_t, \tag{2}$$

where $\alpha_j \in I\!R, j = 1, \ldots, J$, and $\beta_j \in I\!R_+, j = 1, \ldots, J$.

In this model, the mean and variance functions can be considered as stepwise constants. A natural generalization now is to allow for any smooth functions f and σ and estimate both functions nonparametrically. This leads us to the following model:

$$y_t = f(y_{t-1}) + \sigma(y_{t-1})\xi_t. \tag{3}$$

Is is known that ARCH models can be used as approximations of diffusion models, see Gouriéroux (1992). (3) can also be viewed as a general diffusion process in discrete time, allowing for any type of nonlinearity in the mean and variance function.

The use of nonparametric methods in time series analysis has been extensive since Robinson (1983) provided consistency results for α–mixing processes. It is known that stationary Markov chain processes have the α–mixing property, so for the model in (3), where $\{y_t\}$ is a Markov chain, it is sufficient to show that it is also stationary. For a nonlinear model like (3) it is not straightforward to check if the series is stationary. As a complementary result, however, we computed the Augmented Dickey–Fuller (ADF) test statistic for a linear model. The usual result for financial time series is achieved: the log–levels have a unit root and the returns do not. Only for the log–levels of DEM/USD the test just rejects at 1%. To conclude, we can assume the returns to be stationary.

The local linear estimator (LLE) was chosen in favor of the Nadaraya–Watson (NW) or Gasser–Müller (GM) estimator. Under fixed design, the Gasser–Müller estimator is preferrable to NW because of its better bias behaviour. Under random design, however, the variance of GM is higher by the factor 1.5. Asymptotically, local linear estimation combines the advantages of GM and NW, having the same bias as GM and the same variance as NW. For details see Kneip and Engel (1994), who also derive an estimator similar to a Kernel estimator with WARPing but asymptotically equivalent to LLE. The LLE performs better than NW and GM especially at the boundaries.

A more practical reason is that the LLE corresponds to a local least squares problem, and for these kinds of problems easy and fast efficient algorithms are available. Also, not only the regression function, but all of its derivatives up to the $(p-1)^{th}$ order are estimated simultaneously.

Consider again the CHARN model (3). The task is to estimate the mean function $f(x) = E(y_t|y_{t-1} = x)$ and the variance function $\sigma^2(x) = E(y_t^2|y_{t-1} = x) - E^2(y_t|y_{t-1} = x)$, where $y_t \equiv \Delta S_t$ denotes the exchange rate return. For details about assumptions and asymptotics of the

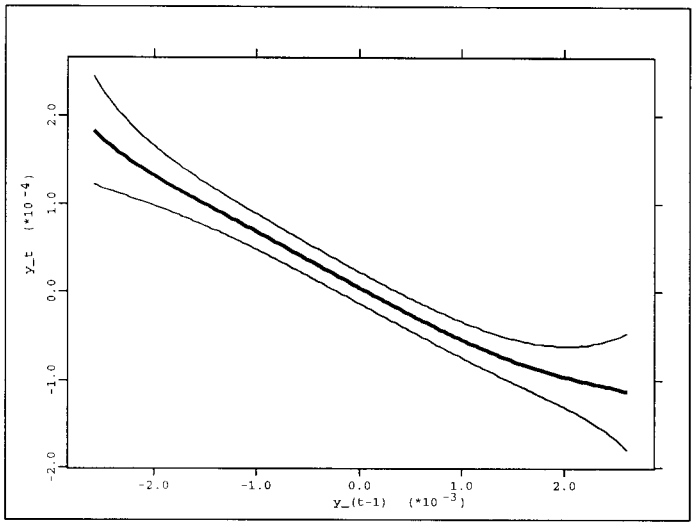


Figure 4: The estimated mean function for DEM/USD with uniform confidence bands. Shown is the truncated range $(-0.0025, 0.0025)$.

LPE procedure used here see Tsybakov (1986) and Härdle and Tsybakov (1995). In general, local polynomial estimation (LPE) is based on computing the following weighted least squares

$$\begin{aligned} \bar{c}_n(x) &= \arg\min_{c \in I\!R^p} \sum_{t=1}^{n} (y_t^2 - c^T U_{tn})^2 K\left(\frac{y_{t-1} - x}{h_n}\right) \\ c_n(x) &= \arg\min_{c \in I\!R^p} \sum_{t=1}^{n} (y_t - c^T U_{tn})^2 K\left(\frac{y_{t-1} - x}{h_n}\right), \end{aligned}$$

where K is a kernel, h_n a bandwidth, $U_{tn} = F(u_{tn})$, $u_{tn} = \frac{y_{t-1}-x}{h_n}$, and

$$F(u) = \left(1, u, \ldots, \frac{u^{p-1}}{(p-1)!}\right)^T.$$

Denoting the true regression function of $E(y_t^2 | y_{t-1} = x)$ by $g(x)$, the estimators of $f(x)$ and $g(x)$ are the first elements of the p–dimensional vectors $c_n(\cdot)$ and $\bar{c}_n(\cdot)$. Consequently, the variance estimate is

$$\hat{\sigma}^2(x) = \hat{g}_n(x) - \hat{f}_n^2(x),$$

with $\hat{f}_n(x) = c_n(x)^T F(0)$ and $\hat{g}_n(x) = \bar{c}_n(x)^T F(0)$.

The estimated functions are plotted together with approximate 95% confidence bands, see e.g. Härdle (1990). The cross-validation optimal bandwidth $h = 0.0028$ is used for the local linear estimation of the mean function in Figure 4. As indicated by the 95% confidence bands, the estimation is not very robust at the boundaries. Therefore, Figure 4 covers a truncated range. The result corresponds to the negative autocorrelation reported and explained above.

Analogously, the variance estimate is shown in Figure 5, using the cross-validation optimal bandwidth.

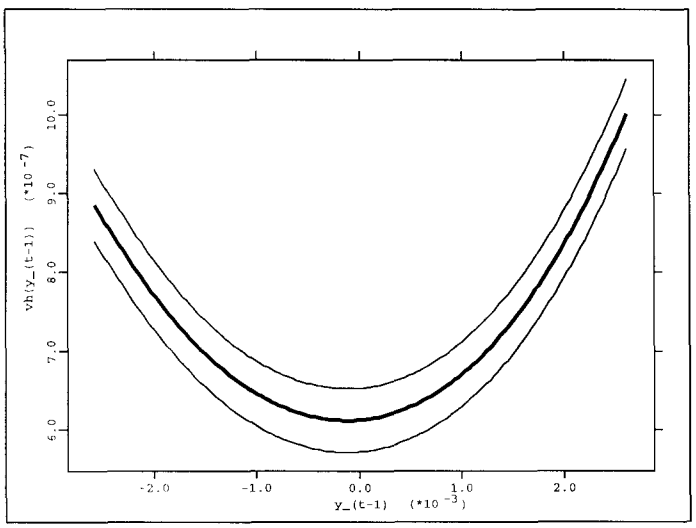


Figure 5: The estimated variance function for DEM/USD with uniform confidence bands. Shown is the truncated range $(-0.0025, 0.0025)$.

To save space, just the plots for DEM/USD are given. The basic results across all rates are the mean reversion (although not very distinct for JPY/DEM) and the "smiling" shape of the conditional variance. Conditional heteroskedasticity appears to be very distinctly. The smile is almost exactly symmetrical for JPY/USD, whereas for DEM/USD and DEM/JPY a "reverted leverage effect" can be observed, meaning that the conditional variance is higher for positive lagged returns than for negative ones of the same size. But the difference is still within the 95% confidence band.

5.1 Residual Analysis, DEM/USD

Table 2 shows the autocorrelations of the residuals and squared residuals of the fitted model.

Especially the second lag reveals some linear dependence. It seems that just the first order autocorrelation has been captured. To see how the ARCH effects behave, regard the autocorrelation analysis of the squared residuals. Indeed, ARCH effects are present, but smaller than in the returns. Because the fitted model, which we will call now S1 model, does not yield satisfactory residuals, another argument was introduced into the volatility function, namely the bid–ask spread. This model will be called S2 model.

5.2 S2 model

It is known that the spread is closely related to "risk". The economic reason for this lies in the nature of bid–ask spreads. Basically, the spread can be considered as a compensation for the market maker, having two components: the transaction costs and the risk component. Risk is higher in less active markets and thus the bid–ask spread widens, because the bank takes the risk of having an open position for a longer time interval than in busy hours.

Let BA_t denote the bid–ask spread at time t. Then the S2 model can be written as:

Bossaerts, P., Hafner, Ch. and Härdle, W. (1996)
A New Method for Volatility Estimation with Applications in Foreign Exchange Rate Series.

	residuals			squared residuals		
Lag	ACF	PACF	Q-Stat	ACF	PACF	Q-Stat
1	-0.003	-0.003	0.1790	0.075	0.075	144.64
2	-0.027	-0.027	18.154	0.099	0.094	396.29
3	-0.005	-0.006	18.919	0.125	0.113	795.34
4	0.015	0.015	24.942	0.083	0.061	971.96
5	0.015	0.015	30.888	0.114	0.087	1303.6
6	-0.001	-0.000	30.909	0.088	0.054	1502.1
7	0.009	0.01	32.786	0.072	0.035	1633.7
8	0.006	0.005	33.563	0.066	0.025	1744.7
9	0.008	0.008	35.213	0.055	0.017	1822.9
10	-0.005	-0.005	35.803	0.054	0.016	1896.5

Table 2: Residual analysis of the estimated model.

$$y_t = f(y_{t-1}) + \sigma(y_{t-1}, BA_{t-1})\xi_t \tag{4}$$

First, f was estimated local linearly, σ with a two–dimensional Nadaraya–Watson estimate with various bandwidths. The results are not reported here, but the improvement to S1 was not very high.

Then, σ was estimated with a two–dimensional local linear estimator. The bandwidths were chosen to be the same as for the "best" Nadaraya–Watson estimator, namely $h_1 = 0.001$ for y_{t-1} and $h_2 = 0.0005$ for the bid–ask spread. Also, the number of bins – 40*40=1600 – was the same. The autocorrelation of the residuals and squared residuals of the S2 model are given in Table 3.

The residuals reveal that at lag 2 some negative autocorrelation remains. As the squared residuals show, ARCH effects still are present, but smaller than for the S1 model. This indicates that the bid–ask spread is a persistence factor for the volatility, although not a sufficient one. Thus, being better able to cope with the long memory in the process, the S2 model improves the S1 model to some extent. It needs to be further investigated how sensitive this result is to the choice of the bandwidths. Our choice of global bandwidths can, of course, be generalized to adaptive bandwidths as in Fan and Gijbels (1995). First results on this topic look promising.

Also, the optimal number of included lags has to be determined by selection criteria. The resulting multi–dimensional model can then be reduced to an additive model.

6 Conclusion

A CHARN model was fitted to three major foreign exchange rates via local linear estimation on the basis of a redefined time scale. The results show for all rates mean reversion and conditional heteroskedasticity. For two rates, the "smile" is slightly skewed, but not significantly.

By adding the bid–ask spread to the conditioning set one is able to improve the squared residual

Bossaerts, P., Hafner, Ch. and Härdle, W. (1996)
A New Method for Volatility Estimation with Applications in Foreign Exchange Rate Series.

	residuals			squared residuals		
Lag	ACF	PACF	Q-Stat	ACF	PACF	Q-Stat
1	-0.002	-0.002	0.12439	0.005	0.005	0.64717
2	-0.034	-0.034	29.2413	0.061	0.060	93.7255
3	-0.007	-0.007	30.3563	0.096	0.096	327.786
4	0.012	0.011	33.9038	0.078	0.075	484.077
5	0.011	0.010	36.8414	0.081	0.072	653.834
6	0.002	0.002	36.9122	0.072	0.057	785.611
7	0.010	0.011	39.3611	0.065	0.046	896.294
8	0.003	0.004	39.6654	0.041	0.017	939.21
9	0.010	0.010	42.0623	0.050	0.024	1003.68
10	-0.007	-0.007	43.2624	0.038	0.012	1039.71

Table 3: Residual analysis of the estimated S2 model.

autocorrelations. Thus, persistence of the variance as another stylized fact is partly captured. Contrary to IGARCH models, where the variance is nonstationary, the degree of persistence is not determined but driven by a stochastic process.

The model is planned to be extended mainly in two directions:

1. More lags can be included in the mean function as well as in the variance function. This, of course, would bring up the "curse of dimensionality" one usually has in nonparametric estimation. A solution could be the additive model class, for which Chen and Tsay (1993) have given algorithms and applications.

2. The bid–ask spread can also be included into the mean function in order to get a relationship between mean and variance (analogously to GARCH–M models).

Further research will concentrate on goodness–of–fit tests of these models and on the predictive power of CHARN–type models. Above all, it is aimed to get a better understanding of the dynamic behaviour of the volatility, which plays a major role in theoretical finance models.

References

Baillie, R. T. and Bollerslev, T. (1990) Intra–day and inter–market volatility in foreign exchange rates, Review of Economic Studies 58: 565–585.

Bossaerts, P., Härdle, W. and Hafner, C. (1995) Foreign exchange rates have surprising volatility, Discussion Paper, Humboldt–Universität zu Berlin.

Chen, R. and Tsay, R. S. (1993) Nonlinear additive arx models, Journal of the American Statistical Association 88: 955–967.

Drost, F. C. and Nijman, T. E. (1993) Temporal Aggregation of GARCH processes, Econometrica 61: 909–927.

Engle, R. F. and González-Rivera, G. (1991) Semiparametric ARCH models, Journal of Business & Economic Statistics 9: 345–359.

Fan, J. and Gijbels, I. (1995) Local polynomial modeling and its application – Theory and methodologies, Chapman and Hall.

Fan, J. and Müller, M. (1995) Density and regression smootheing, in: Härdle, W., Klinke, S. and Turlach, B., XploRe – an interactive statistical computing environment, Springer.

Ghysels, E., Gouriéroux, C. and Jasiak, J. (1994) Market time and asset price movements, theory and estimation, Discussion paper Université de Montréal.

Goodhart, C.A. and Figliuoli, L. (1991) Every minute counts in financial markets, Journal of International Money and Finance 10: 23–52.

Gouriéroux, C. (1992) Modèles ARCH et Applications Financières, Economica.

Gouriéroux, C. and Monfort, A. (1992) Qualitative threshold ARCH models, Journal of Econometrics 52: 159–199.

Guillaume, D.M., Dacorogna, M.M., Davé, R.R., Müller, U.A., Olsen, R.B. and 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.

Härdle, W. (1990) Applied nonparametric regression, Cambridge University Press.

Härdle, W. and Tsybakov, A. (1995) Local polynomial estimation of the volatility function, SFB 373 Discussion paper.

Kneip, A. and Engel, J. (1994) A remedy for Kernel regression under random design, Discussion paper SFB 303, Universität Bonn.

Mandelbrot, B. B. and Taylor, H. M. (1967) On the distribution of stock price differences, Operations Research 15: 1057–1062.

Robinson, P. M. (1983) Nonparametric estimators for time series, Journal of Time Series Analysis 4: 185–207.

Stock, J. H. (1988) Estimating continuous time processes subject to time deformation, Journal of the American Statistical Association 83(401): 77–84.

Tsybakov, A. (1986) Robust reconstruction of functions by the local–approximation method, Problems of Information Transmission 22, 133–146.

Athens Conference on Applied Probability ad Time Series Ted Hannan memory Volume. LNS 115, p 55-72
Editor: P. Robinson, M Rosenblatt Springer Verlag

Foreign Exchange Rates Have Surprising Volatility

Peter Bossaerts Christian Hafner Wolfgang Härdle

June 1995 (First Draft: February 1995)*

*First author's affiliation: California Institute of Technology and Tilburg University; Mailing address: CentER, Tilburg University, PO Box 90153, NL-5000 LE Tilburg, The Netherlands; Phone: +31.13.663.101; e-mail: pbs@rioja.caltech.edu. Second and third author's affiliation: Humboldt–Universität zu Berlin; Mailing address: Institut für Statistik und Ökonometrie, Humboldt–Universität zu Berlin, Spandauer Strasse 1, D-10178 Berlin, Germany; Phone +49.30.246.82.30; e-mail: SFB373@wiwi.hu-berlin.de. This research was financed through contributions from the Sonderforschungsbereich 373 "Quantifikation und Simulation Ökonomischer Prozesse." Comments from participants and discussants at the fifth $(EC)^2$ Conference as well as the first HFDF Conference are gratefully acknowledged. Michel Dacorogna and Christian Gouriéroux gave valuable criticism. Olsen Associates generously provided the data.

Athens Conference on Applied Probability ad Time Series Ted Hannan memory Volume. LNS 115, p 55-72
Editor: P. Robinson, M Rosenblatt Springer Verlag

Abstract

Local Polynomial Estimation (LPE) is implemented on a dataset of high-frequency foreign exchange (FX) quotes. This nonparametric technique is meant to provide a flexible background against which to evaluate parametric time series models. Assuming a conditionally heteroscedastic nonlinear autoregressive (CHARN) model, estimates of the mean and volatility functions are reported. The mean function displays pronounced reversion. Surprisingly, the volatility function exhibits asymmetry. The CHARN model, however, captures only the short-run behavior of conditional volatility. Nevertheless, part of the evidence of persistent conditional volatility appears in reality to be the effect of conditional kurtosis. Stochastic volatility models are ideal to capture this time series feature.

Keywords: Local Polynomial Estimation, Conditional Volatility, Conditional Kurtosis, Nonlinear Autoregressive Models, Foreign Exchange Markets.

Bossaerts, P., Härdle, W. and Hafner, Ch. (1996)
Foreign exchange rates have surprising Volatility.

Athens Conference on Applied Probability ad Time Series Ted Hannan memory Volume. LNS 115, p 55-72
Editor: P. Robinson, M Rosenblatt Springer Verlag

1 Introduction

Conditional volatility of asset prices in general and foreign exchange (FX) rates in particular have been the subject of intense investigation over the last few years. Virtually all estimation has been parametric. Most popular are the GARCH family (e.g., Baillie and Bollerslev [1989]), as well as stochastic volatility models (e.g., Mahieu and Schotman [1994]). In this paper, we report estimates from the application of a new, nonparametric technique to high-frequency FX quote data, namely, Local Polynomial Estimation (LPE). This leads to surprising insights on the dynamics of FX rates.

In their analysis of the shape of the volatility of foreign exchange rates and stock returns as a function of lagged, exogenous information, Pagan and Ullah [1988] and Pagan and Schwert [1990] have already pointed out that nonparametric modelling is urgent. While it is not meant to displace parametric modelling, it reveals important information with which to enhance the parametric estimation. For ease of analysis, the extant parametric models (the ARCH family, as well as models of stochastic volatility) are all linear after suitable transformation. Misspecifications are accomodated for by the addition of explanatory variables, at the cost of making the analysis less parsimonious. What we provide here is a simple nonparametric and nonlinear analysis of the volatility function. The result is not only a descriptive account of the data, but also a framework with which one can evaluate the suitability of the existing parametric models. A similar approach has recently been taken by Fournie [1992], Aït-Sahalia [1994], Hutchinson, Lo and Poggio [1994] and Aït-Sahalia and Lo [1994], in continuous-time modelling, and by Härdle and Mammen [1993], in regression analysis.

One could argue that there are few theoretical reasons to expect nonlinearities in the volatility function of FX rates. In particular, there are no "leverage effects," unlike with common stock prices (see, e.g., Christie [1982], Pagan and Schwert [1990]). Yet, a closer inspection of the mechanics of the FX market reveals a potential for nonlinearities. Foremost, one should mention central bank intervention. Bossaerts and Hillion [1992], for instance, find a pronounced effect of eminent central bank intervention on bid-ask spreads. The nonlinearity of typical policy reaction functions (Neumann [1984], Hsu and Kugler [1994]) is likely to be reflected in the process of FX rate changes.

Moreover, as far as FX rates are concerned, it is attractive to be able to capture the stochastic nature of persistence in conditional volatility. There appear to be two types of events in high-frequency FX rate changes: those that induce volatility and those that have no effect on subsequent volatility (see Bewley, Lowe and Trevor [1988], who analyzed intraday Australian dollar quotes). CHARN modelling allows for such phenomena. Of course, it would ultimately be desirable to identify the nature of these different events.

Bossaerts, P., Härdle, W. and Hafner, Ch. (1996)
Foreign exchange rates have surprising Volatility.

Athens Conference on Applied Probability ad Time Series Ted Hannan memory Volume. LNS 115, p 55-72
Editor: P. Robinson, M Rosenblatt Springer Verlag

We explicitly fit the mean function as well, allowing it to be nonlinear. In some recent analyses, the mean change in the FX rate is even constrained to be zero (e.g., Mahieu and Schotman [1994]). The mean function is not unimportant, even if the purpose of the modelling of time-varying volatility is option pricing. As Lo and Wang [1994] illustrate, misestimation of mean changes in asset prices (let alone failure to account for a mean) leads to substantial biases in option prices and hedges. In continuous time, the mean function is irrelevant. But since estimation necessarily takes place in discrete time, biases enter whenever time-variation in volatility is estimated without accounting for the mean function.

We obtain estimates of the mean and volatility functions of the CHARN model by means of Local Polynomial Estimation (LPE). As its name indicates, LPE is based on locally fitting polynomials. This means that polynomials are estimated by weighted least squares, where the weights depend on the distance of an observation from the values of the arguments of the mean and volatility function at which an estimate is to be obtained. Kernel functions localize the weights in the space of predictor variables. The bandwidth of the kernel function determines the smoothness of the fit. For consistency, the bandwidth must be lowered appropriately as the number of observations increases.

LPE has a long tradition in regression estimation for cross-sectional data. Stone [1977] and Cleveland [1979] seem to have been the first to suggest the technique. Tsybakov [1986] proved asymptotic normality. For an application in finance, see Bossaerts and Hillion [1995], where LPE is used to obtain estimates of dynamic hedge portfolio weights. Recently, Härdle and Tsybakov have analyzed the properties of LPE in the context of CHARN models (Härdle and Tsybakov [1995]). Crucial in the analysis is the concept of geometric ergodicity, which ensures the existence of a time-invariant distribution and sufficiently strong mixing such that laws of large numbers and central limit theorems hold. The conditions for geometric ergodicity are familiar to option pricing theorists, where these are needed to prove existence of solutions to stochastic differential equations (e.g., Karatzas and Shreve [1983]). Duffie and Singleton [1992] also appealed to geometric ergodicity in order to show consistency and asymptotic normality of their simulation estimator of Markov models.

Other nonparametric procedures have been suggested for the analysis of time series, such as standard kernel estimation (see, e.g., Györfi, et al. [1989]). LPE has the advantage, however, of featuring improved smoothing bias, as well as being computationally straightforward (local least squares; also, derivatives are obtained in a straightforward way). Closely related to the nonparametric techniques are the (parametric) threshold ARCH models of Zakoïan [1990] and Gouriéroux and Monfort [1992], which are in fact nonparametric modelling procedures, whereby mean and volatility functions are

Athens Conference on Applied Probability ad Time Series Ted Hannan memory Volume. LNS 115, p 55-72
Editor: P. Robinson, M Rosenblatt Springer Verlag

approximated by step functions with a fixed number of steps.

The LPE technique is implemented on the Olsen high-frequency FX quote data. Substantial mean reversion is discovered, as well as asymmetry in the volatility function. Long-run autocorrelation in squared residuals, however, remained even after introducing additional conditioning variables, such as the bid-ask spread. These were meant to capture volatility persistence. The problem did not disppear after (admittedly timid) implementation of HARCH modeling (Müller, et al. [1995]). Undersmoothing, however, indirectly revealed pronounced evidence of *conditional kurtosis*. Stochastic volatility models (e.g., Mahieu and Schotman [1994]) are well-tailored to capture such time series properties.

The remainder of the paper is organized as follows. The next section briefly introduces LPE of CHARN models. Section 3 discusses the dataset. Section 4 reports the LPE results. Section 5 concludes.

2 LPE of CHARN models

Let $\{y_t\}$ be a Markov time series, satisfying the following stochastic difference equation:

$$y_t = f(y_{t-1}) + s(y_{t-1})\xi_t, \tag{1}$$

where ξ_t are i.i.d. random variables with mean zero and unit variance. Here f and s are unknown mean and volatility functions, respectively, with $s(x) > 0$ for all values of x, and y_0 is a random variable independent of the series ξ_t. The model (1) is a heteroscedastic nonlinear autoregression (CHARN).

We estimate the volatility function $v(x)$ $(= s^2(x))$ from a sample $y_1, ..., y_T$ by means of Local Polynomial Estimation (LPE). The procedure will simultaneously generate an estimate of $f(x)$. Define T vector functions $u_T(z)$, as follows:

$$u_T(z)' = [1 \;\; \frac{z}{h_T} \;\; (\frac{z}{h_T})^2 \;\; ... \;\; (\frac{z}{h_T})^{l-1}/(l-1)!], \tag{2}$$

where h_T is a parameter to be referred to as the bandwidth parameter and l denotes the degree of the polynomial. For consistency, the bandwidth parameter should decrease with the sample size (T). Consider now the minimization problems:

$$\gamma_T(x) = \arg\min_{\gamma} \sum_{t=1}^{T} \left\{ y_t^2 - \gamma' u_T(y_{t-1} - x) \right\}^2 K(\frac{y_{t-1} - x}{h_T}), \tag{3}$$

$$\phi_T(x) = \arg\min_{\phi} \sum_{t=1}^{T} \left\{ y_t - \phi' u_T(y_{t-1} - x) \right\}^2 K(\frac{y_{t-1} - x}{h_T}), \tag{4}$$

Bossaerts, P., Härdle, W. and Hafner, Ch. (1996)
Foreign exchange rates have surprising Volatility.

Athens Conference on Applied Probability ad Time Series Ted Hannan memory Volume. LNS 115, p 55-72
Editor: P. Robinson, M Rosenblatt Springer Verlag

where K is a kernel function and h_T the bandwidth. The estimate of f at x, $\hat{f}(x)$, is then given by:

$$\hat{f}(x) = \phi_T(x)' u_T(0). \tag{5}$$

The estimate of v at x, $\hat{v}(x)$, on the other hand, is given by:

$$\hat{v}(x) = \gamma_T(x)' u_T(0) - \{\phi_T(x)' u_T(0)\}^2 . \tag{6}$$

From these equations, it is clear that the estimates are obtained as the intercepts of polynomials which are fit by weighted least squares, where the weight to be put on an observation is determined by its distance from the target value, x. In other words, our procedure can also be described as local weighted least squares.

There are plenty of valid kernel functions (see Härdle [1990]). The bandwidth parameter, however, should be chosen carefully, in order to avoid overfitting. Crossvalidation is a simple procedure, albeit computationally intensive. The "out of sample" average squared prediction error of the estimated model is minimized with respect to the bandwidth. The "out of sample" prediction error for an observation is obtained from estimates of the mean and volatility functions based on all the data except the observation at hand.[1]

Härdle and Tsybakov [1995] establish the theoretical properties of LPE estimation in the Markov model (1). They show that the estimates of the mean and volatility functions converge and are asymptotically normally distributed. The conditions stated there guarantee asymptotic stationarity. The effect from initial sampling from another distribution than the stationary one thereby dies out fast enough for central limit theorems to continue to hold even in the nonstationary case.

3 Data

The dataset was compiled by Olsen and Associates, and consists of bid and ask quotes from Reuter's FXFX page. The sample covers the period 1 October 1992 at 0:00:00 GMT till 30 September 1993 at 23:59:59 GMT. The data were filtered by Olsen to remove erroneous quotes and other outliers (less than 0.5% of the data). Quotes for two currencies are available: DEM/USD and YEN/USD. Obviously, this is a huge dataset: the DEM/USD file, for instance, contains 1,472,241 records. We focused on transactions ten and twenty minutes apart, so that we could safely use the average of the bid and ask quotes in our analysis. Over shorter intervals, temporary skewness in bid-ask spreads

[1]The estimations were performed with XploRe. For a description, see XploRe Systems [1995].

Bossaerts, P., Härdle, W. and Hafner, Ch. (1996)
Foreign exchange rates have surprising Volatility.

Athens Conference on Applied Probability ad Time Series Ted Hannan memory Volume. LNS 115, p 55-72
Editor: P. Robinson, M Rosenblatt Springer Verlag

because of inventory rebalancing by market making banks become important (see Guillaume, et al. [1994]), so that the average of the bid and the ask is devoid of economic meaning. Profit opportunities are obviously higher at shorter frequencies, because of the possibility to trade against banks' inventory rebalancing. But we set out to focus on the volatility effects net of such short-term inventory redressing. Incidentally, the impact of market making on short-term autocorrelations invalidates the analysis of FX quote changes as diffusion processes, for in that case, the importance of the mean function ought to decrease with the discretization mesh. Because of profit opportunities induced by market making, the mean function becomes actually more important with reductions in the length of the sampling interval.

Hence, if b_t denotes the bid at t and a_t the ask, the price at t, p_t, is defined to be

$$p_t = [\log b_t + \log a_t]/2.$$

We model the time series behavior of the *change* in the price, i.e.,

$$y_t = p_t - p_{t-1}.$$

At one point, however, we also report results involving the bid-ask spread itself, defined as:

$$\log a_t - \log b_t.$$

We did not, however, sample quotes over intervals of ten or twenty minutes in calendar time. There is substantial seasonality in FX data, due to seasonalities in trading intensity. Volatility, for instance, is highest during the period of the 24 hour trading day when the European markets are open. In contrast, there is little activity, and, hence, little volatility, on Sunday mornings (measured relative to GMT). Therefore, time was first deformed, after which sampling took place in this newly-defined measure of time. Effectively, time intervals were shortened during busy periods, while the real-time equivalent of ten or twenty minutes was lenghtened over periods of low activity.

We used our own time deformation, where activity is measured as the kernel fit through the sample average number of quote revisions over twenty-minute intervals of the trading week. Quotes were obviously not always available at exactly twenty-minute marks. Whenever that happened, we took the first subsequent quote. Of course, sometimes no quote change is forthcoming even beyond the next time interval (this obviously occurs often during holidays which fell on an otherwise regular trading day). The empty intervals in deformed time are then skipped in order to match consistently the intervals in real and deformed time. The skipping of intervals in deformed time because of absence of

Bossaerts, P., Härdle, W. and Hafner, Ch. (1996)
Foreign exchange rates have surprising Volatility.

Athens Conference on Applied Probability ad Time Series Ted Hannan memory Volume. LNS 115, p 55-72
Editor: P. Robinson, M Rosenblatt Springer Verlag

quote changes meant that of the theoretical 26,280 quote changes, we sampled only 25,434 (DEM/USD) and 25,247 (YEN/USD).

We also used data sampled in Olsen's own redefinition of time ("theta time"; see, e.g., Dacorogna, et al. [1993]), which the Olsen Research Group graciously sent us. Theirs are quotes sampled over ten minutes in redefined time. Time deformation is not only based on a simple measure of activity (as ours was), but included other conditioning variables which are not directly available from the Reuters FXFX page (which provides only a limited picture of the state of the market). When comparing the CHARN estimation results, however, no differences could be found.[2] Because the dataset has become standard, we decided to report here only the results from LPE estimation on Olsen's series in "theta time."[3]

4 LPE Results

LPE was implemented on the samples discussed in the previous section. We fitted first-order polynomials (i.e., linear functions) locally, using a quartic kernel, with bandwidth selected by means of crossvalidation. Figure 1 displays our estimate of v, the variance function, for the DEM/USD, together with 95% confidence bounds. Most surprising is the asymmetric shape, as if there were a leverage effect (albeit inverted) in FX similar to that found in stock prices, with volatility increasing after increases in the DEM/USD rate.

We suspect that this "leverage" effect is caused by the asymmetric nature of central bank reaction policies, with more uncertainty about imminent intervention after increases in the DEM/USD rate. The asymmetry in the variance function is significant. To make this clear visually, Figure 2 plots the variance function against the absolute value of lagged spot quote changes.

Figure 3 plots the estimate of the mean function for the DEM/USD. It displays substantial mean reversion. The functional relationship is close to linear.

Figures 4, 5 and 6 repeat this exercise for the YEN/USD. Mean reversion is even stronger for this currency; asymmetry in the variance function is less pronounced for large lagged changes in the FX quote.

While CHARN modelling is able to capture interesting short-term mean and volatility patterns, it captures only part of the persistence in absolute values and squares of FX quote changes. Table 1

[2]Only when differencing over longer intervals could clear seasonalities in our own dataset be discovered. These seasonalities are absent in Olsen's data in "theta time".

[3]For some estimation results on the first dataset, see Bossaerts, Härdle and Hafner [1995].

Bossaerts, P., Härdle, W. and Hafner, Ch. (1996)
Foreign exchange rates have surprising Volatility.

Athens Conference on Applied Probability ad Time Series Ted Hannan memory Volume. LNS 115, p 55-72
Editor: P. Robinson, M Rosenblatt Springer Verlag

documents this. It lists serial correlations of the signed change in the exchange rate quotes (y_t), as well as of absolute and squared values. In comparison, it displays the same autocorrelations for ξ_t, the residual in the CHARN model. This residual is white noise, and, hence, should not be autocorrelated. Whereas the CHARN model is able to capture all the serial correlation in signed FX quote changes, it fails to account for most of the autocorrelation of absolute and squared values.

We tried to accomodate this persistence by adding conditioning variables to the volatility function, namely the lagged bid-ask spread (a persistent variable as well, which theory would claim changes systematically with conditional volatility) and more lags of quote changes. This had little effect on the higher-order autocorrelation of the noise. It does indicate that the autocorrelation in conditional volatility is mostly deterministic, reducing the importance of one of our conjectures, namely that large FX quote changes which fail to generate subsequent spells of high volatility are important.

The addition of the bid-ask spread as conditioning variable did generate the expected effects. In particular, a higher bid-ask spread predicted higher conditional volatility. The effect was nonlinear, however: small increases in the bid-ask spread are associated with minimal changes in conditional volatility.[4]

We also implemented Müller, et al. [1995]'s idea of HARCH modelling, whereby the sum of FX quote changes over several lags is used as conditioning variable in the variance function. Unlike Müller, et al. [1995], we only went up to twenty-four lags (four hours in redefined time). This failed to address satisfactorily even low-order serial correlation in absolute and squared residuals of the CHARN model.

In the LPE estimation, bandwidths were selected by means of crossvalidation. This provides optimal smoothness, in the sense that it balances bias against variance. When reducing the bandwidth, a better fit is obviously obtained. Surprisingly, however, it not only reduced first-order serial correlation in absolute and squared values of the residuals, *but also higher-order autocorrelations.* Table 1 illustrates this. The bandwidth was set equal to 0.0001 (about 1/60th of the optimal bandwidth).

In an attempt to discover the cause of this effect, we plotted the estimated volatility function for the reduced bandwidths. The estimates were extremely erratic, indicating that the improved higher-order serial correlation of the residuals was generated by mixing the residual of the original model with a random variable whose distribution depends on the lagged value of the FX quote change. Formally, the resulting model can be written in terms of the original CHARN model, as follows:

$$y_t = f(y_{t-1}) + s(y_{t-1})\eta_t\xi_t, \tag{7}$$

[4]See Bossaerts, Härdle and Hafner [1995] for estimation results.

Athens Conference on Applied Probability ad Time Series Ted Hannan memory Volume. LNS 115, p 55-72
Editor: P. Robinson, M Rosenblatt Springer Verlag

where η_t is a mixing variable with mean one and variance depending on y_{t-1}. Of course, $s(y_{t-1})$ and η_t could be merged to one random variable, thus obtaining a stochastic volatility model (see, e.g., Mahieu and Schotman [1994]).

The main effect of the use of a mixing variable is that it introduces explicitly *conditional kurtosis.* Conditional kurtosis is also present in the original model (Equation (1)), provided the unconditional distribution of ξ_t is not normal. But it changes as a function of y_{t-1} only indirectly, through the effect on the conditional volatility. Equation (7) has the potential of disentangling the impact of y_{t-1} onto future volatility and future kurtosis.

There is evidence in the data that the impact of lagged FX quote changes on future volatility and kurtosis are different. In particular, they are inverted. Whereas higher conditional volatility is associated with large changes in exchange rate quotes (see Figures 1 and 4), conditional kurtosis is higher for small FX quote changes. Figures 8 and 9 show this. They display plots of conditional kurtosis as a function of y_{t-1}. The plots were generated as follows. FX quote changes are allocated to "bins". The bins are formed after sorting the data with respect to the lagged FX quote change. Each bin contained 100 observations, in ascending order. The sample kurtosis was computed per bin, and plotted in Figures 8 and 9 against the sample mean lagged FX quote change. One should smooth the estimates across bins, but we wanted to provide the reader with the raw results. No matter how one smooths, the effect would be the same: conditional kurtosis is far higher for small changes in exchange rates.[5]

5 Conclusion

We presented results from local polynomial estimation of the mean and volatility function in a conditionally heteroscedastic nonlinear autoregressive (CHARN) model of foreign exchange quote changes. The data were sampled over intervals of time that were deformed to remove seasonalities. To estimate the mean and volatility functions, linear functions were fit locally by least squares, using a kernel function to determine the weights to be put on each observation. The resulting estimates show clear (i) mean reversion, (ii) nonlinearity and asymmetry in conditional volatility.

The biggest challenge to the model came from the high persistence in absolute and squared residuals. In part, this appeared to be the result of conditional kurtosis that could not be captured by

[5]The estimates of conditional kurtosis are valid only if fourth moments really exist. Incidentally, the same applies to the estimates of the serial correlation of squared exchange rate changes. For a critical view on this, see Dacorogna, et al. [1992].

Bossaerts, P., Härdle, W. and Hafner, Ch. (1996)
Foreign exchange rates have surprising Volatility.

Athens Conference on Applied Probability ad Time Series Ted Hannan memory Volume. LNS 115, p 55-72
Editor: P. Robinson, M Rosenblatt Springer Verlag

changes in the volatility parameter. It appears that parametric modelling could be improved by disentangling conditional volatility and conditional kurtosis. Stochastic volatility models may be ideal to attain this goal.

References

Aït-Sahalia, Y., 1994, "A Specification Test For Continuous-Time Stochastic Processes," University of Chicago Graduate School of Business Working Paper.

Baillie, R.T. and T. Bollerslev, 1990, "Intra-Day and Inter-Market Volatility in Foreign Exchange Rates," *Review of Economic Studies* 58, 565-85.

Bewley, R., P. Lowe and R. Trevor, 1988, "Exchange Rate Changes: Are They Distributed As Stochastic Mixtures of Normals?" University of New South Wales Discussion Paper.

Bossaerts, P., W. Härdle and C. Hafner, 1995, "A New Method For Volatility Estimation With Applications In Foreign Exchange Rate Series," *Physica Verlag*, forthcoming.

Bossaerts, P. and P. Hillion, 1991, "Market Microstructure Effects of Government Intervention in the Foreign Exchange Market," *Review of Financial Studies* 4, 513-541.

Bossaerts, P. and P. Hillion, 1995, "Local Parametric Analysis of Hedging in Discrete Time," *Journal of Econometrics*, forthcoming.

Christie, A., 1982, "The Stochastic Behavior of Common Stock Variances: Value, Leverage and Interest Rate Effects," *Journal of Financial Economics* 10, 407-432.

Cleveland, W., 1979, "Robust Locally Weighted Regression and Smoothing Scatterplots," *Journal of the American Statistical Association* 74, 823-36.

Dacorogna, M.M., O. Pictet, U.A. Müller and C.G. de Vries, 1994, "The Distribution of Extremal Foreign Exchange Rate Returns in Extermely Large Datasets," Olsen & Associates, working paper.

Dacorogna, M.M., U.A. Müller, R.J. Nagler, R.B. Olsen and O.V. Pictet, 1993, "A Geographical Model for the Daily and Weekly Seasonal Volatility in the FX Market," *Journal of International Money and Finance* 12, 413-438.

Bossaerts, P., Härdle, W. and Hafner, Ch. (1996)
Foreign exchange rates have surprising Volatility.

Athens Conference on Applied Probability ad Time Series Ted Hannan memory Volume. LNS 115, p 55-72
Editor: P. Robinson, M Rosenblatt Springer Verlag

Duffie, D. and K. Singleton, 1993, "Simulated Moments Estimation of Markov Models of Asset Prices," *Econometrica* 61, 929-952.

Fournie, E., 1992, "Un Test de Type Kolmogorov-Smirnov Pour Processes de Diffusion Ergodiques," INRIA Working Paper.

Gouriéroux, C. and A. Monfort, 1992, "Qualitative Threshold ARCH Models," *Journal of Econometrics* 52, 159-99.

Guillaume, D.M., M.M. Dacorogna, R.R. Davé, U.A. Müller, R.B. Olsen and O.V. Pictet, 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.

Györfi, L., W. Härdle, P. Sorda and P. Vieu, 1989, *Nonparametric Curve Estimation From Time Series*, New York: Springer Verlag.

Härdle, W. and A. Tsybakov, 1995, "Local Polynomial Estimation of the Volatility Function," *Journal of Econometrics*, forthcoming.

Härdle, W. and E. Mammen, 1993, "Comparing Nonparametric Versus Parametric Regression Fits," *Annals of Statistics* 21, 1926-47.

Hsu, C.T. and P. Kugler, 1994, "The Term Structure of Interest Rates: A Dynamic Analysis," University of Vienna working paper.

Karatzas, I. and S. Shreve, 1983, *Brownian Motion and Stochastic Calculus*, New York: Springer Verlag.

Lo, A. and J. Wang, 1994, "Implementing Option Pricing Models When Asset Returns Are Predictable," MIT Sloan School Discussion Paper.

Mahieu, R. and P. Schotman, 1994, "Stochastic Volatility and the Distribution of Exchange Rate News," LIFE Discussion Paper.

Müller, U.A., M.M. Dacorogna, R.D. Davé, R.B. Olsen, O.V. Pictet and J.E. von Weizsäcker, 1995, "Volatilities of Different Time Resolutions - Analyzing the Dynamics of Market Components," Olsen & Associates Research Group preprint.

Neumann, M.J.M., 1984, "Intervention in the mark/dollar market: The authorities' reaction function," *Journal of International Money and Finance* 3, 233-39.

Bossaerts, P., Härdle, W. and Hafner, Ch. (1996)
Foreign exchange rates have surprising Volatility.

Athens Conference on Applied Probability ad Time Series Ted Hannan memory Volume. LNS 115, p 55-72
Editor: P. Robinson, M Rosenblatt Springer Verlag

Pagan, A. and A. Ullah, 1988 "The Econometric Analysis of Models with Risk Terms," *Journal of Applied Econometrics* 3, 87-105.

Pagan, A. and G.W. Schwert, 1990 "Alternative Models for Conditional Stock Volatility," *Journal of Econometrics* 45, 267-290.

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

Tsybakov, A., 1986, "Robust Reconstruction of Functions by the Local-Approximation Method," *Problems of Information Transmission* 22, 133-46.

XploRe Systems, 1993, *XploRe 3.2*, Institut für Statistik und Ökonometrie, Humboldt-Universität zu Berlin.

Zakoïan, J.-M., 1991, "Threshold Heteroscedastic Models," INSEE Working Paper.

Athens Conference on Applied Probability ad Time Series Ted Hannan memory Volume. LNS 115, p 55-72
Editor: P. Robinson, M Rosenblatt Springer Verlag

Table 1
Serial correlations of FX quote changes and residuals of the CHARN model

		FX quote changes			CHARN residuals (optimal bandwidth)			CHARN residuals (low bandwidth)		
	order	y_t	$\|y_t\|$	$(y_t)^2$	ξ_t	$\|\xi_t\|$	$(\xi_t)^2$	ξ_t	$\|\xi_t\|$	$(\xi_t)^2$
DEM/USD										
	1	-.007	.220	.204	-.001	.121	.043	-.003	.005	-.004
	2	-.022	.173	.119	-.015	.145	.075	-.011	.113	.051
	3	-.028	.151	.089	-.025	.132	.075	-.021	.102	.055
	6	.003	.121	.071	.001	.111	.068	.002	.088	.049
	12	-.007	.107	.059	-.008	.094	.046	-.008	.079	.038
	24	.001	.080	.033	.003	.072	.029	.005	.064	.032
	48	-.002	.057	.021	-.002	.054	.024	-.003	.045	.019
YEN/USD										
	1	-.062	.295	.238	-.001	.099	.020	-.001	.016	.007
	2	-.010	.206	.146	-.011	.148	.069	-.010	.115	.050
	3	-.015	.164	.074	-.018	.127	.048	-.014	.105	.042
	6	-.006	.133	.061	-.007	.105	.045	-.008	.085	.039
	12	.004	.118	.047	-.000	.097	.037	-.008	.079	.026
	24	-.010	.111	.058	-.008	.091	.044	-.009	.077	.035
	48	.004	.084	.049	-.005	.068	.029	.001	.053	.022

Bossaerts, P., Härdle, W. and Hafner, Ch. (1996)
Foreign exchange rates have surprising Volatility.

Athens Conference on Applied Probability ad Time Series Ted Hannan memory Volume. LNS 115, p 55-72
Editor: P. Robinson, M Rosenblatt Springer Verlag

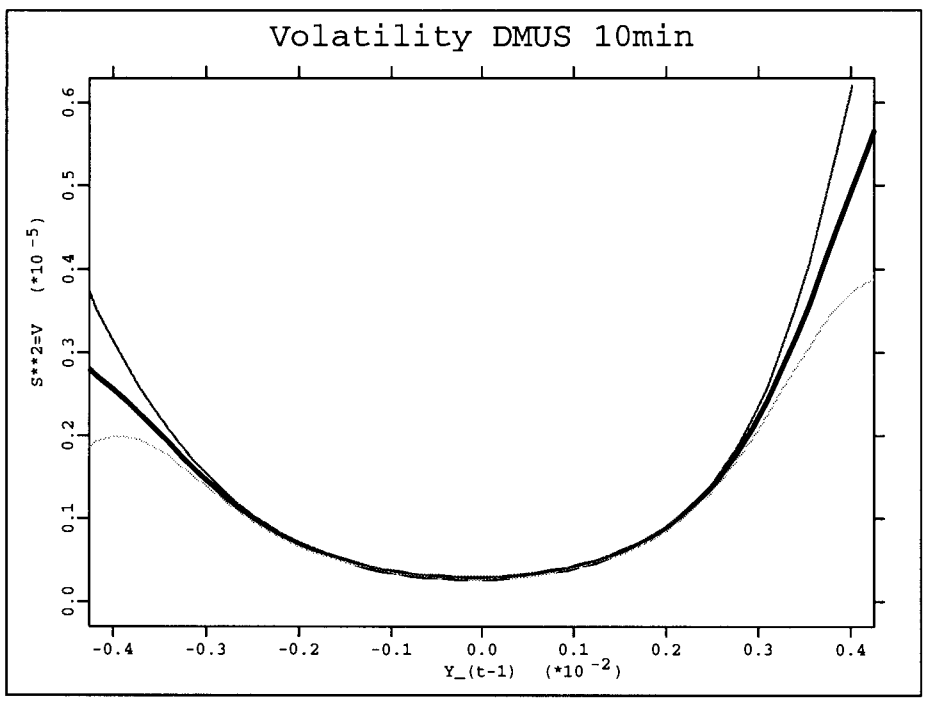


Figure 1: Estimated conditional variance function with 95% confidence bands in a conditionally heteroscedastic nonlinearly autoregressive model of changes in DEM/USD quotes over ten-minute intervals in deformed time during the period 1 Oct 92/30 Sep 93. The estimates were obtained by locally fitting linear functions using a quartic kernel. Crossvalidation determined the bandwidth size.

Bossaerts, P., Härdle, W. and Hafner, Ch. (1996)
Foreign exchange rates have surprising Volatility.

Athens Conference on Applied Probability ad Time Series Ted Hannan memory Volume. LNS 115, p 55-72
Editor: P. Robinson, M Rosenblatt Springer Verlag

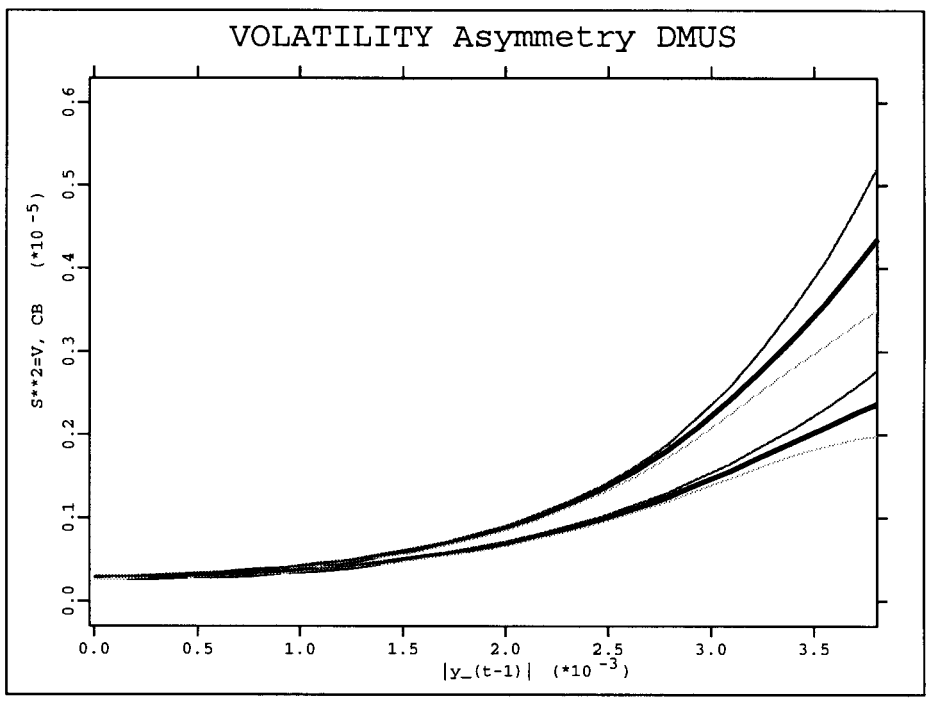


Figure 2: Estimated conditional variance function against absolute values of its argument, with 95% confidence bands in a conditionally heteroscedastic nonlinearly autoregressive model of changes in DEM/USD quotes over ten-minute intervals in deformed time during the period 1 Oct 92/30 Sep 93. The estimates were obtained by locally fitting linear functions using a quartic kernel. Crossvalidation determined the bandwidth size.

Athens Conference on Applied Probability ad Time Series Ted Hannan memory Volume. LNS 115, p 55-72
Editor: P. Robinson, M Rosenblatt Springer Verlag

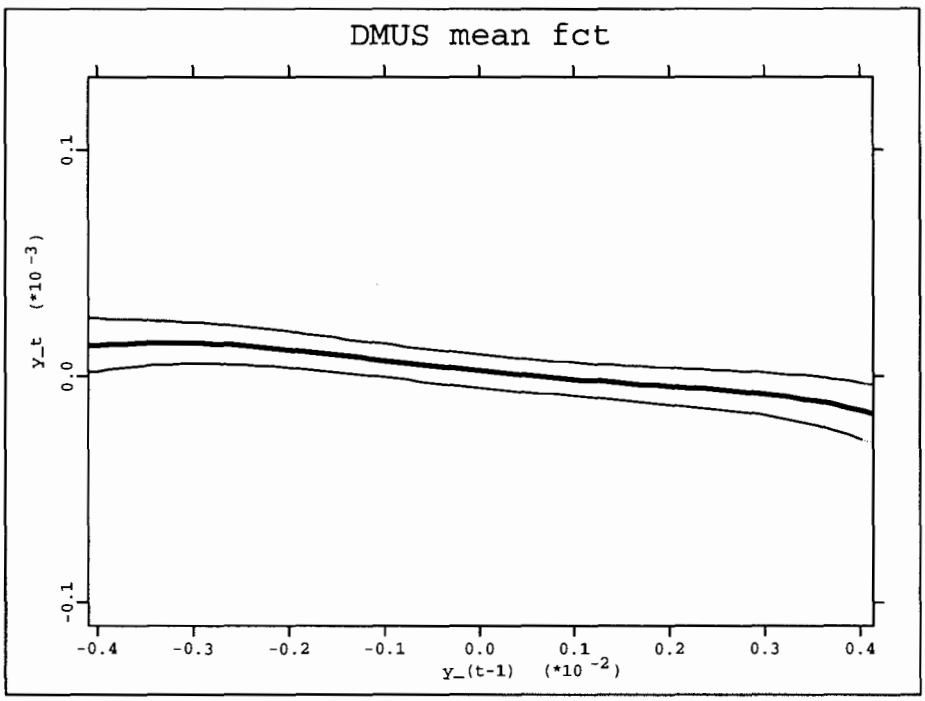


Figure 3: Estimated conditional mean function with 95% confidence bands in a conditionally heteroscedastic nonlinearly autoregressive model of changes in DEM/USD quotes over ten-minute intervals in deformed time during the period 1 Oct 92/30 Sep 93. The estimates were obtained by locally fitting linear functions using a quartic kernel. Crossvalidation determined the bandwidth size.

Bossaerts, P., Härdle, W. and Hafner, Ch. (1996)
Foreign exchange rates have surprising Volatility.

Athens Conference on Applied Probability ad Time Series Ted Hannan memory Volume. LNS 115, p 55-72
Editor: P. Robinson, M Rosenblatt Springer Verlag

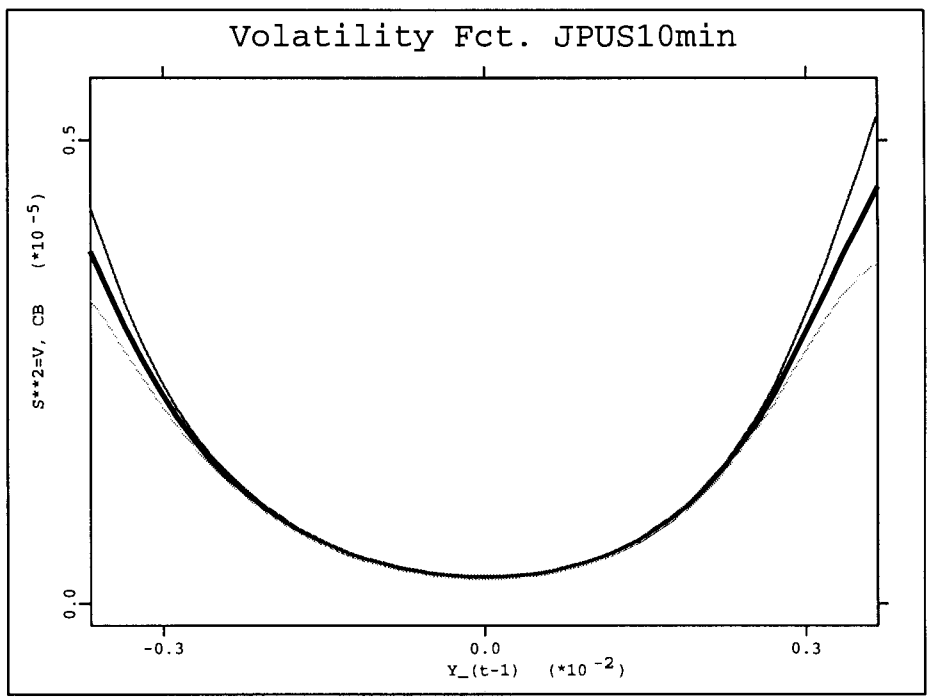


Figure 4: Estimated conditional variance function with 95% confidence bands in a conditionally heteroscedastic nonlinearly autoregressive model of changes in YEN/USD quotes over ten-minute intervals in deformed time during the period 1 Oct 92/30 Sep 93. The estimates were obtained by locally fitting linear functions using a quartic kernel. Crossvalidation determined the bandwidth size.

Bossaerts, P., Härdle, W. and Hafner, Ch. (1996)
Foreign exchange rates have surprising Volatility.

Athens Conference on Applied Probability ad Time Series Ted Hannan memory Volume. LNS 115, p 55-72
Editor: P. Robinson, M Rosenblatt Springer Verlag

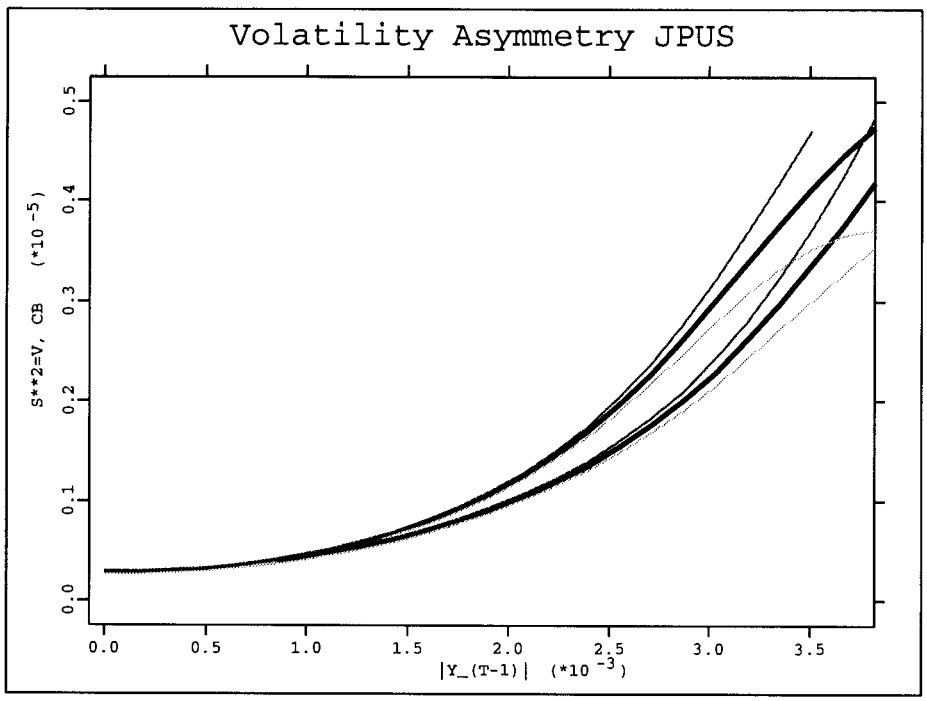


Figure 5: Estimated conditional variance function against absolute values of its argument, with 95% confidence bands in a conditionally heteroscedastic nonlinearly autoregressive model of changes in YEN/USD quotes over ten-minute intervals in deformed time during the period 1 Oct 92/30 Sep 93. The estimates were obtained by locally fitting linear functions using a quartic kernel. Crossvalidation determined the bandwidth size.

Bossaerts, P., Härdle, W. and Hafner, Ch. (1996)
Foreign exchange rates have surprising Volatility.

Athens Conference on Applied Probability ad Time Series Ted Hannan memory Volume. LNS 115, p 55-72
Editor: P. Robinson, M Rosenblatt Springer Verlag

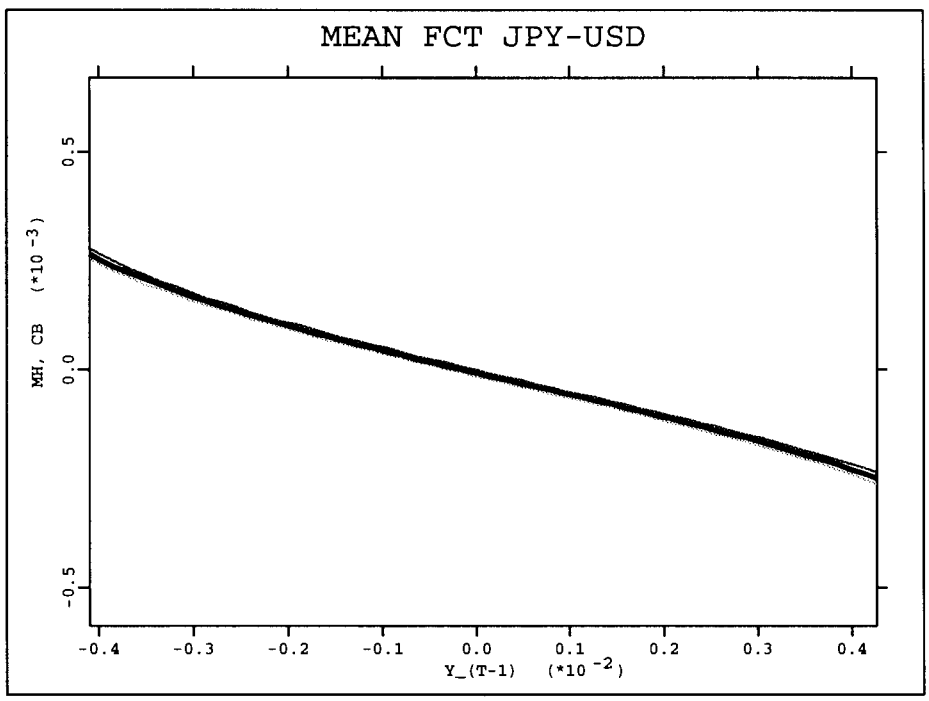


Figure 6: Estimated conditional mean function with 95% confidence bands in a conditionally heteroscedastic nonlinearly autoregressive model of changes in YEN/USD quotes over ten-minute intervals in deformed time during the period 1 Oct 92/30 Sep 93. The estimates were obtained by locally fitting linear functions using a quartic kernel. Crossvalidation determined the bandwidth size.

Bossaerts, P., Härdle, W. and Hafner, Ch. (1996)
Foreign exchange rates have surprising Volatility.

Athens Conference on Applied Probability ad Time Series Ted Hannan memory Volume. LNS 115, p 55-72
Editor: P. Robinson, M Rosenblatt Springer Verlag

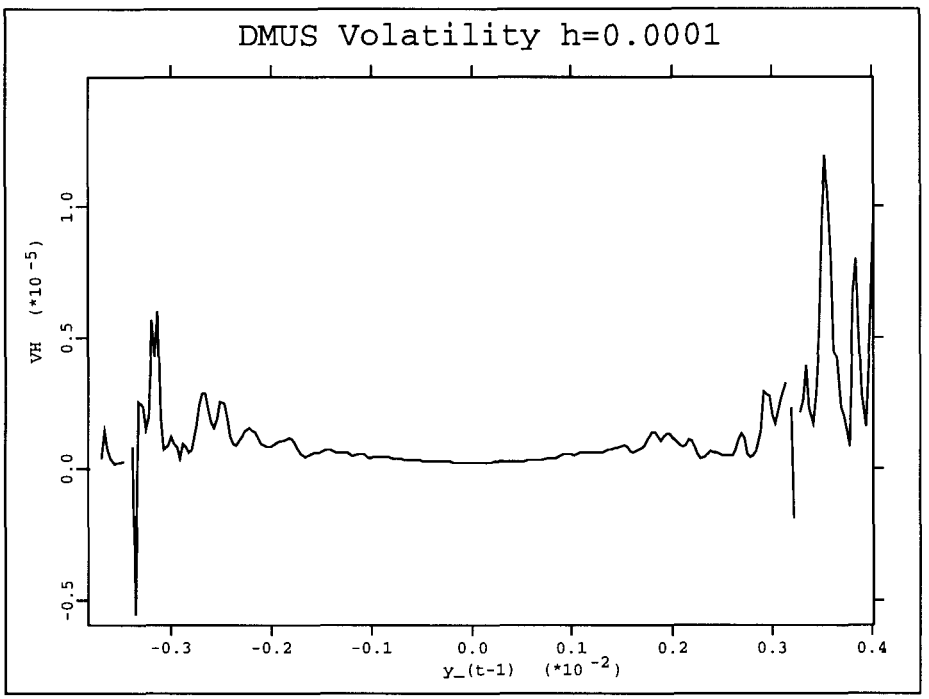


Figure 7: Estimated conditional variance function with 95% confidence bands in a conditionally heteroscedastic nonlinearly autoregressive model of changes in DEM/USD quotes over ten-minute intervals in deformed time during the period 1 Oct 92/30 Sep 93. The estimates were obtained by locally fitting linear functions using a quartic kernel. Bandwidth size (h): 0.0001.

Bossaerts, P., Härdle, W. and Hafner, Ch. (1996)
Foreign exchange rates have surprising Volatility.

Athens Conference on Applied Probability ad Time Series Ted Hannan memory Volume. LNS 115, p 55-72
Editor: P. Robinson, M Rosenblatt Springer Verlag

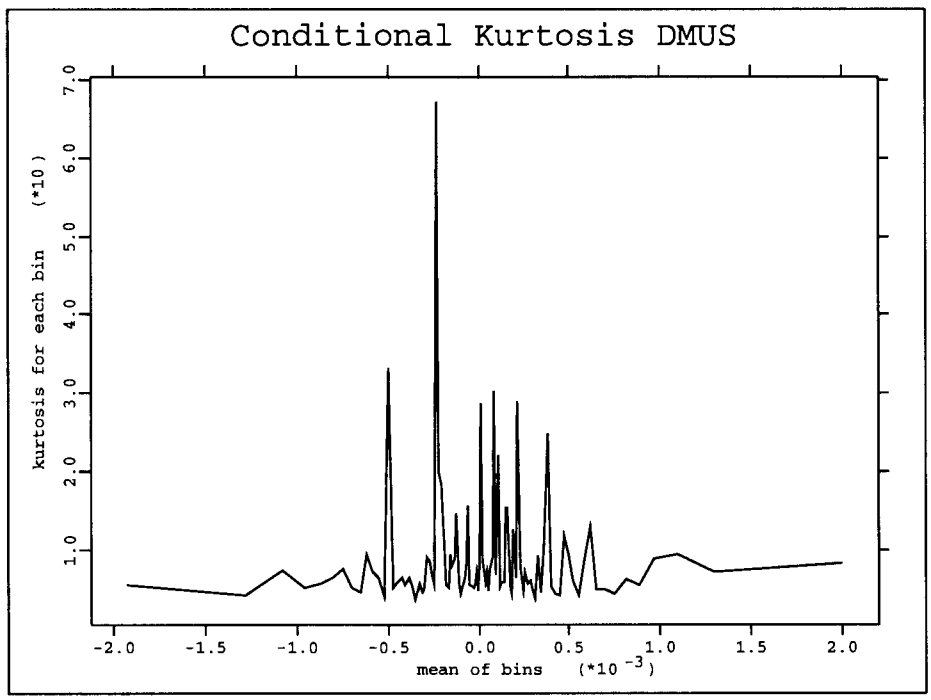


Figure 8: Estimated conditional kurtosis of changes in DEM/USD quotes over ten-minute intervals in deformed time during the period 1 Oct 92/30 Sep 93. The estimates were obtained by binning the data and computing the sample kurtosis per bin.

Bossaerts, P., Härdle, W. and Hafner, Ch. (1996)
Foreign exchange rates have surprising Volatility.

Athens Conference on Applied Probability ad Time Series Ted Hannan memory Volume. LNS 115, p 55-72
Editor: P. Robinson, M Rosenblatt Springer Verlag

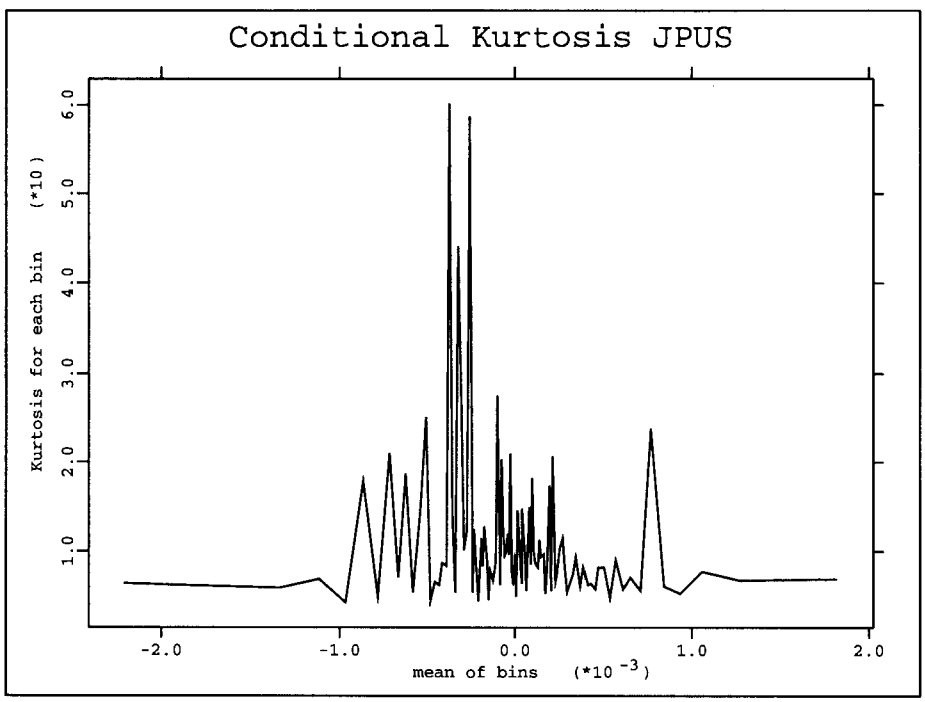


Figure 9: Estimated conditional kurtosis of changes in YEN/USD quotes over ten-minute intervals in deformed time during the period 1 Oct 92/30 Sep 93. The estimates were obtained by binning the data and computing the sample kurtosis per bin.

Bossaerts, P., Härdle, W. and Hafner, Ch. (1996)
Foreign exchange rates have surprising Volatility.

Statistical Theory and Computational Aspects of Smoothing. Härdle, W. and Schimek, M. (eds.)
Physika Verlag, Heidelberg. ISBN 3-7908-0930-6

Nonparametric estimation of additive separable regression models

R. Chen[1], W. Härdle[2], O.B. Linton[3] and E. Severance-Lossin[4]

[1]Department of Statistics, Texas A& M University, College Station, TX 77843, U.S.A.
[2]Humboldt-Universität zu Berlin, Wirtschaftswissenschaftliche Fakultät, Institut für Statistik und Ökonometrie Spandauer Str. 1, D-10178 Berlin, Germany
[3]Cowles Foundation for Research in Economics, Yale University, New Haven, CT 06520, U.S.A.
[4]Humboldt-Universität zu Berlin, Wirtschaftswissenschaftliche Fakultät, Institut für Statistik und Ökonometrie, Spandauer Str. 1, D-10178 Berlin, Germany

Summary

Additive regression models have been shown to be useful in many situations. Numerical estimation of these models is usually done using the iterative back-fitting technique. This paper proposes an estimator for additive models with an explicit 'hat matrix' which does not use iteration. The asymptotic normality of the estimator is proved. We also investigate a variable selection procedure using the proposed estimator and prove that asymptotically the procedure finds the correct variable set with probability 1. A simulation study is presented investigating the practical performance of the procedure.

1 Introduction

An additive nonparametric regression model has the form

$$m(x) = E(Y \mid X = x) = c + \sum_{\alpha=1}^{d} f_\alpha(x_\alpha), \tag{1}$$

where Y is a scalar dependent variable, $X = (X_1, \cdots, X_d)$ is a vector of explanatory variables, c is a constant and $\{f_\alpha(\cdot)\}_{\alpha=1}^{d}$ is a set of unknown functions satisfying $E[f(X_\alpha)] = 0$, and $x = (x_1, \cdots, x_d)$. Additive models of this form have been shown to be useful in practice: they naturally generalize the linear regression models and allow interpretation of marginal changes i.e.

Chen, R., Härdle, W., Linton, O. and Severance-Lossin, E. (1996)
Nonparametric Estimation of Additive Separable Regression Models.

Statistical Theory and Computational Aspects of Smoothing. Härdle, W. and Schimek, M. (eds.)
Physika Verlag, Heidelberg. ISBN 3-7908-0930-6

the effect of one variable on the mean function m holding all else constant. They are also interesting from a theoretical point of view since they combine flexible nonparametric modeling of many variables with statistical precision that is typical for just one explanatory variable. This paper is concerned with variable selection and direct estimation of the functions $f_\alpha(\cdot)$ and $m(\cdot)$ in an additive regression model (1).

To our knowledge model (1) has been first considered in the context of input-output analysis by Leontief (1947) who called it *additive separable*. In the statistical literature the additive regression model has been introduced in the early eighties, and promoted largely by the work of Buja, Hastie and Tibshirani (1989) and Hastie and Tibshirani (1990). It has lead to the development of a variety of theoretical results and to many applications implemented using modern software. Stone (1985, 1986) proved that model (1) can be estimated with a one-dimensional rate of convergence typical for estimating a single function f_α of one regressor only.

Buja, Hastie and Tibshirani (1989, eq (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). Unfortunately, not many statistical measures of this procedure like bias and variance have been fully derived in closed form.

We assume that model (1) holds exactly, i.e. the regression function is additive. For this case, Linton and Nielsen (1995) proposed a method of estimating the additive components f_α. Their method is to estimate a functional of m by marginal integration; under the additive structure this functional is f_α up to a constant. Their analysis is restricted to the case of dimension $d = 2$. Tjøstheim and Auestad (1994) proposed a similar estimator, mistakenly called 'projector', for time series but did not fully derive its asymptotic properties, specifically its bias.

The same model has been examined by Härdle and Tsybakov (1995) for general d under the assumption that the covariates are mutually independent. They introduced a principal component-like procedure for selecting important variables based on the variance of the estimated components, see also Maljutov and Wynn (1994).

The present paper improves upon these earlier results in various ways. First, a direct estimator based on marginal integration is proposed thereby avoiding iteration. Second, the explanatory variables are allowed to be correlated with a joint density p that does not factorize. This improves upon the paper by Härdle and Tsybakov (1995). Third, the dimension of X is

Statistical Theory and Computational Aspects of Smoothing. Härdle, W. and Schimek, M. (eds.)
Physika Verlag, Heidelberg. ISBN 3-7908-0930-6

not restricted to dimension $d = 2$ as in Linton and Nielsen (1995). Fourth, the 'hat matrix' of the proposed estimator is of less complicated form than in backfitting. Fifth, we give the exact asymptotic bias of our estimator thereby improving Tjøstheim and Auestad (1994). In addition to extending results on the estimator a procedure is given for selecting significant regressors.

The 'integration idea' is based on the following observation. If $m(x) = E(Y \mid X = x)$ is of the additive form (1), and the joint density of $X_{i\underline{\alpha}} = X_{i1}, \cdots, X_{i(\alpha-1)}, X_{i(\alpha+1)}, \cdots, X_{id}$ is denoted as $p_{\underline{\alpha}}$, then for a fixed $x_\alpha \in \mathbf{R}$,

$$f_\alpha(x_\alpha) + c = \int m(x_1, \cdots, x_\alpha, \cdots, x_d) p_{\underline{\alpha}}(x_{\underline{\alpha}}) \prod_{\beta \neq \alpha} dx_\beta, \tag{2}$$

where $x_{\underline{\alpha}} = (x_1, \cdots, x_{\alpha-1}, x_{\alpha+1} \cdots, x_d)$, provided $E[f_\beta(X_\beta)] = 0, \beta = 1, \cdots, d$. The idea is to estimate the function $m(\cdot)$ with a multidimensional kernel estimator and then to integrate out the variables other than X_α. We shall establish the asymptotic normal distribution of the estimator for f_α and derive explicitly its bias and variance. In obtaining this result we shall see that the rate of convergence for estimating the mean function m is $n^{2/5}$, typical for regression smoothing with just one explanatory variable.

The variable selection problem is important for practical use of additive regression modeling. It has been addressed by many authors. Often there are many predictor variables and we wish to select those components that contribute much explanation. We analyze here a procedure based on the size of $S_\alpha = E[f_\alpha^2(X_\alpha)]$. A component function f_α will be called significant if $S_\alpha \geq s_0$, where s_0 is a defined threshold level. We give an estimator for the set of significant functions and derive an upper bound (inverse to the sample size n) of the probability of misspecifying this set of significant component functions.

The rest of the paper is organized as follows. In section 2, we introduce the technique of estimating the functions in the additive model. In section 3, we propose a variable selection procedure which uses the estimator proposed in section 2 and prove that asymptotically it finds the correct variable set with probability 1. Section 4 provides a simulation study and a real example. The detailed conditions and proofs of the theorems are given in the appendix.

2 The Estimator

Let $(X_{i1}, \cdots, X_{id}, Y_i)$, $i = 1, \cdots, n$ be a random sample from the following additive model

$$Y_i = c + \sum_{\alpha=1}^{d} f_\alpha(X_{i\alpha}) + \epsilon_i, \tag{3}$$

where the ϵ_i have mean 0, finite variance $\sigma^2(X_i)$, and are mutually independent conditional on the $X's$. The functions $f_\alpha(\cdot)$ are assumed to have zero

Statistical Theory and Computational Aspects of Smoothing. Härdle, W. and Schimek, M. (eds.)
Physika Verlag, Heidelberg. ISBN 3-7908-0930-6

mean $\int f_\alpha(w)p_\alpha(w)dw = 0$, where $p_\alpha(\cdot)$ is the marginal density of X_α. Let $m(x_1, \cdots, x_d) = c + \sum_{\alpha=1}^{d} f_\alpha(x_\alpha)$ be the mean function. Then for a fixed x, the functional

$$\int m(x)p_{\underline{\alpha}}(x_{\underline{\alpha}}) \prod_{\beta \neq \alpha} dx_\beta$$

is $f_\alpha(x_\alpha) + c$. Let $K(\cdot)$ and $L(\cdot)$ be kernel functions with finite support. Let $K_h(\cdot) = h^{-1}K(\cdot/h)$ and define $L_g(\cdot)$ similarily. Using a multidimensional Nadaraya-Watson estimator (Nadaraya 1964, Watson 1964) to estimate the mean function $m(\cdot)$, we average over the observations to obtain the following estimator. For $1 \leq \alpha \leq d$ and any x in the domain of $f_\alpha(\cdot)$, define, for $h > 0$, $g > 0$,

$$\begin{aligned}
\hat{f}_\alpha(x_\alpha) &= \tfrac{1}{n}\textstyle\sum_{i=1}^{n} \tilde{m}(X_{i1}, \cdots, X_{i(\alpha-1)}, x_\alpha, X_{i(\alpha+1)}, \cdots, X_{id}) \\
&= \tfrac{1}{n}\textstyle\sum_{i=1}^{n} \left[\frac{\sum_{l=1}^{n} L_g(X_{i\underline{\alpha}} - X_{l\underline{\alpha}})K_h(X_{i\alpha} - x_\alpha)Y_l}{\sum_{t=1}^{n} L_g(X_{i\underline{\alpha}} - X_{t\underline{\alpha}})K_h(X_{t\alpha} - x_\alpha)} \right] \qquad (4) \\
&= \textstyle\sum_{l=1}^{n} \left\{ \frac{1}{n}\sum_{i=1}^{n} \left[\frac{L_g(X_{i\underline{\alpha}} - X_{l\underline{\alpha}})K_h(X_{i\alpha} - x_\alpha)}{\sum_{t=1}^{n} L_g(X_{i\underline{\alpha}} - X_{t\underline{\alpha}})K_h(X_{t\alpha} - x_\alpha)} \right] \right\} Y_l.
\end{aligned}$$

If the X's were independent, we might use $\frac{1}{n}\sum_{l=1}^{n} Y_l \frac{K_h(X_{i\alpha} - x_\alpha)}{\frac{1}{n}\sum_{t=1}^{n} K_h(X_{t\alpha} - x_\alpha)}$ to estimate $f_\alpha(x_\alpha)$. This is a one-dimensional Nadaraya-Watson estimator. However, this estimator has larger variance in comparison to our estimator even in this restricted situation, see Härdle and Tsybakov (1995).

To illustrate our method, we simulated a data set $(X_{i1}, \cdots, X_{id}, Y_i)$, $i = 1, \cdots, 200$, according to model (3) with ϵ_i distributed as $N(0, 0.5^2)$, $f_1(x_1) = x_1^2 - 1$, $f_2(x_2) = x_2/2$ and $X_1, X_2 \sim N(0, 1)$ with $cov(X_1, X_2) = 0.2$ and $c = 0$. The data points are shown on top of the needles in Figure 1. The parabolic shape of f_1 is quite visible but the linear form of f_2 is less evident from the projection onto the (x_2, y) plane. This becomes more clear from Figure 2 where we apply the estimator (4) to estimate the additive component functions. In the Figure we show the estimated curves dashed lines and the true curves as solid lines. Both $\hat{f}_1$ and $\hat{f}_2$ show some smoothing bias at the boundary of the support but capture the general form of the component functions quite well. In both cases we used the bandwidth $h = 0.5$ and $g = 1.5$ and a Normal kernel. We also applied the backfitting procedure to this data set and obtained almost identical curves, see Härdle, Klinke and Turlach (1995, chapter 1).

The bias and variance of the integration estimator are given in the following theorem. Denote by $p(x_1, \cdots, x_d)$ the joint density of $X_{11}, \cdots, X_{1d}$.

Statistical Theory and Computational Aspects of Smoothing. Härdle, W. and Schimek, M. (eds.)
Physika Verlag, Heidelberg. ISBN 3-7908-0930-6

THEOREM 1. *Suppose that conditions (A1) - (A6) given in section 5 hold. Let* $h = \gamma_0 n^{-1/5}$. *Assume that the bandwidths* g *and* h *satisfy* $nhg^{d-1}/\log n \to \infty$, *and that the order of* L *is* $q > \frac{d-1}{2}$. *Then*

$$n^{2/5}\{\hat{f}_\alpha(x_\alpha) - f_\alpha(x_\alpha) - c\} \xrightarrow{L} N\left\{b_\alpha(x_\alpha), v_\alpha^2(x_\alpha)\right\}$$

where

$$b_\alpha(x_\alpha) = \gamma_0^2 \mu_2(K)\left\{\tfrac{1}{2} f_\alpha''(x_\alpha) + f_\alpha'(x_\alpha)\int \frac{p_{\underline{\alpha}}(x_{\underline{\alpha}})}{p(x)} \frac{\partial p}{\partial x_\alpha}(x)\, dx_{\underline{\alpha}}\right\}$$

$$v_\alpha^2(x_\alpha) = \gamma_0^{-1} \|K\|_2^2 \int \frac{\sigma^2(x) p_{\underline{\alpha}}^2(x_{\underline{\alpha}})}{p(x)} dx_{\underline{\alpha}}.$$

From this theorem, we see that the rate of convergence to the asymptotic normal limit distribution does not suffer from the 'curse of dimensionality.' To achieve this rate of convergence, though, we must impose some restrictions on the bandwidth sequences. This condition is needed for bias reduction of components $\beta \neq \alpha$. Note that the above bandwidth condition does not exclude the 'optimal one dimensional smoothing bandwidth' $g = h = n^{-1/5}$ for $d \leq 4$. More importantly one can take $g = o(n^{-1/5})$, leaving only the first two terms in the bias expression. For higher dimensions, $d \geq 5$, though we can no longer use g at the rate $n^{-1/5}$ and the terms involving g dominate. To avoid this problem and obtain the one-dimensional rate of convergence we must reduce bias in the directions not of interest. This can be done by taking L to be a higher order kernel.

Define $\hat{m}(x_1, \cdots, x_d) = \sum_{\alpha=1}^d \hat{f}_\alpha(x_\alpha) - (d-1)\hat{c}$, where $\hat{c} = n^{-1}\sum_{i=1}^n Y_i$. The following theorem gives the asymptotic distribution of $\hat{m}$ and shows that asymptotically the covariance between $\hat{f}_\alpha(x_\alpha)$ and $\hat{f}_\beta(x_\beta)$ is of smaller order than the variances of each component function.

THEOREM 2. *Under the assumptions of Theorem 1,*

$$n^{2/5}\{\hat{m}(x_1, \cdots, x_d) - m(x_1, \cdots, x_d)\} \xrightarrow{L} N\left(b(x), v^2(x)\right),$$

where $b(x) = \sum_{\alpha=1}^d b_\alpha(x_\alpha)$ *and* $v^2(x) = \sum_{\alpha=1}^d v_\alpha^2(x_\alpha)$.

3 Variable Selection Procedure

To establish a variable selection procedure we first show that $S_\alpha = \int f_\alpha^2(w) p_\alpha(w) dw$ can be estimated $n^{1/2}$ consistently. The following theorem establishes this result.

Statistical Theory and Computational Aspects of Smoothing. Härdle, W. and Schimek, M. (eds.)
Physika Verlag, Heidelberg. ISBN 3-7908-0930-6

THEOREM 3. *Suppose that conditions (A1) - (A6) given in section 5 hold. Let* $h = \gamma_0 n^{-1/4}$. *Assume that the bandwidths* g *and* h *satisfy* $nhg^{d-1}/\log n \to \infty$, *and that the order of* L *is* $q > \frac{d-1}{2}$. *Then,*

$$\hat{S}_\alpha = \frac{1}{n}\sum_{i=1}^{n}\{\hat{f}_\alpha(X_{i\alpha})\}^2 = S_\alpha + O_p(n^{-1/2}).$$

As in the linear regression, it is important to select a suitable subset among all the available predictors for building an additive model. We propose the following variable selection procedure. Let A be a subset of $\{1, \cdots, d\}$ such that $A = \{\alpha : S_\alpha > 0\}$ so that for $\alpha \notin A$, $S_\alpha = 0$. Note that since A is finite $\{S_\alpha \mid \alpha \in A\}$ is bounded away from zero. Since $\hat{S}_\alpha$ estimates the functional S_α a large $\hat{S}_\alpha$ implies that the variable X_α should be included in the model. Our variable selection procedure selects the indices α such that $\hat{S}_\alpha > b_n$ where b_n is some prescribed level. Denote by $\widehat{A} = \{\alpha : \hat{S}_\alpha \geq b_n\}$, the set of estimated coefficients. The following theorem states the asymptotic correctness of this selection procedure.

THEOREM 4. *Under the assumptions of Theorem 3 and* $b_n = O(n^{-1/4})$,

$$P(\widehat{A} \neq A) \leq \frac{C}{n},$$

for some constant C.

Since $\hat{S}_\alpha$ estimates S_α we can view $\hat{S}_\alpha / \sum_{\beta=1}^{d} \hat{S}_\beta$ as an estimate of the portion of variation in Y explained by X_α. This allows for a meaningful finite sample interpretation of the test statistic.

4 Simulation Study and Application

In this section we investigated some small sample properties of our estimator through a simulation study. We concentrated on the following questions. First, how variable is the estimator and how much bias do we have to expect? Second, how does the precision depend on the bandwidth choice? Third, how much more variable is the estimator in higher dimensions. We also applied the additive estimator in an econometric context investigating livestock production of Wisconsin farms.

First we continue with our introductory example with parabolic - linear functions. We simulated 250 data sets of size $n = 200$ with $f_1(x_1) = x_1^2 - 1$ and $f_2(x_2) = x_2/2$. The covariance structure of X and the error distribution were

Statistical Theory and Computational Aspects of Smoothing. Härdle, W. and Schimek, M. (eds.)
Physika Verlag, Heidelberg. ISBN 3-7908-0930-6

as in the introductory example. The simulated 90 % confidence intervals for f_1 and f_2 are shown as thin lines in figure whereas the mean values are shown as thick line. For all simulations we used the bandwidths $h = 0.5$ and $g = 1.5$ and a Normal kernel. In both graphs the true curve has been subtracted for better bias judgment. The smoothing bias becomes visible in figure at the boundaries due to the parabolic shape. It is less a problem for the linear f_2 as Theorem 1 suggests. In both cases the true curve lies well within the computed confidence limits and the shape of the true curve is well reflected by the confidence intervals. The intervals become wider at the boundaries since we have less observations there. We also investigated the effect of bandwidth choice on the above findings and found that, for example, at $x = 0$, the center of the confidence interval increases as the bandwidth increases and the band becomes smaller. This is in full accordance with Theorem 1 showing the dependence of the asymptotic bias on the curvature and the variance as a function of bandwidth. We observe the same phenomenon for dimension $d = 5$, through simulation.

Next we consider a real example. We consider the estimation of a production function for livestock in Wisconsin. A typical economic model in this context is a Cobb-Douglas production function,

$$log(Y) = \alpha_1 log(X_1) + \ldots + \alpha_d log(X_d).$$

The model is additive with parametric linear components. We replace the linear components in the model with arbitrary, up to smoothness conditions, nonlinear functions. This gives a very flexible model of a strongly separable production function.

We use a subset (250 observations) of an original data set of over 1000 Wisconsin farms collected by the Farm Credit Service of St. Paul, Minnesota in 1987. The data were cleaned, removing outliers and incomplete records and selecting only farms that only produce animal outputs. The data consists of farm level inputs and outputs measured in dollars. The output (Y) used in this analysis is livestock, and the input variables used are family labor, hired labor, miscellaneous inputs (repairs, rent, custom hiring, supplies, insurance, gas, oil, and utilities), animal inputs (purchased feed, breeding, and veterinary services), and intermediate run assets (assets with a useful life of one to 10 years) resulting in a five dimensional X variable.

We applied the additive kernel estimator to the farm data set. The results are displayed in Figure 4. The curves are shown together with their marginal scatterplots. They are all quite linear except for the component X_2 (hired labor). Note that the smooth curves do not reflect the form of the scatterplots since our estimator does not use marginal smoothing.

Statistical Theory and Computational Aspects of Smoothing. Härdle, W. and Schimek, M. (eds.)
Physika Verlag, Heidelberg. ISBN 3-7908-0930-6

A Appendix

We assume the following conditions hold:

(A1) *The kernel function $K(\cdot)$ is bounded, nonnegative, compactly supported, Lipschitz continuous, with $\int K(u)du = 1$ and $\int uK(u) = 0$. Let $||K||_2^2 = \int K^2(u)du < \infty$ and $\mu_2(K) = \int u^2K(u)du < \infty$.*

(A2) *The kernel function $L(\cdot)$ is bounded, nonnegative, compactly supported, Lipschitz continuous and $\int L(u)du = 1$. Let $\mu_i(L) = \int u^i L(u)du$, then $\mu_i(L) = 0, i = 1, 2, \cdots, q-1$, while $\mu_q(L) < \infty$ and $||L||_2^2 = \int L^2(u)du < \infty$.*

(A3) *The densities $p_\alpha(\cdot)$, $p_{\underline{\alpha}}(\cdot)$ and $p(\cdot)$ are bounded, Lipschitz continuous and bounded away from zero by a constant p_0.*

(A4) $E(\epsilon_i^4) < \infty$.

(A5) *The variance function, $\sigma^2(\cdot)$, is Lipschitz continuous.*

(A6) *The functions $f_\alpha(\cdot)$ have q Lipschitz continuous derivatives.*

Proof of Theorem 1: We use the following notation. Define $E_i[W] = E[W \mid X_i]$ and $E_*[W] = E[W \mid X_1, \cdots, X_n]$. Let

$$p_i = p(X_{i1}, \cdots, X_{i(\alpha-1)}, x_\alpha, X_{i(\alpha+1)}, \cdots, X_{id}),$$

and

$$\widehat{p_i} = \frac{1}{n}\sum_{l=1}^{n} L_g(X_{i\underline{\alpha}} - X_{l\underline{\alpha}})K_h(X_{l\alpha} - x_\alpha).$$

To simplify the notation we always write the α'th component first.

Note that by (2)

$$\frac{1}{n}\sum_{i=1}^{n} m(x_\alpha, X_{i\underline{\alpha}}) = f_\alpha(x_\alpha) + c + O_p(n^{-1/2}), \tag{5}$$

since $m(x_\alpha, X_{i\underline{\alpha}})$ are i.i.d. random variables with finite second moments. Then we can write

$$\begin{aligned} \widehat{f_\alpha}(x) - f_\alpha(x) - c &= \tfrac{1}{n}\sum_{i=1}^{n} \tfrac{\widehat{r_i} - \widehat{p_i} m(x_\alpha, X_{i\underline{\alpha}})}{\widehat{p_i}} + O_p(n^{-1/2}) \\ &\equiv \tfrac{1}{n}\sum_{i=1}^{n} \tfrac{\widehat{a_i}}{p_i} + O_p(n^{-1/2}), \end{aligned} \tag{6}$$

Statistical Theory and Computational Aspects of Smoothing. Härdle, W. and Schimek, M. (eds.)
Physika Verlag, Heidelberg. ISBN 3-7908-0930-6

where $\widehat{r}_i = \frac{1}{n}\sum_{l=1}^{n} L_g(X_{i\underline{\alpha}} - X_{l\underline{\alpha}})K_h(X_{l\alpha} - x_\alpha)Y_l$. It suffices to work with the first term on the right hand side, ignoring the $O_p(n^{-1/2})$ remainder. We separate this into a systematic "bias" and a stochastic "variance".

$$\frac{1}{n}\sum_{i=1}^{n}\frac{E_i(\widehat{a}_i)}{\widehat{p}_i} + \frac{1}{n}\sum_{i=1}^{n}\frac{\widehat{a}_i - E_i(\widehat{a}_i)}{\widehat{p}_i} \equiv T_{1n} + T_{2n}.$$

Then,

$$\begin{aligned} T_{1n} &= \tfrac{1}{n}\textstyle\sum_{i=1}^{n}\frac{E_i(\widehat{a}_i)}{p_i}\{1 + o_p(1)\} \\ T_{2n} &= \tfrac{1}{n}\textstyle\sum_{i=1}^{n}\frac{\widehat{a}_i - E_i(\widehat{a}_i)}{p_i}\{1 + o_p(1)\}, \end{aligned}$$

since, by Silverman (1986), $\sup\left|\frac{\widehat{p}_i - p_i}{p_i p_i}\right| = o_p(1)$. It remains to work with the first order approximations.

Let

$$\widetilde{T}_{1n} = \frac{1}{n}\sum_{i=1}^{n}\frac{E_i(\widehat{a}_i)}{p_i} \quad ; \quad \widetilde{T}_{2n} = \frac{1}{n}\sum_{i=1}^{n}\frac{\widehat{a}_i - E_i(\widehat{a}_i)}{p_i}.$$

We prove the theorem by showing:

I. $\widetilde{T}_{1n} = b_\alpha(x_\alpha) + o_p(n^{-2/5})$

II. $\widetilde{T}_{2n} = \sum_{j=1}^{n} w_{j\alpha}\epsilon_j + o_p(n^{-2/5})$,

where $w_{j\alpha} = \frac{1}{n}K_h(x_\alpha - X_{j\alpha})\frac{p_{\underline{\alpha}}(X_{j\underline{\alpha}})}{p(x_\alpha, X_{j\underline{\alpha}})}$, and $n^{2/5}\sum_{j=1}^{n} w_{j\alpha}\epsilon_j$ obeys a Central Limit Theorem with asymptotic variance as stated in Theorem 1. To see this note that

$$E\left[\left\{n^{2/5}\sum_{j=1}^{n} w_{j\alpha}\epsilon_j\right\}^2\right] = n^{4/5}\sum_{j=1}^{n} E\left[w_{j\alpha}^2\epsilon_j^2\right] = n^{9/5}E\left[w_{1\alpha}^2\epsilon_1^2\right],$$

since $w_{j\alpha}\epsilon_j$ are mean zero and i.i.d., and

$$E\left[w_{1\alpha}^2\epsilon_1^2\right] = \tfrac{1}{n^2}\int \sigma^2(z,w)K_h^2(x_\alpha - z)\frac{p_{\underline{\alpha}}^2(w)}{p^2(x_\alpha, w)}p(z,w)\,dzdw.$$

Statistical Theory and Computational Aspects of Smoothing. Härdle, W. and Schimek, M. (eds.)
Physika Verlag, Heidelberg. ISBN 3-7908-0930-6

Changing variables to $u = \frac{x_\alpha - z}{h}$ gives

$$\begin{aligned} E\left[w_{1\alpha}^2 \epsilon_1^2\right] &= \frac{1}{n^2 h} \int \sigma^2(x_\alpha + hu, w) K^2(u) \frac{p_{\underline{\alpha}}^2(w)}{p^2(x_\alpha, w)} p\left(x_\alpha + hu, w\right) du dw \\ &= n^{-9/5} \mu_2\left(K\right) \int \sigma^2(x_\alpha, w) \frac{p_{\underline{\alpha}}^2(w)}{p(x_\alpha, w)} dw + o(n^{-9/5}), \end{aligned}$$

by assumption (A5) and the bandwidth conditions. The Lindeberg condition, required for the CLT, follows from the existence of the fourth moments and the compact support of the kernels.

We now establish the approximations in **I** and **II**.

I. Consider $p_i^{-1} E_i(\widehat{a}_i)$, which is, in fact, an approximation of the conditional bias of the Nadaraya-Watson estimator at $(x_\alpha, X_{i\underline{\alpha}})$. This is,

$$\begin{aligned} p_i^{-1} E_i(\widehat{a}_i) &= E_i \left[\frac{1}{p_i} n^{-1} \sum_{l=1}^{n} L_g(X_{l\underline{\alpha}} - X_{i\underline{\alpha}}) K_h(X_{l\alpha} - x_\alpha) \left\{ Y_l - m(x_\alpha, X_{i\underline{\alpha}}) \right\} \right] \\ &= \frac{1}{p_i} \int L_g(w - X_{i\underline{\alpha}}) K_h(z - x_\alpha) \left\{ f_\alpha(z) - f_\alpha(x_\alpha) + \sum_{\beta \neq \alpha} f_\beta(w) - f_\beta(X_{i\beta}) \right\} \\ &\quad \times p_{\underline{\alpha}}(w) p_\alpha(z) dw dz, \end{aligned}$$

since $E_*\left[\epsilon_i\right] = 0$. We now change variables to $u = \frac{z - x_\alpha}{h}$ and $v = \frac{w - X_{i\underline{\alpha}}}{g}$, where v is a $d-1$-dimensional vector with typical component v_β, and find

$$\begin{aligned} p_i^{-1} E_i(\widehat{a}_i) &= \frac{1}{p_i} \int L(v) K(u) \left\{ f_\alpha(x_\alpha + hu) - f_\alpha(x_\alpha) + \sum_{\beta \neq \alpha} f_\beta(x_\beta + g v_\beta) - f_\beta(X_{i\beta}) \right\} \\ &\quad \times p_\alpha(x_\alpha + hu) p_{\underline{\alpha}}(X_{i\underline{\alpha}} + gv) du dv \\ &= h^2 \mu_2\left(K\right) \left\{ \frac{1}{2} f_\alpha''(x_\alpha) + \frac{f_\alpha'(x_\alpha)}{p_i} \frac{\partial p}{\partial x_\alpha}(x_\alpha, X_{i\underline{\alpha}}) \right\} + O_p(g^q), \end{aligned}$$

by assumptions (A1), (A2), (A3), and (A6). Since the $p_i^{-1} E_i(\widehat{a}_i)$ are independent and bounded we have

Statistical Theory and Computational Aspects of Smoothing. Härdle, W. and Schimek, M. (eds.)
Physika Verlag, Heidelberg. ISBN 3-7908-0930-6

$$\begin{aligned}\widetilde{T}_{1n} &= h^2\mu_2(K)\left\{\tfrac{1}{2}f''_\alpha(x_\alpha) + f'_\alpha(x_\alpha)\int \tfrac{p_{\underline{\alpha}}(z)}{p(x_\alpha,z)}\tfrac{\partial p}{\partial x_\alpha}(x_\alpha,z)dz\right\} + o_p\left(h^2\right) + O_p(g^q) + O_p(n^{-1/2}) \\ &= b_\alpha(x_\alpha) + o_p\left(n^{-2/5}\right).\end{aligned}$$

II. We now turn to the stochastic term,

$$\widetilde{T}_{2n} = \frac{1}{n}\sum_{i=1}^{n}\frac{\widehat{a}_i - E_i(\widehat{a}_i)}{p_i}.$$

We further write

$$\widehat{a}_i - E_i(\widehat{a}_i) = \widehat{a}_i - E_*(\widehat{a}_i) + E_*(\widehat{a}_i) - E_i(\widehat{a}_i).$$

II.1. We show that $\frac{1}{n}\sum_{i=1}^{n}\frac{\widehat{a}_i - E_*(\widehat{a}_i)}{p_i} = \sum_{j=1}^{n} w_{j\alpha}\epsilon_j + o_p\left(n^{-2/5}\right)$, where

$$\widehat{a}_i - E_*(\widehat{a}_i) = n^{-1}\sum_{j=1}^{n} K_h(x_\alpha - X_{j\alpha})L_g(X_{i\underline{\alpha}} - X_{j\underline{\alpha}})\epsilon_j.$$

Therefore,

$$\begin{aligned}\tfrac{1}{n}\textstyle\sum_{i=1}^{n}\frac{\widehat{a}_i - E_*(\widehat{a}_i)}{p_i} &= \tfrac{1}{n}\textstyle\sum_{i=1}^{n}\tfrac{1}{p_i}n^{-1}\sum_{j=1}^{n} K_h(x_\alpha - X_{j_\alpha})L_g(X_{i\underline{\alpha}} - X_{j\underline{\alpha}})\epsilon_j \\ &= n^{-1}\textstyle\sum_{j=1}^{n} K_h(x_\alpha - X_{j\alpha})\epsilon_j\left\{\tfrac{1}{n}\sum_{i=1}^{n}\tfrac{1}{p_i}L_g(X_{i\alpha} - X_{j\alpha})\right\} \\ &= n^{-1}\textstyle\sum_{j=1}^{n} w_{j\alpha}\epsilon_j\left\{1 + o_p(1)\right\}.\end{aligned} \tag{7}$$

The last equality is demonstrated as follows. Let

$$\eta_j = \frac{1}{n}\sum_{i=1}^{n}\frac{1}{p_i}L_g(X_{i\underline{\alpha}} - X_{j\underline{\alpha}}),$$

and break η_j into $E_j[\eta_j] + \{\eta_j - E_j[\eta_j]\}$. Then,

$$\begin{aligned}E_j[\eta_j] &= \textstyle\int \frac{1}{p(x_\alpha,z)}L_g(z - X_{j\underline{\alpha}})p_{\underline{\alpha}}(z)\,dz \\ &= \textstyle\int \frac{1}{p(x_\alpha, X_{j\underline{\alpha}}+gu)}L(u)p_{\underline{\alpha}}(X_{j\underline{\alpha}} + gu)du \\ &= \tfrac{p_{\underline{\alpha}}(X_{j\underline{\alpha}})}{p(x_\alpha, X_{j\underline{\alpha}})} + O_p(g^q).\end{aligned}$$

Statistical Theory and Computational Aspects of Smoothing. Härdle, W. and Schimek, M. (eds.)
Physika Verlag, Heidelberg. ISBN 3-7908-0930-6

Also,

$$E_j\left[\{\eta_j - E_j[\eta_j]\}^2\right] \leq \frac{1}{n}\int\left\{\frac{1}{p(x_\alpha,z)}L_g(z-X_{j\underline{\alpha}}) - \frac{p_{\underline{\alpha}}(X_{j\underline{\alpha}})}{p(x_\alpha,X_{j\underline{\alpha}})}\right\}^2 p_{\underline{\alpha}}(z)\,dz + O_p\left(n^{-1}g^{2q}\right)$$

$$= \frac{1}{n}\int\left\{\frac{1}{p(x_\alpha,z)}L_g(z-X_{j\underline{\alpha}})\right\}^2 p_{\underline{\alpha}}(z)\,dz + O_p\left(n^{-1}\right).$$

By a change of variables we get

$$E_j\left[\{\eta_j - E_j[\eta_j]\}^2\right] = \frac{1}{ng^{d-1}}\int\left\{\frac{1}{(px_\alpha,X_{j\underline{\alpha}}+gv)}L(v)\right\}^2 p_{\underline{\alpha}}(X_{j\underline{\alpha}}+gv)dv + O_p\left(n^{-1}\right)$$

$$= \frac{1}{ng^{d-1}}\frac{p_{\underline{\alpha}}(X_{j\underline{\alpha}})}{p^2(x_\alpha,X_{j\underline{\alpha}})}\|L\|_2^2 + O_p\left(n^{-1}\right) = o_p(1),$$

by the assumptions (A1), (A2), (A3) and the conditions on the bandwidths. Thus the last line in (7) is shown.

II.2. Next we show $\frac{1}{n}\sum_{i=1}^n \frac{E_*(\widehat{a}_i)-E_i(\widehat{a}_i)}{p_i} = o_p(n^{-2/5})$. Let

$$U_{0n} = \frac{1}{n}\sum_{i=1}^n \frac{E_*(\widehat{a}_i) - E_i(\widehat{a}_i)}{p_i} = \sum_{i=1}^n\sum_{j=1}^n \widetilde{\zeta}_{ij},$$

where $\widetilde{\zeta}_{ij} = \zeta_{ij} - \overline{\zeta}_i$ and $\overline{\zeta}_i = E_i(\zeta_{ij})$, with

$$\zeta_{ij} = \frac{1}{n^2 p_i}K_h(x_\alpha - X_{j\alpha})L_g(X_{i\underline{\alpha}} - X_{j\underline{\alpha}})\left\{m(X_{j\alpha},X_{j\underline{\alpha}}) - m(x_\alpha,X_{i\alpha})\right\}.$$

The double sum U_{0n} is mean zero. When $i=j$, we have

$$\zeta_{ii} = \frac{1}{n^2 g^{d-1}}L(0)\frac{1}{p_i}K_h(x_\alpha - X_{i\alpha})\left\{f(X_{i\alpha}) - f(x_\alpha)\right\}, \tag{8}$$

and $\sum_{i=1}^n \zeta_{ii} = O_p((nh)^{-1/2}(ng^{d-1})^{-1/2})$. We now calculate the variance of $\sum\sum_{i\neq j}\widetilde{\zeta}_{ij}$; this involves the following calculations

$$\sum\sum_{i\neq j}E(\widetilde{\zeta}_{ij}^2),\ \sum\sum_{i\neq j}E(\widetilde{\zeta}_{ij}\widetilde{\zeta}_{ji}),\ \sum\sum\sum_{i\neq j,i\neq k,j\neq k}E(\widetilde{\zeta}_{ij}\widetilde{\zeta}_{ik}),$$

since all other terms are mean zero by a conditioning argument. Now $E(\widetilde{\zeta}_{ij}\widetilde{\zeta}_{ik}) = E[E_i^2(\widetilde{\zeta}_{ij})]$, for $i\neq j, i\neq k, j\neq k$ using conditional independence. But

Statistical Theory and Computational Aspects of Smoothing. Härdle, W. and Schimek, M. (eds.)
Physika Verlag, Heidelberg. ISBN 3-7908-0930-6

$$E_i(\zeta_{ij}) = \frac{1}{n^2} E_i \left[\frac{1}{p_i} K_h(x_\alpha - X_{j\alpha}) L_g(X_{i\underline{\alpha}} - X_{j\underline{\alpha}}) \left\{ m(X_{j\alpha}, X_{j\underline{\alpha}}) - m(x_\alpha, X_{i\alpha}) \right\} \right]$$
$$= \frac{1}{n^2} O(h^2 + g^q),$$

and so

$$\sum\sum\sum_{i\neq j, i\neq k, j\neq k} E(\tilde{\zeta}_{ij}\tilde{\zeta}_{ik}) = O(n^{-1})O(h^4 + g^{2q}). \tag{9}$$

Also,

$$E(\tilde{\zeta}_{ij}\tilde{\zeta}_{ji}) = \frac{1}{n^4} E\Big(\frac{1}{p_i p_j} L_g^2(X_{i\underline{\alpha}} - X_{j\underline{\alpha}}) K_h(x_\alpha - X_{j\alpha}) K_h(x_\alpha - X_{i\alpha})$$
$$\times \left\{ m(X_{i\alpha}, X_{i\underline{\alpha}}) - m(x_\alpha, X_{j\alpha}) \right\} \left\{ m(X_{j\alpha}, X_{j\underline{\alpha}}) - m(x_\alpha, X_{i\alpha}) \right\} \Big)$$
$$= O(\frac{1}{n^4 g^{d-1}}). \tag{10}$$

Then the total contribution to the variance of U_{0n} from these terms is $O(\frac{1}{n^2 g^{d-1}})$. By similar arguments we get

$$\sum\sum_{i\neq j} E(\tilde{\zeta}_{ij}^2) = O(\frac{1}{n^2 g^{d-1} h}). \tag{11}$$

Then by (8), (9), (10) and (11) and the assumptions on the bandwidths,

$$E_j \left[\{\eta_j - E_j[\eta_j]\}^2 \right] = o_p\left(n^{-2/5}\right).$$

This completes the proof of **II**. ∎

Proof of Theorem 2: To simplify the notation we always write the α^{th} component of the density first and the β^{th} component second. In order to prove the theorem we show that the asymptotic covariance between $n^{2/5}\widehat{f}_\alpha(x_\alpha)$ and $n^{2/5}\widehat{f}_\beta(x_\beta)$ is $o(1)$. By **II** in the proof of Theorem 1 we need to show that

$$E\left[\left\{ \sum_{j=1}^n w_{j\alpha}\epsilon_j \right\} \left\{ \sum_{j=1}^n w_{j\beta}\epsilon_j \right\} \right] = nE\left[w_{1\alpha} w_{1\beta} \epsilon_1^2 \right] = o(n^{-4/5}),$$

Statistical Theory and Computational Aspects of Smoothing. Härdle, W. and Schimek, M. (eds.)
Physika Verlag, Heidelberg. ISBN 3-7908-0930-6

since $E[\epsilon_i \epsilon_j] = 0$ for $i \neq j$ and $w_{j\alpha} w_{j\beta} \epsilon_j^2$ are i.i.d.

$$\begin{aligned} E\left[w_{1\alpha} w_{1\beta} \epsilon_1^2\right] &= \tfrac{1}{n^2} \int \sigma^2(z_\alpha, z_\beta, w) K_h(x_\alpha - z_\alpha) K_h(x_\beta - z_\beta) \\ &\quad \times \tfrac{p_\alpha(z_\beta, w) p_\beta(z_\alpha, w)}{p(x_\alpha, z_\beta, w) p(z_\alpha, x_\beta, w)} p(z_\alpha, z_\beta, w) dz_\alpha dz_\beta dw \\ &= \tfrac{1}{n^2} \int \sigma^2(x_\alpha + hu, x_\beta + hv, w) K_h(u) K_h(v) \\ &\quad \times \tfrac{p_\alpha(x_\beta + hv, w) p_\beta(x_\alpha + hu, w)}{p(x_\alpha, x_\beta + hv, w) p(x_\alpha + hu, x_\beta, w)} p(x_\alpha + hu, x_\beta + hv, w) du dv dw \\ &= O\left(n^{-2}\right), \end{aligned}$$

by a change of variables argument and assumptions (A1), (A2), (A3), and (A5). This establishes the negligible asymptotic covariance of $n^{2/5} \widehat{f}_\alpha(x_\alpha)$ and $n^{2/5} \widehat{f}_\beta(x_\beta)$, thus proving the theorem.

Proof of Theorem 3 : We break $\widehat{S}_\alpha$ into the following terms,

$$\begin{aligned} \widehat{S}_\alpha &= \tfrac{1}{n} \textstyle\sum_{i=1}^n \left\{\widehat{f}_\alpha(X_{i\alpha})\right\}^2 \\ &= \tfrac{1}{n} \textstyle\sum_{i=1}^n \left\{f_\alpha(X_{i\alpha}) + c + \widehat{f}_\alpha(X_{i\alpha}) - f_\alpha(X_{i\alpha}) - c\right\}^2 \\ &= \tfrac{1}{n} \textstyle\sum_{i=1}^n \left\{[f_\alpha(X_{i\alpha}) + c]^2 + 2[\widehat{f}_\alpha(X_{i\alpha}) - f_\alpha(X_{i\alpha}) - c] f_\alpha(X_{i\alpha}) + [\widehat{f}_\alpha(X_{i\alpha}) - f_\alpha(X_{i\alpha}) - c]^2\right\} \\ &\equiv U_{1n} + U_{2n} + U_{3n}. \end{aligned}$$

Since the X_i's are i.i.d.,

$$U_{1n} = S_\alpha + O_p(n^{-1/2}).$$

By Theorem 1, U_{2n} can be approximated by

$$\begin{aligned} U_{2n} &\approx \tfrac{2}{n} \textstyle\sum_{i=1}^n \left\{b_\alpha(X_{i\alpha}) + \sum_{j=1}^n w_{j\alpha}(X_{i\alpha}) \epsilon_j\right\} f_\alpha(X_{i\alpha}) \\ &= \tfrac{2}{n} \textstyle\sum_{i=1}^n b_\alpha(X_{i\alpha}) f_\alpha(X_{i\alpha}) + \tfrac{2}{n} \textstyle\sum_{i=1}^n \sum_{j=1}^n w_{j\alpha}(X_{i\alpha}) \epsilon_j f_\alpha(X_{i\alpha}) \qquad (12) \\ &= O_p\left(h^2\right) + O_p\left(n^{-1/2}\right) = O_p\left(n^{-1/2}\right), \end{aligned}$$

Statistical Theory and Computational Aspects of Smoothing. Härdle, W. and Schimek, M. (eds.)
Physika Verlag, Heidelberg. ISBN 3-7908-0930-6

by the bandwidth conditions. From the proof of Theorem 1 and an application of the Cauchy-Schwarz inequality, $U_{3n} = O_p(n^{-3/4})$. Hence,

$$\widehat{S}_\alpha = S_\alpha + O_p(n^{-1/2}).$$

■

Proof of Theorem 4: If $S_\alpha = 0$, then note that $E\left[\widehat{S}_\alpha\right]^2 = O\left(\frac{1}{n^2h^2}\right) = O\left(n^{-6/4}\right)$ and $E\left[\widehat{S}_\alpha^2\right] = O\left(n^{-6/4}\right)$. Therefore, there exists n such that $b > E\left[\widehat{S}_\alpha\right]$. Then,

$$\begin{aligned}
\Pr\left[\widehat{S}_\alpha > b\right] &= \Pr\left[\widehat{S}_\alpha - E\left[\widehat{S}_\alpha\right] > b - E\left[\widehat{S}_\alpha\right]\right] \\
&\leq \Pr\left[\left|\widehat{S}_\alpha - E\left[\widehat{S}_\alpha\right]\right| > b - E\left[\widehat{S}_\alpha\right]\right] \\
&\leq \frac{E\left[\widehat{S}_\alpha^2\right] - E\left[\widehat{S}_\alpha\right]^2}{\left(b - E\left[\widehat{S}_\alpha\right]\right)^2} \\
&= O\left(n^{-6/4}\right)\frac{1}{\left(b - E\left[\widehat{S}_\alpha\right]\right)^2} = O\left(n^{-1}\right).
\end{aligned} \tag{13}$$

If $S_\alpha > 0$ then there exists an n such that $b < E\left[\widehat{S}_\alpha\right]$. Then,

$$\begin{aligned}
\Pr\left[\widehat{S}_\alpha < b\right] &= \Pr\left[\widehat{S}_\alpha - E\left[\widehat{S}_\alpha\right] < b - E\left[\widehat{S}_\alpha\right]\right] \\
&\leq \Pr\left[\left|\widehat{S}_\alpha - E\left[\widehat{S}_\alpha\right]\right| > E\left[\widehat{S}_\alpha\right] - b\right] \\
&\leq \frac{E\left[\widehat{S}_\alpha^2\right] - E\left[\widehat{S}_\alpha\right]^2}{\left(E\left[\widehat{S}_\alpha\right] - b\right)^2} \\
&= O\left(n^{-1}\right)\frac{1}{\left(E\left[\widehat{S}_\alpha\right] - b\right)^2} = O\left(n^{-1}\right).
\end{aligned} \tag{14}$$

The theorem follows from (13) and

$$\Pr\left[\widehat{A} \neq A\right] \leq d\left\{\Pr\left[\widehat{S}_\alpha < b \mid \alpha \in A\right] + \Pr\left[\widehat{S}_\alpha > b \mid \alpha \notin A\right]\right\}.$$

■

Statistical Theory and Computational Aspects of Smoothing. Härdle, W. and Schimek, M. (eds.)
Physika Verlag, Heidelberg. ISBN 3-7908-0930-6

Acknowledgement: We would like to thank Enno Mammen for pointing out an error in an earlier version of this paper, and Ray Carroll for fruitful discussions. Special thanks goes to Stephen Sperlich for computation help.

References

[1] BUJA, A, HASTIE, T.J. AND TIBSHIRANI, R.J. (1989), Linear smoothers and additive models, *Ann. Statist.*, 17, 453-510.

[2] HÄRDLE, W. (1990). *Applied Nonparametric Regression.* Econometric Society Monograph 19, Cambridge University Press.

[3] HÄRDLE, W. AND HALL, P. (1993) On the backfitting algorithm for additive regression models. *Statistica Neerlandica*, 47, 43-57.

[4] HÄRDLE, W. KLINKE, S. AND TURLACH, B.A. (1995) XploRe - an interactive statistical computing environment. Springer Verlag, New York .

[5] HÄRDLE, W AND MARRON, J.S. (1985), Optimal bandwidth selection in nonparametric regression function estimation. *Ann. Statist*, 13, 1465-1481

[6] HÄRDLE, W. AND TSYBAKOV, A.B. (1995) Additive nonparametric regression on principal components, *J. Nonparametric Statistics*, in press

[7] HASTIE, T. J. AND R. J. TIBSHIRANI (1990), Generalized additive models, Chapman and Hall: London

[8] LINTON, O. AND J. P. NIELSEN (1995) A kernel method of estimating structured nonparametric regression based on marginal integration, *Biometrika* **82**, 93-101.

[9] LEONTIEF, W. (1947) Introduction to a theory of an internal structure of functional relationships. *Econometrika*, 15, 361-373.

[10] MALJUTOV, M.B. AND WYNN, H.P. (1994) Sequential screening of significant variables of an additive model. in : Markov processes and applicazions. Festschrift for Dynkin, Birkhäuser Progress.

[11] MASRY, E. AND TJØSTHEIM, D. (1993), Nonparametric estimation and identification of ARCH and ARX nonlinear time series. Convergence properties and rates. *working paper, University of Bergen, Dept. of Math.*

[12] NADARAYA, E.A. (1964), On estimating regression. *Theor. Probab. Appl.*, 9, 141-142.

Statistical Theory and Computational Aspects of Smoothing. Härdle, W. and Schimek, M. (eds.)
Physika Verlag, Heidelberg. ISBN 3-7908-0930-6

[13] PARZEN, E. (1962), On estimation of a probability density and mode. *Ann. Math. Stat.*, 35,1065-76.

[14] SEVERANCE-LOSSIN, E. (1994), Nonparametric Testing of Production assertions in Data with Random Shocks. , PhD Thesis, University of Wisconsin-Madison .

[15] SILVERMAN, B.W. (1986), Density Estimation for Statistics and Data Analysis, Chapman and Hall, London.

[16] STONE, C.J. (1985), Additive regression and other nonparametric models, *Ann. Statist.* 13, 689 - 705

[17] STONE, C.J. (1986), The dimensionality reduction principle for generalized additive models, *Ann. Statist.* 14, 590- 606

[18] TJØSTHEIM, D. AND AUESTAD, B.H. (1994), Nonparametric identification of nonlinear time series: projections, *J. American Statistical Association,* 89, 1398-1409

[19] VENABLES, W.N. AND RIPLEY, B. (1994) Modern applied statistics with S-Plus. Springer Verlag, New York.

[20] WATSON, G.S. (1964), Smooth regression analysis. *Sankhyā A*, 26, 359-372

Finanzmarktanalyse und -prognose, 329-333

Zinsprognose mit univariater nichtparametrischer Zeitreihenanalyse

Wolfgang Härdle und Christian Hafner

16. März 1995

Die Prognose von Zinsniveaus und Finanzzeitreihen allgemein erweist sich als schwierig, da man nach Differenzenbildung meist keine signifikanten Autokorrelationen und damit nicht mehr viel Struktur *im Mittelwert* der Zeitreihe hat. Der Random Walk ist häufig über kurze Zeiträume eine hinreichend gute Approximation.
Aus diesem Grund sind die klassischen ARMA-Modelle für die Prognose dieser Daten wenig geeignet. Es zeigt sich, daß auch nichtparametrische autoregressive Modelle, von denen hier eines angewandt wird, keine große Verbesserung gegenüber einer naiven Prognose liefern.
Wie in einem anderen Beitrag dieser Autoren beschrieben, ist es häufig für die Praxis ebenso wichtig, die *Volatilität* und damit das Risiko zu prognostizieren. Neben der direkten Anwendung der geschätzten Volatilitäten auf die Bewertung von Derivativen lassen sich die zeitabhängigen Schwankungen für genauere Prognoseintervalle heranziehen. Die Varianzfunktion $s(\cdot)$ in dem Modell

$$y_t = f(y_{t-1}) + s(y_{t-1})\xi_t \tag{1}$$

kann wie in Härdle, Tsybakov (1995) mit nichtparametrischen Methoden geschätzt werden. ξ ist hier eine unabhängig und identisch verteilte Zufallsvariable mit Mittelwert Null und Varianz Eins. Hat die Mittelwertfunktion $f(\cdot)$ keinen großen Einfluß, ist dieses Modell interpretierbar als Verallgemeinerung von klassischen ARCH–Modellen.
Für die Prognose der Zins*niveaus* hat eine geschätzte Varianzfunktion jedoch keinen Einfluß, da der bedingte Erwartungswert $f(\cdot)$ die beste Prognose im Sinne des mittleren quadratischen Prognosefehlers ist. Insofern haben wir uns auf die Schätzung und Prognose eines sehr einfachen Modells beschränkt.
Es geht um die Prognose der 10 Jahres DEM Zinsen r jeweils auf $\Delta t = 60$ Wochentage. Der Prognosezeitraum erstreckt sich vom 23.12.1993 bis zum 23.12.1994. Elf fehlende Werte im Prognosezeitraum wurden eliminiert. Somit bleiben für die unten erwähnten Gütekriterien $T = 262 - 11 = 251$ relevante Prognosen.
Die Zeitreihen der Zinsen und der Zinsdifferenzen, jeweils inklusive Prognosezeitraum, sind in Bild 1 und 2 geplottet.

(1996) Härdle, W. and Hafner, Ch. Zinsprognose mit univariater nichtparametrischer Zeitreihenanalyse.

h	MSD	IR1	MAD	IR2	Bias
0.04	0.5460576	0.994566	0.483374	0.996927	0.38875
0.045	0.54584636	0.9937965	0.4832481	0.996666	0.388266
0.05	0.5457698	0.993518	0.48312	0.996405	0.38815
0.055	0.54607813	0.994641	0.4833675	0.996912	0.388591
0.06	0.5459412	0.994142	0.483255	0.996681	0.388464
0.07	0.5462492	0.995264	0.483497	0.997181	0.388903
0.1	0.546776	0.9972	0.484013	0.99824	0.3898

Table 1: Gütekriterien für Modelle mit verschiedenen Bandweiten

MSD: mean standard deviation
IR1: corresponding information ratio
MAD: mean absolute deviation
IR2: corresponding information ratio

Als nichtparametrische Schätzmethode wurde hier wie in Bossaerts, Härdle und Hafner (1995) der Lokale Polynomschätzer angewendet. Der klassische Nadaraya–Watson Schätzer ist als Spezialfall mit dem Polynomgrad Null enthalten. Die Methode ist interpretierbar als gleitende gewichtete Durchschnittsbildung mit Polynomen, wobei als Gewichte meistens Kernfunktionen genommen werden. Die Glattheit der geschätzten Funktion wird durch die Bandweite h bestimmt. Für kleinere Bandweiten erhöht sich die Varianz des Schätzers und die Funktion wird rauher, umgekehrt erhöht sich der Bias für größere Bandweiten und zunehmend glatterer Schätzfunktion.
Das Modell

$$y_t = f(y_{t-60}) + \varepsilon_t \tag{2}$$

wurde für verschiedene Bandweiten h mit dem Local Linear Estimator (LLE) geschätzt, wobei $y_t \equiv \Delta r_t$. Als Gewichtsfunktion wurde der Quartic Kern verwendet.
In Bild 3 und 4 sind die Datenpaare (y_t, y_{t-60}) und die geschätzte Funktion für $h = 0.05$ dargestellt.

Gütekriterien der Prognose für verschiedene Bandweiten h sind in Tabelle 1 gegeben.

Die Residuenvarianz ist deutlich kleiner als die Varianz der Zinsdifferenzen:

(1996) Härdle, W. and Hafner, Ch. Zinsprognose mit univariater nichtparametrischer Zeitreihenanalyse.

$$\frac{Var(\hat{\varepsilon}_t)}{Var(y_t)} = \frac{0.0018491633}{0.0019678} = 0.9396. \tag{3}$$

Allerdings scheint die Erklärungskraft des Modells für die Prognose keine großen Vorteile zu bringen, wie die Gütekriterien zeigen. Als Erweiterung dieses Ansatzes wird von Chen, Tsay (1993) ein nichtparametrisches additives Modell mit mehreren Lags als erklärende Variablen vorgeschlagen.

Literatur

Bossaerts, P.; Härdle, W.; Hafner, C. (1995) *Foreign Exchange-rates have surprising volatility.* SFB 373 Discussion Paper 45, erhältlich via FTP: amadeus.wiwi.hu-berlin.de unter pub/papers/sfb.

Chen, R. and Tsay, R. S. (1993) Nonlinear additive ARX models, Journal of the American Statistical Association 88: 955–967.

Härdle, W. and Tsybakov, A. (1995) Local polynomial estimators of the volatility function in nonparametric autoregression, SFB 373 Discussion Paper 42.

(1996) Härdle, W. and Hafner, Ch. Zinsprognose mit univariater nichtparametrischer Zeitreihenanalyse.

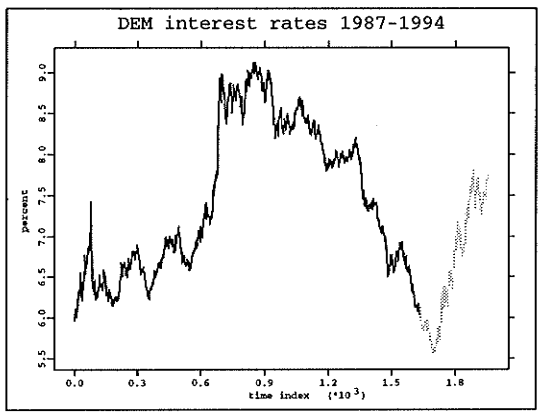


Figure 1: Zeitreihe der 10 Jahres DM-Zinsen über den gesamten Zeitraum 1.7.1987 bis 23.12.1994. Der Prognosebereich ist farblich abgehoben.

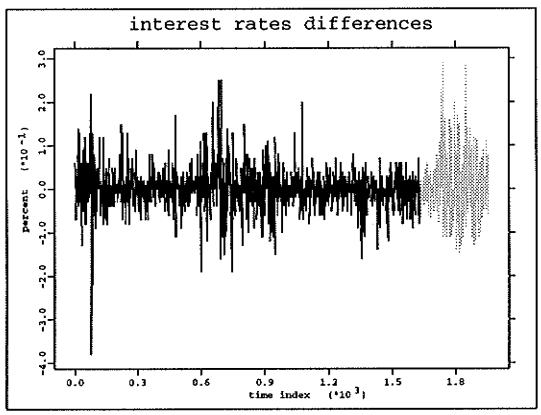


Figure 2: Zeitreihe der ersten Differenzen der 10 Jahres DM-Zinsen über den gesamten Zeitraum 1.7.1987 bis 23.12.1994. Der Prognosebereich ist farblich abgehoben.

(1996) Härdle, W. and Hafner, Ch. Zinsprognose mit univariater nichtparametrischer Zeitreihenanalyse.

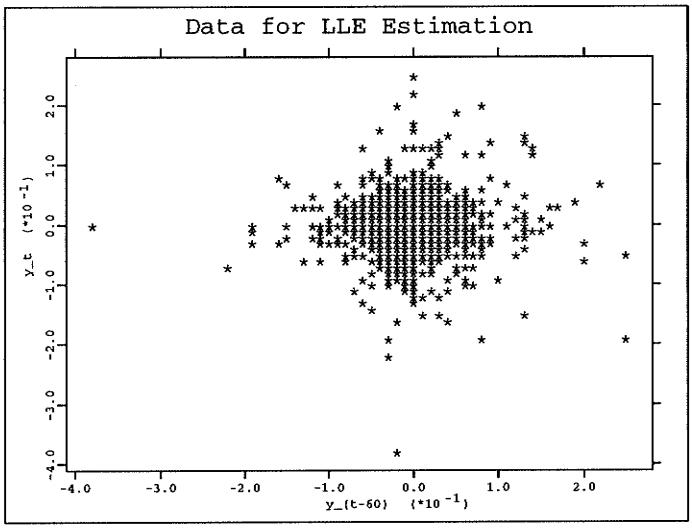


Figure 3: Datenpaare (y_t, y_{t-60}) für den Schätzzeitraum 1.7.1987 bis 30.9.1993, $n = 1631$.

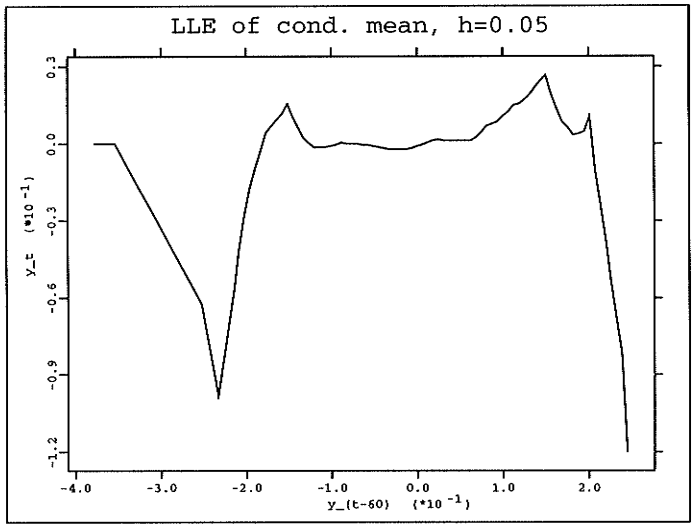


Figure 4: Geschätzte Funktion $\hat{f}_h(y_{t-60})$ für $h = 0.05$ über den Schätzzeitraum 1.7.1987 bis 30.9.1993.

(1996) Härdle, W. and Hafner, Ch. Zinsprognose mit univariater nichtparametrischer Zeitreihenanalyse.

COMPSTAT Proceedings Barcelona Physika Verlag, Heidelberg. ISBN 3-7908-0953-5

A New Generation of a Statistical Computing Environment on the Net

Swetlana Schmelzer, Thomas Kötter, Sigbert Klinke and Wolfgang Härdle

Humboldt-Universität zu Berlin, Wirtschaftswissenschaftliche Fakultät,
Institut für Statistik und Ökonometrie, Spandauer Straße 1,
D-10178 Berlin, Germany
E-Mail: swetlana@wiwi.hu-berlin.de, thomas@wiwi.hu-berlin.de,
sigbert@wiwi.hu-berlin.de, haerdle@wiwi.hu-berlin.de

Abstract. With the availability of the net a new generation of computing environments has to be designed for a large scale of statistical tasks ranging from data analysis to highly interactive operations. It must combine the flexibility of multi window desktops with standard operations and interactive user driven actions. It must be equally well suited for first year students and for high demanding researchers. Its design must has various degrees of flexibility that allow to address different levels of user groups. We present here some ideas how a new generation of a computing environment can be used as a student front end tool for teaching elementary statistics as well as a research device for highly computer intensive tasks, e.g. for semiparametric analysis and bootstrapping.

1 The Net and an Interactive Statistical Computing Environment

First versions of interactive computing environments have been created in the mid eighties. At this time PCs were about to emerge as the standard machine for statisticians and classical systems were either on mainframes or

Schmelzer, S., Kötter, Th., Klinke, S. and Härdle, W. (1996) A New Generation of a Statistical Computing Environment on the Net

copies of these mainframe (batch oriented) programs on PCs. Among the first computing environments on interactive machines were S and ISP, see Becker, Chambers & Wilks (1988), S (1988), ISP (1987) and Härdle (1988). With the universality of these interactive statistical systems on their mainly UNIX machine platforms it was possible to combine research computing (usually dependent on highly parametrized subroutines) with graphically intensive oriented data analysis. The last author used for example in teaching elementary statistical concepts to first year students in a computer classroom (Bowman & Robinson (1989, 1990)).

In Germany the first computer classrooms for statistics and data analysis emerged in the second half of the eighties. The statistics department of Dortmund was the first in our country to use ISP on Apollo machines for teaching graduate courses. Later in the eighties interactive statistical software became available on PCs, the speed though made it more a teaching tool than a research environment. One element of slowness was the hardware design: programs had to switch between graphics and text screens, a factor of speed unknown to Mac based systems, like DataDesk (Velleman 1992), of course. Another disadvantage of PC based statistical packages at that time was the inability to handle and to link windows with different statistical information. On workstations that was less a problem but before Microsoft Windows 3.1 appeared as the standard interface on PCs there was no chance of treating and analysing data simultaneously in parallel viewports or windows.

This was one of the primary motivations to create our own statistical system XploRe. Other motivations were a unified graphics interface, a simple PC based platform and multiple windows. The first version (1986-1988) fulfilled these requirements but did not completely satisfy since the memory was restricted to 640 KB. There was no color available but a variety of smoothing tools for high and low dimensions was implemented in a menu tree structure. Three dimensional dynamic graphics, scatterplot matrices, linking and brushing were available. As a speed enhancing device we used WARPing, see Härdle & Scott (1992), and the FFT.

The next step of development (1988 - 1990) brought color, a menu frame and elementary data manipulations, like transformations on variables. The speed of PCs grew, hardware changed from 286 to 386 chips. We used XploRe 2 in teaching smoothing methods, a branch of statistics where interactive graphics are a conditio sine qua non. The menu was easier to handle but still it was a menu and no user written program. We introduced a toolbox for semiparametric and additive modeling into the system with interactive choice of smoothing parameter.

The Janus head like use of XploRe (for teaching and research) let us think about changing the design. One line of thought was that a language must drive the system. Another design principle was that necessary parallel available information (e.g. the regression line and the residuals) should not be

scattered around in partially overlapping and unlinked windows. On the second thought we realized that a language fulfilling these design principles must be too complex. Students would not like to use it in class. Therefore the XploRe 3 language included menu construction, display mixtures, and the context sensitive "open key". In the years 1990 to 1996 XploRe 3 left the "space of statistical systems" and emerged to an *environment.*

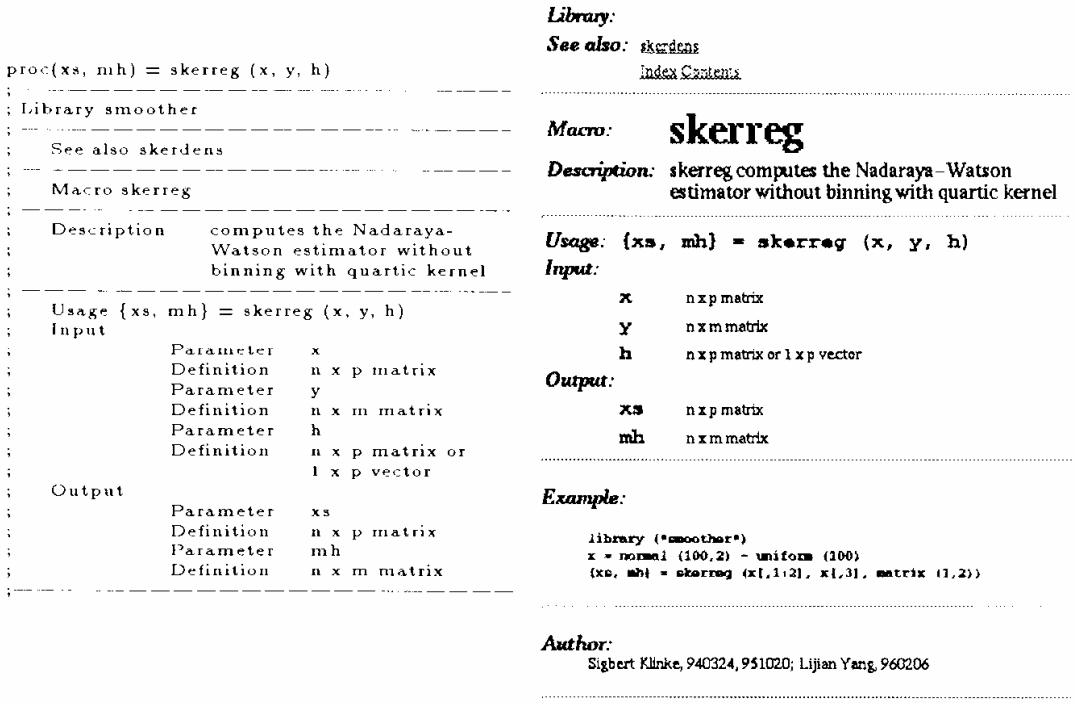
An environment is a computing device that covers a wide range of data manipulations, problem solutions and graphical insights not only over a set of statistical operations (horizontal coverage) but also over a set of user levels (vertical coverage) from first year students to graduates up to researchers. XploRe 3 serves for teaching, too (Proenca 1995). Users can define their customized interface with their own libraries of user written macros and less experienced people can browse through the syntax of operations. This "open key" opens a help page when it is pressed with the cursor on the command name. On user written macros it opens the macro to provide information on the use. User written macros could be started from the built in editor and after execution of the macro the statistician fall back into the editor in order to allow him modifications of the code. This natural idea of interaction with user produced code was forgotten for some time in many statistical systems during this period although this feature was highly praised about 1984 in the TURBO Pascal program development system. The full description of XploRe 3 is in the book by Härdle, Klinke & Turlach (1995).

One feature that is realized now is the partial run of code. In an interactive statistical environment one starts with simple problems (corresponding to a few lines of code), adds more complicated questions and packs after a certain degree of complexity little independent pieces of operations into macros. We may run segments of the log file by copying them into the buffer and pasting the content of the buffer onto the console line. The wide use of HyperText Markup Language (HTML) files makes it reasonable to offer help files as translations of headers of user written macros. From help files segments of code may be pasted onto the console and thus executed. A speed factor is the translator of XploRe 4. XploRe 4 user macros may be called from XploRe commands. A typical example is the computation of the FAST estimator Chen, Härdle, Linton & Severance-Lossin (1996) with a user defined kernel (depending on the dimension of the problem).

A natural action for a beginner is to ask for help. We therefore discuss the help system first in section 2 before we present the internal XploRe data structures are discussed in section 4 and the interactive graphical devices in section 5.

2 The Help System

"Software must be self explaining", this paradigm became more and more important for graphical user interfaces (GUI). The optimism of user control through GUIs was soon followed by the discernment of a trade-off between the variety of possible applications and the limited screen space. Specialized applications may be interesting only for a certain class of specialists. The problem is to make the information on a special method available without making it always visible on a GUI. A well structured help system is asked for here.



```
proc(xs, mh) = skerreg (x, y, h)
;-------------------------------------------------
; Library smoother
;-------------------------------------------------
;   See also skerdens
;-------------------------------------------------
;   Macro skerreg
;-------------------------------------------------
;   Description     computes the Nadaraya-
;                   Watson estimator without
;                   binning with quartic kernel
;-------------------------------------------------
;   Usage {xs, mh} = skerreg (x, y, h)
;   Input
;               Parameter   x
;               Definition  n x p matrix
;               Parameter   y
;               Definition  n x m matrix
;               Parameter   h
;               Definition  n x p matrix or
;                           1 x p vector
;   Output
;               Parameter   xs
;               Definition  n x p matrix
;               Parameter   mh
;               Definition  n x m matrix
;-------------------------------------------------
```

Library:
See also: skerdens
Index Contents

Macro: **skerreg**

Description: skerreg computes the Nadaraya–Watson estimator without binning with quartic kernel

Usage: `{xs, mh} = skerreg (x, y, h)`

Input:

x	n x p matrix
y	n x m matrix
h	n x p matrix or 1 x p vector

Output:

xs	n x p matrix
mh	n x m matrix

Example:

```
library ("smoother")
x = normal (100,2) ~ uniform (100)
{xs, mh} = skerreg (x[,1:2], x[,3], matrix (1,2))
```

Author:
Sigbert Klinke, 940324, 951020; Lijian Yang, 960206

Fig. 1. Excerpts from the source of the `skerreg` macro and the resulting HTML page.

Many help systems show usage information texts to certain keywords. An internal connection of these texts gives a reasonable user support since with related keywords the user may browse through the domain of applications. This concept is realized in HTML which can now be seen as the de-facto standard in distributed documents on the net. In the last two years this domain has found a big acceptance because of the easy use and appealing features of WWW. Now HTML is almost standardized and a variety of browsers and HTML supporters are available. The wide distribution, the high transparency,

and the free choice of HTML browsers like Mosaic, Netscape, Lynx, etc., have been the reasons for their use as the software backbone of the help system of our statistical software XploRe. The network availability of methods and documents in this format makes it a very convenient tool to learn more about the XploRe environment. The location of the document is not important since local copies may be used in the same way as distant information documents.

Furthermore the HTML help system is entirely separated from the software, what means that everybody can obtain first impressions of the statistical computing environment without installing it locally. Besides this we have also an internal help, which can be accessed quickly, e.g., checking the order of parameters of a certain function. In order to provide several instances of the help system, we decided to define a meta code and translators, which generate the different kinds of help pages. In this way we obtain the help system's documents from a single source for HTML, for the short description help inside XploRe, and for the printed manual pages.

There are two different levels of information assistance implemented. First there are the coarse documents giving information about groups of detailed help texts. Then the detailed help texts are constructed in a way that the user may copy parts of these into the console processor and thus verify what the help document states.

Additionally XploRe provides a translator that extracts the standard comments from macros. If the users follow the documentation scheme for macros they can generate their own HTML documents out of their macros. This feature is an important element of XploRe and corresponds to its design as environment. The user customizes his interface to computational statistics by writing own macros and they become tools but *also* documents in the help system available to everybody. Figure 1 shows the help document for the XploRe macro **`skerreg`**, a kernel smoothing block routine of the smoother library.

The design of the help documents has been developed together with psychologists working on software ergonomics, see Hüttner, Wandke & Rätz (1995). One principle in user/computer communication that we learned from our psychological colleagues was that the maximum of response to a user question should be on one screen. Scrolling and mouse movements tire the user and make the use of an interactive software painful and complicated. We have therefore decided to install hyper links not only in the header but also in the ending of the help documents.

3 The WWW Interface

The help system is a first step in experimenting with a new system. The net working facilities are a further stage. For unexperienced users or for demon-

stration or teaching purposes we developed a WWW-interface. Of course, by using a WWW-browser not all environmental features can be supported in the same way as before. Due to the network's bandwidth the interaction has to be limited and dynamic graphs are still hard to realize over the net.

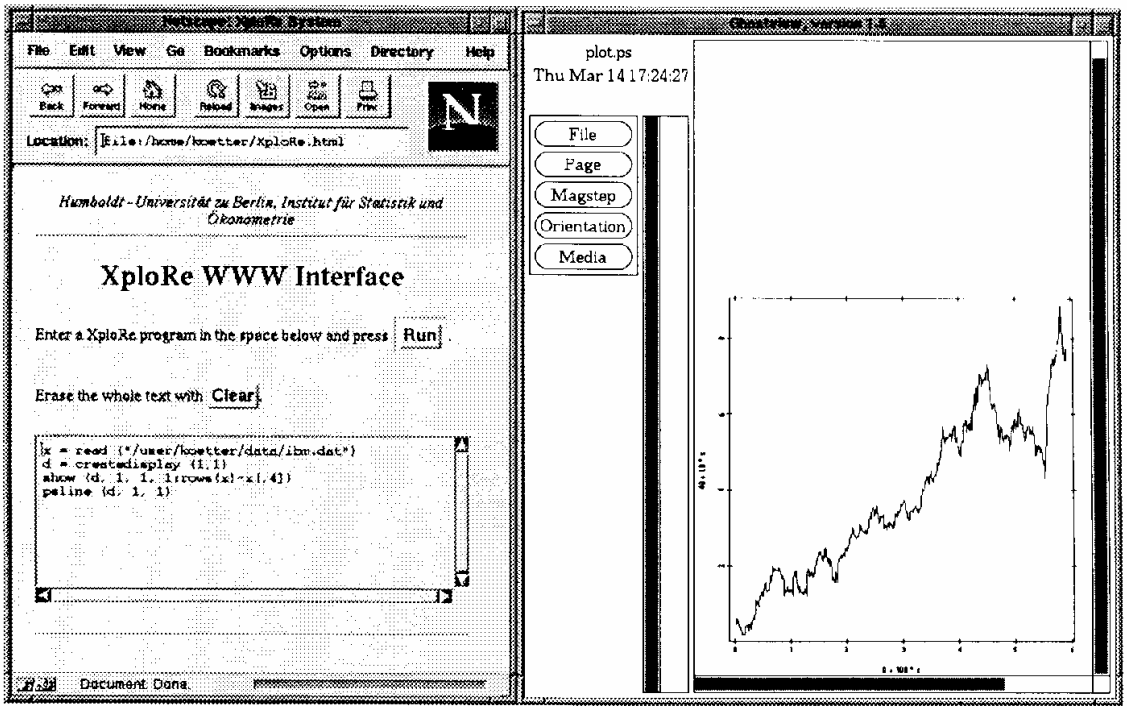


Fig. 2. The XploRe page. The left window shows the WWW-page with the just executed code. On the right side, there is ghostview with the graphical output. The graph shows the value of the IBM stocks from 8/30/93 to 3/14/96.

The state of such a server is another problem. As WWW relies on separated documents and requires no login/logout procedure, each connection is closed after the transfer of the document. I.e., that it is a difficult task to trace different documents belonging to a certain session and to distinguish between several sessions. So far we have only implemented a stateless server, which does not store any information from previous executions, so that each new request for running XploRe results in a new session. Graphs are produced in the PostScript format. This format offers a lot of advantages: first, all XploRe graphs are vector graphs; second, this format is widely spread, so that it can easily be printed or included into documents and with ghostview/ghostscript there exists for most computer systems a PostScript-viewer; third, it consists of plain ASCII-code, which can be exchanged safely between different architectures (e.g., big and little endian), and finally it has been already implemented as export format. Figure 2 shows the WWW-page of XploRe and

a just produced graph.

4 Arrays

4.1 Why arrays instead of matrices ?

Basic elements of statistical data are numbers and strings. In practical statistical work the representation of numbers in vectors (variables) and matrices (variables in columns, observations in rows) is useful. Arrays are collection of matrices.

Arrays are not new elements in statistical programming languages. The APL language worked with array-wise calculations, for example, and proved to be useful in statistical computing tasks (Büning 1983). Later ISP and the statistical software S-Plus provided arrays.

For the implementation of interactive smoothers multidimensional arrays are a desired data structure. They may be used in Exploratory Projection Pursuit (EPP) (Klinke 1995, Jones & Sibson 1987) or in local polynomial regression (Katkovnik (1979, 1983, 1985), Fan & Gijbels (1996)).

Given data $(X_i, Y_i), i = 1, \ldots, n$ the regression of Y on X can be estimated by the local polynomial (LP) estimation technique:

$$\begin{aligned} \hat{b}(x_l) &= (X^T W X)^{-1}\ X^T W Y \\ &= \quad S_n^{-1}(x_l) \quad\ T_n(x_l) \end{aligned}$$

with

$$X = \begin{pmatrix} 1 & (X_1 - x_l) & \ldots & (X_1 - x_l)^p \\ \vdots & \vdots & & \vdots \\ 1 & (X_n - x_l) & \ldots & (X_n - x_l)^p \end{pmatrix}$$

$$Y^T = (Y_1, \cdots, Y_n), \quad W = diag\left\{K\left(\frac{X_i - x_l}{h_n}\right)\right\}$$

A standard task in this regression context is the visualization or crossvalidation of smooth curve estimates. The computation at k datapoints $x_l\,(l = 1, \ldots, k)$, involves the inversion of k matrices $S_n(x_l)$. In order to avoid looping we store all matrices $S_n(x_l)$ in *one* three-dimensional array. By "inverting" it we solve the system for all datapoints simultaneously. The inversion has to follow certain basic operation rules that we describe next.

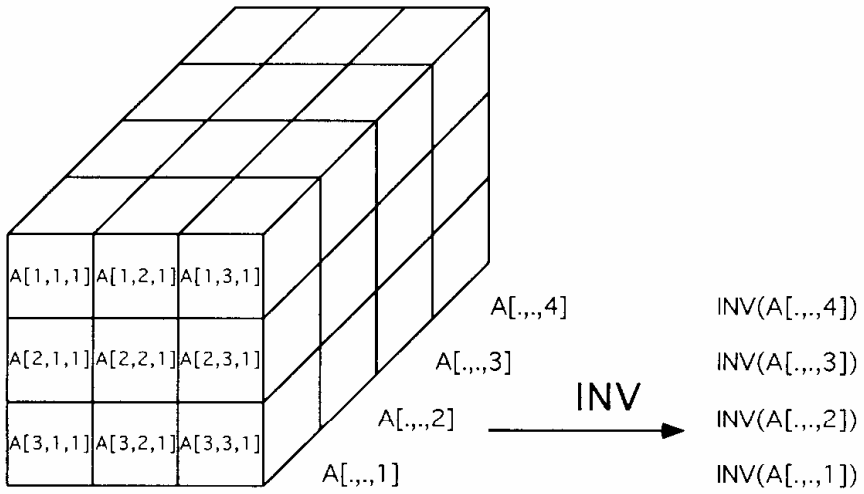


Fig. 3. How the inverse matrix operation will work on a 3-dimensional array.

4.2 Basic Operations

Conformability of elementary operations. Elementary operations are the elementwise mathematical operations (addition, subtractions, multiplication, and division) and the elementwise logical operations (logical or, logical and, less, greater, ...). A typical standard operation is the centering of a data matrix by $Y = X - mean(X)$. On the left side of the minus we have an $n \times p$ matrix, on the right side a $1 \times p$ vector (matrix). Thus for an elementwise operation these matrices are not conformable in strict sense. Nevertheless this operation is necessary.

To achieve conformability for an array we allow elementwise operations only if the size of the operands is the same or equal to 1 in each dimension. We define the resulting size in each dimension as the maximum of the sizes of each operand, see Klinke (1995).

Vector operations. Vector operations on arrays are operations on one variable. Typical operations are the mean, the median, the variance and the sum over observations. The result is an array which has in the working dimension the size 1. It is also of interest to look at conditional means, conditional medians, conditional variances, and conditional sums. The resulting size in the working dimension is k if we have k classes we condition on.

Layer operations. Operations which are specific for matrices are the multiplication, the inversion, the transposition, the calculation of moments etc.

According to Figure 3 these operations are extended by applying them to each layer.

Multiple extensions. An example for multiple extensions is the sorting of vectors. In XploRe 3 we have implemented a sort command that sorts the whole matrix accordingly to a set of parallel sorting vectors. The sorting is extended by the `sort` command in same way as the "inversion" operation and applied to each layer.

Another extension (`sort2`) is given by defining the sort direction and a sequence of parallel vectors to sort after. In Figure 3 the sorting direction can be 3 (the depth) and the sorting vectors may be $A[1,3,]$ and $A[3,1,]$. We do not interpret an array as some repeated or parallel objects, e.g. as in LP regression, but as *one* object.

4.3 Implementation of C++-Classes

The basic object in `XploRe` 4 is an 8-dimensional array. The programming of the arrays in C++ allows us to build up a hierarchy of classes and to reduce the amount of programming.

```
XStringDatabase                database for strings
      xstring                  handling of strings
xplmask                        handling of colors and forms
XplArray                       base class for arrays
      XplNumber                base class for number arrays
            XplInteger         base class for integer arrays
            XplReal            base class for floating point number
                  XplDouble    class for double numbers
      XplChar                  class for texts
      XplMask                  class colors and forms
```

The base class `XplArray` contains basic operation which are common to all arrays. Step by step we speciale the classes. `XplNumber` contains all basic operations which can be done on numbers and so on.

Another class of arrays is designed for string handling. A problem that frequently occurs is that we have nominal variables which are represented as text. For a large dataset which has nominal variables (e.g. yes/no answers) we use references to the text. The string is only stored once in a skip list of the type `XStringDatabase` (Schneider 1994). The class `xstring` offers the standard string operations and the class `XplChar` handles arrays of strings.

A similar design has to be chosen to handle the color, the form and the

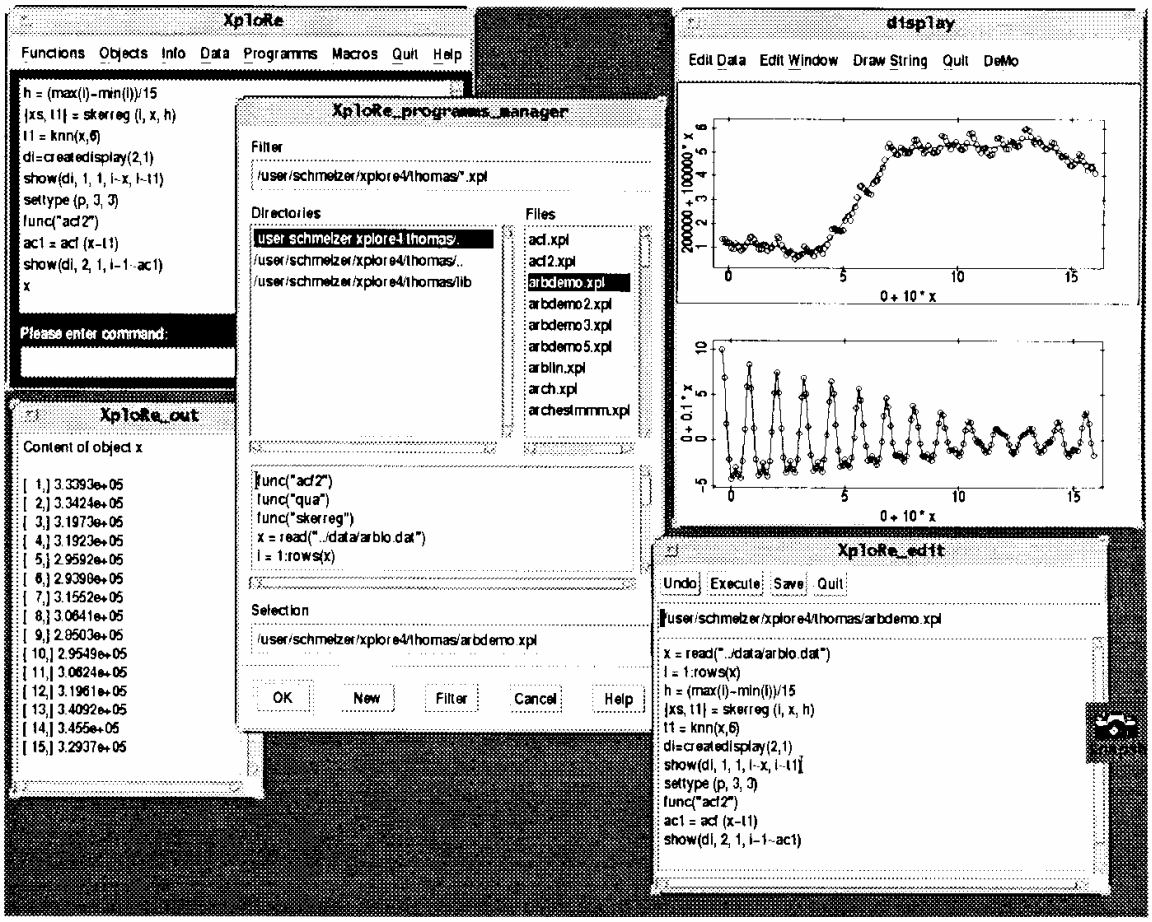


Fig. 4. Snapshot of a XploRe session under UNIX

size of a datapoint or a line. For this purpose we implemented the datatype `xplmask` and arrays of it in `XplMask`.

5 Interactive Graphics and Displays

5.1 Basic Windows

Any statistical interactive environment needs a set of windows to display information. Figure 4 shows a screen shot of a possible XploRe session. We discuss these windows in the following paragraphs in detail.

The console window. The main interaction with the software is done in a console window (the left upper window in Figure 4). It is the window for controlling data and programs. It consists of a set of smaller segments: a linewise input segment, a ten to fifteen lines history segment and the menu bar. All commands in a session can be recalled by scrolling the history window

and they are recorded in a log file. By double clicking a command in the history window it is put into the input segment and automatically executed.

The menu bar offers also short cuts of commands. The user may open data or program files, obtain links to the help system, and receive information about the system load, the objects (data arrays, user defined displays), functions and user loaded macros.

The output window. The textual output is directed to a separate output window (lower left window in Figure 4). The command structure is thus clearly separated from produced numerical and graphical results.

The edit window. The lower right window is an editor and shows the program which was executed shortly before (compare with the history segment in the console window). It is used for editing programs and data. For rapid program development we have included the `Execute` button which shortens turn-around times. Similar approaches can be found in other statistical programs, e.g. GAUSS under DOS, too.

The file browser. In Figure 4 we have activated the file browser (central window). It provides the usual facilities to browse through a file system. In addition we implemented a browser window (above `Selection` box) which shows the optional segments of the selected file. This facility eases much the search for particular files.

5.2 The Display Concept

In a study of user interfaces Hüttner et al. (1995) recommends to minimize mouse movements to icons, buttons etc. Although these mouse movement based commands are less precise than keyboard typed ones, many statistical software relies exclusively on them. As a consequence we rapidly have a full desktop of graphics information with overlaid and hidden windows. The last author calls this situation often an *overmoused* environment. We need too many mouse clicks to recover windows and to see what we have produced in our analysis. Connected plots in a logical and physical coherence is the proposal to group windows and to keep similar things together.

Such a set of non-overlaping windows is called a *display*. In a display different data viewers may be mixed: two dimensional Scatterplots may be grouped together with three dimensional ones, texts or boxplots etc. An example is the graphic window in upper right corner of Figure 4. It shows two plot windows and the upper one with the frame around is the active one. "Active" means here that all operations, e.g. through the menu, will effect just this window. Only if the window or a part of it is linked we may have an effect on other windows.

Standard operations on windows. Stuetzle (1987) recommends standard operations on windows like moving, resizing, iconifing or raising of windows by the user. Nowadays the underlying Graphical User Interface (GUI) will do all these tasks and we have to care about the contents of the window. This includes rescaling of plot windows.

The plot window. A plot window itself may contain several *dataparts*. In Figure 4 we can see in each plot window two objects: a connected line and datapoints plotted with circles. These two dataparts and the appearance of the x-axis, y-axis and the headline have been manipulated by the menu. Another feature is provided by the menu item `Draw String`; it allows us to place a user defined string everywhere in the graph to include, e.g. additional information about the data. The colors, linetypes, etc. of dataparts can be interactively manipulated.

Interrogating datapoints. We can "ask" each datapoint for its values by clicking with the right mouse button close to it. We see then the number of the datapart, the coordinates of that point and if desired coordinates of linked dataparts.

Zooming. Any region of a plot may be zoomed. By selecting a rectangular region all points inside are rescaled to the entire window.

Brushing and linking. A display is a container for visualizing several dataparts together. We may define relations between dataparts and other (imaginary) "members" of the plot. We may choose a set of points which describe a non-rectangular region ("lasso") and the plot uses the datapoints in this region for further computation (e.g. is the correlation coefficient influenced by some observations) or for linking.

5.3 Manipulating a Display by Commands

Figure 5 shows a printout of some variables of the Swiss bank notes data and the corresponding XploRe code.

The first 3 lines load the Swiss Banknote dataset which consists of six measurements (variables 4 to 9) on 100 genuine and 100 forged old swiss 1000-franc bills (Flury & Riedwyl 1981). Then we split the dataset into forged and genuine banknotes.

In line 4 we create the display `di` with 2×2 windows. It has no real windows, just a template without any type. The `show` command puts a window into a display, e.g. in line 5 we show the seventh and nineth variable of both types of banknotes.

Standard operations on windows. Stuetzle (1987) recommends standard operations on windows like moving, resizing, iconifing or raising of windows by the user. Nowadays the underlying Graphical User Interface (GUI) will do all these tasks and we have to care about the contents of the window. This includes rescaling of plot windows.

The plot window. A plot window itself may contain several *dataparts*. In Figure 4 we can see in each plot window two objects: a connected line and datapoints plotted with circles. These two dataparts and the appearance of the x-axis, y-axis and the headline have been manipulated by the menu. Another feature is provided by the menu item `Draw String`; it allows us to place a user defined string everywhere in the graph to include, e.g. additional information about the data. The colors, linetypes, etc. of dataparts can be interactively manipulated.

Interrogating datapoints. We can "ask" each datapoint for its values by clicking with the right mouse button close to it. We see then the number of the datapart, the coordinates of that point and if desired coordinates of linked dataparts.

Zooming. Any region of a plot may be zoomed. By selecting a rectangular region all points inside are rescaled to the entire window.

Brushing and linking. A display is a container for visualizing several dataparts together. We may define relations between dataparts and other (imaginary) "members" of the plot. We may choose a set of points which describe a non-rectangular region ("lasso") and the plot uses the datapoints in this region for further computation (e.g. is the correlation coefficient influenced by some observations) or for linking.

5.3 Manipulating a Display by Commands

Figure 5 shows a printout of some variables of the Swiss bank notes data and the corresponding XploRe code.

The first 3 lines load the Swiss Banknote dataset which consists of six measurements (variables 4 to 9) on 100 genuine and 100 forged old swiss 1000-franc bills (Flury & Riedwyl 1981). Then we split the dataset into forged and genuine banknotes.

In line 4 we create the display `di` with 2×2 windows. It has no real windows, just a template without any type. The `show` command puts a window into a display, e.g. in line 5 we show the seventh and nineth variable of both types of banknotes.

Bowman, A. & Robinson, D. (1990). C.I.T.: Regression & Anova, *Software*, IOP Publishing Ltd.

Büning, H. (1983). Adaptive distribution-free test (german), *Statistische Hefte* pp. 47–67.

Chen, R., Härdle, W., Linton, O. & Severance-Lossin, E. (1996). Nonparametric estimation of additive separable regression models, *in* W. Härdle & M. Schimek (eds), *Statistical Theory and Computational Aspects of Smoothing*, Physika Verlag Heidelberg.

Fan, J. & Gijbels, I. (1996). *Local Polynomial Modeling and Its Application– Theory and Methodologies*, Chapman & Hall.

Flury, B. & Riedwyl, H. (1981). Graphical representation of multivariate data by means of asymmetrical faces, *Journal of the American Statistical Association* **76**(376): 757–765.

Härdle, W. (1988). XploRe - a Computing Environment for eXploratory Regression and density smoothing, *Statistical Software Newsletters* **14**: 113–119.

Härdle, W., Klinke, S. & Turlach, B. (1995). *XploRe - an Interactive Statistical Computing Environment*, Springer, Heidelberg.

Härdle, W. & Scott, D. (1992). Smoothing in low and high dimensions by weighted averaging using rounded points, *Computational Statistics* pp. 97–128.

Hüttner, J., Wandke, H. & Rätz, A. (1995). *Benutzerfreundliche Software*, Bernd-Michael Paschke Verlag Berlin 1995, Berlin.

ISP (1987). ISP is a program for PCs available from Artemis Systems Inc.

Jones, M. & Sibson, R. (1987). What is projection pursuit ?, *Journal of the Royal Statistical Society A* **150**: 1–36.

Katkovnik, V. (1985). *Nonparametric identification and data smoothing: local approximation approach in (in Russian)*, Nauka, Moscow.

Katkovnik, V. Y. (1979). Linear and nonlinear methods for nonparametric regression analysis (in russian), *Avtomatika i Telemehanika* pp. 35–46.

Katkovnik, V. Y. (1983). Convergence of the linear and nonlinear nonparametric kernel estimates (in russian), *Avtomatika i Telemehanika* pp. 108–20.

Klinke, S. (1995). *Data Structures in Computational Statistics*, PhD thesis, Institute of statistics, Catholic university of Louvain.

Proenca, I. (1995). Interactive graphics for teaching simple statistics, *XploRe - an interactive statistical computing environment*, Springer, pp. 113–140.

S (1988). See Becker, Chambers and Wilks, (1988).

Schneider, B. (1994). Querleser - Zerstreutheit mit System: Skip-Listen, *c't* **2**: 204 207.

Stuetzle, W. (1987). Plot windows, *Journal of the American Statistical Association* **82**(398): 466–475.

Velleman, P. (1992). *Data Desk*, Data Description, Ithaca NY.

Annals of Statistics (1993), 21, 157-178

OPTIMAL SMOOTHING IN SINGLE-INDEX MODELS

BY WOLFGANG HÄRDLE, PETER HALL AND HIDEHIKO ICHIMURA[1]

Université Catholique de Louvain, Australian National University and University of Minnesota

Single-index models generalize linear regression. They have applications to a variety of fields, such as discrete choice analysis in econometrics and dose response models in biometrics, where high-dimensional regression models are often employed. Single-index models are similar to the first step of projection pursuit regression, a dimension-reduction method. In both cases the orientation vector can be estimated root-n consistently, even if the unknown univariate function (or nonparametric link function) is assumed to come from a large smoothness class. However, as we show in the present paper, the similarities end there. In particular, the amount of smoothing necessary for root-n consistent orientation estimation is very different in the two cases. We suggest a simple, empirical rule for selecting the bandwidth appropriate to single-index models. This rule is studied in a small simulation study and an application in binary response models.

1. Introduction. A linear regression model for the dependence of a scalar variable Y and a p-vector x has the form $Y = \beta^T x + \varepsilon$, where β is a p-vector of unknown parameters and ε is a random variable with zero mean conditional on x. More generally, we might define $Y = g(\beta^T x) + \varepsilon$, where g is an unknown univariate function. This is a *single-index model*, and is recognized as a particularly useful variation of the linear regression formulation [e.g., Brillinger (1983) and McCullagh and Nelder (1983)]. Of course, the scale of $\beta^T x$ in $g(\beta^T x)$ may be determined arbitrarily, and so we may replace β by the unit vector $\theta = \beta \|\beta\|^{-1}$, where $\|\cdot\|$ denotes the Euclidean metric. The aim is to estimate both θ and g in the equivalent model

$$Y = g(\theta^T x) + \varepsilon. \tag{1.1}$$

In the form (1.1), a single-index model is similar to the first step of projection pursuit regression. There, the model generating the data is usually taken to be

$$Y = g_1(x) + \varepsilon,$$

where g_1 is a p-variate function. The "first projective approximation" to $g_1(x)$ is a function $g(\theta^T x)$, where g is a univariate function, θ is a unit vector and (g, θ) are chosen to minimize $E\{g_1(x) - g(\theta^T x)\}^2$ when x has the distribution of the design variable x. Hall (1989) showed that in the context of this

problem, θ can be estimated root-n consistently. Ichimura (1987) studied the case of single-index models, and also showed that θ can be estimated root-n consistently.

Estimation of either g or θ requires a degree of statistical smoothing. Perhaps the simplest approach is to use kernel methods to construct an approximation $\hat{g}$ of g; thus substitute $\hat{g}$ into an empirical version $\hat{S}(\theta)$ of the mean squared error $S(\theta) = E\{Y - g(\theta^T x)\}^2$; and finally, choose $\hat{\theta}$ to minimize $\hat{S}$. However, performance of this method could depend significantly on the bandwidth chosen for $\hat{g}$. Furthermore, having estimated θ we still need a bandwidth for computing a good estimator for g.

It is not clear, a priori, whether the same bandwidth can be used to construct good estimators of both θ and g. Evidence in Hall [(1989), page 583] suggests that two quite different bandwidths may be necessary—the first to construct a preliminary estimator of g so that θ may be estimated, and the second to construct a final estimator of g. For example, in the projection pursuit version of this problem, a bandwidth of the order which optimizes $\hat{g}$ as an estimator of g will not produce a root-n consistent estimator of θ. Moreover, although Ichimura's (1987) study of single-index models gives a range of bandwidth which enables one to construct a root-n consistent $\hat{\theta}$, that range excludes the size of bandwidth which is optimal for estimating g. Our aim in the present paper is to resolve this problem, and to suggest a practical, empirical way of selecting bandwidth(s) for optimal estimation of both θ and g.

We shall show that, contrary to the projection pursuit case, the *same* bandwidth h can be used for estimating θ and g. We suggest a version of $\hat{S}$ which is a function of both θ and h, and propose that $\hat{S}$ be minimized simultaneously with respect to these variables. An attractive feature of our definition of $\hat{S}(\theta, h)$ is that it can be expanded in the form $\hat{S}(\theta, h) = \tilde{S}(\theta) + T(h) +$ remainder terms, where $\tilde{S}(\theta)$ is an accurate approximation to $S(\theta)$ and does not depend on h, and $T(h)$ is the usual cross-validation criterion for choosing h when estimating $g(\theta_0^T x)$ for known θ_0. Therefore, minimizing $\hat{S}(\theta, h)$ simultaneously with respect to both θ and h is very much like separately minimizing $\tilde{S}(\theta)$ with respect to θ and $T(h)$ with respect to h. It produces a root-n consistent estimator of θ and an asymptotically optimal estimator of h.

We shall address the heteroscedastic case, where the variance of the error term ε can depend on the design variable x. In this context, minimum variance lower bounds for estimating θ require appropriate weights to be introduced into the definition of $\hat{S}$. Those weights might, for example, be proportional to error variances. When that is done, the bandwidth estimator $\hat{h}$ obtained by minimizing $\hat{S}(\theta, h)$ will be asymptotically optimal with respect to a certain weighted version of mean integrated squared error. The particular term of the weight function in the latter may not always be that which one derives—bear in mind that the weights are specially chosen for optimal estimation of θ, not of h—but this difficulty may be remedied by using a two-stage approach.

Our techniques extend to the case of multiple-index models, of the form

$$Y = g(\theta_1^T x, \ldots, \theta_m^T x) + \varepsilon,$$

where again, bandwidth and orientation can be selected by simultaneous minimization of a criterion analogous to $\hat{S}$.

Section 2 describes the methodology behind our approach and states the main theorem. Numerical examples are discussed in Section 3, and Section 4 presents the proof of the main theorem.

2. Methodology.

2.1. *Summary.* Section 2.2 introduces notation and definitions for data generated by a single-index model. Our estimators are proposed in Section 2.3, and their asymptotic behavior is outlined in Section 2.4. The results described there are made rigorous in Section 2.5, which states the main theorem. Finally, Section 2.6 treats the case of weighted least squares, appropriate when the errors are heteroscedastic.

2.2. *Model.* We assume that the recorded data (x_i, Y_i), $1 \le i \le n$, are generated by the model

$$Y_i = g(\theta_0^T x_i) + \varepsilon_i,$$

where g is a smooth univariate function, θ_0 is a p-variate unit vector, $x_1, \ldots, x_n$ represent observed values of a random sequence of p-vectors $X_1, \ldots, X_n$, and $\varepsilon_1, \ldots, \varepsilon_n$ are independent random variables with zero mean and bounded variance. It is supposed that the $(p+1)$-tuples (X_i, ε_i) are independent and identically distributed. Writing x_i for X_i serves to indicate that, in the spirit of regression problems, the X_i's are regarded as fixed. Under this conditioning, the distribution of ε_i (in particular, the variance) may depend on x_i. However, we shall not explicitly consider the impact of this dependence until Section 2.6.

2.3. *Estimators.* Let $A \subseteq \mathbb{R}^p$ be a set chosen so that the denominator in the formulas for kernel estimators does not get too close to 0; details will be given in Section 2.5. Assume that the kernel function K (typically a symmetric probability density) has support $(-1, 1)$, and define $A^{2h} = \{x \in \mathbb{R}^p: \|x - y\| \le 2h \text{ for some } y \in A\}$.

Let (X, Y) have the distribution of a generic pair (X_i, Y_i) and define

$$g(u|\theta) = E(Y|\theta^T X_A = u),$$

where X_A has the distribution of X conditional on $X \in A$. Here and below, θ is always a unit p-vector. The function g is particularly easy to estimate, with

one estimator being

$$\hat{g}(u|\theta) = \left\{ \sum_{j=1}^{n} Y_j K_h(u - \theta^T x_j) \right\} \Big/ \left\{ \sum_{j=1}^{n} K_h(u - \theta^T x_j) \right\},$$

where h is a bandwidth, $K_h(\cdot) = K(\cdot/h)$, and K is a fixed kernel function (typically a symmetric probability density function). If the pair (X_i, Y_i) is omitted from this calculation, then we obtain the estimator

$$\hat{g}_i(u|\theta) = \left\{ \sum_{j \neq i} Y_j K_h(u - \theta^T x_j) \right\} \Big/ \left\{ \sum_{j \neq i} K_h(u - \theta^T x_j) \right\}.$$

Since $g(\cdot|\theta_0) \equiv g$ we may estimate θ by selecting that orientation θ which minimizes a measure of the distance $g(\cdot|\theta) - g$. To this end, define

$$\hat{S}(\theta, h) = \sum_i{}' \{Y_i - \hat{g}_i(\theta^T x_i|\theta)\}^2,$$

where $\sum_i'$ denotes summation over indices i such that $x_i \in A$.

Our aim is to choose θ close to θ_0, and h close to the value h_0 which minimizes the average of $E\{\hat{g}(\theta_0^T x|\theta_0) - g(\theta_0^T x)\}^2$ over $x \in A$. We claim that minimizing $\hat{S}(\theta, h)$ over both variables, simultaneously, achieves this goal. Indeed, we shall prove that

$$\hat{S}(\theta, h) = \tilde{S}(\theta) + T(h) + \text{negligible terms}, \tag{2.1}$$

where

$$\tilde{S}(\theta) = \sum_i{}' \{Y_i - g(\theta^T x_i|\theta)\}^2 \tag{2.2}$$

is the distance measure we would employ instead of $\hat{S}$ if we knew $g(\cdot|\theta)$, and

$$T(h) = \sum_i{}' \{\hat{g}_i(\theta_0^T x_i|\theta_0) - g(\theta_0^T x_i)\}^2 \tag{2.3}$$

is the usual cross-validation estimate of the mean squared distance between $\hat{g}(\cdot|\theta_0)$ and g. Thus, minimizing $\hat{S}(\theta, h)$ simultaneously with respect to both θ and h is very much like separately minimizing $\tilde{S}(\theta)$ with respect to θ and $T(h)$ with respect to h.

A comment on the "negligible terms" in (2.1) is in order. We shall prove that

$$\begin{aligned} \hat{S}(\theta, h) = \tilde{S}(\theta) + T(h) + \{&\text{terms of smaller order than } T(h) \\ &\text{and not depending on } \theta\} \\ + \{&\text{terms of smaller order than either } \tilde{S} \text{ or } T(h)\}. \end{aligned} \tag{2.4}$$

Now, $T(h)$ is of larger size than $\tilde{S}(\theta)$, and there are remainder terms on the right-hand side which are larger than $\tilde{S}(\theta)$ but smaller than $T(h)$. However, as indicated in (2.4), those terms do not depend on θ, and so do not upset the argument recounted in the previous paragraph.

2.4. *Asymptotic behavior of* $\hat{\theta}, \hat{h}$. Let $(\hat{\theta}, \hat{h})$ denote the pair which minimizes $\hat{S}(\theta, h)$. As suggested by the discussion in Section 2.3, $\hat{\theta}$ is (essentially) the minimizer of $\tilde{S}(\theta)$, and $\hat{h}$ is (essentially) the minimizer of $T(h)$; arguing thus we may show that $\hat{\theta}$ is root-n consistent for θ_0, and that $\hat{h}/h_0 \to 1$ in probability, where h_0 is the theoretically optimal bandwidth which minimizes

$$J(h) = \int_A E\{\hat{g}(\theta_0^T x|\theta_0) - g(\theta_0^T x)\}^2 f(x)\,dx, \tag{2.5}$$

and f denotes the design density. In fact, in the case of homoscedastic error with $E(\varepsilon_1^2) = \sigma^2$ and for any unit vector $\omega \neq \pm\theta_0$, $n^{1/2}\omega^T(\hat{\theta} - \theta_0)$ is asymptotically normal $N(0, \sigma^2\omega^T W_0^- \omega)$, where W_0 is a $p \times p$ matrix defined by

$$\begin{aligned} W_0 = \int_A &\{x - E(X_A|\theta_0^T X_A = \theta_0^T x)\}\{x - E(X_A|\theta_0^T X_A = \theta_0^T x)\}^T \\ &\times g'(\theta_0^T x)^2 f(x)\,dx, \end{aligned} \tag{2.6}$$

X_A has the distribution of X conditional on $X \in A$, and W_0^- denotes a generalized inverse of W_0. Note particularly that the first-order asymptotic behavior of $\hat{\theta}$ involves neither the kernel nor the bandwidth. The next section will describe the theory behind these claims.

2.5. *Main theorem*. We impose the following regularity conditions. Assume that $A \subseteq \mathbb{R}^p$ is the union of a finite number of open convex sets. Given $\delta > 0$, let A^δ denote the set of all points in $\mathbb{R}^p$ distant no further than δ from A. Put $\mathscr{U} = \{\theta_0^T x\colon x \in A^\delta\}$, and let γ denote the density of $\theta_0^T X$. Assume that for some $\delta > 0$,

(2.7) f is bounded away from 0 on A^δ and has two bounded derivatives there;

(2.8) g and γ have two bounded, continuous derivatives on $\mathscr{U}$;

(2.9) K is supported on the interval $(-1, 1)$ and is a symmetric probability density, with a bounded derivative;

(2.10) $E(\varepsilon_i|x_i) = 0$, $E(\varepsilon_i^2|x_i) = \sigma^2(x_i)$ for all i, where the function σ^2 is bounded and continuous and $\sup_i E|\varepsilon_i|^m = M_m < \infty$ for all m.

The emphasis on two derivatives in (2.7) and (2.8) is because we are using a second-order kernel; see (2.9). This means that the "optimal" bandwidth h_0, in the sense of minimizing the mean integrated squared error $J(h)$ defined at (2.5), is asymptotic to a constant multiple of $n^{-1/5}$. All our results have analogues for an rth-order kernel [see Härdle (1990), page 135, for a definition], but there we would demand r derivatives of f, g and γ. In (2.7), the restriction that f be bounded away from 0 on A^δ ensures that the denominators in the definitions of $\hat{g}(u|\theta)$ and $\hat{g}_i(u|\theta)$ are, with high probability, bounded away from 0 for $u = \theta^T x$, $x \in A$ and θ near θ_0. The requirement in (2.9) that K be compactly supported can be removed at the expense of a longer

argument; for example, the standard normal kernel is permissible. Finally, the condition that all moments of the ε_i's be bounded [see (2.10)] can be relaxed, to one of boundedness of moments of sufficiently high order. However, our proof at this point, given in step (ii) of Section 4, does not provide a particularly efficient estimate of the "minimum" moment condition, and so we shall not pursue this matter any further.

Let Θ denote the set of all unit p-vectors. Given $C > 0$ and $0 < C_1 < C_2 < \infty$, $\Theta_n = \{\theta \in \Theta: \|\theta - \theta_0\| \le Cn^{-1/2}\}$, $\mathscr{H}_n = \{h: C_1 n^{-1/5} \le h \le C_2 n^{-1/5}\}$. These definitions are motivated by the fact that, since we anticipate that $\hat{\theta}$ is root-n consistent, and we expect $\hat{h}$ to be close to $h_0 \sim \text{const}\, n^{-1/5}$, we should look for a minimum of $\hat{S}(\theta, h)$ which involves θ distant from θ_0 by order $n^{-1/2}$ and h approximately equal to a constant multiple of $n^{-1/5}$. Define

$$\mu(x|\theta) = E(X_A|\theta^T X_A = \theta^T x), \qquad K_1 = \int z^2 K(z)\, dz,$$

$$K_2 = \int K^2(z)\, dz,$$

$$V = \sum_i' \{x_i - \mu(x_i|\theta_0)\} g'(\theta_0^T x_i)\varepsilon_i,$$

$$A_1 = K_2 \int_A \gamma(\theta_0^T x)^{-1} \sigma(x)^2 f(x)\, dx, \qquad A_2 = \tfrac{1}{4} K_1^2 \int_A g''(\theta_0^T x)^2 f(x)\, dx. \tag{2.12}$$

In this notation, $J(h) \sim A_1 h^{-1} + A_2 n h^4$ and $h_0 \sim \{A_1/(4nA_2)\}^{1/5}$ as $n \to \infty$.

THEOREM. *Under the preceding conditions we may write*

$$\hat{S}(\theta, h) = \tilde{S}(\theta) + T(h) + R_1(\theta, h) + R_2(h), \tag{2.13}$$

where $\tilde{S}(\theta)$ and $T(h)$ are given by (2.2) *and* (2.3), *$R_2(h)$ does not depend on θ, and*

$$\sup_{\theta \in \Theta_n,\, h \in \mathscr{H}_n} |R_1(\theta, h)| = o_p(n^{1/5}), \qquad \sup_{h \in \mathscr{H}_n} |R_2(h)| = o_p(1). \tag{2.14}$$

Furthermore,

$$\tilde{S}(\theta) = n\{W_0^{1/2}(\theta - \theta_0) - n^{-1/2}\sigma Z\}^T \{W_0^{1/2}(\theta - \theta_0) - n^{-1/2}\sigma Z\} + R_3 + R_4(\theta), \tag{2.15}$$

$$T(h) = A_1 h^{-1} + A_2 n h^4 + R_5(h), \tag{2.16}$$

where W_0, A_1 and A_2 are given by (2.6) *and* (2.12), *Z is an asymptotically normal $N(0, I)$ random p-vector such that $V = n^{1/2}\sigma W_0^{1/2} Z$, R_3 depends on neither θ nor h, and*

$$\sup_{\theta \in \Theta_n} |R_4(\theta)| = o_p(1), \qquad \sup_{h \in \mathscr{H}_n} |R_5(h)| = o_p(n^{1/5}). \tag{2.17}$$

Formulas (2.13) and (2.14) together provide a rigorous description of (2.4). It follows from (2.15)–(2.17) that with probability tending to 1 as $n \to \infty$, the minimum of $\hat{S}(\theta, h)$ within a radius $O(n^{-1/2})$ of θ_0 for the first variable, and

on a scale of $n^{-1/5}$ for the second variable, satisfies for any unit vector $\omega \neq \pm\theta_0$,

$$\omega^T(\hat\theta - \theta_0) = \omega^T\{n^{-1/2}\sigma(W_0)^{-1/2}Z\} + o_p(n^{-1/2}) \quad \text{and} \quad \hat h = h_0 + o_p(n^{-1/5}),$$

where W_0^- denotes a generalized inverse of W_0. The limit theorems claimed in Section 2.4 for $\hat\theta$ and $\hat h$, that is, $\hat h/h_0 \to 1$ in probability and [in the case where $\sigma(x)^2$ is constant] $n^{1/2}\omega^T(\hat\theta - \theta_0) \to N(0, \sigma^2\omega^T W_0^- \omega)$ in distribution, are immediate consequences.

2.6. *Heteroscedastic errors.* It is clear from the theorem that the estimator $\hat\theta$ is root-n consistent for θ_0, even when the errors ε_i are heteroscedastic. However, in the heteroscedastic case the efficiency of the estimator $\hat\theta$ can be improved by introducing an appropriate weight function, w, when defining the distance criterion $\hat S$. Ichimura (1990) studies this case using a deterministic smoothing parameter. In this section we shall outline the optimal smoothing when w is incorporated and investigate the case where w must be estimated empirically.

We assume throughout that the error variance $\sigma^2(x)$ is actually a function of $\theta_0^T x$, in which case it is appropriate to take w to be also a function of $\theta_0^T x$. Using a weight function can have its disadvantages, as well as its advantages. Aside from the additional computational complexity (particularly if the weights are determined empirically), the weight function alters the definition of $J(h)$ at (2.5), in a way which is not necessarily desirable. However, this problem can be overcome using a multistage approach, as we shall show.

Redefine $\hat S$, $\tilde S$, T and V by

$$\hat S(\theta, h) = \sum_i{}'\{Y_i - \hat g_i(\theta^T x_i|\theta)\}^2 w(x_i), \qquad \tilde S(\theta) = \sum_i{}'\{Y_i - g(\theta^T x_i|\theta)\}^2 w(x_i),$$

$$T(h) = \sum_i{}'\{\hat g_i(\theta_0^T x_i|\theta_0) - g(\theta_0^T x_i)\}^2 w(x_i),$$

$$V = \sum_i{}'\{x_i - \mu(x_i|\theta_0)\}g'(\theta_0^T x_i)w(x_i)\varepsilon_i.$$

Let W_0, A_1 and A_2 be as defined at (2.6) and (2.12), respectively, except that f is replaced by fw throughout. Provided only that w is a bounded, continuous, positive function, the theorem continues to hold, with an identical proof. Now, the variance of $n^{-1/2}V$ is

$$n^{-1}\sum_i{}'\{x_i - \mu(x_i|\theta_0)\}\{x_i - \mu(x_i|\theta_0)\}^T g'(\theta_0^T x_i)^2 w(x_i)^2\sigma(x_i)^2$$

$$\to W_1 = \int_A \{x - E(X_A|\theta_0^T X_A = \theta_0^T x)\}\{x - E(X_A|\theta_0^T X_A = \theta_0^T x)\}^T$$

$$\times g'(\theta_0^T x)^2 w(x)^2\sigma(x)^2\,dx.$$

When $w(x) = \sigma^{-2}(x)$ and $\sigma^2(x)$ is only a function of $\theta_0^T x$, result (2.15) implies that $n^{1/2}(\hat\theta - \theta_0)$ is asymptotically normal $N(0, W_1^-)$, where W_1^- denotes a

generalized inverse of W_1. Furthermore, $\hat{h}/h_0 \to 1$ in probability, where $h_0 \sim \{A_1/(4nA_2)\}^{1/5}$ denotes the bandwidth which minimizes

$$J(h) = \int_A E\{\hat{g}(\theta_0^T x|\theta_0) - g(\theta_0^T x)\}^2 w(x) f(x)\, dx \tag{2.18}$$

[identical to (2.5), except that the weight function has been included].

This particular limit distribution represents the minimum variance lower bound in certain cases of practical importance. For example, in the model $Y_i = g(\theta_0^T x_i) + \varepsilon_i$, where $\sigma(x)$ is a function only of $\theta_0^T x$, and the ε_i's are independent normal $N(0, \sigma(x_i)^2)$, the minimax-optimal estimator of θ_0 computed from the sample of pairs $\{(x_i, Y_i)\colon x_i \in A\}$ has asymptotic variance $n^{-1}W_0$ [where W_0 is defined with $w(x) \equiv \sigma(x)^{-2}$]. Cosslett (1987) treated the case of binary choice models, where $n^{-1}W_0$ is again a minimum variance bound.

In practice, the variance function $\sigma(x)^2$ would usually be unknown, and would require estimation. We shall restrict our attention to the case where

$$\sigma(x)^2 = \tau^2 G\{g(\theta_0^T x)\},$$

where G is a known, smooth function and τ is a (possibly unknown) constant. A two-stage procedure is suggested, as follows.

(I) Conduct inference as in Sections 2.3–2.5, taking the weight function w to be identically 1. Let $(\hat{\theta}, \hat{h}_1)$ denote the resulting estimates, obtained by minimizing the unweighted version of $\hat{S}$.

(II) In the definition of $\hat{S}(\theta, h)$ in Section 2.3, replace $w(x_i)$ by

$$G\{\hat{g}(\hat{\theta}_1^T x_i|\hat{\theta}_1)\}^{-1},$$

in which formula $\hat{h}_1$ replaces h during the computation of $\hat{g}$. Recalculate $(\hat{\theta}, \hat{h}) = (\hat{\theta}_2, \hat{h}_2)$ by minimizing the weighted form of $\hat{S}$. For the two-stage algorithm, minimization should not be taken over the weight function.

It may be shown that if G is a twice-differentiable function, bounded away from 0, then the first-order asymptotics of this algorithm are identical to those which would obtain if we were to take $w(x) \equiv G\{g(\theta_0^T x)\}^{-1}$ in a one-stage weighted procedure. That is, $n^{1/2}(\hat{\theta} - \theta_0) \to N(0, W_0)$, where W_0 admits the definition at (2.6) but with $f(x)$ replaced by $f(x)\tau^{-2}G\{g(\theta_0^T x)\}^{-1}$, and $\hat{h}_2 h_0^{-1} \to 1$ in probability, where H_0 minimizes the function $J(h)$ defined at (2.18), with $w(x)$ replaced by $G\{g(\theta_0^T x)\}^{-1}$. The bandwidth $\hat{h}_1$ from the first stage provides asymptotic minimization of the integrated squared error formula at (2.5), rather than that at (2.18).

3. The method in practice. We examined the practicability of our method in several simulated situations and an application involving weighted cross-validation. The simulations were performed with different size n and with X_1 and X_2 independently uniformly distributed on $[0, 1]^2$. The true

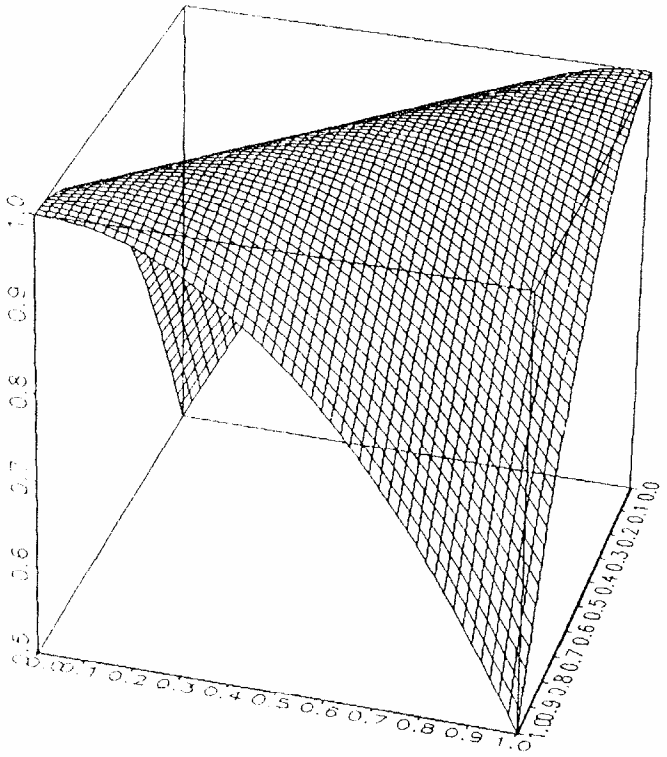


FIG. 1. *The function* $g(\theta_0^T X; 1)$ *on the unit square.*

parameter vector was $\theta_0 = (1, 1)^T/\sqrt{2}$. The link function was $g(u; C) = -C(u - 1/\sqrt{2})^2 + C$ with $C = 1, 4$. An impression of the function $g(u; 1)$ can be gained from Figure 1. We have chosen different "steepness parameters" C to study the performance with different signal-to-noise ratios.

The error distribution was selected to be standard normal with standard deviation $\sigma = 0.2$. All computations were done in GAUSSS 2.0 using the random seed 1678321. The objective function $\hat{S}(\theta, h)$ was computed with a quartic kernel $K(u) = \frac{15}{16}(1 - u^2)^2 I(|u| \leq 1)$ on the projected X-values $X_i^T\theta$.

In order to avoid problems with local minima a grid search was implemented. The grid search was performed for h on the interval $[0.05, 0.45]$ at 10 gridpoints. The projection vector θ on the unit circle was parametrized by an angle $\varphi \in [0, \pi)$. The true parameter θ_0 corresponds to $\varphi_0 = \pi/4$. Preliminary computations showed that $S(\theta, h)$ was very sensitive to $\varphi \notin \phi_0 = [\pi/8, 3\pi/8]$ in the sense that outside ϕ_0 the objective function became very large. Therefore we restricted our grid of 10 points for φ to the interval ϕ_0. In Table 1 we report the results over 100 simulations. In this table the mean (and standard errors) of $\hat{h}$ and $\hat{\varphi}$ [minimizing $S(\theta, h)$] are given as a function of sample size n and curve parameter C.

Table 1 confirms our theoretical results. As the sample size increases the bandwidth h becomes smaller, the direction is more accurately estimated. The shape parameter C has an influence on the selected (h, φ). The direction and the bandwidth are more accurately estimated for $C = 4$ with one exception in the last row of Table 1. There the selected bandwidth for $n = 200$ was on the

TABLE 1

Mean and standard deviations (in parentheses) of estimated direction and bandwidth as a function of sample size n and curve parameter C

n	C	$\hat{h}$	$\hat{\varphi}$
25	1	0.244 (0.136)	0.752 (0.117)
	4	0.153 (0.079)	0.779 (0.098)
50	1	0.208 (0.133)	0.769 (0.110)
	4	0.116 (0.064)	0.766 (0.103)
100	1	0.212 (0.116)	0.784 (0.105)
	4	0.097 (0.045)	0.792 (0.084)
200	1	0.162 (0.046)	0.773 (0.092)
	4	0.156 (0.046)	0.782 (0.045)

average higher than for $n = 100$. The reported standard deriviations though allow us to attribute this phenomenon to sample fluctuations.

A visual impression of what Table 1 means to the data can be obtained from Figure 2. The kernel smoother $g(u|\hat{\theta})$ was computed at the grid $0.1, 0.2, \ldots, 1.3$ for the optimum $\hat{\theta}$ and $\hat{h}$. At each gridpoint we computed a 95% confidence interval. The joined confidence intervals together with the true function $g(u|\theta_0)$ and the mean of $\hat{g}(u|\hat{\theta})$ over the 200 simulations are shown.

As an application we have chosen the side impact example described in Härdle and Stoker (1989), where also a table of the data is given. In this

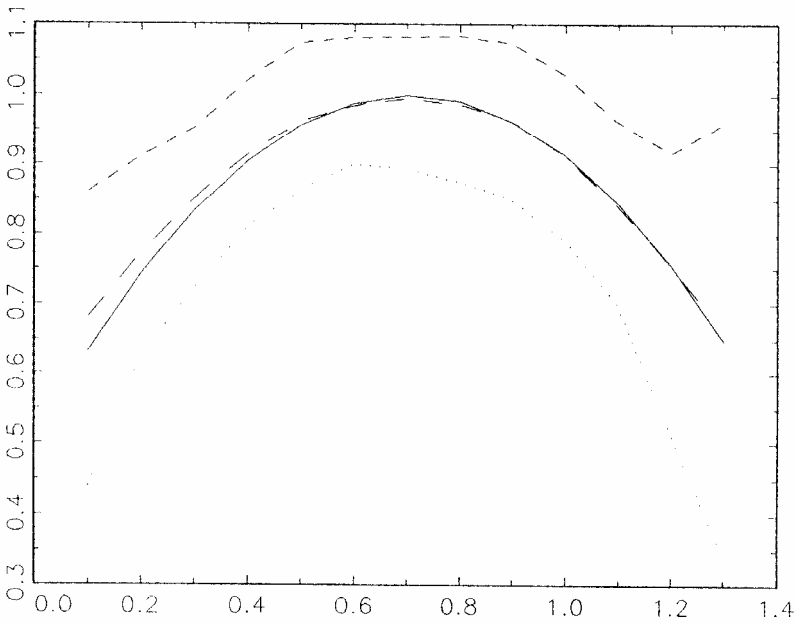


FIG. 2. *The true curve $g(u; 1)$ (solid line), the mean of $\hat{g}_{\hat{h}}(u|\hat{\theta})$ over 200 simulations (long dashes), the upper 95% confidence intervals (short dashes) and the lower 95% confidence intervals (dotted line).*

example Y is binary; $Y \in \{0, 1\}$ and the predictor variable is $p = 3$ dimensional; there are $n = 51$ observations. The first variable corresponds to the age of the subject, the second corresponds to the velocity of the automobile, and the third corresponds to the maximal acceleration (upon impact) measured on the subject's 12th rib. The response variable corresponds to the severity of a side impact accident. It is quite common for these kinds of data to postulate a single-index model; see McCullagh and Nelder (1983).

We standardized the regressors; each variable is centered by its sample mean and divided by its standard deviation. This enables a direct comparison with the results by Härdle and Stoker (1989).

Again we performed a grid search using the quartic kernel and found the optimal parameters to be

$$\hat{\theta} = (0.3, 0.3, 0.9).$$

After normalizing the length of θ vector to be 1, Härdle and Stoker (1989) estimated θ_0 via the average derivative method to be $(0.89, 0.34, 0.30)$. The advantage of our method is that the bandwidth choice is automatic. The advantage of their method is that the estimator has a closed form once the bandwidth is set in advance.

The dependence of $S(\hat{\theta}, h)$ on h can be seen from Figure 3 where we display the objective function as a function of bandwidth. The parameter θ is held

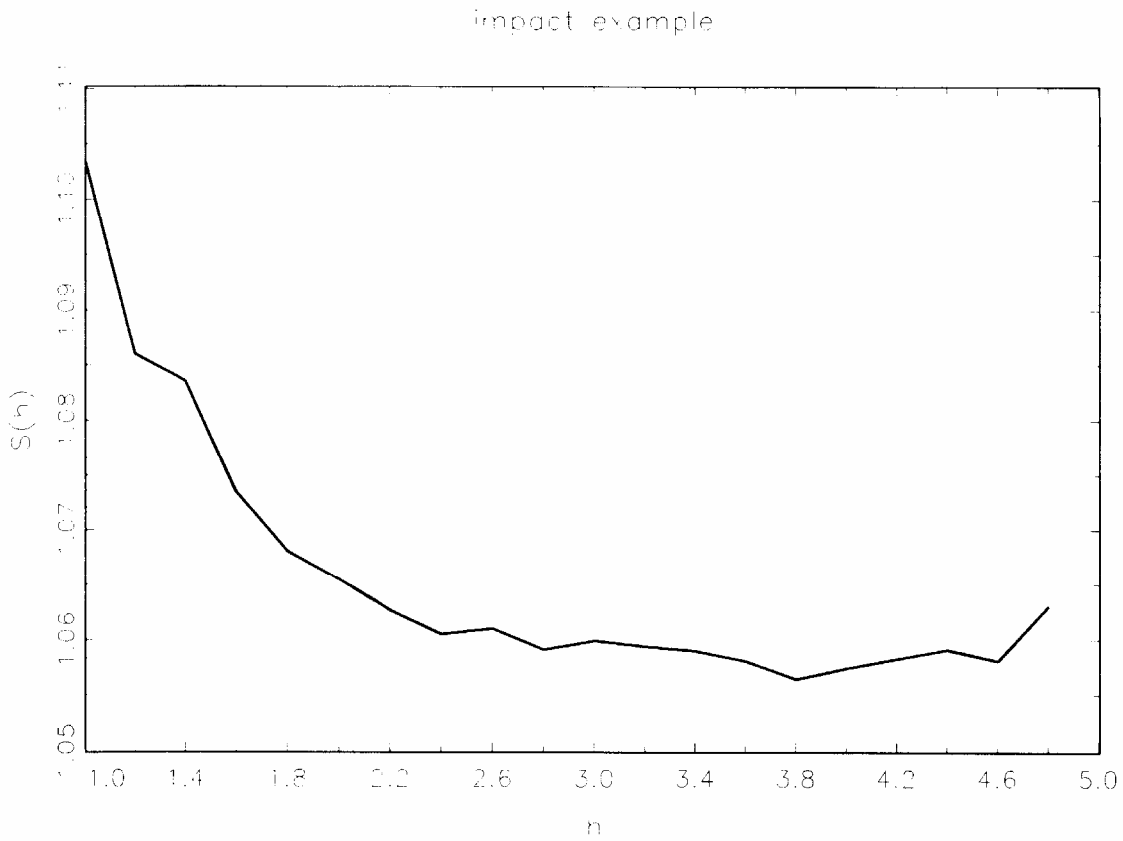


FIG. 3. *The objective function $S(\hat{\theta}, h)$ as a function of h.*

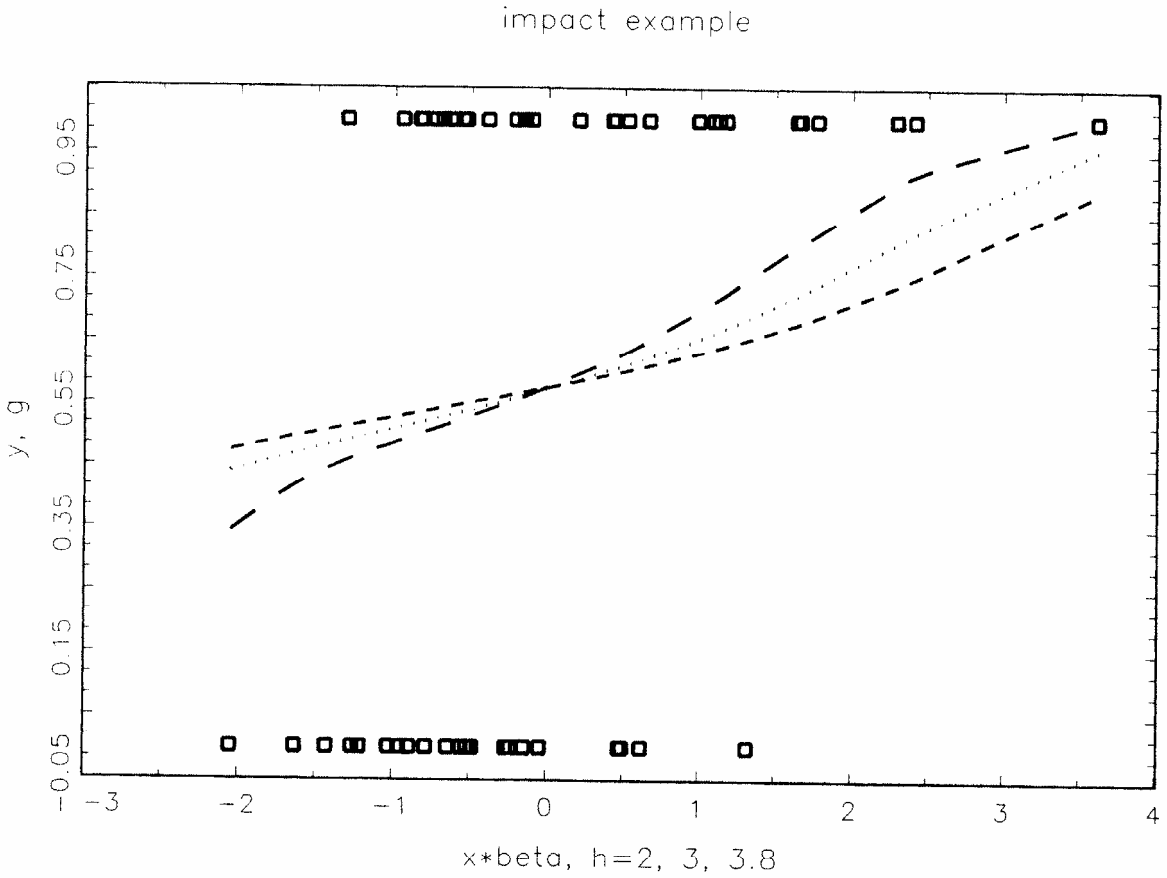


FIG. 4. *The projected data* $\{X_i^T\hat{\theta}, Y_i\}_{i=1}^n$ *and the optimal kernel regression estimate* $\hat{g}(u|\hat{\theta})$ *with* $\hat{\theta} = (0.3, 0.3, 0.9)$ *and* $h = 2$ (*solid line*), $h = 3$ (*dashed line*), $h = 2$ (*dotted line*).

fixed at its respective optimum for that h. One sees that the optimum h is about 3.8 with a flat minimum of $S(\hat{\theta}, h)$. This optimal bandwidth leads to a very smooth estimate of the link function $g(u|\hat{\theta})$. The projected data though is very similar to the indices published in Härdle and Stoker (1989).

Figure 4 shows the projected data $X_i^T\hat{\theta}$ together with the estimated nonparametric link function $\hat{g}_{\hat{h}}(u|\hat{\theta})$. For comparison we also display link functions with smaller bandwidths.

As already noted in Härdle and Stoker (1989), the nonparametric link function shows an asymmetric shape. A bootstrap method for comparing this model with a parametric one (e.g., with a logistic link function) is described in Azzalini, Bowmann and Härdle (1989).

4. Proof of theorem. The proof is given only in outline and is divided into nine steps, of which steps (iii)–(ix) control specific remainder terms. An overview of the entire proof is given in step (i), which draws together the various remainder term estimates from later steps.

If $\mathscr{E}$ denotes an event depending on the design sequence $x_1, x_2, \ldots$, we say that $\mathscr{E}$ occurs "with X-probability 1" if there exists a set E in the σ-field generated by $\mathscr{H} = \{X_1, X_2, \ldots\}$ such that $P_{\mathscr{H}}(E) = 1$ and $E \subseteq \mathscr{E}$.

Step (i): *Preliminaries.* Define $\tilde{S}(\theta) = \sum_i' \{Y_i - g(\theta^T x_i | \theta)\}^2$,

$$D_i = \hat{g}_i(\theta_0^T x_i | \theta_0) - g(\theta_0^T x_i), \qquad \delta_i = g(\theta^T x_i | \theta) - g(\theta_0^T x_i),$$

$$\Delta_i = \hat{g}_i(\theta^T x_i | \theta) - g(\theta^T x_i | \theta) - \{\hat{g}_i(\theta_0^T x_i | \theta_0) - g(\theta_0^T x_i)\}.$$

In this notation,

$$\hat{S}(\theta, h) - \tilde{S}(\theta) = \sum_i'(D_i^2 + \Delta_i^2) + 2\sum_i'(D_i\Delta_i + D_i\delta_i + \Delta_i\delta_i - D_i\varepsilon_i - \Delta_i\varepsilon_i),$$

whence

$$\left|\hat{S}(\theta, h) - \tilde{S}(\theta) - \sum_i' D_i^2 + 2\sum_i' D_i\varepsilon_i\right|$$

$$\leq \sum_i' \Delta_i^2 + 2\left(\sum_i' \Delta_i^2\right)^{1/2}\left\{\left(\sum_i' D_i^2\right)^{1/2} + \left(\sum_i' \delta_i^2\right)^{1/2}\right\} \tag{4.1}$$

$$+ 2\left|\sum_i' D_i\delta_i\right| + 2\left|\sum_i' \Delta_i\varepsilon_i\right|.$$

We assume that $\|\theta - \theta_0\| \leq Cn^{-1/2}$, for a fixed constant $C > 0$. We may write

$$\theta = (1 - \eta^2)^{1/2}\theta_0 + \eta\theta_{00}, \tag{4.2}$$

where $\theta_{00} \perp \theta_0$, and θ_{00} is on the same plane as θ and θ_0.

In outline, our argument from this point runs as follows. We show that with X-probability 1, and for all $\xi > 0$,

$$\sum_i' \Delta_i^2 = O_p(n^{-2/5+\xi}).$$

See steps (iii) and (iv). It is straightforward to prove that $\sum_i' E(D_i^2) = O(n^{1/5})$, whence $T(h) = \sum_i' D_i^2 = O_p(n^{1/5})$. By Taylor expansion from (4.2) it follows that $\delta_i = O(n^{-1/2})$ uniformly in i (meaning, here and below, uniformly in i such that $X_i \in A$). Therefore, $\sum_i' \delta_i^2 = O(1)$, and so

$$\sum_i' \Delta_i^2 + 2\left(\sum_i' \Delta_i^2\right)^{1/2}\left\{\left(\sum_i' D_i^2\right)^{1/2} + \left(\sum_i' \delta_i^2\right)^{1/2}\right\}$$

$$= O_p\left\{n^{-2/5+\xi} + \left(n^{-2/5+\xi}(n^{1/5})\right)^{1/2}\right\} = o_p(1),$$

on choosing $0 < \xi < 1/5$.

Steps (v) and (vi) show that $|\sum_i' D_i\delta_i| = O_p(n^{-3/10+\xi})$, and step (viii) that $|\sum_i' \Delta_i\varepsilon_i| = O_p(n^{-1/5+\xi})$, for all $\xi > 0$. Therefore, the right-hand side of (4.1) equals $o_p(1)$. We prove in step (vii) that the term $\sum_i' D_i\varepsilon_i$, which does not depend on θ, is $O_p(n^{1/10+\xi})$. Hence, by (4.1),

$$\hat{S}(\theta, h) = \tilde{S}(\theta) + T(h) + \{\text{term not depending on } \theta, \text{ of size } o_p(n^{1/5})\}$$

$$+ o_p(1).$$

This formula, with the stated orders of the remainder terms, is available uniformly in $\theta \in \Theta_n$ and $h \in \mathscr{H}_n$, thereby establishing (2.13) and (2.14).

Standard techniques for cross-validation in nonparametric regression [e.g., Härdle, Hall and Marron (1988)] may be used to show that $T(h) = E\{T(h)\} + o_p(n^{1/5})$ and $E\{T(h)\} = J(h) + O(n^{1/5})$ [with $J(h)$ defined at (2.5)] $= A_1 h^{-1} + A_2 n h^4 + O(n^{1/5})$, uniformly in $h \in \mathscr{H}_n$. We show in step (ix) that $\tilde{S}(\theta)$ may be approximated by a quadratic form. Together, these results give (2.15)–(2.17).

Step (ii). For the sake of brevity and clarity our estimation of remainder terms in steps (iii)–(ix) is developed only for (arbitrary) single values $\theta \in \Theta_n$ and $h \in \mathscr{H}_n$. Uniformity is readily established by *straightforward modification* of those arguments, as we show in the present step.

Let $\varphi_n(\theta, h)$ be a (possibly random) quantity for which we show in steps (iii)–(ix) that

$$\varphi_n(\theta, h) = o_p(n^a) \tag{4.3}$$

for arbitrary sequences $\theta \in \Theta_n$ and $h \in H_n$. Examples include $\varphi_n = \sum_i' \Delta_i^2$ [from steps (iii) and (iv); call this Example 1] and $\varphi_n = \sum_i' D_i \varepsilon_i$ [from step (viii); call this Example 2]. We wish to strengthen (4.3) to

$$\sup_{\theta \in \Theta_n,\, h \in \mathscr{H}_n} |\varphi_n(\theta, h)| = o_p(n^a). \tag{4.4}$$

The method of proving (4.3) is, in all cases, based on moment bounds. In the case of Example 1 we show that $E(\varphi_n) = O(n^b)$, and in the case of Example 2, $E(\varphi_n^2) = O(n^{2b})$, where $b < a$. The proofs given in steps (iii)–(ix) in fact establish the moment bounds uniformly on $\theta \in \Theta_n$ and $h \in \mathscr{H}_n$. We claim that the bounds may be strengthened to

$$\sup_{\theta \in \Theta_n,\, h \in \mathscr{H}_n} E\big(\varphi_n / n^b\big)^{2l} = O(1) \tag{4.5}$$

for all integers $l \geq 1$. Accepting this for the time being, observe that if $\Theta_n' \subseteq \Theta_n$ and $H_n' \subseteq H_n$ are discrete sets each containing at most n^c elements, then for any $\alpha > 0$,

$$\begin{aligned} P\Big\{ \sup_{\theta \in \Theta_n',\, h \in \mathscr{H}_n'} |\varphi_n(\theta, h)| > \alpha n^a \Big\} \\ \leq 2n^c \sup_{\theta \in \Theta_n',\, h \in \mathscr{H}_n'} P\big\{ |\varphi_n(\theta, h)| > \alpha n^a \big\} \\ \leq 2n^c \big(n^b / \alpha n^a\big)^{2l} \sup_{\theta \in \Theta_n',\, h \in \mathscr{H}_n'} E\big\{ \varphi_n(\theta, h) / n^b \big\}^{2l} = O(1), \end{aligned}$$

provided only that l is chosen so large that $c < 2l(a - b)$. Therefore,

$$\sup_{\theta \in \Theta_n',\, h \in \mathscr{H}_n'} |\varphi_n(\theta, h)| = o_p(n^a), \tag{4.6}$$

for all sets $\Theta'_n \subseteq \Theta_n$, $H'_n \subseteq H_n$ whose cardinality increases no faster than a polynomial function of n. By making use of the smoothness conditions imposed on f, g and K, we may readily prove that for any given a, if $c = c(a)$ is sufficiently large, if Θ'_n denote regularly spaced sets of n^c points within Θ_n and $\mathscr{H}_n$, respectively, and if for each $(\theta, h) \in \Theta_n \times \mathscr{H}_n$, θ' and h' denote the values in Θ'_n and $\mathscr{H}'_n$ nearest to θ and h, respectively, then

$$\sup_{\theta \in \Theta'_n,\ h \in \mathscr{H}'_n} |\varphi_n(\theta, h) - \varphi_n(\theta', h')| = o_p(n^a). \tag{4.7}$$

Results (4.6) and (4.7) together imply (4.4).

It remains to prove (4.5), which may be done using Rosenthal's inequality [e.g., Hall and Heyde (1980), page 23]. We outline the method below in the case of Example 1; other cases are similar. Write

$$\varphi_n = \sum_i{}' \Delta_i^2 = \sum_i{}' (E\Delta_i)^2 + 2\sum_i{}' (E\Delta_i)(\Delta_i - E\Delta_i) + \sum_i{}' (\Delta_i - E\Delta_i)^2,$$

and further decompose the last series as

$$\sum_i{}' (\Delta_i - E\Delta_i)^2 = \sum_i{}' c_i \{\varepsilon_i^2 - \sigma(x_i)^2\} + \sum_i{}' \sum_{j \neq i}{}' c_{ij}\varepsilon_i\varepsilon_j + \sum_i{}' c_i \sigma(x_i)^2,$$

for constants c_i and c_{ij}. Thus $\varphi_n = \varphi_{n1} + \varphi_{n2} + \varphi_{n3}$, where

$$\varphi_{n1} = \sum_i{}' (E\Delta_i)^2 + \sum_i{}' c_i \sigma(x_i)^2$$

is purely deterministic,

$$\varphi_{n2} = 2\sum_i{}' (E\Delta_i)(\Delta_i - E\Delta_i) + \sum_i{}' c_i \{\varepsilon_i^2 - \sigma(x_i)^2\}$$

is a sum of independent random variables with zero means, and

$$\begin{aligned}
\varphi_{n3} &= \sum_i{}' \sum_{j \neq i}{}' c_{ij}\varepsilon_{ij} = \sum_{j<i}{}' (c_{ij} + c_{ji})\varepsilon_i\varepsilon_j \\
&= \sum_i{}' \sum_{j=1}^{i-1}{}' (c_{ij} + c_{ji})\varepsilon_i\varepsilon_j \\
&= \sum_{i=2}^{n}{}' \varepsilon_i \sum_{j=1}^{i-1}{}' (c_{ij} + c_{ji})\varepsilon_j = \sum_{i=2}^{n}{}' Z_i,
\end{aligned}$$

where

$$Z_i = \varepsilon_i \sum_{j=1}^{i-1}{}' (c_{ij} + c_{ji})\varepsilon_j.$$

Thus φ_{n3} is a martingale with differences Z_i. Arguing thus, and applying

Rosenthal's inequality for moments of martingales and sums of independent random variables, we may prove that

$$E(\varphi_n^{2l}) \le \left\{\varphi_{n1}^{2l} + E(\varphi_{n2}^{2l})^{1/2l} + E(\varphi_{n3}^{2l})^{1/2l}\right\}^{2l} = O(n^{3lb}).$$

Of course, we need only a finite, sufficiently large value of l, and so not all moments of ε need be assumed finite and bounded. However, our approach to the proof does not produce a moderately conservative upper bound to the number of moments required, and so we have asked in the statement of the theorem that all moments be finite.

Step (iii). Here we show that with X-probability 1, and for all $\xi > 0$,

$$\sum_i{}'(E\Delta_i)^2 = O(n^{-2/5+\xi}). \tag{4.8}$$

Define $\mu(x|\theta) = E(X_A|\theta^T X_A = \theta^T x)$. Observe that $E(\Delta_i) = d_i(\theta) - d_i(\theta_0)$, where $d_i(\theta) = E\{\hat{g}_i(\theta^T x_i|\theta)\} - g(\theta^T x_i|\theta)$. In view of the representation (4.2) we may write, for bounded x,

$$g(\theta_0^T x) = g(\theta^T x) - \eta(\theta_{00}^T x)g'(\theta_0^T x) + O(n^{-1}), \tag{4.9}$$

$$g(\theta^T x|\theta) = g(\theta^T x) - \eta\{\theta_{00}^T \mu(x|\theta)\}g'(\theta_0^T x) + O(n^{-1}). \tag{4.10}$$

Therefore,

$$\begin{aligned} d_i(\theta) &= \left[\sum_{j\ne i}\left\{g(\theta_0^T x_j) - g(\theta^T x_i|\theta)\right\}K_h\{\theta^T(x_i - x_j)\}\right] \\ &\quad \times\left[\sum_{j\ne i} K_h\{\theta^T(x_i - x_j)\}\right]^{-1} \\ &= a_i(\theta) + \eta\theta_{00}^T\{\mu(x_i|\theta)g'(\theta_0^T x_i) - V_i(\theta)\} + O(n^{-1}), \end{aligned}$$

where $a_i(\theta) = b_i(\theta)/c_i(\theta)$,

$$b_i(\theta) = (nh)^{-1}\sum_{j\ne i}\left\{g(\theta^T x_j) - g(\theta^T x_i)\right\}K_h\{\theta^T(x_i - x_j)\},$$

$$c_i(\theta) = (nh)^{-1}\sum_{j\ne i} K_h\{\theta^T(x_i - x_j)\},$$

$$V_i(\theta) = \left[(nh)^{-1}\sum_{j\ne i} x_j g'(\theta_0^T x_j)K_h\{\theta^T(x_i - x_j)\}\right]c_i(\theta)^{-1}.$$

Observe next that $\mu(x|\theta) - \mu(x|\theta_0) = O(n^{-1/2})$ and $V_i(\theta) - V_i(\theta_0) = O(n^{-1/2}h^{-1})$ uniformly in i. Therefore,

$$E(\Delta_i) = d_i(\theta) - d_i(\theta_0) = a_i(\theta) - a_i(\theta_0) + O(n^{-1}h^{-1}). \tag{4.11}$$

Furthermore, $b_i(\theta) = O(h^2 n^{\xi})$ for all $\xi > 0$, $c_i(\theta_0)$ is asymptotic to the density of $\theta_0^T X$ evaluated at $\theta_0^T x_i$, and $c_i(\theta) - c_i(\theta_0) = O(n^{-1/2}h^{-1})$. Hence,

$$\begin{aligned} a_i(\theta) - a_i(\theta_0) &= \{b_i(\theta) - b_i(\theta_0)\}c_i(\theta_0)^{-1} \\ &\quad + b_i(\theta)\{c_i(\theta_0) - c_i(\theta)\}\{c_i(\theta)c_i(\theta_0)\}^{-1} \\ &= \{b_i(\theta) - b_i(\theta_0)\}c_i(\theta_0)^{-1} + O(n^{-1/2+\xi}h) \end{aligned} \tag{4.12}$$

uniformly in i, for all $\xi > 0$.

To develop an approximation to $b_i(\theta) - b_i(\theta_0)$, note that $b_i(\theta)$ represents the observed value of $B_i(\theta, x_i)$, where

$$B_i(\theta, x) = (nh)^{-1} \sum_{j \neq i} \{g(\theta^T X_j) - g(\theta^T x)\} K_h\{\theta^T(x - x_j)\}.$$

Now,

$$\begin{aligned} &(1 - n^{-1})hE\{B_i(\theta, x) - B_i(\theta_0, x)\}/P(X \in A) \\ &\quad = E\big[\{g(\theta^T X_A) - g(\theta_0^T X_A) - g(\theta^T x) + g(\theta_0^T x)\}K_h\{\theta^T(x - X_A)\}\big] \\ &\qquad + E\big(\{g(\theta_0^T X_A) - g(\theta_0^T x)\}\big[K_h\{\theta^T(x - X_A)\} - K_h\{\theta_0^T(x - X_A)\}\big]\big) \\ &\quad = h\eta\theta_{00}^T E\big[\{X_A g'(\theta_0^T X_A) - x g'(\theta_0^T x)\}K_h\{\theta_0(x - X_A)\}\big] \\ &\qquad + \eta\theta_{00}^T E\big[\{g(\theta_0^T X_A) - g(\theta_0^T x)\}(x - X_A)K'\{h^{-1}\theta_0^T(x - X_A)\}\big] \\ &\qquad + O(n^{-1}) \\ &\quad = h\eta g'(\theta_0^T x)E\big(\{\theta_{00}^T(X_A - x)\}\big[K\{h^{-1}\theta_0^T(x - X_A)\} \\ &\qquad\qquad + \{h^{-1}\theta_0^T(x - X_A)\}K'\{h^{-1}\theta_0^T(x - X_A)\}\big]\big) + O(n^{-1/2}h^2) \\ &\quad = O(n^{-1/2}h^2). \end{aligned}$$

Therefore, $E\{B_i(\theta, x) - B_i(\theta_0, x)\} = O(n^{-1/2}h)$. More simply,

$$\begin{aligned} \mathrm{Var}\{B_i(\theta, x) - B_i(\theta_0, x)\} &= O\{(nh)^{-2}n(n^{-1/2}h^{-1})^2 h\} \\ &= O(n^{-2}h^{-3}) = O\{(n^{-1/2}h)^2\}, \end{aligned}$$

and so

$$B_i(\theta, x) - B_i(\theta_0, x) = O_p(n^{-1/2}h).$$

An argument based on the Borel–Cantelli lemma may now be used to prove that with X-probability 1, for all $\xi > 0$,

$$b_i(\theta) - b_i(\theta_0) = O(n^{-1/2+\xi}h) = O(n^{-7/10+\xi}).$$

Substituting into (4.12), we deduce that $a_i(\theta) - a_i(\theta_0) = O(n^{-7/10+\xi})$, whence

by (4.11), $E(\Delta_i) = O(n^{-7/10+\xi})$. Since these estimates are available uniformly in i, we obtain (4.8).

Step (iv). We prove that with X-probability 1,

$$\sum_i{}' \operatorname{Var}(\Delta_i) = O(n^{-2/5}). \tag{4.13}$$

Let $c_i(\theta)$ be as in the previous step and observe that

$$\begin{aligned}
\operatorname{Var}(\Delta_i) &= (nh)^{-2} \sum_{j \neq i} \Big[K_h\{\theta^T(x_i - x_j)\} c_i(\theta)^{-1} \\
&\qquad - K_h\{\theta_0^T(x_i - x_j)\} c_i(\theta_0)^{-1} \Big]^2 \sigma(x_j)^2 \\
&\le 2(nh)^{-2} \sum_{j \neq i} \Big[K_h\{\theta^T(x_i - x_j)\} - K_h\{\theta_0^T(x_i - x_j)\} \Big]^2 c_i(\theta_0)^{-2} \sigma(x_j)^2 \\
&\quad + 2(nh)^{-2} \sum_{j \neq i} K_h\{\theta_0^T(x_i - x_j)\}^2 \{c_i(\theta) - c_i(\theta_0)\}^2 \\
&\qquad \times \{c_i(\theta) - c_i(\theta_0)\}^{-2} \sigma(x_j)^2 \\
&= O\Big\{(nh)^{-2} n (n^{-1/2}h^{-1})^2 h\Big\} = O(n^{-2}h^{-3}),
\end{aligned}$$

uniformly in i. The desired result is immediate.

Step (v). We show that with X-probability 1 and for all $\xi > 0$,

$$\left| \sum_i{}' E(D_i)\delta_i \right| = O(n^{-3/10+\xi}). \tag{4.14}$$

We may deduce from (4.9), (4.10) and the fact that $\mu(x|\theta) - \mu(x|\theta_0) = O(n^{-1/2})$, that

$$\delta_i = -\eta\theta_{00}^T\{x_i - \mu(x_i|\theta_0)\} g'(\theta_0^T x_i) + O(n^{-1})$$

uniformly in i. Let $b_i(\theta)$, $c_i(\theta)$ be as in step (iii) of the proof and write γ for the density of $\theta_0^T X_A$. Then for all $\xi > 0$, $c_i(\theta_0) - \gamma(\theta_0^T x_i) = O(h^2 n^{\xi})$, $b_i(\theta_0) = O(h^2 n^{\xi})$, and

$$E(D_i) = b_i(\theta_0) c_i(\theta_0)^{-1} = b_i(\theta_0) \gamma(\theta_0^T x_i)^{-1} + O(h^4 n^{\xi}) = O(h^2 n^{\xi}),$$

uniformly in i. Hence,

$$\sum_i{}' E(D_i)\delta_i = -\eta t + O(n^{-(3/10)+\xi}), \tag{4.15}$$

where

$$t = \sum_i{}' \theta_{00}^T\{x_i - \mu(x_i|\theta_0)\} b_i(\theta_0) g'(\theta_0^T x_i) \gamma(\theta_0^T x_i)^{-1}.$$

Now, t denotes the observed value of

$$T = \sum_i{}' \sum_{j \neq i}{}' a(X_i, X_j),$$

where

$$a(X_i, X_j) = (nh)^{-1}\theta_{00}^T\{X_i - \mu(X_i|\theta_0)\}\{g(\theta_0^T X_j) - g(\theta_0^T X_i)\}$$

$$\times g'(\theta_0^T X_i)\gamma(\theta_0^T X_i)^{-1} K_h\{\theta_0^T(X_i - X_j)\}.$$

Note that $E\{a(X_i, X_j)I(X_i \in A)|\theta_0^T X_i, X_j\} = 0$, whence $E(T) = 0$. Similarly,

$$E\{a(X_i, X_j)a(X_k, X_l)I(X_i, X_k \in A)\} = 0$$

if $i \neq j$, $k \neq l$, $i \neq k$, and $(i, j) \neq (l, k)$. Therefore,

$$E(T^2) = O\Bigg[\Bigg|\sum_j \sum_l \sum_{i \neq j, l} E\{a(X_i, X_j)a(X_i, X_l)I(X_i \in A)\}\Bigg|$$

$$+ \Bigg|\sum \sum_{i \neq j} E\{a(X_i, X_j)a(X_j, X_i)I(X_i, X_j \in A)\}\Bigg|\Bigg]$$

$$= O\{(nh)^{-2}n(nh)^2h^4 + (nh)^{-2}n^2h^3\} = O(nh^4).$$

An argument based on the Borel–Cantelli lemma may now be used to prove that with X-probability 1, for all $\xi > 0$, $T = O(n^{1/2+\xi}h^2)$. Hence, $t = O(n^{1/2+\xi}h^2)$. Substituting into (4.15), we deduce (4.14).

Step (vi). Here we show that with X-probability 1 and for all $\xi > 0$,

$$\mathrm{Var}\Big(\sum_i{}' D_i\delta_i\Big) = O(n^{-4/5+\xi}).$$

Note that

$$\mathrm{Var}\Big(\sum_i{}' D_i\delta_i\Big) = (nh)^{-2}\sum_j{}' u_j^2\sigma(x_j)^2,$$

where

$$u_j = \sum_{i \neq j}{}' \delta_i c_i(\theta_0)^{-1} K_h\{\theta_0^T(x_i - x_j)\}.$$

As in the previous step, we may Taylor-expand δ_i and prove that with X-probability 1, for all $\xi > 0$, and uniformly in $1 \leq j \leq n$, $u_j = -\eta v_j + O(n^{1/2+\xi}h^3)$, where

$$v_j = \sum_{i \neq j}{}' \theta_{00}^T\{x_i - \mu(x_i|\theta_0)\}g'(\theta_0^T x_i)\gamma(\theta_0^T x_i)^{-1} K_h\{\theta_0^T(x_i - x_j)\}.$$

Therefore,

$$\mathrm{Var}\Big(\sum_i{}' D_i\delta_i\Big) \le 2(nh)^{-2}\eta^2\sum_j{}' v_j^2\sigma(x_j)^2 + O(n^{\xi}h^4).$$

Now, v_j equals the observed value of $V_j = \Sigma'_{i\ne j} b(X_i, X_j)$, where

$$b(X_i, X_j) = \theta_{00}^T\{X_i - \mu(X_i|\theta_0)\}g'(\theta_0^T X_i)\gamma(\theta_0^T X_i)^{-1}K_h\{\theta_0^T(X_i - X_j)\}.$$

Methods similar to those in the previous step may be used to prove that $E(V_j) = 0$ and $E(V_j^2) = O(nh)$, whence $\Sigma V_j^2 = O_p(n^2h)$. By an argument based on the Borel–Cantelli lemma, $\Sigma v_j^2 = O(n^{2+\xi}h)$ for all $\xi > 0$, with X-probability 1. Hence,

$$\mathrm{Var}\Big(\sum_i{}' O_i\delta_i\Big) = O\{(nh)^{-2}\eta^2 n^{2+\xi}h + n^{\xi}h^4\} = O(n^{\xi}h^4),$$

as required.

Step (vii). We show that with X-probability 1 and for all $\xi > 0$,

$$E\Big(\sum_i{}' D_i\varepsilon_i\Big)^2 = O(n^{1/5+\xi}). \tag{4.16}$$

Note that $E(D_i) = O(h^2n^{\xi})$ uniformly in i, for all $\xi > 0$. Hence,

$$E\Big\{\sum_i{}' E(D_i)\varepsilon_i\Big\}^2 = \sum_i{}' (ED_i)^2\sigma(x_i)^2 = O(n^{1+\xi}h^4) = O(n^{1/5+\xi}) \tag{4.17}$$

for all $\xi > 0$. Define $s_{ij} = E\{(D_i - ED_i)\varepsilon_i(D_j - ED_j)\varepsilon_j\}$. Then $s_{ii} = \mathrm{Var}(D_i)\sigma(x_i)^2$, and for $i \ne j$,

$$s_{ij} = K_h\{\theta_0^T(x_i - x_j)\}K_h\{\theta_0^T(x_j - x_i)\}\Big[\sum_{k\ne i} K_h\{\theta_0^T(x_i - x_k)\}\Big]^{-1}$$

$$\times\Big[\sum_{k\ne j} K_h\{\theta_0^T(x_j - x_k)\}\Big]^{-1}\sigma(x_i)^2\sigma(x_j)^2.$$

Therefore,

$$E\Big\{\sum_i{}' (D_i - ED_i)\varepsilon_i\Big\}^2 = \sum_i{}' s_{ii} + \sum_i{}'\sum_{i\ne j}{}' s_{ij} \tag{4.18}$$

$$= O\{nh^4 + n\cdot nh(nh)^{-2}\} = O(n^{1/5}).$$

The desired result (4.16) follows from (4.17) and (4.18).

Step (viii). We show that with X-probability 1 and for all $\xi > 0$,

$$E\Big(\sum_i{}' \Delta_i\varepsilon_i\Big)^2 = O(n^{-2/5+\xi}). \tag{4.19}$$

Note that $E(\Delta_i) = O(n^{-7/10+\xi})$ uniformly in i; see step (iii). Therefore,

$$E\Big\{\sum_i{}' E(\Delta_i)\varepsilon_i\Big\}^2 = \sum_i (E\Delta_i)^2\sigma(x_i)^2 = O(n^{-2/5+\xi}). \tag{4.20}$$

Furthermore, much as in the argument leading to (4.18),

$$\begin{aligned} E\Big\{\sum_i{}'(\Delta_i - E\Delta_i)\varepsilon_i\Big\}^2 &= O\Big\{n(n^{-2}h^{-3}) + n\cdot nh(nh)^{-2}(n^{-1/2}h^{-1})^2\Big\} \\ &= O(n^{-2/5}). \end{aligned} \tag{4.21}$$

The claimed result (4.19) is a consequence of (4.20) and (4.21).

Step (ix). Define

$$W = \sum_i{}'\{x_i - \mu(x_i|\theta_0)\}\{x_i - \mu(x_i|\theta_0)\}^T g'(\theta_0^T x_i)^2.$$

We prove that

$$\begin{aligned} \tilde{S}(\theta) &= \sum_i{}'\varepsilon_i^2 - V^T W^{-1} V \\ &\quad + n\big(\theta - \theta_0 - n^{-1}W_0^{-1}V\big)^T W_0\big(\theta - \theta_0 - n^{-1}W_0^{-1}V\big) + o_p(1). \end{aligned} \tag{4.22}$$

By (4.9) and (4.10),

$$g(\theta_0^T x_i) - g(\theta^T x_i|\theta) = \eta\theta_{00}^T\{\mu(x_i|\theta_0) - x_i\}g'(\theta_0^T x_i) + O(n^{-1}),$$

whence

$$\begin{aligned} \tilde{S} &= \sum_i{}'\{\varepsilon_i + g(\theta_0^T x_i) - g(\theta^T x_i|\theta)\}^2 \\ &= \sum_i{}'\varepsilon_i^2 - 2\eta\theta_{00}^T V + \eta^2\theta_{00}^T W\theta_{00} + o_p(1) \\ &= \sum_i{}'\varepsilon_i^2 - nZ'Z + n\big(W_0^{1/2}\eta\theta_{00} - n^{-1/2}\sigma Z\big)^T\big(W_0^{1/2}\eta\theta_{00} - n^{-1/2}\sigma Z\big) \\ &\quad + o_p(1), \end{aligned}$$

where Z is an asymptotically normal $N(0, I)$ random p-vector such that $V = n^{1/2}\sigma W_0^{1/2}Z$. The last line, which follows from the previous one on "completing the squares," implies (4.22).

REFERENCES

AZZALINI, A., BOWMAN, A. and HÄRDLE, W. (1989). On the use of nonparametric regression for model checking. *Biometrika* **76** 1–12.

BRILLINGER, D. R. (1983). A generalized linear model with "Gaussian" regressor variables. In *A Festschrift for Erich L. Lehmann* (P. J. Bickel, K. A. Doksum and J. L. Hodges, eds.) 97–114. Wadsworth, Belmont, CA.

COSLETT, S. R. (1987). Efficiency bounds for distribution-free estimators of the binary choice and the censored regression models. *Econometrica* **55** 559–585.

HALL, P. (1989). On projection pursuit regression. *Ann. Statist.* **17** 573–588.
HALL, P. and HEYDE, C. C. (1980). *Martingale Limit Theory and Its Applications*. Academic, New York.
HÄRDLE, W. (1990). *Applied Nonparametric Regression*. Cambridge Univ. Press.
HÄRDLE, W. (1991). *Smoothing Techniques with Implementation in S*. Springer, Berlin.
HÄRDLE, W., HALL, P. and MARRON, J. S. (1988). How far are automatically chosen regression smoothing parameters from their optimum? (with discussion). *J. Amer. Statist. Assoc.* **83** 86–99.
HÄRDLE, W. and STOKER, T. (1989). Investigating smooth multiple regression by the method of average derivatives. *J. Amer. Statist. Assoc.* **84** 986–995.
ICHIMURA, H. (1987). Estimation of single index models. Ph.D. dissertation, Dept. Economics, MIT.
ICHIMURA, H. (1990). Semiparametric weighted least squares estimation of single-index models. Unpublished manuscript.
MCCULLAGH, P. and NELDER, J. A. (1983). *Generalized Linear Models*. Chapman and Hall, London.

WOLFGANG HÄRDLE
INSTITUTE FÜR STATISTIK
UND ÖKONOMETRIE
FB WIRTSCHAFTSWISSENSCHAFTEN
HUMBOLDT-UNIVERSITÄT ZU BERLIN
0-1020 BERLIN
GERMANY

PETER HALL
DEPARTMENT OF MATHEMATICS
AUSTRALIAN NATIONAL UNIVERSITY
CANBERRA ACT 2601
AUSTRALIA

HIDEHIKO ICHIMURA
DEPARTMENT OF ECONOMICS
UNIVERSITY OF MINNESOTA
MINNEAPOLIS, MINNESOTA 55455

Journal of Econometrics, 58, 31-48.

How sensitive are average derivatives?*

Wolfgang Härdle
Tilburg University, 5000 LE Tilburg, The Netherlands
Université Catholique de Louvain, B-1348 Louvain-la-Neuve, Belgium

A.B. Tsybakov
Tilburg University, 5000 LE Tilburg, The Netherlands

Average derivatives are the mean slopes of regression functions. In practice they are estimated via a nonparametric smoothing technique. Every smoothing method needs a calibration parameter that determines the finite sample performance. In this paper we use the kernel estimation method and develop a formula for the bandwidth that describes the sensitivity of the average derivative estimator. One can determine an optimal smoothing parameter from this formula which tries out to undersmooth the density of the regression variable.

1. Average derivatives in discrete choice analysis

The average derivative is the mean of the slope of a regression function. In a regression setting $Y = m(X) + \varepsilon$ with regression curve $m: \boldsymbol{R}^d \to \boldsymbol{R}$, the average derivative is the mean gradient $\mathrm{E}_X(m'(X))$, or, more generally, the weighted mean gradient

$$\delta = \mathrm{E}_X(m'(X)w(X)), \tag{1.1}$$

where $m'(x)$ is the gradient

$$m'(x) = \left(\frac{\partial m}{\partial x_1}, \ldots, \frac{\partial m}{\partial x_d}\right) \in R^d,$$

$x_1, \ldots, x_d$ are components of the vector x, $w(x)$ is some weight function, and E_X is the expectation with respect to the (marginal) X-distribution.

Correspondence to: Wolfgang Härdle, Institut für Statistik und Ökonometrie, FB Wirtschaftswissenschaften, Humboldt-Universität zu Berlin, D-1020 Berlin, Germany.

*Work of the second author was financially supported by the Department of Econometrics, Tilburg University, The Netherlands.

0304-4076/93/$06.00

The average derivative δ is interesting in the context of discrete choice analysis, where in the case of binary choice we want to infer on the function

$$\mathrm{P}(Y = 1 \mid X = x) = m(x),$$

from observations $\{(X_i, Y_i)\}_{i=1}^n$, $X_i \in \boldsymbol{R}^d$, $Y_i \in \{0, 1\}$. A pure nonparametric approach to estimation of $m(x)$ is possible [see, for example, the recent monographs by Müller (1988), Eubank (1988), Wahba (1990), and Härdle (1990)]. It is well-known though that this approach is not costless: the precision of the estimator is exponentially decreasing as the dimension d increases. In order to avoid this difficulty one could of course fall back into pure parametric models for $m(x)$.

One such model would be

$$m(x) = G(x^T \beta), \tag{1.2}$$

where G, the link function, is of known form, e.g., $G = \Phi$ would postulate a Probit model.

A model comprising the advantages and simplicity of (1.2) and the flexibility of a nonparametric smoothing approach is a single-index model,

$$m(x) = g(x^T \beta), \tag{1.3}$$

with an unknown link function g and index $x^T \beta$.

It is well-known that β in (1.3) can only be identified up to scale [see Härdle and Stoker (1989)]: the (weighted) average derivative (ADE) for this model is

$$\delta = \mathrm{E}_X \left[\frac{\mathrm{d}g}{\mathrm{d}(X^T \beta)} w(X) \right] \beta = \gamma_\beta \cdot \beta, \tag{1.4}$$

so we see that we can estimate β (up to scale) if we know how to estimate δ and if γ_β is different from zero. A simple example for (1.4) is a linear link function $g(\cdot)$: then the coefficients β are multiplied by the slope of $g(\cdot)$ times $\mathrm{E}_X(w(X))$. For general, nonlinear $g(\cdot)$, as in binary choice models, the β coefficients are multiplied by the average slope

$$\gamma_\beta = \mathrm{E}_X \left[\frac{\mathrm{d}g}{\mathrm{d}(X^T \beta)} w(X) \right].$$

We use kernel estimators for the average derivative δ since they are straightforward to implement and easy to understand on an intuitive level. Other possibilities include splines and orthogonal series, but to our knowledge these techniques have not been employed to estimate average derivatives. The main point in this paper is about the selection of the bandwidth, the kernel smoothing parameter, for the d-dimensional case. The one-dimensional case with a focus on estimation of income effects is treated in Härdle, Hart, Marron, and Tsybakov (1991). From an asymptotical viewpoint the choice of bandwidth does not affect the behavior of ADE estimators. It influences only the higher-order terms of asymptotic expansions for mean squared error, not the main term which is of order $O(1/n)$, where n is the number of observations. In practice though, the choice of the smoothing parameter is an important issue as has been pointed out by Hsieh and Manski (1987, p. 551).

In this paper we consider the special choice of weight function: $w(x) = f(x)$, where $f(x)$ is the marginal density of X [cf. Powell, Stock, and Stoker (1989)]. This is motivated by several reasons. First, under such choice of w we avoid the random denominator appearing if $w(x) \equiv 1$ [in fact, for $w(x) \equiv 1$ the ADE estimators contain the density estimator in denominator; see Härdle and Stoker (1989) for details]. Because of the random denominator the necessary asymptotic expansions hold under somewhat restrictive assumptions on the underlying density f [Härdle and Stoker (1989), Härdle, Hart, Marron, and Tsybakov (1991)]. Next, for the multi-dimensional case the $O(1/n)$ rate of the mean-squared error is not attained unless the oscillating higher-order kernels are implemented. This causes a difficulty in treating the case of $w(x) \equiv 1$: the ADE estimator is not well-defined and it requires some truncation [Härdle and Stoker (1989)]. The choice of truncation threshold appears to be crucial in this context. This creates an additional problem which could be easily eliminated if $w(x) = f(x)$.

In section 2 we quantify the sensitivity of ADE via a second-order expansion of mean squared error of a kernel estimator for δ. Section 3 is devoted to the proof of our main theorem. In the appendix we prove some lemmas.

2. The sensitivity of ADE

Assume that independent pairs (X_i, Y_i), $i = 1, \ldots, n$, of random variables, $X_i \in R^d$, $Y_i \in R^1$, are observed and that they have the same distribution as (X, Y), $X \in R^d$, $Y \in R^1$.

Let the regression function $m(x) = E(Y|X = x)$ exist and let X have the density $f(x)$ with respect to Lebesgue measure in R^d. Suppose, moreover, that the regression function m and the density f are continuously differentiable and that $f(x)$ *vanishes outside a compact set*, the support of X.

Using partial integration (over the support of X) we get

$$\begin{aligned}\delta &= \int m'(x) f^2(x)\,dx \\ &= -2\int m(x) f'(x) f(x)\,dx \\ &= -2E(Yf'(X)),\end{aligned} \tag{2.1}$$

where $f'(x) = (\partial f/\partial x_1, \ldots, \partial f/\partial x_d)$ and the expectation is now taken over the joint distribution of (X, Y).

If we knew the marginal density f we could estimate δ by means of the sum $-(2/n)\sum_{i=1}^{n} Y_i f'(X_i)$ which is obtained if one substitutes the expectation in (2.1) by the empirical average.

In our approach we do not know the density function. We shall estimate it from the data via the kernel method. The marginal density $f(\cdot)$ is estimated by

$$\hat{f}_h(x) = n^{-1} \sum_{i=1}^{n} \mathcal{K}_h(x - X_i), \tag{2.2}$$

where $\mathcal{K}_h(u) = h^{-d}\mathcal{K}(u_1/h, \ldots, u_d/h)$ for a multivariate kernel function,

$$\mathcal{K}(u_1, \ldots, u_d) = \prod_{j=1}^{d} K(u_j), \quad u = (u_1, \ldots, u_d) \in R^d, \tag{2.3}$$

based on a one-dimensional kernel K. The scaling of $\mathcal{K}$ is through $h > 0$, the bandwidth, or smoothing parameter.

The gradient $f'(x)$ is estimated by

$$\hat{f}'_h(x) = \frac{1}{n}\sum_{i=1}^{n} \mathcal{K}'_h(x - X_i), \tag{2.4}$$

where

$$\mathcal{K}'_h(u) = h^{-d-1} K'\left(\frac{u_j}{h}\right) \prod_{k \neq j} K\left(\frac{u_j}{h}\right),$$

and K' denotes the derivative of one-dimensional kernel K.

Using (2.4) we can construct an estimate of the average derivative

$$\hat{\delta}_n = -\frac{2}{n}\sum_{i=1}^{n} Y_i \hat{f}'_h(X_i). \tag{2.5}$$

We study the asymptotic mean squared error of $\hat{\delta}_n$ under the following assumptions:

(A1) The kernel K is bounded, continuously differentiable, symmetric with support $[-1,1]$; $K'(0)=0$.

(A2) $\int K(u)\,du = 1$, and there exists a positive integer $k \geq 2$ such that $\int u^j K(u) \times du = 0$, $j = 1, \ldots, k-1$, $\int u^k K(u)\,du = d_K \neq 0$.

(A3) The marginal density $f(x)$ of X is compactly supported and has continuous partial derivatives up to the order $k+1$ on R^d.

(A4) The regression function $m(x)$ has continuous partial derivatives up to the order $k+1$ on R^d.

(A5) The conditional variance $\sigma^2(x) = \mathrm{var}(Y|X=x)$ is bounded on the support of f.

(A6) $h = h_n \to 0$, and $n^2 h_n^{d+2} \to \infty$ as $n \to \infty$.

Later $|\cdot|$ denotes Euclidean norm when applied to vectors.

Theorem. Under the assumptions (A1)–(A6),

$$\mathrm{E}(|\hat{\delta}_n - \delta|^2) = Q_1 n^{-1} + Q_2 n^{-2} h_n^{-d-2} + Q_3 h_n^{2k} + \mathrm{O}\left(\frac{h_n^k}{n} + \frac{1}{n^2}\right) + \mathrm{o}\left(\frac{1}{n^2 h_n^{d+2}} + h_n^{2k}\right), \qquad n \to \infty,$$

where

$$Q_1 = 4[\mathrm{E}(|f(X)m'(X)|^2) - |\mathrm{E}(f(X)m'(X))|^2 + \mathrm{E}(\sigma^2(X)|f'(X)|^2)],$$

$$Q_2 = 4C_K \int \sigma^2(x) f^2(x)\,dx,$$

$$Q_3 = 4\left|\int S_K(x) f(x) m(x)\,dx\right|^2,$$

and

$$C_K = \int |\mathscr{K}'(u)|^2\,du = d\int (K'(u))^2\,du \left(\int K^2(u)\,du\right)^{d-1},$$

$$S_K(x) = d_K \frac{(-1)^k}{k!} \sum_{j=1}^{d} \begin{pmatrix} \partial^{k+1} f(x)/\partial x_1 \partial x_j^k \\ \vdots \\ \partial^{k+1} f(x)/\partial x_d \partial x_j^k \end{pmatrix}.$$

From the Theorem we see that the bandwidth h_n minimizing $E(|\hat{\delta}_n - \delta|^2)$ is given by

$$h_n^* = h_0 n^{-2/(2k+d+2)},$$

where

$$h_0 = \left(\frac{Q_2(d+2)}{2kQ_3}\right)^{1/(2k+d+2)}.$$

For $h_n = h_n^*$, we have

$$E(|\hat{\delta}_n - \delta|^2) = Q_1 n^{-1} + Cn^{-4k/(2k+d+2)} + o(n^{-4k/(2k+d+2)}) + O\left(\frac{1}{n^2}\right), \qquad n \to \infty, \tag{2.6}$$

where

$$C = \left[\left(\frac{2k}{d+2}\right)^{(d+2)/(2k+d+2)} + \left(\frac{d+2}{2k}\right)^{2k/(2k+d+2)}\right] \times Q_2^{2k/(2k+d+2)} Q_3^{(d+2)/(2k+d+2)}.$$

Optimization of k. This in fact is reasonable if one believes that f and m are infinitely many times continuously differentiable. It follows from (2.6) that the best rate for mean squared error equals n^{-1} and it is attained if $k > (d+2)/2$. For example, in one-dimensional case ($d = 1$) it suffices to take $k = 2$. Then the second term in (2.6) equals $Cn^{-8/7}$, and h_n^* is proportional to $n^{-2/7}$ [cf. Härdle, Hart, Marron, and Tsybakov (1991)].

Assumptions (A1) and (A2) entail that the order k of the kernel should be necessarily even. Thus, the condition for choosing k that guarantees the best rate of convergence becomes:

k is the minimal even number such that $k > (d+2)/2$.

Optimization of K. The factor C depends on the kernel K. Optimizing this factor in K leads to the minimization problem (in view of the definition of Q_2 and Q_3):

$$\min_{K \in \mathcal{W}_k} \left(\int u^k K(u)\,du\right)^{d+2} \left(\int (K'(u))^2\,du\right)^{2k} \left(\int K^2(u)\,du\right)^{2(d-1)},$$

where $\mathscr{U}_k$ is the class of kernels satisfying (A1) and (A2). For $d = 1$, this problem was solved by Mammitzsch (1990) who showed that the optimal K is the quartic kernel

$$K(u) = \tfrac{15}{16}(1 - u^2)^2 \, I(|u| \le 1),$$

where $I(\cdot)$ denotes the indicator function. If $d \ge 2$, then $k \ge 4$, and the optimal K is, clearly, an oscillating kernel taking positive and negative values.

3. Proof of the Theorem

Denote

$$\delta_n^* = \frac{\hat{\delta}_n}{2} = -\frac{1}{n}\sum_{i=1}^{n} Y_i \hat{f}_h'(X_i), \qquad \delta^* = \frac{\delta}{2} = -\int m(x) f'(x) f(x)\, dx.$$

Clearly,

$$\mathrm{E}(|\hat{\delta}_n - \delta|^2) = 4\mathrm{E}(|\delta_n^* - \delta^*|^2). \tag{3.1}$$

Write the estimator δ_n^* as

$$\delta_n^* = \frac{1}{n}\sum_{i=1}^{n} (m(X_i) + \varepsilon_i)\hat{f}_h'(X_i),$$

where $\varepsilon_i = Y_i - m(X_i)$. Since $\mathrm{E}(\varepsilon_i | X_i) = 0$, we have

$$\begin{aligned}
&\mathrm{E}(|\delta_n^* - \delta^*|^2) \\
&= \mathrm{E}\left(\left|\frac{1}{n}\sum_{i=1}^{n} \varepsilon_i \hat{f}_h'(X_i)\right|^2\right) + \mathrm{E}\left(\left|\frac{1}{n}\sum_{i=1}^{n} m(X_i)\hat{f}_h'(X_i) - \delta^*\right|^2\right) \\
&= \mathrm{E}\left(\left|\frac{1}{n}\sum_{i=1}^{n} \varepsilon_i \hat{f}_h'(X_i)\right|^2\right) + \mathrm{E}\left(\left|\frac{1}{n}\sum_{i=1}^{n} [\zeta_i - \mathrm{E}(\zeta_i)]\right|^2\right) \qquad (3.2) \\
&\quad + \left|\frac{1}{n}\sum_{i=1}^{n} \mathrm{E}(\zeta_i) - \delta^*\right|^2,
\end{aligned}$$

where $\zeta_i = m(X_i)\hat{f}_h'(X_i)$.

It follows from (A1) that $\mathscr{K}'_h(0) = 0$, and thus

$$\hat{f}'_h(X_i) = \frac{1}{n} \sum_{\substack{j=1 \\ i \neq j}}^{n} \mathscr{K}'_h(X_i - X_j). \tag{3.3}$$

Thus,

$$\mathrm{E}(\zeta_i) = \mathrm{E}(\zeta_1) = \mathrm{E}(m(X_1)\hat{f}'_h(X_1)) = \frac{n-1}{n}\tilde{q}, \tag{3.4}$$

where $\tilde{q} = \mathrm{E}(m(X_1)\mathscr{K}'_h(X_1 - X_2))$. Here we used (2.4) and the fact that $\mathscr{K}'_h(0) = 0$ which follows from (A1). Now, (3.2) and (3.4) entail

$$\mathrm{E}(|\delta_n^* - \delta^*|^2) = V_1 + V_2 + V_3, \tag{3.5}$$

where

$$V_1 = \mathrm{E}\left(\left|\frac{1}{n}\sum_{i=1}^{n} \varepsilon_i \hat{f}'_h(X_i)\right|^2\right),$$

$$V_2 = \mathrm{E}\left(\left|\frac{1}{n}\sum_{i=1}^{n} (\zeta_i - \mathrm{E}(\zeta_1))\right|^2\right),$$

$$V_3 = |\mathrm{E}(\zeta_1) - \delta^*|^2.$$

Let us evaluate the terms V_1, V_2, and V_3.

Clearly,

$$\mathrm{E}(\varepsilon_i \varepsilon_l | X_1, \ldots, X_n) = \begin{cases} \sigma^2(X_i), & i = l, \\ 0, & i \neq l. \end{cases} \tag{3.6}$$

From (3.3) and (3.6) we get

$$V_1 = \mathrm{E}\left(\frac{1}{n^2}\sum_{i,l=1}^{n} \varepsilon_i \varepsilon_l (\hat{f}'_h(X_i), \hat{f}'_h(X_l))\right)$$

$$= \frac{1}{n^2}\sum_{i=1}^{n} \mathrm{E}\left(\sigma^2(X_i)\left|\frac{1}{n}\sum_{\substack{j=1 \\ i \neq j}}^{n} \mathscr{K}'_h(X_i - X_j)\right|^2\right) \tag{3.7}$$

$$= \frac{1}{n^3} \int \sigma^2(X_1)\, \mathrm{E}\left(\left| \sum_{j=2}^{n} \mathscr{K}'_h(X_1 - X_j) \right|^2 \right) f(X_1)\, \mathrm{d}X_1$$

$$= \frac{1}{n^3} \int \sigma^2(x)\, \mathrm{E}\left(\sum_{j,s=2}^{n} (\mathscr{K}'_h(x - X_j), \mathscr{K}'_h(x - X_s)) \right) f(x)\, \mathrm{d}x.$$

Here and later $(\cdot, \cdot)$ denotes the scalar product. We have

$$\mathrm{E}\left(\sum_{j,s=2}^{n} (\mathscr{K}'_h(x - X_j), \mathscr{K}'_h(x - X_s)) \right)$$

$$= \sum_{j=2}^{n} \mathrm{E}(|\mathscr{K}'_h(x - X_j)|^2) + \sum_{\substack{j,s=2 \\ j \neq s}}^{n} |\mathrm{E}(\mathscr{K}'_h(x - X_j))|^2 \tag{3.8}$$

$$= (n-1)\, \mathrm{E}(|\mathscr{K}'_h(x - X_1)|^2)$$

$$+ (n-1)(n-2) |\mathrm{E}(\mathscr{K}'_h(x - X_1))|^2.$$

It follows from Lemmas 1 and 2 and (3.8) that

$$\mathrm{E}\left(\sum_{j,s=2}^{n} (\mathscr{K}'_h(x - X_j), \mathscr{K}'_h(x - X_s) \right)$$

$$= (n-1)(C_K f(x) h^{-d-2} + \beta_3(h, x))$$

$$+ (n-1)(n-2) |f'(x) + h^k S_K(x) + \beta_1(h, x)|^2.$$

Hence

$$V_1 = \frac{1}{n} \int \sigma^2(x) |f'(x)|^2 f(x)\, \mathrm{d}x + \frac{1}{n^2 h^{d+2}} C_K \int \sigma^2(x) f^2(x)\, \mathrm{d}x$$

$$+ \mathrm{o}\left(\frac{1}{n^2 h^{d+2}} \right) + \mathrm{O}\left(\frac{h^k}{n} + \frac{1}{n^2} \right), \qquad n \to \infty, \tag{3.9}$$

where we used the properties of β_1, β_3 and the fact that by (A3) the function $|S_K(x)|$ is uniformly bounded on the support of f.

Next,

$$V_2 = \frac{1}{n^2} \sum_{i,j=1}^{n} \mathrm{E}((\zeta_i, \zeta_j)) - |\mathrm{E}(\zeta_1)|^2 = V_{21} + V_{22} - |\mathrm{E}(\zeta_1)|^2, \tag{3.10}$$

where

$$V_{21} = \frac{1}{n^2} \sum_{i=1}^{n} \mathrm{E}(|\zeta_i|^2), \qquad V_{22} = \frac{1}{n^2} \sum_{\substack{i,j=1 \\ i \neq j}}^{n} \mathrm{E}((\zeta_i, \zeta_j)).$$

We have

$$\mathrm{E}(|\zeta_i|^2) = \frac{1}{n^2} \mathrm{E}\left(m^2(X_i) \sum_{\substack{l,s=1 \\ l,s \neq i}}^{n} (\mathscr{K}'_h(X_i - X_l), \mathscr{K}'_h(X_i - X_s)) \right).$$

Hence

$$V_{21} = \frac{1}{n^3} \mathrm{E}\left(m^2(X_1) \sum_{l,s=2}^{n} (\mathscr{K}'_h(X_1 - X_l), \mathscr{K}'_h(X_1 - X_2)) \right).$$

Using the same argument as in (3.7)–(3.9), we find

$$V_{21} = \frac{1}{n} \int m^2(x) |f'(x)|^2 f(x)\,\mathrm{d}x + \frac{1}{n^2 h^{d+2}} C_K \int m^2(x) f^2(x)\,\mathrm{d}x$$

$$+ \mathrm{o}\left(\frac{1}{n^2 h^{d+2}} \right) + \mathrm{O}\left(\frac{h^k}{n} + \frac{1}{n^2} \right), \qquad n \to \infty. \tag{3.11}$$

Consider the term V_{22} now. Applying (3.3) we obtain

$$V_{22} = \frac{n-1}{n} \mathrm{E}((\zeta_1, \zeta_2)) = \frac{n-1}{n^3}$$

$$\times \left\{ \mathrm{E}\left[m(X_1) m(X_2) \times \sum_{\substack{l,s=1 \\ l \neq 1 \\ s \neq 2}}^{n} (\mathscr{K}'_h(X_1 - X_l), \mathscr{K}'_h(X_2 - X_s)) \right] \right\}$$

$$= \frac{n-1}{n^3} \sum_{\substack{l,s=1 \\ l \neq 1 \\ s \neq 2}}^{n} \Delta_{ls}, \tag{3.12}$$

where

$$\Delta_{ls} = \mathrm{E}[m(X_1)m(X_2)(\mathscr{K}'_h(X_1 - X_l), \mathscr{K}'_h(X_2 - X_s))].$$

Let us treat Δ_{ls} separately in the following four cases:

(i) $s = l$,

(ii) $s \neq l, l \neq 2, s \neq 1$,

(iii) $s \neq l$ and either $l = 2, s \neq 1$ or $l \neq 2, s = 1$,

(iv) $s \neq l, l = 2, s = 1$.

In case (i),

$$\begin{aligned}\Delta_{ls} = \Delta_{ll} &= \mathrm{E}[m(X_1)m(X_2)(\mathscr{K}'_h(X_1 - X_3), \mathscr{K}'_h(X_2 - X_3))] \\ &= \mathrm{E}_{X_3}[|\mathrm{E}_{X_1}(m(X_1)\mathscr{K}'_h(X_1 - X_3))|^2] \stackrel{\text{def}}{=} B_1.\end{aligned} \tag{3.13}$$

In case (ii),

$$\begin{aligned}\Delta_{ls} &= \mathrm{E}[m(X_1)m(X_2)(\mathscr{K}'_h(X_1 - X_l), \mathscr{K}'_h(X_1 - X_s))] \\ &= |\mathrm{E}(m(X_1)\mathscr{K}'_h(X_1 - X_l))|^2 = |\tilde{q}|^2.\end{aligned} \tag{3.14}$$

In case (iii),

$$\begin{aligned}\Delta_{ls} &= \mathrm{E}[m(X_1)m(X_2)(\mathscr{K}'_h(X_1 - X_2), \mathscr{K}'_h(X_2 - X_s))] \\ &= \mathrm{E}[(m(X_1)m(X_2)(\mathscr{K}'_h(X_1 - X_2)), \mathscr{K}'_h(X_2 - X_3))] \stackrel{\text{def}}{=} B_2.\end{aligned} \tag{3.15}$$

In case (iv),

$$\begin{aligned}\Delta_{ls} &= \mathrm{E}[m(X_1)m(X_2)(\mathscr{K}'_h(X_1 - X_2), \mathscr{K}'_h(X_2 - X_1))] \\ &= -\mathrm{E}[(m(X_1)m(X_2)|\mathscr{K}'_h(X_1 - X_2)|^2) \stackrel{\text{def}}{=} B_3.\end{aligned} \tag{3.16}$$

In (3.16) we used the fact that $\mathscr{K}'_h$ is antisymmetric:

$$\mathscr{K}'_h(-u) = -\mathscr{K}'_h(u).$$

It follows from (3.12)–(3.16) that

$$V_{22} = \frac{n-1}{n^3}\left[(n-2)B_1 + (n^2 - 5n + 6)|\tilde{q}|^2 + 2(n-2)B_2 + B_3\right]. \tag{3.17}$$

Now we use Lemma 4 in the appendix, which implies, together with (3.17), that

$$\begin{aligned} V_{22} = \frac{1}{n}\Bigg[& \int f^3(x)|m'(x)|^2\,dx - \int m^2(x)|f'(x)|^2 f(x)\,dx \Bigg] \\ & + |\tilde{q}|^2 + \left(1 - \frac{6}{n}\right) - \frac{C_K}{n^2 h^{d+2}} \int m^2(x) f^2(x)\,dx \\ & + O\left(\frac{h^k}{n} + \frac{1}{n^2}\right) + o\left(\frac{1}{n^2 h^{d+2}}\right), \qquad n \to \infty. \end{aligned} \tag{3.18}$$

From (3.4), (3.10), (3.11), and (3.18), we find

$$\begin{aligned} V_2 &= V_{21} + V_{22} - \left|\frac{n-1}{n}\tilde{q}\right|^2 \\ &= V_{21} + V_{22} - |\tilde{q}|^2\left(1 - \frac{2}{n} + \frac{1}{n^2}\right) \\ &= \frac{1}{n}\Bigg[\int f^3(x)|m'(x)|^2\,dx - 4|\tilde{q}|^2\Bigg] \\ &\quad + O\left(\frac{h^k}{n} + \frac{1}{n^2}\right) + o\left(\frac{1}{n^2 h^{d+2}}\right), \qquad n \to \infty. \end{aligned} \tag{3.19}$$

By substitution on (A.2) into (3.19), we obtain

$$\begin{aligned} V_2 = \frac{1}{n}\Bigg[& \int f^3(x)|m'(x)|^2\,dx - 4|\delta^*|^2\Bigg] \\ & + O\left(\frac{h^k}{n} + \frac{1}{n^2}\right) + o\left(\frac{1}{n^2 h^{d+2}}\right), \qquad n \to \infty. \end{aligned} \tag{3.20}$$

It remains to find the asymptotic expression for V_3.
By (A.2) we have

$$V_3 = \left| \frac{n-1}{n} \hat{q} - \delta^* \right|^2$$

$$= \left| \frac{q}{n} + h^k \int S_K(x) f(x)\, m(x)\, dx + \beta_2(h) \right|^2 \tag{3.21}$$

$$= h^{2k} \left| \int S_K(x) f(x)\, m(x)\, dx \right|^2$$

$$+ O\left(\frac{h^k}{n} + \frac{1}{n^2} \right) + o(h^{2k}), \qquad n \to \infty .$$

From (3.5), (3.9), (3.20), and (3.21) one gets

$$E(|\delta_n^* - \delta^*|^2)$$

$$= \frac{1}{n} \left[\int f^{3\infty}(x) |m'(x)|^2\, dx - 4|\delta^*|^2 + \int \sigma^2(x) |f'(x)|^2 f(x)\, dx \right]$$

$$+ \frac{1}{n^2 h^{d+2}} C_K \int \sigma^2(x) f^2(x)\, dx + h^{2k} \left| \int S_K(x) f(x)\, m(x)\, dx \right|^2$$

$$+ O\left(\frac{h^k}{n} + \frac{1}{n^2} \right) + o\left(\frac{1}{n^2 h^{d+2}} \right), \qquad n \to \infty .$$

This, in view of (3.1), proves the Theorem.

Appendix

Lemma 1. Let assumptions (A1)–(A4) be satisfied. Then

$$E(\mathscr{K}'_h(x - X_1)) = -f'(x) - h^k S_K(x) - \beta_1(h, x), \qquad \forall x \in R^d, \tag{A.1}$$

where $\sup_x|\beta_1(h,x)| = o(h^k)$ *as* $h \to 0$, *and*

$$\tilde{q} = \delta^* - h^k \int S_K(x) f(x) m(x)\,dx + \beta_2(h), \tag{A.2}$$

where $|\beta_2(h)| = o(h^k)$ *as* $h \to 0$.

Proof. By partial integration,

$$\begin{aligned} E(\mathcal{K}_h'(x - X_1)) &= \frac{1}{h^{d+1}} \int \mathcal{K}'\left(\frac{x-z}{h}\right) f(z)\,dz \\ &= -\frac{1}{h} \int \mathcal{K}'(u) f(x - uh)\,du \\ &= \frac{1}{h} \int (f(x-uh))' \mathcal{K}(u)\,du \\ &= -\int f'(x-uh)\mathcal{K}(u)\,du, \end{aligned} \tag{A.3}$$

where we used the fact that $\mathcal{K}$ and f are compactly supported. Assumption (A3) entails that the Taylor expansion is valid, and thus, uniformly in $u \in \operatorname{supp}\mathcal{K}$,

$$\left| f'(x-uh) - \sum_{\alpha:|\alpha|\le k} \frac{1}{\alpha!}(-1)^{|\alpha|} u^\alpha h^{|\alpha|} \begin{pmatrix} \partial^{|\alpha|+1} f(x)/\partial x_1 \partial x^\alpha \\ \vdots \\ \partial^{|\alpha|+1} f(x)/\partial x_d \partial x^\alpha \end{pmatrix} \right|$$

$$\le \beta_1(h,x), \tag{A.4}$$

where $\sup_x|\beta_1(h,x)| = o(h^k)$, $h \to 0$, and $\alpha = (\alpha, \ldots, \alpha_d)$ is multi-index, $\alpha_j \ge 0$, $|\alpha| = \alpha_1 + \cdots + \alpha_d$, $\alpha! = \alpha_1! \cdots \alpha_d!$, $u^\alpha = u_1^{\alpha_1} \cdot \cdots \cdot u_d^{\alpha_d}$ for $u = (u_1, \ldots, u_d) \in R^d$, and $\partial^{|\alpha|}/\partial x^\alpha = \partial^{|\alpha|}/\partial x_1^{\alpha_1} \cdot \cdots \cdot \partial x_d^{\alpha_d}$ for $x = (x_1, \ldots, x_d) \in R^d$.

It follows from (A2) that

$$\int u^\alpha \mathcal{K}(u)\,du = 0, \qquad 0 \le |\alpha| \le k-1, \tag{A.5}$$

$$\int u^\alpha \mathcal{K}(u)\,du = \begin{cases} 0, & |\alpha| = k,\ \operatorname{card}\{j:\ \alpha_j \ne 0\} > 1, \\ d_K, & |\alpha| = k,\ \operatorname{card}\{j:\ \alpha_j \ne 0\} = 1. \end{cases}$$

From (A.3)–(A.5) we get

$$\mathrm{E}(\mathscr{K}' h(x - X_1))$$

$$= -f'(x) - \frac{1}{k!} h^k (-1)^k \sum_{\alpha:|\alpha|=k} \begin{pmatrix} \partial^{|\alpha|+1} f(x)/\partial x_1 \partial x^\alpha \\ \vdots \\ \partial^{|\alpha|+1} f(x)/\partial x_d \partial x^\alpha \end{pmatrix}$$

$$\times \int u^\alpha \mathscr{K}(u)\,\mathrm{d}u - \beta_1(h, x) \tag{A.6}$$

$$= -f'(x) - h^k S_K(x) - \beta_1(h, x),$$

which proves (A.1). To get (A.2) note that by (A.1)

$$\tilde{q} = \mathrm{E}(m(X_1)\mathscr{K}'_h(X_1 - X_2))$$

$$= \mathrm{E}(m(X_1)\mathrm{E}_{X_2}(\mathscr{K}'_h(X_1 - X_2))$$

$$= \mathrm{E}(m(X_1)(-f'(X_1) - h^k S_K(X_1) - \beta_1(h, X_1)))$$

$$= \delta^* - h^k \int S_K(x) f(x)\, m(x)\,\mathrm{d}x - \int m(x) f(x)\, \beta_1(h, x)\,\mathrm{d}x.$$

Now, $\sup_x|\beta_1(h, x)| = o(h^k)$, and $\int m(x)f(x)\,\mathrm{d}x < \infty$ since m and f are bounded on the compact support of f. This proves (A.2).

Lemma 2. Let assumptions (A1)–(A4) be satisfied. Then

$$\mathrm{E}(|\mathscr{K}'_h(x - X_1)|^2) = C_K f(x) h^{-d-2} + \beta_3(h, x),$$

where $\sup_x|\beta_3(h, x)| = o(h^{-d-2})$, $h \to 0$.

Proof.
$$\mathrm{E}(|\mathscr{K}'_h(x - X_1)|^2) = \frac{1}{h^{2d+2}} \int \left|\mathscr{K}'\left(\frac{x - z}{h}\right)\right|^2 f(z)\,\mathrm{d}z$$

$$= \frac{1}{h^{d+2}} \int |\mathscr{K}'(u)|^2 f(x - uh)\,\mathrm{d}u.$$

It follows from (A3) that f is Lipschitz continuous on its support with some Lipschitz constant L_f. Thus,

$$\left| \frac{1}{h^{d+2}} \int |\mathcal{K}'(u)|^2 f(x-uh)\,du - \frac{1}{h^{d+2}} f(x)\, C_K \right| \le \frac{L_f}{h^{d+1}} \int |\mathcal{K}'(u)|^2 |u|\,du.$$

This proves the lemma.

Lemma 3. Let assumptions (A1)–(A4) be satisfied. Then

$$\mathrm{E}(m(X_1)\mathcal{K}'_h(x - X_1)) = -(m(x)f(x))' - h^k S_{1K}(x) - \beta_4(h, x),$$

where

$$S_{1K}(x) = d_K \frac{(-1)^k}{k!} \sum_{j=1}^{d} \begin{pmatrix} \partial^{k+1}(f(x)m(x))/\partial x_1 \partial x_j^k \\ \vdots \\ \partial^{k+1}(f(x)m(x))/\partial x_d \partial x_j^k \end{pmatrix}$$

and $\sup_x |\beta_4(h, x) = o(h^k)$, $h \to 0$.

Proof. The proof is similar to the proof of (A.1). In fact, instead of (A.3) we now have

$$\begin{aligned} \mathrm{E}(m(X_1)\mathcal{K}'_h(x - X_1)) &= \int (f(x-uh)\,m(x-uh))'\,\mathcal{K}(u)\,du \\ &= -\int [f'(x-uh)\,m(x-uh) \\ &\quad + f(x-uh)\,m'(x-uh)]\,\mathcal{K}(u)\,du. \end{aligned}$$

Lemma 4. Let assumptions (A1)–(A4) be satisfied. Then, as $h \to 0$,

$$\begin{aligned} B_1 &= \int |m'(x)|^2 f(x)\,dx + \int m^2(x)|f'(x)|^2 f(x)\,dx \\ &\quad + 2\int m(x)\,(m'(x), f'(x))\,f^2(x)\,dx + O(h^k), \end{aligned} \tag{A.7}$$

$$\begin{aligned} B_2 &= -\int m^2(x)|f'(x)|^2 f(x)\,dx \\ &\quad - \int m(x)\,(m'(x), f'(x))\,f^2(x)\,dx + O(h^k), \end{aligned} \tag{A.8}$$

$$B_3 = \frac{1}{h^{d+2}} \left(C_K \int m^2(x) f^2(x)\,dx + o(1) \right). \tag{A.9}$$

It follows from (A3) that f is Lipschitz continuous on its support with some Lipschitz constant L_f. Thus,

$$\left| \frac{1}{h^{d+2}} \int |\mathcal{K}'(u)|^2 f(x-uh)\,du - \frac{1}{h^{d+2}} f(x)\, C_K \right| \le \frac{L_f}{h^{d+1}} \int |\mathcal{K}'(u)|^2 |u|\,du.$$

This proves the lemma.

Lemma 3. Let assumptions (A1)–(A4) be satisfied. Then

$$\mathrm{E}(m(X_1)\mathcal{K}'_h(x - X_1)) = -(m(x)f(x))' - h^k S_{1K}(x) - \beta_4(h, x),$$

where

$$S_{1K}(x) = d_K \frac{(-1)^k}{k!} \sum_{j=1}^{d} \begin{pmatrix} \partial^{k+1}(f(x)m(x))/\partial x_1 \partial x_j^k \\ \vdots \\ \partial^{k+1}(f(x)m(x))/\partial x_d \partial x_j^k \end{pmatrix}$$

and $\sup_x |\beta_4(h, x) = o(h^k)$, $h \to 0$.

Proof. The proof is similar to the proof of (A.1). In fact, instead of (A.3) we now have

$$\begin{aligned} \mathrm{E}(m(X_1)\mathcal{K}'_h(x - X_1)) &= \int (f(x-uh)\,m(x-uh))'\,\mathcal{K}(u)\,du \\ &= -\int [f'(x-uh)\,m(x-uh) \\ &\quad + f(x-uh)\,m'(x-uh)]\,\mathcal{K}(u)\,du. \end{aligned}$$

Lemma 4. Let assumptions (A1)–(A4) be satisfied. Then, as $h \to 0$,

$$\begin{aligned} B_1 &= \int |m'(x)|^2 f(x)\,dx + \int m^2(x)|f'(x)|^2 f(x)\,dx \\ &\quad + 2\int m(x)\,(m'(x), f'(x))\,f^2(x)\,dx + O(h^k), \end{aligned} \tag{A.7}$$

$$\begin{aligned} B_2 &= -\int m^2(x)|f'(x)|^2 f(x)\,dx \\ &\quad - \int m(x)\,(m'(x), f'(x))\,f^2(x)\,dx + O(h^k), \end{aligned} \tag{A.8}$$

$$B_3 = \frac{1}{h^{d+2}} \left(C_K \int m^2(x) f^2(x)\,dx + o(1) \right). \tag{A.9}$$

References

Eubank, R., 1988, Spline smoothing and nonparametric regression (Marcel Dekker, New York, NY).

Härdle, W., 1990, Applied nonparametric regression, Econometric Society monograph series 19 (Cambridge University Press, Cambridge).

Härdle, W. and T. Stoker, 1989, Investigating smooth multiple regression by the method of average derivatives, Journal of the American Statistical Association 84, 986–995.

Härdle, W., J. Hart, J.S. Marron, and A.B. Tsybakov, 1991, Bandwith choice for average derivative estimation, Journal of the American Statistical Association 87, 218–226.

Hsieh, D.A. and C.F. Manski, 1987, Monte Carlo evidence on adaptive maximum likelihood estimation, Annals of Statistics 15, 541–551.

Mammitzsch, V., 1990, Asymptotically optimal kernels for average derivative estimation, in: Proceedings of the 10th Prague conference on statistical decision functions, information theory and random processes, Prague, Aug. 1990.

Müller, H.G., 1988, Nonparametric regression analysis of longitudinal data, Springer lecture notes in statistics 46 (Springer-Verlag, New York, NY).

Powell, J.L., J.H. Stock, and T.M. Stoker, Semiparametric estimation of index coefficients, Econometrica 57, 1403–1431.

Wahba, G., 1990, Spline models for observational data, CBMS–NSF regional conference series in applied mathematics, SIAM monograph series 59 (Society for Industrial and Applied Mathematics).

Computational Statistics and Data Analysis, 16, 11-18.

On the inconsistency of bootstrap distribution estimators *

Peter Hall
Australian National University, Canberra, ACT 2601, Australia

Wolfgang Härdle
CORE, Université Catholique de Louvain, B-1348 Louvain-la-Neuve, Belgium

Léopold Simar
Facultés Universitaires Saint-Louis, B-1000 Bruxelles, Belgium and CORE, Université Catholique de Louvain, B-1348 Louvain-la-Neuve, Belgium

Received May 1991
Revised January 1992

Abstract: We show that bootstrap distribution estimators of a ranked parameter value are consistent if and only if there are no ties for the rank in question. When inconsistency occurs, the bootstrap distribution estimator does not even converge in probability. This asymptotic result has important implications for small to moderate sample sizes, where poor distribution estimators can result when there are no ties but there are two or more closely spaced parameter values competing to the same rank. Several ways of alleviating the problem of inconsistency are suggested.

Keywords: Bootstrap; Consistency; Iterated bootstrap; Maximum; Rank; Tie

1. Introduction

Several authors have noted that bootstrap estimators of distributions can be inconsistent. For example, in the case of heavy-tailed sampling populations, Athreya (1987), Knight (1989), and Hall (1991) have studied the problem of

Correspondence to: W. Härdle, FB Wirtschaftswissenschaften Humboldt Universität zu Berlin, D-1020 Berlin, Germany.
* This work was supported in part by grant no 26 from the program 'Pôle d'attraction interuniversitaire – Deuxième phase' to CORE, Université Catholique de Louvain. The first author was partly financed by the Département de Mathématiques Appliquées, Université Catholique de Louvain.

consistent bootstrap estimation of a sample mean, and have shown that strong consistency is dependent on the existence of finite variance. For another paper on inconsistency of the bootstrap, see Mammen (1990). However, this is a rather pathological example, since it is often reasonable to assume finiteness of moments higher than the second. In the present paper we discuss an example of inconsistency with more serious implications. It is related to an example discussed by Beran and Srivastava (1985) in connection with ties among eigenvalues, and arises in estimating a ranked parameter value, as follows.

Let $\theta_1, \ldots, \theta_p$ denote unknown parameter values, and let $\hat{\theta}_1, \ldots, \hat{\theta}_p$ represent root-n consistent estimators, where n is sample size. Suppose we wish to estimate the rth largest θ_i, which we denote by ω. A root-n consistent estimator is given by $\hat{\omega}$, the rth largest $\hat{\theta}_i$. If we wished to construct a confidence interval for ω then we could use the bootstrap to estimate the distribution of $\hat{\omega}$. Our point is that the usual bootstrap estimator of this distribution is consistent if and only if there are no ties for the value of rth largest θ_i. In the event of a tie, the bootstrap distribution estimator does not even converge in probability to a constant. However, it does converge weakly to a distribution that we specify.

Practical situations when we want to estimate the extremes of a parameter vector occur for example in frontier curve analysis. In this field of econometrics one is interested in comparing the efficiencies of industries with similar production frontiers. The most inefficient one is the one with a maximal distance (usually the maximal intercept parameter in an ANOVA model) from the most efficient production frontier. For a discussion of different models (Cobb–Douglas, Translog, ...) which use this technique of determining inefficiency we refer to Schmidt and Sickles (1984). For an application of our proposed technique of double bootstrap see Hall, Härdle and Simar (1991). Another motivation comes from testing increasing dispersion in demand analysis.

In Härdle, Hildenbrand and Jerison (1991) positive definiteness of an 'income effects matrix' was related to the so-called law of demand. Positive definiteness was tested using the distribution of the smallest eigenvalue of this matrix. A bootstrap method was used there under the assumption that there are no ties in the eigenvalues. Here we give the reasons for this and provide remedies for these problems.

From a practical viewpoint our result indicates that the bootstrap can produce poor estimators in small to moderate samples, when there are no ties but when two or more close values of θ_i are competing for the rank of rth largest. A simulation study in Section 3 demonstrates that this is in fact the case.

Section 2 describes our results in detail, and provides a short proof. It discusses ways of alleviating the problem of inconsistency. These include testing for a tie before conducting inference, using a resample of smaller size than the actual sample, and employing the iterated bootstrap. We also address the problem of pivoting. Section 3 presents numerical work which confirms our theoretical conclusions. In both Sections 2 and 3 we confine attention to the case $r = 1$, so that $\omega = \max \theta_i$, which we denote by m. The notation for this case is simpler than for general r, but exhibits the main features of the other cases.

2. Main results

Let $\theta_1, \ldots, \theta_p$ denote unknown parameters, and write $\hat{\theta}_1, \ldots, \hat{\theta}_p$ for root-n consistent estimators. Then $\hat{m} = \max \hat{\theta}_i$ is a root-n consistent estimator of $m = \max \hat{\theta}_i$. Assume that the sequence of differences $n^{1/2}(\hat{\theta}_i - \theta_i)$ has an asymptotic p-variate normal distribution with zero mean, and that precisely q of the θ_i's are tied for $\max \theta_i$, where $1 \le q \le p$. Then the limiting distribution of $n^{1/2}(\hat{m} - m)$ is the distribution of $\max Z_j$, where $Z_1, \ldots, Z_q$ are normal random variables with zero mean, not necessarily independent. That is, adopting an unusual notation which will prove useful in a moment, if we define

$$\psi(x \mid z_1, \ldots, z_q) = P\{\max(Z_j + z_j) - \max z_j \le x\}$$

for constants z_j, then

$$P\{n^{1/2}(\hat{m} - m) \le x\} \to \psi(x) \equiv \psi(x \mid 0, \ldots, 0), \quad -\infty < x < \infty. \tag{2.1}$$

The bootstrap estimator of the distribution of $\hat{m} - m$ may be defined as follows. Suppose the estimators $\hat{\theta}_i$ were computed from a random sample $x = \{X_1, \ldots, X_n\}$, and let $x^* = \{X_1^*, \ldots X_n^*\}$ denote a resample obtained by sampling randomly, with replacement, from x. Write $\hat{\theta}_i^*$ for the version of $\hat{\theta}_i$ computed for the resample, and put $\hat{m}^* = \max \hat{\theta}_i^*$. Assume the commonly satisfied regularity condition that the sequence $\{n^{1/2}(\hat{\theta}_i^* - \hat{\theta}_i), 1 \le i \le p\}$ has the same limiting normal distribution, conditional on x, as $\{n^{1/2}(\hat{\theta}_i - \theta_i), 1 \le i \le p\}$ does inconditionally. This regularity condition is fulfilled e.g. in the case of $(\theta_1, \ldots, \theta_p)$ being the mean of a p-variate random variable with existing second moments, see Hall (1992; Section 4.2.) for a more detailed discussion. The bootstrap estimator of $\psi(x)$ is

$$\hat{\psi}(x) = P\{n^{1/2}(\hat{m}^* - \hat{m}) \le x \mid x\}.$$

We claim that $\hat{\psi}(x)$ is consistent for $\psi(x)$ if and only if there are no ties, i.e. $q = 1$. When $q \ge 2$, $\hat{\psi}(x)$ does not even converge in probability, let alone to $\psi(x)$. It does, however, converge in distribution:

$$\hat{\psi}(x) \to \psi(x \mid Z_1', \ldots, Z_q') \tag{2.2}$$

as $n \to \infty$, where $(Z_1', \ldots, Z_q')$ denotes an independent copy of $(Z_1, \ldots, Z_q)$. Note that, in the case $q = 1$, the right-hand side of (2.2) is identical to $\psi(x)$.

Next we sketch proofs of (2.1) and (2.2). Denote by $i_1, \ldots, i_q$ the distinct values of i such that $\theta_i = m$, and put $\hat{m}_{(1)} = \max_j \hat{\theta}_{i_j}$, $\hat{m}_{(1)}^* = \max_j \hat{\theta}_{i_j}^*$, $S = n^{1/2}(\hat{m} - m)$, $S_{(1)} = n^{1/2}(\hat{m}_{(1)} - m)$, $S^* = n^{1/2}(\hat{m}^* - \hat{m})$, $S_{(1)}^* = n^{1/2}(\hat{m}_{(1)}^* - \hat{m}_{(1)})$, $\theta_{(1)} = (m, \ldots, m)^T$, $\hat{\theta}_{(1)} = (\hat{\theta}_{i_j})$ and $\hat{\theta}_{(1)}^* = (\hat{\theta}_{i_j}^*)$; the latter three quantities are q-vectors. Write V for the $p \times p$ asymptotic covariance matrix of the sequence $n^{1/2}(\hat{\theta}_i - \theta_i)$, $1 \le i \le p$, and let $V_{(1)}$ be the $q \times q$ submatrix formed by taking the intersection of those rows and columns of V which have indices $i_1, \ldots, i_q$. It is straightforward to prove that

$$P(\hat{m}_{(j)} \ne \hat{m}) = o(1), \qquad P(\hat{m}_{(j)}^* \ne \hat{m}^* \mid x) = o_p(1). \tag{2.3}$$

Therefore,

$$P(S \le x) = P(S_{(1)} \le x) + o(1), \tag{2.4}$$

$$P(S^* \le x \mid x) = P(S^*_{(1)} \le x \mid x) + o_p(1). \tag{2.5}$$

Furthermore,

$$\begin{aligned} P(S_{(1)} \le x) &= P\left\{ n^{1/2} \max_j \left(\hat\theta_{i_j} - \theta_{i_j}\right) \le x \right\} \\ &\to P\left(\max_j Z_j \le x\right) = \psi(x), \end{aligned} \tag{2.6}$$

where $Z_1, \ldots, Z_q$ have a joint normal $N(0, V_{(1)})$ distribution; and

$$\begin{aligned} P(S^*_{(1)} \le x \mid x) &= P\Big[\max_j \left\{ n^{1/2}\left(\hat\theta^*_{i_j} - \hat\theta_{i_j}\right) + n^{1/2}\left(\hat\theta_{i_j} - \theta_{i_j}\right) \right\} \\ &\qquad - \max\, n^{1/2}\left(\hat\theta_{i_j} - \theta_{i_j}\right) \le x \mid x \Big] \\ &= \psi\left\{ x \mid n^{1/2}\left(\hat\theta_{i_1} - \theta_{i_1}\right), \ldots, n^{1/2}\left(\hat\theta_{i_q} - \theta_{i_q}\right) \right\} + o_p(1) \\ &\to \psi\left(x \mid Z'_1, \ldots, Z'_q\right) \end{aligned} \tag{2.7}$$

in distribution. Result (2.1) follows from (2.4) and (2.6), and (2.2) follows from (2.5) and (2.7)

There is a variety of ways of remedying the problem of inconsistency. Perhaps the most practical is to ascertain, before applying the bootstrap, whether there is empirical evidence of ties for $\max \theta_i$. For example, an *ad hoc* test for equality among the larger θ_i's could be applied. In the event that evidence for a tie was present, or perhaps more correctly, in the case of a test, that evidence for lack of ties was absent, a combined estimator of the largest θ_i should be used.

An alternative approach, which is of technical interest but could be rather difficult to implement, is to use a resample of smaller size than the sample. If the $\hat\theta_i^*$'s are computed from resamples of size n_0 instead of n, where $n_0 \to \infty$ and $n_0/n \to 0$ as $n \to \infty$, then we should re-define $S^* = n_0^{1/2}(\hat m^* - m)$ and $S^*_{(1)} = n_0^{1/2}(\hat m_1^* - \hat m_1)$. We claim that for this new definition of S^* we have $\hat\psi(x) = P(S^* \le x \mid x) \to \psi(x)$. That is the problem of inconsistency is removed. To appreciate why, note that the argument which formely gave us (2.3) and (2.7) now produces

$$\begin{aligned} \hat\psi(x) = P(S^* \le x \mid x) &= P(S^*_{(1)} \le x \mid x) + o_p(1) \\ &= \psi\left\{ x \mid n_0^{1/2}\left(\hat\theta_{i_1} - \theta_{i_1}\right), \ldots, n_0^{1/2}\left(\hat\theta_{i_q} - \theta_{i_q}\right) \right\} + o_p(1) \\ &\to \psi(x \mid 0, \ldots, 0) = \psi(x), \end{aligned}$$

the last line following since each $\hat\theta_i - \theta_i = O_p(n^{1/2})$. One drawback to this approach is that it relies on choosing n_0. An asymptotic analysis based on minimizing the mean squared error of $\hat\psi - \psi$ indicates that $n_0 = (\text{const}) \times n^{1/2}$ is

asymptotically optimal, but the constant depends on a variety of unknown quantities.

The iterated or double bootstrap, which is sometimes recommended for improving the performance of bootstrap approximations to distributions (e.g. Hall 1986, Beran 1987), is not of assistance in eradicating the problem of consistency. An application of the iterated bootstrap produces a consistent approximation to $\psi_1(x) = E\{\psi(x \mid Z_1', \ldots, Z_q')\}$, not an approximation to $\psi(x) = \psi(x \mid 0, \ldots, 0)$. Nevertheless, as the numerical work in the next section shows, the iterated bootstrap does tend to alleviate the problem. This can be ascribed to at least two factors. Firstly, in the case where there are no ties the iterated bootstrap provides particularly accurate estimators of quantiles for constructing confidence intervals, etc, even for small to moderate samples. This follows directly from Hall and Martin (1988). Our numerical study shows that this efficiency extends some distance towards the case where the larger θ_i's are close although not precisely tied. Secondly, $\psi_1(x)$ is a non-random, smoothed version of the weak limit of the conditional distribution of S^*, and as such it represents a better approximation to the non-random distribution function $\psi(x)$ then does the random, heavily sample-dependent distribution function $\psi\{x \mid n^{1/2}(\hat{\theta}_{i_0} - m), \ldots, n^{1/2}(\hat{\theta}_{i_q} - m)\}$, which $\hat{\psi}(x)$ actually approximates.

One final point worth mentioning is that of the potential for pivoting. When there are no ties for m, the asymptotic distribution of $\hat{m} - m$ depends on unknowns only through scale, which may generally be estimated. In this case the use of a pivotal bootstrap method, such as percentile-t, can enhance performance (e.g. Hall 1988). However, no reasonable prospect of pivoting appears to exist in the case where size is used to remove the problem of inconsistency. The reason is of course that the limiting distribution function ψ depends on unknowns among the Z_i's, even after standardization for scale.

3. Numerical results

The simulation study we performed split into two parts. First the effect of smaller resampling sizes was investigated, second the technique of double bootstrap was applied. The simulation setting was a p-variate Normal one with $p = 3$ and mean $\theta = (5, 5, 1)^T$ and identity covariance matrix. In the notation of Section 2, $q = 2$, $m = 5$, $i_1 = 1$ and $i_2 = 2$. All computations were done in GAUSS 2.0.

For the study of smaller resampling sizes we simulated 1000 times in order to obtain $\hat{F}_0$, the empirical distribution function of $\sqrt{n}\,(\hat{m} - m)$. Out of these 1000 simulated samples we randomly selected one and applied the bootstrap also 1000 times to this particular one. The resample size n_0 was fixed at 5 different levels: $n_0 = n$, $0.8n$, $0.6n$, $0.4n$, $0.2n$. The sample size was $n = 50$, 100, 1000.

Let $\hat{F}_1$ be the bootstrap empirical distribution function of $\sqrt{n}\,(\hat{m}^* - \hat{m})$ and $\hat{F}_2 - \hat{F}_5$ the bootstrap distributions with resample size $n_0 = 0.8n, \ldots, 0.2n$. Let $\hat{f}_j$, $j = 0, \ldots, 5$ denote estimated densities of these distributions. We used the

Table 1
The maximal deviations $\| \hat{f}_0 - \hat{f}_j \|\infty$, $j = 1,\ldots,5$ between estimated densities as a function of original sample size n and resampling size n_0

	$n = 50$	100	500	1000
$n_0 = n$	0.185	0.082	0.172	0.200
$n_0 = 0.8n$	0.227	0.067	0.175	0.150
$n_0 = 0.6n$	0.170	0.040	0.202	0.152
$n_0 = 0.4n$	0.190	0.075	0.115	0.135
$n_0 = 0.2n$	0.162	0.057	0.092	0.122

Table 2
The squared distance $\| \hat{f}_0 - \hat{f}_j \|2$, $j = 1,\ldots,5$ between estimated densities as a function of original sample size n and resampling size n_0. Entries in this table are $100 \times \| \hat{f}_0 - \hat{f}_j \|2$

	$n = 50$	100	500	1000
$n_0 = n$	0.747	0.080	0.655	0.476
$n_0 = 0.8n$	1.007	0.048	0.623	0.349
$n_0 = 0.6n$	0.628	0.024	0.619	0.322
$n_0 = 0.4n$	0.377	0.032	0.210	0.209
$n_0 = 0.2n$	0.348	0.038	0.167	0.174

Table 3
The Kolmogorov–Smirnov distance $\| \hat{F}_0 - \hat{F}_j \|\infty$, $j = 1,\ldots,5$ between estimated distribution functions as a function of sample size n and resampling size n_0

	$n = 50$	100	500	1000
$n_0 = n$	0.242	0.062	0.220	0.173
$n_0 = 0.8n$	0.266	0.048	0.208	0.139
$n_0 = 0.6n$	0.210	0.040	0.210	0.138
$n_0 = 0.4n$	0.179	0.034	0.145	0.135
$n_0 = 0.2n$	0.141	0.040	0.109	0.094

histogram method (with binwidth 0.4) to compute these density estimates. In Tables 1–3 the following distance measures are reported as a function of sample size n and resampling size n_0,

$$\text{maximal deviation} \qquad \sup_x | \hat{f}_0(x) - \hat{f}_j(x) |,$$

$$L_2\text{-distance} \qquad \frac{1}{\#\text{bins}} \sum_{\text{bins}} \left(\hat{f}_0 - \hat{f}_j \right)^2,$$

$$\text{Kolmogorov–Smirnov} \qquad \sup_x | \hat{F}_0(x) - \hat{F}_j(x) |.$$

A graphical insight into the inconsistency of the bootstrap can be obtained via Figure 1 showing the estimated densities $\hat{f}_0, \ldots, \hat{f}_5$ for $n = 500$.

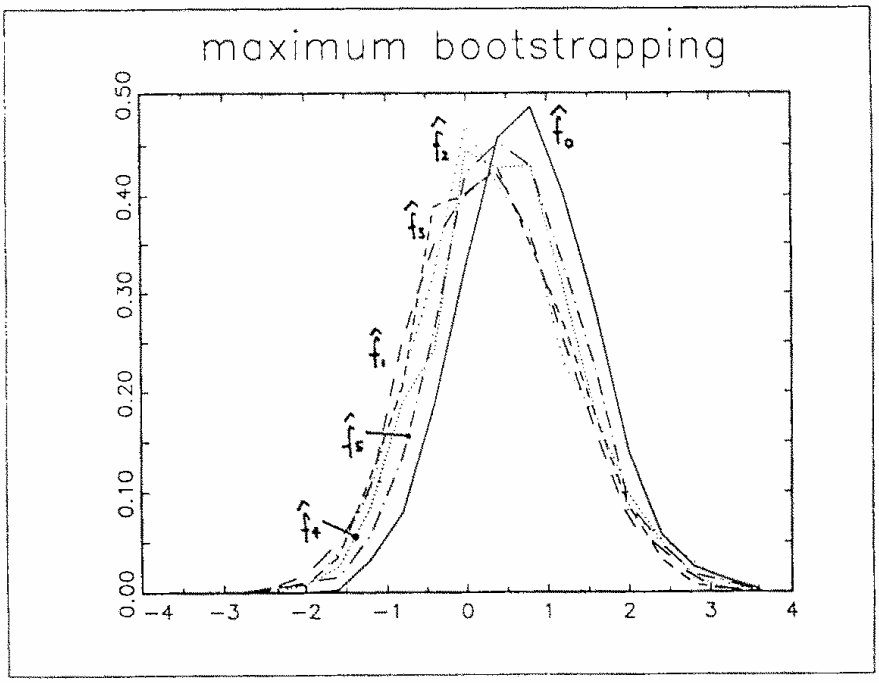


Fig. 1. The estimated densities $\hat{f}_0$ (solid line), $\hat{f}_1$ (long dashed), $\hat{f}_2$ (dots), $\hat{f}_3$ (short dashed), $\hat{f}_4$ (closely spaced dots), $\hat{f}_5$ (dots and dashes) for $n = 500$.

One sees clearly from Figure 1 how the smaller resampling size n_0 ameliorates the approximation to the simulated distribution.

The double bootstrap was applied for the same random mechanism as in the first experiment. We know from our theoretical investigations that they are both inconsistent in the case of a maximum of θ which is attained by more than one element. So we applied the double bootstrap in a narrow neighborhood of the maximum to see how well the double bootstrap ameliorates coverage probability. For this purpose we simulated for several sample sizes a three-dimensional Normal with identity covariance matrix and mean $\theta = (\theta_j, 5, 1)^T$ with $\theta_j = 5, 5.01, 5.02, 5.03, 5.04, 5.05$. The bootstrap and double bootstrap were applied each 100 times and the whole procedure was simulated 100 times for evaluation of true coverage probability for $\alpha_0 = 0.5$.

The double bootstrap works as follows. One determines a bootstrap quantile $\hat{m}_\alpha$ with nominal coverage $\pi(\alpha) = P(m \le \hat{m}_\alpha)$. We say nominal coverage since the $(1-\alpha)$-level sided confidence interval for m, $\mathcal{I} = (-\infty, \hat{m}_\alpha)$, based on bootstrapping is only asymptotically correct. However, for finite n the error $P(m \in \mathcal{I}_\alpha) - (1-\alpha)$ may be significant so one tries to estimate the true coverage $\pi(\alpha)$ by $\hat{\pi}(\alpha)$ from an iterated bootstrap. If $\hat{\pi}(\alpha)$ is our estimator of $\pi(\alpha)$ then we seek the solution $\hat{\alpha}$ of the equation

$$\hat{\pi}(\hat{\alpha}) = 1 - \alpha_0,$$

where α_0 is the predetermined level. Finally we take $\mathcal{I}_{\hat{\alpha}}$ as the improved confidence interval.

The double bootstrap, or iterated bootstrap, is a re-resample from the re-sample. If x denotes the original sample, the resample is given by x^*, i.e.

Table 4
The coverage accuracies for bootstrap (B) and double bootstrap (BB)

θ_j	$n = 20$		50		100		200		500	
	B	BB	B	BB	B	BB	B	BB	B	BB
5	0.85	0.76	0.84	0.66	0.80	0.71	0.81	0.72	0.86	0.72
5.01	0.84	0.75	0.81	0.69	0.86	0.69	0.82	0.60	0.79	0.60
5.02	0.85	0.71	0.79	0.66	0.83	0.71	0.80	0.55	0.86	0.71
5.03	0.86	0.70	0.82	0.65	0.80	0.64	0.74	0.61	0.79	0.63
5.04	0.87	0.72	0.79	0.69	0.81	0.60	0.76	0.62	0.71	0.63
5.05	0.76	0.66	0.76	0.63	0.76	0.60	0.76	0.56	0.62	0.44

$P(X_i^* = X_j) = n^{-1}$, $1 \le i, j \le n$. The re-sample x^{**} is constructed by $P(X_i^{**} = X_j^*) = n^{-1}$, $1 \le i, j \le n$. The estimate of $\pi(a)$ is $\hat{\pi}(\alpha)$, the recorded proportion of times that $\hat{m}_\alpha^*$, the double bootstrap quantile, is greater or equal $\hat{m}$. Then the equation $0.5 = \hat{\pi}(\alpha)$ is solved for α and one takes this adjusted quantile with improved accuracy. The α's were evaluated at 0.26, 0.28, ..., 0.5, 0.52, 0.54.

The true coverage for the two techniques for sample size 20, 50, 100, 200, 500 is shown in Table 4. Of course for θ_j very close to $m = 5$ the coverage accuracy is far from the nominal $\alpha = 0.5$. The double bootstrap though performs quite well in improving the coverage accuracy. Theoretical reasons for this are given in Hall (1992).

References

Athreya, K.U., Bootstrap of the mean in the infinite variance case, *Annals of Statistics*, **15** (1987) 724–731.

Beran, R., Prepivoting to reduce level error of confidence sets, *Biometrika*, **74** (1987) 457–468.

Beran, R. and M.S. Srivastava, Bootstrap tests and confidence regions for functions of a covariance matrix, *Annals of Statistics*, **13** (1985) 95–115.

Hall, P., On the bootstrap and confidence intervals, *Annals of Statistics*, **14** (1986) 1431–1452.

Hall, P., Theoretical comparisons of bootstrap confidence intervals, *Annals of Statistics*, **16** (1988) 927–953.

Hall, P., Asymptotic properties of the bootstrap of heavy-tailed distributions, *Annals of Probability*, **18** (1991) 1342–1360.

Hall, P., *The Bootstrap and Edgeworth Expansions* (Springer-Verlag, New York, 1992).

Hall, P. and M.A. Martin, On bootstrap resampling and iteration, *Biometrika*, **75** (1988) 661–671.

Hall, P., W. Härdle and L. Simar, Iterated bootstrap with applications to frontier models, CORE discussion paper no 9121 (1991).

Härdle, W., W. Hildenbrand and M. Jerison, Empirical evidence on the law of demand, *Econometrica* 59 (1991) 1525–1549.

Knight, K., On the bootstrap of the sample mean in the infinite variance case, *Annals of Statistics*, **17** (1989) 1168–1175.

Mammen, E., On the relation between asymptotic normality and consistency of bootstrap, Unpublished manuscript (1990).

Schmidt, P. and R.E. Sickles, Production frontiers and panel data, *Journal of Business and Economic Statistics*, **2** (1984) 367–374.

Econometric Theory, 8, 276-290

A BOOTSTRAP TEST FOR POSITIVE DEFINITENESS OF INCOME EFFECT MATRICES

WOLFGANG HÄRDLE
CORE, Université Catholique de Louvain
JEFFREY D. HART
Texas A&M University

Positive definiteness of income effect matrices provides a sufficient condition for the *law of demand* to hold. Given cross section household expenditure data, empirical evidence for the law of demand can be obtained by estimating such matrices. Härdle, Hildenbrand, and Jerison [10] used the bootstrap method to simulate the distribution of the smallest eigenvalue of random matrices and to test their positive definiteness. Here, theoretical aspects of this bootstrap test of positive definiteness are considered. The asymptotic distribution of the smallest eigenvalue, $\hat{\lambda}_1$, of the matrix estimate is obtained. This theory applies generally to symmetric, asymptotically normal random matrices. A bootstrap approximation to the distribution of $\hat{\lambda}_1$ is shown to converge in probability to the asymptotic distribution of $\hat{\lambda}_1$. The bootstrap test is illustrated using British family expenditure survey data.

1. INTRODUCTION

The law of demand is a centerpiece of economic theory. It guarantees uniqueness of equilibria and allows static comparison of different economical situations. It is clearly important to be able to empirically verify this law. Härdle, Hildenbrand, and Jerison [10] developed a framework for gaining empirical evidence on the law of demand from observable expenditure data. In particular they show that positive definiteness of an "income effect matrix" is a sufficient condition for the law of demand to hold. They estimated this income effect matrix from data on consumer demand, and proposed a statistical test of the hypothesis that the estimated matrix, A, is positive definite. The smallest eigenvalue, $\hat{\lambda}_1$, of the matrix estimate $\hat{A}$ was used to infer that A is positive definite. Since the sampling distribution of $\hat{\lambda}_1$ could not be readily obtained, the bootstrap method was used to perform a test of the hypothesis that the smallest eigenvalue of A is 0.

The authors are grateful to W. Stute for acquainting them with the Wielandt–Hoffman theorem. The second author expresses his appreciation to the SFB 303 at the Universität Bonn for supporting his research in Bonn during the summer of 1988. The first author thanks CORE for supporting part of this research and Herman Bierens for helpful discussions on the role of bootstrapping.

One of the purposes of the current paper is to provide theoretical justification for this inference scheme. A second aim is to develop some asymptotic distribution theory for eigenvalues of asymptotically normal random matrices. At this point we shall describe the income effects estimation problem in more abstract statistical terms. Consider the joint distribution of a k component random vector $(Y_1, \ldots, Y_k)$ and a one-dimensional random variable X. For $i, j = 1, \ldots, k$, define

$$m_{ij}(x) = E(Y_i Y_j \mid X = x)$$

and

$$a_{ij} = \int_{-\infty}^{\infty} m'_{ij}(x) f(x)\, dx,$$

where f is the density of the random variable X. In the economics setting of the above paper, $Y_1, \ldots, Y_k$ are the expenditures of a randomly selected consumer on k different goods. The quantity X is regarded as the income of the consumer. The (symmetric) matrix A of interest has typical element a_{ij}. Given n independent copies $Z_1, \ldots, Z_n$ of $Z = (X, Y_1, \ldots, Y_k)$, our goal is to infer whether or not A is positive definite. Since A is positive definite if and only if its smallest eigenvalue λ_1 is positive, one may test for positive definiteness of A by testing the hypothesis $H_0 : \lambda_1 = 0$ against the alternative $H_1 : \lambda_1 > 0$.

Härdle and Stoker [9] have shown that, under suitable conditions, the elements of A can be estimated consistently at a $\sqrt{n}$ rate. This result requires smoothness of f and the regression functions m_{ij}, but not any parametric assumptions. Note that each element of A is an *average derivative*. Average derivatives are estimated using the ADE technique of Härdle and Stoker [9] as follows. Suppose $\{(X_r, W_r)\}_{r=1}^n$ is a sample of i.i.d. random variables with $m(x) = E(W_1 \mid X_1 = x)$. The average derivative

$$\delta = E_X[m'(X)]$$

is estimated by

$$\hat{\delta} = \frac{1}{n} \sum_{r=1}^{n} W_r \hat{\ell}(X_r),$$

where

$$\hat{\ell}(x) = -\frac{\hat{f}'_h(x)}{\hat{f}_h(x)} I(\hat{f}_h(x) > b).$$

Here b denotes a cutoff bound,

$$\hat{f}_h(x) = \frac{1}{n} \sum_{r=1}^{n} K_h(x - X_r)$$

is the well-known Rosenblatt–Parzen density estimate, and $K_h(u) = h^{-1}K(u/h)$ is the kernel sequence with bandwidth h.

The matrix $\hat{A}$ used in the Härdle, Hildenbrand, Jerison [10] procedure utilizes the above ADE technique of estimating the a_{ij}. It will be shown that, for this $\hat{A}$, $\sqrt{n}(\hat{\lambda}_1 - \lambda_1)$ converges in distribution to $N(0, \sigma^2)$. In principle this fact would allow one to form an asymptotic test of the hypotheses of interest. However, the quantity σ^2 is a complicated function of A and of the covariance matrix of the elements of $\hat{A}$. Hence, using the bootstrap is an attractive alternative to the asymptotic test since the distribution of $\hat{\lambda}_1$ is easily estimated by means of resampling the data $Z_1, \ldots, Z_n$. This is done as follows: given a bootstrap sample $Z_1^*, \ldots, Z_n^*$, which is drawn at random and with replacement from $Z_1, \ldots, Z_n$, one calculates an estimator $\hat{\lambda}_1^* = \Lambda(Z_1^*, \ldots, Z_n^*)$, where Λ is such that $\hat{\lambda}_1 = \Lambda(Z_1, \ldots, Z_n)$. One then uses Monte Carlo methods to approximate $P\left(\sqrt{n}(\hat{\lambda}_1^* - \hat{\lambda}_1) > y \middle| Z_1, \ldots, Z_n\right)$. The null hypothesis $H_0: \lambda_1 = 0$ is rejected at level α if $\sqrt{n}\hat{\lambda}_1$ exceeds the $(1 - \alpha)$ percentile of the simulated bootstrap distribution.

The rest of the paper will proceed as follows. In Section 2 we derive the asymptotic distribution of the smallest eigenvalue of $\hat{A}$. Theorem 1 in that section is actually independent of the ADE technique and gives the asymptotic distribution of $\hat{\lambda}_1$ for an asymptotically normal random matrix $\hat{A}$. In Section 3 we propose a bootstrap procedure that leads to an asymptotically valid test of $H_0: \lambda_1 = 0$. Finally, we show in Section 4 an application of these ideas to inferring positive definiteness of an income effect matrix. Proofs of most of our results are given in the appendix.

2. ASYMPTOTIC DISTRIBUTION OF SMALLEST EIGENVALUE

In this section we show that the smallest eigenvalue of $\hat{A}$ is asymptotically normal. In the context of estimating covariance matrices, Muirhead [12] and Beran and Srivastava [2] have derived similar asymptotic normality results. Let us introduce the following matrix notation. A column vector of 0's and a $k \times k$ identity matrix will be denoted, respectively, $\mathbf{0}$ and I. The eigenvalues of A are $\lambda_1 < \lambda_2 < \cdots < \lambda_k$, while those of $\hat{A}$ are $\hat{\lambda}_1 < \hat{\lambda}_2 < \cdots < \hat{\lambda}_k$. It is tacitly assumed throughout that the eigenvalues of A are distinct. $C = [c_{ij}]$ will denote a $k \times k$ matrix with typical element c_{ij}. For any $k \times k$ symmetric matrix C, uvec(C) is the $k(k+1)/2$ component column vector $(c_{11}, \ldots, c_{1k}, c_{22}, \ldots, c_{2k}, \ldots, c_{kk})'$. In other words, uvec$(C)$ contains the diagonal and upper triangular elements of C. Finally, diag$(u_1, \ldots, u_k)$ will denote a $k \times k$ diagonal matrix with diagonal elements $u_1, \ldots, u_k$.

Given n i.i.d. random vectors $(X_1, Y_{11}, \ldots, Y_{k1}), \ldots, (X_n, Y_{1n}, \ldots, Y_{kn})$ our estimator $\hat{A}$ is formally defined as follows:

$$\hat{a}_{ij} = \frac{1}{n} \sum_{r=1}^{n} Y_{ir} Y_{jr} \hat{\ell}(X_r),$$

where

$$\hat{\ell}(x) = -\frac{\hat{f}_h'(x)}{\hat{f}_h(x)}\, I(\hat{f}_h(x) > b),$$

and $\hat{f}_h$ is as defined in Section 1. The quantity b is some small positive number, and the indicator in $\hat{\ell}$ is used to produce a more stable estimator of the score function $\ell = -f'/f$.

Before proceeding, we state some conditions that we need to prove our results.

(i) The support of the density f is a convex, possibly unbounded, subset of the real line.

(ii) $f(x) = 0$ for all x at the boundary of f's support. (If the support of f is the whole real line, this just means that $f(x) \to 0$ as $|x| \to \infty$.)

(iii) Each regression function $m_{ij}(x)$ is continuously differentiable for all x in the support of f.

(iv) For each (i,j) the random variable $m_{ij}'(X_1) + \ell(X_1)(Y_{i1}Y_{j1} - m_{ij}(X_1))$ has finite second moment.

(v) The density f is at least four times differentiable.

(vi) The kernel K is differentiable, has support $(-1,1)$, is symmetric about 0, and satisfies $K(u) = 0$ for $|u| \geq 1$. In addition, the kernel is of order 4, meaning that $\int_{-1}^{1} K(u)\,du = 1$, $\int_{-1}^{1} u^j K(u)\,du = 0$, $j = 1,2,3$, and $\int_{-1}^{1} u^4 K(u)\,du \neq 0$.

In addition to conditions (i)–(vi), we require that Assumptions 7, 8, and 9 of Härdle and Stoker [9] hold for f and for each m_{ij}.

(vii) The functions f and $m = m_{ij}$ are smooth in the following sense: There exist functions w_f, $w_{f'}$, w_m, and $w_{\ell m}$ such that for v close to zero

$$|f(x+v) - f(x)| < w_f(x)|v|$$

$$|f'(x+v) - f'(x)| < w_{f'}(x)|v|$$

$$|m'(x+v) - m'(x)| < w_{m'}(x)|v|$$

$$|\ell(x+v)m(x+v) - \ell(x)m(x)| < w_{\ell m}(x)|v|$$

where $E[(\ell W w_f)^2] < \infty$, $E[(W w_{f'})^2] < \infty$, $E[w_{m'}^2] < \infty$, $E[w_{\ell m}^2] < \infty$, and W denotes the random variable $Y_{i1}Y_{j1}$. The expectations are taken with respect to the joint distribution of X_1 and W.

(viii) Let $A_n = \{x | f(x) > b\}$ and $B_n = \{x | f(x) \leq b\}$. For $m = m_{ij}$, as $n \to \infty$, $\int_{B_n} m(x) f'(x)\,dx = o(n^{-1/2})$.

(ix) The fourth derivative of f, $f^{(4)}$ is locally Hölder continuous, that is, $|f^{(4)}(x+v) - f^{(4)}(x)| \leq c(x)|v|^\gamma$ for some $\gamma > 0$. The $4+\gamma$ moment of $K(\cdot)$ exists, and the following integrals are bounded as $n \to \infty$

$$\int_{A_n} m(x) f^{(4)}(x)\,dx; \qquad h^\gamma \int_{A_n} c(x) m(x)\,dx;$$

$$h \int_{A_n} m(x)\ell(x) f^{(4)}(x)\,dx; \qquad h^{\gamma+1} \int_{A_n} c(x) m(x)\ell(x)\,dx.$$

Härdle and Stoker [9] show that under the above conditions one can define sequences h and b such that

$$\sqrt{n}(\hat{a}_{ij} - a_{ij}) \xrightarrow{\mathfrak{D}} N(0, \sigma_{ij}^2)$$

as $n \to \infty$. Using their result and the Cramér–Wold device, it is shown in the appendix that

$$\sqrt{n}(\text{uvec}(\hat{A}) - \text{uvec}(A)) \xrightarrow{\mathfrak{D}} MVN(\mathbf{0}, V), \tag{2.1}$$

where V is the covariance matrix of $\text{uvec}(\chi)$ and χ has typical element

$$m_{ij}'(X_1) + \ell(X_1)(Y_{i1}Y_{j1} - m_{ij}(X_1)).$$

We now state our result on the asymptotic normality of $\hat{\lambda}_1$.

THEOREM 1. *Define $A_{ij}(\lambda_1)$ to be the cofactor of the ijth element of $A - \lambda_1 I$. Let*

$$B = 2[A_{ij}(\lambda_1)] - \text{diag}(A_{11}(\lambda_1), \ldots, A_{kk}(\lambda_1)),$$

and let $D(x) = |A - xI|$. Suppose that conditions (i)–(ix) hold. In addition, suppose that $n \to \infty$, $h \to 0$, $b \to 0$, $b^{-1}h \to 0$, $b^4 n^{1-\epsilon} h^4 \to \infty$ for some $\epsilon > 0$ and $nh^6 \to 0$. Then

$$\sqrt{n}(\hat{\lambda}_1 - \lambda_1) \xrightarrow{\mathfrak{D}} N(0, \sigma^2),$$

where

$$\sigma^2 = \frac{\text{uvec}(B)' V \text{uvec}(B)}{(D'(\lambda_1))^2}.$$

Proof. As a consequence of (2.1), $\text{uvec}(\hat{A}) \xrightarrow{p} \text{uvec}(A)$. Now, the Wielandt–Hoffman theorem (see Wilkinson [16]) implies that λ_1 is a continuous function of the elements of A, and so $\hat{\lambda}_1 \xrightarrow{p} \lambda_1$. Defining $\hat{D}(x) = |\hat{A} - xI|$, we have

$$\hat{D}(\lambda_1) = (\lambda_1 - \hat{\lambda}_1)\hat{D}'(\tilde{\lambda}_1),$$

where $\tilde{\lambda}_1$ is between $\hat{\lambda}_1$ and λ_1. This implies, along with our assumption of distinct eigenvalues for A, that

$$(\lambda_1 - \hat{\lambda}_1) = \frac{\hat{D}(\lambda_1)}{D'(\lambda_1)}\left(1 + \frac{D'(\lambda_1) - \hat{D}'(\tilde{\lambda}_1)}{\hat{D}'(\tilde{\lambda}_1)}\right).$$

From the consistency of $\hat{A}$ and $\hat{\lambda}_1$, it follows that $\hat{D}'(\tilde{\lambda}_1) \xrightarrow{p} D'(\lambda_1)$, and so $(\lambda_1 - \hat{\lambda}_1) = (\hat{D}(\lambda_1)/D'(\lambda_1))(1 + o_p(1))$. The proof is completed by using the fact (see Theorem A, p. 122 of Serfling [14]) that smooth functions of asymptotically normal vectors are also asymptotically normal. Here the vector in question is $\text{uvec}(\hat{A})$, and the smooth function is $|\hat{A} - \lambda_1 I|$. The expression for the limiting variance is obtained using Theorem A, p. 122 of Serfling [14] and the fact that, for a $k \times k$ symmetric matrix C, $\partial|C|/\partial C = 2[C_{ij}] -$

$\operatorname{diag} C_{11}, \ldots, C_{kk})$, where C_{ij} is the cofactor of c_{ij}. (See Graybill [5], p. 356 for proof of the latter fact.) ■

Theorem 1 deserves a pause for some remarks.

(1) The assumptions (i) to (vi) are quite common conditions in the setting of nonparametric regression, see, for example, Härdle [8]. In particular the kernel of order 4 allows one to estimate f at a faster rate of convergence than is possible with a positive kernel. Assumptions (vii)–(ix) are conditions on the tail behavior of mf' and $m\ell$. For a detailed discussion, see the Härdle and Stoker [9] paper.

(2) When the null hypothesis $H_0: \lambda_1 = 0$ is true, Theorem 1 implies that $\sqrt{n}\hat{\lambda}_1$ is approximately normally distributed with mean 0 and variance

$$\sigma^2 = \operatorname{uvec}(B)' V \operatorname{uvec}(B)/(D'(0))^2,$$

where $B = 2[A_{ij}] - \operatorname{diag}(A_{11}, \ldots, A_{kk})$ and A_{ij} is the cofactor of a_{ij}. Härdle and Stoker [9] show how one can construct a consistent estimator, $\hat{V}$, of V. Therefore, σ^2 is consistently estimated by $\hat{\sigma}^2 = \operatorname{uvec}(\hat{B})' \hat{V} \operatorname{uvec}(\hat{B})/(\hat{D}'(0))^2$, where $\hat{B}$ and $\hat{D}$ are defined as are B and D except with A replaced by $\hat{A}$. It follows that the statistic $\sqrt{n}\hat{\lambda}_1/\hat{\sigma}$ is the basis for an asymptotically valid test of positive definiteness of A. Owing to the possibly large dimension of V, though, there are some obvious concerns about the stability of $\hat{\sigma}$ in even moderately large samples. In the analysis of Härdle, Hildenbrand, and Jerison [10], for example, $k = 9$, and so V is a 45×45 matrix.

(3) The proof of Theorem 1 can be modified slightly to obtain the joint asymptotic normality of $\hat{\lambda}_1, \ldots, \hat{\lambda}_k$. Also, we emphasize again that Theorem 1 applies generally to any sequence of symmetric random matrices $\hat{A}_n$ which are asymptotically normal in the sense of (2.1).

(4) The assumption that the eigenvalues are different is essential for the proof of Theorem 1. We suspect that one can prove an analogous result for the case of equal eigenvalues but have been unable to do so.

3. BOOTSTRAPPING THE EIGENVALUE DISTRIBUTION

As pointed out previously, a bootstrap-based test seems an attractive alternative to the asymptotic test of Section 2. The bootstrap test used by Härdle, Hildenbrand, and Jerison [10] is the one described in Section 1. Here we shall analyze a slightly different bootstrap test. Our test is motivated by the fact that $\hat{a}_{ij}$ has the same limiting distribution as the random variable $\tilde{a}_{ij} = n^{-1} \sum_{r=1}^{n} \tilde{\ell}(X_r) Y_{ir} Y_{jr}$, where

$$\tilde{\ell}(x) = \left[\ell(x) - \frac{\hat{f}_h'(x)}{f(x)} - \frac{\hat{f}_h(x)}{f(x)} \ell(x) \right] I(f(x) > b).$$

This fact suggests that it might be reasonable to bootstrap the distribution of a sample analog of $\tilde{a}_{ij}$. To this end, consider a bootstrap sample

$(Z_1^*, \ldots, Z_n^*)$ drawn at random and with replacement from $(Z_1, \ldots, Z_n)$, and calculate the matrix $\tilde{A}^*$ with typical element $\tilde{a}_{ij}^* = n^{-1} \sum_{r=1}^n \tilde{\ell}^*(X_r^*) Y_{ir}^* Y_{jr}^*$, where

$$\tilde{\ell}^*(X_r^*) = \hat{\ell}(X_r^*) - \frac{(\hat{f}_h^*)'(X_r^*)}{\hat{f}_h(X_r^*)} I(\hat{f}_h(X_r^*) > b) - \frac{\hat{f}_h^*(X_r^*)}{\hat{f}_h(X_r^*)} \hat{\ell}(X_r^*)$$

and

$$\hat{f}_h^*(X_r^*) = \frac{1}{nh} \sum_{\substack{s=1 \\ s \neq r}}^{n} K\left(\frac{X_r^* - X_s^*}{h}\right).$$

Having obtained $\tilde{A}^*$, calculate $\tilde{\lambda}_1^*$, the smallest eigenvalue of $\tilde{A}^*$. Repeated sampling allows one to approximate the bootstrap distribution of $(\tilde{\lambda}_1^* - \hat{\lambda}_1)$ and to conduct a test of the relevant hypothesis. For a test of level α, the null hypothesis $H_0: \lambda_1 = 0$ is rejected in favor of $H_1: \lambda_1 > 0$ if $\hat{\lambda}$ is larger than the $(1 - \alpha)$ percentile of the bootstrap distribution of $\tilde{\lambda}_1^* - \lambda_1$. Note that, in $\tilde{a}_{ij}^*$, $\hat{\ell}$ and $\hat{f}_h$ are fixed with respect to the bootstrap distribution. This mimics the fact that ℓ and f are fixed functions in $\tilde{a}_{ij}$. Furthermore, since $\hat{\ell}$ and $\hat{f}_h$ will (in large samples) be close to ℓ and f, it seems intuitively plausible that the bootstrap distribution of $\tilde{a}_{ij}^*$ will be a reasonable approximation to that of $\tilde{a}_{ij}$. The following theorem shows, in fact, that the bootstrap distribution of $\sqrt{n}(\tilde{\lambda}_1^* - \hat{\lambda}_1)$ is asymptotically close to that of $\sqrt{n}(\hat{\lambda}_1 - \lambda_1)$.

THEOREM 2. *Suppose the conditions of Theorem* 1 *hold (including those on b and h), and in addition suppose that the third moment of* $m_{ij}'(X_1) + \ell(X_1)(Y_{i1}Y_{j1} - m_{ij}(X_1))$ *exists for each* (i,j). *Then*

$$\sup_{-\infty < y < \infty} \left| P\left(\sqrt{n}(\tilde{\lambda}_1^* - \hat{\lambda}_1) \leq y \,\middle|\, Z_1, \ldots, Z_n\right) - \Phi(y/\sigma) \right| \xrightarrow{p} 0,$$

where Φ *is the standard normal c.d.f. and* σ *is defined in Theorem* 1. ■

Ideally, Theorem 2 would be based on the bootstrap algorithm employed by Härdle, Hildenbrand, and Jerison [10]. Unfortunately, a proof for this case appears to be beyond the scope of current bootstrap technology. We emphasize again, though, that the modified bootstrap algorithm in Theorem 2 still leads to an asymptotically valid test of $H_0: \lambda_1 = 0$. In general, two basic advantages can result from using a bootstrap procedure. First, the bootstrap allows one to devise a valid inference scheme in problems where classical methodology offers either no solution, or only very complicated ones. This alone is a good motivation for using the bootstrap in the income effect inference problem. A second advantage is that, when applied correctly, bootstrap confidence intervals and tests often have smaller level error than do the competing procedures based on asymptotic theory (see, e.g., Beran [1], Hall [6], and Hall and Hart [7]). Suppose it is known that $\sqrt{n}(T_n - \theta)/\sigma_n$ converges in distribution to $N(0,1)$, where T_n and σ_n are statistics and θ is an unknown parameter. Then Beran [1] shows that, in general, a confidence

interval derived from the bootstrap distribution of $\sqrt{n}(T_n^* - T_n)/\sigma_n^*$ has smaller level error than does the interval $T_n \pm z_{\alpha/2}\sigma_n/\sqrt{n}$ (where z_α denotes a percentile of the standard normal distribution). On the other hand, the level error of an interval based on the bootstrap distribution of $\sqrt{n}(T_n^* - T_n)$ is of the same order as that of $T_n \pm z_{\alpha/2}\sigma_n/\sqrt{n}$. In other words, it is better to bootstrap the approximate pivotal quantity $\sqrt{n}(T_n - \theta)/\sigma_n$ than the nonpivotal $\sqrt{n}(T_n - \theta)$.

Note that our Theorem 2 applies to a nonpivotal quantity. Since the asymptotic variance σ^2 of $\hat{\lambda}_1$ can be estimated consistently, one could apply the bootstrap to the approximate pivot $\sqrt{n}(\hat{\lambda}_1 - \lambda_1)/\hat{\sigma}$. This, however, would be an incredibly computer intensive procedure owing to the complexity of $\hat{\sigma}$. For each bootstrap sample one would have to compute not only $\hat{A}^*$, but the cofactors of all the upper triangular elements of $\hat{A}^*$ and an estimate $\hat{V}^*$ of V. This is especially unattractive when A and V are high dimensional.

An alternative way to reduce the level error of a bootstrap test is to use the prepivoting method of Beran [1]. In our setting, for example, prepivoting requires no estimator of σ. The procedure requires a second level of bootstrapping in which bootstrap samples are drawn from each of the first round (or usual) bootstrap samples. While also computer intensive, using this method in our setting only requires calculating $\hat{\lambda}_1$ for a large number of data sets of size n. To show formally that prepivoting in our problem reduces asymptotic level error, Beran [1] points out that one must obtain an appropriate Edgeworth expansion for the distribution of $\sqrt{n}(\hat{\lambda}_1 - \lambda_1)/\sigma$. This we have not yet undertaken, although the work of Bickel, Götze, and van Zwet [3] may be helpful in this regard.

A final comment here concerns the bandwidth h and the quantity b in the score function estimator $\hat{\ell}$. In the proof of Theorem 2 it is assumed that all bootstrap samples (for a given data set $Z_1, \ldots, Z_n$) use the same values for h and b as were used in constructing $\hat{A}$. In practice one may have a data-based procedure for selecting these two procedure parameters. If so, this extra source of variation in $\hat{\lambda}_1$ could be incorporated into the bootstrap algorithm. However, for most applications it does not seem unreasonable to condition on such parameters when selecting bootstrap samples.

4. APPLICATION

The income effect matrix A was estimated in Härdle, Hildenbrand, and Jerison [10] from the Family Expenditure Survey (1968–1983) [4]. The smoothness of the income density can be justified for the data since recordings are made in pence and no grouping has been performed. The matrix $\hat{A}$ is symmetric and turned out to be positive definite for all years. Table 1 in that paper contains the smallest eigenvalue of $\hat{A}$ estimated from the sample of around 7000 households for each year. The condition of the matrices was

very good, and so their positive definiteness cannot be attributed to numerical errors. The entries in $\hat{A}$ and the eigenvalues were not sensitive to the choice of h and b. Varying these parameters never changed the positive definiteness.

The bootstrap procedure discussed previously was applied to the income effects matrices of several years. For the purpose of illustration we describe here the bootstrap simulated eigenvalue distribution for the years 1971 and 1973. The smallest eigenvalues of $\hat{A}$ were $\hat{\lambda}_{1971} = 0.0031$ and $\hat{\lambda}_{1973} = 0.0026$.

A graphical comparison of the two bootstrap distributions is given in Figure 1 in the form of parallel boxplots. Both distributions are skewed and the distribution for 1973 is closer to zero. Additional insight can be gained from a density estimate.

A density estimate

$$\hat{g}_h(\lambda) = \sum_{i=1}^{100} K_h \lambda - \Lambda_i / 100$$

for the eigenvalue distribution is shown for the year 1971 in Figure 2.

For this graph we used a quartic kernel $K(u) = \frac{15}{16}(1 - u^2)^2 I_{|u| \le 1)}$ and a cross-validated bandwidth. This kernel is optimal for ADE, see Mammitzsch

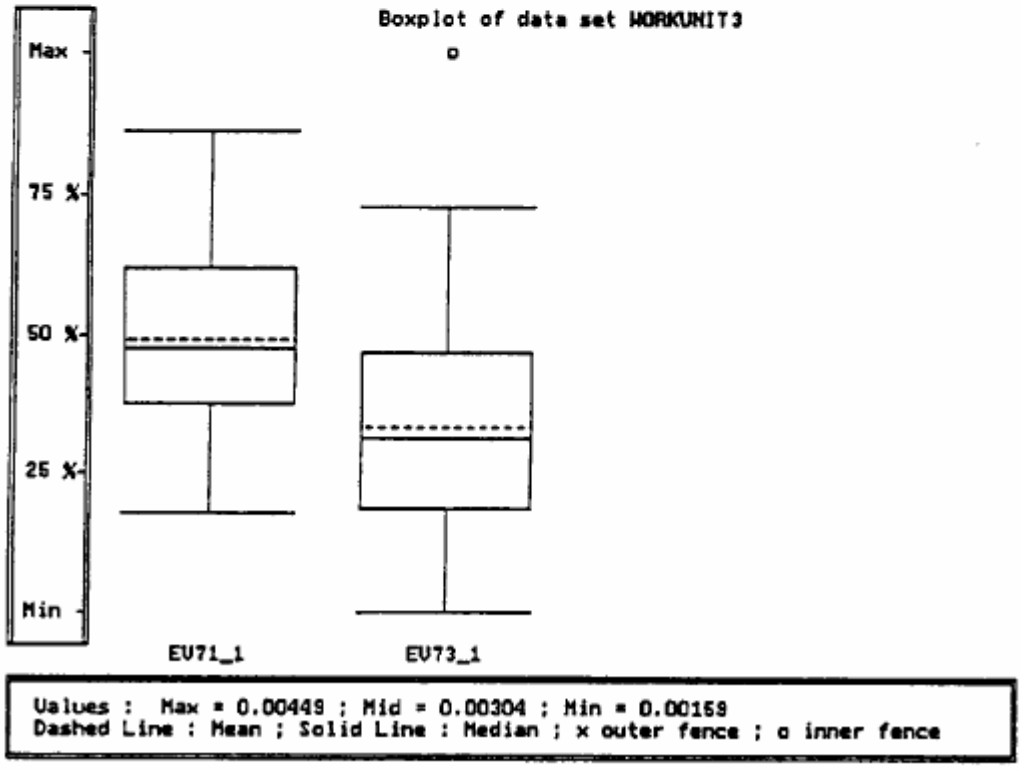


FIGURE 1. Parallel boxplots for the bootstrap distribution of the smallest eigenvalue for 1971 and 1973.

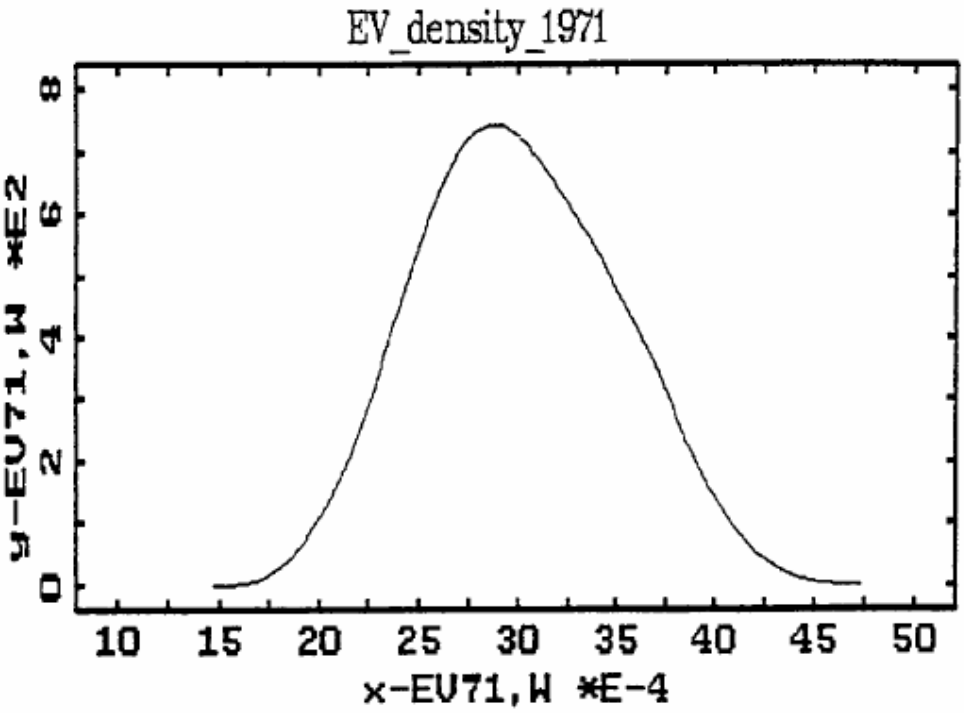


FIGURE 2. Estimated eigenvalue distribution for 1971 using a cross-validated bandwidth and quartic kernel.

[11]. The cross-validation technique yields bandwidths that asymptotically minimize the squared error between the true density and the estimated kernel density, see Stone [15]. The cross-validation function for this year is shown in Figure 3.

The kernel density function in Figure 2 has been computed using the bandwidth $h = \hat{h}$ that minimizes this cross-validation function. Figure 4 finally shows the cross-validated optimized kernel density estimate for the eigenvalue distribution of 1973. The smoothing computations here were done with the package XploRe 2.0 [17].

A bootstrap test for positive definiteness can now be conducted as follows. One determines an interval $[-B^*, C^*]$ from the bootstrap distribution of $\hat{\lambda}_1^* - \hat{\lambda}_1$ which has probability, say, 0.95. Then one computes a confidence interval $[\hat{\lambda}_1 - C^*, \hat{\lambda}_1 + B^*]$ for λ_1. The hypothesis of positive definiteness is rejected if $\hat{\lambda}_1 - C^* > 0$. (Of course, the nominal level of this one-sided test is 0.025.) It is clear from Figures 2 and 4 that the hypothesis $\lambda_1 = 0$ is untenable for both years. Though we have used a smoothed bootstrap distribution in this analysis, the results are essentially the same when we use the raw bootstrap distribution. The smoothed bootstrap is convenient for graphical purposes.

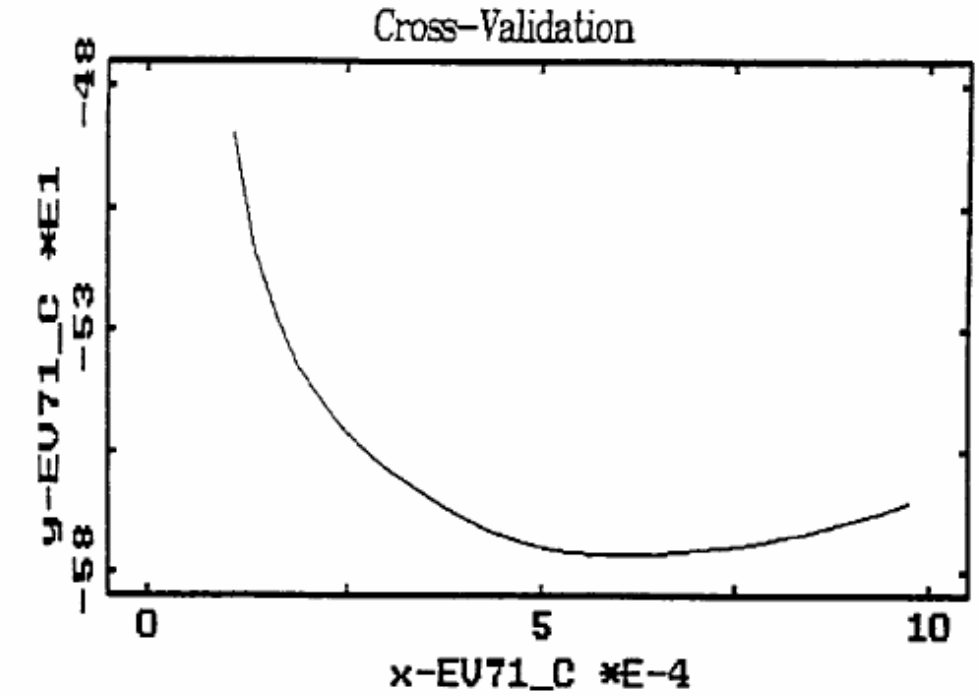


FIGURE 3. The cross-validation function for the year 1971.

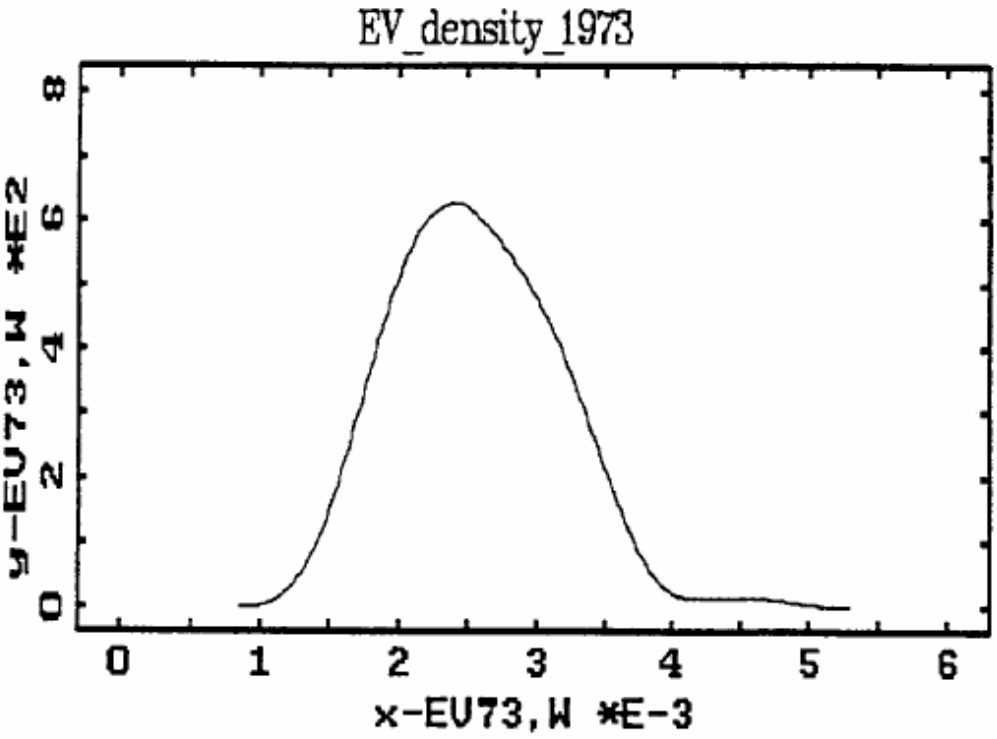


FIGURE 4. Optimally estimated eigenvalue distribution for 1973.

REFERENCES

1. Beran, R. Prepivoting to reduce level error of confidence sets. *Biometrika* 74 (1987):457–468.
2. Beran, R. & M.S. Srivastava. Bootstrap tests and confidence regions for functions of a covariance matrix. *Annals of Statistics* 13 (1985):95–115.
3. Bickel, P.J., F. Götze & W.R. van Zwet. The Edgeworth expansion for *U*-statistics of degree two. *Annals of Statistics* 14 (1986):1463–1484.
4. Family Expenditure Survey, Annual Base Tapes (1968–1983) Department of Employment, Statistics Division, Her Majesty's Stationery Office, London 1968–1983. *The data utilized in this book were made available by the ESRC Data Archive at the Universtiy of Essex.*
5. Graybill, F.A. *Matrices With Applications in Statistics.* Belmont, CA: Wadsworth, 1983.
6. Hall, P. Theoretical comparison of bootstrap confidence intervals (with discussion). *Annals of Statistics* 16 (1988):927–953.
7. Hall, P. & J.D. Hart. Bootstrap test for difference between means in nonparametric regression. *Journal of the American Statistical Association* 85 (1990):1039–1049.
8. Härdle, W. *Applied Nonparametric Regression.* Econometric Society Monograph Series 19, Cambridge University Press, 1990.
9. Härdle, W. & T. Stoker. Investigating smooth multiple regression by the method of average derivatives. *Journal of the American Statistical Association* 84 (1989):986–995.
10. Härdle, W., W. Hildenbrand & M. Jerison. Empirical evidence on the law of demand. *Econometrika* 59(1991):1525–1550.
11. Mammitzsch, V. Asymptotically optimal kernels for average derivative estimation. IMS Lecture, Davis, CA, June 1989.
12. Muirhead, R.C. *Aspects of Multivariate Statistical Theory.* New York: Wiley, 1982.
13. Randles, R.H. & D.A. Wolfe. *Introduction to the Theory of Nonparametric Statistics.* New York: Wiley, 1979.
14. Serfling, R.J. *Approximation Theorems of Mathematical Statistics.* New York: Wiley, 1980.
15. Stone, C.J. Additive regression and other nonparametric models. *Annals of Statistics* 13 (1985):689–705.
16. Wilkinson, J.H. *The Algebraic Eigenvalue Problem.* London: Oxford University Press, 1965.
17. XploRe 2.0 *XploRe—A Computing Environment for eXploratory Regression and Data Analysis.* Available from CORE (1990).

APPENDIX

Proof of (2.1). Define $\bar{a}_{ij}$ by

$$\bar{a}_{ij} = -\frac{1}{n}\sum_{r=1}^{n} Y_{ir}Y_{jr}\frac{\hat{f}'_h(X_r)}{\hat{f}_h(X_r)}\, I(f(X_r) > b).$$

The proof of Theorem 3.1 in Härdle and Stoker [9] shows that

$$\sqrt{n}(\bar{a}_{ij} - a_{ij}) = n^{-1/2}\sum_{r=1}^{n}(\chi_{ijr} - E(\chi_{ij1})) + \epsilon_{ij,n},$$

where $\chi_{ijr} = m'_{ij}(X_r) + (Y_{ir}Y_{jr} - m_{ij}(X_r))\ell(X_r)$ and $\epsilon_{ij,n} = o_p(1)$ for each (i,j).

A nontrivial linear combination of the elements of $\sqrt{n}\,\mathrm{uvec}([\bar{a}_{ij} - a_{ij}])$ has the form $T_n + S_n$, where $T_n = n^{-1/2}\sum_{r=1}^{n}\left(\sum_{i=1}^{k}\sum_{j=i}^{k} c_{ij}(\chi_{ijr} - E(\chi_{ij1}))\right)$ and $S_n =$

$\sum_{i=1}^{k}\sum_{j=i}^{k} c_{ij}\epsilon_{ij,n}$. We have $S_n \xrightarrow{p} 0$ since each $\epsilon_{ij,n} \xrightarrow{p} 0$. The random variable T_n is asymptotically normally distributed by the central limit theorem. Using Slutsky's theorem and the Cramér–Wold device, it follows that $\sqrt{n}\,\mathrm{uvec}([\tilde{a}_{ij} - a_{ij}]) \xrightarrow{\mathfrak{D}} MVN(\mathbf{0}, V)$, where V is the covariance matrix of $\mathrm{uvec}([\chi_{ij1}])$. ■

To prove that $\sqrt{n}\,\mathrm{uvec}([\hat{a}_{ij} - a_{ij}])$ has the same limiting distribution as $\sqrt{n}\,\mathrm{uvec}([\bar{a}_{ij} - a_{ij}])$, one argues as on p. 992 of Härdle and Stoker [9]. The argument is essentially the same as theirs since the same density estimate $\hat{f}_h$ is used in each $\hat{a}_{ij}$.

Before proceeding, we introduce the following notation. By $P(A \mid \mathbf{Z}_n)$ is meant the probability of event A with respect to the bootstrap distribution, that is, conditional on the observations $\mathbf{Z}_n = (Z_1, \ldots, Z_n)$. Expectation with respect to the bootstrap distribution is denoted E^*. We now state a lemma that is used in the proof of Theorem 2. Proof of the lemma is omitted but is available from the authors.

LEMMA. *Let F be a distribution function and suppose that $P(U_n \le u \mid \mathbf{Z}_n) \xrightarrow{p} F(u)$, uniformly in u. Furthermore, suppose that the random variable $P(|V_n| > \delta \mid \mathbf{Z}_n)$ converges in probability to 0 for each $\delta > 0$. Then $P(U_n + V_n \le u \mid \mathbf{Z}_n) \xrightarrow{p} F(u)$, uniformly in u.* ■

Proof of Theorem 2. Our proof begins by applying U-statistic technology to each $\tilde{a}_{ij}^*$. This is in analogy to the technique of Härdle and Stoker [9]. Let us first define the following quantities: $I_r = I(\hat{f}_h(X_r) > b)$, $I_r^* = I(\hat{f}_h(X_r^*) > b)$,

$$\hat{m}'_{ij}(x) = \frac{(n-1)}{n^2h^2}\sum_{r=1}^{n} K'\left(\frac{x - X_r}{h}\right)\frac{Y_{ir}Y_{jr}I_r}{\hat{f}_h(X_r)},$$

$$(\widehat{\ell m})_{ij}(x) = \frac{(n-1)}{n^2h}\sum_{r=1}^{n} K\left(\frac{x - X_r}{h}\right)\hat{\ell}(X_r)\frac{Y_{ir}Y_{jr}}{\hat{f}_h(X_r)},$$

$$\chi_n^{ij}(Z_r^*) = \hat{\ell}(X_r^*)Y_{ir}^*Y_{jr}^* + \hat{m}'_{ij}(X_r^*) - (\widehat{\ell m})_{ij}(X_r^*),$$

$$\tilde{a}_{1ij}^* = \frac{1}{\binom{n}{2}}\sum_{r=1}^{n-1}\sum_{s=r+1}^{n} -\frac{1}{2h^2}K'\left(\frac{X_r^* - X_s^*}{h}\right)\left(\frac{Y_{ir}^*Y_{jr}^*I_r^*}{\hat{f}_h(X_r^*)} - \frac{Y_{is}^*Y_{js}^*I_s^*}{\hat{f}_h(X_s^*)}\right)$$

and

$$\tilde{a}_{2ij}^* = -\frac{1}{\binom{n}{2}}\sum_{r=1}^{n-1}\sum_{s=r+1}^{n}\frac{1}{2h}K\left(\frac{X_r^* - X_s^*}{h}\right)\left(\frac{Y_{ir}^*Y_{jr}^*\hat{\ell}(X_r^*)}{\hat{f}_h(X_r^*)} + \frac{Y_{is}^*Y_{js}^*\hat{\ell}(X_s^*)}{\hat{f}_h(X_s^*)}\right).$$

After a bit of algebra, we have

$$\tilde{a}_{ij}^* - \hat{a}_{ij} = \left(\frac{1}{n}\sum_{r=1}^{n}\chi_n^{ij}(Z_r^*) - \hat{a}_{ij}\right) + \left(\frac{n-1}{n}\right)(\tilde{a}_{1ij}^* - \hat{a}_{ij} - P_{1ij}^*)$$
$$+ \left(\frac{n-1}{n}\right)(\tilde{a}_{2ij}^* + \hat{a}_{ij} + P_{2ij}^*), \qquad \textbf{(A.1)}$$

where P_{1ij}^* and P_{2ij}^* are the projections (with respect to the bootstrap distribution) of, respectively, $\tilde{a}_{1ij}^*$ and $\tilde{a}_{2ij}^*$ onto the class of sums of the form $\sum_{r=1}^{n} k_n(Z_r^*)$. Two salient facts for future reference are $E^*[\chi_n^{ij}(Z_r^*)] = \hat{a}_{ij}$ and $\chi_n^{ij}(Z_1^*), \ldots, \chi_n^{ij}(Z_n^*)$ are

i.i.d. with respect to the bootstrap distribution. Now, using the result on p. 83 of Randles and Wolfe [13], $nE^*(\tilde{a}^*_{1ij} - \hat{a}_{ij} - P^*_{1ij})^2 = -\beta_{1ij}/(n-1) - 2\beta_{2ij}/(n-1)$, where

$$\beta_{1ij} = \frac{1}{n}\sum_{r=1}^{n}\left(\hat{\ell}(X_r)Y_{ir}Y_{jr} + \hat{m}'_{ij}(X_r)\right)^2 - \hat{a}_{ij}^2$$

and

$$\beta_{2ij} = \frac{1}{n^2}\sum_{r=1}^{n}\sum_{s=1}^{n}\left[\frac{1}{2h^2}K'\left(\frac{X_r - X_s}{h}\right)\left(\frac{Y_{ir}Y_{jr}I_r}{\hat{f}_h(X_r)} - \frac{Y_{is}Y_{js}I_s}{\hat{f}_h(X_s)}\right)\right]^2 - \hat{a}_{ij}^2.$$

It should be clear that both β_{1ij} and β_{2ij} are bounded in probability as $n \to \infty$. (This can be shown rigorously by arguing as in the proof of Theorem 3.2 in Härdle and Stoker [9].) It now follows that

$$nE^*(\tilde{a}^*_{1ij} - \hat{a}_{ij} - P^*_{1ij})^2 = O_p(1/n). \tag{A.2}$$

Similarly one can show that

$$nE^*(\tilde{a}^*_{2ij} + \hat{a}_{ij} + P^*_{2ij})^2 = O_p(1/n). \tag{A.3}$$

The last two facts will subsequently be used in conjunction with our lemma.

To finish the proof of Theorem 2 we proceed more or less as in the proof of Theorem 1. Letting $\tilde{D}^*(x) = |\tilde{A}^* - xI|$, it follows that

$$(\hat{\lambda}_1 - \bar{\lambda}^*_1) = \frac{\tilde{D}^*(\hat{\lambda}_1)}{\hat{D}'(\hat{\lambda}_1)} + \frac{\tilde{D}^*(\hat{\lambda}_1)}{\hat{D}'(\hat{\lambda}_1)} \cdot \left(\frac{\hat{D}'(\hat{\lambda}_1) - (\tilde{D}^*)'(\xi_n)}{(\tilde{D}^*)'(\xi_n)}\right),$$

where ξ_n is a number between $\bar{\lambda}^*_1$ and $\hat{\lambda}_1$. Now, for any symmetric matrix T, let $g(\text{uvec}(T)\ (\hat{\lambda}_1) = |T - \hat{\lambda}_1 I|$. By Taylor's theorem

$$\tilde{D}^*(\hat{\lambda}_1) = \sum_{i=1}^{k}\sum_{j=i}^{k} b_{ij,n}(\tilde{a}^*_{ij} - \hat{a}_{ij}) + \sum_{i=1}^{k}\sum_{j=i}^{k}(\tilde{a}^*_{ij} - \hat{a}_{ij})\left[\frac{\partial g(\text{uvec}(T);\hat{\lambda}_1)}{\partial t_{ij}}\bigg|_{\text{uvec}(T)=\text{uvec}(A_n)} - b_{ij,n}\right], \tag{A.4}$$

where $b_{ij,n} = \partial g(\text{uvec}(T);\hat{\lambda}_1)/\partial t_{ij}\big|_{\text{uvec}(T)=\text{uvec}(\hat{A})}$, $\text{uvec}(A_n)$ lies on the line segment connecting $\text{uvec}(\hat{A})$ and $\text{uvec}(\tilde{A}^*)$, and we have used the fact that $g(\text{uvec}(\hat{A});\hat{\lambda}_1) = 0$. The first of the two double sums in (A.4) is, by (A.1), of the form

$$\frac{1}{n}\sum_{r=1}^{n}\left(\sum_{i=1}^{k}\sum_{j=i}^{k} b_{ij,n}(\chi_n^{ij}(Z_r^*) - \hat{a}_{ij})\right) + \gamma_n.$$

Using Markov's inequality along with (A.2) and (A.3), we can bound $P\left(\sqrt{n}|\gamma_n/\hat{D}'(\hat{\lambda}_1)| > \delta \,|\, \mathbf{Z}_n\right)$ by a random variable which is $O_p(1/n)$. In doing so, we make use of the fact that $\hat{D}'(\hat{\lambda}_1)$ and each $b_{ij,n}$ depend only on the data $\mathbf{Z}_n$, and that $\hat{D}'(\hat{\lambda}_1)$ and $b_{ij,n}$ are consistent estimators of, respectively, $D'(\lambda_1)$ and b_{ij}, where the matrix B is defined in Theorem 1.

Turning now to the second term in (A.4), we note that, for each (i,j),

$$\theta_{ij} = \frac{\partial g(\text{uvec}(T);\hat{\lambda}_1)}{\partial t_{ij}}\bigg|_{\text{uvec}(T)=\text{uvec}(A_n)} - b_{ij,n}$$

is a $(k-1)$st order polynomial in $\hat{\lambda}_1$ with coefficients of the form $\sum$ [(product of elements of A_n) − (product of the corresponding elements of $\hat{A}$)]. Using this and the fact (verified as in (A.6) below) that $P\left(\sqrt{n}(\tilde{a}_{ij}^* - \hat{a}_{ij}) \le u \middle| \mathbf{Z}_n\right) \xrightarrow{p} \Phi(u/\sigma_{ij})$, uniformly in u, it is straightforward to show that, for each δ,

$$P\left(\sqrt{n} \left| \sum_{i=1}^{k} \sum_{j=i}^{k} \theta_{ij}(\tilde{a}_{ij}^* - \hat{a}_{ij}) / \hat{D}'(\hat{\lambda}_1) \right| > \delta \middle| \mathbf{Z}_n \right) \xrightarrow{p} 0.$$

Using a very similar argument, we can show that, for each δ,

$$P\left(\sqrt{n} \left| \frac{\tilde{D}^*(\hat{\lambda}_1)}{\hat{D}'(\hat{\lambda}_1)} \cdot \frac{\hat{D}'(\hat{\lambda}_1) - (\tilde{D}^*)'(\xi_n)}{(\tilde{D}^*)'(\xi_n)} \right| > \delta \middle| \mathbf{Z}_n \right) \xrightarrow{p} 0.$$

Because of our lemma, all that is now left to do is to argue that

$$\sup_{-\infty<y<\infty} |G_n(y) - \Phi(y/\sigma)| \xrightarrow{p} 0,$$

where $G_n(y) = P\left(n^{-1/2} \sum_{r=1}^{n} \left(\sum_{i=1}^{k} \sum_{j=i}^{k} b_{ij,n}(\chi_n^{ij}(Z_r^*) - \hat{a}_{ij})\right) / \hat{D}'(\hat{\lambda}_1) \le y \middle| \mathbf{Z}_n\right)$. We have

$$\sup_{-\infty<y<\infty} |G_n(y) - \Phi(y/\sigma)| \le \sup_{-\infty<y<\infty} |G_n(y) - \Phi(y|\hat{D}'(\hat{\lambda}_1)|/s_n)| + \sup_{-\infty<y<\infty} |\Phi(y|\hat{D}'(\hat{\lambda}_1)|/s_n) - \Phi(y/\sigma)|, \qquad \textbf{(A.5)}$$

where s_n^2 is the bootstrap variance of $\sum_{i=1}^{k} \sum_{j=i}^{k} b_{ij,n} \chi_n^{ij}(Z_1^*)$. The quantity s_n^2 has the form $\text{uvec}(\hat{B})' \hat{V} \text{uvec}(\hat{B})$, where $\hat{B} = [b_{ij,n}]$ and $\hat{V}$ has elements $n^{-1} \sum_{r=1}^{n} (\chi_n^{ij}(Z_r) - \hat{a}_{ij})(\chi_n^{kl}(Z_r) - \hat{a}_{kl})$. By arguing as do Härdle and Stoker ([9], pp. 992–993) in proving the consistency of their estimator of the covariance matrix of the ADE, it can be shown that $\hat{V}$ is a consistent estimator of V defined in Theorem 1. As mentioned before, $\hat{B}$ and $\hat{D}'(\hat{\lambda}_1)$ are consistent for B and $D'(\lambda_1)$, and so $s_n/|\hat{D}'(\hat{\lambda}_1)| \xrightarrow{p} \sigma$. Now, the second term in (A.5) is bounded by $|(y/\sigma^*)\phi(y/\sigma^*)| \cdot |\sigma^*(|\hat{D}'(\hat{\lambda}_1)|s_n^{-1} - \sigma^{-1})|$, where ϕ is the standard normal density and σ^* is between $s_n/|\hat{D}'(\hat{\lambda}_1)|$ and σ. Since the function $z\phi(z)$ is bounded uniformly in z, and since $s_n/|\hat{D}'(\hat{\lambda}_1)|$ is consistent for σ, it follows that the second term to the right of the inequality in (A.5) converges in probability to 0.

The first term to the right of the inequality in (A.5) is bounded by

$$(33/4)E^*\left\{ \left| \sum_{i=1}^{k} \sum_{j=i}^{k} b_{ij,n}(\chi_n^{ij}(Z_1^*) - \hat{a}_{ij}) \right|^3 \right\} \Big/ \left(s_n^3 \sqrt{n}\right), \qquad \textbf{(A.6)}$$

this following from the Berry–Esséen theorem. Using Minkowski's inequality, the cube root of the expectation in this bound is no more than

$$\sum_{i=1}^{k} \sum_{j=i}^{k} |b_{ij,n}| \left[n^{-1} \sum_{r=1}^{n} |\chi_n^{ij}(Z_r) - \hat{a}_{ij}|^3 \right]^{1/3}.$$

Using our assumption of 3 moments for each χ_{ij1} and an argument as in the proof of consistency of s_n, it follows that the expectation in (A.6) is bounded in probability. Combining all the previous results and applying our lemma completes the proof of Theorem 2. ■

Allg. Statistisches Archiv, 77, 9-31.

Nichtparametrische Glättungsmethoden in der alltäglichen statistischen Praxis

Von WOLFGANG HÄRDLE und MARLENE MÜLLER, Berlin

Zusammenfassung: Nichtparametrische Glättungsmethoden sind ein oft eingesetztes Mittel der Statistik zur Darstellung der Zusammenhänge von Daten. Eine sehr bekannte Darstellungsform ist das Histogramm. Eine Mittelung von Histogrammen führt zum Kernschätzer. Wir stellen diese flexible Klasse von Schätzern vor, die außer bei der Schätzung von Verteilungsdichten auch eine große Rolle bei der nichtparametrischen Schätzung von Regressionsfunktionen spielt. Ein Hauptproblem bei der Anwendung von Kernschätzern ist die Bestimmung des optimalen Bandweitenparameters. Wir erläutern Möglichkeiten dazu, einschließlich der Konstruktion von Konfidenzintervallen und -bändern. Alle vorgestellten Methoden können auf einfache Weise in Computerprogrammen realisiert werden. Wir verwenden *XploRe 3.0 - An interactive statistical computing environment*. Die entsprechenden Routinen sind im Appendix angegeben.

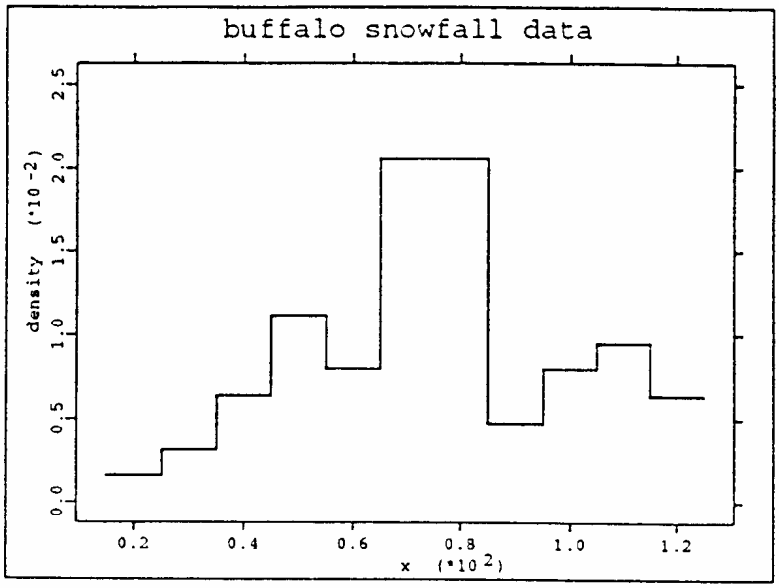
Summary: Nonparametric smoothing methods are often used in statistics for explaining connections within data. A rather well-known means to do that is the histogram. Averaging of histograms leads to the kernel estimator. We present this very flexible class of estimators which play an important role in estimating densities as well as in nonparametric estimation of regression functions. A great problem in using kernel estimators is to find the optimal bandwidth parameter. We show possibilities to do this, including the construction of confidence intervals and confidence bands. All presented methods can be easily implemented in computer programs. We use *XploRe 3.0 - An interactive statistical computing environment*; the corresponding codes are given in the appendix.

I. Immer Ärger mit Histogrammen !

Eine der Standardaufgaben der alltäglichen statistischen Praxis ist es, die Verteilung oder die Dichte von Daten darzustellen. Dazu wird zumeist das Histogramm herangezogen: Die Daten $X_1, \ldots, X_n$ werden diskretisiert, d. h. der Datenbereich wird in eine Folge von Intervallen („Bins“) gleicher Länge h (Binweite) unterteilt. Die Intervalle beginnen an einem Punkt x_0, dem „Ursprung“ der Binfolge. Die Festlegung des Ursprungs x_0 erfolgt in vielen Fällen automatisch, verbreitete Festlegungen sind z. B. $x_0 = 0$ oder $x_0 = x_{min}$. Wir wollen dieses Vorgehen an den sogenannten Buffalo-Snowfall-Daten demonstrieren. Die Daten stammen von E. PARZEN und stellen die jährlichen Schneefallhöhen (in Zoll) von Buffalo, New York, in den Wintern von 1910/11 bis 1972/73 dar (HÄRDLE, 1991; SILVERMAN, 1986). Dazu wählen wir zunächst willkürlich als Binweite (Klassenbreite) $h = 10$ und als Ursprung $x_0 = 0$. Abbildung 1 zeigt das so entstandene Histogramm für die Buffalo-Snowfall-Daten.

Härdle, W. and Müller, M. (1993) Nichtparametrische Glättungsmethoden in der alltäglichen statistischen Praxis

Abb. 1: Histogramm für Buffalo-Snowfall-Daten, $n = 63$, $h = 10$, $x_0 = 0$

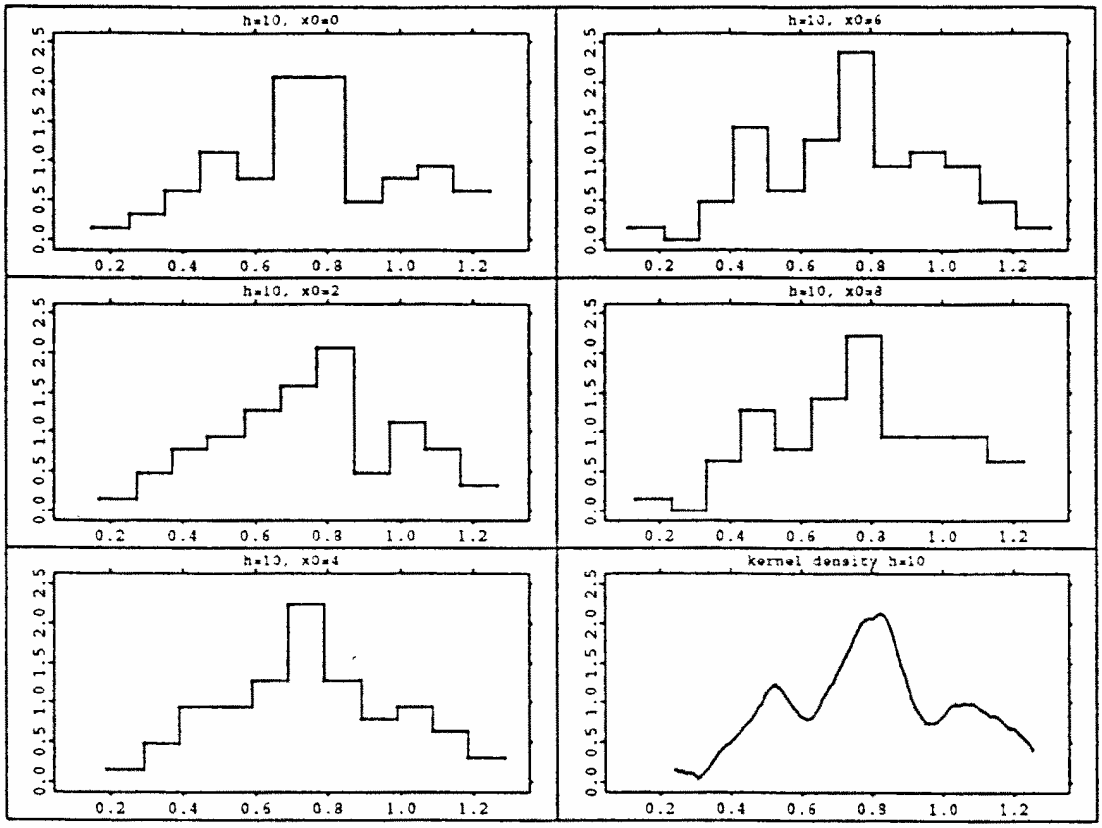


Das Histogramm scheint bei dieser Festlegung der Parameter h und x_0 für eine Verteilung der Daten mit zwei Modi zu sprechen und leicht linksschräg zu sein. Oft endet an dieser Stelle schon die statistische Analyse! Man hat ein Paar (x_0, h) gewählt und interpretiert das Ergebnis auf Grund der entstehenden Graphik.

Der Einfluß der Binweiten ist vielen Statistikern bekannt; die Abhängigkeit des Histogramms vom Parameter x_0 wird jedoch in Anwendungsfällen oft unterschätzt. Variiert man den Wert für x_0, so können sich verschiedene Histogramme ergeben. Abbildung 2 zeigt Histogramme für die Buffalo-Snowfall-Daten zur Binweite $h = 10$ auf der Grundlage der Ursprungswerte $x_0 = 0, 2, 4, 6, 8$. Das linke obere Bild entspricht Abbildung 1. Noch deutlicher für zwei Modi und Linkschrägheit scheint das linke mittlere Bild zu sprechen. Dagegen erscheinen das rechte obere und das rechte mittlere Bild deutlich rechtsschräg, während das linke untere Bild eher typisch für eine symmetrische Verteilung aussieht. Um auf die Eigenschaften der im folgenden vorgestellten Methode hinzuweisen, haben wir zum Vergleich schon einmal eine Kerndichteschätzung mit gleicher Bandweite hinzugefügt (untere Zeile rechts).

Alle Berechnungen und graphischen Darstellungen in diesem Artikel sind mit *XploRe 3.0 - An interactive statistical computing environment* erstellt worden. Der zugehörige XploRe-Code zu Abbildung 2 ist im Appendix angegeben (Programm 1). Aus der bisherigen Darstellung ergibt sich insbesondere die Frage nach einer optimalen und durch ein Computerprogramm automatisierten Wahl der Parameter des Histogramms. In den folgenden Abschnitten werden wir zeigen, wie man durch den Übergang von Histogrammen zu geglätteten Dichteschätzungen den Einfluß des Ursprungs x_0 ausschalten kann.

Abb. 2: Histogramme für Buffalo-Snowfall-Daten,
$n = 63,\ h = 10,\ x_0 = 0, 2, 4, 6, 8$



Damit reduziert sich das Problem auf die Bestimmung der Binweite h, die z. B. durch asymptotische Überlegungen oder die Kreuzvalidierungsmethode optimal gewählt werden kann.

II. WARPing

Wie eingangs erwähnt, hat das Histogramm zwei Parameter, nämlich x_0 und h. Dieser Abschnitt soll nun ein Verfahren zur Ausschaltung des Einflusses von x_0 erläutern. Die Methode des WARPing (HÄRDLE, 1991; SCOTT, 1992) besteht darin, Histogramme mit verschiedenen Ursprungswerten x_0 zu mitteln. Im einzelnen ergeben sich folgenden Schritte.

1. Diskretisiere die Daten in kleine Bins der Länge $\delta_0 = h/M$.
2. Konstruiere M Histogramme mit Binweite h und $x_0 = 0, \delta_0, 2\delta_0, \ldots, (M-1)\delta_0$.
3. Mittle über diese.

Als Ergebnis dieses Verfahrens erhält man eine genauere Aproximation an die den Daten zugrundeliegende Dichte. Der Begriff „WARPing“ stellt ein Abkürzung für *W*eighted *A*verag*ing* of *R*ounded *P*oints dar. Schritt 1 umfaßt das Runden (Diskretisieren) der Daten, im zweiten Schritt werden M Histogramme mit verschiedenen Ursprungswerten berechnet. Die einzelnen gerundeten Datenpunkte werden durch die Mittelung in Schritt 3 gewichtet.

Da die Mittelung von Histogrammen einer Bewichtung der gerundeten Datenpunkte durch die Dreiecksdichte („Dreieckskern“)

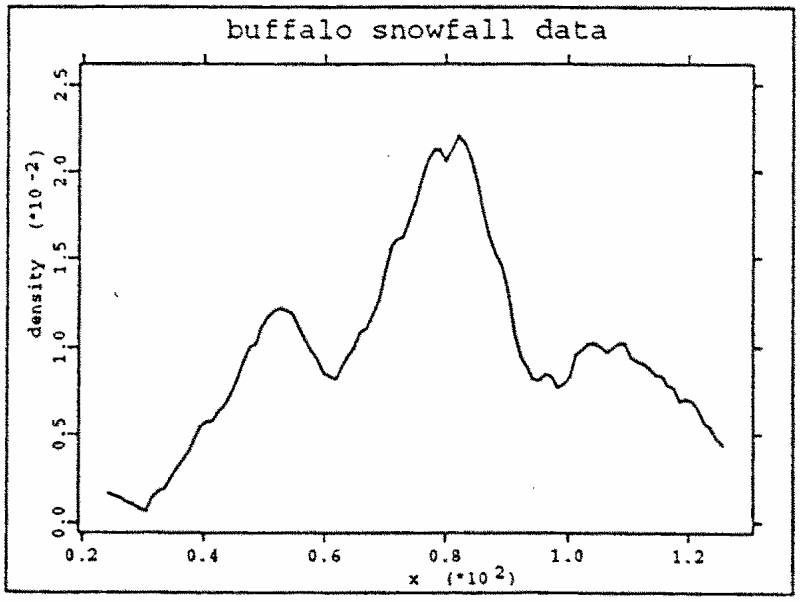
$$K(u) = (1 - |u|)\mathrm{I}(|u| \leq 1) \tag{1}$$

äquivalent ist, läßt sich das Ergebnis der WARPing-Methode (im Fall sehr kleiner Bins) einfacher in der Form

$$\hat{f}_h(x) = n^{-1} \sum_{i=1}^{n} K_h(x - X_i) \tag{2}$$

mit $K_h(\bullet) = K(\bullet/h)/h$ angeben. Abbildung 3 zeigt das "WARPed" Histogramm für die Buffalo-Snowfall-Daten, wieder zur Binweite (Bandweite) $h = 10$. Diese Abbildung scheint fast identisch zu Abbildung 2, rechts unten; der Unterschied besteht in der gewählten Funktion K, die offenbar zu einer etwas weniger glatten Gestalt der Funktion führt. Auf den Einfluß verschiedener möglicher Dichtefunktionen K und deren Vergleichbarkeit wird Abschnitt V näher eingehen. Die WARPing-Methode hat also zunächst einmal die Frage der x_0-Abhängigkeit beantwortet. Die Wahl von h und K (der effektiven Gewichte) ist noch offen.

Abb. 3: „WARPed“ Histogramm für Buffalo-Snowfall-Daten, $n = 63$, $h = 10$ (Programm 2)



III. Kerndichteschätzer

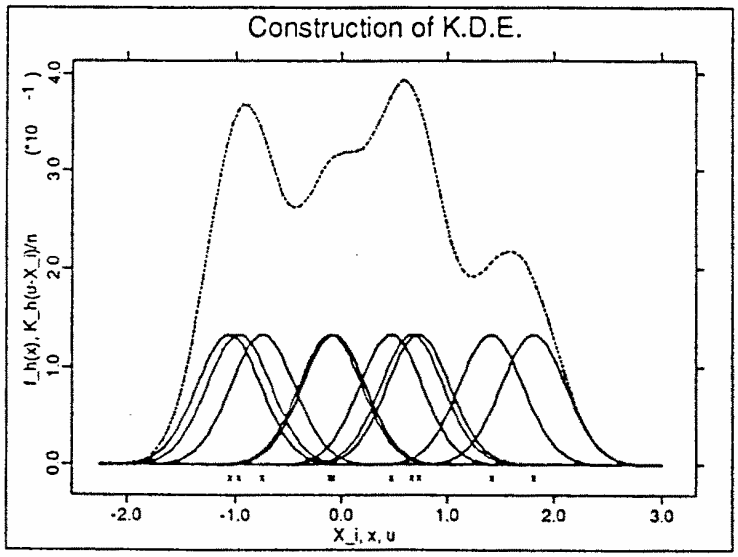
Wir erinnern noch einmal an unsere Ausgangssituation: Gegeben sind Daten $X_1, X_2, \ldots, X_n$, für deren Verteilung oder Dichte f wir uns interessieren. Eine Klasse von wirkungsvollen Verfahren zur Ermittlung bzw. Schätzung von f stellen die sogenannten Kerndichteschätzer dar, die sich in Verallgemeinerung der vorgestellten WARPing-Methode durch Formel (2) beschreiben lassen, wobei als mögliche Kerne oder Kernfunktionen in der Literatur auch die Funktionen aus Tabelle 1 verwendet werden.

Tabelle 1: Kernfunktionen

Kernfunktion	Bezeichnung	XploRe-Name
$K(u) = \frac{1}{2}\mathrm{I}(\lvert u\rvert \leq \frac{1}{2})$	Rechteckskern	`uni`
$K(u) = (1 - \lvert u\rvert)\mathrm{I}(\lvert u\rvert \leq 1)$	Dreieckskern	`trian`
$K(u) = \frac{3}{4}(1 - u^2)\mathrm{I}(\lvert u\rvert \leq 1)$	Epanechnikov-Kern	`epa`
$K(u) = \frac{15}{16}(1 - u^2)^2\mathrm{I}(\lvert u\rvert \leq 1)$	Quartic-Kern	`qua`
$K(u) = \frac{35}{32}(1 - u^2)^3\mathrm{I}(\lvert u\rvert \leq 1)$	Triweight-Kern	`qua`
$K(u) = \frac{1}{\sqrt{2\pi}} exp(-\frac{u^2}{2}) = \varphi(u)$	Normal- od. Gauss-Kern	`gau`
$K(u) = \frac{\pi}{4} cos(\frac{\pi}{2}u)\mathrm{I}(\lvert u\rvert \leq 1)$	Kosinuskern	`cosi`

Der Kerndichteschätzer wird so konstruiert, indem über jeden Datenpunkt die mit h skalierte Kernfunktion K_h gelegt und dann gemittelt wird. Die Form des Kerns bestimmt das lokale Gewicht. In der folgenden Abbildung ist ein Normalkern verwendet worden (TURLACH, 1992).

Abb. 4: Der Kerndichteschätzer ist ein Mittel über K_h an den Beobachtungen



Die Wahl der Bandweite h ist sehr wichtig. Der Effekt, der durch unterschiedliche Wahl von h entsteht, soll an einer durch Zufallszahlen erzeugten bimodalen Verteilung gezeigt werden (Abb. 5, 6). Die durchgehende Linie stellt jeweils die bimodale Dichte dar (Mischung zweier Normalverteilungen), während die gestrichelten Linien die geschätzte Dichte mit einem sehr großen Wert für h (Abb. 5) bzw. einem sehr kleinen Wert für h (Abb. 6) zeigen. Generell läßt sich die Wirkung des h-Wertes wie folgt zusammenfassen: Ist h zu groß, so ist die Dichteschätzung zu glatt, ist dagegen h zu klein, so ist die Dichteschätzung zu rauh.

Abb. 5: Der Effekt des Überglättens

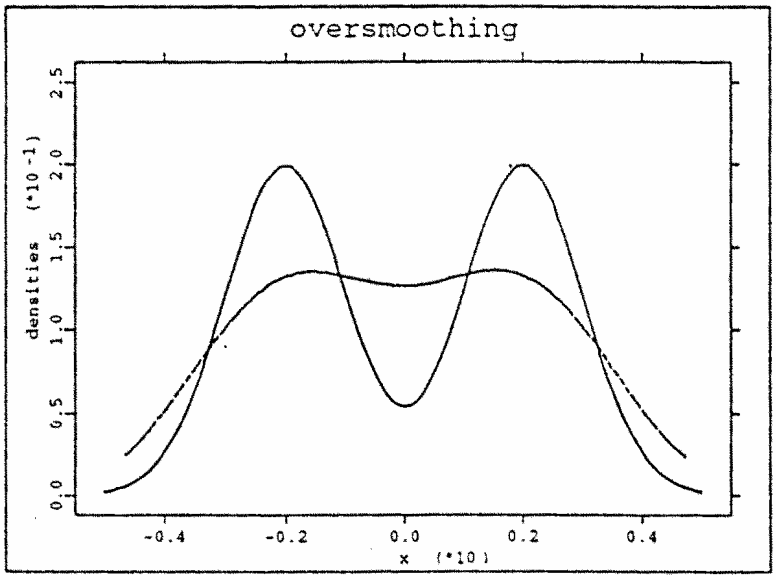
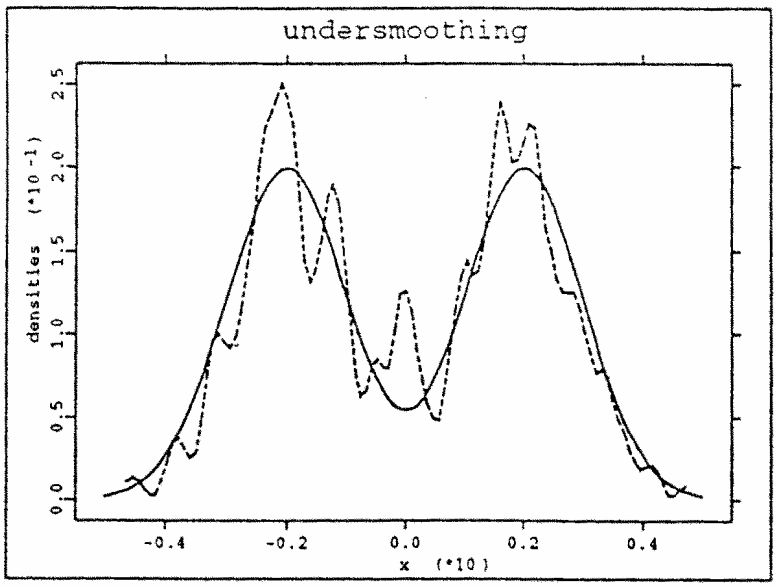


Abb. 6: Der Effekt des Unterglättens



Da mit der Bandweite h die Glattheit der Dichteschätzung wächst, wird h auch oft als Glättungsparameter bezeichnet. Der folgende Abschnitt wird Möglichkeiten zur optimalen Wahl von h erörtern.

IV. Optimierung der Dichteschätzer

Aus dem dargestellten Problem der Wirkungen des Parameters h ergibt sich die Frage nach einem Verfahren zur Bestimmung eines optimalen Wertes für den Glättungsparameter. In der Literatur wird eine ganze Reihe verschiedener Verfahren zur Wahl von h vorgeschlagen. Wir wollen uns auf die Skizzierung zweier sehr verbreiteter Methoden beschränken. Weitere Möglichkeiten und Erläuterungen findet man u. a. in den Arbeiten von HÄRDLE/HALL/MARRON sowie SCOTT und TURLACH.

Ein möglicher Ansatz ist, h so zu bestimmen, daß es einen zu definierenden Abstand $d(h)$ zwischen der zugrundeliegenden Dichte f und der geschätzten Dichte $\hat{f}_h$ minimiert. Ein vielfach verwendetes Abstandsmaß ist hier der integrierte quadratische Fehler (integrated squared error)

$$ISE(h) = \int (\hat{f}_h(x) - f(x))^2 dx, \tag{3}$$

der die Eigenschaft hat, sowohl für zu großes als auch für zu kleines h groß zu werden. Das Optimum liegt daher „irgendwo in der Mitte". $ISE(h)$ ist im Normalfall nicht exakt berechenbar. Für den Erwartungswert von $ISE(h)$

$$\mathbf{E}\, ISE(h) = \int Var(\hat{f}_h(x))dx + \int [Bias(\hat{f}_h(x))]^2 dx \tag{4}$$

läßt sich eine asymptotische gültige Formel angeben:

$$\mathbf{E}\, ISE(h) \approx C_1 n^{-1} h^{-1} + C_2 h^4, \tag{5}$$

wobei hier die Konstanten C_1, C_2 die Ausdrücke $C_1 = \int K^2(u)du$ und $C_2 = \frac{1}{4}\,\mu_2^2(K) \int [f''(x)]^2 dx$ bezeichnen, mit $\mu_2(K) = \int u^2 K(u)du$. Bei dem Versuch, (5) über h zu minimieren, ergibt sich jedoch ein circulus vitiosus:

Um $\hat{f}_h$ zu optimieren, muß man bereits $\int [f'']^2$ kennen!

Statistische Ansätze zur Auflösung des Teufelskreises bestehen in der Schätzung von $\int [f'']^2$ oder $ISE(h)$ selbst. Im letzteren Fall ist eine Schätzung nur bis auf eine Konstante genau zu anzugeben, da nur das Minimum von $ISE(h)$ interessiert.

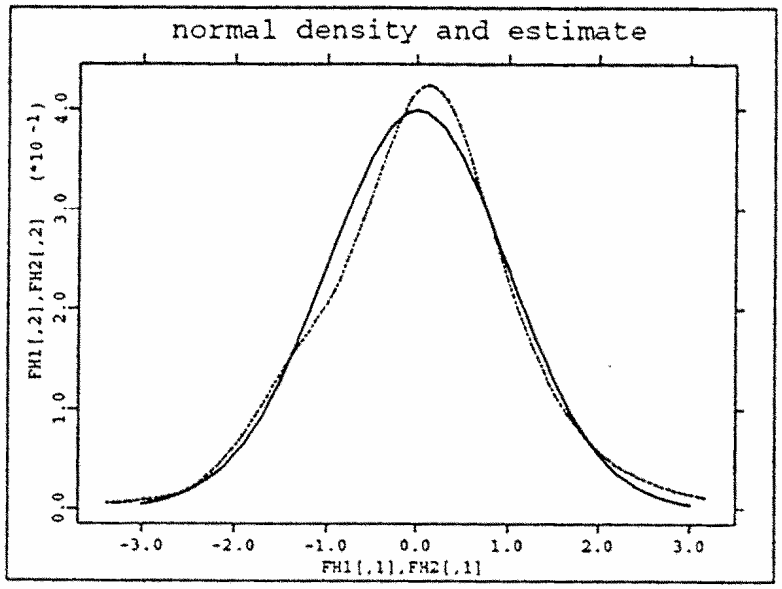
Wir werden als Lösungen insbesondere die Plug-In-Methode und die Kreuzvalidierungs-Methode betrachten. Als Plug-in-Methoden werden Schätzverfahren bezeichnet, die unbekannte Parameter durch deren Schätzungen ersetzen. Die einfachste „quick & dirty" Plug-in-Methode geht auf SILVERMAN zurück

und besteht in der Wahl des passenden (optimalen) Parameters $\hat{h}$ für nach $N(0, \sigma^2)$ normalverteilte Daten unter Verwendung des Normalkerns. Hier ergibt sich

$$\hat{h} = 1.06\ \hat{\sigma}\ n^{-1/5} \tag{6}$$

mit $\hat{\sigma}$ als Schätzung der Standardabweichung. Dieses Verfahren ist im XploRe-Macro `denauto.xpl` realisiert (Programm 3). Als Beispiel zeigen wir in Abbildung 7 die geschätzte Dichte für eine normalverteilte Stichprobe.

Abb. 7: Dichteschätzung und Dichte der Normalverteilung



Die Methode der Kreuzvalidierung (cross-validation) ist ein universell einsetzbares Schätzverfahren. Wir verwenden es zur Schätzung von $ISE(h)$. Auf Grund der Darstellung

$$ISE(h) = \int \hat{f}_h{}^2(x)dx - 2\mathbf{E}_f[\hat{f}_h(X)] + \int f^2(x)dx \tag{7}$$

ist hier nur der zweite Term zu schätzen. Die Idee der Kreuzvalidierung besteht darin, zur Ermittlung von $\mathbf{E}_f[\hat{f}_h(X)]$ über Dichteschätzungen an den Punkten X_i zu mitteln und dabei in der jeweiligen Dichteschätzung den Punkt X_i selbst nicht zu verwenden. In Formeln ausgedrückt heißt das

$$\widehat{\mathbf{E}_f[\hat{f}_h(X)]} = n^{-1} \sum_{i=1}^{n} \hat{f}_{h,-i}(X_i),$$

wobei

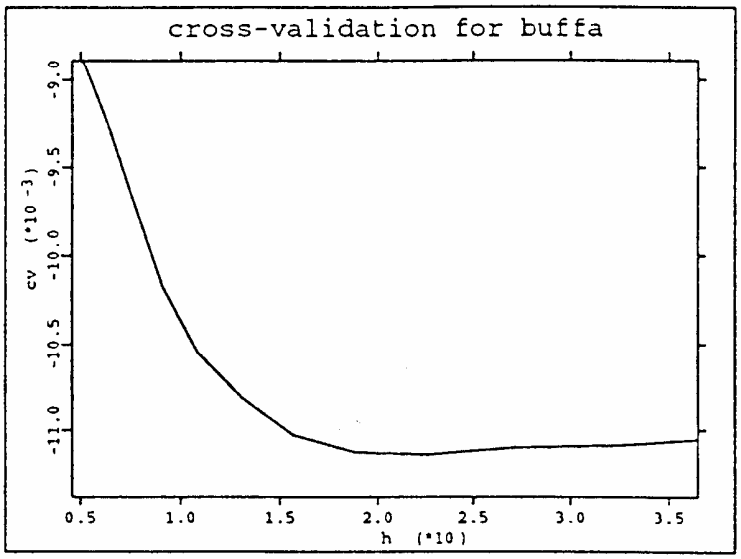
$$\hat{f}_{h,-i}(x) = (n-1)^{-1} \sum_{j=1 i \neq j}^{n} K_h(x - X_j)$$

eine Kerndichteschätzung für f unter Auslassung des i-ten Datenpunktes ist. Daher läßt sich $ISE(h)$ bis auf eine Konstante c durch

$$\widehat{ISE}(h)+c = \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}(K*K)_h(X_i-X_j) - \frac{2}{n(n-1)}\sum_{i=1}^{n}\sum_{\substack{j=1\\ i\neq j}}^{n}K_h(X_i-X_j) \quad (8)$$

schätzen. $K * K$ bezeichnet die Faltung des Kerns K mit sich selbst, d. h. $K * K(\bullet) = \int K(u)K(\bullet - u)du$. Dieses Verfahren ist für den Quartic-Kern in dem XploRe-Macro **dencvl.xpl** (Programm 5) implementiert. Abbildung 8 zeigt die durch dieses Macro berechnete Kurve der Werte von $\widehat{ISE}(h) + c$ gemäß Formel (8).

Abb. 8: Kreuzvalidierung für die Buffalo-Snowfall Daten

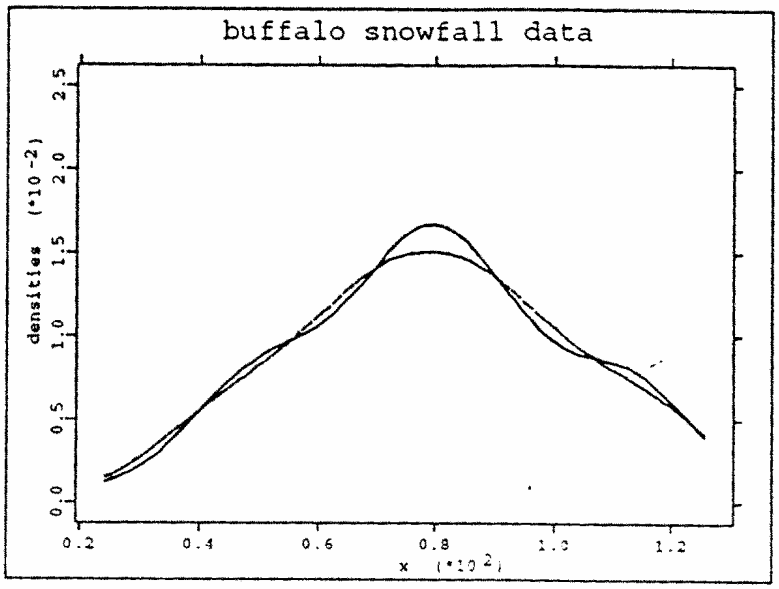


Die Minimierung der von **dencvl.xpl** errechneten Kurve ermittelt einen optimalen Bandweiten-Wert $h \approx 22$. Wir stellen die zwei mittels Plug-In-Verfahren und Kreuzvalidierung optimierten Kerndichteschätzungen in Abb. 9 dar. Die offensichtlich überglättende, etwas flachere Kurve (gestrichelte Linie) stellt die mittels Plug-In-Verfahren optimierte Schätzung dar. Hier wird deutlich, daß dieses Verfahren im Fall einer zur Normalverteilung wenig ähnlichen Dichte ein schlechteres Ergebnis erbringen wird.

V. Kanonische Kerne

Es sei noch einmal an die beiden Dichteschätzungen in Abb. 2, unten rechts, und in Abb. 3 erinnert. Beide zeigten Kerndichtschätzungen zur Bandweite $h = 10$, wobei in Abb. 2 der Quartic-Kern und in Abb. 3 der Dreieckskern verwendet wurde. Deutlich sichtbar ist die rauhere Struktur der Dichteschätzung in Abb. 3, die zeigt, daß die Bandweite h zu niedrig gewählt wurde.

Abb. 9: Kerndichteschätzer mit optimierten Bandweiten



In Abschnitt III wurde erläutert, welchen Einfluß der Wert h auf die Glattheit der entstehenden Schätzung hat, wobei sich zeigte, daß die Dichteschätzungen mit wachsendem h glatter werden. Daher ergibt sich die Frage, ob man für den in Abb. 3 verwendeten Dreieckskern die Bandweite h so weit erhöhen kann, daß die dabei entstehende Dichteschätzung identisch zur Quartic-Kern-Dichteschätzung wird. Allgemeiner formuliert: Welche Beziehungen gibt es zwischen den verschiedenen Kernfunktionen aus Tabelle 1?

Die Antwort gibt das von MARRON/NOLAN (1989) eingeführte Konzept der kanonischen Kerne. Diese Konzept bringt die in Tabelle 1 angegebenen Kerne auf eine gemeinsame Skala, so daß die Kernfunktionen zueinander äquivalent werden. Genauer gesagt, ergibt sich für zwei dieser Kerne

$$\hat{f}_{h_1,K_1}(x) = n^{-1} \sum_{i=1}^{n} (K_1)_{h_1}(x - X_i) \approx n^{-1} \sum_{i=1}^{n} (K_2)_{h_2}(x - X_i) = \hat{f}_{h_2,K_2}(x), \quad (9)$$

falls die Bandweiten h_1, h_2 die Formel

$$h_2 = h_1 \frac{\delta_2^*}{\delta_1^*} \quad (10)$$

erfüllen. Die Faktoren δ_1^*, δ_2^* berechnen sich nach

$$\delta_i^* = \left(\frac{\int K_i^2(u)du}{\mu_2^2(K_i)} \right)^{1/5}. \quad (11)$$

Dieser Zusammenhang zwischen verschiedenen Kernfunktionen läßt sich auf einfache Weise durch die asymptotische Formel (5) für $\mathbf{E}ISE(h)$ erklären.

$\int K_i^2(u)du$ und $\mu_2^2(K_i)$ sind diejenigen Faktoren in Bias und Varianz, die vom Kern abhängen. Verschiedene Kerne ergeben bei gleicher Bandweite h verschiedene Werte von $\mathbf{E}ISE(h)$, so daß die Umrechnung nach Formel (10) gerade asymptotisch gleiche Werte von $\mathbf{E}ISE(h)$ erreicht. Da $\mathbf{E}ISE(h)$ ebenfalls ein Abstandsmaß zwischen Dichteschätzung $\hat{f}_h$ und der zugrundeliegenden Dichte f ist, erhält man daher auch die asymptotische Gleichheit in (9). Tabelle 2 gibt die Umrechnungsfaktoren für die in Tabelle 1 aufgeführten Kernfunktionen an.

Tabelle 2: Umrechnungsfaktoren für Kernfunktionen

δ_j^*/δ_i^*	Rechteck	Dreieck	Epanechnikov	Quartic	Triweight	Normal	Kosinus
Rechteck	1.000	0.715	0.786	0.663	0.584	1.740	0.761
Dreieck	1.398	1.000	1.099	0.927	0.817	2.432	1.063
Epanechnikov	1.272	0.910	1.000	0.844	0.743	2.214	0.968
Quartic	1.507	1.078	1.185	1.000	0.881	2.623	1.146
Triweight	1.711	1.225	1.345	1.136	1.000	2.978	1.302
Normal	0.575	0.411	0.452	0.381	0.336	1.000	0.437
Kosinus	1.315	0.941	1.033	0.872	0.768	2.288	1.000

Damit ergibt sich etwa die zur Quartic-Kerndichteschätzung in Abb. 2 äquivalente Bandweite für den Dreieckskern als $h = 10.78$, d. h. die Bandweite muß beim Dreieckskern vergrößert werden, um eine zu Abb. 2 äquivalente Kerndichteschätzung zu erhalten.

VI. Konfidenzintervalle und -bänder

Da sich für die Kerndichteschätzungen in einem Punkt x eine asymptotische Normalverteilung für gegen unendlich wachsenden Stichprobenumfang n herleiten läßt, kann man für $\hat{f}_h(x)$ auch Konfidenzintervalle konstruieren. Genauer gilt (SILVERMAN, 1986; HÄRDLE, 1991) unter der Vorraussetzung, daß die zweite Ableitung $f''(x)$ existiert, für eine Bandweite $h = cn^{-1/5}$ die asymptotische Beziehung

$$n^{2/5}(\hat{f}_h - f(x)) \xrightarrow{\mathcal{L}} N\left(\frac{c^2}{2}f''(x)\mu_2(K)\ ,\ c^{-1}f(x)C_1\right), \tag{12}$$

wobei C_1 wie zuvor $\int K^2(u)du$ bezeichne. Damit ergibt sich als asymptotisches Konfidenzintervall für $f(x)$

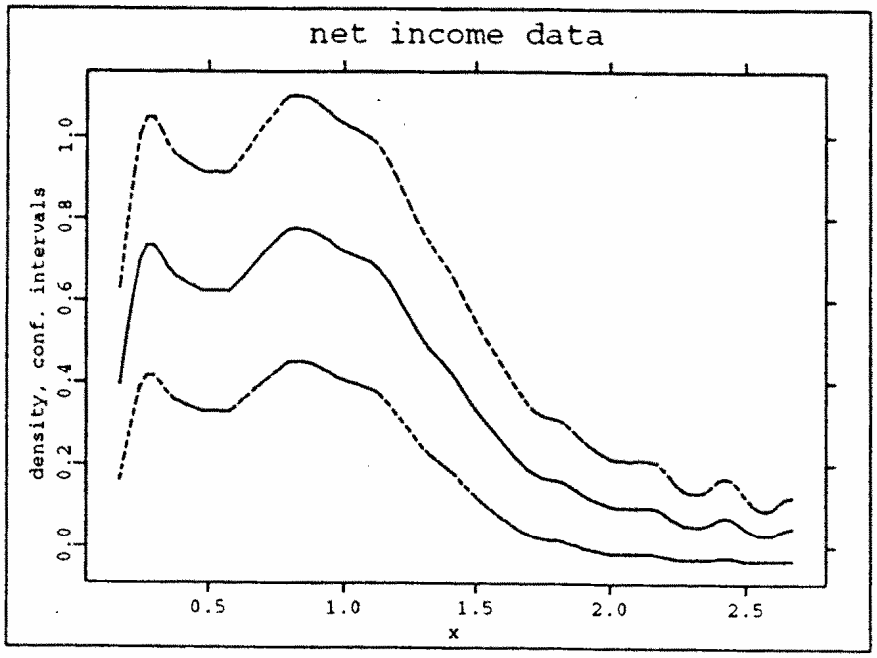
$$\left[\hat{f}_h - \frac{h^2}{2}f''(x)\mu_2(K) - u_{1-\alpha/2}\sqrt{\frac{f(x)C_1}{nh}},\right.$$
$$\left.\hat{f}_h - \frac{h^2}{2}f''(x)\mu_2(K) + u_{1-\alpha/2}\sqrt{\frac{f(x)C_1}{nh}}\right] \tag{13}$$

mit $u_{1-\alpha/2}$ als dem $(1-\frac{\alpha}{2})$-Quantil der Standard-Normalverteilung. Für kleines h (im Verhältnis zu $n^{-1/5}$) ist der jeweils zweite Term in den Konfidenzgrenzen vernachlässigbar, so daß mit der Plug-In-Schätzung $\hat{f}_h(x)$ für $f(x)$ das Konfidenzintervall aus (13) durch das Intervall

$$\left[\hat{f}_h - u_{1-\alpha/2}\sqrt{\frac{\hat{f}_h(x)C_1}{nh}} \;,\; \hat{f}_h + u_{1-\alpha/2}\sqrt{\frac{\hat{f}_h(x)C_1}{nh}} \right] \tag{14}$$

approximiert werden kann. Die entsprechenden Intervalle für die Nettoeinkommens-Daten britischer Familien von 1973 (Family Expenditure Survey, 1968-83) sind in Abb. 10 dargestellt, der dazugehörige XploRe-Code ist in Programm 6 zu finden.

Abb. 10: Konfidenzintervalle für die Dichte der Nettoeinkommens-Daten, $h = 0.15$, $\alpha = 0.05$



Im Unterschied zu den vorangehend betrachteten Konfidenzintervallen, die jeweils nur für einen Funktionswert $f(x)$ gültig sind, lassen sich auch Formeln zur Berechnung von Konfidenzbändern für die ganze Funktion f herleiten. Wir wollen hierzu die Formeln von BICKEL/ROSENBLATT (1973) für eine auf das Intervall $[0,1]$ beschränkte Dichte f angeben. BICKEL/ROSENBLATT verwenden eine geringere Ordnung als $n^{-1/5}$ für die Bandweite h, was zur Biasreduktion und damit zur asymptotischen Erwartungstreue der Schätzung $\hat{f}_h(x)$ führt. Unter gewissen Regularitätsvoraussetzungen an f erhält man für $h_n = n^{-\delta}, \delta \in (\frac{1}{5}, \frac{1}{2})$ und alle $x \in [0,1]$ die Formel

$$\lim_{n\to\infty} P\Bigg[\hat{f}_h(x) - \left(\frac{\hat{f}_h(x)C_1}{n^{1-\delta}} \right)^{1/2} \left(\frac{z}{(2\delta \log n)^{1/2}} + d_n \right)^{1/2}$$
$$\leq f(x) \leq \hat{f}_h(x) + \left(\frac{\hat{f}_h(x)C_1}{n^{1-\delta}} \right)^{1/2} \left(\frac{z}{(2\delta \log n)^{1/2}} + d_n \right)^{1/2} \Bigg]$$
$$= \exp\left(-2\exp(-z) \right) \tag{15}$$

mit

$$d_n = (2\delta \log n)^{1/2} + (2\delta \log n)^{-1/2} \log\left(\frac{1}{\pi} \frac{\int (K'(u))^2 du}{2\sqrt{C_1}} \right).$$

Ein Konfidenzband zu einem gegebenen Signifikanzniveau α ergibt sich daher aus der Berechnung des Wertes z, der

$$\exp\left(-2\exp(-z)\right) = 1 - \alpha.$$

erfüllt. So erhält man z. B. für $\alpha = 0.05$ die Lösung $z \approx 3.663$.

VII. Nichtparametrische Regression

Das Regressionsmodell geht üblicherweise von Messungen zweier Größen X und Y, d. h. Daten der Form $(X_1, Y_1), \ldots, (X_n, Y_n)$ aus, die durch eine unbekannte Regressionsfunktion $m(\bullet)$ in folgender Weise verbunden sind:

$$Y_i = m(X_i) + \epsilon_i \ , \ i = 1, \ldots, n, \tag{16}$$

ϵ_i bezeichnen die Fehler. Das Ziel der Regressionsanalyse besteht in der Schätzung der Funktion $m(\bullet)$. Ist außer den Daten auch eine bestimmte Struktur dieser Funktion bekannt, so ermittelt man $m(\bullet)$ in der Regel über einen parametrischen Ansatz, im einfachsten Fall setzt man $m(\bullet)$ als lineare Funktion von X an und erhält das bekannte lineare Regressionsmodell. Ist die Funktion $m(\bullet)$ a priori nicht in ihrer funktionalen Form spezifiziert, bietet sich ein nichtparametrisches Verfahren an. Die Analogie zur nichtparametrischen Schätzung von Dichtefunktionen über Kerndichteschätzungen, wie sie in den vorangehenden Abschnitten beschrieben wurde, ist offensichtlich. Anstelle der Dichte f ist nun die Funktion m zu schätzen. Aus

$$m(x) = \mathbf{E}(Y \mid X = x) = \frac{\int y f(x, y) dy}{f(x)} \tag{17}$$

und

$$\hat{f}_h(x, y) = n^{-1} \sum_{i=1}^{n} K_h(x - X_i) K_h(y - Y_i) \tag{18}$$

als Schätzung für die gemeinsame Dichte $f(x,y)$ erhält man mit

$$\int y\hat{f}_h(x,y)dy = n^{-1}\sum_{i=1}^{n} K_h(x-X_i)Y_i \tag{19}$$

die folgende von NADARAYA (1964) und WATSON (1964) vorgeschlagene Schätzung:

$$\hat{m}_h(x) = \frac{n^{-1}\sum_{i=1}^{n} K_h(x-X_i)Y_i}{n^{-1}\sum_{j=1}^{n} K_h(x-X_j)}. \tag{20}$$

Allgemeiner kann diese Schätzung auch in der Form

$$\hat{m}_h(x) = \sum_{i=1}^{n} W_{hi}(x)Y_i \tag{21}$$

geschrieben werden, d. h. die geschätzte Funktion $\hat{m}_h(\bullet)$ errechnet sich als Summe der mit gewissen Größen $W_{hi}(\bullet)$ bewichteten Y_i-Werte. Dabei werden Datenpunkte Y_i, deren zugehörige X_i-Werte nahe an x liegen, stärker bewichtet als „weiter entfernte" Y_i-Punkte.

Es gibt unter den nichtparametrischen Regressionsverfahren weitere, auf dem gleichen Prinzip beruhende Methoden wie z. B. *Nearest-Neighbour-Schätzungen*, *Splines*, *Polynomialregression* und *Wavelets*. Zu Wavelets soll hier nur auf die Literatur (MEYER, 1990) verwiesen werden.

Die Nearest-Neighbour-Schätzung (Nächste-Nachbarn-Schätzung) für $m(\bullet)$ verwendet zur Schätzung von $m(x)$ genau die k zu x nächsten Datenpunkte X_i, die alle mit gleichem Gewicht in die Schätzung eingehen, d. h. $\hat{m}_k(x) = \sum_{i=1}^{n} W_{ki}(x)Y_i$ mit

$$W_{ki}(x) = \begin{cases} \frac{1}{k} & \text{falls } X_i \text{ einer der } k \text{ nächsten Punkte ist,} \\ 0 & \text{sonst.} \end{cases} \tag{22}$$

Anstelle der Bandweite h ist in diesem Fall die Anzahl k der einbezogenen Nachbarn als Glättungsparameter zu schätzen. Die Spline-Schätzung $\hat{m}_\lambda(x)$ ergibt sich aus der Minimierung des Ausdrucks

$$\sum_{i=1}^{n}(Y_i - g(X_i))^2 + \lambda\int (g''(x))^2 dx \tag{23}$$

über eine Klasse von Funktionen g. Hier spielt der Parameter λ die Rolle des Glättungsparameters h bei den Kernschätzern. Die Lösung von (23) ist eine Funktion, die sich stückweise aus kubischen Polynomen zusammensetzt.

als Schätzung für die gemeinsame Dichte $f(x,y)$ erhält man mit

$$\int y\hat{f}_h(x,y)dy = n^{-1}\sum_{i=1}^{n} K_h(x-X_i)Y_i \tag{19}$$

die folgende von NADARAYA (1964) und WATSON (1964) vorgeschlagene Schätzung:

$$\hat{m}_h(x) = \frac{n^{-1}\sum_{i=1}^{n} K_h(x-X_i)Y_i}{n^{-1}\sum_{j=1}^{n} K_h(x-X_j)}. \tag{20}$$

Allgemeiner kann diese Schätzung auch in der Form

$$\hat{m}_h(x) = \sum_{i=1}^{n} W_{hi}(x)Y_i \tag{21}$$

geschrieben werden, d. h. die geschätzte Funktion $\hat{m}_h(\bullet)$ errechnet sich als Summe der mit gewissen Größen $W_{hi}(\bullet)$ bewichteten Y_i-Werte. Dabei werden Datenpunkte Y_i, deren zugehörige X_i-Werte nahe an x liegen, stärker bewichtet als „weiter entfernte" Y_i-Punkte.

Es gibt unter den nichtparametrischen Regressionsverfahren weitere, auf dem gleichen Prinzip beruhende Methoden wie z. B. *Nearest-Neighbour-Schätzungen*, *Splines*, *Polynomialregression* und *Wavelets*. Zu Wavelets soll hier nur auf die Literatur (MEYER, 1990) verwiesen werden.

Die Nearest-Neighbour-Schätzung (Nächste-Nachbarn-Schätzung) für $m(\bullet)$ verwendet zur Schätzung von $m(x)$ genau die k zu x nächsten Datenpunkte X_i, die alle mit gleichem Gewicht in die Schätzung eingehen, d. h. $\hat{m}_k(x) = \sum_{i=1}^{n} W_{ki}(x)Y_i$ mit

$$W_{ki}(x) = \begin{cases} \frac{1}{k} & \text{falls } X_i \text{ einer der } k \text{ nächsten Punkte ist,} \\ 0 & \text{sonst.} \end{cases} \tag{22}$$

Anstelle der Bandweite h ist in diesem Fall die Anzahl k der einbezogenen Nachbarn als Glättungsparameter zu schätzen. Die Spline-Schätzung $\hat{m}_\lambda(x)$ ergibt sich aus der Minimierung des Ausdrucks

$$\sum_{i=1}^{n}(Y_i - g(X_i))^2 + \lambda\int (g''(x))^2 dx \tag{23}$$

über eine Klasse von Funktionen g. Hier spielt der Parameter λ die Rolle des Glättungsparameters h bei den Kernschätzern. Die Lösung von (23) ist eine Funktion, die sich stückweise aus kubischen Polynomen zusammensetzt.

Abb. 12: Spline-Schätzung für die Motorrad-Daten

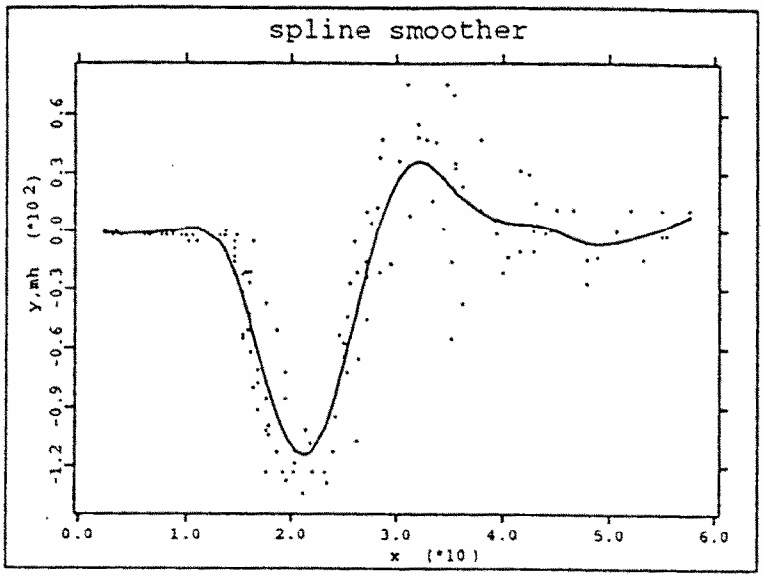
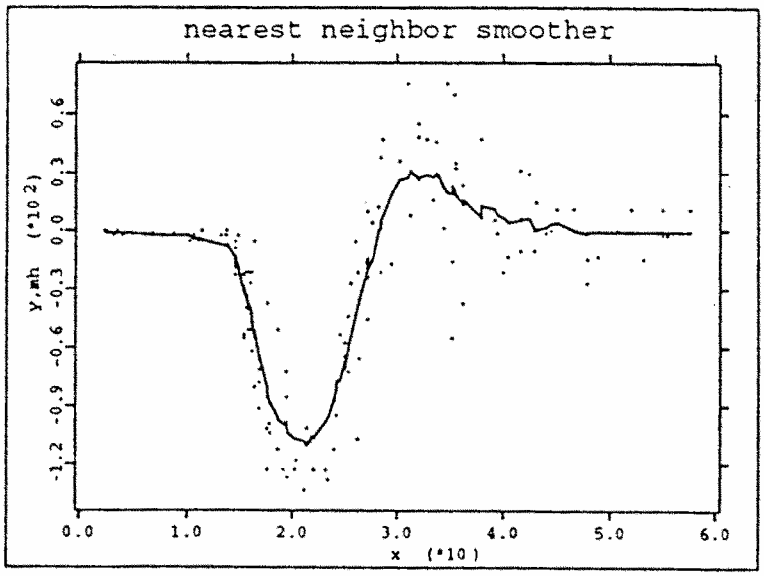


Abb. 13: Nearest-Neighbour-Schätzung für die Motorrad-Daten, $k = 15$



Die optimale Wahl des Glättungsparameters im Fall der nichtparametrischen Schätzung von Regressionsfunktionen kann analog zur Dichtenschätzung durch Plug-In- oder Kreuzvalidierungsmethoden erfolgen. Einen guten Überblick über diesen Fall geben insbesondere die Monographien von EUBANK (1988), HÄRDLE (1990, 1991), MÜLLER (1988) und SCOTT (1992).

Appendix

XploRe 3.0 ist ein interaktives, offenes Statistik-Programmsystem, das dem Nutzer erlaubt, mit einfachen Mitteln selbst komplexe statistische Berechnungen zu programmieren, in Programmbibliotheken einzubinden und damit

das System seinen eigenen Bedürfnissen optimal anzupassen. Insbesondere wird die Anwendung multivariater nichtparametrischer Methoden durch bereits vorgefertigte Macros und Macro-Bibliotheken unterstützt. Die XploRe-Programmiersprache ist matrixorientiert und verfügt über umfangreiche Möglichkeiten interaktiver Graphikdarstellung.

Programm 1. Histogramme für die Buffalo-Snowfall-Daten

```
proc ()=main()
  x=read(buffa)                    ; Einlesen der Buffalo-Snowfall-Daten
  library(smoother)                ; Laden der notwendigen Macro-Bibliothek
                                   ;
  fh1=histogrm(x 10 0)             ; Histogramm, Ursprung=0, Binweite=10
  fh2=histogrm(x 10 2)             ; Histogramm, Ursprung=2, Binweite=10
  fh3=histogrm(x 10 4)             ; Histogramm, Ursprung=4, Binweite=10
  fh4=histogrm(x 10 6)             ; Histogramm, Ursprung=6, Binweite=10
  fh5=histogrm(x 10 8)             ; Histogramm, Ursprung=6, Binweite=10
  fh6=denest(x 10)                 ; Dichteschaetzung, Bandweite=10
                                   ;
  createdisplay(fig2,2 3,s2d)      ; erzeugt gleichzeitige Graphik-
                                   ; darstellung in 6 Fenstern
                                   ;
  show(fh1 s2d1,fh2 s2d2,fh3 s2d3,fh4 s2d4,fh5 s2d5,fh6 s2d6)
                                   ;
                                   ; zeigt die 6 Funktionen in den
                                   ; 6 Fenstern
endp
```

Programm 2. "WARPed" Histogramm für Buffalo-Snowfall-Daten

```
proc ()=main()
  x=read(buffa)                    ; Einlesen der Buffalo-Snowfall-Daten
  library(smoother)                ; Laden der Macro-Bibliothek
                                   ;
  h=10                             ; Bandweite
  d=max(x)-min(x))./100            ; Binweite fuer kleine Bins
  (xb yb)=bindata(x d)             ; Diskretisieren der Daten
  wy=symweigh(0 d/h h/d &trian)    ; Berechnen der Gewichte fuer
                                   ; den Dreieckskern
  wx=aseq(0 rows(wy))
  (xc yc or)=conv((xb yb wx wy)    ; Dichteschaetzung als Faltung
  fh=(xc*d)~(yc/rows(x)*d)         ;
  show(fh s2d)                     ; graphische Darstellung
endp
```

Programm 3. Kerndichteschätzung nach der „quick & dirty“ Plug-in-Methode

```
; ***************************************************************************
; * DENAUTO macro ***********************************************************
; ***************************************************************************
; *                                                                         *
; * Example:  LIBRARY(smoother)                                             *
; *           x = READ(buffa)                                               *
; *           y = DENAUTO(x)                                                *
; *           gives the automatic density estimate using the quartic        *
; *           kernel for the buffalo snowfall data.                         *
; *                                                                         *
; ***************************************************************************
; ** Wolfgang Haerdle, 910426 ***********************************************
; ***************************************************************************
;
proc (fh)=denauto(x)
  error(cols(x)<>1 "DENAUTO: COLS(X) <> 1")
  d=(max(x)-min(x))./100
  n=rows(x)
  (xb yb)=bindata(x d)                ; bin data
  sigma=sqrt(cov(x))
  h=2.62*1.06*sigma*n^(-0.2)          ; determine h by rule of thumb
  wy=symweigh(0 d/h h/d &qua)         ; create weights
  wx=aseq(0 rows(wy))
  (xc yc or)=conv(xb yb wx wy)        ; calc density func
  fh=(xc*d)~(yc/(n*d))
endp
;
; ***************************************************************************
```

Programm 4. Macro zur Kerndichteschätzung

```
; ************************************************************************
; * DENEST macro *********************************************************
; ************************************************************************
; *                                                                      *
; * Example:  LIBRARY(smoother)                                          *
; *           x = NORMAL(100)                                            *
; *           y = DENEST(x 0.4)                                          *
; *           gives the kernel density estimate using the quartic        *
; *           kernel for the buffalo snowfall data using the             *
; *           bandwidth h=0.4                                            *
; *                                                                      *
; ************************************************************************
; ** Wolfgang Haerdle, 910426 ********************************************
; ************************************************************************
;
proc (fh)=denest(x h)
  error(cols(x)<>1 "DENEST: COLS(X) <> 1")
  d=(max(x)-min(x))./100
  error(h .<=d "DENEST: h smaller than d")
  (xb yb)=bindata(x d)                ; bin data
  wy=symweigh(0 d/h h/d &qua)         ; create weights
  wx=aseq(0 rows(wy))
  (xc yc or)=conv(xb yb wx wy)        ; calc density func
  fh=(xc*d)~(yc/(rows(x)*d))
endp
;
; ************************************************************************
```

Programm 5. Kreuzvalidierung

```
; ***************************************************************************
; * DENCVL macro ************************************************************
; ***************************************************************************
; *                                                                         *
; * Example:  LIBRARY(smoother)                                             *
; *           x = READ(buffa)                                               *
; *           y = DENCVL(x)                                                 *
; *           gives the cross validation function using the quartic         *
; *           kernel for the buffalo snowfall data                          *
; *                                                                         *
; ***************************************************************************
; ** Wolfgang Haerdle, 910709 ***********************************************
; ***************************************************************************
;
proc (cv)=dencvl(x)
  error(cols(x)<>1 "DENCVL: COLS(X) <> 1")
  d=(max(x)-min(x))./100
  n=rows(x)
  (xb yb)=bindata(x d)              ; bin data
  wx=0
  wy=1
  (xc nz or)=conv(xb yb wx wy)      ; nz contains # pts in bin
  m=3
  h=m*d
  cv=matrix(16 2)
  while (m<=33)
    h=h*1.2
    wy=symweigh(0 d/h m &qua)       ; create weights
    wx=aseq(0 rows(wy))
    (xc yc or)=conv(xb yb wx wy)    ; calc density func
    fh=yc/(n*d)
    wm0=15*m^4/(16*m^4-1)
    i=(m-1)/2
    cv[i,1]=h
    cv[i,2]=d.*(fh'*fh)-2.*nz'*fh/(n-1)+2*wm0/((n-1)*h)
    m=m+2
  endo
endp
;
; ***************************************************************************
```

Programm 6. Konfidenzintervalle für Nettoeinkommensdaten

```
proc ()=main()
  x=read(netinc)                          ; Einlesen der Nettoeinkommens-Daten
  library(smoother)                       ; Laden der Macro-Bibliothek
                                          ;
  h=0.15                                  ; Bandweite
  fh=denest(x h)                          ; Dichteschaetzung, h=0.15
  ci=2*sqrt((5/7).*fh[,2]./(rows(x)*h))
  cup=fh[,1]~(fh[,2].+ci)                 ; obere Konfidenzgrenzen
  clo=fh[,1]~(fh[,2].-ci)                 ; untere Konfidenzgrenzen
  show(fh s2d)                            ; graphische Darstellung
endp
```

Programm 7. Kernschätzung für Regressionsfunktion

```
; ***************************************************************************
; * REGEST macro ************************************************************
; ***************************************************************************
; *                                                                         *
; * Example:  LIBRARY(smoother)                                             *
; *           x = READ(geyser)                                              *
; *           y = REGEST(x 0.4)                                             *
; *           gives the kernel regession estimate using the quartic         *
; *           kernel for the geyser data using the bandwidth h=0.4          *
; *                                                                         *
; ***************************************************************************
; ** Wolfgang Haerdle, 910426 ***********************************************
; ***************************************************************************
;
proc (mh)=regest(x h)
  error(cols(x).<>2 "REGEST: COLS(X) <> 2")
  d=(max(x[,1])-min(x[,1])).*/100
  error(h.<=d  "REGEST: h smaller than d")
  (xb yb)=bindata(x[,1] d 0 x[,2])    ; bin data in x and sum of y's
  wy=symweigh(0 d/h h/d &qua)         ; create weights for quartic kernel
  wx=aseq(0 rows(wy))
  (xc yc or)=conv(xb yb wx wy)        ; smooth x's and y's
  y=(d.*xc)~(yc[,2]./yc[,1])
  mh=paf(y y[,2].<>NAN)               ; remove missings at the end
endp

; ***************************************************************************
```

Programm 8. Nearest-Neighbour-Schätzung für Regressionsfunktion

```
; ************************************************************************
; * KNN macro ************************************************************
; ************************************************************************
; *                                                                      *
; * Example:  LIBRARY(smoother)                                          *
; *           x = READ(geyser)                                           *
; *           y = KNN(x 11)                                              *
; *           gives the kNN regression estimate for the geyser data      *
; *           using the parameter k=11.                                  *
; *                                                                      *
; ************************************************************************
; ** Sigbert Klinke, 901012 ************************************************
; ************************************************************************
;
proc (y)=knn(x k)
  error(cols(x)<>2 "KNN: COLS <> 2")
  z=sort(x)
  a=z[,1]
  b=z[,2]
  y=z[,1]~sknn(a b k)
endp
;
; ************************************************************************
```

Literatur

BICKEL, P.J. / ROSENBLATT, M. (1973): On some global measures of the deviations of density function estimates; *Annals of Statistics* **1**, 1071-1095.

EUBANK, R.L. (1988): *Spline smoothing and nonparametric regression.* Marcel Dekker, New York.

Family expenditure survey, Annual Base Tapes (1968-1983). Department of Employment, Statistics Division, Her Majesty's Stationary Office, London. (zur Verfügung gestellt von ESRC Data Archive, University of Essex)

HÄRDLE, W. (1990): *Applied nonparametric regression.* Cambridge University Press, Cambridge.

HÄRDLE, W. (1991): *Smoothing Techniques. With Applications in S.* Springer, New York.

HÄRDLE, W. / HALL, P. / MARRON, J.S. (1988): How far are automatically chosen regression smoothing parameters from their optimum? (with discussion). *Journal of the American Statistical Association*, **83**, 86-99.

HALL, P. (1982): Cross-validation in density estimation. *Biometrika*, **69**, 383-390.

HALL, P. (1983): Large sample optimality of least-squares-cross-validation in density estimation. *Annals of Statistics*, **11**, 1156-1174.

HALL, P. / MARRON, J.S. (1991): Lower bounds for bandwidth selection in density estimation. *Probability Theory and Related Fields*, **90**, 149-173.

HALL, P. / SHEATHER, S.J. / JONES, M.C. / MARRON, J.S. (1991): On optimal data-based bandwidth selection in kernel density estimation. *Biometrika*, **78**, 263-269.

MARRON, J.S. (1988): Automatic smoothing parameter selection: a survey. *Empirical Economics*, **13**, 187-208.

MARRON, S. / NOLAN, D. (1989): Canonical kernels for density estimation. *Statistics and Probability Letters*, **7**, 191-195.

MÜLLER, H.-G. (1988): *Nonparametric regression analysis of longitudinal data.* Springer, Berlin.

MEYER, Y. (1990): *Ondelettes et Operateurs I. Ondelettes.* Hermann, Paris.

PARK, B. / MARRON, J.S. (1990): Comparison of data-driven bandwidth selectors. *Journal of the American Statistical Association*, **85**, 66-72.

PARK, B. / TURLACH, B. (1992): Practical performance of several data driven bandwidth selectors (with discussion). *Computational Statistics*, **7**, 251-271.

SCOTT, D. (1992): *Multivariate density estimation.* Wiley, New York.

SILVERMAN, B.W. (1985): Some aspects of the spline smoothing approach to nonparametric regression curve fitting (with discussion). *Journal of the Royal Statistical Society, Series B*, **47**, 1-52.

SILVERMAN, B.W. (1986): *Density estimation for statistics and data analysis.* Chapman and Hall, London.

TURLACH, B. (1992): *On bandwidth selection in kernel density estimation.* Diplomarbeit, C.O.R.E., Universität Louvain-la-Neuve, Belgien.

XploRe (1992): *XploRe 3.0 - an interactive statistical computing environment.* Beziehbar von Xplore Systems, Institut für Statistik und Ökonometrie, Wirtschaftswissenschaftliche Fakultät der Humboldt-Universität Berlin.

Prof. Dr. Wolfgang Härdle
Dipl.-Math. Marlene Müller
Institut für Statistik und Ökonometrie
Humboldt-Universität

Spandauer Str. 1
O-1020 Berlin

Statistica Neerlandica (1993) Vol. 47, nr. 1, pp. 43-57

On the backfitting algorithm for additive regression models

W. Härdle

Wirtschaftswissenschaftliche Fakultät
Humboldt Universität zu Berlin, D 1020 Berlin, Germany

P. Hall

Department of Mathematics
Australian National University,
Canberra, ACT 2601, Australia

We analyse additive regression model fitting via the backfitting algorithm. We show that in the case of a large class of curve estimators, which includes regressograms, simple step-by-step formulae can be given for the backfitting algorithm. The result of each cycle of the algorithm may be represented succinctly in terms of a sequence of d projections in n-dimensional space, where d is the number of design coordinates and n is sample size. It follows from our formulae that the limit of the algorithm is simply the projection of the data onto that vector space which is orthogonal to the space of all n-vectors fixed by each of the projections. The formulae also provide the convergence rate of the algorithm, the variance of the backfitting estimator, consistency of the estimator, and the relationship of the estimator to that obtained by directly minimizing mean squared distance.

Key Words: Gauss-Seidel algorithm, generalized additive models, numerical analysis, projection, backfitting.

1 Introduction

It is well known that regression smoothing in high dimensions faces the problem of data sparseness. Additive models have been introduced to overcome this problem. The idea of additive model fitting is to approximate in a regression relation $Y = g(X) + \text{error}$, $X \in \mathbb{R}^d$, the *high dimensional* function $g : \mathbb{R}^d \to \mathbb{R}$ by a sum of *one-dimensional* functions $g_j : \mathbb{R} \to \mathbb{R}$, i.e.

$$g(x) \simeq \mu + \sum_{j=1}^{d} g_j(x_j). \tag{1}$$

If the regression function g lies in this class of additive separable functions so that the approximation in (1) is in fact an equality for some $\{g_j\}_{j=1}^{d}$, then better convergence rates for estimating g are obtainable (Stone, 1985). In order to obtain this dimension reduction one needs to find the additive approximation in (1).

In the numerical literature the algorithm is usually called the Gauss-Seidel iterative method (RALSTON, 1965; GOLUB and VAN LOAN, 1983, Ch. 10). Recent work on this dimension reduction principle and generalized additive models (FRIEDMAN and STUETZLE, 1981; BREIMAN and FRIEDMAN, 1985; HASTIE and TIBSHIRANI, 1986, 1987) promoted the use of the backfitting algorithm. Establishing convergence of the algorithm in cases of statistical interest has proved quite difficult (BUJA, HASTIE and TIBSHIRANI, 1989, section 3.4).

In the present paper we analyse a class of curve estimators, for which the backfitting algorithm can easily be studied step by step.

The backfitting algorithm prescribes iterated fitting of residuals from earlier steps over components of the x-vector until convergence is reached. This can be quite a computer-intensive task if the dimension d is high. Therefore, fast one-dimensional regression smoothers are necessary as elementary building blocks of this algorithm. Kernel smoothers usually require a number of operations much larger than sample size, n. A faster method based on discretization into bins (like a regressogram) and weighting these bins (also called WARPing) has been proposed by HÄRDLE and SCOTT (1990). In this paper we give insight into the functioning of such estimators inside the backfitting procedure.

We demonstrate that the result of each step of the algorithm may be represented in terms of a sequence of projections $H_1, \ldots, H_d$ on n-dimensional space. This representation addresses technical problems reported in the above papers, arising from asymmetry and non-invertibility of statistical smoothing operators. The limit of the algorithm equals the projection P of the data onto that vector space orthogonal to the space of all n-vectors which are held fixed by each H_j. The rate of the convergence of the algorithm after ν cycles is $O(|\lambda_0|^{\nu})$, where $|\lambda_0|$ denotes the largest non-unit absolute eigenvalue of the product of the projections H_j. The sum of the variances of the backfitting estimator equals $(\mathrm{tr}P - 1)\sigma^2 = (\Sigma m_j - d)\sigma^2$, where σ^2 is the error variance and m_j is the number of knots of the fitted estimator on the jth coordinate.

The projections H_j depend on the order of the regressogram and the number of knots. In a simple case, to be discussed below, H_j may be expressed in terms of cell frequencies. After ν cycles of the backfitting algorithm we obtain the following expression for the backfitting estimator at the ith design point and in the jth coordinate:

$$\sum_{r=0}^{\nu-1} \{Z^T (H_1 \ldots H_d)^r H_1 \ldots H_j (I - H_j)\}_i, \tag{2}$$

where $Z = (Y_i - \bar{Y})$ is the column vector of centred data values. This series converges as $\nu \to \infty$, and the sum over j of the limit equals $(PZ)_i$ where P is the projection defined earlier. We shall show that, under a mild additional condition, this quantity is the one which would be obtained by directly minimizing the squared-error distance of an additive model from the centred data values $Y_i - \bar{Y}$, and prove that it is consistent for the unique additive approximation which minimizes mean squared error.

To make our account of the algorithm reasonably self-contained, we shall outline the operation of the backfitting method in Section 2. Section 3 will analyze the method, and

draw the conclusions described above. Section 4 will give proofs of two theorems from Section 3.

2 The model and the backfitting algorithm

Assume that $(d+1)$-variate data (x_i, Y_i) are observed, with $x_i = (x_{i1}, ..., x_{id})$. We suppose that $x_1, ... x_n$ represent independent observations of a random vector $X = (X_1, ..., X_d)$, and that x_i and Y_i are connected by the model

$$Y_i = g(x_i) + e_i, \qquad 1 \le i \le n,$$

where g is an unknown smooth function from $\mathbb{R}^d$ to $\mathbb{R}$, and $e_1, ..., e_n$ are observation errors. It would be common to assume that the e_i's were independent and identically distributed with zero mean, but distributional assumptions such as this are not required for most of our work.

We seek an additive approximation to g, of the form

$$g(X) \simeq \mu + \sum_{j=1}^{d} g_j(X_j),$$

where μ is a constant (usually $Eg(X)$) and the g_j's are smooth univariate functions. If the g_j's were linear then the approximation would be one of linear regression, but we shall make no structural assumptions about either g_j or g. It causes no ambiguity to write $g_j(X_j)$ as $g_j(X)$, and so we shall usually adopt this simpler notation.

The problem of choosing the approximating functions g_j is identifiable if we demand that

$$E\{g_j(X)\} = 0, \qquad 1 \le j \le d, \tag{3}$$

and that $g_1, ..., g_d$ be chosen to minimize $E\{g_j(X) - \mu - \Sigma g_j(X)\}^2$ subject to (3); see STONE (1985) and HASTIE and TIBSHIRANI (1986). The backfitting algorithm may be thought as an attempt to estimate $g_1, ..., g_d$, or more importantly to estimate Σg_j. The limit of the algorithm applied to the jth coordinate, assuming the algorithm converges, is an estimator $\hat{g}_j$ of g_j.

The backfitting algorithm is started with all functions g_j set equal to zero. Suppose that after s steps we have obtained the approximation $\hat{g}_{j(s)}(x_i)$ to $\hat{g}_j(x_i)$, for $1 \le i \le n$ and $1 \le j \le d$, and have just updated $\hat{g}_{j0(s-1)}$ to $\hat{g}_{j0(s)}$. If $1 \le j_0 \le d-1$ then we next update $\hat{g}_{j0+1(s)}$ to $\hat{g}_{j0+1(s+1)}$, choosing to minimize

$$\sum_{i=1}^{n} \left\{ Y_i - \bar{Y} - \sum_{j=1}^{j_0} \hat{g}_{j(s)}(x_i) - \hat{g}_{j0+1(s+1)}(x_i) - \sum_{j=j_0+2}^{d} \hat{g}_{j(s)}(x_i) \right\}^2 \tag{4}$$

subject to

$$\sum_{i=1}^{n} \hat{g}_{j0+1(s+1)}(x_i) = 0.$$

This is the backfitting step since we are regressing on residuals from earlier steps. Minimization of (4) can be done with different techniques, such as that $\hat{g}_{j0+1(s+1)}$ be a running lines smoother (HASTIE and TIBSHIRANI, 1986), or a spline of specified type. (If $j_0 = d$, take $j_0 = 0$ in (4) and (5).) The $(s+1)$'st approximation is finally

$$\hat{g}_{j(s+1)} = \begin{cases} \hat{g}_{j(s)} & \text{if } j \neq j_0 + 1 \\ \hat{g}_{j0+1(s+1)} & \text{if } j = j_0 + 1. \end{cases} \tag{5}$$

Often the criterion for smoothing does not change from one step of the algorithm to the next. If we are fitting a univariate regressogram to each coordinate, then the centres of the regressogram blocks will most likely stay fixed throughout the algorithm; if we are fitting a univariate histospline, then the positions of knots in the splines will be unchanged. In such cases, it is particularly convenient to think of each step in the algorithm as supplying an additive correction γ_j as follows. Take

$$U_i = Y_i - \bar{Y} - \sum_{j=1}^{d} \hat{g}_{j(s)}(x_i),$$

and choose the univariate function γ_j in the specified way, for example to be a regressogram with specified block centres, to minimize

$$\sum_{i=1}^{n} \{U_i - \gamma_j(x_{ij})\}^2 \text{ subject to } \sum_{i=1}^{n} \gamma_j(x_{ij}) = 0. \tag{6}$$

Let $\hat{\gamma}_j$ denote the result of this procedure. Then a prescription equivalent to (5) is

$$\hat{g}_{j(s+1)} = \hat{g}_{j(s)} + \delta_{jj0}\hat{\gamma}_{j0},$$

where δ_{ij0} is the Kronecker delta.

The backfitting algorithm starts with all functions set equal to zero. At the jth step, $1 \leq j \leq d$, we obtain an approximation $\hat{g}_{j1}$ to $\hat{g}_j$. In terms of the functions $\hat{g}_{j1}$, and for $1 \leq s \leq d$, we have $\hat{g}_{j(s)} = \hat{g}_{j1}$ if $1 \leq j \leq s$, $\hat{g}_{j(s)} = 0$ otherwise. At the $(d+j)$th step, $1 \leq j \leq d$, we obtain a function $\hat{\gamma}_j = \hat{g}_{j2}$, and for $d < s \leq 2d$ our approximation to $\hat{g}_j$ is $\hat{g}_{j(s)} = \hat{g}_{j1} + \hat{g}_{j2}$ if $d + j \leq s$, $\hat{g}_{j(s)} = \hat{g}_{j1}$ otherwise. After $s = vd + l$ steps, where $-(d-j) \leq l < j$, our approximation to $\hat{g}_j$ is

$$\hat{g}_{j(s)} = \sum_{r=1}^{v} \hat{g}_{jr};$$

our approximation to $\Sigma\hat{g}_j$ after $vd + l$ steps, for $0 \leq l \leq d$ and $v \geq 0$, is

$$\sum_{j=1}^{d} \hat{g}_{j(s)} = \sum_{r=1}^{v} \sum_{j=1}^{d} \hat{g}_{jr} + \sum_{j=1}^{l} \hat{g}_{j,v(1}.$$

In the next section we develop explicit formulae for the successive additive corrections $\hat{g}_{jr}$.

3 Analysis of the backfitting algorithm

3.1 *Summary*

In Section 3 we give a careful and detailed account of convergence properties of the backfitting algorithm, in the case where the basic estimators are regression versions of histosplines with given knots. In fact, our estimator class is even larger than this – we consider estimators which are linear in the response variable, with the weights chosen by backfitting. Section 3.2 defines these estimators. Section 3.3 defines certain projection matrices K_j in terms of which the result of each step in the backfitting algorithm may be written down explicitly. A formula for the limit of the algorithm is given in Section 3.4, both for the individual coordinates and for the sum of the coordinatewise functions. Section 3.5 shows that the backfitting algorithm is equivalent to directly minimizing the sum of squared errors of an additive approximation, and Section 3.6 points out that the rate of convergence of the backfitting algorithm is geometrically fast. Finally, Section 3.7 derives very simple formulae for the variance of the backfitting estimator, and shows that the estimator is mean-square consistent.

3.2 *Construction of estimators*

We begin by considering the regressogram, which we assume is constructed for the jth coordinate using blocks over consecutive intervals B_{jk}, $1 \le k \le m_j$. Such a function is given by

$$r_j(u) = \sum_{k=1}^{m_j} c_{jk} I(u \in B_{jk}),$$

where c_{jk} denotes the height of the regressogram block on interval B_{jk}, $I(\cdot)$ is the usual indicator function and m_j denotes the number of blocks in the jth coordinate. The value of r_j at the jth coordinate of x_i is

$$r_j(x_{ij}) = \sum_{k=1}^{m_j} c_{jk} J_{jk}(i), \tag{7}$$

where $J_{jk}(i) = 1$ if $x_{ij} \in B_{jk}$ and $J_{jk}(i) = 0$ otherwise.

If r_j is constructed by interpolating among neighbouring regressogram blocks, instead of assuming a fixed value over the kth interval B_{jk}, then formula (7) continues to hold although with an altered definition of $J_{jk}(i)$. For example, if the intervals B_{jk} are of constant width h_j, if b_{jk} denotes the centre of B_{jk}, and if r_j is obtained by interpolating linearly between centres of consecutive blocks, then (7) remains valid provided we redefine

$$J_{jk}(i) = h_j^{-1}\{I(b_{j,k-1} < x_{ij} \le b_{jk})(x_{ij} - b_{j,k-1}) = I(b_{jk} < x_{ij} \le b_{j,k+1})(b_{j,k+1} - x_{ij})\}$$

for $2 \le k \le j-1$, with similar formulae for $k = 1$ and $k = m_j$. Here it is assumed that $b_{j1} \le \ldots \le b_{j,m_j}$. The regressogram is a first-order spline, the linearly interpolated

regressogram is a second-order spline with knots at the points b_{jk}, and so on. Such estimators are analogues of the well-known histosplines in nonparametric density estimation.

We assume that the curve estimate fitted to the jth coordinate is of the form (7). In this prescription of the "basis-functions" $J_{jk}(i)$ may depend on the data in any manner, but they should be chosen outside the backfitting algorithm. Only the c_{jk}'s are subject to selection by the algorithm. Define J_{jk} to be the column vector of length n with ith element $J_{jk}(i)$. We assume that

$$\textit{for each } j\textit{, the vectors } J_{jk},\ 1 \le k \le m_j,\ \textit{are linearly independent.} \tag{8}$$

This condition is satisfied in cases of interest, such as the regressogram, and polynomial interpolations of the regressogram.

3.3 *Definition of matrices K_j*

Let A_{jk}, $1 \le k \le m_j$, be n-vectors such that

$$c_{jk} = z^T A_{jk}$$

gives a minimum of

$$S = \sum_{i=1}^{n} \left\{ z_i - \sum_{k=1}^{m_j} c_{jk} J_{jk}(i) \right\}^2, \tag{9}$$

where $z = (z_1, \ldots, z_n)^T$. Define

$$K_j = \sum_{k=1}^{m_j} A_{jk} J_{jk}^T, \tag{10}$$

an $n \times n$ matrix. (The matrix H_j discussed in Section 1 is simply $I - K_j$.) The normal equations which any minimizer of S must satisfy are

$$\sum_{k=1}^{m_j} c_{jk} J_{jk}^T J_{jl} = z^T J_{jl}.$$

Substituting from (9), and using (10), we deduce that $z^T K_j J_{jl} = z^T J_{jl}$. Since this must be true for all z then $K_j J_{jl} = J_{jl}$, $1 \le l \le m_j$.

Thus, K_j has m_j linearly independent eigenvectors each with eigenvalue 1. By the definition of K_j, the rank of K_j can be at most m_j. Therefore all other eigenvalues of K_j are zero, repeated $n - m_j$ times. It follows that K_j is idempotent, and hence is symmetric. In particular, $J_{jl}^T K_j = J_{jl}^T$, $1 \le l \le m_j$.

In the case of the regressogram, the (i_1, i_2) element of the projection operator K_j is

$$K_j(i_1, i_2) = v_{jk_j(i_1)}^{-1} I\{k_j(i_1) = k_j(i_2)\},$$

where $k_j(i)$ denotes the index k of the interval B_{jk}, $1 \le k \le m_j$, into which the jth coordinate of x_i falls, and v_{jk} denotes the number of x_i's, $1 \le i \le n$, which fall into B_{jk}.

3.4 *The limit of the backfitting algorithm*

Let V denote the space of column vectors u such that $K_j u = 0$ for each j, $1 \le j \le d$, and let $V_\perp$ be the space complementary to V, of dimension n minus the dimension of V. Write P for the matrix of the projection of $\mathbb{R}^n$ onto V, and put $Z^T = (Z_1, ..., Z_n)$ where $Z_i = Y_i - \bar{Y}$. Our first theorem shows that the limit of the backfitting algorithm equals the projection of Z onto $V_\perp$.

THEOREM 1. *Assume condition (8). Then the s-step backfitting approximation to* $g(x_i) - \mu$ *converges, as* $s \to \infty$, *to*

$$\sum_{j=1}^{d} \hat{g}_j(x_i) = (PZ)_i. \tag{11}$$

The individual components of the approximation are given by

$$\hat{g}_j(x_i) = \left(Z^T \left[\sum_{r=0}^{\infty} \{(I - K_1) \dots (I - K_d)\}^r (I - K_1) \dots (I - K_{j-1}) K_j \right] \right)_i, \tag{12}$$

in which the infinite series converges absolutely.

Note that P is defined only in terms of the collection of transformations $K_1, ..., K_d$, and does not depend on the order of this sequence. It follows that the limit $\Sigma_j \hat{g}_j(x_i)$ of the algorithm does not depend on the order in which we listed the coordinates when embarking on the algorithm. It is not clear that $\hat{g}_j(x_i)$ does not depend on the initial order of the coordinate list, although it will be invariant if the constrained minimizer of $S' = \Sigma_i \{Z_i - \Sigma_j \Sigma_k c_{jk} J_{jk}(i)\}^2$ is uniquely defined. That requires a more stringent condition than (8); see (15) below. When (15) holds, invariance of $\hat{g}_j(x_i)$ under different orderings of the initial coordinate list follows from the fact that the $\hat{g}_j(x_i)$'s uniquely minimize S'.

The overall additive approximation to $g(x_i)$ is of course

$$\bar{Y} + \sum_{j=1}^{d} \hat{g}_j(x_i) = \bar{Y} + \{P(Y - \bar{Y}1)\}_i,$$

where $Y^T = (Y_1, ..., Y_n)$ and $1^T = (1, ..., 1)$; here we have used (11). In most cases of interest, in particular for histospline estimators, we have $1 = \Sigma_k J_{jk}$ for each j (see the discussion following (15) below), and so $1 \in V_\perp$. Therefore $P1 = 1$, and so our formula for the additive approximation to $g(x_i)$ reduces to $(PY)_i$ – that is, to the ith coordinate of the projection of the data onto $V_\perp$.

In principle the backfitting algorithm could be circumvented by direct calculation of P, by first constructing a basis of the vector space $V_\perp$ which is generally of dimension $\Sigma m_j - d + 1$. However, from a computational point of view the backfitting algorithm is usually easier to implement since no large matrices have to be inverted.

3.5 *Direct minimization*

The backfitting algorithm is a sequence of componentwise fits of the marginal functions r_j; defined at (7), subject to the constraints

$$\sum_{i=1}^{n} \sum_{k=1}^{m_j} c_{jk} J_{jk}(i) = 0, \qquad 1 \le j \le d. \tag{13}$$

We could instead attempt to fit directly an additive model in which the jth component had the form (7) and the optimization was constrained by (13). In other words, we could try to minimize

$$S' = \sum_{i=1}^{n} \left\{ Z_i - \sum_{j=1}^{d} \sum_{k=1}^{m_j} c_{jk} J_{jk}(i) \right\}^2 \tag{14}$$

subject to (13); here $Z_i = Y_i - \bar{Y}$. The solution to this problem is unique, provided we strengthen condition (8) to *the vectors* J_{jk}, $1 \le k \le m_j$, $1 \le j \le d$, *generate a vector space of dimension* $\mathrm{S}m_j - d + 1$, *the extra constraints on the vectors being*

$$\mathrm{S}_{k=1}^{m_j} J_{jk} = 1 \equiv (1, \ldots, 1)^T, \qquad 1 \le j \le d. \tag{15}$$

Again this condition is trivially satisfied in cases of interest, the additional constraints following from the fact that the estimator r_j at (7) would generally be constructed to take the value c when each c_{jk} was equal to c. It is a straightforward matter to check that (15) holds for the regressogram and for polynomial interpolations of the regressogram, provided only that the extreme edges of the interpolation are defined appropriately.

Of course, the problem of minimizing S' without constraints does not admit a unique solution. Nevertheless, any one of the solutions produces the same minimum, and the solution subject to (13) is identical to the estimator produced by the backfitting algorithm, as our next theorem shows.

THEOREM 2. *Assume condition (15). Then the minimum of (9) subject to (13) is unique, and is identical to the solution obtained by the backfitting algorithm. That is, if* $\mathrm{S}_j \mathrm{S}_k \bar{c}_{jk} J_{jk}(\cdot)$ *minimizes (9) subject to (13), then* $\mathrm{S}_j \mathrm{S}_k \bar{c}_{jk} J_{jk}(i) = \mathrm{S}_j \hat{g}_j(x_i)$, *the latter given by (12). Furthermore, the global minimum of (14) is identical to the constrained minimum subject to (13).*

If one attempts to minimize S' subject to (13) in the usual manner, using Lagrange multipliers, one obtains the equivalent problem of minimizing

$$\sum_{i=1}^{n} \left\{ Z_i - \sum_{j=1}^{d} \sum_{k=1}^{m_j} c_{jk} J_{jk}(i) \right\}^2 + \sum_{j=1}^{d} \lambda_j \sum_{i=1}^{n} \sum_{k=1}^{m_j} J_{jk}(i).$$

After a little algebra it emerges that each of the multipliers must be zero, which is to be expected given that the unconstrained minimum is identical to the constrained minimum.

When condition (15) holds, the vector space $V_\perp$ introduced in Section 3.4 may be

represented as the set of all linear combinations of the vectors J_{jk}, $1 \leq k \leq m_j$ and $1 \leq j \leq d$. A proof is given in Appendix (1).

3.6 *Convergence rate*

The rate of convergence of the backfitting algorithm, for example the rate of convergence of the infinite series in (12), is geometrically fast. This follows from the fact that any eigenvalue of a product of projections is either equal to 1, or strictly less than 1 in absolute value. To see how this applies to (12), consider the infinite series

$$\sum_{r=0}^{\infty} z^T M^r (I - K_1) \ldots (I - K_{j-1}) K_j, \tag{16}$$

where $M = (I - K_1) \ldots (I - K_d)$ and z^T is a row vector. If z^T is a left eigenvector of M corresponding to an eigenvalue λ whose absolute value is strictly less than 1, the rate of convergence in (16) is $O(|\lambda|^r)$. If the eigenvalue equals 1 then z^T is a fixed point of each of the projections $I - K_j$, and so $z^T K_j = 0$. It follows that each term in (16) vanishes. Hence for a general z, not necessarily an eigenvector of M, the rate of convergence on (16) equals $O(|\lambda_0|)^r)$ where λ_0 is the eigenvalue of M whose absolute value is largest, subject to being strictly less than 1.

3.7 *Distribution of the backfitting estimator*

Except for very special cases, the expected values of $\hat{g}_j(x_i)$ and $\Sigma_j \hat{g}_j(x_i)$ cannot be simplified much beyond the obvious formulae obtainable from (11) and (12). However, the variances can be simplified. In particular, it is generally true that

$$\operatorname{cov}\left\{\sum_{j=1}^{d} \hat{g}_j(x_{i_1}), \sum_{j=1}^{d} \hat{g}_j(x_{i_2})\right\} = \{P(i_1, i_2) - n^{*1}\}\sigma^2, \tag{17}$$

where $P(i_1, i_2)$ denotes the (i_1, i_2)th element of P and $\sigma^2 = \operatorname{var}(Y_i)$.

To appreciate why, observe both $\hat{g}_j(x_i)$ and $\Sigma_j \hat{g}_j(x_i)$ have the form $(Z^T Q)_i$, where Q is an $n \times n$ matrix depending only on the J_{jk}'s. Generally the vectors J_{jk} will depend only on the design points x_i; not on the Y_i's. In this event, assuming that the Y_i's are uncorrelated with common variance σ^2,

$$\begin{aligned} \sigma^{*2} \operatorname{cov}\{(Z^T Q)_{i_1}, (Z^T Q)_{i_2}\} &= \{Q^T (I - n^{*1} 1 1^T) Q\}_{i_1 i_2} \\ &= (Q^T Q)_{i_1 i_2} - n^{*1} (Q^T 1)_{i_1} (Q^T 1)_{i_2}. \end{aligned} \tag{18}$$

In the case of $\Sigma_j \hat{g}_j(x_i)$ we have $Q = Q^T = P$. When (15) holds, $P1 = 1$ and also $P^2 = P$ since P is a projection. Result (17) now follows from (18).

One consequence of (17) is that the sum of the variances of $\Sigma_j \hat{g}_j(x_i)$ satisfies

$$E\left[\sum_{i=1}^{n}\left\{\sum_{j=1}^{d} \hat{g}_j(x_i) - E\sum_{j=1}^{d} \hat{g}_j(x_i)\right\}^2\right] = (\operatorname{tr} P - 1)\sigma^2 = \left(\sum_{j=1}^{d} m_j - d\right)\sigma^2$$

the second identity following from (15).

Variance formulae analogous to (18) may be written down for the result of stopping the backfitting algorithm after a given number of steps. From those results it is clear that even the first-order asymptotics (as $n \to \infty$) of the estimator change if the estimator is stopped short rather than run to its limit. Thus, the algorithm does not converge, in some asymptotic sense, after a fixed number of steps.

STONE (1985) proved that under general conditions, for any given constant μ, there exist uniquely determined univariate, continuous functions $g_1, \ldots, g_d$ such that $E\{g(X) - \mu - \Sigma_j g_j(X_j)\}^2$ is minimized subject to $Eg_j(X_j) = 0$ for $1 \le j \le d$. Assuming that those conditions are satisfied, it is a straightforward matter to prove that if $\mu = Eg(X)$ then the backfitting estimator $\Sigma_j \hat{g}_j$ is consistent to $\Sigma_j g_j$, in the sense that

$$n^{-1} \sum_{i=1}^{n} \left[\sum_{j=1}^{d} \{\hat{g}_j(x_i) - g_j(x_i)\} \right]^2 \to 0$$

in probability. We shall give an outline of the proof in Appendix (2).

4 Proofs

PROOF OF THEOREM 1

The proof of Theorem 1 is given in three parts.

4.1 *A general step in the backfitting algorithm*

In the backfitting algorithm described at (6), take γ_j to be the function r_j of (7) with as yet undetermined weights c_{jk}. For given j we wish to choose the c_{jk}'s to minimize

$$\sum_{i=1}^{n} \left\{ U_i - \sum_{k=1}^{m_j} c_{jk} J_{jk}(i) \right\}^2 \text{ subject to } \sum_{i=1}^{n} \sum_{k=1}^{m_j} c_{jk} J_{jk}(i) = 0. \tag{19}$$

Application of the method of Lagrange multipliers produces the equations

$$\sum_{k=1}^{m_j} \hat{c}_{jk} J_{jk}^T J_{jl} = W^T J_{jl}, \qquad 1 \le l \le m_j,$$

where $W = (W_i)$ and $W_i = U_i - \bar{U}$. We know from (9) that the solution of these equations is given by

$$\hat{c}_{jk} = W^T A_{jk}, \qquad 1 \le k \le m_j. \tag{20}$$

The resulting estimate of γ_j is

$$\hat{\gamma}_j(x_i) = \sum_{k=1}^{m_j} \hat{c}_{jk} J_{jk}(i) = \sum_{k=1}^{m_j} W^T A_{jk} J_{jk}(i),$$

that is

$$\hat{\gamma}_j(x_{i_0}) = (W^T K_j)_i; \qquad 1 \le i_0 \le n. \tag{21}$$

Here and below we write $\hat{\gamma}_j(x_i), \hat{g}_j(x_i), \ldots$ for $\hat{\gamma}_j(x_{ij}), \hat{g}_j(x_{ij}), \ldots$, when no ambiguity can arise.

4.2 *An arbitrary number of steps in the backfitting algorithm*

In the first step of the algorithm we take $U_i = Y_i - \bar{Y}$ and $j = 1$ in (21), obtaining for $\hat{\gamma}_1(x_i)$:

$$\hat{g}_{11}(x_i) = (Z^T K_1)_i,$$

where $Z = (Z_i)$ and $Z_i = Y_i - \bar{Y}$. In the second step we take $U_i = Y_i - \bar{Y} - \hat{g}_{11}(x_i) = Z_i - \hat{g}_{11}(x_i)$ and $j = 2$, obtaining for $\gamma_2(x_{i_0})$:

$$\begin{aligned}\hat{g}_{21}(x_{i_0}) &= \sum_{i=1}^{n} \{Z_i - \hat{g}_n(x_i)\} K_2(i, i_0) \\ &= \{(Z^T - Z^T K_1) K_2\}_{i_0} = \{Z^T (I - K_1) K_2\}_{i_0}.\end{aligned}$$

Here we have used the fact that $\Sigma \hat{g}_{11}(x_i) = 0$, which property follows from the constraint in (7). After j steps of the algorithm, $1 \leq j \leq d$, we obtain

$$\hat{g}_{j1}(x_i) = (Z^T L_j)_i,$$

where

$$L_j = \left\{ \prod_{l=1}^{j-1} (I - K_l) \right\} K_j$$

and $\Pi_{1 \leq l \leq m} M_j$ means $M_1, M_2 \quad \ldots \quad M_m$, in that order. Define

$$L = \sum_{j=1}^{d} L_j = \sum_{j=1}^{d} \left\{ \prod_{l=1}^{j-1} (I - K_l) \right\} K_j. \tag{22}$$

Then d steps of the backfitting algorithm produce as our approximation to $g(x_i) - \mu$,

$$\hat{g}_{(1)}(x_i) = \sum_{j=1}^{d} \hat{g}_{j1}(x_i) = (Z^T L)_i. \tag{23}$$

In the next cycle of the algorithm we replace $Y_i - \bar{Y}$ by

$$Y_i - \bar{Y} - \hat{g}_{(1)}(x_i) = \{Z^T (I - L)\}_i,$$

and run through the sequence again, obtaining after d more steps,

$$\sum_{j=1}^{d} \hat{g}_{j2}(x_i) = \{Z^T (I - L) L\}_i. \tag{24}$$

Our approximation to $g(x_i) - \mu$ after a total of $2d$ steps is the sum of the functions at (23) and (24).

Arguing thus we find that for any $0 \leq j \leq d$ and $v \geq 0$, $vd + j$ steps of the backfitting algorithm produce the following approximation to $g(x_i) - \mu$:

$$\begin{aligned}\sum_{r=1}^{v} \sum_{l=1}^{d} \hat{g}_{lr}(x_i) + \sum_{l=1}^{j} \hat{g}_{l,v+1}(x_i) &= \left[Z^T \left\{ \sum_{r=0}^{v} (I - L)^r L + (I - L)^v \sum_{l=1}^{j} L_l \right\} \right]_i \\ &= \left[Z^T \left\{ I - (I - L)^v + (I - L)^v \sum_{l=1}^{j} L_l \right\} \right]_i . \end{aligned} \tag{25}$$

Note that $\Sigma_{0 \le r \le \nu-1}(I-L)^r L = I-(I-L)^\nu$. If $-(d-j) \le l \le j$ then the approximation to $g_j(x_i)$ obtained after $s = \nu d + l$ steps of the backfitting algorithm is

$$\hat{g}_j(s)(x_i) = \sum_{r=1}^{\nu} \hat{g}_{jr}(x_i) = \left[Z^T \left\{ \sum_{r=0}^{\nu-1} (I-L)^r L_j \right\} \right]_i . \tag{26}$$

When interpreting (25) and (26) it is helpful to note that by (22),

$$I - L = \prod_{j=1}^{d} (I - K_j).$$

4.3 *Convergence of the backfitting algorithm*

In the work which follows we shall interpret an $n \times n$ matrix M as a linear transformation on the space $\mathbb{R}^n$ of row vectors having length n, taking v^T into $v^T M$. Let V denote the vector space of n-vectors v^T which are fixed by each of the transformations $I - K_1, \ldots, I - K_d$. Write $V_\perp$ for the space orthogonal to V, of dimension equal to n minus the dimension of V. For any $1 \le j \le d$ and any n-vector w, we may uniquely write $w = v + v_\perp$, where $v^T \in V$ and $v_\perp^T \in V_\perp$. Since $v^T(I - K_j) = 0$ then $v^T K_j = 0$, whence it follows that for $r \ge 0$,

$$v^T (I-L)^r L_j = v^T \left\{ \prod_{l=1}^{d} (I - K_l) \right\}^r \left\{ \prod_{l=1}^{j-1} (I - K_l) \right\} K_j = v^T K_j = 0.$$

For any $v_\perp^T \in V_\perp$, $v_\perp^T (J-L)^r \to 0$ as $r \to \infty$; see Appendix (3) for a proof. Combining these results we deduce that $w^T (I-L)^r L_j \to 0$ as $r \to \infty$, and so $(I-L)^r L_j \to 0$. In fact the convergence is geometrically fast; see Section 3.6.

We may now deduce from (26) that $\hat{g}_{j(s)}(x_i)$ converges as $s \to \infty$, to $\hat{g}_j$ given by

$$\hat{g}_j(x_i) = \left[Z^T \left\{ \sum_{r=0}^{\infty} (I-L)^r L_j \right\} \right]_i .$$

Furthermore the second part of formula (25),

$$\left[Z^T \left\{ (I-L)^\nu \sum_{l=1}^{j} L_l \right\} \right]_i ,$$

converges to zero as $\nu \to \infty$. Therefore the backfitting s-step approximation to $g(x_i) - \mu$ converges, as $s \to \infty$, to

$$\sum_{j=1}^{d} \hat{g}_j(x_i) = \{ Z^T (I - R) \}_i ,$$

where $R = \lim (I-L)^\nu$ is the matrix of the projection which takes $w = v + v_\perp$ into v. Likewise, $P = I - R$ is the projection which takes w into $v_\perp$.

PROOF OF THEOREM 2

The normal equations which arise from minimizing S', defined at (15), are

$$\sum_{j=1}^{d}\sum_{k=1}^{m_j} c_{jk} J_{jk}^T J_{j_0 k_0} = Z^T J_{j_0 k_0}, \qquad 1 \le k_0 \le m_{j_0}, \qquad 1 \le j_0 \le d. \tag{27}$$

Let A_{jk} be the vector defined in Section 3.3, multiply (27) on the right by $A_{j_0 k_0}^T$, and add over k_0, obtaining

$$\sum_{j=1}^{d}\sum_{k=1}^{m_j} c_{jk} J_{jk}^T K_{j_0}^T = Z^T K_{j_0}^T;$$

that is,

$$K_{j_0}\left(Z - \sum_{j=1}^{d}\sum_{k=1}^{m_j} c_{jk} J_{jk}\right) = 0, \qquad 1 \le j_0 \le d. \tag{28}$$

Now, J_{jk}, $1 \le k \le m_j$, are linearly independent eigenvectors of K_j each with eigenvalue 1, and all the other eigenvalues of K_j equal 0. Hence by (28), the c_{jk}'s are the coefficients of the projection of Z onto the space $V_\perp$ spanned by $\{J_{jk}, 1 \le k \le m_j, 1 \le j \le d\}$, see Appendix (1). In particular, $\Sigma_j \Sigma_k c_{jk} J_{jk}(i)$ equals the ith component of the projection of Z^T onto this space. This establishes uniqueness of $\Sigma_j \Sigma_k c_{jk} J_{jk}(i)$, even though the c_{jk}'s are not unique. Under the additional constraints (13), which are equivalent to $\Sigma_k c_{jk} J_{jk}^T 1 = 0$ for $1 \le j \le d$, uniqueness of the c_{jk}'s follows from (15). We already know from Theorem 1 that the solution of the backfitting algorithm equals the projection of Z onto $V_\perp$.

Appendix 1. Identification of $V_\perp$

When (16) holds, we may characterize $V_\perp$ as follows: $w \in V_\perp$ if and only if

$$w = v + \sum_{j=1}^{d} v_j, \tag{A.1}$$

where $K_j v = v$ for each j and $K_{j_0} v_j = \delta_{jj_0} v_j$, δ_{jj_0} being the Kronecker delta. We shall prove that this implies $V_\perp$ equals the space W spanned by J_{jk}, $1 \le k \le m_j$ and $1 \le j \le d$. If w admits the canonical expansion (A.1) then, since $K_j v = v$ and J_{jk}, for $1 \le k \le m_j = \text{rank } K_j$, represent linearly independent eigenvectors of K_j all with eigenvalue 1, then v can be expressed as a linear combination of $J_{j1}, \ldots, J_{jm_j}$ (for any j). Similarly, since $K_j(w - v) = v_j$ then v_j can be expressed as a linear combination of $J_{j1}, \ldots, J_{jm_j}$. It now follows from (A.1) that $w \in W$. Conversely, suppose $w \in W$. Then w equals a linear combination of the J_{jk}'s which by (16) can be expressed in terms of a constant multiple of 1 (this term playing the role of v in (A.1)) plus $v_1, \ldots, v_d$, where each v_j is orthogonal to 1 and is a linear combination of J_{jk}, $1 \le k \le m_j$. Therefore w admits the expansion (A.1).

Appendix 2. Consistency

Let $\mu = E\{g(X)\}$, and assume that $S_I = E\{g(X) - \mu - \text{S}g_j(X_j)\}^2$ is uniquely minimized by univariate, continuous functions $g_1, \ldots, g_d$, subject to $E\{g_j(X_j)\} = 0$ for $1 \le j \le d$.

Then for any sequence of univariate functions $h_1, ..., h_d$ satisfying $E\{h_j(X_j)\} = 0$ for $1 \le j \le d$,

$$E\left[g(X) - \mu - \sum_{j=1}^{d} \{g_j(X_j) + h_j(X_j)\}\right]^2 = S_1 + E\left\{\sum_{j=1}^{d} h_j(X_j)\right\}^2.$$

Approximating to integrals by series we may deduce that for large n, and with $\bar{\mu} = n^{-1}\Sigma g(x_i)$, $S_2 = n^{-1}\Sigma_i\{g(x_i) - \bar{\mu} - \Sigma_j \check{g}_j(x_i)\}^2$ is uniquely minimized by univariate functions $\check{g}_1, \check{g}_d$, subject to $\Sigma_i g_j(x_i) = 0$ for $1 \le j \le d$, and that $n^{-1}\Sigma_i[\Sigma_j\{g_j(x_i) - \check{g}_j(x_i)\}]^2 \to 0$. Hence for any sequence of univariate functions $\check{h}_1, ..., \check{h}_d$ satisfying $\Sigma_i \check{h}_j(x_i) = 0$ for $1 \le j \le d$,

$$n^{-1}\sum_{i-1}^{n}\left[g(x_i) - \mu - \sum_{j=1}^{d}\{\check{g}_j(x_i) + \check{h}_j(x_i)\}\right]^2 = S_2 + n^{-1}\sum_{i-1}^{n}\left\{\sum_{j=1}^{d}\check{h}_j(x_i)\right\}^2.$$

(We write $g_j(x_i)$ instead of $g_j(x_{ij})$, etc, to simplify notation.) Let $\hat{y} = \Sigma\hat{g}_j$ be the backfitting estimator, and put $\bar{e} = n^{-1}\Sigma e_i$. It may be proved after a little algebra that if $Y_i = g(x_i) + e_i$ where the x_i's represent independent observations of X and the e_i's are independent with bounded fourth moments, and if the fitted estimators are polynomial interpolations of the regressogram, then with $b(x_i)$ denoting either $\hat{y}(x_i)$ or $\Sigma_j \hat{g}_j(x_i)$,

$$n^{-1}\sum_{i=1}^{n}\{Y_i - \bar{Y} - b(x_i)\}^2 = n^{-1}\sum_{i=1}^{n}\{g(x_i) - \bar{\mu} - b(x_i)\}^2 + n^{-1}\sum_{i=1}^{n}(e_i - \bar{e})^2 + o_p(1).$$

It follows that

$$\begin{aligned}
0 &\le n^{-1}\sum_{i=1}^{n}\left[\left\{Y_i - \bar{Y} - \sum_{j=1}^{d}\check{g}_j(x_i)\right\}^2 - \{Y_i - \bar{Y} - \hat{y}(x_i)\}^2\right] \\
&= n^{-1}\sum_{i=1}^{n}\left[\left\{g(x_i) - \bar{\mu} - \sum_{j=1}^{d}\check{g}_j(x_i)\right\}^2 - \{g(x_i) - \bar{\mu} - \hat{y}(x_i)\}^2\right] + o_p(1) \\
&= -n^{-1}\left\{\sum_{j=1}^{d}\check{h}_j(x_i)\right\}^2 + o_p(1),
\end{aligned}$$

where $\check{h}_j(x_i) = \hat{g}_j(x_i) - \check{g}_j(x_i)$. Therefore $n^{-1}\Sigma_i[\Sigma_j\{\hat{g}_j(x_i) - \check{g}_j(x_i)\}]^2 \xrightarrow{p} 0$ where $n^{-1}\Sigma_i$ $[\Sigma_j\{\hat{g}_j(x_i) - g_j(x_i)\}]^2 \xrightarrow{p} 0$. This establishes consistency.

Appendix 3. Limits of products of projections

The following result may be proved by standard methods.

THEOREM A.1. *Let $T_1, ..., T_d$ denote linear transformations from $\mathbb{R}^n$ to $\mathbb{R}$, each having the property that each eigenvalue whose absolute value is not strictly less than unity, equals unity. Write S for a product of all the T_j's in any order. Let V denote the vector space of points which are fixed by each T_j, and let R be the linear transformation which projects $\mathbb{R}^n$ onto V. Write $|\lambda_0|$ for the value of the largent absolute eigenvalue of S strictly less than unity. Then $S^\nu \to R$ as $\nu \to \infty$, and in fact $S^\nu(w) = Rw + 0(|\lambda_0|^\nu)$ as $\nu \to \infty$, for each $w \in \mathbb{R}^n$.*

Acknowledgements

Research of P. HALL was supported by a British Science and Engineering Research Council Visiting Fellow grant to the University of Glasgow. Research of W. HÄRDLE was supported by CORE, Louvain-la-Neuve and a grant from the Statistics Department in Glasgow. Both authors are most grateful to A. W. BOWMAN, J. W. KAY and D. M. TITTERINGTON for helpful discussions.

References

BREIMAN, L. and J. H. FRIEDMAN (1985), Estimating optimal transformations for multiple regression and correlation, *Journal of the American Statistical Association 80,* 580–619.

BUJA, A., HASTIE, T. J. and R. J. TIBSHIRANI (1989), Linear Smoothers and Additive Models (with discussion), *Annals of Statistics 17,* 453–555.

FRIEDMAN, J. H. and W. STUETZLE (1981), Projection pursuit regression, *Journal of the American Statistical Association 76,* 817–823.

GOLUB, G. H. and C. F. VAN LOAN (1983), *Matrix computations,* Johns Hopkins Press, Baltimore.

HÄRDLE, W. and D. SCOTT (1992), Smoothing by Weighted Averaging of Rounded Points, *Computation Statistics 7,* 97–128.

HASTIE, T. and R. TIBSHIRANI (1986), Generalized additive models. (With discussion), *Statistical Science 1,* 297–318.

HASTIE, T. and R. TIBSHIRANI (1987), Generalized additive models: some applications, *Journal of the American Statistical Association 82,* 371–386.

RALSTON, A. (1965), *A first course in numerical analysis,* McGraw-Hill, New York.

STONE, C. J. (1985), Additive regression and other non-parametric models, *Annals of Statistics 13,* 689–705.

ДОКЛАДЫ АКАДЕМИИ НАУК, 1993, том 331, № 5

МАТЕМАТИКА

УДК 519.2

ОБ ЭФФЕКТИВНОМ ОЦЕНИВАНИИ УСРЕДНЕННОЙ ПРОИЗВОДНОЙ

© 1993 г. Член-корреспондент РАН И. А. Ибрагимов,
В. Хэрдле (Бельгия), А. Б. Цыбаков

Поступило 18.01.93 г.

1. Пусть (X, Y) – случайный вектор, принимающий значения в R^{d+1} и имеющий плотность распределения вероятностей $h(x, y)$ по отношению к $\lambda \times \mu$, λ – мера Лебега в R^d. Положим

$$m(x) = \mathbf{E}\{Y \mid X = x\}$$

и определим усредненную производную $\delta(1)$ функции m равенством

$$\delta(1) = \mathbf{E}m'(X)$$

и взвешенную усредненную производную равенством

$$\delta(w) = \mathbf{E}m'(X)w(X).$$

Ниже мы рассматриваем задачу оценивания величины $\delta = \delta(f)$; относительно оценивания $\delta(w)$ и происхождения задачи см. [1 - 3]. Нас прежде всего будет интересовать вопрос, при каких априорных предположениях относительно h возможно асимптотически эффективное оценивание δ; при этом асимптотическая эффективность понимается так, как она определена в [4].

Переходя к более точной постановке задачи, обозначим: $H(\beta, L)$ – класс функций $\varphi(x_1, \ldots, x_d)$ из $L_2(R^d)$, имеющих по каждой переменной x_j в $L_2(R^d)$ производную порядка $r = [\beta]$, $r = \beta - 1$, если β целое, удовлетворяющую условию Гёльдера порядка $\alpha = \beta - r$ в L_2 с константой L. Обозначим $DH(\beta, L)$ класс плотностей $h(x, y)$, для которых:

1) функции

$$f(x) = \int h(x, y)\mu(dy), \quad f(x)m(x) = \int yh(x, y)\mu(dy)$$

принадлежат $H(\beta + 1, L)$;

2) функция $\sigma f' \in L_2(R^d)$, $\sigma^2(x) = \mathrm{Var}(Y \mid X = x)$.

Здесь и ниже Var ξ означает дисперсию ξ.

Рассмотрим задачу оценивания усредненной производной

$$\delta = \mathbf{E}_h m'(X)f(X) = -2\mathbf{E}_h Yf'(X) \qquad (1)$$

по наблюдениям $(X_1, Y_1), \ldots, (X_n, Y_n)$ в предложении, что неизвестная плотность h принадлежит известному множеству плотностей Θ.

Теорема 1. *Допустим, что множество* $\Theta \subseteq DH(\beta, L)$ *и* $\beta > d/4$. *Тогда существует последовательность оценок* $\hat{\delta}_n$ *величины* δ *такая, что нормированная последовательность* $\sqrt{n}(\hat{\delta}_n - \delta)$ *асимптотически нормальна со средним "нуль" и корреляционной матрицей* Σ, *где* Σ *есть корреляционная матрица вектора*

$$\int (yf'(x) - (m(x)f(x))')\sqrt{h(x, y)}dw -$$

$$-\int (yf'(x) - (m(x)f(x))')\, d\lambda\, d\mu \int \sqrt{h(x, y)}dw,$$

w – *ортогональная гауссовская мера с* $\mathbf{E}|dw|^2 = d\lambda d\mu$. *Кроме того, равномерно в* Θ

$$\lim_n n\mathbf{E}_h\left|\hat{\delta}_n - \delta\right|^2 = Q_1 = 4\left[\mathbf{E}_h|f(X)m'(X)|^2 - \right.$$

$$\left. -|\mathbf{E}_h(f(X)m'(X))|^2 + \mathbf{E}_h\sigma^2(X)|f'(X)|^2\right].$$

С точки зрения работы [4], задача оценивания δ есть задача оценивания значения $\Phi(h)$ функционала

$$\Phi(h) = \mathbf{E}_h\{m'(X) \cdot f(X)\} =$$

$$= \int \left\{\int yh'_x(x, y)d\mu - \frac{\int yh(x, y)d\mu}{\int h(x, y)d\mu}\int h'_x(x, y)d\mu\right\} \times$$

$$\times h(x, y)d\lambda\, d\mu$$

и в силу теоремы 5.2 из [2] (см. также примеры 5.2) оценка $\hat{\delta}_n$ теоремы 1 асимптотически эффективна.

Теорема 2. *Допустим, что* $\Theta = DH(\beta, L)$, $\beta < d/4$. *Тогда*

$$\lim_n \inf_{\delta_n} \sup_h n\mathbf{E}_h|\delta_n - \delta|^2 = \infty.$$

Ниже излагается схема доказательства теорем 1, 2.

Петербургское отделение Математического института им. В.А. Стеклова Российской Академии наук
CORE, Лувенский университет, Бельгия
Институт проблем передачи информации Российской Академии наук, Москва

Journal Russian Academy of Sciences, 48, 550-552

2. Обозначим: $D_\nu(x)$ – ядро Дирихле,

$$D_\nu(x) = \pi^{-d}\prod_1^d \frac{\sin \nu_j x_j}{x_j},\quad x = (x_1, \ldots, x_d),$$
$$\nu = (\nu_1, \ldots, \nu_d),$$

и, отправляясь от (1), рассмотрим оценки

$$\delta_n(\nu) = -2n^{-1}\sum_1^n Y_j \hat f_n'(X_j),$$
$$\hat f_n'(x) = n^{-1}\sum D_\nu'(x - X_j). \tag{2}$$

Условимся обозначать через $\mathscr{E}_\nu(f)$ величину наилучшего приближения в L_2 функции f целыми функциями степени не выше $\nu = (\nu_1, \ldots, \nu_d)$ (см. [6]).

Теорема 3. *Для оценок δ_n, определенных* (2), *величина смещения*

$$|\delta - \mathbf{E}\delta_n(\nu)| \le 2\,[\mathscr{E}_\nu(f')\mathscr{E}_\nu(mf) + n^{-1}\|mf\| \cdot \|f'\|].$$

Доказательство проводится прямыми вычислениями с учетом ортогональности в L_2 φ_ν и $\psi - \psi_\nu$, где φ_ν означает интеграл Дирихле функции φ.

Теорема 4. *Если $\Theta \subseteq DH(\beta, L)$, то*

$$\mathrm{Var}\ \delta_n(\nu) = \frac{Q_1}{n} + O(\nu_1 \ldots \nu_d \cdot n^{-2}\sum \nu_j^2) + o(n^{-1}),$$

где постоянные под знаком O и o зависят лишь от $\beta, L, \int \sigma^2(x)|f(x)|^2 d\lambda$.

Доказательство проводится прямыми вычислениями. Из теорем 1 и 2 следует, что, выбирая $\nu_j = \exp\{2[(2\beta + 1) + d + 2]^{-1}\ln n\}$ и обозначая через $\hat\delta_n$ оценку $\delta_n(\nu)$ с таким ν, получим при $\beta > d/4$, что

$$\mathrm{Var}\ \hat\delta_n = n^{-1}Q_1 + o(n^{-1}).$$

Теорема 5. *В условиях теоремы* 1 *разность $\sqrt{n}(\hat\delta_n - \delta)$ асимптотически нормальна со средним "нуль" и корреляционной матрицей Σ.*

Доказательство. Асимптотическое поведение корреляционной матрицы вектора $\sqrt{n}(\hat\delta_n - \delta)$ исследуется прямыми вычислениями, как в предыдущей теореме. Далее, $\hat\delta_n$ представляют собой U-статистики и для доказательства асимптотической нормальности $\sqrt{n}(\hat\delta_n - \delta)$ можно обратиться к общим предельным теоремам для U-статистики (см. [7]).

Из теорем 3 - 5, очевидно, следует теорема 1.

3. Здесь мы укажем **схему доказательства теоремы 2.** Обозначим: Γ – куб $|x_j| \le 1/2$, $j = 1, \ldots, d$, $|y| \le 1/2$. и рассмотрим векторы (X, Y) с плотностью распределения вида

$$h(x, y) = \begin{cases} 1 + s(x, y), & (x, y) \in \Gamma, \\ 0, & (x, y) \notin \Gamma, \end{cases} \tag{3}$$

где

$$s(x, y) = \sum_{\nu_1, \ldots, \nu_d = 1}^N a_{\nu 0}\cos 2\pi\nu_1 x_1 \ldots \cos 2\pi\nu_d x_d + \sum_{j=1}^d \sum_{\nu_1, \ldots, \nu_d = 1}^N a_{\nu j}\cos 2\pi\nu_1 x_1 \ldots \sin 2\pi\nu_j x_j \ldots \cos 2\pi\nu_d x_d \sin 2\pi y. \tag{4}$$

Для таких плотностей $\delta = \mathbf{E}m'(X)f(X) = -2\mathbf{E}Yf'(X)$ есть вектор с компонентами $2^{-d+1}\sum \nu_j a_{\nu 0}a_{\nu j}$. Пусть далее ζ, ξ_ν, $\nu = (\nu_1, \ldots, \nu_d)$, $\nu_j = 1, \ldots, N$, – независимые случайные величины, причем ζ принимает с равными вероятностями значения 0, 1, а ξ_ν – значения ± 1. Пусть ρ_0, ρ_1 – положительные постоянные. Определим случайные функции $h(x, y)$ равенствами (3), (4), полагая $a_{\nu 0} = \rho_0\zeta\xi_\nu$, $a_{\nu j} = \rho_1\zeta\xi_\nu$. Через I обозначим индикатор случайного события $G = \{h(x, y) \ge 0\}$, так что на G функции $h(x, y)$ суть плотности распределения.

В зависимости от того, какое значение примет ζ, 0 или 1, вектор δ или нулевой, или вектор с компонентами $2^{-d}\rho_0\rho_1 N(N + 1)$. Поэтому достаточно рассматривать оценки $\hat\delta$, принимающие лишь эти два значения. Условимся обозначать $\mathbf{E}(\cdot)$ математическое ожидание по отношению к (ζ, ξ_ν). Пусть

$$\hat\delta = \begin{cases} 0, & \text{если } Z = (X_1, Y_1, \ldots, X_n, Y_n) \in A, \\ (\ldots, 2^{-d}\rho_0\rho_1 N(N+1), \ldots), & \text{если } Z \notin A. \end{cases}$$

Тогда

$$\sup_h \mathbf{E}_h|\delta - \hat\delta|^2 \ge cN^4\rho_0^2\rho_1^2(\mathrm{mes}\,\bar A + \mathbf{E}\{I \cdot \int_A \prod_1^n (1 + s(x_i, y_i))\,dx_1 \ldots dy_n\}, \tag{5}$$

здесь и ниже c означает строго положительные постоянные. Дальнейший анализ (5) основан на следующих утверждениях:

Лемма 1. *Пусть $\rho_0 \le \rho_1 = n^{-\gamma}$, $\gamma > 1$, $N = n^{2\gamma'}$, $\gamma' < \gamma$. Если* $\mathrm{mes}\,A \ge 3/4$, *то*

$$\mathbf{E}\int_A \prod_1^n (1 + s(x_i, y_i))\,dx_1 \ldots dy_n \ge c.$$

Härdle, W., Ibragimov, I. and Tsybakov, A.B. (1993)
Ob Effektivnom ozenivani usrednennoi proiswodnoi.

Journal Russian Academy of Sciences, 48, 550-552

Лемма 2. *Пусть* ρ_0, ρ_1, N *те же, что выше. Тогда*

$$\mathbf{P}\left\{\left|\int_A \prod_1^n (1 + s(x_i, y_i))\, dx_1 \ldots dy_n\right| > L\right\} \le BL^{-4}.$$

Лемма 3. *Пусть* ρ_0, ρ_1, N *те же, что и выше. Тогда*

$$\mathbf{P}\{\|s\|_\infty \ge 1\} \le Be^{-n^\varepsilon}.$$

Рассматривая всевозможные реализации последовательностей $(\zeta, \xi_\nu) \in G$, мы получим некоторое (конечное) множество плотностей Θ_1, и в силу (5) и лемм 1 - 3

$$\inf_{\hat\delta} \sup_{h \in \Theta_1} \mathbf{E}_h \left|\hat\delta - \delta\right|^2 \ge cN^4 \rho_0^2 \rho_1^2.$$

Несложные выкладки показывают теперь, что, если $\beta < d/4$, можно подобрать N, ρ_0, ρ_1 так, чтобы $N^4\rho_0^2\rho_1^2 > n^{-\alpha}$, $\alpha > 1$, и

$$\sum_j \sum \nu_j^{2(\beta+1)} |a_\nu|^2 \le B.$$

Домножив определенные выше h на подходящим образом выбранный множитель, уничтожающий скачки h, мы превратим Θ_1 в $\Theta_2 \subset \Theta = LH_\beta^2$, что и доказывает теорему 2.

4. Разумеется, вместо ядра Дирихле при построении оценок можно употреблять другие ядра. Скажем, ядро Валле Пуссена приведет к сходным результатам, ядро Фейера в определенных ситуациях может дать худшее смещение и т.д. Ниже мы приведем результат, показывающий зависимость $\mathbf{E}\left|\hat\delta - \delta\right|^2$ от ядра. Обозначим: $\mathcal{K}_l$ – класс ядер $R(x) = \prod_1^d K(x_j)$, где функция K непрерывно дифференцируема, имеет носитель, лежащий внутри $[-1, 1]$, $K'(0) = 0$. Кроме того,

$$\int K(u)du = 1, \quad \int u^j K(u)du = 0, \quad j = 1, \ldots, l-1,$$

$$\int u^l K(u)du \ne 0.$$

Теорема 6. *Допустим, что*: 1) *плотность* $f(x)$ *величин* X_j *имеет компактный носитель и непрерывно дифференцируема* $l + 1$ *раз по каждому аргументу*; 2) *функция регрессии* $m(x)$ *непрерывно дифференцируема* $l + 1$ *раз по каждому аргументу*; 3) *условная дисперсия* $\sigma^2(x)$ *ограничена на носителе* f. *Определим ядерную оценку* $\hat\delta$:

$$\hat\delta_n = -2n^{-1}\sum_{i=1}^n Y_i \hat f_h'(X_i), \quad f_h'(x) = n^{-1}\sum_1^n R_h'(x - X_i),$$

где ядро $R \in \mathcal{K}_l$, $R_h(x) = h^{-d}R(xh^{-1})$, $h = h_n \to 0$, $n_2 h_n^{d+2} \to \infty$. *Тогда*

$$\mathbf{E}\left|\hat\delta_n - \delta\right|^2 = n^{-1}Q_1 + n^{-2}h^{-d-2}Q_2 + h^{2l}Q_3 + o(n^{-2}h^{-d-2} + h^{2l} + n^{-1}).$$

Здесь Q_1 *определено в п.* 2,

$$Q_2 = 4C(K)\int \sigma^2(x) f^2(x)dx,$$

$$Q_3 = 4\left|\int S_K(x) f(x) m(x)dx\right|^2,$$

$$C(K) = d\int (K'(u))^2 du \left(\int K^2(u)du\right)^{d-1},$$

$$S = d_K \frac{(-1)^l}{l!}\sum_1^d \begin{pmatrix} \partial^{l+1} f/\partial x_1 \partial x_j^l \\ \ldots\ldots\ldots \\ \partial^{l+1} f/\partial x_d \partial x_j^l \end{pmatrix}.$$

Эта работа была выполнена, когда первый и третий авторы находились в Центре исследования операций и эконометрики (CORE) Лувенского университета, Бельгия. Оба автора весьма признательны администрации Центра за предоставленные им прекрасные возможности для работы.

СПИСОК ЛИТЕРАТУРЫ

1. *Härdle W., Stoker Th.* // J. Amer. Statist. Assoc. 1989. V. 84. No. 408. P. 986 - 995.
2. *Stoker Th.* // Econometrica. 1986. V. 54. No. 6. P. 1461 -1481.
3. *Powell J., Stock J., Stoker Th.* // Ibid. 1989. V. 57. P. 1240 - 1252.
4. *Ibragimov I., Khasminskii R.* // Ann. Stat. 1991. V. 19. P. 1681 - 1724.
5. *Никольский С.М.* Приближение функций многих переменных и теоремы вложения. М.: Наука, 1969. 480 с.
6. *Королюк В.С., Боровских Ю.* Мартингальная аппроксимация. Киев: Наук. думка, 1988. 248 с.

Härdle, W., Ibragimov, I. and Tsybakov, A.B. (1993)
Ob Effektivnom ozenivani usrednennoi proiswodnoi.

Annals of Statistics, 1993, Vol. 21, No. 4, 1926-1947

Comparing Nonparametric Versus Parametric Regression Fits *

W. Härdle
Humboldt-Universität zu Berlin,
Institut für Statistik und Ökonometrie
Fachbereich Wirtschaftswissenschaften
Spandauer Strasse 1, O-1020 Berlin

E. Mammen
Institut für Angewandte Mathematik,
Ruprecht-Karls Universität, Heidelberg,
Im Neuenheimer Feld 294, W-6900 Heidelberg

November 1990
Revised June 1992
Revised November 1992

Abstract

In general, there will be visible differences between a parametric and a nonparametric curve estimate. It is therefore quite natural to compare these in order to decide whether the parametric model could be justified. An asymptotic quantification is the distribution of the integrated squared difference between these curves. We show that the standard way of bootstrapping this statistic fails. We use and analyse a different form of bootstrapping for this task. We call this method the wild bootstrap and apply it to fitting Engel curves in expenditure data analysis.

short title : Parametric vs. Nonparametric Fits

MSC 1991 Subject Classification : primary 62607, secondary 62609, 62919. Keywords and phrases: kernel estimate, bootstrap, wild bootstrap, goodness-of-fit test

* Research supported by Deutsche Forschungsgemeinschaft, Sonderforschungsbereich 123, "Stochastische Mathematische Modelle" and Sonderforschungsbereich 303, "Information und Koordination Wirtschaftlicher Aktivitäten", CORE, Université Catholique de Louvain and CentER, Katholieke Universiteit Brabant.

Annals of Statistics, 1993, Vol. 21, No. 4, 1926-1947

1. Motivation

The appropriateness of parametric modelling of regression data may be judged by comparison with a nonparametric smoothing estimator. For this purpose one may use a squared deviation measure between the two fits. The integrated squared deviation can be used as a test statistic for testing the parametric model where the critical value is determined by the asymptotic distribution of this statistic. The convergence to the asymptotic normal distribution is quite slow so that it seems more appropriate not to use the asymptotic critical values. A way of computing critical values could possibly be based on resampling from the entire set of observations. It is shown here that this method of bootstrapping fails. Instead we propose in this setting a new variant of the bootstrap (due to Wu, 1986) which we call the *wild bootstrap*. The classical bootstrap — to resample from the entire data — does not work. The bootstrapped statistic has not the same limit behaviour. It will be shown that the wild bootstrap works. The fact that classic resampling does not work here is theoretically appealing and can be understood via a Hoeffding decomposition. Indeed the quadratic term of the proposed test statistic dominates asymptotically the linear term.

It is surprising that although the nonparametric approach in modelling regression relationships has received a lot of attention recently (see Collomb, 1981), there are only a few theoretical results on how to compare parametric with nonparametric fits. In practical studies the importance of comparing parametric with nonparametric curves has been pointed out in the analysis of growth curves by Gasser, Müller, Köhler, Molinari and Prader (1985). Another example stems from the analysis of the income distribution of British households (Family Expenditure Survey, 1968–1983). Hildenbrand and Hildenbrand (1986) found that the widely used lognormal fit for the income distribution was not able to model the seemingly bimodal distribution.

Theoretical results in this direction are offered by Yanagimoto and Yanagimoto (1987), Cleveland and Devlin (1988), Cox, Koh, Wahba and Yandell (1988), Azzalini, Bowman and Härdle (1989), Cox and Koh (1989), Munson and Jernigan (1989), Eubank and Spiegelman (1990), Härdle and Marron (1990), le Cessie and van Houwelingen (1991), Staniswalis and Severini (1991). LaRiccia (1991) used the idea of comparing a parametric model for the quantile function against a nonparametric alternative for testing a composite goodness-of-fit null hypothesis. An asymptotic χ^2 distribution was derived.

In Section 2 we derive the asymptotic distribution of the squared deviation between the parametric and the nonparametric fit. The asymptotic distribution could be estimated with the "plug-in" method, the involved functionals though seem to

be rather complicated. Section 3 is devoted to the question of how to bootstrap in this setting. Section 4 gives several simulations and an application to Engel curve estimation. The proofs are given in Section 5.

2. How Far is the Nonparametric from the Parametric Model?

We consider the following model. Given are n i.i.d. observations $\{(X_i, Y_i)\}_{i=1}^n$ ($X_i \in \mathbb{R}^d$, $Y_i \in \mathbb{R}$) with unknown regression function $m(\bullet) = E(Y_i|X_i = \bullet)$. We write also $Y_i = m(X_i) + \varepsilon_i$ with $E(\varepsilon_i \mid X_i) = 0$. We do not assume that the ε_i are conditional i.i.d. as in Eubank and Spiegelman (1990), Härdle and Marron (1990). In particular, this contains the case of conditional heteroscedasticity. We are interested in the following testing problem. We wish to test the parametric model $\{m_\theta : \theta \in \Theta\}$ against the nonparametric alternative which only assumes that $m(\bullet)$ is "smooth". A natural approach is to plot a parametric regression estimator $m_{\hat\theta}$ with bandwidth $h = h_n$ and kernel K (Nadaraya, 1964; Watson, 1964).

$$\hat m_h(\bullet) = \frac{\sum_{i=1}^n K_h(\bullet - X_i)Y_i}{\sum_{i=1}^n K_h(\bullet - X_i)},$$

$$K_h(\bullet) = h^{-d}K(\bullet/h).$$

For simplification of notation the dependence of $\hat\theta$ and h on n will be dropped.

The question arises if visible differences between $m_{\hat\theta}$ and $\hat m_h$ can be explained by stochastic fluctuations or if they suggest to use nonparametric instead of parametric methods. One way to proceed is to measure the distance between $m_{\hat\theta}$ and $\hat m_h$ and to use this distance as the test statistic for testing the parametric model. Here we study the L_2-distance between the nonparametric and parametric fits. The use of this distance is motivated by mathematical convenience. Certainly from a more data analytic point of view distances would be more satisfactory which reflect similarities in the shape of the regression functions, but nevertheless we will restrict ourselves to the treatment of the weighted L_2-distance $\int(\hat m_h - m_{\hat\theta})^2\pi$ where π is a weight function.

Let $\mathcal{K}_{h,n}$ denote the (random) smoothing operator

$$\mathcal{K}_{h,n}g(\bullet) = \frac{\sum_{i=1}^n K_h(\bullet - X_i)g(X_i)}{\sum_{i=1}^n K_h(\bullet - X_i)}.$$

Because of $E(\hat m_h(\bullet)|X_1, \ldots, X_n) = \mathcal{K}_{h,n}m(\bullet)$ we consider the following modification of the squared deviation between $\hat m_h$ and $m_{\hat\theta}$:

$$T_n = nh^{d/2}\int(\hat m_h(x) - \mathcal{K}_{h,n}m_{\hat\theta}(x))^2\pi(x)dx.$$

In this definition $\hat{m}_h$ is compared with the parametric "estimate" $\mathcal{K}_{h,n} m_{\hat{\theta}}$ of the conditional expectation of $\hat{m}_h$.

We propose to use T_n as a test statistic to test the parametric hypothesis:

$$m \in \{m_\theta : \theta \in \theta\}.$$

On the hypothesis, T_n is asymptotically equivalent to the sum of a constant and a purely quadratic form (see the proof of Proposition 1). This corresponds to symmetries in the asymptotic power of T_n. On the contrary, the test statistic $\int (\hat{m}_h - m_{\hat{\theta}})^2 \pi$ contains asymptotically also a linear term. This linear term makes this test only sensitive against certain "smooth" deviations from the hypothesis (see Proposition 2).

Alternative definitions of a test statistic are possible. The integral may be replaced by a sum (e.g. over the design points). Furthermore the integrand may be multiplied by a power of $\hat{f}_h(\bullet) = n^{-1} \sum_{i=1}^n K_h(\bullet - X_i)$. Under our assumptions the asymptotic arguments for these modifications are applicable but constants might have to be changed.

A related test for testing a parametric form of a density has been proposed by Neuhaus (1986, 1988). For an approximate calculation of critical values we determine the asymptotic distribution of T_n for a parametric $m = m_{\theta_0}$. Furthermore for a comparison of T_n with other goodness-of-fit tests we calculate the asymptotic power of T_n if m (possibly depending on n) lies in the alternative: say $m(x) = m_n(x) = m_{\theta_0}(x) + c_n \Delta_n(x)$ for certain sequences c_n and Δ_n. It is most appropriate to choose c_n such that the asymptotic power of T_n is bounded away from one and from the level. We will show that for instance for "regular" constant $\Delta_n(x) = \Delta(x)$ this will be the case for $c_n = n^{-1/2} h^{-d/4}$.

3. Assumptions

We make the following assumptions on the stochastic nature of the observations and the parametric estimator of the regression function.

(A1) With probability one X_i lies in a compact set (w.l.o.g. $[0,1]^d$). The marginal density $f(\bullet)$ of X_i is bounded away from zero.

(A2) $m(\bullet)$ and $f(\bullet)$ are twice continuously differentiable. π is continously differentiable.

(A3) $\Delta_n(\bullet)$ is bounded (uniformly in x and n) and $c_n = n^{-1/2}h^{-d/4}$. Especially this contains the parametric case because $\Delta_n \equiv 0$ is possible.

(A4) $\sigma^2(\bullet) = \mathrm{var}(Y_i|X_i = x)$ is bounded away from zero and from infinity.

(A5) $E\exp(t\varepsilon_i)$ is uniformly bounded in i and n for $|t|$ small enough (where $\varepsilon_i = Y_i - m(X_i)$).

Before stating our assumptions for the parametric model let us consider the special case of a k-dimensional linear model:

$$m_\theta(\bullet) = \theta_1 g_1(x) + \cdots + \theta_k g_k(x) =< \theta, g(\bullet) >$$

where g is a $\mathbb{R}^k$-valued function (for some k). With a smooth weight function w the weighted least squares estimator $\hat{\theta}_n = \hat{\theta}$ is defined by

$$\hat{\theta} = \arg\min_\theta \sum_{i=1}^n w(X_i)(Y_i - m_\theta(X_i))^2.$$

In the linear model $\hat{\theta}$ can easily be calculated

$$\hat{\theta} = (\sum_{i=1}^n w(X_i)g(X_i)g(X_i)^T)^{-1} \sum_{i=1}^n w(X_i)g(X_i)Y_i.$$

Consider now a regression function $m = m_n$ which may lie in the hypotheses or in the alternative. We want to write m as $m(\bullet) = m_{\theta_0}(\bullet) + c_n\Delta_n(\bullet)$ for some θ_0 (which may also depend on n) and Δ_n. θ_0 and $\Delta_n(\bullet)$ may be chosen as follows

$$\theta_0 = \arg\min_\theta \int w(x)(m(x) - m_\theta(x))^2 dx$$

$$\Delta_n(\bullet) = \frac{1}{c_n}(m(\bullet) - m_{\theta_0}(\bullet)).$$

With this choice of m_{θ_0} and Δ_n, Δ_n is orthogonal to $\{m_\theta(x) : \theta \in \Theta\}$ in the following sense:

$$\int w(x)f(x)\Delta_n(x)g_j(x)dx = 0, \quad j = 1, \ldots, k.$$

This implies that the expectation of $\hat{\theta}$ is approximately θ_0 as can be seen by the

following stochastic expansion of $\hat{\theta}$,

$$\begin{aligned}\hat{\theta} &= \theta_0 + \left(\int w(x)f(x)g(x)g(x)^T dx\right)^{-1} \\ &\quad \left\{\frac{1}{n}\sum_{i=1}^{n} w(X_i)g(X_i)\varepsilon_i + c_n \int w(x)f(x)\Delta_n(x)g(x)dx\right\} \\ &\quad + O_p\left(\frac{c_n}{\sqrt{n}}\right) \\ &= \theta_0 + \frac{1}{n}\sum_{i=1}^{n} h(X_i)\varepsilon_i + O_p\left(\frac{c_n}{\sqrt{n}}\right)\end{aligned}$$

where

$$h(\bullet) = \left(\int w(x)f(x)g(x)g(x)^T dx\right)^{-1} w(\bullet)g(\bullet).$$

If g, h are bounded functions then especially this implies

(P1) $m_{\hat{\theta}}(\bullet) - m_{\theta_0}(\bullet) = \frac{1}{n}\sum_{i=1}^{n} < g(\bullet), h(X_i) > \varepsilon_i + o_p\left(\frac{1}{\sqrt{\log n}\sqrt{n}}\right)$ (uniformly in x), where g and h are bounded functions taking values in $\mathbb{R}^k$ for some k.

For the parametric model we will assume in this section only (P1). By linearization it can be shown that (P1) holds also for weighted least squares estimators $\hat{\theta}$ in nonlinear models if $m(\bullet)$ and $w(\bullet)$ are "smooth" and Δ_n and θ_0 are chosen similarly with $g_j(\bullet) = \frac{\partial}{\partial\theta_j} m_{\theta_0}(\bullet)$

For the kernel K we make the following assumptions.

(K1) The kernel K is a symmetric, twice continuously differentiable function with compact support, furthermore $\int K(u)du = 1$.

(K2) The bandwidth h fulfills $h = h_n \sim n^{-1/(d+4)}$.

Especially (K2) is fulfilled for every choice of the bandwidth h which is asymptotically optimal for the class of twice continuously differentiable regression functions. For simplicity of notation we do not consider bandwidths which are asymptotically optimal for other smoothness classes.

4. The asymptotic behaviour of T_n.

In the following proposition we will approximate the distribution of T_n by a Gaussian distribution with a mean which converges to infinity (also for the null hy-

pothesis). We will measure the distance between these distributions by the following modification of the Mallows distance

$$d(\mu,\nu) = \inf_{X,Y}(E\|X-Y\|^2 \wedge 1 : \mathcal{L}(X)=\mu, \mathcal{L}(Y)=\nu).$$

Convergence in this metric is equivalent to weak convergence.

Proposition 1. *Assume (A1)-(A5), (P1), (K1), (K2). Then*

$$d\left(\mathcal{L}(T_n), N(b_h + \int (\mathcal{K}_h\Delta_n)^2\pi, V)\right) \to 0$$

where

$$b_h = h^{-d/2}K^{(2)}(0)\int \frac{\sigma^2(x)\pi(x)}{f(x)}dx$$

$$V = 2h^d \int \frac{\sigma^2(x)\sigma^2(y)\pi(x)\pi(y)}{f(x)f(y)}(K_h^{(2)}(x-y))^2 dxdy.$$

$\mathcal{K}_h$ *denotes the following smoothing operator*

$$\mathcal{K}_h g(\bullet) = \int K_h(\bullet - t)g(t)dt.$$

$K_h^{(j)}$ *denotes the* j*-times convolution product of* K_h*. If* $\sigma^2(\bullet)$ *is continuous,* V *can be chosen as*

$$V = 2K^{(4)}(0)\int \frac{\left[\sigma^2(x)\right]^2\pi(x)^2}{f^2(x)}dx.$$

Because of the slow order of convergence, we do not recommend to use Proposition 1 for the approximate calculation of critical values (see also the simulations in Section 6). We will use Proposition 1 to study consistency of properties of different bootstrap procedures in the next section. The proposition gives also a rough impression on the power of T_n. It shows that for $d = 1$ the power of the goodness of fit test based on T_n is asymptotically constant on regions of the form $\{m_{\theta_0} + n^{-9/20}\Delta : \int(\mathcal{K}_h\Delta)^2\pi = const.\}$. This can be compared with the behaviour of other goodness of fit tests. The accuracy of the parametric model may also be checked by testing against a higher dimensional parametric model or by test statistics which are asymptotically of Cramer–von Mises type or of Kolmogorov–Smirnov type. These tests have nontrivial power on points contiguous to the parametric model (i.e., $m = m_{\theta_0} + n^{-1/2}\Delta$) but they are of more parametric nature — in the sense that they look into certain one-dimensional directions, Durbin and Knott (1972), Milbrodt and Strasser (1990). The nonparametric behaviour of T_n (nearly the same power for all deviations of fixed

weighted L_2-norm) must be paid by the larger distance ($n^{-9/20}$ instead of $n^{-1/2}$) at which the test works.

We expect that the proposition holds true also for data adapted bandwidth $\hat{h}$ as long as $\hat{h}/h \xrightarrow{p} 1$ but we do not investigate this further. In a related context (two-sample tests based on kernel density estimates) this has been shown in Chapter 3 of Mammen (1992b) using tightness arguments.

Tests based on the statistic $\int(\hat{m}_h - m_{\hat{\theta}})^2\pi$ behave quite differently. This can be seen from Proposition 2. This test behaves asymptotically like an asymptotic linear test. It is sensitive to deviations from the hypothesis in one direction of order $n^{-1/2}$.

Proposition 2. *Suppose (A1), (A2), (A4), (A5), (P1), (K1), (K2) and (A3') $\Delta_n(\bullet)$ is bounded (uniformly in x and n) and $c_n = n^{-1/2}$.*

Furthermore, suppose that $\sigma^2(\bullet)$ is continuous and for simplicity, that the dimonsion d is one. Then

$$d\left(\mathcal{L}(nh^{1/2}\int(\hat{m}_h - m_{\hat{\theta}})^2\pi), N(b'_h + n^{1/2}h^{5/2}c_K\int\rho(x)\Delta_n(x)\pi(x)dx, V')\right) \to 0,$$

where

$$\rho(x) = m''(x) + \frac{2f'(x)}{f(x)}m'(x),$$

$$c_K = \int u^2K(u)du,$$

$$b'_h = b_h + \frac{1}{4}c_K^2nh^{9/2}\int\rho^2(x)\pi(x)dx,$$

$$V' = V + c_K^2nh^5\int[\rho(x)\pi(x)$$

$$- h(x)f(x)\int\rho(u)g(u)\pi(u)du]^2\sigma^2(x)f^{-1}(x)dx$$

5. How to Bootstrap.

The proof of Proposition 1 is based on a stochastic expansion with error terms of order $n^{-1/10}$. Therefore the theorem can only give a rough idea of the stochastic

behaviour of T_n if the sample is small. In the simulations given in the next section we will see that the Normal approximation does not work very well for moderate sample sizes. We will study in this section bootstrap methods as an alternative to asymptotics. We consider three different possibilities of bootstrapping,

- the naive resampling method;
- the adjusted residual bootstrap;
- the wild bootstrap.

We show that only the third type of bootstrap will work. The *naive bootstrap* consists of simple resampling of the original observations. That is the bootstrap sample $\{(X_i^*, Y_i^*)\}_{i=1}^n$ is drawn (with replacement) out of the set $\{(X_i, Y_i)\}_{i=1}^n$. Then create $T^{*,N}$ like T_n by the squared deviation between the parametric fit $m_{\hat{\theta}^*}$ and the nonparametric fit $\hat{m}_h^*$ (both computed from the bootstrap sample $\{(X_i^*, Y_i^*)\}_{i=1}^n$):

$$T^{*,N} = nh^{d/2} \int (\hat{m}_h^*(x) - \mathcal{K}_{h,n} m_{\hat{\theta}^*}(x))^2 \pi(x) dx.$$

The conditional distribution $\mathcal{L}^*(T^{*,N}) = \mathcal{L}(T^{*,N} | \{(X_i, Y_i)\}_{i=1}^n)$ can be approximated by Monte Carlo simulations. From this Monte Carlo approximation define the $(1-\alpha)$ quantile $\hat{t}_\alpha^N$ and reject the parametric hypothesis if $T_n > \hat{t}_\alpha^N$. We call this the naive bootstrap because this resampling does not correctly reflect the stochastic structure of our model as we will see below.

The bootstrap estimate should suffice two conditions. If m lies in the parametric model $\mathcal{L}^*$ (T^*) should consistently estimate the distribution of T_n. But on the alternative $\mathcal{L}^*(T^*)$ should approximate a distribution of T_n under the null hypothesis. This is important for a good power performance. For the second aim we consider the following modification. The bootstrap with adjusted residuals is defined by resampling from the observations $\{(X_i, Y_i - \hat{m}_h(X_i) + m_{\hat{\theta}}(X_i))\}_{i=1}^n$. $T^{*,A}$ might now be created like T_n by the squared deviance between the parametric fit and the nonparametric fit. As above the conditional distribution $\mathcal{L}^*(T^{*,A}) = \mathcal{L}(T^{*,A} | \{(X_i, Y_i)\}_{i=1}^n)$ can be approximated by Monte Carlo simulations. From $\mathcal{L}^*(T^{*,A})$ define the $(1-\alpha)$ quantile $\hat{t}_\alpha^A$ and reject the parametric hypothesis if $T_n > \hat{t}_\alpha^A$. We call this the *adjusted residual bootstrap*. In the following theorem we show that on the null hypothesis both $\mathcal{L}^*(T^{*,A})$ and $\mathcal{L}^*(T^{*,N})$ have variance larger than that of $\mathcal{L}(T_n)$. This implies that naive bootstrap and adjusted residual bootstrap do not work. Both procedures lead to very conservative tests.

Theorem 1. *Assume (A1),...,(A5),(K1),(K2) and $\Delta_n = 0$. (This implies that m lies on the hypothesis i.e. $m = m_{\theta_0}$ for a θ_0.) Define $\hat{\theta}^{*,N}$ and $\hat{\theta}^{*,A}$ but with the bootstrap data $\{(X_i^*, Y_i^*)\}_{i=1}^n$ instead of $\{(X_i, Y_i)\}_{i=1}^n$. Assume for $\hat{\theta}^{*,N}$ (or $= \theta^{*,A}$*

resp.)

$$(P1') \qquad m_{\hat{\theta}^*}(\bullet) - m_{\theta_0}(\bullet) = \frac{1}{n}\sum_{i=1}^{n} < g(\bullet), h(X_i) > \varepsilon_i^* + o_{p^*}\left(\frac{1}{\sqrt{n \log n}}\right)$$

where $\varepsilon_i^* = Y_i^* - m(X_i^*)$. *Assume also that the variance function* $\sigma^2(\bullet)$ *is continous. Then the bootstrap estimates of the variance of* T_n *converge as follows:*

$$\begin{aligned} &\mathrm{var}^*(T^{*,N}) - 3\mathrm{var}(T_n) \xrightarrow{p} 0 \\ &\mathrm{var}^*(T^{*,A}) - \mathrm{var}(T_n) \\ &\qquad - 4\int \frac{\left[\sigma^2(x)\right]^2 \pi^2(x)}{f^2(x)} dx \int (K^{(2)}(t) - K^{(3)}(t))^2 dt \\ &\qquad \xrightarrow{p} 0. \end{aligned}$$

As above it can easily be seen that (P1′) is fulfilled for weighted least squares estimators $\hat{\theta}$ in linear models. Furthermore it can be shown that (P1′) holds for nonlinear models under standard regularity conditions. Theorem 1 shows that the above proposed bootstrap procedures (also after a bias correction) are inconsistent. At first sight this result seems to be surprising and against the intuition of the bootstrap. The deeper reason lies in the fact that under the bootstrap distribution the regression function is NOT the conditional expectation of the observation. Under our assumptions we get almost surely $E^*(Y_i^*|X_i^*) = Y_i^*$, which is typically different from $m_{\hat{\theta}}(X_i^*)$. Here E^* denotes the conditional expectation $E(\bullet|\{(X_i, Y_i)\}_{i=1}^n)$. Wu (1986) has pointed out inconsistency for bootstrap estimates of the least-squares estimator in linear models for nonconstant conditional variance function. In our case the bootstrap will also break down, even for homoscedastic errors. As an alternative we recommend the *wild bootstrap* which is related to proposals of Wu (1986) (see also Beran (1986), Liu (1988), Mammen (1992a)).

This approach does not mimic the i.i.d. structure of (X_i, Y_i). It is rather constructed so that

$$E^*(Y_i^*|X_i^*) = m_{\hat{\theta}}(X_i^*)$$

For this purpose define

$$\tilde{\varepsilon}_i = Y_i - \hat{m}_h(X_i).$$

Since we are going to use this *single residual* $\tilde{\varepsilon}_i$ to estimate the conditional distribution $\mathcal{L}(Y_i - m(X_i)|X_i)$ by an $\hat{F}_i$ we are calling it the *wild bootstrap*. More precisely define

an arbitrary distribution $\hat{F}_i$ such that

$$E_{\hat{F}_i} Z = 0,$$
$$E_{\hat{F}_i} Z^2 = (\tilde{\varepsilon}_i)^2,$$
$$E_{\hat{F}_i} Z^3 = (\tilde{\varepsilon}_i)^3.$$

We use a two-point distribution which is uniquely determined by these requirements. For other constructions see Liu (1988).

Now construct independent $\varepsilon_i^* \sim \hat{F}_i$ and use $(X_i, Y_i^* = m_{\hat{\theta}}(X_i) + \varepsilon_i^*)$ as bootstrap observations. Then create $T^{*,W}$ like T_n by the squared deviation between the parametric fit and the nonparametric fit. From the Monte Carlo approximation of $\mathcal{L}^*(T^{*,W})$ construct the $(1-\alpha)$ quantile $\hat{t}_\alpha^W$ and reject the parametric hypothesis if $T_n > \hat{t}_\alpha^W$. In the following Theorem we show that this procedure works. On the null hypothesis it estimates consistently the distribution of T_n. On the alternative the wild bootstrap estimate converges to a distribution under the null hypothesis. Note that the wild bootstrap mimics correctly the conditional expectation: $E^*(Y_i^*) = m_{\hat{\theta}}(X_i)$.

Theorem 2. *Assume (A1),...,(A5),(P1),(K1),(K2). Furthermore suppose for the parametric estimator $\hat{\theta}^*$ (based on the bootstrap sample)*

$$(P1'') \qquad m_{\hat{\theta}^*}(\bullet) - m_{\hat{\theta}}(\bullet) = \frac{1}{n}\sum_{i=1}^{n} < g(x), h(X_i) > \varepsilon_i^* + o_{p^*}\left(\frac{1}{\sqrt{n \log n}}\right)$$

Then

$$d(\mathcal{L}^*(T^{*,W}), N(b_h, V)) \xrightarrow{p} 0$$

where b_h and V are defined in Theorem 1.

Like (P1′) condition (P1″) is also fulfilled under standard regularity conditions. Note that the proposed wild bootstrap procedure requires rather weak assumptions on the conditional distributions of the errors. If one had prior information on the error structure one could modify the resampling scheme. One such prior information is smoothness of the variance function $\sigma^2(x)$. In this case we propose to define the variance of $\hat{F}_i$ as $\hat{\sigma}^2(X_i)$ where $\hat{\sigma}^2(\bullet)$ is a nonparametric estimator of $\sigma^2(\bullet)$, see Carroll (1982). For another bootstrap procedure where the conditional error distributions vary smoothly with x, see Cao-Abad and Gonzales-Manteiga (1990). For the special case where the errors are conditionally i.i.d. one could bootstrap from the entire set

of (scaled) residuals. We expect that these schemes would work better under the additional model assumptions.

6. Simulations and Applications

We have checked the validity of our asymptotic results in a Monte Carlo experiment. In a first simulation we generated $\{X_i\}_{i=1}^n, n = 100$, uniformly in $[0,1]$ and $Y_i = m(X_i) + \varepsilon_i$ with $\varepsilon_i \sim N(0, \sigma^2), \sigma = 0.1$, independent of X_i. The regression function has been put $m(x) \equiv 0$. For construction of the kernel smooth we have used the quartic kernel

$$K(u) = \frac{15}{16}(1 - u^2)^2 I(|u| \leq 1).$$

The integral of the test statistic T_n has been numerically approximated. The bootstrap resampling was performed $B = 100$ times for each sample, i.e. the Monte Carlo approximation to $\mathcal{L}^*(T^{*,W})$ has been performed by B repetitions of the Wild Bootstrap algorithm. In order to study the distribution of T_n the whole sampling mechanism procedure was carried out $M = 1000$ times.

At first we consider the parametric model of polynomials of degree $k, k = 0, 1, 2, 3$. The true regression curve $m(\bullet)$ is in this model class for each k. For the kernel estimator the bandwidth $h = 0.2$ has been chosen. In Figures 1–4 we present four curves for each k.

Figures 1–4. Monte Carlo density of T_n, bootstrap density estimate, and normal density. The thin line 1 is the Monte Carlo density of T_n , the line 2 is the kernel density of T_n^* from ONE bootstrap sample. Line 3 is the Normal density with the asymptotic mean and variance from Proposition 1, line 4 is the Normal density approximation with estimated mean and variance. The parametric model consist of polynomials of degree k=0 (Figure 1), . . . , k=3 (Figure 4).

The thin line denotes the Monte Carlo kernel estimate of the density of T_n from the M runs. The medium thin line is the kernel density of one bootstrap sample out of the M runs (taken at random). The thick line corresponds to the Normal theory density as given in Proposition 1 based on the true b_h and V. The dashed line finally shows the Normal theory density based on estimated b_h and V. The quantities b_h and V have been estimated by the so-called "plug–in" method, i.e. consistent estimators $\hat{f}_h(\bullet)$ and $\hat{\sigma}^2(\bullet)$ (also based on the kernel technique) have been used. The stability (under variation of the smoothing parameter) of these estimators was satisfactory and did not affect the Figures 1–4.

In all four cases wild bootstrap estimates the distribution of the T_n distance quite

well. The normal approximation with estimated b_h and V is totally misleading. The normal densities have considerable mass on the negative axis. The inaccuracy of the normal approximations increases with the dimension of the parametric model. To study the power of our bootstrap test we have chosen the parametric model

$$m_\theta(x) = \theta_1 + \theta_2 x + \theta_3 x^2$$

and for different c the regression function

$$m(x) = 2x - x^2 + c(x - \frac{1}{4})(x - \frac{1}{2})(x - \frac{3}{4}).$$

Monte Carlo estimates of the power are summarized in Table 1 for different c and different bandwidth h.

The bandwidth has an influence on the level. We have on purpose selected a range of bandwidth wider than the fluctuation of crossvalidated bandwidth which was on average 0.23. For bandwidths around this value the level was hold at 0.05, the more extreme smoothing parameters led to an under- and over-estimation of the level respectively.

In case of using a data adaptative bandwidth $\hat{h}$ the randomness of $\hat{h}$ might also affect the level. To capture this in the bootstrap resampling one could also use a data adaptive bandwidth $\hat{h}^*$, based on the bootstrap sample, in every bootstrap loop. For a discussion of this procedure in a related context see chapter 3 in Mammen (1992b). In our simulations the bandwidth h is fixed and nonrandom.

Table 1:
Monte Carlo estimates of the
power for the regression function
$m(x) = 2x - x^2 + c(x - 1/4)(x - 1/2)(x - 3/4)$.
The level has been chosen to be 0.05.

h, c	0.0	0.5	1.0	2.0
0.10	0.105	0.157	0.325	0.784
0.20	0.054	0.120	0.252	0.795
0.25	0.053	0.099	0.263	0.765
0.30	0.039	0.078	0.225	0.714

Figure 5. Working, Linear and kernel smoother fit for a food expenditure Engel curve. The linear fit has label 1, the working curve has label 2 and the nonparametric kernel smoother has label 3.

Figure 5 shows a linear fit and a working fit and a nonparametric smoothing estimate for the Engel curve for food as a function of total expenditure. The Engel curve is the mean expenditure curve for a certain good. Theoretical economists and econometricians are interested in the form of this regression curve since this form has consequences on theoretical and social questions, see Engel (1895). For the particularly chosen parametric form we refer to Leser (1963). The data came from the Family Expenditure Survey (1968-1983). The data used for this figure was from 1969, the X and the Y data have been rescaled by the mean of X. The quartic kernel has been used with a bandwidth of $h = 0.2$. The bootstrap test rejected the linear regression model for all considered bandwidths for both variables. For food the working curve has been rejected only for some small bandwidths. This is of course due to the fact of an inflated value of the statistic T_n. The crossvalidated bandwidth for this data set lies at $h \approx 0.2$. For a picture of the crossvalidation function see Härdle (1990, Section 5). A summary of the bootstrap estimates of the observed critical values is given in Table 2.

Table 2:
Observed critical values for two parametric fits for the Family Expenditure data set. The number of bootstrap simulations is 100.

	Working	Working	Linear	Linear
h	Fuel	Food	Fuel	Food
0.05	.0	.0	.0	.0
0.10	.0	.0	.0	.0
0.15	.02	.08	.0	.0
0.20	.05	.38	.0	.0
0.25	.08	.57	.0	.0
0.30	.06	.62	.0	.0
0.35	.05	.55	.0	.0
0.40	.06	.55	.0	.0
0.45	.07	.54	.0	.0
0.50	.08	.51	.0	.0

(7.1) and (7.2) give

$$T_n = n\sqrt{h}\int_0^1 (\hat{m}_h(x) - \mathcal{K}_{h,n}m_{\hat{\theta}}(x))^2 dx$$
$$= n\sqrt{h}\int_0^1 (\hat{m}_h(x) - \mathcal{K}_{h,n}m_{\hat{\theta}}(x))^2 (\frac{\hat{f}_h(x)}{f(x)})^2 dx + o_p(1)$$
$$= n\sqrt{h}\int_0^1 \frac{(\frac{1}{n}\sum_{i=1}^n K_h(X_i - x)(m(X_i) + \varepsilon_i - m_{\hat{\theta}}(X_i)))^2}{f^2(x)} dx + o_p(1).$$

Now apply (P1) and $m(\bullet) = m_{\theta_0}(\bullet) + n^{-1/2}h^{1/4}\Delta_n(\bullet)$:
Then one gets

$$T_n = n\sqrt{h}\int_0^1 (U_{n,1}(x) + U_{n,2}(x) + U_{n,3}(x))^2 dx + o_p(1)$$

where

$$U_{n,1}(x) = \frac{\frac{1}{n}\sum_{i=1}^n K_h(X_i - x)n^{-\frac{1}{2}}h^{-\frac{1}{4}}\Delta_n(X_i)}{f(x)}$$
$$U_{n,2}(x) = \frac{\frac{1}{n}\sum_{i=1}^n K_h(X_i - x)\varepsilon_i}{f(x)}$$
$$U_{n,3}(x) = \frac{-\frac{1}{n}\sum_{i=1}^n K_h(X_i - x)\frac{1}{n}\sum_{j=1}^n g(X_i)^T h(X_j)\varepsilon_j}{f(x)}.$$

By straightforward calculations one gets

$$E \quad n\sqrt{h}\int_0^1 U_{n,3}^2(x)dx \to 0$$
$$E \quad n^2 h[\int_0^1 U_{n,i}(x)U_{n,j}(x)dx]^2 \to 0$$

for $1 \le i < j \le 3$.
This implies:

$$T_n = n\sqrt{h}\int_0^1 U_{n,1}^2(x) + U_{n,2}^2(x)dx + o_p(1)$$
$$= T_{n,1} + T_{n,2} + T_{n,3} + o_p(1)$$

where

$$T_{n,1} = n\sqrt{h}\int_0^1 [U_{n,1}^2(x)]dx$$
$$T_{n,2} = \frac{\sqrt{h}}{n}\int_0^1 \frac{\sum_{i=1}^n K_h(X_i - x)^2\varepsilon_i^2}{f^2(x)}dx$$
$$T_{n,3} = \frac{\sqrt{h}}{n}\int_0^1 \frac{\sum_{i\neq j} K_h(X_i - x)K_h(X_j - x)\varepsilon_i\varepsilon_j}{f^2(x)}dx \quad .$$

7. Proofs

W.L.O.G. we will give the proofs only for $d = 1$ and $\pi(x) \equiv 1$.

Proof of Proposition 1. First note that

$$\begin{aligned}\hat{f}_h(x) &= \frac{1}{n}\sum_{i=1}^{n} K_h(X_i - x) \\ &= f(x) + O_p(n^{-2/5}\sqrt{\log n}) \text{ (uniformly in } x),\end{aligned} \tag{7.1}$$

(see Collomb and Härdle (1986)) and that

$$\hat{m}_h(x) = m(x) + O_p(n^{-2/5}\sqrt{\log n}) \text{ (uniformly in } x), \tag{7.2}$$

(see Mack and Silverman (1982)) for $d = 1$. For $d > 1$ one shows for every $\eta > 0$ that

$$\hat{m}_h(x) = m(x) + O_p(n^{-2/(4+d)+\eta}) \text{ (uniformly in } x).$$

This can be proved as in Härdle (1990, Section 4) calculating the moments of $\hat{m}_h\hat{f}_h$ using (A5) and the Lipschitz continuity of the kernel.

First we show for $q > 0$

$$P(n^{-1}\sum_{n=1}^{n} K_h(X_i - x)(Y_i - m(X_i)) \geq Cn^{-2/(4+d)}\sqrt{\log n}) = O(n^{-q})$$

for C large enough. This follows by the following simple application of the Markov inequality. Choose $k > 2 + d/2$. Then for a constant $c > 0$ it holds with $A_n = n^{2/(4+d)-1}\sqrt{\log n}K_h(X_1 - x)\varepsilon_1$

$$\begin{aligned}&P(n^{-1}\sum_{i=1}^{n} K_h(X_i - x)(Y_i - m(X_i)) \geq Cn^{-2/(4+d)}\sqrt{\log n}) \\ &\leq \{E\exp([n^{2/(4+d)}\sqrt{\log n}][n^{-1}K_h(X_1 - x)\varepsilon_i])\}^n \exp(-C(\log n)) \\ &\leq [1 + E[A_n]^2 + \cdots + E[A_n]^k[1 + \exp(A_n)]]^n n^C \\ &\leq [1 + c(\log n)/n + o((\log n)/n)]^n n^{-C} \\ &\leq \exp(c\log n + o(\log n)n^{-C} \\ &\leq n^{c-C+o(1)} = O(n^{-q})\end{aligned}$$

Similarly one can treat $n^{-1}\sum_{i=1}^{n} K_h(X_i - x)(m(X_i) - Y_i)$, $\hat{f}_h(x) - f(x)$ and $f(x) - \hat{f}_h(x)$. Now note that $\hat{f}_h(x)$, f and $\partial/\partial x[n^{-1}\sum_{i=1}^{n} K_h(X_i - x)m(X_i)]$ are bounded by deterministic constants with polynomial growth and that with (A5)

$$\sup_x[\partial/\partial x[n^{-1}\sum_{i=1}^{n} K_h(X_i - x)Y_i] = O_p(h^{-2}\log n).$$

We will show

$$T_{n,1} = \int_0^1 \mathcal{K}_h \Delta_n(x)^2 dx + o_p(1) \tag{7.3}$$

$$T_{n,2} = b_h + o_p(1) \tag{7.4}$$

$$\mathcal{L}(T_{n,3}) \Rightarrow N(0, V) \quad \text{(weakly)} \quad . \tag{7.5}$$

But (7.3)–(7.5) entail the statement of Proposition 1. It remains to show (7.3)–(7.5).

Proof of (7.3): Arguing as above one sees that:

$$\begin{aligned} T_{n,1} &= n\sqrt{h} \int_0^1 \frac{(\frac{1}{n}\sum_{i=1}^n K_h(X_i - x) n^{-\frac{1}{2}} h^{-\frac{1}{4}} \Delta_n(X_i))^2}{f^2(x)} dx \\ &= \int_0^1 \frac{(\frac{1}{n}\sum_{i=1}^n K_h(X_i - x)\Delta_n(X_i))^2}{\hat{f}^2(x)} dx + o_p(1) \\ &= \int_0^1 \mathcal{K}_h \Delta_n(x)^2 dx + o_p(1) \quad . \end{aligned}$$

Proof of (7.4): First note

$$\begin{aligned} E \;\; T_{n,2} &= E\;\sqrt{h} \int_0^1 \frac{K_h(X_1 - x)^2}{f^2(x)} \sigma^2(x) dx \\ &= \int_0^1 \int_0^1 \sqrt{h} \frac{K_h(u - x)^2}{f^2(x)} f(u)\sigma^2(x) dx \\ &= b_h + o(1) \quad . \end{aligned}$$

Because of $\text{var}(T_{n,2}|X_1, \ldots, X_n) = O_p(\frac{1}{h^3 n}) = o_p(1)$ this implies (5.2).

Proof of (7.5): Put

$$W_{ijn} = \begin{cases} \sqrt{h}n \int_0^1 \frac{K_h(X_i - x)K_h(X_j - x)}{f^2(x)} dx \varepsilon_i \varepsilon_j, & \text{if i} \neq \text{j} \;, \\ 0, & \text{otherwise} \;. \end{cases}$$

Then

$$T_{n,3} = \sum_{i,j} W_{ijn} \quad .$$

According to Theorem 2.1 in de Jong (1987) for (7.5) it suffices to prove :

$$\text{var}(T_{n,3}) \to V \tag{7.6}$$

$$\max_{1 \le i \le n} \sum_{j=1}^n \text{var}(W_{ijn})/\text{var}(T_{n,3}) \to 0 \tag{7.7}$$

$$E \;\; T_{n,3}^4/(\text{var}(T_{n,3}))^2 \to 3 \quad . \tag{7.8}$$

The proof of (7.6) is straightforward. (7.7) follows from

$$\mathrm{var}(W_{ijn}) = O(\frac{1}{n^2 h}) \quad .$$

For the proof of (7.8) note that

$$\begin{aligned} E\ \ T_{n,3}^4 &= 12\sum\nolimits^{\neq} EW_{ijn}^2 W_{kln}^2 \\ &\quad + 8\sum\nolimits^{\neq} EW_{ijn}^4 + 48\sum\nolimits_{i,j,k,l}^{\neq} EW_{ijn}W_{jkn}W_{kln}W_{lin} \\ &\quad + 192\sum\nolimits^{\neq} EW_{ijn}W_{ikn}^2 W_{jkn} \\ &= 3\mathrm{var}(T_{n,3})^2 + o(1) \quad , \end{aligned}$$

(Here $\sum^{\neq}$ denotes summation over only all pairwise different indices) because of

$$E\ \ W_{12n}^4 = O(\frac{1}{n^4 h^2}) = o(\frac{1}{n^2}) \quad .$$

$$\begin{aligned} &E\ \ W_{12n}W_{23n}W_{34n}W_{41n} \\ &= \frac{h^2}{n^4}\int \frac{K_h(u_1-x_1)K_h(u_2-x_1)K_h(u_2-x_2)\ldots K_h(u_1-x_4)}{f^2(x_1)\ldots f^2(x_4)} \\ &\cdot f(u_1)\ldots f(u_4)dx_1\ldots dx_4 du_1\ldots du_4 \ \cdot\ O(1) \\ &= O(\frac{h^2}{n^4})\int K_h(u_1-u_2)K_h(u_2-u_3)K_h(u_3-u_4)K_h(u_4-u_1)du_1\ldots du_4 \\ &= O(\frac{h^2}{n^4})\int (K_h^{(2)}(u))^2 du = O(\frac{h}{n^4}) = o(\frac{1}{n^4}) \end{aligned}$$

and

$$\begin{aligned} E\ \ W_{12n}W_{23n}^2 W_{31n} &= O(\frac{h^2}{n^4})\int K_h(u)^2 K_h^{(2)}(u)du \\ &= O(\frac{1}{n^4}) = o(\frac{1}{n^3}). \end{aligned}$$

$$\mathrm{var}(T_{n,3}) = 2\sum\nolimits^{\neq} E\ \ W_{ijn}^2 \quad .$$

We will give only indications of the proofs of Proposition 2 and Theorems 1 and 2.

Proof of Proposition 2:

First note that

$$nh^{1/2}\int (\hat{m}_h - m_{\hat{\theta}})^2 \pi = nh^{1/2}\int \left(V_{n,1}(x)+\cdots+V_{n,4}(x)\right)^2 \pi(x)dx,$$

where

$$V_{n,1}(x) = \hat{m}_h(x) - \mathcal{K}_{h,n}m(x),$$
$$V_{n,2}(x) = \mathcal{K}_{h,n}m(x) - m(x),$$
$$V_{n,3} = m(x) - m_\theta(x),$$
$$V_{n,4}(x) = m_\theta(x) - m_{\hat{\theta}}(x).$$

$nh^{1/2}\int V_{n,1}(x)^2\pi(x)dx$ can be expanded as $nh^{1/2}\int U_{n,1}(x)^2\pi(x)dx$ in the proof of Proposition 1. The other terms can be treated similarly. For instance, one gets

$$2nh^{1/2}\int V_{n,2}(x)V_{n,3}(x)\pi(x)dx$$

$$= nh^{1/2}c_K c_n \int h^2\rho(x)\Delta(x)\pi(x)dx + o_p(1)$$

$$= n^{1/2}h^{5/2}c_K \int \rho(x)\Delta(x)\pi(x)dx + o_p(1).$$

Proof of Theorem 1:

As in the proof of Proposition 1 (but by a little bit finer arguments) one shows first for $T^* = T^{*,N}$(or $T^* = T^{*,A}$) that (note that $\Delta_n = 0$):

$$T^* = \frac{\sqrt{h}}{n}\sum_{1\le i,j\le n}\frac{K^{(2)}(X_i^* - X_j^*)\eta_i^*\eta_j^*}{f(X_i^*)f(X_j^*)} + \Gamma$$

for a random variable Γ with

$$E^*\Gamma^2 = o_p(1).$$

Here E^* denotes the conditional expectation given $\{(X_i, Y_i)\}_{i=1}^n$.$\{(X_i^*, \eta_i^*)\}$ are drawn (with replacement) out of the set $\{(X_i, \eta_i)\}$ where

$$\eta_i = \varepsilon_i - \frac{1}{n}\sum_{j=1}^n g^T(X_i)h(X_j)\varepsilon_j$$
$$\text{if} \quad T^* = T^{*,N}$$

$$\eta_i = \varepsilon_i + (m - \mathcal{K}_{h,n}m)(X_i) - \frac{1}{n}\sum_{j=1}^n \frac{K_h(X_i - X_j)\varepsilon_j}{\hat{f}_h(X_j)}$$
$$\text{if} \quad T^* = T^{*,A}.$$

Define

$$A_{ij} = \frac{K_h(X_i^* - X_j^*)\eta_i^*\eta_j^*}{f(X_i^*)f(X_j^*)}.$$

Put $a = E^*(A_{ij})$ for $i \neq j$.
Then one gets with $a = O_p(n^{-1}h^{-1})$:

$$\begin{aligned}
\mathrm{var}^*(T^*) &= \frac{h}{n^2}\mathrm{var}^*(\sum_{i,j} A_{ij}) + o_p(1) \\
&= \frac{h}{n^2}\mathrm{var}^*(\sum_{i\neq j} A_{ij}) + o_p(1) \\
&= \frac{h}{n^2}E^*(\sum_{i\neq j} A_{ij} - a)^2 + o_p(1) \\
&= \frac{h}{n^2}E^*2\sum_{i\neq j}(A_{ij} - a)^2 \\
&+ \frac{h}{n^2}E^*4\sum_{i\neq j\neq k\neq i}(A_{ij} - a)(A_{jk} - a) + o_p(1) \\
&= 2hE^*A_{12}^2 + 4hnE^*A_{12}A_{23} + o_p(1) \\
&= 2\frac{h}{n^2}\sum_{i,j}\frac{K_h(X_i - X_j)^2}{f^2(X_i)f^2(X_j)}\eta_i^2\eta_j^2 \\
&+ 4\frac{h}{n^2}\sum_{i,j,k}\frac{K_h(X_i - X_j)K_h(X_j - X_k)}{f(X_i)f^2(X_j)f(X_k)}\eta_i\eta_j^2\eta_k \\
&+ o_p(1).
\end{aligned}$$

This gives for $T^* = T^{*,N}$ by straightforward calculations:

$$\begin{aligned}
\mathrm{var}^*(T^{*,N}) &= 6\frac{h}{n^2}\sum_{i,j}\frac{K_h(X_i - X_j)^2}{f^2(X_i)f^2(X_j)}\varepsilon_i^2\varepsilon_j^2 \\
&+ o_p(1).
\end{aligned}$$

But the right term converges in probability to its expectation. This proves the first statement of Theorem 1. The second statement follows by a very lengthy evaluation of the above approximation of $\mathrm{var}^*(T^{*,A})$.

Proof of Theorem 2:

The proof goes along the lines of Proposition 1. Especially (A5) entails $\sup_i \varepsilon_i^2 = O_p(\log n)$ and $E|\varepsilon_i|^8 < const.$ (unif. in i and n). This can be used to prove the two conditions of the theorem of de Jong (1987).

Acknowledgement. We would like to express our gratitude to the Deutsche Forschungsgemeinschaft (SFB 303, SFB 123), CentER and CORE for financial support. We also thank B. Turlach for assistance in programming and are grateful to an associate editor and a referee for number of remarks and suggestions.

References

Azzalini, A., Bowman, A.W. and Härdle, W. (1989). On the use of nonparametric regression for model checking. *Biometrika*, 76, 1–11.

Beran, R. (1986). Discussion to Wu, C.F.J.: Jackknife, bootstrap and other resampling methods in regression analysis. *Annals of Statistics*, 14, 1295–1298.

Carroll, R.J. (1982). Adapting for heteroscedasticity in linear models. *Annals of Statistics*, 10, 1224–1233.

Cao-Abad, R. and Gonzalez-Manteiga, W. (1990). On bootstrapping regression curves. Manuscript.

Cessie, S. le and van Houwelingen, J.C. (1991). A goodness-of-fit test for binary regression models, based on smoothing methods. *Biometrics*, 47, 1267-1282.

Cleveland, W.S. and S.J. Devlin, (1988). Locally weighted regression: An approach to regression analysis by local fitting. *Journal of the American Statistical Association*, 83, 596–610.

Collomb, G., (1981). Estimation non-paramétrique de la régression: Revue Bibliographique. *International Statistical Review*, 49, 75–93.

Collomb, G. and 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.

Cox, D., Koh, E., Whaba, G. and Yandell, B. (1988). Testing the (parametric) null model hypothesis in (semiparametric) partial and generalized spline models. *Annals of Statistics*, 16, 113–119.

Cox, D. and Koh, E. (1989). A smoothing spline based test of model adequacy in polynomial regression. *Annals of the Institute of Statistical Mathematics*, 41, 383–400.

de Jong, P. , (1987). A central limit theorem for generalized quadratic forms. *Probability Theory and Related Fields* 75, 261–277.

Ducpetiaux, E., (1855). *Budgets économiques des classes ouvrières en Belgique.* Brussels.

Durbin, J. and Knott, M. (1972). Components of Cramer–von Mises Statistics I. *Journal of the Royal Statistical Society, Series B*, 34, 290–307.

Engel, E. (1895). Die Produktions- und Consumptionsverhältnisse des Königsreichs Sachsen. Abdruck in "*Bulletin de l'Institut International de Statistique*", 9, 1–54.

Eubank, R. and Spiegelman, C. (1990). Testing the goodness-of-fit of linear models via nonparametric regression techniques. *Journal of the American Statistical Association*, 85, 387–392.

Family Expenditure Survey, Annual Base Tapes. (1968-1983). Department of Employment, Statistics Division, Her Majesty's Stationery Office, London 1968-1983. The data utilized in this paper were made available by the ESRC Data Archive at the University of Essex.

Gasser, T., Köhler, W., Müller, H.G., Largo, R., Molinari, L. and Prader, A. (1985). Human height growth: Correlational and multivariate structure of velocity and acceleration. *Annals of Human Biology*, 12, 501–515.

Härdle, W. (1990).. *Applied nonparametric regression.* Econometric Society Monograph Series 19, Cambridge University Press.

Härdle, W. and Marron, J.S. (1990). Semiparametric comparison of regression curves. *Annals of Statistics*, 18, 63–89.

Hildreth, C. (1954). Point estimates of ordinates of concave functions. *Journal of the American Statistical Association*, 49, 598–619.

Hildenbrand, K. and Hildenbrand, W. (1986). On the mean income effect: a data analysis of the U.K. family expenditure survey. In: *Contributions to Mathematical Economics.* (W.Hildenbrand , A.Mas-Colell, eds.) North-Holland.

Konakov, V.D. and Piterbarg, V.I. (1984). On the convergence rate of maximal deviation distribution for kernel regression estimators. *Journal of Multivariate*

Analysis, 15, 279–294.

LaRiccia, V.N. (1991). Smooth Goodness-of-Fit Tests: A Quantile Function Approach. Journal of the American Statistical Association, 86, 427–431.

Leser, C.E. (1963). Forms of Engel functions. *Econometrica*, 31, 694–703.

Liu, R. (1988). Bootstrap procedures under some non i.i.d. models. *Annals of Statistics*, 16, 1696–1708.

Mack, Y.P. and Silverman, B.W. (1982). Weak and strong uniform consistency of kernel regression estimates. *Zeitschrift für Wahrscheinlichkeitstheorie und verwandte Gebiete*, 61, 405–415.

Mammen, E. (1992a). Bootstrap and wild bootstrap for high dimensional linear models. *Annals of Statistics*, to appear.

Mammen, E. (1992b). When does bootstrap work: asymptotic results and simulations. Lecture Notes in *Statistics* 77, Springer Verlag, Berlin, Heidelberg and New York.

Milbrodt, H. and Strasser, H. (1990). On the asymptotic power of the two-sided Kolmogorov–Smirnov test. *Journal of Statistical Planning and Inference*, 26, 1-23.

Munson, P.J. and Jernigan, R.W. (1989). A cubic spline extension of the Durbin–Watson test. *Biometrika*, 76, 39 47.

Nadaraya, E.A. (1964). On estimating regression. *Theory of Probability and its Applications*, 10, 186 -190.

Neuhaus, G. (1986). A class of quadratic goodness of fit tests. Preprint.

Neuhaus, G. (1988). Addendum to "Local asymptotics for linear rank statistics with estimated score functions". *Annals of Statistics*, 16, 1342–1343.

Preece, M.A. and Baines, M.J. (1978). A new family of mathematical models describing the human growth curve. *Annals of Human Biology*, 5, 1–24.

Staniswalis, J.G. and Severini, T.A. (1991). Diagnostics for assessing regression models. *Journal of American Statistical Association* 86, 684-692.

Tukey, J.W. (1961). Curves as parameters and touch estimation. *Proceedings of the 4th Berkeley Symposium*, 681–694.

Watson, G.S. (1964). Smooth regression analysis. *Sankhyā, Series A*, 26, 359–372.

Wu, C.F.J. (1986). Jackknife, bootstrap and other resampling methods in regression analysis (with discussion). *Annals of Statistics*, 14, 1261–1295.

Yanagimoto, T. and Yanagimoto, M. (1987). The use of marginal likelihood for a diagnostic test for the goodness of fit of the simple linear regression model. *Technometrics*, 29(1), 95–101.

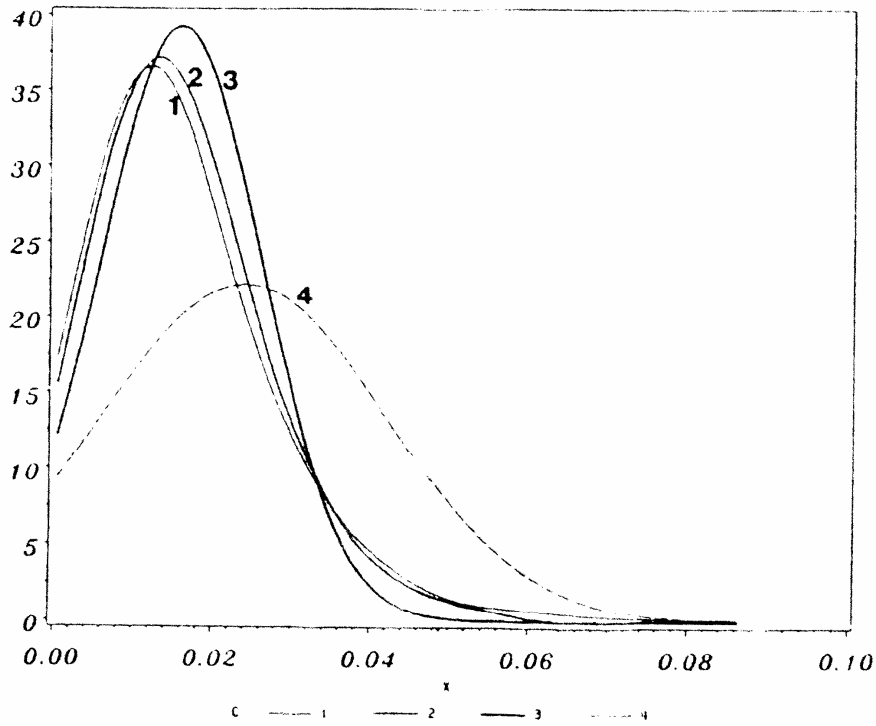


FIG. 1. *Monte Carlo density of T_n, bootstrap density estimate and normal density. The thin line 1 is the Monte Carlo density of T_n; line 2 is the kernel density of T_n^* from ONE bootstrap sample. Line 3 is the normal density with asymptotic mean and variance, given in Proposition 1; line 4 is the normal density approximation with estimated mean and variance. The parametric model consists of constant functions.*

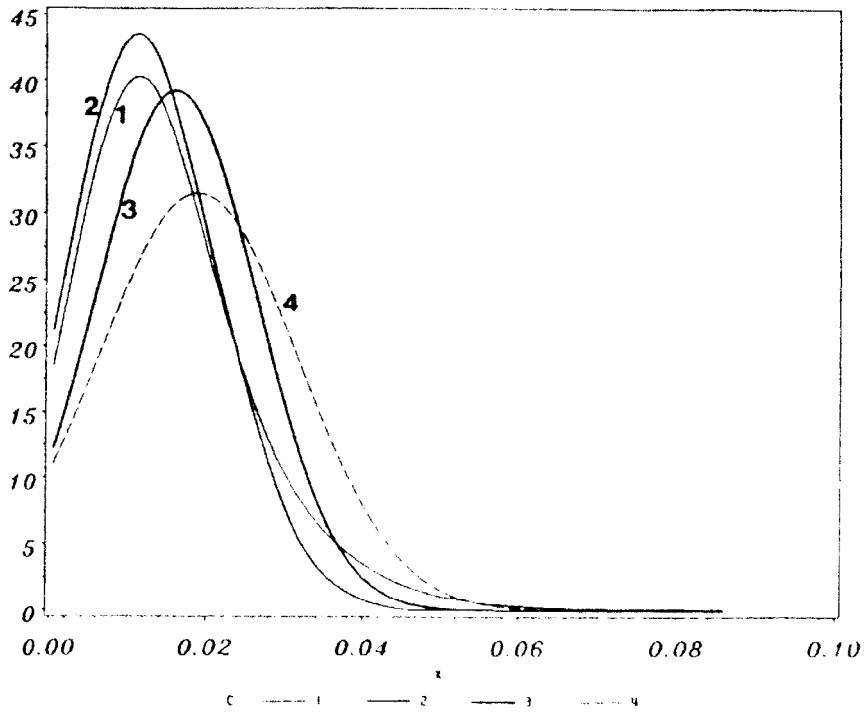


FIG. 2. *Monte Carlo density of T_n, bootstrap density estimate and normal density. The thin line 1 is the Monte Carlo density of T_n; line 2 is the kernel density of T_n^* from ONE bootstrap sample. Line 3 is the normal density with asymptotic mean and variance, given in Proposition 1; line 4 is the normal density approximation with estimated mean and variance. The parametric model consists of linear functions.*

Härdle, W. and Mammen, E. (1993) Comparing Nonparametric versus Parametric Regression Fits

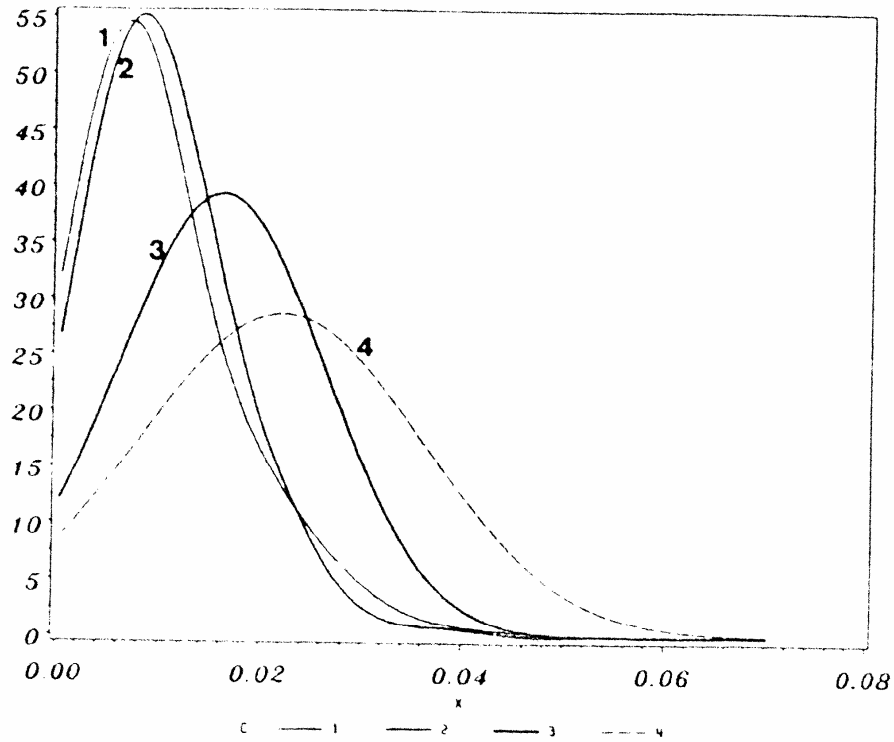


FIG. 3. *Monte Carlo density of T_n, bootstrap density estimate and normal density. The thin line 1 is the Monte Carlo density of T_n; line 2 is the kernel density of T_n^* from ONE bootstrap sample. Line 3 is the normal density with asymptotic mean and variance, given in Proposition 1; line 4 is the normal density approximation with estimated mean and variance. The parametric model consists of quadratic polynomials.*

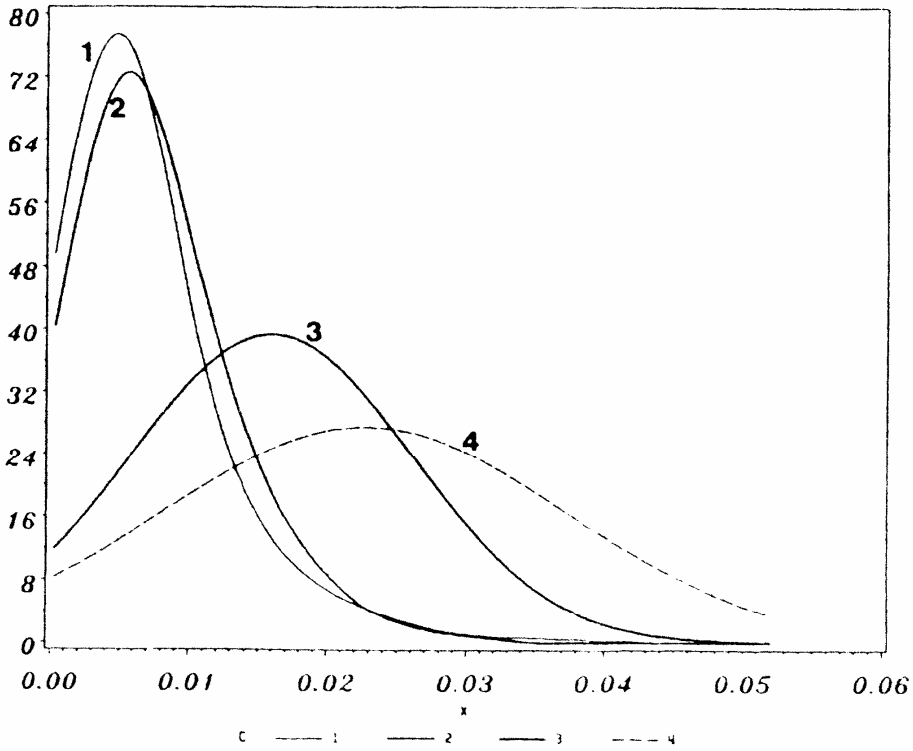


FIG. 4. *Monte Carlo density of T_n, bootstrap density estimate and normal density. The thin line 1 is the Monte Carlo density of T_n; line 2 is the kernel density of T_n^* from ONE bootstrap sample. Line 3 is the normal density with asymptotic mean and variance, given in Proposition 1; line 4 is the normal density approximation with estimated mean and variance. The parametric model consists of cubic polynomials.*

Härdle, W. and Mammen, E. (1993) Comparing Nonparametric versus Parametric Regression Fits

Econometric Theory, 10, 1994, 821–848. Printed in the United States of America.

TESTING A PARAMETRIC MODEL AGAINST A SEMIPARAMETRIC ALTERNATIVE

JOEL L. HOROWITZ
University of Iowa

WOLFGANG HÄRDLE
Institut für Statistik und Ökonometrie
Humboldt Universität zu Berlin

This paper describes a method for testing a parametric model of the mean of a random variable Y conditional on a vector of explanatory variables X against a semiparametric alternative. The test is motivated by a conditional moment test against a parametric alternative and amounts to replacing the parametric alternative model with a semiparametric model. The resulting semiparametric test is consistent against a larger set of alternatives than are parametric conditional moments tests based on finitely many moment conditions. The results of Monte Carlo experiments and an application illustrate the usefulness of the new test.

1. INTRODUCTION

Consider a parametric model for the mean of a scalar random variable Y conditional on a random variable $X \in \mathcal{R}^L$ $(L \geq 1)$:

$$E(Y|X = x) = f(x, \theta), \tag{1}$$

where f is a known function and $\theta \in \mathcal{R}^K$ $(K \geq 1)$ is a parameter whose value must be estimated from data. For example, f might be the mean function in a linear or nonlinear regression model, or it might be the probability that $Y = 1$ conditional on $X = x$ in a parametric binary response model. The problem addressed in this paper is to test the hypothesis that (1) is true for the specified function f and some θ.

One way of testing (1) is to specify a parametric alternative to it and test $f(x, \theta)$ against the alternative. Most familiar methods for testing (1) against a parametric alternative belong to a large class called conditional moments tests [14]. These tests can have high power against specific alternatives, but a parametric conditional moments test based on finitely many moment con-

We thank Donald Andrews, Whitney Newey, Theo Nijman, and Tom Stoker for helpful comments. Parts of this research were carried out while both authors were visitors at the Center for Economic Research, Catholic University of Brabant, Tilburg, The Netherlands. The research of Joel L. Horowitz was supported in part by NSF grants DMS-9208820 and SBR-9307677.

 0266-4666/94 $5.00 + .00

Econometric Theory,V 10, No. 5, 821 - 848

ditions is not consistent against all alternatives. In particular, a test of $f(x,\theta)$ against a parametric alternative model may be inconsistent if the alternative is misspecified.

A second possibility is to compare the parametric model with a nonparametric estimate of $E(Y|X=x)$. Let $\hat{\theta}_n$ denote a $n^{1/2}$-consistent estimator of θ in (1) based on a random sample of the distribution of (Y,X). If (1) is true, the nonparametric estimate and $f(x,\hat{\theta}_n)$ are equal up to random sampling error. See Eubank and Spiegelman [3], Gozalo [4], Härdle and Mammen [8], Hong and White [9], le Cessie and van Houwelingen [12], Whang and Andrews [19], Wooldridge [20], Yatchew [21], and Zheng [22] for specification tests based on this idea. Bierens [2] gives a conditional moments test of a parametric model against a nonparametric alternative. These tests are consistent in all directions. However, some apply only to restricted classes of functions $f(x,\theta)$. For example, Eubank and Spiegelman [3] assume that $f(x,\theta)$ is linear and x is scalar. Other tests have characteristics that can cause them to have low power or other kinds of poor behavior in finite samples. For example, the tests of Gozalo [4], Härdle and Mammen [8], Hong and White [9], and le Cessie and van Houwelingen [12] lose power through the so-called curse of dimensionality [11] if $L > 1$. The tests of Whang and Andrews [19] and Yatchew [21] require splitting the sample into two equal parts, which reduces power and can result in poor small-sample behavior.

This paper describes a test that aims at avoiding these problems while achieving consistency against a larger set of alternatives than is the case with parametric conditional moments tests based on finitely many moment conditions. The intuition behind the test is simple. If $E(Y|X=x)=f(x,\theta)$, then

$$E[Y|f(x,\theta)=f]=f. \tag{2}$$

Therefore, a nonparametric estimate of $E[Y|f(X,\hat{\theta}_n)=f]$, considered as a function of f, differs from a 45° line only by random sampling error. One can test (1) by determining whether the difference between the nonparametric estimate and the 45° line is larger than can be explained by random sampling error.

More generally, consider the model

$$E(Y|X=x)=F[v(x,\theta)], \tag{3}$$

where F and v are known functions. If (3) is correct, nonparametric estimation of $E[Y|v(X,\hat{\theta}_n)=v]$ gives an estimate of $F(v)$. Thus, (3) can be tested by comparing the nonparametric estimate of $E[Y|v(X,\hat{\theta}_n)=v]$ with $F(v)$. One way of obtaining (3) is to specify $v(x,\theta)=f(x,\theta)$ and $F(v)=v$, but other specifications may be useful in applications. For example, suppose the parametric model to be tested has the form (3) with F a nonmonotonic function. If the model is misspecified, it is possible that

$$E\{Y|F[v(X,\theta)]=f\}=f, \tag{4a}$$

Econometric Theory,V 10, No. 5, 821 - 848

whereas

$$E[Y|v(X,\theta) = v] \neq F(v). \tag{4b}$$

In this case, comparison of a nonparametric estimate of $E\{Y|F[v(X,\theta)] = f\}$ with f yields an inconsistent test, whereas comparison of a nonparametric estimate of $E[Y|v(X,\theta) = v]$ with $F(v)$ yields a consistent test.

A test of (1) obtained by comparing a nonparametric estimate of $E[Y|v(X,\hat{\theta}_n) = v]$ with $F(v)$ avoids the curse of dimensionality by using the index function $v(x,\theta)$ to aggregate a multidimensional x. Because one can always set $v(x,\theta) = f(x,\theta)$ and $F(v) = v$, any model of the form (1) can be placed into the single-index form of (3). Thus, the test is not restricted to models that can be estimated in single-index form.

The remainder of this paper describes a formal test of (1) that consists of comparing a nonparametric estimate of $E[Y|v(X,\hat{\theta}_n) = v]$ with $F(v)$. We call this a test of the parametric model (1) against a semiparametric alternative because the alternatives against which the parametric model is tested and against which the test is consistent have the form $E[Y|v(X,\theta) = v] = H(v)$, where H is an unknown function but $v(x,\theta)$ is known up to the finite-dimensional parameter θ. Because the semiparametric alternative may not include the true mean of Y conditional on X, there are directions in which the semiparametric test is inconsistent. However, in a sense that is defined in Section 2, the test is consistent against a larger set of alternatives than are parametric conditional moments tests based on finitely many moments. The results of Monte Carlo experiments and an application based on real data illustrate the usefulness of the semiparametric test.

The paper is organized as follows. The test statistic is presented in Section 2, and its asymptotic distributions under the null hypothesis and local alternatives are derived. Section 3 presents the results of the Monte Carlo experiments and the application. Concluding comments are presented in Section 4. The proofs of theorems are in the Appendix.

2. THE TEST STATISTIC AND ITS ASYMPTOTIC DISTRIBUTION

2.1. The Null and Alternative Hypotheses

Formally the null hypothesis that we test is

$$H_0\colon E[Y|v(X,\theta) = v] = F(v), \tag{5}$$

where Y is a scalar random variable, $X \in \mathcal{R}^L$, F and $v(\cdot,\cdot)$ are known real functions, and $\theta \in \mathcal{R}^K$ is a parameter whose value is unknown and estimated from data. For example, if Y follows a linear-index binary probit model under H_0, F and $v(x,\theta)$, may be specified as the cumulative normal

Horowitz, J. and Härdle, W. (1994)
Testing a Parametric Model against Semiparametric Alternatives.

Econometric Theory,V 10, No. 5, 821 - 848

distribution function and $\theta'x$, respectively. As was discussed in Section 1, (1) can always be put into the form (5). The alternative hypothesis is

$$H_1\colon E[Y|v(X,\theta) = v] = H(v), \tag{6}$$

where H is an unknown function.

$E[Y|v(X,\theta) = v] = F(v)$ is a necessary but not sufficient condition for $E(Y|X = x) = F[v(x,\theta)]$. It is possible that $E[Y|v(x,\theta) = v] = F(v)$ but $E(Y|X = x) \neq F[v(x,\theta)]$, in which case the test of (1) presented here is inconsistent. This possibility is discussed further in Section 2.4.

2.2. Motivation

Suppose for the moment that H and θ were known. Consider a conditional moments test of H_0 against H_1 based on the following moment condition, which is assumed to hold under H_0:

$$E\rho(X,\theta)\{Y - F[v(X,\theta)]\} = 0,$$

where ρ is a scalar function. Let $\{Y_i, X_i : i = 1, \ldots, n\}$ be a random sample of (Y,X). Following Newey [14], the conditional moments test statistic is proportional to

$$S_n = n^{-1/2} \sum_{i=1}^{n} \rho(X_i,\theta)\{Y_i - F[v(X_i,\theta)]\}.$$

Under H_0, $E(S_n) = 0$. Under H_1, $E(S_n) = n^{1/2}E\rho(X,\theta)\{H[v(X,\theta)] - F[v(X,\theta)]\} \equiv \mu$. The test can be expected to have power against H_1 only if $\mu \neq 0$. This happens if

$$\rho(x,\theta) = w[v(x,\theta)]\{H[v(x,\theta)] - F[v(x,\theta)]\},$$

where $w(\cdot)$ is a nonnegative weight function that is chosen so that

$$Ew[v(X,\theta)]\{H[v(X,\theta)] - F[v(X,\theta)]\}^2 > 0. \tag{7}$$

Thus, the conditional moments test in this simple case can be based on the statistic

$$S_n^* = n^{-1/2} \sum_{i=1}^{n} w[v(X_i,\theta)]\{Y_i - F[v(X_i,\theta)]\}\{H[v(X_i,\theta)] - F[v(X_i,\theta)]\}. \tag{8}$$

Since H and θ are unknown, one might consider forming a test of H_0 against H_1 by replacing H and θ in (8) with consistent estimators. This is the approach taken here. We replace θ with the $n^{1/2}$-consistent estimator $\hat{\theta}_n$ and $H[v(X_i,\theta)]$ with a kernel nonparametric regression estimator of

Econometric Theory,V 10, No. 5, 821 - 848

$E[Y|v(X,\hat{\theta}_n) = v(X_i,\hat{\theta}_n)]$. Denote this estimator by $\hat{F}_{ni}[v(X_i,\hat{\theta}_n)]$. The test statistic is

$$T_n = h^{1/2}\sum_{i=1}^{n} w[v(X_i,\hat{\theta}_n)]\{Y_i - F[v(X_i,\hat{\theta}_n)]\}$$
$$\times\{\hat{F}_{ni}[v(X_i,\hat{\theta}_n)] - F[v(X_i,\hat{\theta}_n)]\},$$

where h is the bandwidth used in the kernel nonparametric regression. The normalization factor of $h^{1/2}$ is needed because $(\hat{F}_{ni} - F_{ni}) = O_p[(nh)^{-1/2}]$ rather than $O_p(n^{-1/2})$ as in parametric models. It is shown below that, like the test based on (8), T_n is consistent against H_1 if (7) holds. In contrast to (8), however, T_n does not require a priori knowledge of H and θ.[1]

2.3. The Kernel Nonparametric Regression Estimator

We require $\hat{F}_{ni}(\cdot)$ to be independent of Y_i and asymptotically unbiased. Independence is achieved by omitting the observation (Y_i,X_i) from the computation of $\hat{F}_{ni}$. Asymptotic unbiasedness is achieved through the use of the jackknife-like method proposed by Schucany and Sommers [16] for nonparametric density estimation and Bierens [1] and Härdle [6] for nonparametric regression.[2] The resulting estimator is as follows.

Let $K(\cdot)$ be the kernel function used in the nonparametric regression. Assume that K is an order r kernel. That is, for each integer i between 0 and $r \geq 2$,

$$\int_{-\infty}^{\infty} u^i K(u)\,du = \begin{cases} 1 & \text{if } i = 0, \\ 0 & \text{if } 1 \leq i \leq r-1, \\ d_K \neq 0 & \text{if } i = r. \end{cases}$$

Let $h = cn^{-1/(2r+1)}$, where $c > 0$. Let $s = cn^{-\delta/(2r+1)}$, where $0 < \delta < 1$. Define

$$\hat{F}_{nhi}(v) = \sum_{\substack{j=1 \\ j\neq i}}^{n} Y_j K\left[\frac{v - v(X_j,\hat{\theta}_n)}{h}\right] \Big/ \sum_{\substack{j=1 \\ j\neq i}}^{n} K\left[\frac{v - v(X_j,\hat{\theta}_n)}{h}\right] \tag{9}$$

and

$$\hat{F}_{nsi}(v) = \sum_{\substack{j=1 \\ j\neq i}}^{n} Y_j K\left[\frac{v - v(X_j,\hat{\theta}_n)}{s}\right] \Big/ \sum_{\substack{j=1 \\ j\neq i}}^{n} K\left[\frac{v - v(X_j,\hat{\theta}_n)}{s}\right]. \tag{10}$$

The kernel nonparametric regression estimator used in T_n is

$$\hat{F}_{ni}(v) = [\hat{F}_{nhi}(v) - (h/s)^r \hat{F}_{nsi}(v)]/[1 - (h/s)^r]. \tag{11}$$

Bierens [1] derives the properties of this estimator and proves that it is asymptotically unbiased and has the optimal rate of convergence.

Econometric Theory,V 10, No. 5, 821 - 848

2.4. The Asymptotic Distribution of T_n

Define $\sigma^2(v) = \text{var}[Y|v(X,\theta) = v]$. The following theorem gives the asymptotic distribution of T_n under H_0.

THEOREM 1. *Under H_0 and Assumptions* 1–8 *of the Appendix, T_n is asymptotically distributed as $N(0, \sigma_T^2)$, where*

$$\sigma_T^2 = 2C_K \int_{-\infty}^{\infty} w(v)^2[\sigma^2(v)]^2\, dv$$

and

$$C_K = \int_{-\infty}^{\infty} K(u)^2\, du.$$

The proof of Theorem 1 is lengthy, but the concepts on which it is based are easily described. First, the rate of convergence in probability of the $n^{1/2}$-consistent estimator $\hat{\theta}_n$ is faster than the rate of convergence in probability of $\hat{F}_{ni}$, which is $(nh)^{-1/2}$. As a result, the asymptotic distribution of T_n is unaffected by replacing $\hat{\theta}_n$ with θ. Thus, $T_n = T_n^* + o_p(1)$, where

$$T_n^* = h^{1/2} \sum_{i=1}^{n} w[v(X_i,\theta)]\{Y_i - F[v(X_i,\theta)]\}\{F_{ni}[v(X_i,\theta)] - F[v(X_i,\theta)]\}$$

and $F_{ni}[v(X_i,\theta)]$ is the nonparametric regression estimator obtained by replacing $v(X_i,\hat{\theta}_n)$ with $v(X_i,\theta)$ in (9)–(11). See Lemmas 1–6 of the Appendix for the proof of this result. Second, it can be shown that T_n^* is asymptotically equivalent to a certain degenerate U statistic. See Lemma 7 and the proof of Theorem 1 in the Appendix. Although degenerate U statistics ordinarily are asymptotically distributed as linear combinations of χ^2 variates (see Serfling [17], for example), the one corresponding to T_n^* has a special form that causes it to be asymptotically normally distributed by a central limit theorem of Hall [5]. Theorem 1 is a consequence of the asymptotic normality of this U statistic.

Let $\hat{\sigma}_T^2$ be a consistent estimator of σ_T^2, and let $\hat{\sigma}_T = (\hat{\sigma}_T^2)^{1/2}$. It follows from Theorem 1 that H_0 can be accepted or rejected at the ζ level according to whether $T_n/\hat{\sigma}_T$ exceeds the $1 - \zeta$ quantile of the standard normal distribution. The proposed test is one-sided because, as is shown in Theorem 2 later, T_n diverges to $+\infty$ under alternative hypotheses against which it is consistent. In addition, as is shown in Theorem 3 later, the mean of the asymptotic distribution of T_n is nonnegative under local alternative hypotheses. Let $\hat{\sigma}^2(v)$ be a consistent estimator of $\sigma^2(v)$. Then, under Assumptions 1–8 of the Appendix σ_T^2 is estimated consistently by

$$\hat{\sigma}_T^2 = (2C_K/n) \sum_{i=1}^{n} w[v(X_i,\hat{\theta}_n)]^2\{\hat{\sigma}^2[v(X_i,\hat{\theta}_n)]\}^2/\hat{p}_{nhi}[v(X_i,\hat{\theta}_n)],$$

Econometric Theory,V 10, No. 5, 821 - 848

where $\hat{p}_{nhi}$ is the following nonparametric estimator of the probability density function of $v(X_i,\theta)$:

$$\hat{p}_{nhi}(v) = (nh)^{-1} \sum_{\substack{j=1 \\ j \neq i}}^{n} K\left[\frac{v - v(X_j,\hat{\theta}_n)}{h}\right].$$

Methods for estimating $\sigma^2(v)$ are discussed in Section 2.5.

The next theorem establishes consistency of T_n under H_1.

THEOREM 2. *Let Assumptions* 1–8 *of the Appendix hold. Suppose that* H_1 *is true and that* $Ew(V)\{[H(V) - F(V)]^2\} > 0$, *where* $V = v(X,\theta)$ *and* θ *is the probability limit of* θ_n. *Then,* $\operatorname{plim}_{n\to\infty} T_n = \infty$.

Suppose that H_0 is false and that $E(Y|X = x) = H^*(x)$ for some function H^*. Let $E_{X|v}$ denote expectation relative to the distribution of X conditional on $v(X,\theta) = v$. It follows from Theorem 2 that the test based on T_n is consistent if $E_{X|v}H^*(X) \neq F(v)$ on a subset of the support of $w[v(X,\theta)]$ that has positive probability. The test is inconsistent if $P\{H^*(x) = F[v(x,\theta)]\} < 1$ but $E_{X|v}H^*(X) = F(v)$ almost everywhere on the support of $w(\cdot)$.[3]

Although T_n is not consistent against all alternatives, there is a sense in which it is consistent against a larger set of alternatives than are parametric conditional moments tests based on finitely many moment conditions. Specifically, T_n is consistent against all alternatives $H[v(x,\theta)]$ such that $Ew[v(X,\theta)]\{H[v(X,\theta)] - F[v(X,\theta)]\}^2 > 0$, whereas a parametric conditional moments test is not. To see this, observe that if (3) is true, then

$$E\rho(X,\theta)\{Y - F[v(X,\theta)]\} = 0 \tag{12}$$

for any function $\rho \in \mathcal{R}^q$ for some finite $q > 0$. Accordingly, consider using the moment conditions of (12) to test (1). Suppose that $E[Y|v(X,\theta) = v] = H(v)$, where H satisfies $Ew[v(X,\theta)]\{H[v(X,\theta)] - F[v(X,\theta)]\}^2 > 0$ and $E\rho(X,\theta)\{H[v(X,\theta)] - F[v(X,\theta)]\} \neq 0$. Then, T_n and the conditional moments test based on (12) are both consistent against the alternative H. Now let $\Delta(v)$ be a scalar-valued function such that $E\rho(X,\theta)\Delta[v(X,\theta)] = 0$. Assume that $Ew[v(X,\theta)]\Delta[v(X,\theta)]^2 > 0$. Because ρ is finite-dimensional, there are infinitely many such functions Δ. Set $H^*(v) = F(v) + \Delta(v)$. Then, T_n is consistent against H^* but the conditional moments test based on (12) is not.

We now consider the distribution of T_n under local alternative hypotheses. Define the sequence of local alternatives $H_n[v(x,\theta)]$, by

$$E(Y|X = x) = H_n[v(x,\theta)] \equiv F[v(x,\theta)] + n^{-1/2}h^{-1/4}\Delta_n[v(x,\theta)],$$

where $\{\Delta_n : n = 1,2,\dots\}$ is a sequence of uniformly bounded functions that converges uniformly to a limit function $\Delta(v)$. Note that in this sequence $|H_n(v) - F(v)| = O(n^{-1/2}h^{-1/4})$ uniformly over v, whereas in tests of parametric models against local parametric alternatives the "distance"

Econometric Theory,V 10, No. 5, 821 - 848

between the null and local alternative hypotheses is $O(n^{-1/2})$. We assume that there are an estimator $\hat{\theta}_n$ of θ and a nonstochastic sequence $\{\bar{\theta}_n\}$ such that $n^{1/2}(\hat{\theta}_n - \bar{\theta}_n) = O_p(1)$ and $n^{1/2}h^{1/4}(\bar{\theta}_n - \theta)$ has a finite limit, γ, as $n \to \infty$. For example, if $\hat{\theta}_n$ is the least-squares estimator of θ under the false model $E(Y|X = x) = F[v(x,\theta)]$, $\bar{\theta}_n$ minimizes the expected value of the error sum of squares.

THEOREM 3. *Let Assumptions* 1–4 *and* 8–12 *of the Appendix hold as well as the parts of Assumption* 5 *that pertain to F. Define*

$$\Delta^*(x,\theta) = \Delta[v(x,\theta)] - F'[v(x,\theta)]E[\partial v(X,\theta)/\partial\theta' | v(X,\theta) = v(x,\theta)]\gamma.$$

Under the sequence of local alternative models H_n, T_n is asymptotically distributed as $N(\mu,\sigma_T^2)$, where $\mu = E[w[v(X,\theta)]\Delta^*(X,\theta)^2]$.

Theorem 3 implies that T_n has power against alternatives whose distance from H_0 is $O(n^{-1/2}h^{-1/4})$. If K is a second-order kernel, this distance is $O(n^{-9/20})$, which is close to the distance $O(n^{-1/2})$ that holds in tests against parametric alternative hypotheses. Subject to the regularity conditions given in the Appendix, the distance $O(n^{-1/2}h^{-1/4})$ can be made arbitrarily close to $O(n^{-1/2})$ by using a kernel K of sufficiently high order.

2.5. Choosing $w(\cdot)$ and $\hat{\sigma}^2(v)$

The regularity conditions in the Appendix require $w(\cdot)$ to be continuous and independent of the sample $\{Y_i, X_i\}$. They also require the support of $w(\cdot)$ to be contained within that of $v(X,\theta)$. The continuity requirement is not important in applications; with a finite sample there is no difference between the values of T_n obtained with a $w(\cdot)$ that has jump discontinuities and a $w(\cdot)$ in which the discontinuities have been "slightly" smoothed. The restriction on the support of $w(\cdot)$ can be important. Depending on how σ_T^2 is estimated, use of a $w(\cdot)$ whose support exceeds that of $v(X,\theta)$ may cause substantial overestimation of σ_T^2 and a corresponding loss of power. In practice, it can be difficult to choose a $w(\cdot)$ that satisfies the condition on support without looking at the data. We suggest using the observed values of $v(X_i,\hat{\theta}_n)$ to choose the support of $w(\cdot)$ but not otherwise adjusting $w(\cdot)$ to the data. In the Monte Carlo experiments and application described in Section 3, we found that the T_n test works well if w is chosen to be 1 over an interval that contains 95–99% of the observed values of $v(X,\hat{\theta}_n)$ and 0 elsewhere.

Another possibility is to choose w to maximize power against a specified sequence of local alternatives. There seems to be little advantage in doing this, however. If high power against a specific alternative is desired, one should use a parametric conditional moments test that has high power against this alternative.

The main consideration involved in estimating $\sigma^2(v)$ is that the estimator must be consistent under H_0 and, to avoid loss of power, should not become excessively large under H_1. For example, suppose that Y is homoskedastic so that $\text{var}[Y|v(X,\theta) = v] = \sigma^2$, where σ^2 is a constant. Two possible estimators of σ^2 are

$$\hat{\sigma}_1^2 = n^{-1} \sum_{i=1}^{n} \{Y_i - F[v(X_i, \hat{\theta}_n)]\}^2 \tag{13}$$

and

$$\hat{\sigma}_2^2 = n^{-1} \sum_{i=1}^{n} \{Y_i - \hat{F}_{ni}[v(X_i, \hat{\theta}_n)]\}^2. \tag{14}$$

Both of these estimators are consistent under H_0, but $\hat{\sigma}_1^2$ may be very large under H_1. Accordingly, the test based on T_n is likely to have higher power if $\hat{\sigma}_2^2$ is used.

If Y has heteroskedasticity of unknown form, $\sigma^2(v)$ can be estimated by the nonparametric regression of $\{Y_i - \hat{F}_{ni}[v(X_i, \hat{\theta}_n)]\}^2$ on $v(X_i, \hat{\theta}_n)$. In some cases, the form of heteroskedasticity of Y may be known, and this information can be used to estimate $\sigma^2(v)$. For example, if Y is a binary variable, $\text{var}[Y|v(X,\theta) = v] = P[Y = 1|v(X,\theta) = v]\{1 - P[Y = 1|v(X,\theta) = v]\}$. Therefore, $\sigma^2(v)$ can be estimated by $\hat{F}_{ni}[v(X_i, \hat{\theta}_n)]\{1 - \hat{F}_{ni}[v(X_i, \hat{\theta}_n)]\}$.

3. MONTE CARLO EXPERIMENTS AND AN APPLICATION

3.1. Monte Carlo Experiments

The purpose of the Monte Carlo experiments was to investigate the small-sample size and power of the test based on T_n. To provide a basis for judging whether the performance of the test is good or bad, we also computed the sizes and powers of Bierens' [2] test against a nonparametric alternative, the RESET test, and the most powerful test against the correct parametric alternative model.[4]

The hypothesis H_0 tested in the Monte Carlo experiments is

$$E(Y|X = x) = \beta_0 + v(x,\theta), \tag{15}$$

where X is a $L \times 1$ random variable, $L = 1$ or 3, $v(x,\theta) = \theta' x$, and β_0 is a constant. The data were generated by random sampling from the model

$$Y = \beta_0 + v(X,\theta) + 10b\phi[10v(X,\theta)] + u, \tag{16}$$

where ϕ is the standard normal density function, $\beta_0 = 1$, $\theta_i = L^{-1/2}$ $(i = 1, \ldots, L)$, b is a parameter whose value varies according to the experiment, $X \sim N(0,1)$, and $u \sim N(0,0.25)$. If $b = 0$, H_0 is true. Otherwise, H_0 is false, and $E[Y|v(X,\theta) = v]$ has the shape of a straight line with a bump centered

Econometric Theory,V 10, No. 5, 821 - 848

at $v = 0$. The height of the bump is governed by the value of b. Figure 1 illustrates the shape of $E[Y|v(X,\theta) = v]$ for $b = 1$ or 2. The mean function $E[Y|v(X,\theta) = v]$ in (16) is poorly approximated by the parametric models typically used in applications (e.g., low-order polynomials in v), so it is unlikely that a most powerful or nearly most powerful parametric test of (15) would be carried out in an application if (16) were the true data-generation process. Härdle [7] gives several applications in which the shape of $E[Y|v(X,\theta) = v]$ is similar to Figure 1.

The experiments were carried out at the nominal 0.05 level using sample sizes of $n = 50$ for $L = 1$ and $n = 50$ and 100 for $L = 3$. There were 500 replications in each experiment. Random numbers were generated with the pseudo-random number generators of GAUSS.

In the computation of T_n, K is the standard normal density, $F(v) = v$, $(h,s) = (0.1,0.8)$ if $L = 1$, $(0.3,1.0)$ if $L = 3$ and $n = 50$, and $(0.2,0.9)$ if $L = 3$ and $n = 100$. β_0 and θ were estimated from (15) by ordinary least squares (OLS), $w(\cdot) = 1$ on an interval containing 98% of observed values of $\beta_0 + \theta'X$ and 0 elsewhere, and $\sigma^2(v)$ is given by (14).[5] The semiparametric test is one-sided for the reasons discussed in Section 2.4.

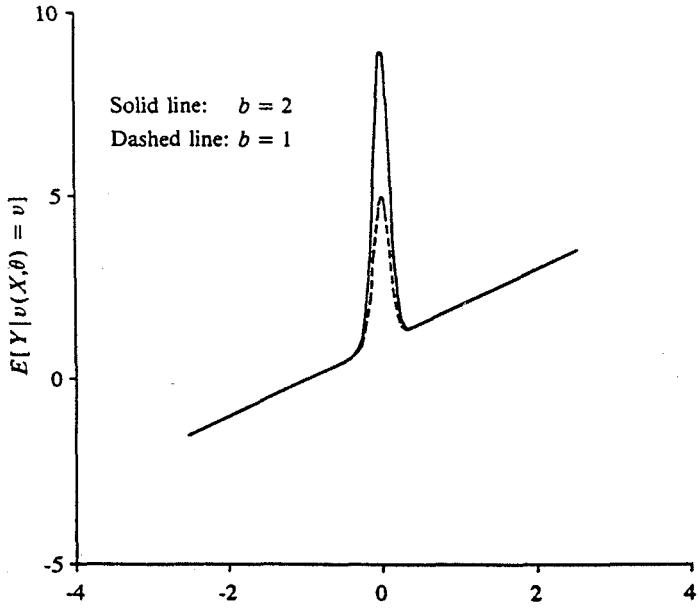


FIGURE 1. $E[Y|v(X,\theta) = v]$ in (16).

Horowitz, J. and Härdle, W. (1994)
Testing a Parametric Model against Semiparametric Alternatives.

Econometric Theory,V 10, No. 5, 821 - 848

Implementation of the test of Bierens [2] requires choosing several parameters of the test statistic and a function. We made choices similar to those used in Bierens' [2] Monte Carlo experiments. In his notation, we set $\gamma = \rho = 0.5$, $T = [1,5]$, $K_n = 10$, and $\phi(x) = \tan^{-1}(x/2)$. Note that Bierens' ϕ is different from ϕ in (16).

RESET consists of first estimating β_0 and the θ's in (15) by OLS and computing the predicted value of Y for each observed x. Denote the prediction by $\hat{Y}$. Then, one of the following two models is estimated by OLS: $E(Y|X = x) = \beta_0 + \theta'x + \alpha\hat{Y}^2$ (RESET2) or $E(Y|X = x) = \beta_0 + \theta'x + \alpha_1\hat{Y}^2 + \alpha_2\hat{Y}^3$ (RESET3). Finally, in RESET2 the hypothesis $H_\alpha: \alpha = 0$ is tested with a t-test. In RESET3, the hypothesis $H'_\alpha: \alpha_1 = \alpha_2 = 0$ is tested with a Wald test. Model (15) is rejected by RESET2 if H_α is rejected and by RESET3 if H'_α is rejected.

The most powerful parametric test of the null hypothesis (15) against the alternative (16) is the t-test of $b = 0$ based on least-squares estimation of β_0, θ, and b in (16). We assumed that the argument of ϕ in (16) is known when computing the power of this test.

Table 1 shows the results of the experiments with $L = 1$. The empirical sizes of the tests are close to the nominal sizes of 0.05. The test based on T_n is considerably more powerful than Bierens' test and both versions of RESET. Not surprisingly, T_n has less power than the most powerful parametric test. Of course, the power of the parametric test would be available in an application only in the unlikely event that (16) were known to be the correct alter-

TABLE 1. Results of the Monte Carlo experiments with $L = 1$

	Pr(reject H_0 at nominal 0.05 level)				
b	Most powerful parametric test	T_n[a]	Bierens' test	RESET2	RESET3
0	0.05	0.03	0.05	0.06	0.07
0.25	0.90	0.37	0.18	0.19	0.17
0.50	0.99	0.86	0.42	0.33	0.23
0.75	1.00	0.98	0.56	0.41	0.33
1.00	1.00	0.98	0.73	0.51	0.36
1.25	1.00	1.00	0.78	0.58	0.46
1.50	1.00	0.99	0.78	0.60	0.47
1.75	1.00	0.99	0.81	0.62	0.43
2.00	1.00	1.00	0.87	0.64	0.49
2.25	1.00	0.99	0.89	0.66	0.47
2.50	1.00	1.00	0.89	0.66	0.47

[a]The fluctuations in the rejection probability when $b \geq 1.25$ are not statistically significant at the 0.10 level.

Econometric Theory,V 10, No. 5, 821 - 848

native model, whereas T_n does not require a priori knowledge of the alternative.

Table 2 shows the results of the experiments with $L = 3$. One striking feature of these results is that the powers of the semiparametric and RESET tests are nonmonotonic functions of b. The reason for this is that the precision with which the θ's are estimated decreases as b increases. As a result, the values of $v(X,\hat{\theta})$ and $v(X,\theta)$ tend to be ordered differently, which causes the "bump" in (16) to spread and, if θ is very imprecisely estimated, shatter into isolated spikes. Shattering makes it difficult for the semiparametric and RESET tests to detect the difference between the null hypothesis model (15) and the true data-generation process (16). As can be seen by comparing the

TABLE 2. Results of the Monte Carlo experiments with $L = 3$

	Pr(reject H_0 at nominal 0.05 level)				
b	Most powerful parametric test	T_n[a]	Bierens' test	RESET2	RESET3
		$n = 50$			
0	0.08	0.02	0.04	0.07	0.08
0.25	0.88	0.16	0.07	0.18	0.16
0.50	0.99	0.47	0.15	0.32	0.28
0.75	1.00	0.72	0.27	0.43	0.36
1.00	1.00	0.71	0.31	0.41	0.34
1.25	1.00	0.70	0.34	0.43	0.31
1.50	1.00	0.63	0.41	0.44	0.35
1.75	1.00	0.57	0.41	0.41	0.31
2.00	1.00	0.51	0.43	0.40	0.31
2.25	1.00	0.45	0.45	0.36	0.27
2.50	1.00	0.40	0.48	0.34	0.20
		$n = 100$			
0	0.06	0.01	0.03	0.07	0.07
0.25	0.99	0.56	0.08	0.28	0.24
0.50	1.00	0.96	0.28	0.53	0.45
0.75	1.00	1.0	0.45	0.71	0.58
1.00	1.00	0.99	0.59	0.82	0.73
1.25	1.00	0.98	0.68	0.89	0.81
1.50	1.00	0.97	0.72	0.90	0.80
1.75	1.00	0.94	0.78	0.89	0.78
2.00	1.00	0.94	0.81	0.91	0.82
2.25	1.00	0.86	0.85	0.86	0.77
2.50	1.00	0.83	0.85	0.84	0.74

[a]The fluctuations in the rejection probability when $b \geq 1.25$ are not statistically significant at the 0.10 level.

results in Table 2 for $n = 50$ and $n = 100$, this problem diminishes as n increases since the θ's are estimated more precisely at large values of n. Non-monotonic power functions are not unusual in econometrics. See Nelson and Savin [13] for further discussion and examples. Another noteworthy feature of the results in Table 2 is that the empirical size of the semiparametric test is well below nominal whereas the empirical sizes of the RESET tests are somewhat larger than nominal. Despite these difficulties, the power of the semiparametric test exceeds the powers of Bierens' test and the RESET tests for all but the largest values of b.

3.2. An Application

Horowitz [10] estimated a binary probit model of the choice between automobile and transit for the trip to work. The estimation data set consisted of 842 trip records drawn from the Washington, D.C., area transportation study. The specification of the probit model is

$$P(\text{Auto} \mid X = x) = \Phi(\theta' x), \tag{17}$$

where Φ is the cumulative normal distribution function, X is a vector of explanatory variables, and θ is a conformable vector of estimated parameters. The components of X are an intercept, the number of automobiles owned by the traveler's household, the difference between automobile and transit out-of-vehicle travel times, the difference between automobile and transit in-vehicle travel times, and the difference between automobile and transit travel costs. Horowitz [10] carried out parametric likelihood ratio, Wald and Lagrangian multiplier tests of (17) against a random coefficients probit model. This model is obtained from (17) by replacing $\Phi(\theta' x)$ with $\Phi[\theta' x/(x'\Sigma x)^{1/2}]$, where Σ is a positive-definite matrix. All of the tests rejected (17) ($p < 0.01$).

To investigate the performance of T_n in an application, we tested (17) using both T_n and Bierens' [2] test. Bierens' test was carried out using the parameter and function choices described in Section 3.1. The value of the test statistic was 0.43. Under the hypothesis that (17) is correctly specified, Bierens' test statistic is asymptotically distributed as χ^2 with 1 degree of freedom. Therefore, Bierens' test does not reject (17) and, thus, does not detect the misspecification of (17) found by the tests against the random coefficients probit model.

In computing the T_n test statistic, $\hat{\theta}_n$ was estimated by maximum likelihood using (17), $v(x,\theta) = \theta' x$, $w(\cdot) = 1$ on an interval containing 98% of the observed values of $\hat{\theta}_n' X$ and 0 elsewhere, and $\hat{\sigma}^2(v) = \hat{F}_{ni}(v)[1 - \hat{F}_{ni}(v)]$. As is explained in note 5, there is no known systematic method for selecting bandwidth values for $\hat{F}_{ni}$. We used several bandwidths that were found through graphical examination of $\hat{F}_{ni}$ to span the range of reasonable choices. Values outside of this range caused the graph of $\hat{F}_{ni}$ to be either excessively wiggly or excessively flat. The value of $T_n/\hat{\sigma}_T$ was in the range

2.45–3.26, depending on the bandwidth. Thus, the T_n test rejects (17) ($p < 0.007$). This is consistent with the results of the tests against the random coefficients probit model.

4. CONCLUSIONS

This paper has described a method for testing a parametric model of the conditional mean against a semiparametric alternative. The test is motivated by a parametric conditional moments test and amounts to replacing the parametric alternative model in the conditional moments test with a semiparametric model. The resulting semiparametric test is not consistent against all alternatives, but in a sense that has been explained it is consistent against a larger set of alternatives than are parametric conditional moments tests based on finitely many moment conditions. The results of Monte Carlo experiments and an application using real data illustrate the usefulness of the semiparametric test.

NOTES

1. The tests in Hong and White [9], Whang and Andrews [19], and Yatchew [21] are also motivated by (8) but aim at consistency against all alternatives and do not use a parametric index function to reduce the dimension of the nonparametric model.

2. At the cost of lower asymptotic local power, asymptotic unbiasedness also could be achieved by using an undersmoothed "ordinary" kernel estimator for $\hat{F}_{ni}$.

3. A referee has pointed out that the test also can be inconsistent if $v(x,\theta)$ is a constant, in which case Assumption 2 of the Appendix is violated.

4. We originally intended to include the test of Whang and Andrews [19] in the comparison. This test is based on comparing the mean square residual from parametric and nonparametric estimates of the conditional mean of Y. We dropped the test from consideration after finding that, in our Monte Carlo experiments, its empirical size at the nominal 0.05 level was between 0.24 and 0.50 for a wide range of bandwidths in the nonparametric regression.

5. A systematic procedure for choosing h and s for $\hat{F}_{ni}$ with finite samples has not been developed. Because the estimator is asymptotically unbiased, the tradeoff between asymptotic bias and variance that underlies bandwidth selection methods such as cross validation does not exist. We selected h and s graphically. With the values we used, the graph of $\hat{F}_{ni}$ is neither excessively wiggly, as happens when h and s are too small, nor excessively flat, as happens when they are too large. The regularity conditions in the Appendix require K to have bounded support, but this is not essential, as is noted there.

REFERENCES

1. Bierens, H.J. Kernel estimators of regression functions. In T.F. Bewley (ed.), *Advances in Econometrics: Fifth World Congress*, vol. 1, pp. 99–144. New York: Cambridge University Press, 1987.
2. Bierens, H.J. A consistent conditional moment test of functional form. *Econometrica* 58 (1990): 1443–1458.
3. Eubank, R.L. & C.H. Spiegelman. Testing the goodness of fit of a linear model via nonparametric regression. *Journal of the American Statistical Association* 85 (1990): 387–392.
4. Gozalo, P.L. A consistent model specification test for nonparametric estimation of regression function models. *Econometric Theory* 9 (1993): 451–477.

Econometric Theory,V 10, No. 5, 821 - 848

5. Hall, P. Central limit theorem for integrated square error of multivariate nonparametric density estimators. *Journal of Multivariate Analysis* 14 (1984): 1-16.
6. Härdle, W. A note on jackknifing kernel regression function estimators. *IEEE Transactions of Information Theory* 32 (1986): 298-300.
7. Härdle, W. *Applied Nonparametric Regression.* New York: Cambridge University Press, 1990.
8. Härdle, W. & E. Mammen. Comparing nonparametric versus parametric regression fits. *Annals of Statistics* 21 (1993): 1926-1947.
9. Hong, Y. & H. White. Consistent Specification Testing via Nonparametric Series Regression. Discussion paper, Department of Economics, University of California, San Diego, CA, 1992.
10. Horowitz, J.L. Semiparametric estimation of a work-trip mode choice model. *Journal of Econometrics* 58 (1993): 49-70.
11. Huber, P.J. Projection pursuit. *The Annals of Statistics* 13 (1985): 435-475.
12. le Cessie, S. & J.C. van Houwelingen. A goodness-of-fit test for binary regression models, based on smoothing methods. *Biometrics* 47 (1991): 1267-1282.
13. Nelson, F.D. & N.E. Savin. The danger of extrapolating asymptotic local power. *Econometrica* 58 (1990): 977-981.
14. Newey, W.K. Maximum likelihood specification testing and conditional moment tests. *Econometrica* 53 (1985): 1047-1070.
15. Pollard, D. *Convergence of Stochastic Processes.* New York: Springer-Verlag, 1984.
16. Schucany, W.R. & J.P. Sommers. Improvement of kernel type density estimators. *Journal of the American Statistical Association* 72 (1977): 420-423.
17. Serfling, R.J. *Approximation Theorems of Mathematical Statistics.* New York: John Wiley & Sons, 1980.
18. Silverman, B.W. Weak and strong uniform consistency of the kernel estimate of a density and its derivatives. *The Annals of Statistics* 6 (1978): 177-184.
19. Whang, Y.-J. & D.W.K. Andrews. Tests of Specification for Parametric and Semiparametric Models. Cowles Foundation discussion paper #968, Yale University, New Haven, CT, 1991.
20. Wooldridge, J.M. A test for functional form against nonparametric alternatives. *Econometric Theory* 8 (1992): 452-475.
21. Yatchew, A.J. Nonparametric regression tests based on least squares. *Econometric Theory* 8 (1992): 435-451.
22. Zheng, X. A consistent test of functional form via nonparametric estimation techniques. Discussion paper, Department of Economics, Princeton University, Princeton, NJ, 1991.

APPENDIX

This appendix contains regularity conditions and the proofs of theorems.

A1. NOTATION

In addition to the notation defined in the text:

N_θ = a neighborhood of θ;

Econometric Theory,V 10, No. 5, 821 - 848

S_X = the support of X;

$\tilde{S}_v$ = an open subset of the support of $v(X,\theta)$;

S_v = a compact subset of $\tilde{S}_v$;

$\tilde{S}_X = \{x: v(x,\theta) \in S_v\}$;

$h = cn^{-1/(2r+1)}$, where $c > 0$ and $r \geq 2$ is an integer;

$s = cn^{-\delta/(2r+1)}$, where $0 < \delta < 1$;

$K(\cdot)$ = an rth order kernel function.

For $\xi = h$ or s,

$$p_{n\xi i}(v) = (n\xi)^{-1} \sum_{\substack{j=1 \\ j\neq i}}^{n} K\left[\frac{v - v(X_j,\theta)}{\xi}\right]$$

$$p_{n\xi}(v) = (n\xi)^{-1} \sum_{j=1}^{n} K\left[\frac{v - v(X_j,\theta)}{\xi}\right]$$

$$F_{n\xi i}(v) = \sum_{\substack{j=1 \\ j\neq i}}^{n} Y_j K\left[\frac{v - v(X_j,\theta)}{\xi}\right] \Bigg/ \sum_{\substack{j=1 \\ j\neq i}}^{n} K\left[\frac{v - v(X_j,\theta)}{\xi}\right]$$

$$F_{ni}(v) = [F_{nhi}(v) - (h/s)^r F_{nsi}(v)]/[1 - (h/s)^r]$$

$$g_{n\xi i}(v) = p_{n\xi i}(v) F_{n\xi i}(v).$$

A2. ASSUMPTIONS

1. S_X is compact. At least one component of X has a probability distribution that is absolutely continuous with respect to Lebesgue measure.
2. For every $\tau \in N_\theta$ and $x \in S_X$, v satisfies the following:
   a. $|v(x,\tau)| < M$ for some $M < \infty$ that does not depend on τ or x.
   b. The probability distribution of $v(X,\tau)$ is absolutely continuous with respect to Lebesgue measure.
   c. $v(x,\tau)$ is differentiable with respect to τ. $\partial v(x,\tau)/\partial \tau_k$ $(k = 1,\ldots,K)$ is uniformly bounded and Lipschitz continuous with respect to τ and the continuous components of x.
3. Let p_τ denote the probability density function of $v(X,\tau)$. For each $\tau \in N_\theta$:
   a. $m_p \leq p_\tau(v) \leq M_p$ for some $m_p > 0$ and $M_p < \infty$ that do not depend on τ.
   b. p_τ has r continuous derivatives that are uniformly bounded over $\tau \in N_\theta$ and $v \in \tilde{S}_v$.
4. $w(\cdot)$ has compact support $S_w \subset \text{int}(S_v)$ and satisfies the following:
   a. $0 \leq w(v) < M_w$ for some $M_w < \infty$ and all $v \in S_w$.
   b. $|w(v_2) - w(v_1)| \leq M_w^* |v_2 - v_1|$ for some $M_w^* < \infty$ and all v_2, v_1.
5. a. $|F[v(x,\tau)]|$ and $|H[v(x,\tau)]|$ are uniformly bounded over $x \in S_x$ and $\tau \in N_\theta$.
   b. $F(v)$ and $H(v)$ have r continuous derivatives that are uniformly bounded over $v \in S_v$.

6. Let E_X denote the expectation over the distribution of X. Define

$$\Gamma(x,v,\tau) = E_X\{F[v(X,\theta)][\partial v(x,\tau)/\partial\tau - \partial v(X,\tau)/\partial\tau] | v(X,\tau) = v\}.$$

Let Γ_k $(k = 1,\ldots,K)$ denote the kth component of Γ. There is a finite number M_Γ, not depending on τ or x, such that for all $\tau \in N_\theta$, $x \in S_X$, $v_1, v_2 \in \bar{S}_v$, and $k = 1,\ldots,K$,

$$|\Gamma_k(x,v_2,\tau) - \Gamma_k(x,v_1,\tau)| \le M_\Gamma |v_2 - v_1|.$$

7. $\sigma^2(v) \equiv \mathrm{Var}[Y | v(X,\theta) = v]$ is a uniformly bounded, continuous function of $v \in S_v$. $E\{Y - E[Y|v(x,\theta)]\}^4$ is uniformly bounded over $v \in S_v$.
8. K is an rth order kernel ($r \ge 2$) with bounded support. Also, K is uniformly bounded, continuous, and symmetrical about 0. The derivative of K, K', is uniformly bounded and has an absolutely integrable Fourier transform, $\psi(\cdot)$.

These assumptions are mainly boundedness and smoothness conditions. The requirement that K has bounded support can be removed at the cost of additional technical complexity in the proofs.

A3. THE ASYMPTOTIC DISTRIBUTION OF T_n UNDER H_0

Lemmas 1-6 show that asymptotically $\hat{\theta}_n$ can be replaced by θ in T_n. Lemma 7 gives a result that is used in deriving the U-statistic form of T_n.

LEMMA 1. *Define*

$$G_{nhi}(v) = [g_{nhi}(v) - F(v)p_{nhi}(v)]/p_\theta(v) \tag{A.1}$$

and

$$J_{nhi}(v) = [p_{nhi}(v) - p_\theta(v)]\{[g_{nhi}(v) - F(v)p_\theta(v)] - F(v)[p_{nhi}(v) - p_\theta(v)]\}/[p_\theta(v)]^2. \tag{A.2}$$

As $n \to \infty$,

$$\sup_{1\le i\le n}\sup_{v\in\bar{S}_v} |p_{nhi}(v) - p_\theta(v)| = O\{[(\log n)/(nh)]^{1/2}\}, \tag{A.3}$$

almost surely,

$$\sup_{1\le i\le n}\sup_{v\in\bar{S}_v} |g_{nhi}(v) - p_\theta(v)F(v)| = O_p[(nh^2)^{-1/2}] \tag{A.4}$$

and

$$\sup_{1\le i\le n}\sup_{v\in\bar{S}_v} |F_{nhi}(v) - F(v) - G_{nhi}(v) + J_{nhi}(v)| = O_p[(\log n)/(n^{3/2}h^2)]. \tag{A.5}$$

These relations also hold if h is replaced by s.

Proof. Only (A.3)-(A.5) are proved. The proofs with h replaced by s are identical. (A.3) follows from

$$\sup_{1\le i\le n}\sup_{v\in\bar{S}_v} |p_{nhi}(v) - p_{nh}(v)| \le M/nh$$

Econometric Theory,V 10, No. 5, 821 - 848

for some $M < \infty$ (by Assumption 8) and

$$\sup_{v \in S_v} |p_{nh}(v) - p_\theta(v)| = 0\{[(\log n)/(nh)]^{1/2}\}$$

almost surely [18].

To prove (A.4), define

$$g_{nh}(v) = (nh)^{-1} \sum_{j=1}^{n} Y_j K\left[\frac{v - v(X_j, \theta)}{h}\right].$$

Observe that

$$\sup_{1 \le i \le n} \sup_{v \in S_v} |g_{nhi}(v) - p_\theta(v)F(v)|$$

$$\le \sup_{v \in S_v} |g_{nh}(v) - p_\theta(v)F(v)| + M(nh)^{-1} \sup_{1 \le i \le n} |Y_i| \tag{A.6}$$

for some $M < \infty$ because K is bounded uniformly. The first term on the right-hand side of (A.6) is $O_p[(nh^2)^{-1/2}]$ [1]. Now consider the second term. Let P_Y denote the marginal c.d.f. of $|Y|$. Given any $\epsilon > 0$,

$$\log P[\sup_{1 \le i \le n} n^{-1/2}|Y_i| < \epsilon] = n \log[1 - [1 - P_Y(n^{1/2}\epsilon)]] \tag{A.7}$$

$$= -(n^{1/2}\epsilon)^2 \zeta_n^{-1}[1 - P_Y(n^{1/2}\epsilon)]/\epsilon^2, \tag{A.8}$$

by a Taylor series expansion, where ζ_n is between $P_Y(n^{1/2}\epsilon)$ and 1. $E(Y^2) < \infty$. Therefore, $\lim_{u \to \infty} u^2[1 - P_Y(u)] = 0$, the right-hand side of (A8) converges to 0 as $n \to \infty$, and

$$\sup_{1 \le i \le n} |Y_i| = o_p(n^{1/2}). \tag{A.9}$$

(A.4) follows from (A.6) and (A.9).

To prove (A.5), expand $F_{nhi} = g_{nhi}/p_{nhi}$ in a Taylor series about $g_{nhi} = Fp_\theta$ and $p_{nhi} = p_\theta$. Then apply (A.3) and (A.4). ■

LEMMA 2. *For any positive $L < \infty$, define $\Theta_L = \{t \in \mathcal{R}^K : n^{1/2}\|t - \theta\| \le L\}$. For each $\ell = 1, \ldots, n$,*

$$\sup_{\theta_n \in \Theta_L} \sup_{x \in S_X} (nh)^{1/2} |\hat{F}_{nh\ell}[v(x, \theta_n)] - F_{nh\ell}[v(x, \theta)]| = O_p(h^{1/2}) \tag{A.10}$$

as $n \to \infty$. The same relation holds when h is replaced by s.

Proof. Only (A.10) is proved. The proof with s in place of h is identical. Define $\tilde{g}_{nh\ell}(\cdot)$ and $\tilde{p}_{nh\ell}(\cdot)$, respectively, by replacing θ with θ_n in the definitions of $g_{nh\ell}(\cdot)$ and $p_{nh\ell}(\cdot)$. It suffices to prove that

$$\sup_{x \in S_X} (nh)^{1/2} |\tilde{g}_{nh\ell}[v(x, \theta_n)] - g_{nh\ell}[v(x, \theta)]| = O_p(h^{1/2}) \tag{A.11}$$

and

$$\sup_{x \in S_X} (nh)^{1/2} |\tilde{p}_{nh\ell}[v(x, \theta_n)] - p_{nh\ell}[v(x, \theta)]| = O_p(h^{1/2}). \tag{A.12}$$

We prove only (A.11). The proof of (A.12) is similar.

Horowitz, J. and Härdle, W. (1994)
Testing a Parametric Model against Semiparametric Alternatives.

Econometric Theory,V 10, No. 5, 821 - 848

Define

$$D_n(x) = (nh)^{-1/2} \sum_{\substack{j=1 \\ j \neq \ell}}^{n} Y_j \left\{ K\left[\frac{v(x,\theta_n) - v(X_j,\theta_n)}{h}\right] - K\left[\frac{v(x,\theta) - v(X_j,\theta)}{h}\right] \right\}, \tag{A.13}$$

$\Delta_{n1}(x) = D_n(x) - ED_n(x)$, and $\Delta_{n2}(x) = ED_n(x)$. Then

$(nh)^{1/2}[\hat{g}_{nh\ell}(x) - g_{nh\ell}(x)] = \Delta_{n_1}(x) + \Delta_{n2}(x)$.

Consider Δ_{n1}. By a Taylor series expansion of the summand of (A.13),

$D_n(x) = n^{1/2}(\theta_n - \theta)'(nh^3)^{-1/2} B_{n\ell}(x)$,

where

$$B_{n\ell}(x) = n^{-1/2} \sum_{\substack{j=1 \\ j \neq \ell}}^{n} Y_j Z_{nj} K'\left[\frac{v(x,\theta_n^*) - v(X_j,\theta_n^*)}{h}\right],$$

θ_n^* is between θ and θ_n, and

$Z_{nj} = Z_{nj}(v) = [\partial v(x,\tau)/\partial\tau - \partial v(X_j,\tau)/\partial\tau]_{\tau=\theta_n^*}$.

By Assumption 8, K' has an absolutely integrable Fourier transform, ψ, so

$$B_{n\ell}(x) = (h/2\pi) \int_{-\infty}^{\infty} \exp[-itv(x,\theta_n^*)]\psi(ht) n^{-1/2} \sum_{\substack{j=1 \\ j \neq \ell}}^{n} Y_j Z_{nj} \{\exp[itv(X_j,\theta_n^*)]\}\, dt.$$

Let $B_{n\ell k}(x)$ and Z_{njk}, respectively, denote the kth components of $B_{n\ell}(x)$ and Z_{nj}. Then for each k,

$$\begin{aligned} &|B_{n\ell k}(x) - EB_{n\ell k}(x)| \\ &\quad \leq (h/2\pi) \int_{-\infty}^{\infty} |\psi(ht)| n^{-1/2} \\ &\qquad \times \left| \sum_{\substack{j=1 \\ j \neq \ell}}^{n} \{Y_j Z_{njk} \exp[itv(X_j,\theta_n^*)] - EY_j Z_{njk} \exp[itv(X_j,\theta_n^*)]\} \right| dt. \end{aligned} \tag{A.14}$$

By Lemma 2.37 of Pollard [15] with $\delta_n = 1$ and $\alpha_n = n^{-1/2} \log n$ in Pollard's notation, the sum in the integrand is $o_p(n^{1/2} \log n)$ uniformly over $x \in \bar{S}_X$ and $\{\theta_n^*: n^{1/2}\|\theta_n^* - \theta\| \leq L\}$. Therefore,

$$|B_{n\ell k}(x) - EB_{n\ell k}(x)| \leq (h/2\pi) \int_{-\infty}^{\infty} |\psi(ht)| dt \cdot o_p(\log n) = o_p(\log n).$$

because $|\psi|$ is integrable. Since $n^{1/2}(\theta_n - \theta) = O(1)$,

$$|\Delta_{n1}(x)| \leq O[(nh^3)^{-1/2}] o_p(\log n) = O(h^{r-1}) o_p(\log n) = o_p(h^{1/2}) \tag{A.15}$$

uniformly over $x \in \bar{S}_X$ and $\{\theta_n: n^{1/2}\|\theta_n - \theta\| \leq L\}$.

Now consider Δ_{n2}. By the Taylor series expansion of (A.13),

Econometric Theory,V 10, No. 5, 821 - 848

$$\Delta_{n2}(x) = [(n-1)/[(nh^3)^{1/2}]](\theta_n - \theta)'E\left\{YZ_n \cdot K'\left[\frac{v(x,\theta_n^*) - v(X,\theta_n^*)}{h}\right]\right\}$$

$$= [(n-1)/[(nh^3)^{1/2}]](\theta_n - \theta)'\int_{-\infty}^{\infty} \Gamma(x,u,\theta_n^*)K'\left[\frac{v(x,\theta_n^*) - u}{h}\right] p_{\theta_n^*}(u)\,du$$

$$= -[(n-1)/[(nh)^{1/2}]](\theta_n - \theta)'$$

$$\times \int_{-\infty}^{\infty} \Gamma[x, v(x,\theta_n^*) + h\xi, \theta_n^*]K'(\xi)p_{\theta_n^*}[v(x,\theta_n^*) + h\xi]\,d\xi \qquad \text{(A.16)}$$

by a change of variables and symmetry of K. By Assumption 2, $v(x,\theta_n^*) \in \bar{S}_v$ and $v(x,\theta_n^*) + h\xi \in \bar{S}_v$ for any ξ, any $x \in \bar{S}_X$, and all sufficiently large n. Therefore, by Assumptions 2, 3, and 6,

$$\|\Gamma[x, v(x,\theta_n^*) + h\xi, \theta_n^*]p_{\theta_n^*}[v(x,\theta_n^*) + h\xi]$$

$$- \Gamma[x, v(x,\theta_n^*), \theta_n^*]p_{\theta_n^*}[v(x,\theta_n^*)]\| \le Mh|\xi| \qquad \text{(A.17)}$$

for each ξ, all sufficiently large n, and some $M < \infty$, where $|\cdot|$ denotes the Euclidean norm. By symmetry of K, (A.16), (A.17), and Lebesgue's dominated convergence theorem, $\Delta_{n2}(x) = O(h^{1/2})$ as $n \to \infty$ uniformly over $x \in S_X$ and $\{\theta_n : n^{1/2}\|\theta_n - \theta\| \le L\}$. (A.11) follows from this result and (A.15). ■

LEMMA 3. *For any $L > 0$ and as $n \to \infty$,*

$$\sup_{1 \le i \le n} \sup_{x \in \bar{S}_X} (nh)^{1/2}|\hat{F}_{nhi}[v(x,\theta_n)] - F_{nhi}[v(x,\theta)]| = O_p(h^{1/2})$$

uniformly over $\theta_n \in \Theta_L = \{\theta_n : n^{1/2}\|\theta_n - \theta\| \le L\}$. The same relation holds with h replaced by s.

Proof. It suffices to prove that uniformly over $\theta_n \in \Theta_L$

$$\sup_{1 \le i \le n} \sup_{x \in \bar{S}_X} (nh)^{1/2}|\tilde{g}_{nhi}[v(x,\theta_n)] - g_{nhi}[v(x,\theta)]| = O_p(h^{1/2}), \qquad \text{(A.18)}$$

$$\sup_{1 \le i \le n} \sup_{x \in \bar{S}_X} (nh)^{1/2}|\tilde{p}_{nhi}[v(x,\theta_n)] - p_{nhi}[v(x,\theta)]| = O_p(h^{1/2}), \qquad \text{(A.19)}$$

and that (A.18) and (A.19) hold with h replaced by s, where $\tilde{g}$ and $\tilde{p}$ are defined as in Lemma 2. The proof is given only for (A.18). The proofs of (A.19) and the relations for s are identical. Define

$$d_{nhi}(x) = (nh)^{-1/2}Y_i\left\{K\left[\frac{v(x,\theta_n) - v(X_i,\theta_n)}{h}\right] - K\left[\frac{v(x,\theta) - v(X_i,\theta)}{h}\right]\right\}.$$

Because of (A.11) and (A.12), to prove (A.18) it suffices to show that

$$\sup_{1 \le i \le n} \sup_{\theta_n \in \Theta_L} \sup_{x \in \bar{S}_X} |d_{nhi}(x)| = o_p(h^{1/2}). \qquad \text{(A.20)}$$

By a Taylor series expansion,

$$d_{nhi}(x) = (nh^3)^{-1/2}(\theta_n - \theta)'Y_iZ_{ni}K'\left[\frac{v(x,\theta_n^*) - v(X_i,\theta_n^*)}{h}\right],$$

Horowitz, J. and Härdle, W. (1994)
Testing a Parametric Model against Semiparametric Alternatives.

Econometric Theory,V 10, No. 5, 821 - 848

where Z_{ni} is defined as in Lemma 2 and θ_n^* is between θ_n and θ. Therefore, by Assumptions 2 and 8,

$$\sup_{1\le i\le n}\sup_{\theta_n\in\Theta_L}\sup_{x\in S_X}|h^{-1/2}d_{nhi}(x)| \le M(nh^2)^{-1}\sup_{1\le i\le n}|Y_i| \tag{A.21}$$

for some $M < \infty$. But $\sup_{1\le i\le n}|Y_i| = o_p(n^{1/2})$ by (A.9). Therefore, the right-hand side of (A.21) is $o_p(1)$, and (A.20) holds. ■

LEMMA 4. *Define Θ_L as in Lemma 2. Define G_{nsi} and J_{nsi} by replacing h with s in the definitions of G_{nhi} and J_{nsi}. Define*

$$G_{ni}(v) = [G_{nhi}(v) - (h/s)^r G_{nsi}(v)]/[1 - (h/s)^r]$$

and

$$J_{ni}(v) = [J_{nhi}(v) - (h/s)^r J_{nsi}(v)]/[1 - (h/s)^r].$$

As $n \to \infty$ and uniformly over $\theta_n \in \Theta_L$,

$$\sup_{1\le i\le n}\sup_{v\in S_v}|F_{ni}(v) - F(v) - G_{ni}(v) + J_{ni}(v)| = O_p[(\log n)/(n^{3/2}h^2)]$$

and

$$\sup_{1\le i\le n}\sup_{x\in S_X}(nh)^{1/2}|\hat{F}_{ni}[v(x,\theta_n)] - F_{ni}[v(x,\theta)]| = O_p(h^{1/2}).$$

Proof. These results follow by combining the definitions of $\hat{F}_{ni}$ and F_{ni} with the results of Lemmas 1 and 3. ■

LEMMA 5. *Let $\{\theta_n : n = 1,2,\dots\}$ be a sequence in $\mathcal{R}^K$ that converges to θ. For all sufficiently large n and $x \in S_X$, $v(x,\theta_n) \in S_w$ implies that $v(x,\theta) \in S_v$.*

Proof. By Assumption 2, $|v(x,\theta_n) - v(x,\theta)| \le M\|\theta_n - \theta\|$ for some $M < \infty$ that does not depend on x. The result follows from the fact that $S_w \subset \text{int}(S_v)$. ■

LEMMA 6. *Define*

$$T_n^* = h^{1/2}\sum_{i=1}^n w[v(X_i,\theta)]\{Y_i - F[v(X_i,\theta)]\}\{F_{ni}[v(X_i,\theta)] - F[v(X_i,\theta)]\}.$$

Then as $n \to \infty$, $T_n = T_n^* + o_p(1)$.

Proof. Some algebra shows that $T_n = \sum_{j=1}^6 R_{nj}$, where

$$R_{n1} = h^{1/2}\sum_{i=1}^n w[v(X_i,\hat{\theta}_n)]\{Y_i - F[v(X_i,\theta)]\}$$
$$\times\{\hat{F}_{ni}[v(X_i,\hat{\theta}_n)] - F_{ni}[v(X_i,\theta)]\}$$

$$R_{n2} = h^{1/2}\sum_{i=1}^n \{w[v(X_i,\hat{\theta}_n)] - w[v(X_i,\theta)]\}$$
$$\times\{Y_i - F[v(X_i,\theta)]\}\{F_{ni}[v(X_i,\theta)] - F[v(X_i,\theta)]\}$$

$$R_{n3} = -h^{1/2}\sum_{i=1}^n w[v(X_i,\hat{\theta}_n)]\{Y_i - F[v(X_i,\theta)]\}$$
$$\times\{F[v(X_i,\hat{\theta}_n)] - F[v(X_i,\theta)]\}$$

Horowitz, J. and Härdle, W. (1994)
Testing a Parametric Model against Semiparametric Alternatives.

Econometric Theory,V 10, No. 5, 821 - 848

$$R_{n4} = -h^{1/2}\sum_{i=1}^{n} w[v(X_i,\hat{\theta}_n)]\{F[v(X_i,\hat{\theta}_n)] - F[v(X_i,\theta)]\}$$
$$\times \{\hat{F}_{ni}[v(X_i,\hat{\theta}_n)] - F_{ni}[v(X_i,\theta)]\}$$

$$R_{n5} = -h^{1/2}\sum_{i=1}^{n} w[v(X_i,\hat{\theta}_n)]\{F[v(X_i,\hat{\theta}_n)] - F[v(X_i,\theta)]\}$$
$$\times \{F_{ni}[v(X_i,\theta)] - F[v(X_i,\theta)]\}$$

$$R_{n6} = h^{1/2}\sum_{i=1}^{n} w[v(X_i,\hat{\theta}_n)]\{F[v(X_i,\hat{\theta}_n)] - F[v(X_i,\theta)]\}^2.$$

Let $\{\theta_n\}$ be an arbitrary nonstochastic sequence in $\mathcal{R}^K$ satisfying $n^{1/2}(\theta_n - \theta) = O(1)$ as $n \to \infty$. Define $\tilde{R}_{ni}$ $(i = 1,\ldots,5)$ by replacing $\hat{\theta}_n$ with θ_n in R_{ni}. It suffices to show that $\tilde{R}_{ni} = o_p(1)$ for $i = 1,\ldots,5$, and $R_{n6} = o_p(1)$.

a. $\tilde{R}_{n1}$: Given any $\epsilon > 0$, let A_{nie} denote the intersection of the events $\tilde{p}_{nhi}[v(x,\theta_n)] > \epsilon$ uniformly over $x \in \tilde{S}_X$, $\tilde{p}_{nsi}[v(x,\theta_n)] > \epsilon$ uniformly over $x \in \tilde{S}_X$, and

$$\sup_{x\in\tilde{S}_X} (nh)^{1/2}|\hat{F}_{ni}[v(x,\theta_n)] - F_{ni}[v(x,\theta)]| \le \epsilon,$$

where $\tilde{p}_{nhi}$ is as defined in Lemma 2. Define

$$A_{ne} = \bigcap_{i=1}^{n} A_{nie}.$$

Let $1(\cdot)$ be the indicator of the event in parentheses. By (A.3), (A.19), and Lemma 4, $P(A_{ne}^c) = o(1)$. By Lemma 5, $x \in \tilde{S}_X$ if $w[v(x,\theta_n)] > 0$ and n is sufficiently large. Therefore,

$$\tilde{R}_{n1} = R_{n1}^* + o_p(1), \tag{A.22}$$

where

$$R_{n1}^* = h^{1/2}\sum_{i=1}^{n} 1(A_{nie})w[v(X_i,\theta_n)]\{Y_i - F[v(X_i,\theta)]\}$$
$$\times \{\hat{F}_{ni}[v(X_i,\theta_n)] - F_{ni}[v(X_i,\theta)]\}.$$

$E(R_{n1}^*) = 0$ because A_{nie} does not depend on Y_i or X_i. Define $U = Y - F[v(X,\theta)]$. Then

$$\mathrm{Var}(R_{n1}^*) = n^{-1}E\sum_{i=1}^{n} 1(A_{nie})w[v(X_i,\theta_n)]^2\sigma^2[v(X_i,\theta_n)](nh)$$
$$\times \{\hat{F}_{ni}[v(X_i,\theta_n)] - F_{ni}[v(X_i,\theta)]\}^2$$
$$+ n^{-1}E\left(\sum_{i=1}^{n}\sum_{\substack{j=1\\ j\neq i}}^{n} 1(A_{nie})1(A_{nje})w[v(X_i,\theta_n)]\,w[v(X_j,\theta_n)]U_iU_j\right.$$
$$\times (nh)\{\hat{F}_{ni}[v(X_i,\theta_n)] - F_{ni}[v(X_i,\theta)]\}$$
$$\left.\times \{\hat{F}_{nj}[v(X_j,\theta_n)] - F_{nj}[v(X_j,\theta)]\}\right).$$

Econometric Theory,V 10, No. 5, 821 - 848

where

$$R_{n5}^* = -h^{1/2}\sum_{i=1}^{n} w[v(X_i,\theta_n)]\{F[v(X_i,\theta_n)] - F[v(X_i,\theta)]\}G_{ni}[v(X_i,\theta)].$$

Let F' denote the derivative of F. By a Taylor series expansion,

$$R_{n5}^* = -h^{1/2}(\theta_n - \theta)'\sum_{i=1}^{n} w[v(X_i,\theta_n)]F'[v(X_i,\theta_n^*)]v_\theta(X_i,\theta_n^*)G_{ni}[v(X_i,\theta)],$$

where θ_n^* is between θ and θ_n. By arguments identical to those of Bierens [1], $EG_{ni}[v(x,\theta)] = o[(nh)^{-1/2}]$ uniformly over x. Therefore, $E(R_{n5}^*) = o(1)$ by Assumptions 2, 4, and 5. In addition,

$$(R_{n5}^*)^2 \le Mh\|\theta_n - \theta\|^2 E\sum_{i=1}^{n}\sum_{j=1}^{n}\{G_{ni}[v(X_i,\theta)]G_{nj}[v(X_j,\theta)]\}.$$

By the arguments of Bierens [1], the expectation is $O(h^{-1})$ uniformly over X. Therefore, $E(R_{n5}^*)^2 = o(1)$. It follows from Chebyshev's inequality that $R_{n5}^* = o_p(1)$ and from this result and (A.26) that $\tilde{R}_{n5} = o_p(1)$.

f. R_{n6}: By Assumptions 2 and 5, $\{F[v(X,\hat{\theta}_n] - F[v(X,\theta)]\}^2 = O_p(n^{-1})$ uniformly over $\{X: v \in S_v\}$. Since, in addition, w is bounded uniformly, $R_{n6} = O_p(h^{1/2})$. ■

LEMMA 7. *Define* $V_i = v(X_i,\theta)$. *Then*

$$T_n = h^{1/2}\sum_{i=1}^{n} w(V_i)[Y_i - F(V_i)]G_{ni}(V_i) + o_p(1). \tag{A.27}$$

Proof. By Lemmas 4 and 6,

$$T_n = h^{1/2}\sum_{i=1}^{n} w(V_i)[Y_i - F(V_i)]G_{ni}(V_i) - T_{n1} + o_p(1),$$

where

$$T_{n1} = h^{1/2}\sum_{i=1}^{n} w(V_i)[Y_i - F(V_i)]J_{ni}(V_i).$$

It suffices to show that $T_{n1} = o_p(1)$. $E(T_{n1}) = 0$ because $J_{ni}(V_i)$ does not depend on Y_i. In addition, since $EU_iJ_{nj}(v) = o(n^{-1})$ uniformly over v,

$$E(T_{n1}^2) = h\sum_{i=1}^{n} E\{w(V_i)^2\sigma^2(V_i)J_{ni}(V_i)^2\} + o(1). \tag{A.28}$$

But for any $v \in S_v$,

$$\begin{aligned}
[1 - (h/s)^r]J_{ni}(v) &\le |g_{nhi}(v) - p_\theta(v)F(v)||p_{nhi}(v) - p_\theta(v)|/p_\theta(v)^2 \\
&\quad + F(v)|p_{nhi}(v) - p_\theta(v)|^2/p_\theta(v)^2 + (h/s)^r \\
&\quad \times [|g_{nsi}(v) - p_\theta(v)F(v)||p_{nsi}(v) - p_\theta(v)|/p_\theta(v)^2 \\
&\quad\quad + F(v)|p_{nsi}(v) - p_\theta(v)|^2/p_\theta(v)^2] \\
&\le [|g_{nhi}(v) - p_\theta(v)F(v)|/p_\theta(v)^2 + (h/s)^{r+1/2} \\
&\quad \times |g_{nsi}(v) - p_\theta(v)F(v)|/p_\theta(v)^2] \\
&\quad \times O\{[(\log n)/(nh)]^{1/2}\} + O[(\log n)/(nh)]
\end{aligned} \tag{A.29}$$

Econometric Theory,V 10, No. 5, 821 - 848

almost surely by (A.3) and $h/s < 1$. By arguments similar to those of Bierens [1], $E[g_{nhi}(v) - p_\theta(v)F(v)]^2 = O[1/(nh)]$ uniformly over $v \in S_v$. By the Cauchy-Schwartz inequality, $E|g_{nhi}(v) - p_\theta(v)F(v)| = O[1/(nh)^{1/2}]$ uniformly over $v \in S_v$. Therefore, squaring (A.29) and taking expected values yields

$$[1 - (h/s)^r]^2 E[J_{ni}(v)^2] = O\{[(\log n)/(nh)]^2\} \tag{A.30}$$

uniformly over $v \in S_v$. Substituting (A.30) into (A.28) and using Assumption 4 yields $E[T_{n1}^2] = o(1)$. $T_{n1} = o_p(1)$ follows from Chebyshev's inequality. ■

Proof of Theorem 1. For $i = 1, \ldots, n$, define $U_i = Y_i - F(V_i)$ and $Z_i = (U_i, V_i)$. Also, for $v \in S_v$ and $i, j = 1, \ldots, n$, define

$$K_{hs}(v) = [1 - (h/s)^r]^{-1}[K(v/h) - (h/s)^{r+1}K(v/s)],$$

$$A_n(Z_i, Z_j) = [1/(nh^{1/2})]\, w(V_i)U_i[p_\theta(V_i)]^{-1} \times [U_j + F(V_j) - F(V_i)]K_{hs}(V_j - V_i),$$

$$\mu(Z_i) = E[A_n(Z_i, Z_j) + A_n(Z_j, Z_i)|Z_i],$$

$$H_n(Z_i, Z_j) = A_n(Z_i, Z_j) + A_n(Z_j, Z_i) - \mu(Z_i),$$

and

$$\Psi_n = \sum_{1 \le i < j \le n}\sum H_n(Z_i, Z_j).$$

It follows from (A.27) that

$$T_n = \Psi_n + \sum_{1 \le i < j \le n}\sum \mu(Z_i) + o_p(1). \tag{A.31}$$

Since $E(U_i) = 0$ for all $i = 1, \ldots, n$, and the U_i are independent, $E[\mu(Z_i)] = 0$, and $E[\mu(Z_i)\mu(Z_j)] = 0$ if $i \neq j$. Moreover, arguments similar to those of Bierens [1] yield

$$\mu(Z_i) = [1/(nh^{1/2})]\, w(V_i)U_i[p_\theta(V_i)]^{-1}o(h^{r+1})$$

uniformly over Z_i. Therefore, $E[\mu(Z_i)^2] = o(h^{2r+1}/n^2)$, so the second term on the right-hand side of (A.31) has mean 0 and variance $o(h^{2r+1})$. It follows from Chebyshev's inequality that this term $o_p(1)$ so that $T_n = \Psi_n + o_p(1)$. Therefore, to prove the theorem it suffices to show that $\Psi_n \xrightarrow{d} N(0, \sigma_T^2)$. Define

$$Q_n(Z_i, Z_j) = E[H_n(Z_\ell, Z_i)H_n(Z_\ell, Z_j)|Z_i, Z_j].$$

Lengthy but straightforward calculations show that

$$E[Q_n(Z_i, Z_j)^2]/\{E[H_n(Z_i, Z_j)^2]\}^2 \to 0,$$

$$n^{-1}E[H_n(Z_i, Z_j)^4]/\{E[H_n(Z_i, Z_j)^2]\}^2 \to 0,$$

and

$$(\tfrac{1}{2})n^2 E[H_n(Z_i, Z_j)^2] \to \sigma_T^2$$

as $n \to \infty$. Therefore, $\Psi_n \xrightarrow{d} N(0, \sigma_T^2)$ by Theorem 1 of Hall [5]. ■

A4. PROPERTIES OF T_n UNDER H_1

Proof of Theorem 2. Let $\{\theta_n\}$ be a nonstochastic sequence such that $n^{1/2}(\theta_n - \theta) = O(1)$. Let $\tilde{T}_n$ be defined as T_n with $\hat{\theta}_n$ replaced by θ_n. It suffices to show that $\operatorname{plim}_{n\to\infty}$

Horowitz, J. and Härdle, W. (1994)
Testing a Parametric Model against Semiparametric Alternatives.

Econometric Theory,V 10, No. 5, 821 - 848

$\hat{T}_n/(nh^{1/2}) > 0$. To do this, let $U = Y - H[v(X,\theta)]$ and $v_\theta = \partial v/\partial\theta$. Let θ_n^* denote a point between θ_n and θ (not necessarily the same point in each usage). Some algebra and Taylor series expansions yield

$$\hat{T}_n/(nh^{1/2}) = \sum_{\ell=1}^{12} R_{n\ell},$$

where

$$R_{n1} = n^{-1}\sum_{i=1}^{n} w[v(X_i,\theta_n)]U_i\{\hat{F}_{ni}[v(X_i,\theta_n)] - F_{ni}[v(X_i,\theta)]\},$$

$$R_{n2} = n^{-1}\sum_{i=1}^{n} w[v(X_i,\theta_n)]U_i\{F_{ni}[v(X_i,\theta)] - H[v(X_i,\theta)]\},$$

$$R_{n3} = n^{-1}\sum_{i=1}^{n} w[v(X_i,\theta_n)]U_i\{H[v(X_i,\theta)] - F[v(X_i,\theta)]\},$$

$$R_{n4} = -[(\theta_n - \theta)'/n]\sum_{i=1}^{n} w[v(X_i,\theta_n)]U_i F'[v(X_i,\theta_n^*)]v_\theta(X_i,\theta_n^*),$$

$$R_{n5} = n^{-1}\sum_{i=1}^{n} w[v(X_i,\theta_n)]\{H[v(X_i,\theta) - F[v(X_i,\theta)]\}$$
$$\times\{\hat{F}_{ni}[v(X_i,\theta_n)] - F_{ni}[v(X_i,\theta)]\},$$

$$R_{n6} = n^{-1}\sum_{i=1}^{n} w[v(X_i,\theta_n)]\{H[v(X_i,\theta)] - F[v(X_i,\theta)]\}$$
$$\times\{F_{ni}[v(X_i,\theta)] - H[v(X_i,\theta)]\},$$

$$R_{n7} = n^{-1}\sum_{i=1}^{n} w[v(X_i,\theta_n)]\{F[v(X_i,\theta)] - H[v(X_i,\theta)]\}^2,$$

$$R_{n8} = -[(\theta_n - \theta)'/n]\sum_{i=1}^{n} w[v(X_i,\theta_n)]F'[v(X_i,\theta_n^*)]v_\theta(X_i,\theta_n^*)$$
$$\times\{H[v(X_i,\theta)] - F[v(X_i,\theta)]\},$$

$$R_{n9} = -[(\theta_n - \theta)'/n]\sum_{i=1}^{n} w[v(X_i,\theta_n)]F'[v(X_i,\theta_n^*)]v_\theta(X_i,\theta_n^*)$$
$$\times\{\hat{F}_{ni}[v(X_i,\theta_n)] - F_{ni}[v(X_i,\theta)]\},$$

$$R_{n,10} = -[(\theta_n - \theta)'/n]\sum_{i=1}^{n} w[v(X_i,\theta_n)]F'[v(X_i,\theta_n^*)]v_\theta(X_i,\theta_n^*)$$
$$\times\{F_{ni}[v(X_i,\theta)] - H[v(X_i,\theta)]\},$$

$$R_{n,11} = [(\theta_n - \theta)'/n]\sum_{i=1}^{n} w[v(X_i,\theta_n)]$$
$$\times F'[v(X_i,\theta_n^*)]^2 v_\theta(X_i,\theta_n^*)v_{\theta'}(X_i,\theta_n^*)(\theta_n - \theta),$$

Econometric Theory,V 10, No. 5, 821 - 848

and

$$R_{n,12} = -[(\theta_n - \theta)'/n] \sum_{i=1}^{n} w[v(X_i,\theta_n)]F'[v(X_i,\theta_n^*)]v_\theta(X_i,\theta_n^*)$$
$$\times \{H[v(X_i,\theta)] - F[v(X_i,\theta)]\}.$$

R_{n1} is n^{-1} times the analogous quantity in Lemma 6. Therefore, $R_{n1} = o_p(1)$. By Lemmas 4 and 6,

$$R_{n2} = n^{-1} \sum_{i=1}^{n} w[v(X_i,\theta_n)]U_i\{G_{ni}[v(X_i,\theta)] - J_{ni}[v(X_i,\theta)]\} + o_p(1).$$

$R_{n2} = o_p(1)$ now follows from a proof similar to that of Lemma 7. It is not difficult to show that $R_{n7} \xrightarrow{p} E\{w(V)[H(V) - F(V)]^2\}$ and that the remaining R_{ni} are $o_p(1)$ as $n \to \infty$. Therefore, $\bar{T}_n/(nh^{1/2}) \xrightarrow{p} Ew(V)\{[H(V) - F(V)]^2\} > 0$. ■

Additional Assumptions and a Lemma Used in Proving Theorem 3. Under the sequence of local alternative models specified in Theorem 3, define $\bar{H}_n(v) = E[Y|v(X,\bar{\theta}_n) = v]$.

9. $\bar{H}_n$ has r continuous derivatives that are uniformly bounded over $v \in S_v$. Also, $p_\tau(v)$ has r derivatives that are bounded, continuous functions of $\tau \in N_\theta$ and $v \in \bar{S}_v$.
10. Define

$$\Gamma_n(x,v,\tau) = E_X\{\bar{H}_n[v(X,\bar{\theta}_n)][\partial v(x,\tau)\partial\tau - \partial v(X,\tau)/\partial\tau]|v(X,\tau) = v\}.$$

Let Γ_{nk} $(k = 1,\ldots,K)$ denote the kth component of Γ_n. There is a finite number M_Γ, not depending on τ or x, such that for all $\tau \in N_\theta$, $x \in S_X$, $v_1, v_2 \in S_v$, and $k = 1,\ldots,K$,

$$|\Gamma_{nk}(x,v_2,\tau) - \Gamma_{nk}(x,v_1,\tau)| \le M_\Gamma|v_1 - v_2|.$$

11. $\sigma^2(v)$ is a bounded, continuous function of $v \in S_v$ and $\theta \in N_\theta$ for all sufficiently large n. $E[Y - E[Y|v(x,\theta)] = v]^4$ is bounded uniformly over $v \in S_v$ and $\theta \in N_\theta$ for all sufficiently large n.
12. Define

$$Q_{nr}(v,h\xi) = (d^r/d\xi^r)\{[\bar{H}_n(v + h\xi) - \bar{H}_n(v)]p_{\bar{\theta}_n}(v + h\xi)\}.$$

For some $\alpha > 1/(4\delta)$, finite constant $C > 0$, and all sufficiently large n,

$$|Q_{nr}(v,h\xi) - Q_{nr}(v,s\xi)| \le C|h\xi - s\xi|^\alpha.$$

Assumptions 9–11 extend Assumptions 2 and 5–7 to the local alternative mean functions $\bar{H}_n$ and the density of $v(X,\bar{\theta}_n)$. Assumption 12 ensures that the bias of the kernel estimator of $E[Y|v(X,\bar{\theta}_n) = v]$ relative to its asymptotic distribution is $o(n^{-1/2}h^{-1/4})$. This property is needed for the result given in Theorem 3. ■

LEMMA 8. *Let the assumptions of Theorem* 3 *hold. Under the sequence of models* H_n, *the conclusions of Lemmas* 1–4 *hold when* F *and* θ *are replaced by* $\bar{H}_n$ *and* $\bar{\theta}_n$.

Proof. It may be verified that each step of the proofs of Lemmas 1–4 can be carried out under the assumptions of Lemma 8. ■

Econometric Theory,V 10, No. 5, 821 - 848

Proof of Theorem 3. Define $\{\theta_n\}$, θ_n^*, and $\bar{T}_n$ as in the proof of Theorem 2 but with $\bar{H}_n$ and $\bar{\theta}_n$ in place of H and θ. It suffices to show that the conclusion of Theorem 3 holds for $\bar{T}_n$. To do this, let $V = v(X, \bar{\theta}_n)$, $v_\theta = \partial v/\partial \bar{\theta}_n$, $U_i = Y_i - E[Y|V = v(X_i, \bar{\theta}_n)]$, and $w_{ni} = w[v(X_i, \theta_n)]$. Some algebra and Taylor series expansions yield

$$\bar{T}_n = (nh^{1/2}) \sum_{i=1}^{12} R_{n\ell},$$

where $R_{n\ell}$ ($\ell = 1, \ldots, 12$) is obtained by replacing H and θ with $\bar{H}_n$ and $\bar{\theta}_n$ in the corresponding terms in the proof of Theorem 2.

$(nh^{1/2})R_{n1} = o_p(1)$ by a proof identical to that for R_{n1} in Lemma 6. Convergence in distribution of $(nh^{1/2})R_{n2}$ follows from Lemma 8 and arguments identical to those used in proving Theorem 1. By using Chebyshev's inequality, it can be shown that $nh^{1/2}R_{n3} = o_p(1)$ and $nh^{1/2}R_{n4} = o_p(1)$. $nh^{1/2}R_{n5} = o_p(1)$ follows directly from Lemma 8. Also by Lemma 8,

$$nh^{1/2}R_{n6} = h^{1/2} \sum_{i=1}^{n} w_{ni} n^{-1/2} h^{-1/4} \Delta_n(V_i) G_{ni}(V_{ni}) + o_p(1),$$

where $V_{ni} = v(X_i, \bar{\theta}_n)$, and $\bar{H}_n$ and $\bar{\theta}_n$ replace F and θ in the definition of G_{ni}. Under Assumption 12, $E[G_{ni}(v)] = o[h^{1/4}/(nh^{1/2})]$ and $\mathrm{Var}[G_{ni}(v)] = O[(nh)^{-1}]$ uniformly over $v \in S_v$. Also, $\mathrm{Cov}[G_{ni}(v_i), G_{nj}(v_j)] = O(1/n)$. Therefore, $(nh^{1/2})R_{n6} = o_p(1)$. A Taylor series expansion may be used to show that $(n^{1/2}h^{1/4})\{E[Y|v(X, \bar{\theta}_n) = v(x, \bar{\theta}_n)] - F[v(x, \bar{\theta}_n)]\} \to \Delta^*[v(x, \bar{\theta}_n)]$ uniformly over $x \in \bar{S}_X$. Therefore, $(nh^{1/2})R_{n7} = \mu + o_p(1)$ by the strong law of large numbers. It is easily seen that $(nh^{1/2})R_{n8}$, $(nh^{1/2})R_{n9}$, $(nh^{1/2})R_{n,11}$, and $(nh^{1/2})R_{n,12}$ are all $o_p(1)$. $(nh^{1/2})R_{n,10} = R^*_{n,10} + o_p(1)$, by Lemma 8, where

$$R^*_{n,10} = -(\theta_n - \theta)'h^{1/2} \sum_{i=1}^{n} w_{ni} F'[v(X_i, \theta_n^*)] v_\theta(X_i, \theta_n^*)[G_{ni}(V_i) - J_{ni}(V_i)].$$

Also by Lemma 8, $E(R^*_{n,10}) = o(1)$. Arguments similar to those made for $n^{1/2}h^{1/4}R_{n6}$ yield $\mathrm{Var}(R^*_{n,10}) = o(1)$, so $R_{n,10} = o_p(1)$ by Chebyshev's inequality. ■

Horowitz, J. and Härdle, W. (1994)
Testing a Parametric Model against Semiparametric Alternatives.

Chapter 38

APPLIED NONPARAMETRIC METHODS

WOLFGANG HÄRDLE*

Humboldt-Universität Berlin

OLIVER LINTON†

Oxford University

Contents

*This work was prepared while the first author was visiting CentER, KUB Tilburg, The Netherlands. It was financed, in part, by contract No 26 of the programme "Pôle d'attraction interuniversitaire" of the Belgian government.

†We would like to thank Don Andrews, Roger Koenker, Jens Perch Nielsen, Tom Rothenberg and Richard Spady for helpful comments. Without the careful typewriting of Mariette Huysentruit and the skillful programming of Marlene Müller this work would not have been possible.

Handbook of Econometrics, Volume IV, Edited by R.F. Engle and D.L. McFadden

4. Handbook of Econometrics, North Holland, 38, 2295 - 2339

Härdle, W. and Linton, O. (1994) Applied Nonparametric Methods.

4. Handbook of Econometrics, North Holland, 38, 2295 - 2339

Abstract

We review different approaches to nonparametric density and regression estimation. Kernel estimators are motivated from local averaging and solving ill-posed problems. 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 semiparametric estimation.

1. Nonparametric estimation in econometrics

Although economic theory generally provides only loose restrictions on the distribution of observable quantities, much econometric work is based on tightly specified parametric models and likelihood based methods of inference. Under regularity conditions, maximum likelihood estimators consistently estimate the unknown parameters of the likelihood function. Furthermore, they are asymptotically normal (at convergence rate the square root of the sample size) with a limiting variance matrix that is minimal according to the Cramer–Rao theory. Hypothesis tests constructed from the likelihood ratio, Wald or Lagrange multiplier principle have therefore maximum local asymptotic power. However, when the parametric model is not true, these estimators may not be fully efficient, and in many cases – for example in regression when the functional form is misspecified – may not even be consistent. The costs of imposing the strong restrictions required for parametric estimation and testing can be considerable. Furthermore, as McFadden says in his 1985 presidential address to the Econometric Society, the parametric approach

> "interposes an untidy veil between econometric analysis and the propositions of economic theory, which are mostly abstract without specific dimensional or functional restrictions."

Therefore, much effort has gone into developing procedures that can be used in the absence of strong a priori restrictions. This survey examines nonparametric smoothing methods which do not impose parametric restrictions on functional form. We put emphasis on econometric applications and implementations on currently available computer technology.

There are many examples of density estimation in econometrics. Income distributions – see Hildenbrand and Hildenbrand (1986) – are of interest with regard to welfare analysis, while the density of stock returns has long been of interest to financial economists following Mandelbrot (1963) and Fama (1965). Figure 1 shows a density estimate of the stock return data of Pagan and Schwert (1990) in comparison with a normal density. We include a bandwidth factor in the scale parameter to correct for the finite sample bias of the kernel method.

Härdle, W. and Linton, O. (1994) Applied Nonparametric Methods.

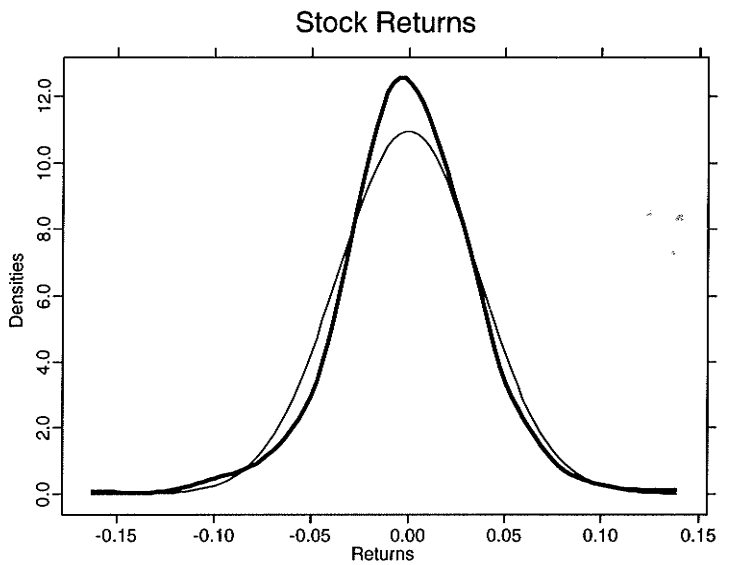


Figure 1 Density estimator of stock returns of Pagan and Schwert data compared with a mean zero normal density (thin line) with standard deviation $\sqrt{\hat{\sigma}^2 + \hat{h}^2}$, $\hat{\sigma} = 0.035$ and $\hat{h} = 0.009$, both evaluated at a grid of 100 equispaced points. Sample size was 1104. The bandwidth $\hat{h}$ was determined by the XploRe macro denauto according to Silverman's rule of thumb method.

Regression smoothing methods are used frequently in demand analysis – see for example Deaton (1991), Banks et al. (1993) and Hausman and Newey (1992). Figure 2 shows a nonparametric kernel regression estimate of the statistical Engel curve for food expenditure and total income. For comparison the (parametric) Leser curve is also included.

There are four main uses for nonparametric smoothing procedures. Firstly, they can be employed as a convenient and succinct means of displaying the features of a dataset and hence to aid practical parametric model building. Secondly, they can be used for diagnostic checking of an estimated parametric model. Thirdly, one may want to conduct inference under only the very weak restrictions imposed in fully nonparametric structures. Finally, nonparametric estimators are frequently required in the construction of estimators of Euclidean-valued quantities in semiparametric models.

By using smoothing methods one can broaden the class of structures under which the chosen procedure gives valid inference. Unfortunately, this robustness is not free. Centered nonparametric estimators converge at rate $\sqrt{nh}$, where $h \to 0$ is a smoothing parameter, which is slower than the $\sqrt{n}$ rate for parametric estimators in correctly specified models. It is also sometimes suggested that the asymptotic

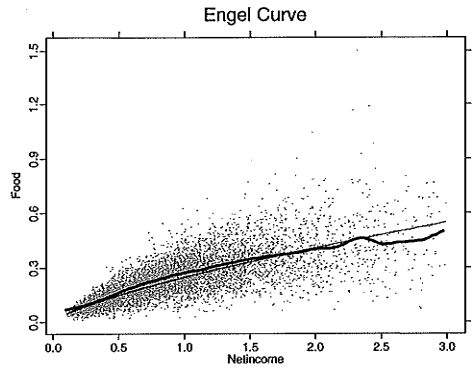


Figure 2. A kernel regression smoother applied to the food expenditure as a function of total income. Data from the Family Expenditure Survey (1968–1983), year 1973, Quartic kernel, bandwidth $h = 0.2$. The data have been normalized by mean income. Standard deviation of net income is 0.544. The kernel has been computed using the XploRe macro regest.

distributions themselves can be poor approximations in small samples. However, this problem is also found in parametric situations. The difference is quantitative rather than qualitative: typically, centered nonparametric estimators behave similarly to parametric ones in which n has been replaced by nh. The closeness of the approximation is investigated further in Hall (1992).

Smoothing techniques have a long history starting at least in 1857 when the Saxonian economist Engel 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. Rosenblatt (1956) proved asymptotic consistency of the kernel density estimator.

These methods have developed considerably in the last ten years, and are now frequently used by applied econometricians – see the recent survey by Deaton (1993). 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 (1993).

We base our survey of these methods around kernels. All the techniques we review for nonparametric regression are linear in the data, and thus can be viewed as kernel estimators with a certain equivalent weighting function. Since smoothing parameter selection methods and confidence intervals have been mostly studied for kernels,

4. Handbook of Econometrics, North Holland, 38, 2295 - 2339

we feel obliged to concentrate on these methods as the basic unit of account in nonparametric smoothing.

2. Density estimation

It is simplest to describe the nonparametric approach in the setting of density estimation, so we begin with that. Suppose we are given iid real-valued observations $\{X_i\}_{i=1}^n$ with density f. Sometimes – for the crossvalidation algorithm described in Section 4 and for semiparametric estimation – it is required to estimate f at each sample point, while on other occasions it is sufficient to estimate at a grid of points $x_1, \ldots, x_M$ for M fixed. We shall for the most part restrict our attention to the latter situation, and in particular concentrate on estimation at a single point x.

Below we give two approaches to estimating $f(x)$.

2.1. *Kernels as windows*

If f is smooth in a small neighborhood $[x-h, x+h]$ of x, we can justify the following approximation,

$$2h \cdot f(x) \approx \int_{x-h}^{x+h} f(u)\,du = P(X \in [x-h, x+h]), \tag{1}$$

by the mean value theorem. The right-hand side of (1) can be approximated by counting the number of X_i's in this small interval of length $2h$, and then dividing by n. This is a histogram estimator with *bincenter* x and *binwidth* $2h$. Let $K(u) = \frac{1}{2} I(|u| \leqslant 1)$, where $I(\cdot)$ is the indicator function taking the value 1 when the event is true and zero otherwise. Then the histogram estimator can be written as

$$\hat{f}_h(x) = n^{-1} \sum_{i=1}^{n} K_h(x - X_i), \tag{2}$$

where $K_h(\cdot) = h^{-1} K(\cdot/h)$. This is also a kernel density estimator of $f(x)$ with kernel $K(u) = \frac{1}{2} I(|u| \leqslant 1)$ and *bandwidth* h.

The step function kernel weights each observation inside the window equally, even though observations closer to x should possess better information than more distant ones. In addition, the step function estimator is discontinuous in x, which is unattractive given the smoothness assumption on f. Both objectives can be satisfied by choosing a smoother "window function" K as kernel, i.e. one for which $K(u) \to 0$ as $|u| \to 1$. One example is the so-called quartic kernel

$$K(u) = \tfrac{15}{16}(1 - u^2)^2 I(|u| \leqslant 1). \tag{3}$$

In the next section we give an alternative motivation for kernel estimators. The less technically able reader may skip this section.

Härdle, W. and Linton, O. (1994) Applied Nonparametric Methods.

2.2. *Kernels and ill-posed problems*

An alternative approach to the estimation of f is to find the best smooth approximation to the empirical distribution function and to take its derivative.

The distribution function F is related to f by

$$Af(x) = \int_{-\infty}^{\infty} I(u \leqslant x) f(u)\,\mathrm{d}u = F(x), \tag{4}$$

which is called a Fredholm equation with integral operator $Af(x) = \int_{-\infty}^{x} f(u)\,\mathrm{d}u$. Recovering the density from the distribution function is the same as finding the inverse of the operator A. In practice, we must replace the distribution function by the empirical distribution function (edf) $F_n(x) = n^{-1}\sum_{i=1}^{n} I(X_i \leqslant x)$, which converges to F at rate $\sqrt{n}$. However, this is a step function and cannot be differentiated to obtain an approximation to $f(x)$. Put another way, the Fredholm problem is ill-posed since for a sequence F_n tending to F, the solutions (satisfying $Af_n = F_n$) do not necessarily converge to f: the inverse operator in (4) is not continuous, see Vapnik (1982, p. 22).

Solutions to ill-posed problems can be obtained using the Tikhonov (1963) regularization method. Let $\Omega(f)$ be a lower semicontinuous functional called the *stabilizer*. The idea of the regularization method is to find indirectly a solution to $Af = F$ by use of the stabilizer. Note that the solution of $Af = F$ minimizes (with respect to $\hat{f}$)

$$\int_{-\infty}^{\infty}\left[\int_{-\infty}^{\infty} I(u \geqslant x)\hat{f}(u)\,\mathrm{d}u - F(x)\right]^2 \mathrm{d}x.$$

The stabilizer $\Omega(\hat{f}) = \|\hat{f}\|^2$ is now added to this equation with a Lagrange parameter λ,

$$R_\lambda(\hat{f}, F) = \int_{-\infty}^{\infty}\left[\int_{-\infty}^{\infty} I(x \geqslant u)\hat{f}(u)\,\mathrm{d}u - F(x)\right]^2 \mathrm{d}x + \lambda\int_{-\infty}^{\infty}\hat{f}^2(u)\,\mathrm{d}u. \tag{5}$$

Since we do not know $F(x)$, we replace it by the edf $F_n(x)$ and obtain the problem of minimizing the functional $R_\lambda(\hat{f}, F_n)$ with respect to $\hat{f}$.

A necessary condition for a solution $\hat{f}$ is

$$\int_{-\infty}^{\infty} I(x \geqslant u)\left[\int_{-\infty}^{\infty} I(x \geqslant s)\hat{f}(s)\,\mathrm{d}s - F_n(x)\right]\mathrm{d}x + \lambda\hat{f}(u) = 0.$$

Applying the Fourier transform for generalized functions and noting that the

Fourier transform of $I(u \geqslant 0)$ is $(\mathrm{i}/\omega) + \pi\delta(\omega)$ (with $\delta(\cdot)$ the delta function), we obtain

$$\left(\frac{1}{\mathrm{i}\omega}\right)\left[\left(-\frac{1}{\mathrm{i}\omega}\right)\Gamma(\omega) - n^{-1}\sum_{i=1}^{n}\left(-\frac{\mathrm{e}^{\mathrm{i}\omega X_i}}{\mathrm{i}\omega}\right)\right] + \lambda\Gamma(\omega) = 0,$$

where Γ is the Fourier transform of $\hat{f}$. Solving this equation for Γ and then applying the inverse Fourier transform, we obtain

$$\hat{f}(x) = n^{-1}\sum_{i=1}^{n}\frac{1}{2\sqrt{\lambda}}\mathrm{e}^{|x-X_i|/\sqrt{\lambda}}.$$

Thus we obtain a kernel estimator with kernel $K(u) = \frac{1}{2}\exp(-|u|)$ and bandwidth $h = \sqrt{\lambda}$. More details are given in Vapnik (1982, p. 302).

2.3. *Properties of kernels*

In the first two sections we derived different approaches to kernel smoothing. Here we would like to collect and summarize some properties of kernels. A *kernel* is a piecewise continuous function, symmetric around zero, integrating to one:

$$K(u) = K(-u); \qquad \int K(u)\,\mathrm{d}u = 1. \tag{6}$$

It need not have bounded support, although many commonly used kernels live on $[-1, 1]$. In most applications K is a positive probability density function, however for theoretical reasons it is sometimes useful to consider kernels that take on negative values. For any integer j, let

$$\mu_j(K) = \int u^j K(u)\,\mathrm{d}u; \qquad \nu_j(K) = \int K(u)^j\,\mathrm{d}u.$$

The order p of a kernel is defined as the first nonzero moment,

$$\mu_j = 0, \qquad j = 1, \ldots, p-1; \qquad \mu_p \neq 0. \tag{7}$$

We mostly restrict our attention to positive kernels which can be at most of order 2. An example of a higher order kernel (of order 4) is

$$K(u) = \tfrac{15}{32}(7u^4 - 10u^2 + 3)I(|u| \leqslant 1).$$

A list of common kernel functions is given in Table 1. We shall comment later on the values in the third column.

Table 1
Common kernel functions.

Kernel	$K(u)$	$D(K_{opt}, K)$
Epanechnikov	$\frac{3}{4}(1-u^2)I(\|u\| \leqslant 1)$	1
Quartic	$\frac{15}{16}(1-u^2)^2 I(\|u\| \leqslant 1)$	1.005
Triangular	$(1-\|u\|)I(\|u\| \leqslant 1)$	1.011
Gauss	$(2\pi)^{-1/2}\exp(-u^2/2)$	1.041
Uniform	$\frac{1}{2}I(\|u\| \leqslant 1)$	1.060

2.4. *Properties of the kernel density estimator*

The kernel estimator is a sum of iid random variables, and therefore

$$E[\hat{f}_h(x)] = \int K_h(x-z)f(z)\,dz = K_h * f(x), \tag{8}$$

where $*$ denotes convolution, assuming the integral exists. When f is $N(0,\sigma^2)$ and K is standard normal, $E[\hat{f}_h(x)]$ is therefore the normal density with standard deviation $\sqrt{\sigma^2+h^2}$ evaluated at x, see Silverman (1986, p. 37). This explains our modification to the normal density in Figure 1.

More generally, it is necessary to approximate $E[\hat{f}_h(x)]$ by a Taylor series expansion. Firstly, we change variables

$$E[\hat{f}_h(x)] = \int K(u)f(x-uh)\,du. \tag{9}$$

Then expanding $f(x-uh)$ about $f(x)$ gives

$$E[\hat{f}_h(x)] = f(x) + \tfrac{1}{2}h^2\mu_2(K)f''(x) + o(h^2), \tag{10}$$

provided $f''(x)$ is continuous in a neighborhood of x. Therefore, the bias of $\hat{f}_h(x)$ is $O(h^2)$ as $h \to 0$.

By similar calculation,

$$\mathrm{Var}[\hat{f}_h(x)] \approx \frac{1}{nh}v_2(K)f(x), \tag{11}$$

see Silverman (1986, p. 38). Therefore, provided $h \to 0$ and $nh \to \infty$, $\hat{f}_h(x) \xrightarrow{p} f(x)$. Further asymptotic properties of the kernel density estimator are given in Prakasa Rao (1983).

The statistical properties of $\hat{f}_h(x)$ depend closely on the bandwidth h: the bias

increases and the variance decreases with h. We investigate how the estimator itself depends on the bandwidth using the income data of Figure 2. Figure 3a shows a kernel density estimate for the income data with bandwidth $h = 0.2$ computed using the quartic kernel in Equation 3 and evaluated at a grid of 100 equispaced points. There is a clear bimodal structure for this implementation. A larger bandwidth $h = 0.4$ creates a single model structure as shown in Figure 3b, while a smaller $h = 0.05$ results in Figure 3c where, in addition to the bimodal feature, there is considerable small scale variation in the density.

It is therefore important to have some method of choosing h. This problem has been heavily researched – see Jones et al. (1992) for a collection of recent results and discussion. We take up the issue of automatic bandwidth selection in greater detail for the regression case in Section 4.2. We mention here one method that is frequently used in practice – Silverman's rule of thumb. Let $\hat{\sigma}^2$ be the sample variance of the data. Silverman (1986) proposed choosing the bandwidth to be

$$h = 1.364\left\{\frac{\nu_2(K)}{\mu_2^2(K)}\right\}^{1/5}\hat{\sigma}n^{-1/5}.$$

This rule is optimal (according to the IMSE – see Section 4 below) for the normal density, and is not far from optimal for most symmetric, unimodal densities. This procedure was used to select h in Figure 1.

2.5. *Estimation of multivariate densities, their derivatives and bias reduction*

A multivariate (d-dimensional) density function f can be estimated by the kernel estimator

$$\hat{f}_H(x) = \frac{1}{n}\sum_{i=1}^{n} k_H(x - X_i), \tag{12}$$

where $k_H(\cdot) = \{\det(H)\}^{-1}k(H^{-1}\cdot)$, where $k(\cdot)$ is a d-dimensional kernel function, while H is a d by d bandwidth matrix. A convenient choice in practice is to take $H = hS^{1/2}$, where S is the sample covariance matrix and h is a scalar bandwidth sequence, and to give k a product structure, i.e. let $k(u) = \prod_{j=1}^{d} K(u_j)$, where $u = (u_1, \ldots, u_d)^T$ and $K(\cdot)$ is a univariate kernel function.

Partial derivatives of f can be estimated by the appropriate partial derivatives of $\hat{f}_H(x)$ (providing $k(\cdot)$ has the same number of nonzero continuous derivatives). For any d-vector $r = (r_1, \ldots, r_d)$ and any function $g(\cdot)$ define

$$g^{(r)}(x) = \frac{\partial^{|r|}}{\partial^{r_1}x_1 \cdots \partial^{r_d}x_d}g(x),$$

where $|r| = \sum_{j=1}^{d} r_j$, then $\hat{f}_H^{(r)}(x)$ estimates $\hat{f}(x)$.

4. Handbook of Econometrics, North Holland, 38, 2295 - 2339

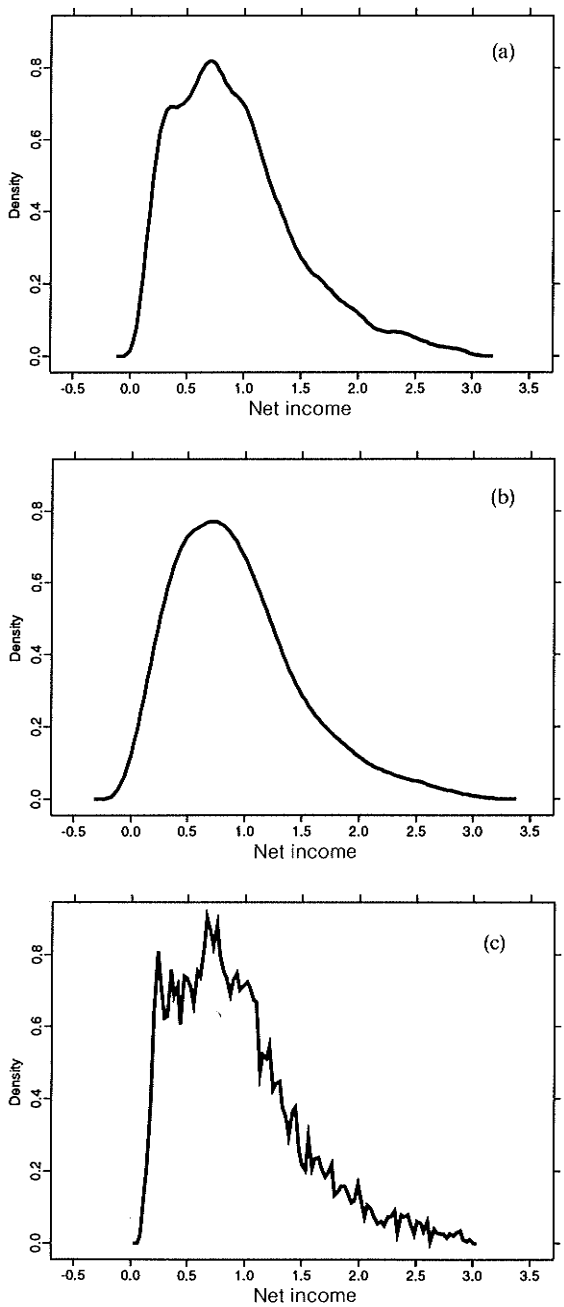


Figure 3. Kernel density estimates of net income distribution: (a) $h = 0.2$, (b) $h = 0.4$, (c) $h = 0.05$. Family Expenditure Survey (1968–1983). XploRe macro denest. Year 1973.

Härdle, W. and Linton, O. (1994) Applied Nonparametric Methods.

The properties of multivariate derivative estimators are described in Prakasa Rao (1983, p. 237). In fact, when a bandwidth $H = hA$ is used, where h is scalar and A is any fixed positive definite d by d matrix, then $\mathrm{Var}[\hat{f}_H^{(r)}(x)] = \mathrm{O}(n^{-1}h^{-(2|r|+d)})$, while the bias is $\mathrm{O}(h^2)$. For a given bandwidth h, the variance increases with the number of derivatives being estimated and with the dimensionality of X. The latter effect is well known as the *curse of dimensionality*.

It is possible to improve the order of magnitude of the bias by using a pth order kernel, where $p > 2$. In this case, the Taylor series expansion argument shows that $E[\hat{f}_h(x)] - f(x) = \mathrm{O}(h^p)$, where p is an even integer. Unfortunately, with this method there is the possibility of a negative density estimate, since K must be negative somewhere. Abramson (1982) and Jones et al. (1993) define *bias reduction* techniques that ensure a positive estimate. Jones and Foster (1993) review a number of other bias reduction methods.

The merits of bias reduction methods are based on asymptotic approximations. Marron and Wand (1992) derive exact expressions for the first two moments of higher order kernel estimators in a general class of mixture densities and find that unless very large samples are used, these estimators may not perform as well as the asymptotic approximations suggest. Unless otherwise stated, we restrict our attention to second order kernel estimators.

2.6. *Fast implementation of density estimation*

Fast evaluation of Equation 2 is especially important for optimization of the smoothing parameter. This topic will be treated in Section 4.2. If the kernel density estimator has to be computed at each observation point for k different bandwidths, the number of calculations are $\mathrm{O}(n^2hk)$ for kernels with bounded support. For the family expenditure dataset of Figure 1 with about 7000 observations this would take too long for the type of interactive data analysis we envisage. To resolve this problem we introduce the idea of discretization. The method is to map the raw data onto an equally spaced grid of smaller cardinality. All subsequent calculations are performed on this data summary which results in considerable computational savings.

Let $H_l(x; \Delta)$, $l = 0, 1, \ldots, M-1$, be the lth histogram estimator of $f(x)$ with origin l/M and small binwidth Δ. The sensitivity of histograms with respect to choice of origin is well known, see, e.g. Härdle (1991, Figure 1.16). However, if histograms with different origins are then repeatedly averaged, the result becomes independent of the histograms' origins. Let $\hat{f}_{M,\Delta}(x) = (1/M)\sum_{l=0}^{M} H_l(x; \Delta)$ be the averaged histogram estimator. Then

$$\hat{f}_{M,\Delta}(x) = \frac{1}{nh} \sum_{j \in \mathcal{I}} I(x \in B_j) \sum_{i=-M}^{M} n_{j-i} w_i, \tag{13}$$

4. Handbook of Econometrics, North Holland, 38, 2295 - 2339

where $\mathscr{Z} = \{\ldots, -1, 0, 1, \ldots\}$, $B_j = [b_j - \frac{1}{2}h, b_j + \frac{1}{2}h]$ with $h = \Delta/M$ and $b_j = jh$, while $n_j = \sum_{i=1}^{n} I(X_i \in B_j)$ and $w_i = (M - |i|/M)$. At the bincenters

$$\hat{f}_{M,\Delta}(b_j) = \frac{1}{nh} \sum_{i=-M}^{M} n_{j-i} w_i.$$

Note that $\{w_i\}_{i=-M}^{M}$ is, in fact, a discrete approximation to the (rescaled) triangular kernel $K(u) = (1 - |u|)I(|u| \leqslant 1)$. More generally, weights w_i can be used that represent the discretization of any kernel K. When K is supported on $[-1, 1]$, w_i is the rescaled evaluation of K at the points $-i/M$ $(i = -M, \ldots, M)$. If a kernel with non-compact support is used, such as the Gaussian for example, it is necessary to truncate the kernel function. Figure 4 shows the weights chosen from the quartic kernel with $M = 5$.

Since Equation 13 is essentially a convolution of the discrete kernel weights w_i with the bincounts n_j, modern statistical languages such as GAUSS or XploRe that supply a convolution command are very convenient for computation of Equation 13. Binning the data takes exactly n operations. If C denotes the number of nonempty bins, then evaluation of the binned estimator at the nonempty bins requires $O(MC)$ operations. In total we have a computational cost of $O(n + kM_{\max}C)$ operations for evaluating the binned estimator at k bandwidths, where $M_{\max} = \text{Max}\{M_j; j = 1, \ldots, k\}$, This is a big improvement.

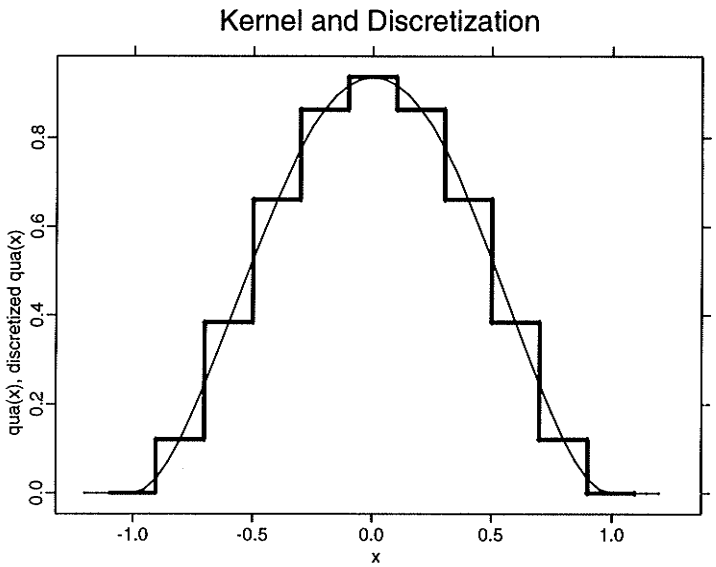


Figure 4. The quartic kernel $\text{qua}(u) = \frac{15}{16}(1 - u^2)^2 I(|u| \leqslant 1)$. Discretizing the kernel (without rescaling) leads to $w_{-i} = \text{qua}(i/M)$, $i = -M, \ldots, M$. Here $M = 5$ was chosen. The weights are represented by the thick step function.

The discretization technique also works for estimating derivatives and multivariate densities, see Härdle and Scott (1992) and Turlach (1992). This method is basically a time domain version of the Fast Fourier Transform computational approach advocated in Silverman (1986), see also Jones (1989).

3. Regression estimation

The most common method for studying the relationship between two variables X and Y is to estimate the conditional expectation function $m(x) = E(Y|X = x)$. Suppose that

$$Y_i = m(X_i) + \varepsilon_i, \qquad i = 1, \ldots, n, \tag{14}$$

where ε_i is an independent random error satisfying $E(\varepsilon_i|X_i = x) = 0$, and $\mathrm{Var}(\varepsilon_i|X_i = x) = \sigma^2(x)$. In this section we restrict our attention to independent sampling, but some extensions to the dependent sampling case are given in Section 5. The methods we consider are appropriate for both *random design*, where the (X_i, Y_i) are iid, and *fixed design*, where the 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 make inferences conditional on the sample $\{X_i\}_{i=1}^n$. However, many papers in the literature prove theoretical properties unconditionally, and we shall, for ease of exposition, present results in this form. We quote most results only for the case where X is scalar, although where appropriate we describe the extension to multivariate data.

In some cases, it is convenient to restrict attention to the equispaced design sequence $X_i = i/n$, $i = 1, \ldots, n$. Although this is unsuitable for most econometric applications, there are situations where it is of interest; specifically, time itself is conveniently described in this way. Also, the relative ranks of any variable (within a given sample) are naturally equispaced – see Anand et al. (1993).

The estimators of $m(x)$ we describe are all of the form $\sum_{i=1}^n W_{ni}(x) Y_i$ for some weighting sequence $\{W_{ni}(x)\}_{i=1}^n$, but arise from different motivations and possess different statistical properties.

3.1. *Kernel estimators*

Given the technique of kernel density estimation, a natural way to estimate $m(\cdot)$ is first to compute an estimate of the joint density $f(x, y)$ of (X, Y) and, then, to integrate it according to the formula

$$m(x) = \frac{\int y f(x, y)\, \mathrm{d}y}{\int f(x, y)\, \mathrm{d}y}. \tag{15}$$

4. Handbook of Econometrics, North Holland, 38, 2295 - 2339

The kernel density estimate $\hat{f}_h(x, y)$ of $f(x, y)$ is

$$\hat{f}_h(x, y) = n^{-1} \sum_{i=1}^{n} K_h(x - X_i) K_h(y - Y_i),$$

and by Equation 6

$$\int \hat{f}_h(x, y)\, dy = n^{-1} \sum_{i=1}^{n} K_h(x - X_i); \qquad \int y \hat{f}_h(x, y)\, dy = n^{-1} \sum_{i=1}^{n} K_h(x - X_i) Y_i.$$

Plugging these into the numerator and denominator of Equation 15 we obtain the Nadaraya–Watson kernel estimate

$$\hat{m}_h(x) = \frac{\sum_{i=1}^{n} K_h(x - X_i) Y_i}{\sum_{i=1}^{n} K_h(x - X_i)}. \tag{16}$$

The bandwidth h determines the degree of smoothness of $\hat{m}_h$. This can be immediately seen by considering the limits for h tending to zero or to infinity, respectively. Indeed, at an observation X_i, $\hat{m}_h(X_i) \to Y_i$, as $h \to 0$, while at an arbitrary point x, $\hat{m}_h(x) \to \bar{Y}$, 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. Conditions for consistency of $\hat{m}_h$ are given in the following theorem, proved in Schuster (1972):

Theorem 1

Let $K(\cdot)$ satisfy $\int |K(u)|\, du \leqslant \infty$ and $\mathrm{Lim}_{|u| \to \infty} uK(u) = 0$. Suppose also that $m(x)$, $f(x)$, and $\sigma^2(x)$ are continuous at x, and $f(x) > 0$. Then, provided $h = h(n) \to 0$ and $nh \to \infty$ as $n \to \infty$, we have

$$\hat{m}_h(x) \xrightarrow{\mathrm{P}} m(x).$$

The kernel estimator is asymptotically normal, as was first shown in Schuster (1972).

Theorem 2

Suppose in addition to the conditions of Theorem 1 that $\int |K(u)|^{2+\eta}\, du < \infty$, for some $\eta > 0$. Suppose also that $m(x)$ and $f(x)$ are twice continuously differentiable at x and that $E(|Y|^{2+\eta}|x)$ exists and is continuous at x. Finally, suppose that

$\overline{\text{Lim}}\, h^5 n < \infty$. Then

$$\sqrt{nh}[\hat{m}_h(x) - m(x) - h^2 B_{nw}(x)] \Rightarrow N(0, V_{nw}(x)),$$

where

$$B_{nw}(x) = \tfrac{1}{2}\mu_2(K)\left[m''(x) + 2m'(x)\frac{f'}{f}(x) \right]$$

$$V_{nw}(x) = \nu_2(K)\sigma^2(x)/f(x).$$

The Nadaraya–Watson estimator has an obvious generalization to d-dimensional explanatory variables and pth order kernels. In this case, assuming a common bandwidth h is used, the (asymptotic) bias is $O(h^p)$, when p is an even integer, while the (asymptotic) variance is $O(n^{-1}h^{-d})$.

3.2. *k-Nearest neighbor estimators*

3.2.1. *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 $\mathcal{N}(x) = \{i: X_i \text{ is one of the } k\text{-}NN \text{ to } x\}$ 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 $\mathcal{N}(x)$,

$$\hat{m}_k(x) = \frac{1}{k}\sum_{i \in \mathcal{N}(x)} Y_i. \tag{17}$$

Connections to kernel smoothing can be made by considering Equation 17 as a kernel smoother with uniform kernel $K(u) = \frac{1}{2}I(|u| \leqslant 1)$ and variable bandwidth $h = R(k)$, the distance between x and its furthest k-NN,

$$\hat{m}_k(x) = \frac{\sum_{i=1}^{n} K\left(\dfrac{x - X_i}{R}\right) Y_i}{\sum_{i=1}^{n} K\left(\dfrac{x - X_i}{R}\right)}. \tag{18}$$

Note that in Equation 18, for this specific kernel, the denominator is equal to (k/nR) the k-NN density estimate of $f(x)$. The formula in Equation 18 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 3

Let the conditions of Theorem 2 hold, except that $k \to \infty$, $k/n \to 0$ and $\overline{\lim} k^5/n^4 < \infty$ as $n \to \infty$. Then

$$\sqrt{k}[\hat{m}_k(x) - m(x) - (k/n)^2 B_{nn}(x)] \Rightarrow N(0, V_{nn}(x)),$$

where

$$B_{nn}(x) = \mu_2(K)\left[\frac{m''(x) + 2m'(x)\dfrac{f'}{f}(x)}{8f^2(x)}\right]$$

$$V_{nn}(x) = 2\sigma^2(x)v_2(K).$$

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 2. An approximate identity between *k-NN* and kernel smoothers can be obtained by setting

$$k = 2nhf(x), \tag{19}$$

or equivalently $h = k/[2nf(x)]$. For this choice of k or h respectively, the asymptotic mean squared error formulas of Theorem 2 and Theorem 3 are identical.

3.2.2. Symmetrized k-NN estimators

A computationally useful modification of $\hat{m}_k$ 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). The bias formulas are slightly different, see Härdle and Carroll (1990), but Equation 19 remains true.

3.3. Local polynomial estimators

The Nadaraya–Watson estimator can be regarded as the solution of the minimization problem

$$\hat{m}_h(x) = \arg_\theta \min \sum_{i=1}^{n} K_h(x - X_i). \tag{20}$$

This motivates the local polynomial class of estimators. Let $\hat{\theta}_0, \ldots, \hat{\theta}_p$ minimize

$$\sum_{i=1}^{n} K_h(x - X_i)\left[Y_i - \theta_0 - \theta_1(X_i - x) - \cdots - \theta_p \frac{(X_i - x)^p}{p!} \right]^2. \tag{21}$$

Then $\hat{\theta}_0$ serves as an estimator of $m(x)$, while $\hat{\theta}_j$ estimates the jth derivative of m. Clearly, $\hat{\theta}_0$ is linear in Y. 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 $\hat{m}_{h,1}(x)$.

Theorem 4

Let the conditions of Theorem 2 hold. Then

$$\sqrt{nh}[\hat{m}_{h,1}(x) - m(x) - h^2 B_1(x)] \Rightarrow N(0, V_1(x)),$$

where

$$B_1(x) = \tfrac{1}{2}\mu_2(K)m''(x)$$

$$V_1(x) = \nu_2(K)\sigma^2(x)/f(x).$$

The local linear estimator is unbiased when m is linear, while the Nadaraya–Watson estimator may be biased depending on the marginal density of the design.

We note here that fitting higher order polynomials can result in bias reduction, see Fan and Gijbels (1992) and Ruppert and Wand (1992) – who also extend the analysis to multidimensional explanatory variables.

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 Equation 21. Fan, Heckman and Wand (1992) developed a theory for a nonparametric estimator in a GLIM (Limited Dependent Variable) model in which, for example, a probit likelihood function replaces the polynomial in Equation 21. An advantage of this procedure is that low bias results when the parametric model is true (Linton and Nielsen 1993).

3.4. *Spline estimators*

For any estimate $\hat{m}$ of m, the residual sum of squares (RSS) is defined as $\sum_{i=1}^{n}[Y_i - \hat{m}(X_i)]^2$, which is a widely used criterion, in other contexts, for generating estimators of regression functions. However, the RSS is minimized by $\hat{m}$ interpolating the data, assuming no ties in the X's. To avoid this problem it is necessary to add a stabilizer. Most work is based on the stabilizer $\Omega(\hat{m}) = \int[\hat{m}''(u)]^2\,du$, although see

Ansley et al. (1993) and Koenker et al. (1993) for alternatives. The cubic spline estimator $\hat{m}_\lambda$ is the (unique) minimizer of

$$R_\lambda(\hat{m}, m) = \sum_{i=1}^{n} [Y_i - \hat{m}(X_i)]^2 + \lambda \int [\hat{m}''(u)]^2 \, du. \tag{22}$$

The spline $\hat{m}_\lambda$ has the following properties. It is a cubic polynomial between two successive X-values at the observation points $\hat{m}_\lambda(\cdot)$ and its first two derivatives are continuous; at the boundary of the observation interval the spline is linear. This characterization of the solution to Equation 22 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 $\hat{m}_\lambda$. As $\lambda \to 0, \hat{m}_\lambda$ interpolates the observations, while if $\lambda \to \infty, \hat{m}_\lambda$ tends to a least squares regression line. Although $\hat{m}_\lambda$ is linear in the Y data, see Härdle (1990, pp 58–59), 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

$$K(u) = \frac{1}{2} \exp\left(-\frac{|u|}{\sqrt{2}} \right) \sin\left(\frac{|u|}{\sqrt{2}} + \frac{\pi}{4} \right), \tag{23}$$

which is of fourth order, since its first three moments are zero, while the equivalent bandwidth $h = h(\lambda; X_i)$ is

$$h(\lambda; X_i) = \lambda^{1/4} n^{-1/4} f(X_i)^{-1/4}. \tag{24}$$

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. However, from Section 2.2 we conclude that this interpretation can also be given to kernel estimators.

3.5. *Series estimators*

Series estimators have received considerable attention in the econometrics literature, following Elbadawi et al. (1983). This theory is very much tied to the structure of

Hilbert space. Suppose that m has an expansion for all x:

$$m(x) = \sum_{j=0}^{\infty} \beta_j \varphi_j(x), \tag{25}$$

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), and the *Fourier* series used in Gallant and Souza (1991).

A simple method of estimating $m(x)$ involves firstly selecting a basis system and a truncation sequence $t(n)$, where $t(n)$ is an integer less than n, and then regressing Y_i on $\varphi_{ti} = (\varphi_0(X_i), \ldots, \varphi_t(X_i))^T$. Let $\{\hat{\beta}_j\}_{j=0}^{t(n)}$ be the least squares "parameter" estimates, then

$$\hat{m}_{t(n)}(x) \sum_{j=0}^{t(n)} \hat{\beta}_j \varphi_j(x) = \sum_{i=1}^{n} W_{ni}(x) Y_i, \tag{26}$$

where $W_n(x) = (W_{n1}, \ldots, W_{nn})^T$, with

$$W_n(x) = \varphi_{tx}^T (\boldsymbol{\Phi}_t^T \boldsymbol{\Phi}_t)^{-1} \boldsymbol{\Phi}_t^T, \tag{27}$$

where $\varphi_{tx} = (\varphi_0(x), \ldots, \varphi_t(x))^T$ and $\boldsymbol{\Phi}_t = (\varphi_{t1}, \ldots, \varphi_{tn})^T$.

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, provided $t(n)$ grows at a sufficiently fast rate, the optimal (given the smoothness of m) rate of convergence can be established – see Stone (1982), while fixed window kernels achieve at best a rate of convergence (of MSE) of $n^{4/5}$. However, the same effect can be achieved by using a kernel estimator, where the order of the kernel 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 $t(n)$.

3.6. *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, pp 352–355 in Chambers et al. (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 5a we have plotted together with the raw data a kernel smoother $\hat{m}_h$ for which a quartic kernel was used with $h = 2000$. Very similar to this is the spline smoother shown in Figure 5b ($\lambda = 10^9$). In this example, the X's are not too far from uniform. The effective local bandwidth for the spline smoother from Equation 24 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 5d) with $t = 8$ is quite close to the spline estimator.

In contrast to these regression estimators stands the *k-NN* smoother ($k = 11$) in Figure 5c. We used the symmetrized *k-NN* estimator for this plot. By formula (19) 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.

3.7. *Confidence intervals*

The asymptotic distribution results contained in Theorems 2–4 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.

In this section we restrict our attention to the Nadaraya–Watson regression estimator. In this case, we suppose that $hn^{1/5} \to 0$, which ensures that the bias term does not appear in the limiting distribution. Let

$$\mathrm{CLO}(x) = \hat{m}_h(x) - c_{\alpha/2}\hat{s}$$

$$\mathrm{CUP}(x) = \hat{m}_h(x) + c_{\alpha/2}\hat{s},$$

where $\Phi(c_\alpha) = (1 - \alpha)$ with $\Phi(\cdot)$ the standard normal distribution, while $\hat{s}^2$ is a consistent estimate of the asymptotic variance of $\hat{m}_h(x)$. Suitable estimators include

(1) $\hat{s}_1^2 = n^{-1}h^{-1}v_2(K)\hat{\sigma}_h^2(x)/\hat{f}_h(x)$

(2) $\hat{s}_2^2 = \hat{\sigma}_h^2(x)\sum_{i=1}^{n} W_{ni}^2(x)$

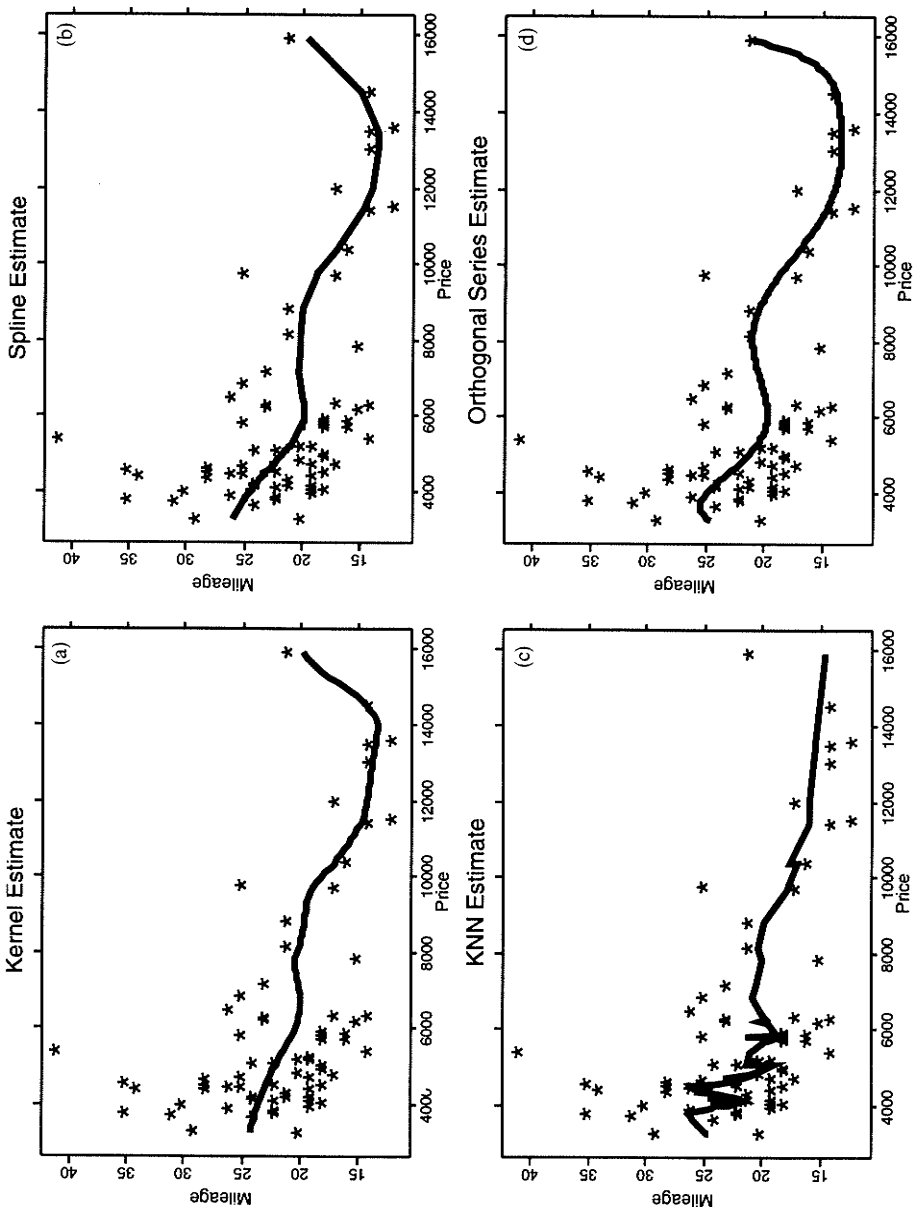


Figure 5(a–d). Scatterplot of car price (x) and miles per gallon (y) with four different smooth approximations ($n = 74, h = 2000, k = 11, \lambda = 10^9, t = 8$). Standard deviation of car price is 2918.

$$(3)\quad \hat{s}_3^2 = \sum_{i=1}^{n} W_{ni}^2(x)\hat{\varepsilon}_i^2,$$

where $\hat{f}_h(x)$ is defined in Equation 2, $\hat{\varepsilon}_i = Y_i - \hat{m}_h(X_i)$ are the nonparametric residuals and $\hat{\sigma}_h^2(x) = \sum_{i=1}^{n} W_{ni}(x)\hat{\varepsilon}_i^2$ is a nonparametric estimator of $\sigma^2(x)$ – see Robinson (1987) and Hildenbrand and Kneip (1992) for a discussion of alternative conditional variance estimators and their application.

With the above definitions.

$$P\{m(x)\in[\mathrm{CLO}(x), \mathrm{CUP}(x)]\} \to 1-\alpha. \tag{28}$$

These confidence intervals are frequently employed in econometric applications, see for example Bierens and Pott-Buter (1990), Banks et al. (1993) and Gozalo (1989). This 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 $\mathrm{CLO}^*(x)$ and $\mathrm{CUP}^*(x)$ such that

$$P\{m(x)\in[\mathrm{CLO}^*(x), \mathrm{CUP}^*(x)] \quad \forall x\in\chi\} \to 1-\alpha. \tag{29}$$

Let

$$\mathrm{CLO}^*(x) = \hat{m}_h(x) - \left\{\frac{c_\alpha^*}{\delta} + \delta + \frac{1}{2\delta}\ln\left[\frac{v_2(K')}{4\pi^2 v_2(K)}\right]\right\}\hat{s}_1,$$

$$\mathrm{CUP}^*(x) = \hat{m}_h(x) + \left\{\frac{c_\alpha^*}{\delta} + \delta + \frac{1}{2\delta}\ln\left[\frac{v_2(K')}{4\pi^2 v_2(K)}\right]\right\}\hat{s}_1,$$

where $\delta = \sqrt{2\log(1/h)}$, and $\exp[-2\exp(-c_\alpha^*)] = (1-\alpha)$. Then (29) is satisfied under the conditions given in Härdle (1990, Theorem 4.3.1). See also Prakasa Rao (1983, Theorem 2.1.17) for a treatment of the same problem for density estimators.

In Figure 6 we show the uniform confidence bands for the income data of Figure 2.

Hall (1993) advocates using the bootstrap to construct uniform confidence bands. He argues that the error in (29) is $O(1/\log n)$, which can be improved to $O((\log h^{-1})^3/nh)$ by the judicious use of this resampling method in the random design case. See also Hall (1992) and Härdle (1990) for further applications of the bootstrap in nonparametric statistics.

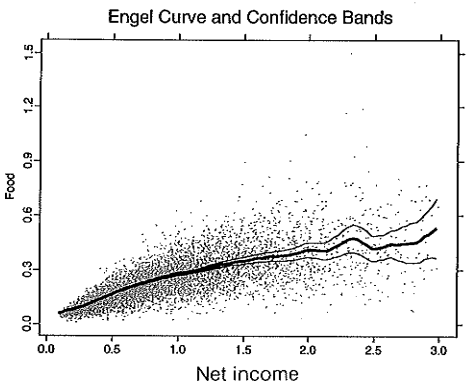


Figure 6. Uniform confidence bands for the income data. Food versus net income. Calculated using XploRe macro reguncb.

3.8. *Regression derivatives and quantiles*

There are a number of other functionals of the conditional distribution that are of interest for applications. The first derivative of the regression function measures the strength of the relationship between Y and X, while second derivatives can quantify the concavity or convexity of the regression function. Let $\hat{m}(x)$ be any estimator of $m(x)$ that has at least r non-zero derivatives at x. Then $m^{(r)}(x)$ can be estimated by the rth derivative of $\hat{m}(x)$, denoted $\hat{m}^{(r)}(x)$. Müller (1988) describes kernel estimators of $m^{(r)}(x)$ based on the convolution method of Gasser and Müller (1984); their method gives simpler bias expressions than the Nadaraya–Watson estimator. An alternative technique is to fit a local polynomial (of order r) estimator, and take the coefficient on the rth term in (21), see Ruppert and Wand (1992). In each case, the resulting estimator is linear in Y_i, with bias of order h^2 and variance of order $n^{-1}h^{-(2r+1)}$.

Quantiles can also be useful. The median is an alternative – and robust – measure of location, while other quantiles can help to describe the spread of the conditional distribution. Let $f_{Y|X=x}(y)$ denote the conditional distribution of Y given $X = x$, and let $c_\alpha(x)$ be the αth conditional quantile, i.e.

$$\alpha = \int_{-\infty}^{C_\alpha(x)} f_{Y|X=x}(y)\,dy, \tag{30}$$

where for simplicity we assume this is unique. There are several methods for estimating $c_\alpha(x)$.

Firstly, let $Z_j = [W_{nj}(x), Y_j]^T$, where $W_{nj}(x)$ are kernel or nearest neighbor weights. We first sort $\{Z_j\}_{j=1}^n$ on the variable Y_j, and find the largest index J such that

$$\sum_{j=1}^{J} W_{nj}(x) \leqslant \alpha.$$

Then let

$$\hat{c}_\alpha(x) = Y_J. \tag{31}$$

Stute (1986) shows that $\hat{c}_\alpha(x)$ consistently estimates $c_\alpha(x)$, with the same convergence rates as in ordinary nonparametric regression, see also Bhattacharya and Gangopadhyay (1990). When K is the uniform kernel and $\alpha = \frac{1}{2}$, this procedure corresponds to the running median discussed in Härdle (1990, pp 69–71). A smoother estimator is obtained by also smoothing in the y direction, i.e.

$$\hat{c}_\alpha(x) = \frac{1}{n}\sum_{j=1}^{n} K_h\left(\frac{J-j}{n}\right) Y_j.$$

Provided K has at least r non-zero derivatives, the rth derivative of $c_\alpha(x)$ can be estimated by the rth derivative of $\hat{c}_\alpha(x)$. See Anand et al. (1993) and Robb et al. (1992) for applications.

An alternative method of estimating conditional quantiles is through minimizing an appropriate loss function. This idea originated in Koenker and Bassett (1978). In particular,

$$\hat{c}_\alpha(x) = \arg_\theta \min \sum_{i=1}^{n} K_h(x - X_i)\rho_\alpha(Y_i - \theta), \tag{32}$$

where $\rho_\alpha(y) = |y| + (2\alpha - 1)y$, consistently estimates $c_\alpha(x)$. Computation of the estimator can be carried out by linear programming techniques. Chaudhuri (1991) provides asymptotic theory for this estimator in a general multidimensional context and for estimators of the derivatives of $c_\alpha(x)$.

In neither (31) nor (32) is the estimator linear in Y_i, although the asymptotic distribution of the estimators are determined by a linear approximation to them, i.e. the estimators are asymptotically normal.

4. Optimality and bandwidth choice

4.1. *Optimality*

Let $Q(h)$ be a performance criterion. We say that a bandwidth sequence h^* is asymptotically optimal if

$$\frac{Q(h^*)}{\inf_{h \in H_n} Q(h)} \xrightarrow{p} 1, \tag{33}$$

as $n \to \infty$, where H_n is the range of permissible bandwidths. There are a number of alternative optimality criteria in use. Finally, we may be interested in the quadratic

loss of the estimator at a single point x, which is measured by the *Mean squared error*, $\mathrm{MSE}\{\hat{m}_h(x)\}$. Secondly, we may be only concerned with a global measure of performance. In this case, we may consider the *Integrated mean squared error*, $\mathrm{IMSE} = \int \mathrm{MSE}[\hat{m}_h(x)]\pi(x)f(x)\,dx$ for some weighting function $\pi(\cdot)$. An alternative is the in-sample version of this, the *averaged squared error*

$$d_A(h) = n^{-1}\sum_{j=1}^{n}[\hat{m}_h(X_j) - m(X_j)]^2\pi(X_j). \tag{34}$$

The purpose of $\pi(\cdot)$ may be to downweight observations in the tail of X's distribution, and thereby to eliminate boundary effects – see Müller (1988) for a discussion. When $h = \mathrm{O}(n^{-1/5})$, 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 $\mathrm{O}(n^{-4/5})$ 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 2 we obtain an approximate mean squared error expansion,

$$\mathrm{MSE}[\hat{m}_h(x)] \approx n^{-1}h^{-1}V(x) + h^4B^2(x). \tag{35}$$

and the bandwidth minimizing Equation 35 is

$$h_0(x) = \left[\frac{V(x)}{4B^2(x)}\right]^{1/5} n^{-1/5}. \tag{36}$$

Similarly, the optimal bandwidth with respect to IMSE is the same as in (36) 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 4.2 below how one can obtain empirical versions of (36).

The optimal local bandwidths can vary considerably with x, a point which is best illustrated for density estimation. Suppose that the density is standard normal and a standard normal kernel is used. In this case, as $x \to \infty$, $h_0(x) \to \infty$: when data is sparse a wider window is called for. Also at $x = \pm 1$, $h_0(x) = \infty$, which reflects the fact that $\phi'' = 0$ at these points. Elsewhere, substantially less smoothing is called for: at ± 2.236, $h_0(x) = 0.884n^{-1/5}$ (which is the minimum value of $h_0(x)$). The optimal global bandwidth is $1.06n^{-1/5}$.

Although allowing the bandwidth to vary with x dominates over the strategy of throughout choosing a single bandwidth, in practice this requires considerably more computation, and is rarely used in applications.

By substituting h_0 in (35), we find that the optimal MSE and IMSE depend on

Table 2
Kernel exchange rate.

s_j^*/s_i^*	Uniform	Triangle	Epanechnikov	Quartic	Gaussian
Uniform	1.000	0.715	0.786	0.663	1.740
Triangle	1.398	1.000	1.099	0.927	2.432
Epanechnikov	1.272	0.910	1.000	0.844	2.214
Quartic	1.507	1.078	1.185	1.000	2.623
Gaussian	0.575	0.411	0.452	0.381	1.000

K only through

$$T(K) = v_2^2(K)\mu_2(K). \tag{37}$$

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 et al. (1985). A kernel is said to be optimal if it minimizes (37). The optimal kernel of order 2 is the Epanechnikov kernel given in Table 1. The third column of this table shows the loss in efficiency of other kernels in relation to this optimal one. 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.

Any kernel can be rescaled as $K^*(\cdot) = s^{-1}K(\cdot/s)$ which of course changes the value of the kernel constants and hence h_0. In particular,

$$v_2(K^*) = s^{-1}v_2(K); \qquad \mu_2^2(K^*) = s^2\mu_2(K).$$

We can uncouple the scaling effect by using for each kernel K, that K^* with scale

$$s^* = \left[\frac{v_2(K^*)}{\mu_2^2(K)}\right]^{1/5}$$

for which $\mu_2^2(K^*) = v_2(K^*)$. Now suppose we wish to compare two smooths with kernels K_j and bandwidths h_j respectively. This can be done by transforming both to their canonical scale, see Marron and Nolan (1989), and then comparing their s_j^*. In Table 2 we give the exchange rate between various commonly used kernels. For example, the bandwidth of 0.2 used with a quartic kernel in Figure 2, translates into a bandwidth of 0.133 for a uniform kernel and 0.076 for a Gaussian kernel.

4.2. *Choice of smoothing parameter*

For each nonparametric regression method, one has to choose how much to smooth for the given dataset. In Section 3 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.

4.2.1. Plug-in

The asymptotic approximation given in (36) can be used to determine an optimal local bandwidth. We can calculate an estimated optimal bandwidth $\hat{h}_{pl}$ in which the consistent estimators $\hat{m}''_{h^*}(x)$, $\hat{\sigma}^2_{h^*}(x)$, $\hat{f}_{h^*}(x)$ and $\hat{f}'_{h^*}(x)$ replace the unknown functions. We then use $\hat{m}_{\hat{h}_{pl}}(x)$ 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 et al. (1992a).

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.

4.2.2. 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 $m(X_j)$ by the observations Y_j:

$$p(h) = n^{-1} \sum_{j=1}^{n} [\hat{m}_h(X_j) - Y_j]^2 \pi(X_j).$$

This is called the resubstitution estimate.

However, this quantity makes use of each observation twice – the response variable Y_j is used in $\hat{m}_h(X_j)$ to predict itself. Therefore, $p(h)$ can be made arbitrarily small by taking $h \to 0$ (when there are no tied X observations). This fact can be expressed via asymptotic expressions for the moments of p. Conditional on $X_1, \ldots, X_n$, we have

$$E[p(h)] = E[d_A(h)] + \frac{1}{n} \sum_{i=1}^{n} \sigma^2(X_i)\pi(X_i) - 2\frac{1}{n} \sum_{i=1}^{n} W_{ni}(X_i)\sigma^2(X_i)\pi(X_i), \tag{38}$$

and the third term is of the same order of magnitude as $E[d_A(h)]$, but with a 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 jth observation

$$\hat{m}_{h,j}(X_j) = \frac{\sum_{j \neq i} K_h(X_j - X_i) Y_i}{\sum_{j \neq i} K_h(X_j - X_i)}. \tag{39}$$

This leave-one-out estimate is used to form the so-called crossvalidation function

$$\mathrm{CV}(h) = n^{-1} \sum_{j=1}^{n} [\hat{m}_{h,j}(X_j) - Y_j]^2, \tag{40}$$

which is to be minimized with respect to h. For technical reasons, the minimum must be taken only over a restricted set of bandwidths such as $H_n = [n^{-(1/5-\zeta)}, n^{-(1/5+\zeta)}]$, for some $\zeta > 0$.

Theorem 5

Assume that the conditions given in Härdle (1990, Theorem 5.1.1) hold.

Then the bandwidth selection rule, "Choose $\hat{h}$ to minimize CV(h)" is asymptotically optimal with respect to $d_A(h)$ and IMSE.

Proof

See Härdle and Marron (1985).

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 the plug-in procedure, m and f need not be differentiable (a Lipschitz condition is required, however).

4.2.3. *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 h's which are too small. The general form of this selector is

$$G(h) = n^{-1} \sum_{j=1}^{n} [\hat{m}_h(X_i) - Y_i]^2 \pi(X_i) \Xi[W_{ni}(X_i)],$$

where Ξ is the correction function with first-order Taylor expansion

$$\Xi(u) = 1 + 2u + \mathrm{O}(u^2), \tag{41}$$

as $u \to 0$. Some well known examples are:

(i) Generalized crossvalidation (Craven and Wahba 1979; Li 1985),

$$\Xi_{\mathrm{GCV}}(u) = (1-u)^{-2};$$

(ii) Akaike's information criterion (Akaike 1970)

$$\Xi_{\mathrm{AIC}}(u) = \exp 2u;$$

(iii) Finite prediction error (Akaike 1974).

$$\Xi_{\mathrm{FPE}}(u) = (1+u)/(1-u);$$

(iv) Shibata's (1981) model selector,

$$\Xi_S(u) = 1 + 2u;$$

(v) Rice's (1984) bandwidth selector,

$$\Xi_T(u) = (1-2u)^{-1}.$$

Härdle et al. (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.

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 7.

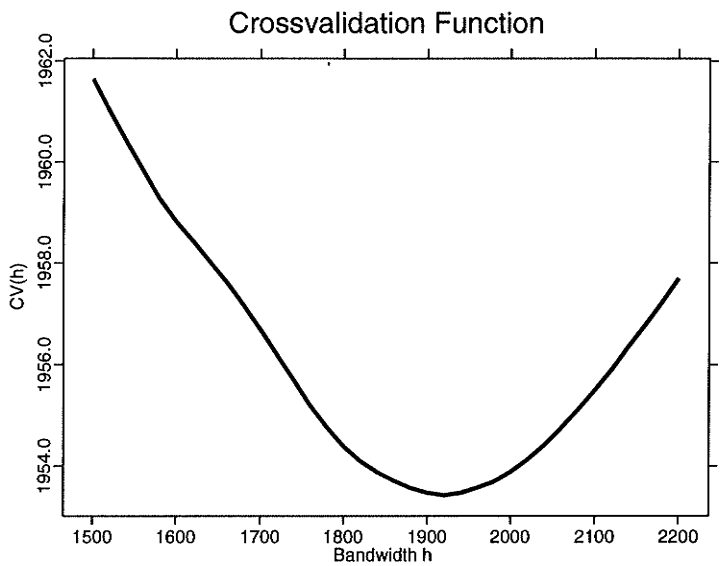


Figure 7. The crossvalidation function CV(h) for the car data. Quartic kernel. Computation made with XploRe macro regcvl.

The computation is for the quartic kernel using the WARPing method, see Härdle and Scott (1992). The minimal $\hat{h} = \arg\min \mathrm{CV}(h)$ is at 1922 which shows that in Figure 5a we used slightly too large a bandwidth.

Härdle et al. (1988) investigate how far the crossvalidation optimal $\hat{h}$ is from the true optimum, $\hat{h}_0$ (that minimizes $d_A(h)$). They show that for each optimization method,

$$n^{1/10}\left(\frac{\hat{h}-\hat{h}_0}{\hat{h}_0}\right) \Rightarrow N(0,\sigma^2), \tag{42}$$

$$n[d_A(\hat{h}) - d_A(\hat{h}_0)] \Rightarrow C_1\chi_1^2, \tag{43}$$

where σ^2 and C_1 are both positive. The above methods are all asymptotically equivalent at this higher order of approximation. Another interesting result is that the estimated $\hat{h}$ and optimum $\hat{h}_0$ 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 (42). Most of this work is in density estimation – see the recent review of Jones et al. (1992) for references. In this case, various $\sqrt{n}$ 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 et al. (1992a) construct a $\sqrt{n}$ consistent bandwidth selector for regression based on a bias reduction technique.

5. 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. Considerable resources are devoted to providing forecasts of macroeconomic entities such as GNP, unemployment and inflation, while the benefits of predicting asset prices are obvious. In many cases linear models have been the basis of econometric prediction, while more recently nonlinear models such as ARCH have become popular. Nonparametric methods can also be applied in this context, and provide a model free basis of predicting future outcomes. 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), see also Phillips (1991) and Robinson (1991).

Suppose that we observe the vector time series $\{Z_i\}_{i=1}^n$, where $Z_i = (Y_i, X_i)$, and X_i is strictly exogenous in the sense of Engle et al. (1983). It is convenient to assume

that the process is stationary and mixing is as defined in Gallant and White (1988), which includes most linear processes, for example, 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, we want to predict Y_i from X_i. This problem we call regression with correlated errors.

5.1. *Autoregression*

For convenience we restrict our 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

$$M_k(y) = E(Y_{i+k} | Y_i = y). \tag{44}$$

More generally, one may wish to estimate the conditional variance of Y_{i+k} from lagged values,

$$V_k(y) = \mathrm{Var}(Y_{i+k} | Y_i = y).$$

One can also estimate 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.

Diebold and Nason (1990), Meese and Rose (1991) and Mizrach (1992) estimate $M(\cdot)$ for use in predicting asset prices over short horizons. In each case, a locally weighted regression estimator was employed with a nearest neighbor type window, while bandwidth was chosen subjectively (except in Mizrach (1992) where cross-validation was used). Not surprisingly, their published results concluded that there was little gain in predictive accuracy over a simple random walk. Pagan and Hong (1991), Pagan and Schwert (1990) and Pagan and Ullah (1988) estimate $V(\cdot)$ in order to evaluate the risk premium of asset returns. They used a variety of nonparametric methods including Fourier series and kernels. Their focus was on estimation rather than prediction, and their procedures relied on some parametric estimation. See also Whistler (1988) and Gallant et al. (1991).

A scientific basis can also be found for choosing bandwidth in this sampling scheme. Härdle and Vieu (1991) showed that crossvalidation also works in the autoregression problem – "choose" $\hat{h} = \arg\min \mathrm{CV}(h)$ gives asymptotically optimal estimates.

To illustrate this result we simulated an autoregressive process $Y_i = M(Y_{i-1}) + \varepsilon_i$ with

$$M(y) = y \exp(-y^2), \tag{45}$$

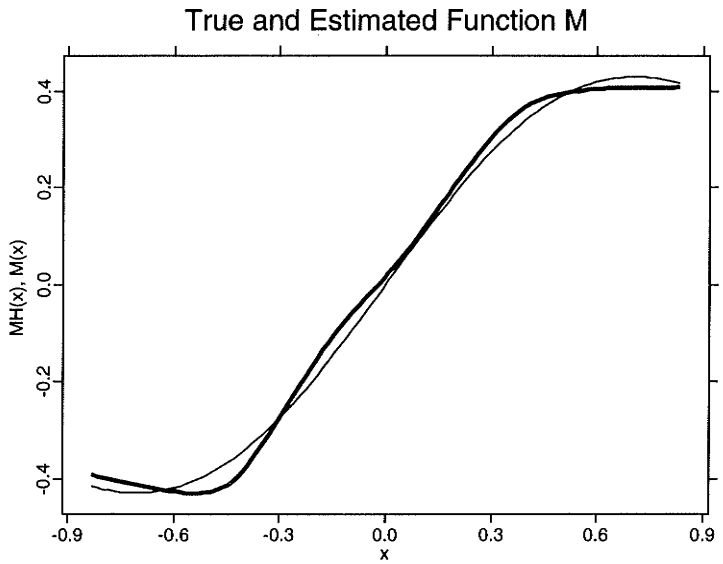


Figure 8. The time regression function $M(y) = y\exp(-y^2)$ for the simulated example (thin line) and the kernel smoother (thick line).

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 in (3) was used. The minimum of $\mathrm{CV}(h)$ was $\hat{h} = 0.43$, while the maximum of $d_A(h)$ was at $h = 0.52$. The curve of $d_A(h)$ was very flat for this example, since there was very little bias present. In Figure 8 we compare the estimated curve with the autoregression function and find good agreement.

5.2. *Correlated errors*

We now consider the regression model

$$Y_i = m(X_i) + \varepsilon_i,$$

where X_i is fixed in repeated samples and the errors ε_i satisfy $E(\varepsilon_i/X_i) = 0$, but are autocorrelated. The kernel estimator $\hat{m}_h(x)$ of $m(x)$ is consistent under quite general conditions. In fact, its bias is the same as when the ε_i are independent. However, the variance is generally affected by the dependency structure. Suppose that the error

process is MA(1), i.e.

$$\varepsilon_i = u_i + \theta u_{i-1},$$

where u_i are iid with zero mean and variance σ^2. In this case,

$$\operatorname{Var}[\hat{m}_h(x)] = \sigma^2\left[(1+\theta^2)\sum_{i=1}^{n} W_{ni}^2 + 2\theta\sum_{i=1}^{n-1} W_{ni}W_{ni+1}\right] \tag{46}$$

which is $O(n^{-1}h^{-1})$, but differs from Theorem 2. If the explanatory variable were time itself (i.e. $X_i = i/n, i = 1,\ldots,n$), then a further approximation is possible:

$$\operatorname{Var}[\hat{m}_h(x)] \approx \frac{1}{nh}\sigma^2(1+\theta^2+2\theta)\nu_2(K).$$

Hart and Wehrly (1986) develop MSE approximations in a regression model in which the error correlation is a general function $\rho(\cdot)$ of the time between observations.

Unfortunately, crossvalidation fails in this case. Suppose that the errors are AR(1) with autoregression parameter close to one. The effect on the crossvalidation technique described in Section 4 must be drastic. The error process stays a long time 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 and 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 also sometimes appealing in an independent setting. See the discussion of Härdle et al. (1988) and Hart and Vieu (1990). It may also be possible to correct for this effect by "whitening" the residuals in (40), although this has yet to be shown.

6. Applications to semiparametric estimation

Semiparametric models offer a compromise between parametric modeling and the nonparametric approaches we have discussed. When data are high dimensional or if it is necessary to account for both functional form and correlation of a general nature, fully nonparametric methods may not perform well. In this case, semiparametric models may be preferred.

By a semiparametric model we mean that the density of the observable data, conditional on any ancillary information, is completely specified by a finite

dimensional parameter θ and an unknown function $G(\cdot)$. The exhaustive monograph of Bickel et al. (1992) develops a comprehensive theory of inference for a large number of semiparametric models, although mostly within iid sampling. There are a number of reviews for econometricians including Robinson (1988b), Newey (1990) and Powell (this volume).

In many cases, θ is of primary interest. Andrews (1989) provides asymptotic theory for a general procedure designed to estimate θ when a preliminary estimate $\hat{G}$ of G is available. The method involves substituting $\hat{G}$ for G in an estimating equation derived, perhaps, from a likelihood function. Typically, the dependence of the estimated parameters $\hat{\theta}$ on the nonparametric estimators disappears asymptotically, and

$$\sqrt{n}(\hat{\theta} - \theta) \Rightarrow N(0, \Omega_0), \tag{47}$$

where $\Omega_0 > 0$.

Nevertheless, the small sample properties of $\hat{\theta}$ can depend quite closely on the way in which this preliminary step is carried out – see the Monte Carlo evidence contained in Engle and Gardiner (1976), Hsieh and Manski (1987), Stock (1989) and Delgado (1992). Some recent work has investigated analytically the small sample properties of semiparametric estimators. Carroll and Härdle (1989), Cavanagh (1989), Härdle et al. (1992b), Linton (1991, 1992, 1993) and Powell and Stoker (1991) develop asymptotic expansions of the form

$$\mathrm{MSE}[\sqrt{n}(\hat{\theta} - \theta)] \approx \Omega_0 + \frac{\Omega_1}{q_1(n,h)} + \frac{\Omega_2}{q_2(n,h)}, \tag{48}$$

where q_1 and q_2 both increase with n under restrictions on $h(n)$. These expansions yield a formula for the optimal bandwidth similar to (36). An important finding is that different amounts of smoothing are required for $\hat{\theta}$ and for $\hat{G}$; in particular, it is often optimal to undersmooth $\hat{G}$ (by an order of magnitude) when the properties of $\hat{\theta}$ are at stake.

The MSE expansions can be used to define a plug-in method of bandwidth choice for $\hat{\theta}$ that is based on second order optimality considerations.

6.1. *The partially linear model*

Consider

$$Y_i = \beta^T X_i + \phi(Z_i) + \varepsilon_i; \qquad X_i = g(Z_i) + \eta_i, \qquad i = 1, 2, \ldots, n \tag{49}$$

where $\phi(\cdot)$ and $g(\cdot)$ are of unknown functional form, while $E(\varepsilon_i | Z_i) = E(\eta_i | Z_i) = 0$. If an inappropriate parametric model is fit to $\phi(\cdot)$, the resulting MLE of β may be

inconsistent. This necessitates using nonparametric methods that allow a more general functional form, when it is needed. Engle et al. (1986) uses this model to estimate the effects of temperature on electricity demand, while Stock (1991) models the effect of the proximity of toxic waste on house prices. In both cases, the effect is highly nonlinear as the large number of covariates make a fully nonparametric analysis infeasible. See also Olley and Pakes (1991). This specification also arises from various sample selection models. See Ahn and Powell (1990) and Newey et al. (1990).

Notice that

$$Y_i - E(Y_i|Z_i) = \beta^T[X_i - E(X_i|Z_i)] + \varepsilon_i.$$

Robinson (1988a) constructed a semiparametric estimator of β replacing $g(Z_i) = E(X_i/Z_i)$ and $m(Z_i) = E(Y_i/Z_i)$ by nonparametric kernel estimators $\hat{g}_h(Z_i)$ and $\hat{m}_h(Z_i)$ and then letting

$$\hat{\beta} = \left[\sum_{i=1}^{n} \{X_i - \hat{g}_h(Z_i)\}\{X_i - \hat{g}_h(Z_i)\}^T\right]^{-1} \sum_{i=1}^{n} [X_i - \hat{g}_h(Z_i)][Y_i - \hat{m}_h(Z_i)].$$

In fact, Robinson modified this estimator by trimming out observations for which the marginal density of Z was small. Robinson's estimator satisfies (47), provided the dimensions of Z are not too high relative to the order of the kernel being used (provided m and g are sufficiently smooth).

Linton (1992) establishes that the optimal bandwidth for $\hat{\beta}$ is $O(n^{-2/9})$, when Z is scalar, and the resulting correction to the (asymptotic) MSE of the standardized estimator is $O(n^{-7/9})$.

6.2. *Heteroskedastic nonlinear regression*

Consider the following nonlinear regression model:

$$Y_i = \tau(X_i; \beta) + \varepsilon_i, \qquad i = 1, 2, \ldots, n, \tag{50}$$

where $\tau(\cdot; \beta)$ is known, while $E(\varepsilon_i|X_i) = 0$ and $\mathrm{Var}(\varepsilon_i|X_i) = \sigma^2(X_i)$, where $\sigma^2(\cdot)$ is of unknown functional form. Efficient estimation of β can be carried out using the *pseudo-likelihood* principle. Assuming that the ε_i are iid and normally distributed, the sample log-likelihood function is proportional to

$$\mathscr{L}[\beta; \sigma^2(\cdot)] = \sum_{i=1}^{n} [Y_i - \tau(X_i; \beta)]^2 \sigma^2(X_i)^{-1}, \tag{51}$$

where $\sigma^2(\cdot)$ is known. In the semiparametric situation we replace $\sigma^2(X_i)$ by a nonparametric estimator $\hat{\sigma}^2(X_i)$, and then let $\hat{\beta}$ minimize $\mathscr{L}[\beta; \hat{\sigma}^2(\cdot)]$.

Carroll (1982) and Robinson (1987) examine the situation where $\tau(X;\beta) = \beta^T X$ in which case

$$\hat{\beta} = \left[\sum_{i=1}^{n} X_i X_i^T \hat{\sigma}^2(X_i)^{-1} \right]^{-1} \sum_{i=1}^{n} X_i Y_i \hat{\sigma}^2(X_i)^{-1}. \tag{52}$$

They establish (under iid sampling) that $\hat{\beta}$ is asymptotically equivalent to the infeasible GLS estimator based on (51). Remarkably, Robinson allows X to have unbounded support, yet did not need to trim out contributions from its tails: he used nearest neighbor estimators of $\sigma^2(\cdot)$ that always average over the same number of observations. Extensions of this model to the multivariate nonlinear $\tau(\cdot;\beta)$ case are considered in Delgado (1992), while Hidalgo (1992) allows both heteroskedasticity and serial correlation of unknown form. Applications include Melenberg and van Soest (1991), Altug and Miller (1992) and Whistler (1988).

Carroll and Härdle (1989), Cavanagh (1989) and Linton (1993) develop second order theory for these estimators. In this case, the optimal bandwidth is $O(n^{-1/5})$ when X is scalar, making the correction to the (asymptotic) MSE $O(n^{-4/5})$.

6.3. *Single index models*

When the conditional distribution of a scalar variable Y, given the d-dimensional predictor variable X, depends on X only through the index $\beta^T X$, we say that this is a single index model.

One example is the single index regression model in which $E[Y|X = x] = m(x) = g(x^T\beta)$, but no other restrictions are imposed. Define the vector of average derivatives

$$\delta = E[m'(X)] = E[g'(X^T\beta)]\beta, \tag{53}$$

and note that δ determines β up to scale – as shown by Stoker (1986). Let $f(x)$ denote the density of X and l be its vector of the negative log-derivatives (partial), $l = -\partial \log f/\partial x = -f'/f$ (l is also called the *score vector*). Under the assumptions on f given in Powell et al. (1989), we can write

$$\delta = E[m'(X)] = E[l(X)Y], \tag{54}$$

and we estimate δ by $\hat{\delta} = n^{-1}\sum_{i=1}^{n} \hat{l}_H(X_i)Y_i$, where $\hat{l}_H(x) = -\hat{f}'_H/f_H(x)$ is an estimator of $l(x)$ based on a kernel density smoother with bandwidth matrix H. Furthermore, $g(\cdot)$ is estimated by a kernel estimator $\hat{g}_h(\cdot)$ for which $[\hat{\delta}^T X_i]_{i=1}^n$ is the right-hand side data.

Härdle and Stoker (1989) show that

$$\sqrt{n}(\hat{\delta} - \delta) \Rightarrow N(0, \Sigma_\delta),$$

where $\Sigma_\delta = \mathrm{Var}\{l(X)[Y - m(X)] + m'(X)\}$, while $\hat{g}_h$ converges at rate $\sqrt{nh}$ – i.e. like a one dimensional function. Stoker (1991) proposed alternative estimators for δ based on first estimating the partial derivatives $m'(x)$ and then averaging over the observations. A Monte Carlo comparison of these methods is presented in Stoker and Villas-Boas (1992). Härdle et al. (1992b) develop a second order theory for $\hat{\delta}$: in the scalar case, the optimal bandwidth h is $O(n^{-2/7})$ and the resulting correction to the MSE is $O(n^{-1/7})$.

Another example is the *binary choice* model

$$Y_i = I(\beta^T X_i + u_i \geqslant 0), \tag{55}$$

where (X, u) are iid. There are many treatments of this specification following the seminal paper of Manski (1975) – in which a slightly more general specification was considered. We assume also that u is independent of X with unknown distribution function $F(\cdot)$, in which case $\Pr[Y_i = 1 | X_i] = F(\beta^T X_i) = E(Y_i | \beta^T X_i)$, i.e. $F(\cdot)$ is a regression function. In fact, (55) is a special case of (53). Applications include Das (1990), Horowitz (1991), and Melenberg and van Soest (1991).

Klein and Spady (1993) use the *profile likelihood* principle (see also Ichimura and Lee (1991)) to obtain (semiparametric) efficient estimates of β. When F is known, the sample log-likelihood function is

$$\mathscr{L}\{F(\beta)\} = \sum_{i=1}^{n} \{Y_i \ln[F(\beta^T X_i)] + (1 - Y_i)\ln[1 - F(\beta^T X_i)]\}. \tag{56}$$

For given β, let $\hat{F}(\beta^T X)$ be the nonparametric regression estimator of $E(Y | \beta^T X)$. A feasible estimator $\hat{\beta}$ of β is obtained as the minimizer of

$$\mathscr{L}[\hat{F}(\beta)] = \sum_{i=1}^{n} \{Y_i \ln[\hat{F}(\beta^T X_i)] + (1 - Y_i)\ln[1 - \hat{F}(\beta^T X_i)]\}. \tag{57}$$

This can be found using standard numerical optimization techniques. The average derivative estimator can be used to provide initial consistent estimators of β, although it is not in general efficient, see Cosslett (1987). Note that to establish $\sqrt{n}$-consistency, it is necessary to employ bias reduction techniques such as higher order kernels as well as to trim out contributions from sparse regions. Note also that $\hat{\beta}$ is not as efficient as the MLE obtained from (56).

We examined the performance of the average derivative estimator on a simulated dataset, where

$$X \sim N(0, I_2)$$

$$\Pr(Y = 1 | X = x) = \Lambda(\beta^T x) + 0.6\phi'(\beta^T x)$$

$$\beta = (1, 1)^T,$$

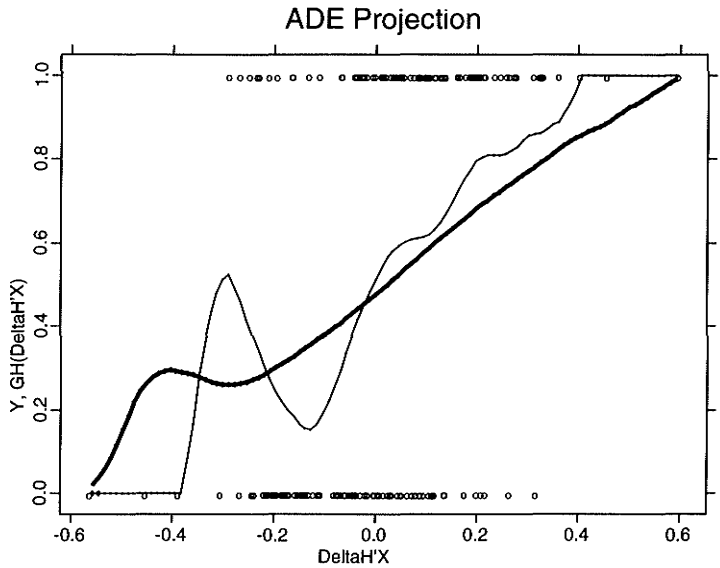


Figure 9. For the simulated dataset $\hat{\delta}^T, X$ versus Y and two estimates of $g(\hat{\delta}^T X_i)$ are shown. The thick line shows the Nadaraya–Watson estimator with a bandwidth $h = 0.3$, while for the thin line $h = 0.1$ was chosen.

while Λ and ϕ are the standard logit and normal density functions respectively. A sample of size $n = 200$ was generated, and the bivariate density function was estimated using a Nadaraya–Watson estimator with bandwidth matrix $H = \text{diag}(0.99, 0.78)$. This example is taken from Härdle and Turlach (1992). The estimation of δ and its asymptotic covariance matrix $\hat{\Sigma}_\delta$ was done with XploRe macro adefit. For this example $\delta = (0.135, 0.135)^T$, and

$$\hat{\delta} = \begin{pmatrix} 0.124 \\ 0.118 \end{pmatrix}, \qquad \hat{\Sigma}_\delta = \begin{pmatrix} 0.188 & 0.036 \\ 0.036 & 0.206 \end{pmatrix}.$$

Figure 9 shows the estimated regression function $\hat{g}_h(\hat{\delta}^T X_i)$.

These results allow us to test some hypotheses formally using a Wald statistic (see Stoker (1992), pp 53–54). In particular, to test the restriction $R\delta = r_0$, the Wald statistic

$$W = n(R\hat{\delta} - r_0)^T (R\hat{\Sigma}_\delta R^T)^{-1} (R\hat{\delta} - r_0)$$

is compared to a χ^2 (rank R) critical value. Table 3 gives some examples for this technique.

Table 3
Wald statistics for some restrictions on δ.

Restriction	Value W	d.f.	$P[\chi^2(\text{d.f.}) > W]$
$\delta^1 = \delta^2 = 0$	25.25	2	0
$\delta^1 = \delta^2 = 0.135$	0.365	2	0.83
$\delta^1 = \delta^2$	0.027	1	0.869

7. Conclusions

The nonparametric methods we have examined are especially useful when the variable over which the smoothing takes place is one dimensional. In this case, the relationship can be plotted and evaluated, while the estimators converge at rate $\sqrt{nh}$.

For higher dimensions these methods are less attractive due to the slower rate of convergence and the lack of simple but comprehensive graphs. In these cases, there are a number of restricted structures that can be employed including the nonparametric additive models of Hastie and Tibshirani (1990), or semiparametric models like the partially linear and index models examined in Section 6.

References

Abramson, I. (1982) "On bandwidth variation in ,kernel estimates – a square root law", *Annals of Statistics*, 10, 1217–1223.

Ahn, H. and J.L. Powell (1990) "Estimation of Censored Selection Models with a Nonparametric Selection Mechanism", Unpublished Manuscript, University of Wisconsin.

Akaike, H. (1970) "Statistical predictor information", *Annals of the Institute of Statistical Mathematics*, 22, 203–17.

Akaike, H. (1974) "A new look at the statistical model identification", *IEEE Transactions of Automatic Control, AC* 19, 716–23.

Altug, S. and R.A. Miller (1992) "Human capital, aggregate shocks and panel data estimation", Unpublished manuscript, University of Minnesota.

Anand, S., C.J. Harris and O. Linton (1993) "On the concept of ultrapoverty", Harvard Center for Population Studies Working paper, 93–02.

Andrews, D.W.K. (1989) "Semiparametric Econometric Models: I Estimation", Cowles Foundation Discussion paper 908.

Andrews, D.W.K. (1991) "Asymptotic Normality of Series Estimators for Nonparametric and Semiparametric Regression Models", *Econometrica*, 59, 307–346.

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.

Ansley, C.F., R. Kohn and C. Wong (1993) "Nonparametric spline regression with prior information", *Biometrika*, 80, 75–88.

Banks, J., R. Blundell and A. Lewbel (1993) "Quadratic Engel curves, welfare measurement and consumer demand", Institute for Fiscal Studies, 92–14.

Bhattacharya, P.K. and A.K. Gangopadhyay (1990) "Kernel and Nearest-Neighbor Estimation of a Conditional Quantile", *Annals of Statistics*, 18, 1400–15.

Bickel, P.J., C.A.J. Klaassen, Y. Ritov and J.A. Welner (1992) *Efficient and Adaptive Inference in Semiparametric Models*. Johns Hopkins University Press: Baltimore.

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.
Bierens, H.J. and H.A. Pott-Buter (1990) "Specification of household Engel curves by nonparametric regression", *Econometric Reviews*, 9, 123–184.
Brillinger, D.R. (1980) *Time Series, Data Analysis and Theory*, Holden–Day.
Carroll, R.J. (1982) "Adapting for Heteroscedasticity in Linear Models", *Annals of Statistics*, 10, 1224–1233.
Carroll, R.J. and W. Härdle (1989) "Second Order Effects in Semiparametric Weighted Least Squares Regression", *Statistics*, 20, 179–186.
Cavanagh, C.L. (1989) "The cost of adapting for heteroskedasticity in linear models", Unpublished manuscript, Harvard University.
Chambers, J.M., W.S. Cleveland, B. Kleiner and P.A. Tukey (1983) *Graphical Methods for Data Analysis*. Duxburry Press.
Chaudhuri, P. (1991) "Global nonparametric estimation of conditional quantile functions and their derivatives", *Journal of Multivariate Analysis*, 39, 246–269.
Cleveland, W.S. (1979) "Robust Locally Weighted Regression and Smoothing Scatterplots", *Journal of the American Statistical Association*, 74, 829–836.
Cosslett, S.R. (1987) "Efficiency bounds for Distribution-free estimators of the Binary Choice and the Censored Regression model", *Econometrica*, 55, 559–587.
Cox, D.R. and D.V. Hinkley (1974) *Theoretical Statistics*. Chapman and Hall.
Craven, P. and Wahba, G. (1979) "Smoothing noisy data with spline functions", *Numer. Math.*, 31, 377–403.
Daniell, P.J. (1946) "Discussion of paper by M.S. Bartlett", *Journal of the Royal Statistical Society Supplement*, 8:27.
Das, S. (1990) "A Semiparametric Structural Analysis of the Idling of Cement Kilns", *Journal of Econometrics*, 50, 235–256.
Deaton, A.S. (1991) "Rice-prices and income distribution in Thailand: a nonparametric analysis", *Economic Journal*, 99, 1–37.
Deaton, A.S. (1993) "Data and econometric tools for development economics", *The Handbook of Development Economics*, Volume III, Eds J. Behrman and T.N. Srinavasan.
Delgado, M. (1992) "Semiparametric Generalised Least Squares in the Multivariate Nonlinear Regression Model", *Econometric Theory*, 8, 203–222.
Diebold, F., and J. Nason (1990) "Nonparametric exchange rate prediction?", *Journal of International Economics*, 28, 315–332.
Doukhan, P. and Ghindès, M. (1980) "Estimation dans le processus $X_n = f(X_{n-1}) + \varepsilon_n$", *Comptes Rendus, Académie des Sciences de Paris*, 297, Série A, 61–4.
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.
Engle, R.F. and R. Gardiner (1976) "Some Finite Sample Properties of Spectral Estimators of a Linear Regression", *Econometrica*, 44, 149–165.
Engle, R.F., D.F. Hendry and J.F. Richard (1983) "Exogeneity", *Econometrica*, 51, 277–304.
Engle, R.F., C.W.J. Granger, J. Rice and A. Weiss (1986) "Semiparametric Estimates of the Relationship Between Weather and Electricity Sales", *Journal of the American Statistical Association*, 81, 310–320.
Eubank, R.L. (1988) *Smoothing Splines and Nonparametric Regression*. Marcel Dekker.
Fama, E.F. (1965) "The behavior of stock prices", *Journal of Business*, 38, 34–105.
Family Expenditure Survey, Annual Base Tapes (1968–1983). Department of Employment, Statistics Division, Her Majesty's Stationary Office, London, 1968–1983.
Fan, J. (1992) "Design-Adaptive Nonparametric Regression", *Journal of the American Statistical Association*, 87, 998–1004.
Fan, J. and I. Gijbels (1992) "Spatial and Design Adaptation: Variable order approximation in function estimation", *Institute of Statistics Mimeo Series*, no 2080, University of North Carolina at Chapel Hill.
Fan, J., N.E. Heckman and M.P. Wand (1992) "Local Polynomial Kernel Regression for Generalized Linear Models and Quasi-Likelihood Functions", University of British Columbia Working paper 92–028.
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.

Gallant, A.R. and G. Souza (1991) "On the asymptotic normality of Fourier flexible form estimates", *Journal of Econometrics*, 50, 329–353.
Gallant, A.R. and H. White (1988) *A Unified Theory of Estimation and Inference for Nonlinear Dynamic Models*. Blackwell: Oxford.
Gallant, A.R., D.A. Hsieh and G.E. Tauchen (1991) "On Fitting a Recalcitrant Series: The Pound/Dollar Exchange Rate, 1974–1983", in *Nonparametric and Semiparametric Methods in Econometrics and Statistics*. Eds Barnett, Powell, and Tauchen. Cambridge University Press.
Gasser, T. and H.G. Müller (1984) "Estimating regression functions and their derivatives by the kernel method", *Scandinavian Journal of Statistics*, 11, 171–85.
Gasser, T., H.G. Müller and V. Mammitzsch (1985) "Kernels for nonparametric curve estimation", *Journal of the Royal Statistical Society Series B*, 47, 238–52.
Gozalo, P.L. (1989) "Nonparametric analysis of Engel curves: estimation and testing of demographic effects", Brown University, Department of Economics Working paper 92–15.
Györfi, L., W. Härdle, P. Sarda and P. Vieu (1990) *Nonparametric Curve Estimation from Time Series*. Lecture Notes in Statistics, 60. Springer–Verlag: Heidelberg, New York.
Hall, P. (1992) *The Bootstrap and Edgeworth Expansion*. Springer–Verlag: New York.
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.
Hall, P. and I. Johnstone (1992) "Empirical functional and efficient smoothing parameter selection", *Journal of the Royal Statistical Society Series B*, 54, 475–530.
Härdle, W. (1990) *Applied Nonparametric Regression*. Econometric Society Monographs 19, Cambridge University Press.
Härdle, W. (1991) *Smoothing Techniques with Implementation*. Springer-Verlag: Heidelberg, New York, Berlin.
Härdle, W. and R.J. Carroll (1990) "Biased cross-validation for a kernel regression estimator and its derivatives", *Österreichische Zeitschrift für Statistik und Informatik*, 20, 53–64.
Härdle, W. and M. Jerison (1991) "Cross Section Engel Curves over Time", *Recherches Economiques de Louvain*, 57, 391–431.
Härdle, W. and J.S. Marron (1985) "Optimal bandwidth selection in nonparametric regression function estimation", *Annals of Statistics*, 13, 1465–81.
Härdle, W. and M. Müller (1993) "Nichtparametrische Glättungsmethoden in der alltäglichen statistischen Praxis", *Allgemeines Statistiches Archiv*, 77, 9–31.
Härdle, W. and D.W. Scott (1992) "Smoothing in Low and High Dimensions by Weighted Averaging Using Rounded Points", *Computational Statistics*, 1, 97–128.
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.
Härdle, W. and B.A. Turlach (1992) "Nonparametric Approaches to Generalized Linear Models", In: Fahrmeir, L., Francis, B., Gilchrist, R., Tutz, G. (Eds.) *Advances in GLIM and Statistical Modelling*, Lecture Notes in Statistics, 78. Springer–Verlag: New York.
Härdle, W. and P. Vieu (1991) "Kernel regression smoothing of time series", *Journal of Time Series Analysis*, 13, 209–232.
Härdle, W., P. Hall and J.S. Marron (1988) "How far are automatically chosen regression smoothing parameters from their optimum?", *Journal of the American Statistical Association*, 83, 86–99.
Härdle, W., P. Hall and J.S. Marron (1992a) "Regression smoothing parameters that are not far from their optimum", *Journal of the American Statistical Association*, 87, 227–233.
Härdle, W., J. Hart, J.S. Marron and A.B. Tsybakov (1992b) "Bandwidth Choice for Average Derivative Estimation", *Journal of the American Statistical Association*, 87, 218–226.
Härdle, W., P. Hall and H. Ichimura (1993) "Optical Smoothing in Single Index Models", *Annals of Statistics*, 21, to appear.
Hart, J. and P. Vieu (1990) "Data-driven bandwidth choice for density estimation based on dependent data", *Annals of Statistics*, 18, 873–890.
Hart, D. and T.E. Wehrly (1986) "Kernel regression estimation using repeated measurements data", *Journal of the American Statistical Association*, 81, 1080–8.
Hastie, T.J. and R.J. Tibshirani (1990) *Generalized Additive Models*. Chapman and Hall.
Hausman, J.A. and W.K. Newey (1992) "Nonparametric estimation of exact consumer surplus and deadweight loss", MIT, Department of Economics Working paper 93–2, Massachusetts.

Hidalgo, J. (1992) "Adaptive Estimation in Time Series Models with Heteroscedasticity of Unknown Form", *Econometric Theory*, 8, 161–187.
Hildenbrand, K. and W. Hildenbrand (1986) "On the mean income effect: a data analysis of the U.K. family expenditure survey", in *Contributions to Mathematical Economics*, ed W. Hildenbrand and A. Mas-Colell. North-Holland: Amsterdam.
Hildenbrand, W. and A. Kneip (1992) "Family expenditure data, heteroscedasticity and the law of demand", Universität Bonn Discussion paper A-390.
Horowitz, J.L. (1991) "Semiparametric estimation of a work-trip mode choice model", University of Iowa Department of Economics Working paper 91–12.
Hsieh, D.A. and C.F. Manski (1987) "Monte Carlo Evidence on Adaptive Maximum Likelihood Estimation of a Regression", *Annals of Statistics*, 15, 541–551.
Hussey, R. (1992) "Nonparametric evidence on asymmetry in business cycles using aggregate employment time series", *Journal of Econometrics*, 51, 217–231.
Ichimura, H. and L.F. Lee (1991) "Semiparametric Least Squares Estimation of Multiple Index Models: Single Equation Estimation", in *Nonparametric and Semiparametric Methods in Econometrics and Statistics*. Eds Barnett, Powell, and Tauchen. Cambridge University Press.
Jones, M.C. (1985) "Discussion of the paper by B.W. Silverman", *Journal of the Royal Statistical Society Series B*, 47, 25–26.
Jones, M.C. (1989) "Discretized and interpolated Kernel Density Estimates", *Journal of the American Statistical Association*, 84, 733–741.
Jones, M.C. and P.J. Foster (1993) "Generalized jacknifing and higher order kernels", Forthcoming in *Journal of Nonparametric Statistics*.
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.
Jones, M.S., O. Linton and J.P. Nielsen (1993) "A multiplicative bias reduction method", Preprint, Nuffield College, Oxford.
Klein, R.W. and R.H. Spady (1993) "An Efficient Semiparametric Estimator for Binary Choice Models", *Econometrica*, 61, 387–421.
Koenker, R. and G. Bassett (1978) "Regression quantiles", *Econometrica*, 46, 33–50.
Koenker, R., P. Ng and S. Portnoy (1993) "Quantile Smoothing Splines", Forthcoming in *Biometrika*.
Lewbel, A. (1991) "The Rank of Demand Systems: Theory and Nonparametric Estimation", *Econometrica*, 59, 711–730.
Li, K.-C. (1985) "From Stein's unbiased risk estimates to the method of generalized cross-validation", *Annals of Statistics*, 13, 1352–77.
Linton, O.B. (1991) "Edgeworth Approximation in Semiparametric Regression Models", PhD thesis, Department of Economics, UC Berkeley.
Linton, O.B. (1992) "Second Order Approximation in the Partially Linear Model", Cowles Foundation Discussion Paper no 1065.
Linton, O.B. (1993) "Second Order Approximation in a linear regression with heteroskedasticity of unknown form", Nuffield College Discussion paper no 75.
Linton, O.B. and J.P. Nielsen (1993) "A Multiplicative Bias Reduction Method for Nonparametric Regression", Forthcoming in *Statistics and Probability Letters*.
McFadden, D. (1985) "Specification of econometric models", Econometric Society, Presidential Address.
Mack, Y.P. (1981) "Local properties of *k-NN* regression estimates", *SIAM J. Alg. Disc. Meth.*, 2, 311–23.
Mandelbrot, B. (1963) "The variation of certain speculative prices", *Journal of Business*, 36, 394–419.
Manski, C.F. (1975) "Maximum Score Estimation of the Stochastic Utility Model of Choice", *Journal of Econometrics*, 3, 205–228.
Marron, J.S. and D. Nolan (1989) "Canonical kernels for density estimation", *Statistics and Probability Letters*, 7, 191–195.
Marron, J.S. and M.P. Wand (1992) "Exact Mean Integrated Squared Error", *Annals of Statistics*, 20, 712–736.
Meese, R.A. and A.K. Rose (1991) "An empirical assessment of nonlinearities in models of exchange rate determination", *Review of Economic Studies*, 80, 603–619.
Melenberg, B. and A. van Soest (1991) "Parametric and semi-parametric modelling of vacation expenditures", CentER for Economic Research, Discussion paper no 9144, Tilburg, Holland.

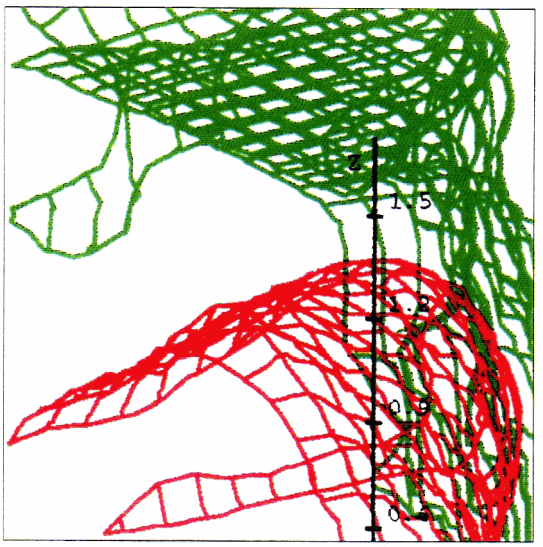
Mizrach, B. (1992) "Multivariate nearest-neighbor forecasts of EMS exchange rates", *Journal of Applied Econometrics*, 7, 151–163.
Müller, H.G. (1987) "On the asymptotic mean square error of L_1 kernel estimates of C_∞ functions", *Journal of Approximation Theory*, 51, 193–201.
Müller, H.G. (1988) *Nonparametric Regression Analysis of Longitudinal Data.* Lecture Notes in Statistics, Vol. 46. Springer–Verlag: Heidelberg/New York.
Nadaraya, E.A. (1964) "On estimating regression", *Theory of Probability and its Applications*, 10, 186–190.
Newey, W.K. (1990) "Semiparametric Efficiency Bounds", *Journal of Applied Econometrics*, 5, 99–135.
Newey, W.K., J.L. Powell and J.R. Walker (1990) "Semiparametric Estimation of Selection Models: Some Empirical Results", *American Economic Review Papers and Proceedings*, 80, 324–328.
Olley, G.S. and A. Pakes (1991) "The Dynamics of Productivity in the Telecommunications Equipment Industry", Unpublished manuscript, Yale University.
Pagan, A.R. and Y.S. Hong (1991) "Nonparametric Estimation and the Risk Premium", in *Nonparametric and Semiparametric Methods in Econometrics and Statistics.* Eds Barnett, Powell, and Tauchen. Cambridge University Press.
Pagan, A.R. and W. Schwert (1990) "Alternative models for conditional stock volatility", *Journal of Econometrics*, 45, 267–290.
Pagan, A.R. and A. Ullah (1988) "The econometric analysis of models with risk terms", *Journal of Applied Econometrics*, 3, 87–105.
Park, B.U. and B.A. Turlach (1992) "Practical performance of several data-driven bandwidth selectors (with discussion)", *Computational Statistics*, 7, 251–271.
Phillips, P.C.B. (1991) "Spectral Regression for Cointegrated Time Series" in *Nonparametric and Semiparametric Methods in Econometrics and Statistics.* Eds Barnett, Powell, and Tauchen. Cambridge University Press.
Powell, J.L. and T.M. Stoker (1991) "Optimal Bandwidth Choice for Density-Weighted Averages", Unpublished manuscript, Princeton University.
Powell, J.L., J.H. Stock and T.M. Stoker (1989) "Semiparametric Estimation of Index Coefficients", *Econometrica*, 57, 1403–1430.
Prakasa Rao, B.L.S. (1983) *Nonparametric Functional Estimation.* Academic Press.
Rice, J.A. (1984) "Bandwidth choice for nonparametric regression", *Annals of Statistics*, 12, 1215–30.
Robb, A.L., L. Magee and J.B. Burbidge (1992) "Kernel smoothed consumption-age quantiles", *Canadian Journal of Economics*, 25, 669–680.
Robinson, P.M. (1983) "Nonparametric Estimators for Time Series", *Journal of Time Series Analysis*, 4, 185–208.
Robinson, P.M. (1987) "Asymptotically Efficient Estimation in the Presence of Heteroscedasticity of Unknown Form", *Econometrica*, 56, 875–891.
Robinson, P.M. (1988a) "Root-N-Consistent Semiparametric Regression", *Econometrica*, 56, 931–954.
Robinson, P.M. (1988b) "Semiparametric Econometrics: A Survey", *Journal of Applied Econometrics*, 3, 35–51.
Robinson, P.M. (1991) "Automatic Frequency Domain Inference on Semiparametric and Nonparametric Models", *Econometrica*, 59, 1329–1364.
Rosenblatt, M. (1956) "Remarks on some nonparametric estimates of a density function", *Annals of Mathematical Statistics*, 27, 642–669.
Ruppert, D. and M.P. Wand (1992) "Multivariate Locally Weighted Least Squares Regression", Rice University, Technical Report no. 92–4.
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.
Sentana, E. and S. Wadhwani (1991) "Semi-parametric Estimation and the Predictability of Stock Returns: Some Lessons from Japan", *Review of Economic Studies*, 58, 547–563.
Shibata, R. (1981) "An optimal selection of regression variables", *Biometrika*, 68, 45–54.
Silverman, B.W. (1984) "Spline smoothing: the equivalent variable kernel method", *Annals of Statistics*, 12, 898–916.
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.
Silverman, B.W. (1986). *Density estimation for statistics and data analysis.* Chapman and Hall: London.
Stock, J.H. (1989) "Nonparametric Policy Analysis", *Journal of the American Statistical Association*, 84, 567–576.

Stock, J.H. (1991) "Nonparametric Policy Analysis: An Application to Estimating Hazardous Waste Cleanup Benefits", in *Nonparametric and Semiparametric Methods in Econometrics and Statistics*. Eds Barnett, Powell, and Tauchen. Cambridge University Press.
Stoker, T.M. (1986) "Consistent Estimation of Scaled Coefficients", *Econometrica*, 54, 1461–1481.
Stoker, T.M. (1991) "Equivalence of direct, indirect, and slope estimators of average derivatives", in *Nonparametric and Semiparametric Methods in Econometrics and Statistics*. Eds Barnett, Powell, and Tauchen. Cambridge University Press.
Stoker, T.M. (1992) *Lectures on Semiparametric Econometrics*. CORE Lecture Series. Université Catholique de Louvain, Belgium.
Stoker, T.M. and J.M. Villas-Boas (1992) "Monte Carlo Simulation of Average Derivative Estimators", Unpublished manuscript, MIT: Massachusetts.
Stone, C.J. (1982) "Optical global rates of convergence for nonparametric regression", *Annals of Statistics*, 10, 1040–1053.
Strauss, J. and D. Thomas (1990) "The shape of the calorie-expenditure curve", Unpublished manuscript, Rand Corporation, Santa Monica.
Stute, W. (1986) "Conditional Empirical Processes", *Annals of Statistics*, 14, 638–647.
Tibshirani, R. (1984) "Local likelihood estimation", PhD Thesis, Stanford University, California.
Tikhonov, A.N. (1963) "Regularization of incorrectly posed problems", *Soviet Math.*, 4, 1624–1627.
Turlach, B.A. (1992) "On discretization methods for average derivative estimation", CORE Discussion Paper no. 9232, Université Catholique de Louvain, Louvain–la–Neuve, Belgium.
Vapnik, V. (1982). *Estimation of Dependencies Based on Empirical Data*. Springer–Verlag: Heidelberg, New York, Berlin.
Wahba, G. (1990) *Spline Models for Observational Data*. CBMS-NSF Regional Conference Series in Applied Mathematics, no. 59.
Watson, G.S. (1964) "Smooth regression analysis", *Sankhya Series A*, 26, 359–372.
Whistler, D. (1988) "Semiparametric ARCH Estimation of Intra-Daily Exchange Rate Volatility", Unpublished manuscript, London School of Economics.
Whittaker, E.T. (1923) "On a new method of graduation", *Proc. Edinburgh Math. Soc.*, 41, 63–75.
XploRe (1993) An interactive statistical computing environment. Available from XploRe Systems, Institute für Statistik und Ökonometrie, Wirtschaftswissenschaftliche Fakultät, Humboldt-Universität zu Berlin, D 10178 Berlin, Germany.

Wolfgang Härdle

ÖKONOMIE *dynamisch*

Der Sonderforschungsbereich Quantifikation und Simulation ökonomischer Prozesse



WIRTSCHAFTSWISSENSCHAFTEN

Ökonomische Prozesse sind sich im Wandel befindliche wirtschaftliche Systeme. Die Kenntnis der Quantifizierung und Modellierung einer solchen wirtschaftlichen Dynamik erlaubt etwa die Analyse der Entwicklung des Arbeitsmarktes, der Innovationsfähigkeit verzerrter ökonomischer Strukturen oder die Migration von Arbeitskräften innerhalb Europas. Die Dynamik des Preiswettbewerbs, Entscheidungen über Fusionen und Firmenübernahmen, die Preisbildung auf den Finanzmärkten und die Stabilität der Geldnachfrage sind wichtige Determinanten wirtschaftlicher Prozesse. Die flexible statistische Modellierung solcher Daten und die Erfassung dieser Modellierungsinstrumente in Daten- und Methodenbanken sind Voraussetzungen für eine empirische Analyse ökonomischer Prozesse. Eine solche quantitativ orientierte Analyse kann nur im Dialog mit ökonomischen Konzepten, mit mathematisch-statistischen Methoden und durch computergestützte Simulation durchgeführt werden. Hierzu ist ein verstärkter Einsatz von vernetzten und parallelen Hochleistungsrechnern notwendig.

Der Sonderforschungsbereich ermöglicht diesen Dialog: Die ökonomische Theorie stellt Ideen für meßbare Hypothesen bereit, die quantitativ-statistischen Fächer entwickeln die Werkzeuge zur empirischen Überprüfung, die angewandte Mathematik hilft bei der Bewertung der Ergebnisse und der Entwicklung neuer Methoden.

Die Projekte

Das Gesamtkonzept des Sonderforschungsbereiches als Dialog theoretischer und praktischer Fächer kommt in der Gliederung der drei Projektbereiche zum Ausdruck:
Projektbereich A: Quantitative Verfahren
Projektbereich B: Mathematische Methoden
Projektbereich C: Ökonomische Modellierung.

Im folgenden werden die einzelnen Projekte kurz vorgestellt. Konkrete Zusammenhänge der drei Projektbereiche werden anschließend am Beispiel »Aktienrenditen« und »Frauenbeschäftigung« verdeutlicht.

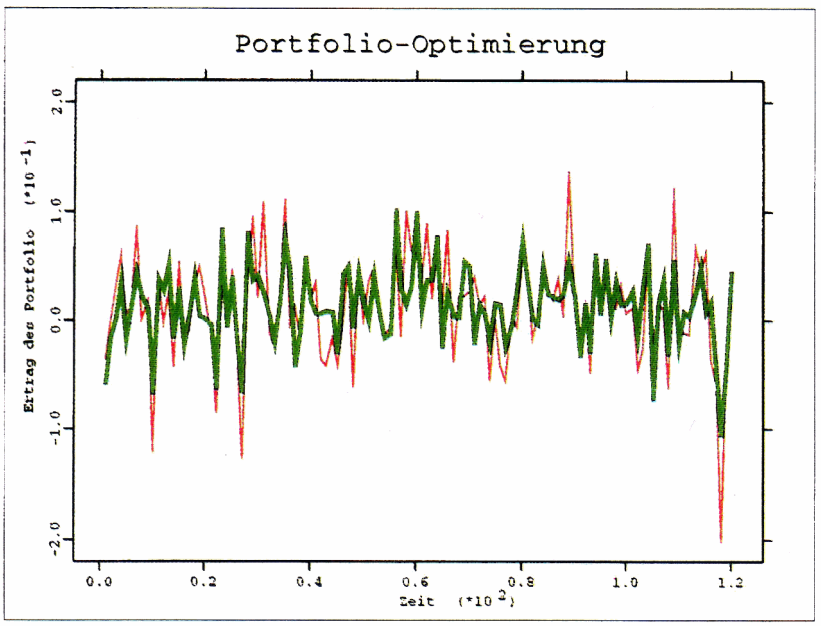


Abb. 1
Beispiel für empirische Kapitalmarktforschung: Die rote Kurve stellt den Kursverlauf eines gleichartig bestückten Portfolios der Notierungen von IBM, PanAm Delta Airlines, Consoldated Edition, Gerber, Texaco von Januar 1978 bis Dezember 1987 dar. Die hohen Ausschläge zeigen erhöhte Risiken an. Die grüne Kurve ist der Kursverlauf eines optimierten Portfolios mit offensichtlich geringeren Kursrisiken.

Wandel auf dem Arbeitsmarkt (Projekt C2)
Arbeitnehmer müssen sich neuen Arbeitsbedingungen anpassen, sei es durch Weiterbildung, Berufswechsel, Migration oder Austritt aus dem Erwerbsleben. Die empirischen Befunde in diesem Bereich deuten bisher darauf hin, daß die Dynamik der entstehenden Arbeitsmärkte Osteuropas schon (nach knapp drei Jahren) mehr und mehr wie die im Westen funktioniert. Der *Matching*-Ansatz (s. Glossar) liefert eine befriedigende Erklärung für diese Dynamik, ist aber theoretisch nicht ausreichend fundiert, vor allem in bezug auf die Trennung von Matches und den Zusammenhang zwischen Arbeitsplätzen und Beschäftigung.

Dynamik des Wettbewerbs und der Preisbildung bei Kapazitätsbeschränkungen (Projekt C4)
Unter simulierten Bedingungen (Laborexperimente am Rechner) werden nichtkooperative Auktionsspiele durchgeführt.

Empirische Kapitalmarktforschung (Projekt C1)
Hier werden klassische Modelle zur erwarteten Rendite von Aktien, wie etwa das Capital Asset Pricing Modell (CAPM), mehr und mehr in Frage gestellt. Die Arbeitsgruppe des Projektes C1 befaßt sich mit der Entwicklungund Überprüfung neuer Hypothesen zur Erklärung der empirisch beobachteten Anomalien (s. dazu Abb. 1 und das Beispiel »Aktienrenditen« auf Seite 20).

Stabilität der Geldnachfragefunktion (Projekt C3)
Mit unterschiedlichen methodischen Ansätzen wird unter anderem die Stabilität der Geldnachfrage in

Deutschland für alternative Geldmengendefinitionen sowie die Modellierung der Geldnachfrage nach der deutschen Wiedervereinigung untersucht.

Mathematische Theorie der Finanzmärkte (Projekt B2)
In diesem Projekt werden in breitem Maße stochastische Prozesse mit stetiger Zeit eingesetzt. Die entsprechenden Modelle, z.B. für Börsenkurse, zeichnen sich dadurch aus, daß man viele der einschneidenden Voraussetzungen, wie Vollständigkeit des Marktes, auflokkert. Im Ergebnis erhält man Existenz- und Eindeutigkeitsaussagen für Preise von sog. contingent claims (s. Glossar), insbesondere Optionen (Termingeschäfte) und Absicherungsstrategien, die Risiken dieser derivativen Instrumente (Aktien, Einlagen etc.) minimieren.

Interoperable Informationssysteme (Projekt A3)
Im Zusammenhang mit Projekt B2 ist der effiziente, integrierte Zugriff auf Daten und Methoden, die an verschiedenen Orten auf unterschiedlichen Rechnerarchitekturen und unter verschiedenen Betriebssystemen abgelegt sein können, von zentraler Bedeutung. In geplanten Forschungsarbeiten werden Überlegungen zur semantischen Systemintegration mit Hilfe von sog. »Metainformation« angestellt, die die Bedeutung eines vorliegenden Datenbestandes beschreibt. Die meisten vorliegenden Ansätze sind jedoch auf die Verwaltung relativ einfach strukturierter Dateien beschränkt; für die Modellverwaltung im Bereich der ökonomischen Analyse sind noch wesentliche Erweiterungen notwendig.

Ein **Sonderforschungsbereich** (SFB)
ist ein langfristig angelegtes Forschungsvorhaben mit mehreren fachübergreifenden Projekten. Seit Beginn dieses Jahres wird an der Humboldt-Universität ein SFB (373) von der Deutschen Forschungsgemeinschaft (DFG) zum Thema »Quantifikation und Simulation Ökonomischer Prozesse« gefördert. Der SFB umfaßt insgesamt 10 Projekte (s. Übersicht auf S. 21), wovon eines an der FU Berlin und eines an der Universität Potsdam durchgeführt wird.
Berlin eignete sich besonders für dieses Forschungsthema, da sich in den letzten Jahren die große Chance ergab, gemeinsame Forschungsinteressen von bislang getrennt voneinander arbeitenden Forschergruppen zu bündeln. Beteiligt sind die wirtschaftswissenschaftliche Fakultät und das Institut für Mathematik der Humboldt-Universität, das Institut für Angewandte Analysis und Stochastik (IAAS), der Fachbereich Wirtschaftswissenschaft der FU Berlin sowie das Institut für Mathematik der Universität Potsdam

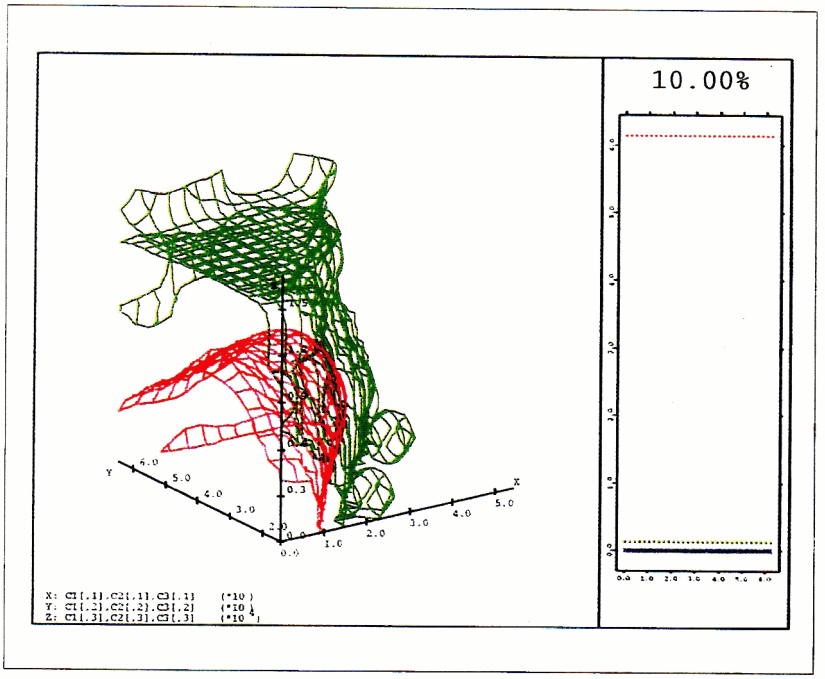


Abb. 2
Beispiel für semiparametrische Modellierung. Zur Interpretation des Diagramms s. Beschreibung des Projekts A1.

Nichtparametrische Zeitreihenanalyse (Projekt A2)
In der nichtparametrischen Zeitreihenanalyse geht man davon aus, daß die vorliegenden Zeitreihen (z.B. Aktienrenditen, Zinssätze, Geldmengenaggregate, Bruttosozialprodukt) von stochastischen Prozessen aus einer sehr allgemeinen Modellklasse generiert werden. Diese Prozesse oder spezifischen Charakteristika werden auf sehr flexible Weise aus den vorliegenden Daten modelliert bzw. geschätzt und beispielsweise für Prognosezwecke benutzt.
Wenngleich mit diesem Verfahren bereits gewisse Erfolge bei der Prognose erzielt worden sind, so bleiben derzeit doch eine Reihe von Problemen, wie z.B. Spezifikationen und Modellauswahl, offen. Insbesondere ist die asymptotische Theorie der hier interessierenden Schätzer und Testwertverfahren noch unzureichend entwickelt.

Semiparametrische Modelle (Projekt A1)
werden in der quantitativen ökonomischen Forschung vielfältig (wie etwa in Marketing und der Arbeitsmarktforschung) angewandt. Das Instrument der semiparametrischen Modellierung erlaubt es, ökonomische Strukturen unter schwachen Annahmen über die Datenverteilung zu verstehen. Der Preis, der für diese (oftmals explorative) Herangehensweise zu zahlen ist, ist die enorme Rechenintensität und die Notwendigkeit hochinteraktiver Graphiksysteme.
Ein Beispiel für die explorative graphische Modellierung ist in Abb. 2 dargestellt. Die grüne und die rote Oberfläche zeigen zwei Oberflächen gleicher Dichte von etwa 20 000 Daten der Studie Qualifikation und Berufsverlauf 85/86 des Zentralinstituts für europäische Sozialforschung. Die Achsen sind X=Betriebszugehörigkeit, Y=Alter und Z=Bruttomonatseinkommen. Die externe gekrümmte Form der Oberfläche zeigt die hohe Nichtlinearität dieses Datensatzes. Die klassische quantitative Theorie würde hier ein etwa trichterförmiges Gebilde annehmen. Sehr deutlich erkennt man jedoch mit dieser interaktiv graphischen Technik das Plateau konstanter Einkommen für einen etwa dreieckigen Bereich der (x, y) Ebene. Offensichtlich sind in die-

Glossar
Contingent Claim: bezeichnet eine zustandsbedingte Zahlung, z.B. wird eine Option (Termingeschäft) fällig, wenn zum Auszahlungszeitpunkt der Kurs des zugrundeliegenden Wertpapiers einen vorher festgelegten, sog. Ausübungspreis überschreitet.
Matching: beschreibt den Prozeß der ökonomischen »Zusammenführung« von Wirtschaftsagenten. Es könnte auf dem Markt auch für das Suchen von Ehepartnern, Immobilien oder Arbeitsverhältnissen verwendet werden. Auf aggregierter oder »makroökonomischer« Ebene bezieht sich Matching auf eine »black-box«, durch die z.B. eine positive Anzahl von Arbeitslosen (Arbeitsuchenden) mit einer positiven Anzahl von offenen Stellen koexistieren kann.

ser Kombination von Betriebszugehörigkeit und Alter keine Einkommensänderungen vorgekommen. Auf Seite 22 geben wir ein Beispiel für semiparametrische Single Index Modelle für die Frage der Beschäftigung von Frauen.

Theorie und Anwendung von Resamplingverfahren (Projekt B1)
Dieser Bereich hat in den letzten Jahren eine rasante Entwicklung genommen, wobei Verfahren für die Schätzung von Charakteristiken von Regressionsschätzungen bei identischen Fehlerverteilungen vorgeschlagen werden. Dabei sind die exakten Eigenschaften der Bootstrapverfahren (s. Glossar) und der zugehörigen adaptiven Schätzungen kaum untersucht, obwohl sie für die bei ökonomischen Anwendungen vorkommenden kleinen und mittleren Beobachtungsumfänge von entscheidendem Interesse sind.

Das Projekt B1 befaßt sich auch mit *Wavelet-Methoden*. Damit können in verbesserter Qualität Zeitreihenphänomene erfaßt und bearbeitet werden, die starkes räumlich oder zeitlich inhomogenes Verhalten aufweisen. Dies findet insbesondere Anwendung bei abrupten Oszillationen oder Sprüngen, wie sie typisch sind für Ökonomien, die sich im Wandel befinden.

Nichtlineare Modelle und Verfahren (Projekt B3)
Dieses Projekt befaßt sich mit der Entwicklung geeigneter Methoden zur Lösung nichtlinearer statistischer Probleme. Zusammen mit dem Projekt C4 *Dynamik des Wettbewerbs* sollen tragfähige Modelle für die Managerkompensation, z.B. als Funktion der Unternehmensgröße, entwickelt werden.

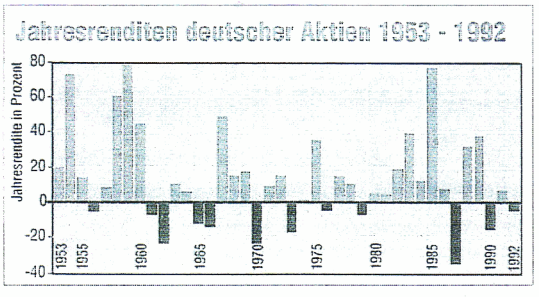
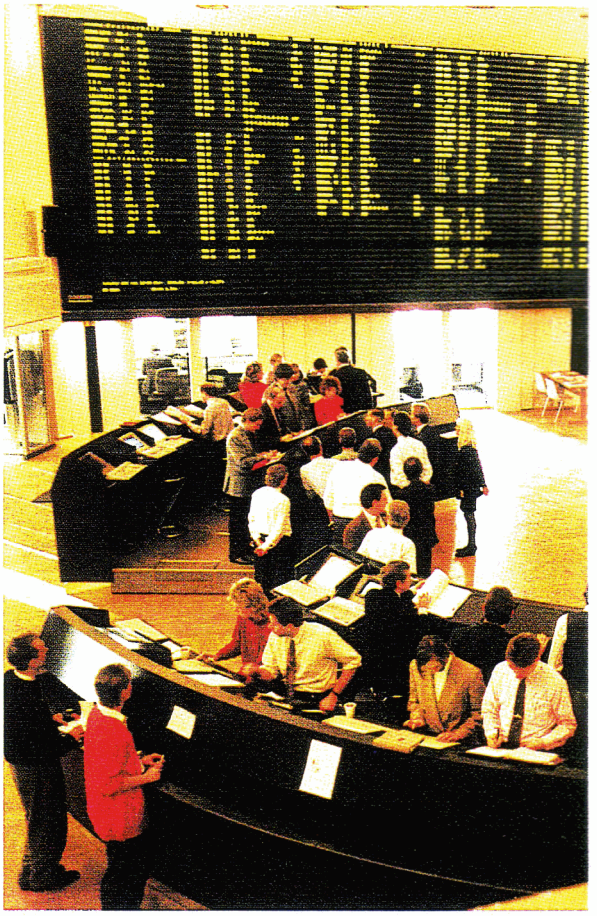


Abb. 3
zeigt die durchschnittliche Rendite der in Frankfurt notierten deutschen Aktien für die Jahre 1953 bis 1992.

Glossar
Bootstrapping: Damit wird versucht, die Eigenschaften von Statistiken (Varianzen von Regressionsschätzern, Verteilungen von Teststatistiken u.ä.) im Simulationsexperiment zu ermitteln. Aus einer anfänglich zu schätzenden Verteilungsfunktion werden dabei durch »Ziehen und Zurücklegen« ständig neue Stichproben zufällig gezogen, die dann zur Ermittlung der interessierenden Statistik verwendet werden. Auf diese Weise können unter bestimmten Bedingungen effizientere Ergebnisse erzielt werden als auf Basis einer nur asymptotisch gültigen stochastischen Theorie.
Resampling: bedeutet das mehrmalige Ziehen aus ein und derselben Stichprobe. Von der Stichprobe (Daten) wird zuerst eine Schätzung (etwa des Risikos eines Portfolios) gebildet. Dann wird die Genauigkeit dieser Schätzung durch Simulation, d.h. mehrmaliges Ziehen einer neuen Stichprobe aus den vorhandenen Daten, untersucht. Dieses Verfahren benötigt hochleistungsfähige Rechner, da sich mit der Anzahl der Simulationen auch die Genauigkeit des Schätzers verbessert.



Beispiel I: Aktienrenditen
Steuerfreie Anleger, die Anfang 1953 insgesamt 10 000 DM in Frankfurter Aktien anlegten, breit streuten, nicht umschichteten (»buy and hold«) und Dividenden, Bezugsrechte und ähnliche Vermögensvorteile reinvestierten, hätten Ende 1992 nach unseren Berechnungen über einen Kapitalbetrag von 777 986 DM verfügen können. Anleger, die Anfang 1953 ihr Kapital auf ein Sparbuch eingezahlt hätten, das für den gesamten Zeitraum mit einem jährlichen Zins von 11,5% ausgestattet gewesen wäre, hätten es auf den gleichen Betrag gebracht. Sparbücher bzw. festverzinsliche Wertpapiere mit einer solchen Verzinsung existieren aber leider nicht.

Monatsgeld hätte im betrachteten Zeitraum im Schnitt 5,6% pro Jahr erbracht. 10 000 DM wären bei einer vollen Reinvestition der Zinszahlungen in den genannten 40 Jahren auf 88 506 DM angestiegen. Mit Bundesanleihen oder ähnlichen langfristigen festverzinslichen Wertpapieren hätte eine Verzinsung von 7,6% pro Jahr erzielt werden können.
Als langfristige Kapitalanlage waren Aktien im genannten Zeitraum deshalb weitaus vorteilhafter als Festverzinsliche. Aktienbesitz ist natürlich mit einem nicht unbeträchtlichen Risiko verbunden (vgl. Abb 3).

Nach dem nun schon klassischen Modell der modernen Kapitalmarkttheorie, dem Sharpe/Lintner Capital Asset Pricing Modell (CAPM), für das Sharpe 1992 den Nobelpreis erhielt, sollte die langfristige Durchschnittsrendite einzelner Aktien auf lineare Weise mit ihrem (nichtdiversifizierbarem) Risiko zusammenhängen. Im vergangenen Jahrzehnt wurden nun eine Reihe

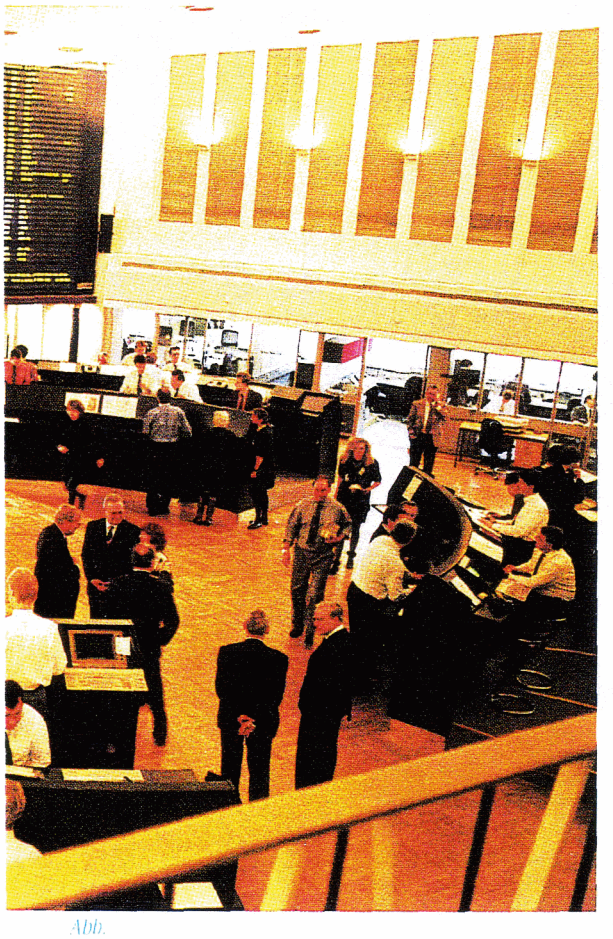


Abb.
Berliner Börse

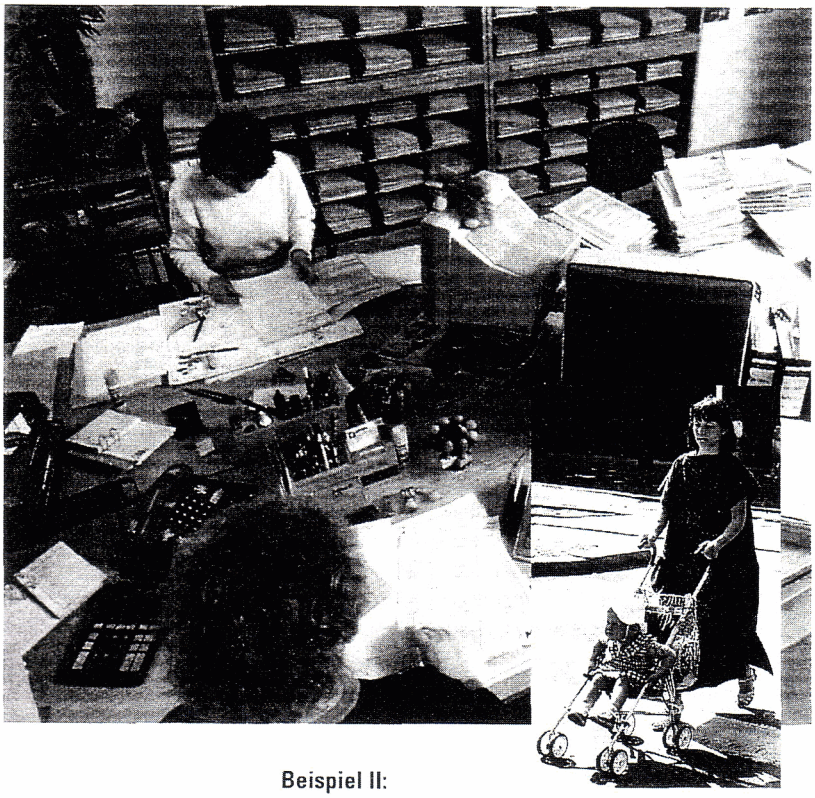
von Anomalien entdeckt, die auf fast allen großen Kapitalmärkten und bereits seit langer Zeit existieren und die diesem Modell widersprechen. Dazu zählen: der Größen-Effekt, der Januar-Effekt, der Buchwert/Marktwert-Effekt und der Winner/Loser-Effekt.

Größen-Effekt: Damit wird die empirische Regelmäßigkeit bezeichnet, daß Aktien mit niedriger Marktkapitalisierung in der Vergangenheit in Deutschland eine um durchschnittlich zwei Prozent höhere Rendite erzielten als Aktien von Unternehmen mit hoher Marktkapitalisierung. Es ist also sinnvoll, auch Aktien mittlerer und kleiner Unternehmen ins Portefeuille aufzunehmen und nicht nur die Aktien der größten und bekanntesten Gesellschaften.
Januar-Effekt: Er bezeichnet die Erscheinung, daß Aktien im Schnitt in den letzten Dezembertagen sowie im Januar/Februar ungewöhnlich stark steigen. Also, mit Aktienkäufen nicht wie beim Kauf von Weihnachtsgeschenken bis März warten, sondern noch möglichst Anfang Dezember ordern.
Buchwert/Marktwert-Effekt: Er bezieht sich auf die leicht höheren Durchschnittsrenditen von Aktien, bei denen dieser Quotient relativ hoch ist, also nahe bei Eins liegt.
Winner/Loser-Effekt: Danach ist die Durchschnittsrendite von Aktien, deren Kurs relativ stark gefallen ist, höher als die von Aktien, deren Kurs im letzten Jahr gestiegen ist.

Hauptziel des Projektes C1 ist es, derartige Anomalien zu überprüfen und gegebenenfalls Erklärungshypothesen für ihre Existenz zu erarbeiten und zu testen.

Teil-/Projekte des Sonderforschungsbereichs Quantifikation und Simulation ökonomischer Prozesse

Projektbereich		engeres Fach	Leiter
A	**Quantitative Verfahren**		
A1	Semiparametrische Modelle	Angewandte Statistik	Wolfgang Härdle
A2	Nichtparametrische Zeitreihenanalyse	Ökonometrie	Helmut Lütkepohl
A3	Eine Methodenbank zur interoperablen Modellierung ökonomischer Prozesse	Wirtschaftsinformatik	Oliver Günther
B	**Mathematische Methoden**		
B1	Kurvenschätzung und Resamplingverfahren	Mathematische Statistik	Olaf Bunke
B2	Statistik stochastischer Prozesse und Stochastik der Finanzmärkte	Stochastik	Uwe Küchler
B3	Nichtlineare Modelle und Verfahren	Mathematische Statistik	Henning Läuter
C	**Ökonomische Modellierung**		
C1	Bestimmungsfaktoren von erwarteten Aktienrenditen	Finanzierung/Kapitalmärkte	Richard Stehle
C2	Wandel auf dem Arbeitsmarkt	Arbeitsökonomie	Michael Burda
C3	Stabilität der Geldnachfrage in der Bundesrepublik Deutschland	Makro-/Ökonometrie	Helmut Lütkepohl
C4	Dynamik des Wettbewerbs	Wirtschaftstheorie/ Industrieökonomie	Elmar Wolfstetter



Beispiel II:
Frauenbeschäftigung

Die Bestimmung von Einflußfaktoren für die Beschäftigung von Frauen ist ein wichtiges Thema der quantitativen Wirtschaftsforschung. Diese Thematik wird in den Projekten A1, B1, B3 und C2 mit verschiedenen Methoden behandelt. Als Faktoren für die Beschäftigung werden etwa die allgemeine Arbeitslosenrate oder der Steuersatz bei Beschäftigung zu einem bestimmten Zeitpunkt betrachtet. Ebenso werden für die Frauenbeschäftigung im allgemeinen Größen wie Alter, Ausbildungszeit, Gehalt des Ehepartners, Ausbildung von Mutter und Vater, die Zeiten früherer Beschäftigung und die Indikatorvariable »Kinder unter sechs Jahren« betrachtet. Eine Indikatorvariable ist gleich 1, falls die Bedingung (hier »eine Frau hat Kinder unter sechs Jahren«) erfüllt ist, sonst 0. Die Tatsache der Nichtbeschäftigung der Frauen wird ebenfalls als Indikatorvariable codiert (Y=1=»beschäftigt«; Y=0=»nicht beschäftigt«). Die Frage an den quantitativen Ökonomen ist nun die der Modellierung der Frauenbeschäftigung Y als Funktion von $X = (X_1, X_2, ..., X_8, Z)$, wobei X_1=»Alter«, X_2=»Ausbildung«, X_3=»Gehalt des Ehepartners«, X_4=»Steuersatz«, X_5=»Ausbildung der Mutter«, X_6=»Ausbildung des Vaters«, X_7=»Arbeitslosenrate«, X_8=»Berufserfahrung«, Z=»Kinder« bezeichnen.
Das Interesse der Modellierung wird sich insbesondere auf das Studium des Einflusses von Einzelfaktoren konzentrieren. Exemplarisch möchten wir hier der Variablen Z (»Kinder unter sechs Jahren«) besonderes Augenmerk schenken. Die Aufgabe der Modellierung kann folgendermaßen formuliert werden: Finde eine Approximation an die Chance (Wahrscheinlichkeit) der Beschäftigung von Frauen als Funktion der Faktoren X. Dies geschieht wie auch bei der Bestimmung anderer ökonomischer Größen (wie Inflationsrate, DAX, Kreditglaubwürdigkeit, ...) durch die Bildung eines Index. Dieser Index ist eine Wichtung der Einflußfaktoren.

$$\text{index} = \beta_1 X_1 + \beta_2 X_2 + ... + \beta_8 X_8 + \alpha Z.$$

Die Gewichte β_i und α bestimmen insbesondere durch ihr Vorzeichen die größere oder kleinere Chance für die Beschäftigung der einzelnen Frau. Ein positives Gewicht β_2 etwa besagt, daß die Ausbildungzeit die Chancen der Beschäftigung erhöht. Es ist klar, daß dieser Index nur eine Approximation darstellen kann, selbst zwei Individuen mit den gleichen Faktoren X können verschiedene Beschäftigungsverhältnisse haben. Diese im Modell enthaltene Unsicherheit wird durch einen Fehlerterm aufgefangen. Dieser Fehlerterm wurde in bisherigen Studien mit einer theoretisch vorgegebenen Verteilung modelliert und führte zu sog. Probit- oder Logit-Modellen. Die Gründe dafür waren zuallererst praktischer Natur: Die Rechenkapazität war nicht ausreichend für die Analyse komplexerer und datentreuerer Modelle. Empirische Befunde der letzten Jahre deuten jedoch darauf hin, daß diese (parametrischen) Modelle zu stark verzerrten Aussagen über die Indexgewichtung führen. Am SFB werden nun unter Einsatz von Hochleistungsrechnern (Workstations) diese restriktiven Annahmen aufgelöst und damit eine realistischere Modellierung durchgeführt. Erste Forschungsergebnisse gibt es für das dargestellte Beispiel der Modellierung der Frauenbeschäftigung.

Der Indexvektor (Gewichtsvektor) $(\beta_1, ..., \beta_8)$ kann durch die semiparametrische »Average Derivative Estimation (ADE)«–Methode bestimmt werden. Im vorliegenden Beispiel einer Stichprobe aus den USA vom Umfang n = 193 ergab sich für $(\beta_1, ..., \beta_8)$:

Alter β_1	Gehalt Ehepartner β_3	Ausbildung Mutter β_5	Arbeitslosenrate β_7
-0.188	-0.176	0.163	0.054
Ausbildung β_2	Steuersatz β_4	Ausbildung Vater β_6	Berufserfahrung β_8
0.445	-0.385	-0.269	0.697

Dies zeigt z.B. einen negativen Einfluß der Faktoren Alter, Gehalt des Ehepartners und Steuersatz. Relativ großen, positiven Einfluß haben dagegen erwartungsgemäß Ausbildung und Berufserfahrung. Die Ausbildung von Vater und Mutter der Frau haben gegensätzliche Vorzeichen. Die (exogene) Arbeitslosenrate hat für die betrachtete Stichprobe kaum Bedeutung.

Die Marginalanalyse von Z kann nun durch Aufspaltung der Gesamtstichprobe in zwei Teilstichproben entsprechend der Ausprägung von Z (0 oder 1) behandelt werden. Die folgende Abb. 4 (oben) zeigt nichtparametrische Schätzungen der Beschäftigungswahrschein-

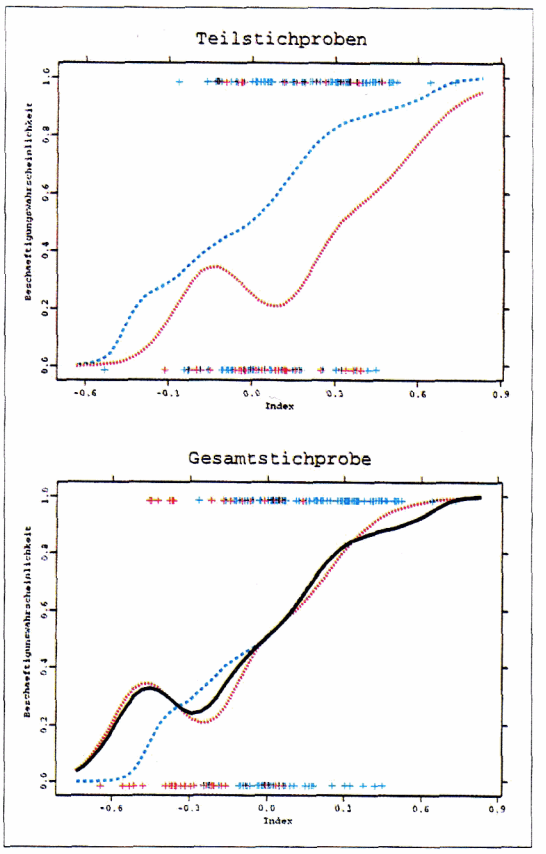


Abb. 4

Die Grafik zeigt die Beschäftigungswahrscheinlichkeit von Frauen in den Teilstichproben (blau=keine Kinder unter sechs Jahren; rot=Kinder unter sechs Jahren) und in der Gesamtstichprobe.

lichkeiten in Abhängigkeit vom Index $\beta_1 X_1 + \ldots \beta_8 X_8$. Die blaue Kurve zu den blauen Datenpunkten entspricht der Teilstichprobe »keine Kinder unter sechs Jahren« ($Z = 0$), die rote Kurve zu den roten Datenpunkten gehört zur anderen Teilstichprobe ($Z = 1$). Der Koeffizient α kann dann als »horizontale Differenz« der beiden Kurven bestimmt werden. Im vorliegenden Beispiel ergab sich $\alpha = -0.332$. Dies drückt den offensichtlich negativen Einfluß von Kindern unter sechs Jahren auf die Chance zur Beschäftigung einer Frau aus.

Verschiebt man die rote Kurve um den Wert α, müßte sie in etwa deckungsgleich mit der blauen Kurve sein. Diese Verschiebung für die rote Kurve und die zugehörigen Datenpunkte ist in Abb. 4 (unten) dargestellt. Die schwarze, durchgezogene Kurve zeigt die nichtparametrische Schätzung der Beschäftigungswahrscheinlichkeit in Abhängigkeit vom endgültigen Index $\beta_1 X_1 + \ldots \beta_8 X_8 + \alpha Z$, d.h. jetzt unter Einbeziehung der Indikatorvariablen Z.
Die resultierende Kurve zeigt einen deutlichen »Bukkel« im linken Teil, der durch die oben genannten parametrischen Verfahren nicht entdeckt werden würde. Dieser Effekt wird durch einige Frauen erzeugt, die trotz »schlechter Voraussetzungen« (im Sinne eines kleinen Indexwertes) beschäftigt sind. Dies kann z.B. durch die Einbeziehung weiterer Faktoren in die Indexbildung analysiert werden und damit Basis beschäftigungspolitischer Maßnahmen werden.

Aktivitäten des Sonderforschungsbereiches 373

Alle Aktivitäten des Sonderforschungsbereiches, wie die Einladung von Gästen, Veranstaltung von Tagungen und Workshops, werden vom Vorstand konzipiert und entschieden. Mitglieder des Vorstandes sind alle Teilprojektleiter (ex officio), die Geschäftsführerin (Dr. Sibylle Schmerbach) und vier wissenschaftliche Mitarbeiter. Der Vorstand wählt den Sprecher, der den SFB nach außen vertritt. Zu den wichtigsten Aktivitäten des SFB gehören:

Projektübergreifende Workshops

1993 »Wirtschaftsstatistik« (S. Schmerbach)
1993 »Computeraided Semiparametric Modelling« (W. Härdle/B. Rönz)
1994 »Managerkompensation« (J. Schwalbach/E. Wolfstetter)
1994/1996 »Statistics and numerics of stochastic processes with applications in finance« (U. Küchler)
1994 »Nonparametric Dynamic Modelling« (EC2-Tagung, H. Lütkepohl)
1995 »Model Management and Metadata« (O. Günther)
1995 »Curve estimation and resampling« (O. Bunke)
1996 »Applied Semiparametrics economics« (M. Burda/W. Härdle)

»Discussion papers«

Die Publikations-Reihe fördert durch projektbereichübergreifendes Referieren den wissenschaftlichen Austausch der Projekte. Die »discussion papers« sind über elektronische Maildienste abrufbar: anonymous FTP 141.20.100.2, Directory pub/papers/sfb; Postscriptfiles sfbdp001.ps.Z,... .

Seminare

Jede Woche wird in verschiedenen Seminaren für Empirische Wirtschaftsforschung an der Humboldt-Universität ein projektübergreifendes Programm realisiert, das allen Teilprojekten eine öffentliche Plattform zur Darstellung und oft auch kontroversen Diskussion ihrer Forschungsergebnisse bietet.

SFB-Newsletter

Diese wöchentlich erscheinende Publikation macht alle wissenschaftlichen Aktivitäten des SFB einer interessierten Öffentlichkeit bekannt. Redaktion: Dr. Sibylle Schmerbach, Eimear Kelly, B.A., D.B.S.

Die Aktivitäten sind ein wesentliches Element der Forschungstätigkeit des SFB. Sie werden maßgeblich dazu beitragen, Ost-Berlin als Wissenschaftsstandort auch im internationalen Bereich wieder zu etablieren.

Prof. Dr. Wolfgang Härdle

Nach Promotion (Universität Heidelberg) und Habilitation (Universität Bonn) war W. Härdle zunächst Visiting Professor (1989–90), dann Professeur Ordinaire (C4), CORE, an der Université Catholique de Louvain (1990–92). 1992 nahm er den Ruf als Ordentlicher Professor (C4) für Wirtschaftswissenschaften an der Humboldt-Universität an. W. Härdle publiziert u.a. zu den Bereichen Angewandte Statistik und Ökonometrie und ist (Mit-)Herausgeber zahlreicher Sammelbände und Schriftenreihen.

Kontakt

Humboldt-Universität zu Berlin
Wirtschaftswissenschaftliche Fakultät
Institut für Statistik und Ökonometrie
Spandauer Str. 1
D-10178 Berlin
Tel.: 030/2468-231/230
Fax: 030/2468-249
Bitnet: haerdle @ wiwi.hu-berlin.de

Publications de l'Institute de Statistique Université de Paris, XXXVIII, 61-86

Pub. Inst. Stat. Univ. Paris
XXXVIII, fasc. 3, 1994, 61 à 86

Kernel Estimation: the Equivalent Spline Smoothing Method *

Wolfgang Härdle
Humboldt-Universität zu Berlin
Wirtschaftswissenschaftliche Fakltät
Institut für Statistik und Ökonometrie
Spandauer Str. 1
D - 10178 Berlin
Germany

Michael Nussbaum
Weierstraß-Institut für Angewandte
Analysis und Stochastik
Mohrenstr. 39
D - 10117 Berlin
Germany

12. Juli 1994

* This paper was originally started as a "deutsch–deutsche Zusammenarbeit" financed by Sonderforschungsbereich 303, later the research on this paper was carried out within the Sonderforschungsbereich 373 at Humboldt-University Berlin and was printed using funds made available by the Deutsche Forschungsgemeinschaft.

(1994) Härdle, W. and Nussbaum, M.
Kernel Estimation: the Equivalent Spline Smoothing Method.

Publications de l'Institute de Statistique Université de Paris, XXXVIII, 61-86

ABSTRACT

Among nonparametric smoothers, there is a well-known correspondence between kernel and Fourier series methods, pivoted by the Fourier transform of the kernel. This suggests a similar relationship between kernel and spline estimators. A known special case is the result of Silverman (1984) on the effective kernel for the classical Reinsch-Schoenberg smoothing spline in the nonparametric regression model. We present an extension by showing that a large class of kernel estimators have a spline equivalent, in the sense of identical asymptotic local behaviour of the weighting coefficients. This general class of spline smoothers includes also the minimax linear estimator over Sobolev ellipsoids. The analysis is carried out for piecewise linear splines and equidistant design.

Keywords: Kernel estimator, spline smoothing, filtering coefficients, differential operator, Green's function approximation, asymptotic minimax spline.

(1994) Härdle, W. and Nussbaum, M.
Kernel Estimation: the Equivalent Spline Smoothing Method.

Publications de l'Institute de Statistique Université de Paris, XXXVIII, 61-86

1. Introduction

It is part of the basic knowledge about smoothing methods that there is a correspondence between kernel and orthogonal series methods. Loosely speaking, and supposing a circular setting on the unit interval, we can say that a kernel estimator is equivalent to a tapered orthogonal series estimator, where the tapering coefficients are the Fourier coefficients of the kernel scaled with bandwidth parameter h. This is just a way of saying that convolution (which is what a kernel smoother does) is equivalent to multiplication of Fourier transforms. Such a relationship, which is elementary in the classical Fourier series context, can also be established between kernel and spline estimators. It is the purpose of the present paper to make this precise, and thus to contribute to a better understanding of smoothing methods in nonparametric estimation.

Our starting point is the result of Silverman (1984) who proved such a correspondence for the classical Reinsch-Schoenberg smoothing spline. Consider the nonparametric regression problem of estimating a curve m given observations

$$Y_i = m(x_i) + \varepsilon_i, \qquad i = 1, \dots n.$$

Assume that the design points $x_i \in [0,1]$ are known and nonrandom, and the ε_i are random errors. The standard cubic spline smoother is defined to be the minimizer over functions g of

$$n^{-1} \sum_{i=1}^{n} (Y_i - g(x_i))^2 + \lambda \int (g''(x))^2 \, dx \tag{1.1}$$

where λ is a smoothing parameter. It was shown that this procedure is equivalent to using a certain kernel estimator, where in addition the bandwidth varies locally on $[0,1]$ in dependence on the design density. It should be stressed that, although the theorem was proved in a statistical context, that result is *of purely analytic nature*. Indeed the smoothing philosophy can be developed in a deterministic framework, and the methods have been studied thoroughly. For other

(1994) Härdle, W. and Nussbaum, M.
Kernel Estimation: the Equivalent Spline Smoothing Method.

Publications de l'Institute de Statistique Université de Paris, XXXVIII, 61-86

approximation-theoretic results on splines connected specifically with statistics see Utreras (1983) and Cox (1984a).

In our generalization we establish that, essentially, to each kernel estimator based on a kernel K there corresponds a certain spline estimator with 'effective kernel' K. This correspondence is analogous to the one between kernel and orthogonal series smoothers, and is based on the fact that there is a basis in the space of splines which is some way close to the classical Fourier basis. The Fourier transform of the Kernel K determines the shape of the spline smoother, and Silverman's (1984) result appears as a special case.

Let us introduce the following notations. By $(\cdot,\cdot)$ and $\|\cdot\|$ we denote the scalar product and norm in $L_2(0,1)$, respectively. For natural p, let D^p be the derivative of $f \in L_2(0,1)$ in the distributional sense, and let

$$W_2^p(0,1) = \{f \in L_2(0,1) \quad ; \quad \mathrm{D}^p f \in L_2(0,1)\}$$

be the Sobolev space of order p on the unit interval. For functions f and g we define the 'design inner product'

$$\langle f, g\rangle_n = n^{-1} \sum_{i=1}^{n} f(x_i) g(x_i)$$

and the differential bilinear form

$$(f, g)_p = (\mathrm{D}^p f, \mathrm{D}^p g).$$

The spline basis we have in mind is the Demmler-Reinsch basis, i. e. the n-tuple of functions ψ_{in}, $i = 1, \ldots, n$ in $W_2^p(0,1)$ which simultaneously diagonalize the bilinear forms $\langle\cdot,\cdot\rangle_n$ and $(\cdot,\cdot)_p$:

$$\langle \psi_{in}, \psi_{jn}\rangle_n = \delta_{ij} \qquad , \qquad (\psi_{in}, \psi_{jn})_p = \gamma_{in}\delta_{ij} \quad , \quad i, j = 1, \ldots, n$$

and where $\gamma_{1n} \leq \ldots \leq \gamma_{nn}$ are minimal for all such n-tuples. It is well known that, for $p = 2$, the minimizer of (1.1), $\tilde{g}$ say, is of the form

$$\tilde{g} = \sum_{i=1}^{n} c_i \psi_{in} \tilde{Y}_i \qquad , \qquad \tilde{Y}_i = \langle Y, \psi_{in}\rangle_n, \tag{1.2}$$

(1994) Härdle, W. and Nussbaum, M.
Kernel Estimation: the Equivalent Spline Smoothing Method.

Publications de l'Institute de Statistique Université de Paris, XXXVIII, 61-86

see Craven and Wahba (1979). To obtain the explicit form of the coefficients c_i, we have to minimize

$$\sum_{i=1}^{n}\{(1-c_i)^2\tilde{Y}_i^2 + \lambda\gamma_{in}c_i^2\tilde{Y}_i^2\}$$

which yields $c_i = (1+\lambda\gamma_{in})^{-1}$. For the spectral numbers γ_{in} asymptotic relations are known, see e.g. Speckman (1985), Nussbaum (1985). If the design is equidistant then

$$\gamma_{in} \;=\; (\pi i)^{2p}\,(1+o(1)), \quad i,\; n \;\to\; \infty. \tag{1.3}$$

Define $h = \lambda^{1/2p}$; then from (1.3) we infer

$$c_i \;\approx\; (1+(\pi i h)^{2p})^{-1}. \tag{1.4}$$

For $p = 2$ the function $\varphi(x) = (1+(2\pi x)^4)^{-1}$ is known as the 'Butterworth filter'; we have thus

$$c_i \;\approx\; \varphi(ih/2). \tag{1.5}$$

It turns out that Silverman's effective spline kernel function K_S is the inverse Fourier transform of the Butterworth filter:

$$\begin{aligned} K_S(t) &= \int_{-\infty}^{\infty} \exp(-2\pi i t x)\;\varphi(x)dx \\ &= \frac{1}{2}\exp(-|u|/\sqrt{2})\;\sin(|u|/\sqrt{2}+\pi/4). \end{aligned} \tag{1.6}$$

From the form of the coefficients (1.5) and the orthogonal expansion (1.2) we understand why K_S should be the 'effective kernel' of the classical spline smoother; we shall amplify on this below. Our recipe, to obtain an equivalent spline smoother for a kernel estimator with kernel K, is now obvious: take φ in (1.5) as the Fourier transform $\hat{K}$ of K. The correspondence will be made rigorous by a theorem on the local behaviour of the newly defined spline smoother. However we have been able to carry out this program on the rigorous level as yet only for

(1994) Härdle, W. and Nussbaum, M.
Kernel Estimation: the Equivalent Spline Smoothing Method.

Publications de l'Institute de Statistique Université de Paris, XXXVIII, 61-86

piecewise linear splines. Hence Silverman's result, which refers to cubic splines, is not a special case, but rather its analog for the piecewise linear case. Still we believe this result to be instructive and pointing to the validity for splines of arbitrary degree.

A standard assumption in this context is that the nonrandom design points $x_1, \ldots, x_n$ behave regularly as $n \to \infty$, in the sense that the associated empirical distribution function L_n tends to a limit L which has a density ℓ. Then, according to Silverman (1984), the equivalent kernel estimator is one in which the bandwidth varies locally on $[0,1]$, in dependence on the limiting design density ℓ. For our result on the general class of spline smoothers, we confine ourselves to uniform ℓ; more specifically, an equidistant design will be assumed. It is easy to see that the local variability of the bandwidth of the equivalent kernel estimator in the case of nonuniform ℓ is a phenomenon which is independent of the kernel shape, and should hold in our general framework.

Using the terminology of time series analysis, the function φ in (1.5) may be termed a *filter*. It has been established that the Pinsker filter

$$\varphi(x) \;=\; (1 - |2\pi x|^p)_+$$

(cf. Pinsker (1980)) is connected with the minimax-among-linear estimator over Sobolev classes

$$W_2^p(Q) = \{f \in W_2^p(0,1) \quad ; \quad \|\mathrm{D}^p f\|^2 \le Q\}$$

when the loss is the squared norm deviation induced by the design inner product $\langle \cdot, \cdot \rangle_n$, see Speckman (1985). Also it is known that, for independent identically normally distributed ε_i, this spline estimator attains the best possible constant in the L_2-risk asymptotics, in a minimax sense over the Sobolev class; cf. Nussbaum (1985). In this setting the Butterworth filter, i. e. the classical spline smoother is not optimal, and this is one of the motivations for our extended class.

(1994) Härdle, W. and Nussbaum, M.
Kernel Estimation: the Equivalent Spline Smoothing Method.

Publications de l'Institute de Statistique Université de Paris, XXXVIII, 61-86

Dealing with the classical spline smoother, Cox (1983), (1984a) developed an effective framework for approximating it by the continuous analog, i. e. by a method-of-regularization operator. Our approach is inspired by these results; however, due to the particular simplicity of the selected special case, we are able to apply more direct methods. It should be noted that the conditions of Cox (1983), (1984a) exclude the piecewise linear case (a priori smoothness 1); thus our result seems to indicate a possible weakening of those regularity conditions.

Messer (1991) and Messer and Goldstein (1993) elaborate the result on the classical spline smoother, obtaining considerable analytic insight, but their analysis is still limited to Silverman's particular case. An important contribution to the general equivalence problem has been made by Thomas-Agnan (1991); we discuss this in the remarks at the end of the paper.

2. The Spline Kernel

To shed some more light on the equivalence which is the subject of this paper, we will follow Cox (1984a) in considering *the associated continuous smoothing problem.* In (1.1), put aside the randomness of the data Y_i for a moment, and assume that $Y_i = m(x_i),\ i = 1, \ldots, n$, where m is a continuous function on $[0,1]$. Then as $n \to \infty$, the minimization criterion (1.1) will be close to

$$\int_0^1 (m(x) - g(x))^2\ \ell(x)dx + \lambda\ \|\mathrm{D}^p g\|^2 \tag{2.1}$$

(for $p = 2$; in the sequel p will be general). Similarly, the Demmler-Reinsch spline basis will tend to a limiting orthogonal system $\psi_i,\ i = 1, 2, \ldots$ in $L_2(0,1)$ which may be characterized as follows. We have

$$\int_0^1 \psi_i \psi_j\ dL = \delta_{ij} \qquad , \qquad (\psi_i, \psi_j)_p = \gamma_i \delta_{ij} \quad , \quad i, j = 1, 2, \ldots$$

where $\gamma_1 \leq \gamma_2 \leq \ldots$, and the basis $\{\psi_i\}$ is extremal in the sense that the spectral values γ_i are minimal. The continuous analog of the smoothing operator

(1994) Härdle, W. and Nussbaum, M.
Kernel Estimation: the Equivalent Spline Smoothing Method.

Publications de l'Institute de Statistique Université de Paris, XXXVIII, 61-86

(1.2) then is

$$\tilde{g} = \sum_{i=1}^{\infty} c_i m_i \psi_i \quad , \quad m_i = (m, \psi_i) \tag{2.2}$$

where $c_i = (1+\lambda\gamma_i)^{-1}$. For our general class of smoothers, we put $c_i = \varphi(ih/2)$ for some filter φ (remind $\lambda = h^{2p}$). Thus the analysis of spline smoothing operators may be broken up into two parts:

- approximate the discrete problem by the continuous one, as $n \to \infty$, uniformly over a range of h
- study the continuous problem for smoothing parameter $h \to 0$.

Let us further examine the continuous problem, to see why a relationship like (1.6) should be expected between the filter function φ and the effective kernel K. For simplicity let us first assume that the limiting design density is uniform: $\ell \equiv 1$. It is well known that the basis functions ψ_i are eigenfunctions of the differential operator $(-\mathrm{D}^2)^p$ defined on functions in W_2^{2p} which satisfy natural (Neumann) boundary conditions:

$$(-\mathrm{D}^2)^p \psi_j = \gamma_j \psi_j \tag{2.3a}$$

$$\mathrm{D}^k \psi_j(0) = \mathrm{D}^k \psi_j(1) = 0, \; k = p, \ldots, 2p-1. \tag{2.3b}$$

The smoothing procedure (2.2) is an integral operator on $[0,1]$ with kernel

$$H(x,y) = \sum_{j=1}^{\infty} c_j \psi_j(x) \psi_j(y).$$

In the case $c_j = (1+\lambda\gamma_j)^{-1}$ which corresponds to the method of regularization criterion (2.1) H is the Green's function for the elliptic boundary value problem

$$(-\mathrm{D}^2)^p g + \lambda g = f \tag{2.4}$$

with boundary conditions (2.3*b*) on g. In our more general case $c_j = \varphi(jh/2)$ the function $H = H_h$ may be seen as a *generalized Green's function*. Silverman's

(1994) Härdle, W. and Nussbaum, M.
Kernel Estimation: the Equivalent Spline Smoothing Method.

Publications de l'Institute de Statistique Université de Paris, XXXVIII, 61-86

result, if translated to the continuous smoothing case, says that the classical Green's function behaves locally like a kernel K_S:

$$h\, H_h(y+ht,y) \;\to\; K_S(t) \;\text{ as } h \to 0 \tag{2.5}$$

for every $y \in [0,1]$.

This relationship may be very easily derived when we consider the *circular smoothing problem*. Suppose we seek the minimizer g of (2.1) subject to periodic boundary conditions on $\mathrm{D}^k g$. This will lead to the Green's function of the problem (2.4) with boundary conditions

$$\mathrm{D}^k g(0) = \mathrm{D}^k g(1),\; k = 0, \ldots, 2p-1 \tag{2.6}$$

which can also be expressed in terms of eigenfunctions. In the periodic case these are

$$\psi_0(x) = 1, \quad \psi_j(x) = \sqrt{2}\cos(2\pi j x),\; j = 1,2,\ldots$$
$$\psi_j(x) = \sqrt{2}\sin(2\pi j x),\; j = -1,-2,\ldots$$

with corresponding eigenvalues $(2\pi j)^{2p}$. Hence for the Green's function we have, with $\varphi(x) = (1+(2\pi x)^{2p})^{-1}$

$$\begin{aligned} H_h(x,y) &= 1 + 2\sum_{j=1}^{\infty} \varphi(jh)\{\cos(2\pi j x)\cos(2\pi j y) + \sin(2\pi j x)\sin(2\pi j y)\} \\ &= 1 + 2\sum_{j=1}^{\infty} \varphi(jh)\;\cos(2\pi j(x-y)) \end{aligned}$$

since φ is symmetric about 0. Consequently we have

$$\begin{aligned} h\, H_h(y+ht,y) &= h + 2h\sum_{j=1}^{\infty} \varphi(jh)\;\cos(2\pi j t h) \\ \approx 2\int_0^{\infty} \varphi(x)\cos(2\pi x t)dx &= \int_{-\infty}^{\infty} \exp(-2\pi \mathbf{i} x t)\varphi(x)dx \;=\; K(t) \end{aligned} \tag{2.7}$$

if K is the inverse Fourier transform of φ. This relationship will carry over to general φ provided the last set of displays remains true,which will be the

(1994) Härdle, W. and Nussbaum, M.
Kernel Estimation: the Equivalent Spline Smoothing Method.

Publications de l'Institute de Statistique Université de Paris, XXXVIII, 61-86

case under appropriate smoothness and integrability conditions on φ. Thus in the periodic case we readily obtain our result on the local behaviour of the generalized Green's function

$$H_h(x,y) = \sum_{j=1}^{\infty} \varphi(jh/2)\, \psi_j(x)\psi_j(y). \tag{2.8}$$

However, to deal with the original spline smoothing problem we have to consider the nonperiodic case. Here the functions ψ_i are eigenfunctions of $(-\mathrm{D}^2)^p$ *under a different set of boundary conditions*, namely the Neumann set (2.3b). The heuristics then is clear: since we look at the local behaviour of the generalized Green's function in a neighborhood of a fixed point y in the interior of the interval, we can expect that the boundary conditions matter less and less as $h \to 0$, and the behaviour will be as in the periodic case. This interpretation is supported by the well known eigenvalue asymptotics in the Neumann case:

$$\gamma_j = (\pi j)^{2p}\,(1 + o(1)) \text{ as } j \to \infty$$

(see Agmon (1968), compare also the discrete analog (1.3)). This means that for large j the eigenvalues are close to those of the periodic problem (remind that those were $(2\pi j)^{2p}$, $j = \pm 1, \pm 2, \ldots$, with the same asymptotics under rearrangement). In (2.8), small values of j matter less as $h \to 0$, so if the eigenfunctions ψ_j have a similar tendency to approach those of the periodic problem we can expect the convergence (2.7). This is confirmed for the classical Green's function ($\varphi(x) = (1 + (2\pi x)^{2p})^{-1}$) by Silverman's result; we shall have to deal with the case of general φ fulfilling appropriate conditions.

We remark that Huber (1979) considered the discrete periodic smoothing problem in the case of an equidistant design $\{x_i\}$ on the unit interval, and obtained another approximation to the effective kernel of the procedure. It is shown to be equivalent to Silverman's result by Härdle (1989), chap. 3.4.

3. The continuous smoothing problem

(1994) Härdle, W. and Nussbaum, M.
Kernel Estimation: the Equivalent Spline Smoothing Method.

Publications de l'Institute de Statistique Université de Paris, XXXVIII, 61-86

We now proceed to derive the asymptotic relation (2.7) for the generalized Green's function (2.8) for the limiting continuous smoothing problem, in the nonperiodic case. Here the functions ψ_j figuring in (2.8) are the eigenfunctions in the problem (2.3) on the interval $[0,1]$. We are able to obtain the desired result as yet only in the case $p = 1$ and $\ell \equiv 1$ (uniform design density). The eigenfunctions in this case are

$$\psi_1(x) = 1, \quad \psi_j(x) = \sqrt{2}\cos(\pi(j-1)x), \; j = 2, 3, \ldots \tag{3.1}$$

with corresponding eigenvalues $(\pi(j-1))^2$ (see Triebel (1972), theorem 23.3, p. 301).

Let us now fix appropriate conditions on the filter function φ and the kernel K. We shall use the following notations. By $L_q(a,b)$, $q = 1, 2$ we denote the L_q-space of complex-valued functions on an interval (a,b); when $(a,b) = \mathbb{R}$ we write L_q. Furthermore consider the Sobolev spaces $W_2^1(a,b)$ as defined in section 1; we write W_2^1 if $(a,b) = \mathbb{R}$. Integrals without limits extend over $\mathbb{R}$.

Now let K be a real-valued function on $\mathbb{R}$ with

$$K \in L_1 \;, \quad \int K(x)dx \;=\; 1 \;, \; K(x) = K(-x). \tag{3.2}$$

For any $g \in L_1$ let $\hat{g}$ be the Fourier transform of g:

$$\hat{g}(t) \;=\; \int \exp(2\pi i t x)\; g(x)\; dx.$$

Define the filter function φ as $\varphi = \hat{K}$. Then we can state the following elementary result.

Publications de l'Institute de Statistique Université de Paris, XXXVIII, 61-86

Proposition 3.1. *Let K be a kernel satisfying conditions (3.2). Then $\varphi = \hat{K}$ has properties*

(i) φ *is real and symmetric about* 0

(ii) $\varphi(0) = 1$, φ *is bounded and continuous.*

Furthermore, assume that $K \in L_2$ and understand the Fourier transform as defined on L_2. Then φ is also in L_2, and K is the inverse Fourier transform of φ:

$$K(u) = \hat{\varphi}(-u) = \hat{\varphi}(u).$$

At this point let us introduce tail and smoothness conditions on K. Define the set V_2^1 of complex valued functions on $\mathbb{R}$ as

$$V_2^1 = \{f \in L_2 \ , \ \int (1 + |x|^2)|f(x)|^2 \, dx < \infty\}.$$

It is well known that $f \in W_2^1$ is equivalent to $\hat{f} \in V_2^1$, and $\widehat{DK}(t) = 2\pi i t \hat{f}(t)$. Our additional condition on K is

$$K \in W_2^1 \ , \ K' \in V_2^1. \tag{3.3}$$

Define the operator J by $(Jf)(x) = xf(x)$.

Proposition 3.2. *Let K be a kernel satisfying conditions (3.3). Then*

(iii) $\varphi \in V_2^1$

(iv) $J\varphi \in W_2^1$.

Lemma 3.1. *Let K be a kernel satisfying conditions (3.2), (3.3). Then for $\varphi = \hat{K}$ we have*

$$\sup_{h>0} \ h \sum_{j=1}^{\infty} (jh)^2 \, \varphi^2(jh) < \infty.$$

Proof. Define an interval $A_{jh} = ((j-1)h, jh)$. By standard imbedding theorems

$$(jh)^2 \varphi^2(jh) = (J\varphi(jh))^2 \leq C \left\{h^{-1} \|J\varphi\|^2(A_{jh}) + h\|(J\varphi)'\|^2(A_{jh})\right\}.$$

(1994) Härdle, W. and Nussbaum, M.
Kernel Estimation: the Equivalent Spline Smoothing Method.

Publications de l'Institute de Statistique Université de Paris, XXXVIII, 61-86

Now sum over j and use property (iv) of φ .

□

We are now in a position to define our generalized Green's function: for any $x, y \in [0,1]$ and functions ψ_j from (3.1) we set

$$H_h(x,y) \;=\; 1 + 2\sum_{j=1}^{\infty} \varphi(jh/2)\, \cos(\pi j x) \cos(\pi j y). \tag{3.4}$$

Lemma 3.1 ensures convergence of the series uniformly over x, y. Putting

$$\cos(\pi j x) \;=\; \frac{1}{2}\left(\exp(\pi \mathbf{i} j x) + \exp(-\pi \mathbf{i} j x)\right)$$

and $x = y + th$, we obtain, using the symmetry of φ,

$$\begin{aligned} H_h(y+th, y) \;&=\; 1 + \frac{1}{2} \sum_{j=-\infty, j\neq 0}^{\infty} \varphi(jh/2)\, \left\{\exp(\pi \mathbf{i} j h t) \exp(2\pi \mathbf{i} j y) + \exp(\pi \mathbf{i} j h)\right\} \\ &=\; \frac{1}{2} \sum_{-\infty}^{\infty} \varphi(jh/2) \exp(\pi \mathbf{i} j h t)\, \left\{1 + \exp(2\pi \mathbf{i} j y)\right\}. \end{aligned}$$

Lemma 3.2. *For any t we have as $h \to 0$*

$$\frac{h}{2} \sum_{j=-\infty}^{\infty} \varphi(jh/2) \exp(\pi \mathbf{i} j h t) \;\to\; \int \exp(2\pi \mathbf{i} u t)\, \varphi(u)\, du.$$

Proof. Define

$$\varphi_t(x) \;=\; \varphi(x) \exp(2\pi \mathbf{i} x t).$$

For simplicity we substitute $h/2$ by h in the lemma. Consider intervals A_{jh} as in lemma 3.1. The difference of the two sides in the present lemma is

$$\begin{aligned} &\sum_{j=-\infty}^{\infty} \int_{A_{jh}} (\varphi_t(x) - \varphi_t(jh))\, dx \\ \leq\; &\sum_{j=-\infty}^{\infty} \int_{A_{jh}} |\varphi_t(x) - \varphi_t(jh)|\, dx. \end{aligned} \tag{3.5}$$

(1994) Härdle, W. and Nussbaum, M.
Kernel Estimation: the Equivalent Spline Smoothing Method.

Publications de l'Institute de Statistique Université de Paris, XXXVIII, 61-86

Consider first intervals A_{jh} which do intersect with $[-2,2]$. The corresponding sum of terms in (3.5) is $o(1)$, since φ_t is continuous. For the other intervals, the expression under the integral sign is bounded by

$$\Big(\int_{A_{jh}} (x\varphi_t'(x))^2\, dx\Big)^{1/2} \Big(\int_{A_{jh}} x^{-2} dx\Big)^{1/2}.$$

The Cauchy-Schwartz inequality then gives an upper bound for (3.5)

$$\Big(\int_{|x|>1} x^{-2} dx\Big)^{1/2} \|J(\varphi_t')\|\ h\ +\ o(1). \tag{3.6}$$

Now we have

$$(J\varphi)' \ =\ \varphi + J\varphi',$$

hence

$$\|J\varphi'\| \ \leq\ \|\varphi\| + \|(J\varphi)'\|.$$

Furthermore

$$\varphi_t' \ =\ (2\pi \mathbf{i} t\varphi + \varphi')\ \exp(2\pi \mathbf{i} t).$$

Consequently

$$\|J\varphi_t'\| \ \leq\ 2\pi t\|J\varphi\| + \|J\varphi'\| \leq\ 2\pi t\|J\varphi\| + \|\varphi\| + \|(J\varphi)'\|.$$

By proposition (3.2) all these terms are finite, hence (3.6) is $o(1)$.

□

Lemma 3.3. *For any $\delta > 0$, we have as $h \to 0$*

$$\frac{h}{2} \sum_{j=-\infty}^{\infty} \varphi(jh/2) \exp(\pi \mathbf{i} jht)\ \exp(2\pi \mathbf{i} jy) \ =\ o(1)$$

uniformly over $y + ht \in (\delta, 1-\delta)$.

Proof. Let k be a natural, and observe that

$$\begin{aligned} & h \textstyle\sum_{|j|>k} \varphi(jh/2) \exp(2\pi \mathbf{i} j(y + ht/2)) \\ \leq\ & \big(\textstyle\sum_{|j|>k} h^{-1} j^{-2}\big)^{1/2}\ \big(\textstyle\sum_{|j|>k} h(hj)^2 \varphi^2(jh/2)\big)^{1/2}. \end{aligned}$$

(1994) Härdle, W. and Nussbaum, M.
Kernel Estimation: the Equivalent Spline Smoothing Method.

Publications de l'Institute de Statistique Université de Paris, XXXVIII, 61-86

According to lemma 3.1 the second factor is bounded, uniformly over h and k. The first factor is

$$\left(h^{-1}\ O(k^{-1})\right)^{1/2}.$$

Suppose that $k \sim Mh^{-1}$; then for sufficiently large M the above term can be made less than $\varepsilon/2$. The remaining sum over terms $|hj| \leq M$ in the series is estimated as follows. This sum can be construed as being a series as in the assertion, with φ having support on $[-M, M]$ and being continuous there. Take a finite partition of $[-M, M]$ into intervals of equal length. Since φ can be approximated by corresponding step functions, uniformly on $[-M, M]$ if the partition becomes finer, it suffices to prove the lemma for each such step function. Each such step function is a linear combination of functions which are indicators of symmetric intervals $[-a, a]$, $a < M$. Hence it suffices to prove the lemma for each $\varphi = \chi_{(-a,a)}$, the indicator of some symmetric interval. In this case, for $r = [h^{-1}a]$ we have

$$h \sum_{|j| \leq h^{-1}a} \exp(2\pi \mathbf{i} j(y + ht)) \ = \ h\ D_r(y + ht), \tag{3.7}$$

where $D_r(\cdot)$ is the Dirichlet kernel

$$D_r(x) \ = \ \frac{\sin(\pi(2r+1)x)}{\sin(\pi x)}.$$

Now for $x \in (\delta, 1-\delta)$ the numerator is bounded away from 0, hence $D_r(x)$ is uniformly bounded for $r \geq 1$, $x \in (\delta, 1-\delta)$. As $h \to 0$, (3.7) proves the lemma. □

The final result on the generalized Green's function H_n can now be stated as follows. Observe beforehand that the convergence of lemma 3.2 holds uniformly over $|t| \leq C$, and also uniformly in h over any range $h \leq \overline{h}$ such that $\overline{h} \to 0$. The convergence of lemma 3.3 holds uniformly over $y + ht \in (\delta, 1-\delta)$ and $h \leq \overline{h}$.

(1994) Härdle, W. and Nussbaum, M.
Kernel Estimation: the Equivalent Spline Smoothing Method.

Publications de l'Institute de Statistique Université de Paris, XXXVIII, 61-86

Lemma 3.4. *We have for any* $y \in (0,1)$, $t \in \mathbb{R}$

$$h\, H_h(y+ht, y) \;\to\; K(t) \;\; as \;\; h \to 0,$$

and the convergence is uniform over $y \in (\delta, 1-\delta)$, $(\delta > 0)$, $|t| \le C$ *and* $h \in (0, \overline{h})$ *where* $\overline{h} \to 0$.

4. The spline basis

Having treated the limiting continuous smoothing problem for degree of differentiability $p = 1$ and uniform limiting design ($\ell \equiv 1$), we now look at the discrete analog, i. e. the problem with data observed at points $x_1, \ldots, x_n$. For this we assume that the regression design is of a particular uniformly spaced kind:

$$x_i \;=\; (i - 1/2)/n \;,\; i = 1, \ldots, n.$$

It is well known that the natural interpolation and smoothing splines for $p = 1$ are piecewise linear. For given $\{x_i\}$ as above and a function f defined on $[0,1]$, let $f^{(n)} = (f(x_1), \ldots, f(x_n))'$ be the trace of f on $\{x_i\}$. Let $S(f^{(n)})$ be the piecewise linear interpolant of f, uniquely defined on $[0,1]$ by the requirement to be constant on the marginal intervals $[0, 1/2n]$, $[1 - 1/2n, 1]$. The following fact is well known; see e.g. Laurent (1972), theorem 4.1.3.

Lemma 4.1. *For* $f \in W_2^1(0,1)$, *the function* $S(f^{(n)})$ *is in* $W_2^1(0,1)$, *and is the solution of*

$$\min\{\; \|Dg\|^2 \;;\; g^{(n)} = f^{(n)} \;,\; g \in W_2^1(0,1) \;\}.$$

Let $\mathcal{S}_n = S(\mathbb{R}^n)$ be the n-dimensional linear space of such piecewise linear spline functions. It is clear that there is a basis ψ_{jn}, $j = 1, \ldots, n$ in $\mathcal{S}_n$

(1994) Härdle, W. and Nussbaum, M.
Kernel Estimation: the Equivalent Spline Smoothing Method.

Publications de l'Institute de Statistique Université de Paris, XXXVIII, 61-86

which simultaneously diagonalizes the bilinear forms $\langle \cdot, \cdot \rangle_n$ and $(\cdot, \cdot)_1$. Lemma 4.1 implies that $\{\psi_{jn}\}$ coincides with the Demmler-Reinsch basis (for $p = 1$) introduced in section 1. Obviously the standard smoothing spline for $p = 1$, i. e. the minimizer over functions in $W_2^1(0,1)$ of

$$n^{-1} \sum_{i=1}^{n} (Y_i - g(x_i))^2 + \lambda \int (\mathrm{D}g(x))^2 \, dx \tag{4.1}$$

is in $\mathcal{S}_n$, and hence can be expressed in terms of the basis $\{\psi_{jn}\}$ according to (1.2). Here the filtering coefficients c_j are $c_j = (1 + \lambda \gamma_{jn})^{-1}$; the interpolation spline $S(Y)$ is obtained for $c_j = 1$ (no smoothing).

It turns out that in our particularly simple setting the functions ψ_{jn} are just the spline interpolants of the ψ_j from the limiting continuous problem, i. e. of the cosine functions given by (3.1).

Lemma 4.2. *The functions ψ_{jn} defined by*

$$\psi_{jn} \;=\; S(\psi_j^{(n)}) \;,\; j = 1, \ldots, n$$

ψ_j being given by (3.1), satisfy

$$\langle \psi_{in}, \psi_{jn} \rangle_n = \delta_{ij} \qquad , \qquad (\psi_{in}, \psi_{jn})_1 = \gamma_{jn} \delta_{ij} \quad , \quad i, j = 1, \ldots, n$$

where

$$\gamma_{jn} \;=\; 4n^2 \sin^2(\pi(j-1)/2n) \;,\; j = 1, \ldots, n \;.$$

Proof. Consider a set of points: $x_k = (k - 1/2)/n$, $k = 1, \ldots, 2n$. Then for any natural r, $1 \le r \le 2n - 1$, the set of points $\exp(\pi \mathbf{i} r x_k)$, $k = 1, \ldots, 2n$ is evenly spaced on the unit circle in the complex plane. Hence

$$\sum_{k=1}^{2n} \exp(\pi \mathbf{i} r x_k) \;=\; 0. \tag{4.2}$$

(1994) Härdle, W. and Nussbaum, M.
Kernel Estimation: the Equivalent Spline Smoothing Method.

Publications de l'Institute de Statistique Université de Paris, XXXVIII, 61-86

Observe that each function $\cos(\pi r x)$, for $1 \le r \le 2n-1$ is symmetric on the interval $(0,2)$ with symmetry center 1. Hence

$$\sum_{k=1}^{n} \cos(\pi r x_k) \;=\; \frac{1}{2}\sum_{k=1}^{2n} \cos(\pi r x_k) \;=\; 0 \tag{4.3}$$

as a consequence of (4.2). Now we have for $i,j \ge 2$

$$\langle \psi_{in}, \psi_{jn} \rangle_n \;=\; n^{-1} \sum_{k=1}^{n} \{\cos(\pi(i-j)x_k) + \cos(\pi(i+j-2)x_k)\}.$$

This expression vanishes if $i \neq j$, according to (4.3), and equals 1 if $i = j$. The case where one of the ψ_{jn} is ψ_{1n}, i. e. identically 1, can be treated analogously. Thus the first orthogonality relation is proved. For the second, suppose first that either i or j is 1. Then, as $D\psi_{1n} \equiv 0$ and $\gamma_{1n} = 0$, the claim about $(\cdot,\cdot)_1$ is clear. Suppose now that $i,j \ge 2$. Consider a set of points $z_k = k/n,\ k = 0,\ldots,2n$. Analogously to (4.2) it can be shown that for $1 \le r \le 2n-1$

$$\sum_{k=1}^{2n} \exp(\pi \mathbf{i} r z_k) \;=\; 0. \tag{4.4}$$

Observe that each function $\sin(\pi r x)$, for $1 \le r \le 2n-1$ is antisymmetric on the interval $(0,2)$ about 1 and vanishes in 0 and 1. Hence for $1 \le i,j \le n$

$$\begin{aligned} \frac{2}{n} \sum_{i=1}^{n-1} \sin(\pi i z_k)\sin(\pi j z_k) \;&=\; n^{-1} \sum_{i=1}^{2n} \sin(\pi i z_k)\sin(\pi j z_k) \\ &=\; \delta_{ij} \end{aligned} \tag{4.5}$$

as a consequence of (4.4). Now

$$(\psi_{in}, \psi_{jn})_1 \;=\; n^{-1} \sum_{k=2}^{n} \big((\psi_{in}(x_k) - \psi_{in}(x_{k-1}))\ (\psi_{jn}(x_k) - \psi_{jn}(x_{k-1}))\ n^2.$$

Furthermore, writing $x_{k-1} = x_k - n^{-1}$, we obtain

$$\begin{aligned} (\psi_{in}(x_k) - \psi_{in}(x_{k-1}) \;&=\; \sqrt{2}\ \sin(\pi(i-1)(x_k - 1/2n))\ 2\sin(\pi(i-1)/2n) \\ &=\; 2\sqrt{2}\ \sin(\pi(i-1)z_{k-1})\ \sin(\pi(i-1)/2n). \end{aligned}$$

(1994) Härdle, W. and Nussbaum, M.
Kernel Estimation: the Equivalent Spline Smoothing Method.

Publications de l'Institute de Statistique Université de Paris, XXXVIII, 61-86

This yields in view of (4.5)

$$(\psi_{in}, \psi_{jn})_1 \;=$$

$$\frac{2}{n} \sum_{k=1}^{n-1} \sin(\pi(i-1)z_k)\sin(\pi(j-1)z_k)\; 4n^2 \sin(\pi(i-1)/2n)\sin(\pi(j-1)/2n)$$

$$= \delta_{ij}\; 4n^2\; \sin^2(\pi(j-1)/2n)$$

which proves the lemma.

□

Remark. The lemma describes the eigenvalues and eigenvectors of the $n \times n$ band matrix

$$\begin{pmatrix} 1 & -1 & & & & & \\ -1 & 2 & -1 & & & & \\ & -1 & 2 & & & & \\ & & & \ddots & & & \\ & & & & 2 & -1 & \\ & & & & -1 & 2 & -1 \\ & & & & & -1 & 1 \end{pmatrix}$$

as γ_{jn} and $\psi_j^{(n)}$, $j = 1, \ldots, n$. Note that

$$\gamma_{jn} \;=\; (\pi j)^2\; (1 + o(1))$$

uniformly over $k_1(n) \le j \le k_2(n)$, for any $k_1(n) \to \infty$, $k_2(n) = o(n)$ as $n \to \infty$, which is a special case of (1.3).

Let us now describe the approximation property of the ψ_{jn} for the basis $\{\psi_j\}$.

Lemma 4.3. *We have*

$$\sup_{x \in [0,1]} |\psi_{jn}(x) - \psi_j(x)| \;\le\; n^{-1}\pi j\;,\; j = 1, \ldots, n.$$

(1994) Härdle, W. and Nussbaum, M.
Kernel Estimation: the Equivalent Spline Smoothing Method.

Publications de l'Institute de Statistique Université de Paris, XXXVIII, 61-86

Proof. Set $x_0 = 0$. We have for $x \in [x_{k-1}, x_k]$, $k = 1, \ldots, n+1$

$$|\psi_{jn}(x) - \psi_j(x)| \ \leq \ \sup_{x \in [x_{k-1}, x_k]} |\psi_j'(x)| n^{-1} \ \leq \ n^{-1} \pi (j-1).$$

□

This result can be immediately applied to describe the closeness of the generalized Green's function H_h and its discrete (spline) analog. Observe that given an observation vector Y, our spline estimator is the function of $x \in [0,1]$

$$\sum_{j=1}^{n} \varphi((j-1)h/2) \psi_{jn}(x) \langle Y, \psi_{jn} \rangle_n \ = \ n^{-1} \sum_{k,j=1}^{n} \varphi((j-1)h/2) \psi_{jn}(x) \psi_{jn}(x_k) Y_k. \tag{4.6}$$

Define for $x \in [0,1]$, $k \in \{1, \ldots, n\}$

$$H_{hn}(x, x_k) \ = \ \sum_{j=1}^{n} \varphi((j-1)h/2) \psi_{jn}(x) \psi_{jn}(x_k). \tag{4.7}$$

Clearly this is the analog of the generalized Green's function (3.4).

Lemma 4.4. *Let $\underline{h}_n$, $\overline{h}_n$ be sequences: $\underline{h}_n \leq \overline{h}_n$, $\overline{h}_n \to 0$, $\underline{h}_n n \to \infty$ as $n \to \infty$. Then we have*

$$h \ |H_h(x, x_k) - H_{hn}(x, x_k)| \ \to \ 0$$

uniformly over $h \in [\underline{h}_n, \overline{h}_n]$, $x \in [0,1]$, $k = 1, \ldots, n$.

Proof. Since $\psi_j(x)$, $\psi_{jn}(x)$ are uniformly bounded, we can use the method used in the proof of lemma (3.3) to show that in both H_h and H_{hn} we need only consider summation terms for $j \leq Mh^{-1}$ for some M. It then remains to show that

$$h \ \sum\nolimits_{j \leq Mh^{-1}} \ \varphi((j-1)h/2) \ |\psi_j(x)\psi_j(x_k) - \psi_{jn}(x)\psi_{jn}(x_k)|$$

tends to zero uniformly. According to lemma (4.3), for $j \leq Mh^{-1}$

(1994) Härdle, W. and Nussbaum, M.
Kernel Estimation: the Equivalent Spline Smoothing Method.

Publications de l'Institute de Statistique Université de Paris, XXXVIII, 61-86

$$\sup_{x\in[0,1]} |\psi_{jn}(x) - \psi_j(x)| \;\le\; n^{-1}\pi M h^{-1} \;\le\; C\underline{h}_n^{-1}n^{-1} \;=\; o(1).$$

This proves the lemma.

□

Collecting the results of lemmas 3.4 and 4.4 we obtain the following result.

Theorem . *Suppose that in the regression model the design points x_i are $x_i = (i-1/2)/n$, $i = 1,\ldots,n$. Let K be a kernel function satisfying conditions (3.2), (3.3), and let $\varphi = \hat{K}$ be its Fourier transform. Let ψ_{jn}, $j = 1,\ldots,n$ be the Demmler-Reinsch basis in the space $\mathcal{S}_n$ of piecewise linear splines with knots at x_i. Consider the spline estimator given by (4.6) for smoothing parameter h, and let H_{hn} be the corresponding weight function given by (4.7). Let $\underline{h}_n$, $\overline{h}_n$ be sequences: $\underline{h}_n \le \overline{h}_n$, $\overline{h}_n \to 0$, $n\underline{h}_n \to \infty$. Then*

$$h\; H_{hn}(x_k + th, x_k) \;\to\; K(t)\;, \quad n \to \infty$$

uniformly over $x_k \in (\delta, 1-\delta)$ $(\delta > 0)$, $h \in (\underline{h}_n, \overline{h}_n)$ and $|t| \le C$.

5. Remarks

Having carried out our analysis for smoothness $p = 1$ (piecewise linear splines), it remains to include the classical spline smoother for $p = 1$ into this framework. Consider the minimizer of (4.1); as in (1.2)-(1.4) it can be seen that it corresponds to a filter function

$$\varphi(x) \;=\; (1 + (2\pi x)^2)^{-1}.$$

(1994) Härdle, W. and Nussbaum, M.
Kernel Estimation: the Equivalent Spline Smoothing Method.

Publications de l'Institute de Statistique Université de Paris, XXXVIII, 61-86

This filter function clearly satisfies conditions (i)-(iv) of section 3; hence its Fourier transform $K = \hat{\varphi}$ satisfies the condition of the theorem. We have

$$K(u) = \hat{\varphi}(u) = \frac{1}{2}\exp(-|u|)$$

so the double exponential density is the analog of Silverman's kernel K_S for $p = 1$. We conjecture that our main result can be generalized to arbitrary degree of smoothnes p and to a general limiting design density ℓ, provided the design tends to its limit sufficiently quickly. This is of course suggested by the results on the classical smoothing spline. We believe that more analytic results on the spectrum of differential operators and their approximation e. g. by Galerkin methods should be drawn upon for this. A useful reference is Chatelin (1983).

Let us stress again that so far our results did not involve stochastics, though they were obtained with a view to statistical smoothing. An interesting statistical result related to the subject of this paper was obtained by Cox (1984b). It was shown that the spline smoother applied to pure noise (i. e. to data ε_i) yields a random function on $[0, 1]$ which, when appropriately scaled, is close to a Gaussian process. This central limit theorem holds for general (nonnormal) noise distribution, and was used to show that the method of generalized cross-validation for choosing the smoothing parameter is asymptotically optimal. In turn, this study was motivated by a result of Speckman (1985) on the minimax linear spline, who established optimality of the bandwidth selector *under normality of the noise*. The normality assumption was removed by Cox (1984b), but the classical smoothing spline was substituted for the minimax linear one. Thus it appears a natural idea to generalize the limit theorem for spline estimators to our class. As the corresponding class of filters includes the Pinsker one, one should be able to infer optimality of the adaptive bandwidth choice for the minimax linear spline in the nongaussian case. This would complement a recently established lower asymptotic risk bound (see Golubev and Nussbaum,

(1994) Härdle, W. and Nussbaum, M.
Kernel Estimation: the Equivalent Spline Smoothing Method.

Publications de l'Institute de Statistique Université de Paris, XXXVIII, 61-86

1990), which showed that the minimax linear spline is a candidate for attainment also under nonnormal noise. That appears to be one way to confirm that this bound, which involves optimal rate *and constant*, is attainable adaptively by a spline estimator, without knowledge of the derivative bound Q and of the noise variance σ^2.

Thomas-Agnan (1991) defines a general class of spline-type smoothers, called α-splines, starting from the following observation. It is well known that in (1.1) the integral may be extended over $(0,1)$ or over the whole real line; in both cases the same spline minimizer results. If the whole line is used then (1.1) may be written in terms of the Fourier transform $\hat{g}$ of g

$$n^{-1}\sum_{i=1}^{n}(Y_i - g(x_i))^2 + \lambda \int \mid (2\pi t)^p \hat{g}(t) \mid^2 dt$$

Let α be a complex-valued function function defined on $\mathbb{R}$ fulfilling some regularity conditions; consider the minimizer g of

$$n^{-1}\sum_{i=1}^{n}(Y_i - g(x_i))^2 + \lambda \int \mid \alpha(t)(2\pi t)^p \hat{g}(t) \mid^2 dt$$

The solution is called an α-spline. For $\alpha \equiv 1$ and $p = 2$ one obtains the classical smoothing spline. The α-splines represent a large class of linear smoothers; in particular, they should be equivalent to kernel estimators. To see this heuristically, consider the corresponding continuous smoothing problem on the whole real line:

$$\int (m(x) - g(x))^2 dx + \lambda \int \mid \alpha(t)(2\pi t)^p \hat{g}(t) \mid^2 dt$$

Substituting the first integral by $\int(\hat{m}(t) - \hat{g}(t))^2 dt$ and arguing similarly to (1.2) we obtain a minimizer

$$\hat{g}(t) = (1 + \lambda\alpha(t)(2\pi t)^p)^{-1}\hat{m}(t) \qquad (5.1).$$

(1994) Härdle, W. and Nussbaum, M.
Kernel Estimation: the Equivalent Spline Smoothing Method.

Publications de l'Institute de Statistique Université de Paris, XXXVIII, 61-86

The Fourier transform expression for a general kernel smoother on the whole real line would be, using a filter function φ and bandwidth parameter h as before,

$$\hat{g}(t) = \varphi(ht/2)\hat{m}(t). \tag{5.2}$$

A choice $\alpha(t) = (2\pi ht)^{-p}\varphi^{-1}(ht/2) - 1, \lambda = h^p$ yields equality of (5.1) and (5.2). Though in the original concept α was assumed fixed, we see that a bandwidth-dependent choice of α makes the method sufficiently flexible to yield a spline-type optimization problem corresponding to the general kernel estimator. It is not essential in this connection that the α-splines are not necessarily polynomial splines. Thomas-Agnan (1991) discusses solution of the optimization problem via reproducing kernel Hilbert space methods. A rigorous proof of equivalence in the sense considered in this paper might be easier than for our estimator since Fourier transform methods are more directly at hand.

(1994) Härdle, W. and Nussbaum, M.
Kernel Estimation: the Equivalent Spline Smoothing Method.

Publications de l'Institute de Statistique Université de Paris, XXXVIII, 61-86

References

Agmon, S. (1968). Asymptotic formulas with remainder estimates for eigenvalues of elliptic operators. *Arch. Rational Mech. Anal. , 28, 165–183.*

Chatelin, F. (1983). *Spectral approximation to linear operators. Academic Press, New York.*

Cox, D. D. (1983). Asymptotics for $M-$type smoothing splines. *Annals of Statistics, 11, 530–551.*

Cox, D. D. (1984a). Multivariate smoothing spline functions. *SIAM J. Numer. Anal. 21, 789-813*

Cox, D. D. (1984b). Gaussian approximation of smoothing splines. *manuscript*

Cox, D. D. (1988). Approximation of method of regularization estimators. *Annals of Statistics, 16, 694-712*

Golubev, G. K. and Nussbaum, M. (1990). A risk bound in Sobolev class regression. *Annals of Statistics, 18, 758–778.*

Härdle, W. (1990). *Applied Nonparametric Regression. Cambridge University Press, Econometric Society Monograph Series 19.*

Huber, P.J. (1979). Robust smoothing. *in: Robustness in Statistics (E.Launer and G.Wilkinson, eds.) Academic Press, New York.*

Laurent, P. (1972). *Approximation et Optimisation. Hermann, Paris.*

Nussbaum, M. (1985). Spline smoothing in regression models and asymptotic efficiency in L_2. *Annals of Statistics, 13, 984–997.*

Messer, K. (1991). A comparison of a spline estimate to its "equivalent" kernel estimate. *Annals of Statistics, 19, 817–829.*

Messer, K. and Goldstein, L. (1993). A new class of kernels for nonparametric curve estimation. *Annals of Statistics, 21, 179–195.*

Pinsker, M. S (1980). Optimal filtering of square integrable signals in Gaussian white noise. *Problems Inform. Transmission, 16, No. 2, 52-68.*

Silverman, B.W. (1984). Spline smoothing : the equivalent variable kernel method. *Annals of Statistics, 12, 898–916.*

(1994) Härdle, W. and Nussbaum, M.
Kernel Estimation: the Equivalent Spline Smoothing Method.

Publications de l'Institute de Statistique Université de Paris, XXXVIII, 61-86

Speckman, P. (1985). Spline smoothing and optimal rates of convergence in nonparametric regression models. *Annals of Statistics, 13, 970–983.*

Thomas-Agnan, C. (1991). Spline functions and stochastic filtering. *Annals of Statistics, 19, 1512–1527.*

Triebel, H. (1972). *Höhere Analysis. Deutscher Verlag der Wissenschaften, Berlin.*

Utreras, F. (1983). Natural spline functions, their associated eigenvalue problem. *Numer. Math. 42, 107–118.*

(1994) Härdle, W. and Nussbaum, M.
Kernel Estimation: the Equivalent Spline Smoothing Method.

English translation of "Ob Effektivnom ozenivani usrednennoi proiswodnoi"). Russian Acad. Sci. Dokl. Math. Vol. 48, No.1, 148-152.

ON EFFICIENT ESTIMATION OF AN AVERAGED DERIVATIVE

UDC 519.2

I. A. IBRAGIMOV, V. KHERDLE [W. HÄRDLE], AND A. B. TSYBAKOV

1

Let (X, Y) be a random vector taking values in R^{d+1} and having probability distribution density $h(x, y)$ with respect to $\lambda \times \mu$, where λ is Lebesgue measure in R^d. We let

$$m(x) = \mathbf{E}\{Y|X = x\}$$

and define the *averaged derivative* $\delta(1)$ of the function m by

$$\delta(1) = \mathbf{E}m'(X)$$

and the *weighted averaged derivative* by

$$\delta(w) = \mathbf{E}m'(X)w(X).$$

Below we consider the problem of estimating the quantity $\delta = \delta(f)$; see [1]–[3] about estimation of $\delta(w)$ and the origin of the problem. We are interested first and foremost in the question of a priori assumptions about h that make possible asymptotically efficient estimation of δ; here asymptotic efficiency is understood as defined in [4].

Passing to a more precise formulation of the problem, we use the following notation: $H(\beta, L)$ is the class of $L_2(R^d)$-functions $\varphi(x_1, \ldots, x_d)$ that have with respect to each variable x_j a derivative in $L_2(R^d)$ of order $r = [\beta]$, or $r = \beta - 1$ if β is an integer, that satisfies a Hölder condition of order $\alpha = \beta - r$ in L_2 with constant L. Denote by $DH(\beta, L)$ the class of densities $h(x, y)$ for which

1) the functions

$$f(x) = \int h(x, y)\mu(dy), \qquad f(x)m(x) = \int yh(x, y)\mu(dy)$$

belong to $H(\beta + 1, L)$, and

2) $\sigma f' \in L_2(R^d)$, where $\sigma^2(x) = \mathrm{Var}(Y|X = x)$. Here and below, $\mathrm{Var}\,\xi$ denotes the variance of ξ.

We consider the problem of estimating the averaged derivative

$$\delta = \mathbf{E}_h m'(X)f(X) = -2\mathbf{E}_h Y f'(X) \tag{1}$$

from observations $(X_1, Y_1), \ldots, (X_n, Y_n)$ under the assumption that the unknown density h belongs to a known set Θ of densities.

Theorem 1. *Assume that $\Theta \subseteq DH(\beta, L)$ and $\beta > d/4$. Then there exists a sequence of estimators $\hat{\delta}_n$ of δ such that the normalized sequence $\sqrt{n}(\hat{\delta}_n - \delta)$ is asymptotically*

1991 *Mathematics Subject Classification.* Primary 62F12.

normal with mean zero and correlation matrix Σ, *where* Σ *is the correlation matrix of the vector*

$$\int (yf'(x) - (m(x)f(x))')\sqrt{h(x,y)}\,dw - \int (yf'(x) - (m(x)f(x))')\,d\lambda\,d\mu \int \sqrt{h(x,y)}\,dw,$$

and w *is the orthogonal Gaussian measure with* $\mathbf{E}|dw|^2 = d\lambda\,d\mu$. *Moreover,*

$$\lim_n n\mathbf{E}_h|\hat{\delta}_n - \delta|^2 = Q_1 = 4[\mathbf{E}_h|f(X)m'(X)|^2 - |\mathbf{E}_h(f(X)m'(X))|^2 + \mathbf{E}_h\sigma^2(X)|f'(X)|^2]$$

uniformly in Θ.

From the point of view of [4] the problem of estimating δ is the problem of estimating the value $\Phi(h)$ of the functional

$$\Phi(h) = \mathbf{E}_h\{m'(X)\cdot f(X)\} = \int\left\{\int yh'_x(x,y)\,d\mu - \frac{\int yh(x,y)\,d\mu}{\int h(x,y)\,d\mu}\int h'_x(x,y)\,d\mu\right\}h(x,y)\,d\lambda\,d\mu,$$

and in view of Theorem 5.2 in [2] (see also Example 5.2) the estimator $\hat{\delta}_n$ in Theorem 1 is asymptotically efficient.

Theorem 2. *Assume that* $\Theta = DH(\beta, L)$, $\beta < d/4$. *Then*

$$\underline{\lim_n}\inf_{\delta_n}\sup_h n\mathbf{E}_h|\delta_n - \delta|^2 = \infty.$$

The scheme of proof of Theorems 1 and 2 is presented below.

2

We let $D_\nu(x)$ be the Dirichlet kernel

$$D_\nu(x) = \pi^{-d}\prod_1^d \frac{\sin \nu_j x_j}{x_j}, \qquad x = (x_1, \ldots, x_d),\ \nu = (\nu_1, \ldots, \nu_d),$$

and, starting from (1), we consider the estimators

$$\delta_n(\nu) = -2n^{-1}\sum_1^n Y_j\hat{f}'_n(X_j), \qquad \hat{f}'_n(x) = n^{-1}\sum D'_\nu(x - X_j). \tag{2}$$

Denote by $\mathscr{E}_\nu(f)$ the best approximation in L_2 of a function f by entire functions of degree at most $\nu = (\nu_1, \ldots, \nu_n)$ (see [5]).

Theorem 3. *The size of the bias of the estimators* δ_n *defined in* (2) *is*

$$|\delta - \mathbf{E}\delta_n(\nu)| \le 2[\mathscr{E}_\nu(f')\mathscr{E}_\nu(mf) + n^{-1}\|mf\|\cdot\|f'\|].$$

The proof is by direct computation, with use of the L_2-orthogonality of φ_ν and $\psi - \psi_\nu$, where φ_ν denotes the Dirichlet integral of the function φ.

Theorem 4. *If* $\Theta \subseteq DH(\beta, L)$, *then*

$$\operatorname{Var}\delta_n(\nu) = \frac{Q_1}{n} + O(\nu_1\cdots\nu_d\cdot n^{-2}\sum\nu_j^2) + o(n^{-1}),$$

where the constants in O *and* o *depend only on* β, L, *and* $\int\sigma^2(x)|f'(x)|^2\,d\lambda$.

The proof is by direct computation. It follows from Theorems 1 and 2 that, choosing $\nu_j = \exp\{2[(2\beta+1)+d+2]^{-1}\ln n\}$ and letting $\hat{\delta}_n$ be the estimator $\delta_n(\nu)$ with this ν, we get for $\beta > d/4$ that

$$\operatorname{Var}\hat{\delta}_n = n^{-1}Q_1 + o(n^{-1}).$$

Theorem 5. *Under the conditions of Theorem 1 the difference $\sqrt{n}(\hat{\delta}_n - \delta)$ is asymptotically normal with mean zero and correlation matrix Σ.*

Proof. The asymptotic behavior of the correlation matrix of the vector $\sqrt{n}(\hat{\delta}_n - \delta)$ can be investigated by direct computation as in the preceding theorem. Further, the $\hat{\delta}_n$ are U-statistics, and to prove the asymptotic normality of $\sqrt{n}(\hat{\delta}_n - \delta)$ we can turn to general limit theorems for U-statistics (see [6]).

Theorem 1 obviously follows from Theorems 3–5.

3

Here we indicate a scheme for proving Theorem 2. Let Γ be the cube $|x_j| \le 1/2$, $j = 1, \dots, d$, $|y| \le 1/2$, and consider vectors (X, Y) with distribution density of the form

$$h(x, y) = \begin{cases} 1 + s(x, y), & (x, y) \in \Gamma, \\ 0, & (x, y) \notin \Gamma, \end{cases} \tag{3}$$

where

$$\begin{aligned} s(x, y) = &\sum_{\nu_1, \dots, \nu_d = 2}^{N} a_{\nu 0} \cos 2\pi\nu_1 x_1 \cdots \cos 2\pi\nu_d x_d \\ &+ \sum_{j=1}^{d} \sum_{\nu_1, \dots, \nu_d = 1}^{N} a_{\nu j} \cos 2\pi\nu_1 x_1 \cdots \sin 2\pi\nu_j x_j \cdots \cos 2\pi\nu_d x_d \sin 2\pi y. \end{aligned} \tag{4}$$

For such densities $\delta = \mathbf{E}m'(X)f(X) = -2\mathbf{E}Yf'(X)$ is the vector with components $2^{-d+1}\sum \nu_j a_{\nu 0} a_{\nu j}$. Further, let ζ and ξ_ν, $\nu = (\nu_1, \dots, \nu_d)$, $\nu_j = 1, \dots, N$, be independent random variables, with ζ taking the values 0 and 1 with equal probabilities, and ξ_ν the values ± 1. Let ρ_0 and ρ_1 be positive constants. We define random functions $h(x, y)$ by (3) and (4), setting $a_{\nu 0} = \rho_0 \zeta \xi_\nu$ and $a_{\nu j} = \rho_1 \zeta \xi_\nu$. Denote by I the indicator set of the random event $G = \{h(x, y) \ge 0\}$, so that $h(x, y)$ is a distribution density on G.

The vector δ is either zero or the vector with components $2^{-d}\rho_0\rho_1 N(N+1)$, depending on whether ζ takes the value 0 or 1. Therefore, it suffices to consider estimators $\hat{\delta}$ taking only these two values. Let $\mathbf{E}(\cdot)$ denote the expectation with respect to (ζ, ξ_ν). Let

$$\hat{\delta} = \begin{cases} 0 & \text{if } Z = (X_1, Y_1, \dots, X_n, Y_n) \in A, \\ (\dots, 2^{-d}\rho_0\rho_1 N(N+1), \dots) & \text{if } Z \notin A. \end{cases}$$

Then

$$\sup_h \mathbf{E}_h|\delta - \hat{\delta}|^2 \ge cN^4\rho_0^2\rho_1^2(\operatorname{meas}\overline{A} + \mathbf{E}\left\{I \cdot \int_A \prod_1^n (1 + s(x_i, y_i))\, dx_1 \cdots dy_n\right\}; \tag{5}$$

here and below, c denotes strictly positive constants. Further analysis of (5) is based on the following assertions.

Lemma 1. *Let* $\rho_0 \le \rho_1 = n^{-\gamma}$, $\gamma > 1$, *and* $N = n^{2\gamma'}$, $\gamma' < \gamma$. *If* meas $A \ge 3/4$, *then*

$$\mathbf{E}\int_A \prod_1^n (1 + s(x_i, y_i))\, dx_1 \cdots dy_n \ge c.$$

Lemma 2. *Let* ρ_0, ρ_1, *and* N *be the same as above. Then*

$$\mathbf{P}\left(\left|\int_A \prod_1^n (1 + s(x_i, y_i))\, dx_1 \cdots dy_n\right| > L\right) \le BL^{-4}.$$

Lemma 3. *Let* ρ_0, ρ_1, *and* N *be the same as above. Then*

$$\mathbf{P}\{\|s\|_\infty \ge 1\} \le Be^{-n^c}.$$

Considering all possible realizations of the sequence $(\zeta, \xi_\nu) \in G$, we get some (finite) set Θ_1 of densities, and by (5) and Lemmas 1–3,

$$\inf_{\hat\delta} \sup_{h\in\Theta_1} \mathbf{E}_h|\hat\delta - \delta|^2 \ge cN^4\rho_0^2\rho_1^2.$$

Simple computations now show that if $\beta < d/4$, then it is possible to choose N, ρ_0, and ρ_1 such that $N^4\rho_0^2\rho_1^2 > n^{-\alpha}$, $\alpha > 1$, and

$$\sum_j \sum \nu_j^{2(\beta+1)} |a_\nu|^2 \le B.$$

Multiplying that h defined above by a suitably chosen factor annihilating the jumps of h, we convert Θ_1 into $\Theta_2 \subset \Theta = LH_\beta^2$, which proves Theorem 2.

4

Of course, instead of the Dirichlet kernel it is possible to use other kernels for the construction of estimators. For example, the Vallée-Poussin kernel leads to similar results, the Fejér kernel can give a worse bias in certain situations, and so on. Below we present a result showing the dependence of $\mathbf{E}|\hat\delta - \delta|^2$ on the kernel. Let $\mathscr{K}_l$ be the class of kernels $R(x) = \prod_1^d K(x_j)$, where the function K is continuously differentiable and has support in $[-1, 1]$, and $K'(0) = 0$. Moreover,

$$\int K(u)\, du = 1, \qquad \int u^j K(u)\, du = 0, \quad j = 1, \ldots, l-1, \qquad \int u^l K(u)\, du \ne 0.$$

Theorem 6. *Assume that*: 1) *the density* $f(x)$ *of the variables* X_j *has compact support and is continuously differentiable* $l + 1$ *times with respect to each argument*; 2) *the regression function* $m(x)$ *is continuously differentiable* $l + 1$ *times with respect to each argument*; *and* 3) *the conditional variance* $\sigma^2(x)$ *is bounded on the support of* f. *Define the kernel estimator*

$$\hat\delta_n = -2n^{-1}\sum_{i=1}^n Y_i \hat f_h'(X_i), \qquad f_h'(x) = n^{-1}\sum_1^n R_h'(x - X_i)$$

of δ, *where* $R \in \mathscr{K}_l$, $R_h(x) = h^{-d}R(xh^{-1})$, $h = h_n \to 0$, *and* $n_2 h_n^{d+2} \to \infty$. *Then*

$$\mathbf{E}|\hat\delta_n - \delta|^2 = n^{-1}Q_1 + n^{-2}h^{-d-2}Q_2 + h^{2l}Q_3 + o(n^{-2}h^{-d-2} + h^{2l} + n^{-1}).$$

Here Q_l is defined in §2, and

$$Q_2 = 4C(K)\int \sigma^2(x)f^2(x)\,dx\,, \qquad Q_3 = 4\left|\int S_K(x)f(x)m(x)\,dx\right|^2\,,$$

$$C(K) = d\int (K'(u))^2\,du\left(\int K^2(u)\,du\right)^{d-1}\,,$$

$$S_k(x) = d_K\frac{(-1)^l}{l!}\sum_1^d \begin{pmatrix} \partial^{l+1}f/\partial x_1\partial x_j^l \\ \cdots\cdots\cdots\cdots \\ \partial^{l+1}f/\partial x_d\partial x_j^l \end{pmatrix}\,.$$

This work was carried out while the first and third authors were at the Center for Operations Research and Econometrics (CORE) of Louvain University in Belgium. Both of these authors are very grateful to the Center administration for the excellent conditions provided to them for this work.

Bibliography

1. Wolfgang Härdle and Thomas M. Stoker, J. Amer. Statist. Assoc. **84** (1989), 986–995.
2. Thomas M. Stoker, Econometrica **54** (1986), 1461–1481.
3. James L. Powell, James H. Stock, and Thomas M. Stoker, Econometrica **57** (1989), 1403–1430.
4. I. A. Ibragimov and R. Z. Khas'minskii, Ann. Statist. **19** (1991), 1681–1724.
5. S. M. Nikol'skii, *Approximation of functions of several variables and embedding theorems*, "Nauka", Moscow, 1969; English transl., Springer-Verlag, Berlin, 1975.
6. V. S. Korolyuk and Yu. Borovskikh, *Martingale approximation*, "Naukova Dumka", Kiev, 1988. (Russian)

St. Petersburg Branch, Steklov Mathematical Institute, Russian Academy of Sciences

Center for Operations Research and Econometrics (CORE), Louvain University, Belgium

Institute of Information Transfer Problems, Russian Academy of Sciences, Moscow

Received 18/JAN/93

Translated by H. H. McFADEN

Computational Statistics and Data Analysis, 20, 1-18.

Fast and Simple Scatterplot Smoothing

by

W. Härdle* and J.S. Marron**

March 1994

Abstract

An important element of both exploratory data analysis and many dimensionality reduction techniques is a scatterplot smoother. In both areas there is a strong need for fast and simple procedures. A major hurdle is choice of the amount of smoothing. In this paper we propose a fast and simple choice of the scatterplot smoothing parameter, based on blockwise least squares parabolic fitting. Our method provides both global and location adaptive smoothing methods. The "local method" meets the needs of different trade offs between variability of the errors and curvature of the underlying regression function, as tempered by the design.

* Institut für Statistik und Ökonometrie , Wirtschaftswissenschaftliche Fakultät, Spandauer Str. 1, Humboldt-Universität zu Berlin

** CORE, Université Catholique de Louvain and on leave of Department of Statistics, University of North Carolina,
Research partially supported by NSF Grant DMS-9203|35

Computational Statistics and Data Analysis, 20, 1-18.

Fast and Simple Scatterplot Smoothing

by W. Härdle and S. Marron

1. INTRODUCTION

Scatterplot smoothing is a very useful exploratory tool for data analysis as well as an essential ingredient in many new fitting techniques for high dimensional data. In projection pursuit, single index models and generalized additive models among others, the basis of the model is composed of low dimensional smooth functions. Applications of smoothing methods in both of these areas have the need for fast and simple techniques. A first attempt at meeting these goals in the related area of density estimation were the "rules of thumb" described in Silverman (1986: Section 3.4.2).

Speed is important for exploratory analysis since a too long lag between conception of the ideas of the experimenter and their visual realisation hinders the analytic process. Speed becomes an essential element in additive modelling and related methods because of their use of massive iterations of scatterplot smoothing, see the recent monograph of Hastie and Tibshirani (1991).

Simplicity is always desirable as a general feature. In smoothing this manifests itself through ease of implementation and direct understandable interpretation. One benefit of the ease of implementation is that smoothing methods become widely accessible. Another benefit is that the implementation of highly efficient algorithms (Scott, 1985; Silverman, 1982) is more straightforward. Ease of interpretation is essential to any data analyst who wants proper insights into the data.

All scatterplot smoothing methods amount to local averaging of the response variables. The kernel method is perhaps the simplest and most easily understood local averaging method. Competing methods have their advantages (Eubank, 1988; Wahba, 1991) but often lack simplicity and ease of computation. For example the best intuitive insight into the smoothing spline method comes from the asymptotic approximation by a kernel smoother, Silverman (1984).

We propose a kernel smoother with data dependent smoothing parameter. The first method uses the same bandwidth for each location and hence is termed "global".

The second uses a location adaptive bandwidth and is called "local". Both methods are based on "plug-in" ideas, i.e. the technique of estimating unknowns in an asymptotic formula for the mean squared error optimal smoothing parameter. The needed unknown quantities are the curvature, the residual variance and the marginal density of the design variables. We propose simple estimates of these. One possibility is to estimate the regression function by a *high degree* polynomial. Another is based on partitioning the design space into disjoint blocks and fitting a *low degree* polynomial in each block.

For the latter approach the regression function is approximated via a least squares parabolic fit within each block. The smoothing bias is then approximated by an appropriate functional of this parabola and the design distribution. The error variances in each block are estimated via the residual variances from the local parabolic fit. The global smoothing parameter is constructed by combining blockwise variances and squared biases to form an estimate of the globally optimal smoothing parameter. The local smoothing parameter function is taken as a smooth of blockwise estimates of the bandwidths.

An application of these methods is illustrated in Figure 1 with the motorcycle data set of Schmidt, Mattern and Schüler (1981) also analysed by Silverman (1986). The observations measure acceleration (in g) as a function of time (in milliseconds).

Figure 1. Motorcycle data set with global smooth (Dashed-Dotted line), local smooth (solid line). Raw data with blockwise parabolic fits shown in lower right. Local bandwidth in upper left (numbers are explained in section 3.

In this example we use four blocks. The picture in the lower right corner called the "raw data plot" shows the four piecewise parabolic fits together with the raw data. The main picture shows the global (dotted and dashed) and the local (solid) smooth. For easy comparison the horizontal bars in the lower right picture correspond to the top and bottom of the main figure. The important differences between the global and local fits occur in the main valley at x roughly equal to 22 and on the right of the following peak. Because the global method uses the same bandwidth at each location it oversmooths in the valley where a small bandwidth is more appropriate. The local smooth gives better performance in both of these areas. The reason becomes apparent from the picture in the upper left corner called the "bandwidth plot". This picture

shows the bandwidth function used at each location. The global method uses the same bandwidth at each location which appears as a constant function (dotted and dashed). The bandwidth function used for the local method is shown as a solid curve which is a kernel smooth of the closely dotted step function. The heights of each step of the latter is the estimated optimal bandwidth in each block. Note that the local bandwidth is small in the neighborhood of the valley and much larger on the right side which explains the improvement of the local over the global method.

2. The blocking method

A convenient mathematical structure for understanding scatterplot smoothing starts with an i.i.d sequence $\{X_i\}_{i=1}^n$ of design variables assumed to lie in an interval $[a, b]$. The response variables $\{Y_i\}_{i=1}^n$ are of the form:

$$Y_i = m(X_i) + \varepsilon_i, \qquad i = 1, \ldots, n$$

with independent errors $\{\varepsilon_i\}_{i=1}^n$ and a smooth regression function $m(x)$. Scatterplot smoothing may be viewed as an attempt to estimate the function m. We illustrate our ideas in the context of the Nadaraya–Watson kernel smoother, defined by

$$\hat{m}_h(x) = n^{-1} \sum_{i=1}^{n} K_h(x - X_i) Y_i / n^{-1} \sum_{i=1}^{n} K_h(x - X_i) \tag{2.1}$$

where $K_h(\bullet) = h^{-1} K(\bullet / h)$ is the rescaled kernel function K with bandwidth h. It is straightforward and in fact simpler, to apply our ideas to other smoothers, e.g. local linear methods. For sensible conditions on the kernel function see Härdle (1990). Behavior of the scatterplot smoother is crucially dependent on the choice of h. When the bandwidth is too small the resulting estimate is too wiggly, when the bandwidth is too big important features of the data are smoothed away. A simple and useful quantification of this tradeoff follows from analysis of the asymptotic mean integrated squared error. The variance of the kernel smoother $\hat{m}_h(x)$ is approximated by

$$n^{-1} h^{-1} V(x) = n^{-1} h^{-1} \int K^2(u) du \frac{\sigma^2(x)}{f(x)} \tag{2.2}$$

where $\sigma^2(x)$ denotes the variance function $E(Y^2 \mid x) - m^2(x)$, and $f(x)$ is the marginal density of the X variables. The bias is approximated by

$$h^2 B(x) = \frac{h^2}{2} \int u^2 K(u) du \left[m''(x) + 2 \frac{m'(x) f'(x)}{f(x)} \right]. \tag{2.3}$$

This indicates that the best bandwidth at each location x is well represented by

$$h_0(x) = n^{-1/5}\left[\frac{V(x)}{4B^2(x)}\right]^{1/5}$$

A good global bandwidth, i.e. one suitable in an average sense, is given by

$$h_1 = n^{-1/5}\left[\frac{\int V(x)}{4\int B^2(x)}\right]^{1/5} \tag{2.4}$$

This bandwidth is obtained by minimizing the approximate integrated mean squared error, $n^{-1}h^{-1}\int V(x)dx + h^4\int B^2(x)dx$.

2.1 Blocks for a constant global smoothing parameter

Practical use of the above representation for h_1 in formula (2.4) requires estimates of $\int V(x)$ and $\int B^2(x)$, which in turn can be built up, using formulas (2.2) and (2.3), from estimates of $m(x)$ and $f(x)$. Weighted versions could be used as well. In particular $\int V(x)f(x)dx$ and $\int B^2(x)f(x)dx$ would be easier to estimate below. Classical fast and simple estimates of these functions are based on polynomials and histograms respectively.

Histograms are constructed by first partitioning the design interval $[a, b]$ into blocks $\mathcal{B}_j, \quad j = 1, \ldots, N$. We had good success with N quite small, often 3 or 4. This could probably be improved with larger N, but data based choice of N would no longer give a simple method. For simplicity we work explicitly here with equal length intervals

$$\mathcal{B}_j = \left[a + \frac{(j-1)(b-a)}{N},\ a + \frac{j(b-a)}{N}\right].$$

Let $\mathcal{B}$ denote a generic block $\mathcal{B}_j$, and r and l denote right and left boundaries of this block. The proportion of X_i falling in each interval reflects the height of the density near the center of the block (bin). Let $c = \frac{r+l}{2}$ denote the blockcenter and $r_b = \frac{r-l}{2}$ denote the block radius. The histogram density estimate is

$$\hat{f}(c) = \frac{1}{2nr_b}\sum_{i=1}^{n} I\left(|c - X_i| \le r_b\right). \tag{2.5}$$

To estimate the derivative of f on $\mathcal{B}$ we use a simple differencing method. This requires two estimates of f so we split the block into two left and right halves. Denote the

frequencies on each half of the block by

$$n_l = \sum_{i=1}^{n} I(l \le X_i < c),$$

$$n_r = \sum_{i=1}^{n} I(c \le X_i < r).$$

Forming a difference quotient based on histograms at the center of the two subblocks gives the derivative estimate

$$\hat{f}'(c) = \frac{(n_r - n_l)/(nr_b)}{r_b}. \tag{2.6}$$

These two estimates are combined into the score function estimate

$$\widehat{\left(\frac{f'}{f}\right)}(c) = \frac{2(n_r - n_l)}{r_b(n_r + n_l)}. \tag{2.7}$$

The score function together with estimates of m' and m'' are used to construct an estimate of $\int B^2(x)$. The estimation of $V(x)$ in (2.4) is constructed from a sum of squared residuals (RSS) about an estimate $\hat{m}(x)$ of $m(x)$, normalized by an estimate of f. In particular, in the generic block $\mathcal{B}$ define

$$RSS = \sum_{X_i \in \mathcal{B}} (Y_i - \hat{m}(X_i))^2.$$

An estimate of $\left(\frac{\sigma^2}{f}\right)(c)$ is then given by

$$\widehat{\left(\frac{\sigma^2}{f}\right)}(c) = \frac{2nr_b}{(n_l + n_r)^2} RSS,$$

which leads to

$$\hat{V}(c) = \frac{5}{7}\widehat{\left(\frac{\sigma^2}{f}\right)}(c) \tag{2.8}$$

for the quartic kernel,

$$K(u) = \frac{15}{16}(1 - u^2)^2 I(|u| \le 1). \tag{2.9}$$

This kernel satisfies

$$\begin{aligned} \int K^2(u)du &= 5/7 \\ \int u^2 K(u)du &= 1/7 \end{aligned} \tag{2.10}$$

A natural and straight forward method for estimating m, m' and m'' is least-squares polynomial regression. The simplest version of this is a parabola but this often does not have sufficient flexibility, with the ability to model different curvature at different locations. This is clearly seen from Figure 2a where we show a kernel estimate for the motorcycle data with bandwidth based on a parabolic fit shown in the raw data plot in the lower right inset of this picture. Density estimates for (2.7) are constructed with $N = 1$.

Figure 2a. Motorcycle data set with a quartic kernel estimate. Bandwidth chosen by formula (2.4) based on parabolic regression estimate shown in right inset.

Note that the parabolic pilot estimate provides a very poor reflection of the curvature of the underlying function. For example the curvature of the valley near time 20 is drastically underestimated while the estimated curvature of the initial flat segment is too large. A consequence of the poor fit of this model is the estimated variance is extremely large and the overall estimated bias is relatively small. So our automatic global bandwidth of 24.49, used to construct the main curve estimate (Dashed-Dotted line), is way too large.

One means of addressing this problem is the use of a higher degree polynomial. Figure 2b shows the application of a fifth degree polynomial to the motorcycle data.

Figure 2b. Motorcycle data set with a fifth degree polynomial pilot estimate. Bandwidth chosen by formula (2.4).

The added flexibility of the higher degree polynomial shown in the raw data plot is still not enough to effectively model the initial flat segment or the heavy curvature in the first valley. The net effect is similar to Figure 2a but to a lesser extent. Note that the resulting automatic bandwidth of 9.16 is still too large. For example the valley is not deep enough and the "corner" near time 14 is smoothed away. Increasing the polynomial degree to 16 gives a much better visual fit of the pilot polynomial in the raw data plot, with resulting improvement in the final kernel estimate as seen in Figure 2c.

Figure 2c. Motorcycle data set with a sixteenth degree polynomial pilot estimate. Bandwidth chosen by formula (2.4).

Despite the fact that there is always some degree which gives a good polynomial fit, we are not satisfied with this approach because of the necessity to choose the degree of the polynomial. Data based choice of this degree is possible (Shibata, 1981), but would leave the realm of fast and simple scatterplot smoothing.

An alternative approach, which duplicates the flexibility of a higher degree polynomial is to fit blockwise parabolas. We chose parabolas here because they are the simplest polynomials which allow nontrivial estimation of second derivatives. In a sense we are replacing the degree of the polynomial by the number of blocks. A rough correspondence between these approaches, based on degree of freedom considerations, is

$$(\text{degree of polynomial} + 1) = 3N.$$

This correspondence fits well with the relationship between Figure 1 and Figure 2c. Figure 1 uses 12 degrees of freedom (3 parameters to fit over the 4 blocks), which results in a slightly larger bandwidth than the one based on 17 degrees of freedom in Figure 2c. We also tried several other polynomial degrees. The resulting bandwidth and kernel smooth were virtually identical to those in Figure 1, for the polynomial of degree 8.

In our opinion the number of blocks is a more practical parameter to work with than high polynomial degrees for which we have also observed numerical instability. In particular the number of blocks is easily visually selected from pictures as in the lower right raw data plot inset in Figure 1. When the number of blocks is inappropriate this type of figure usually suggests a more sensible number of blocks. On the other hand pictures like the raw data plot in Figure 2a do not suggest an effective degree of polynomial.

The blockwise approach based on local parabolic pilot fits leads to the following estimate of $V = \int V(x)$,

$$\widehat{V} = \sum_{j=1}^{N} 2r_b \widehat{V}(c_j) \tag{2.11}$$

using notation from (2.8) and the quartic kernel given in (2.9). The bias $B(c_j)$ in

each block $\mathcal{B}$ is estimated via

$$\widehat{B}(c_j) = \frac{1}{2}\,\frac{1}{7}2\hat{\beta}_{3j} + 2\left[2\hat{\beta}_{3j}(c_j) + \hat{\beta}_{2j}\right]\widehat{\left(\frac{f'}{f}\right)}(c_j) \tag{2.12}$$

where $\hat{\beta}_{2j}, \hat{\beta}_{3j}$ are least squares estimates in the model

$$\beta_{1j} + \beta_{2j}\,x + \beta_{3j}\,x^2 \tag{2.13}$$

over block $\mathcal{B}_j$. Again here the quartic kernel has been used with the numbers given in (2.10). The final estimate of $B_2 = \int B^2(x)dx$ is obtained by summing up the squares of these quantities,

$$\widehat{B}_2 = \sum_{j=1}^{N} 2r_b\widehat{B}^2(c_j). \tag{2.14}$$

Putting things together leads then to

$$\hat{h}_1 = n^{-1/5}\left[\frac{\widehat{V}}{4\widehat{B}_2}\right]^{1/5}$$

A fast and simple alternative to the local parabolic fit is to consider zero-th, first and second order differences of a regressogram estimate of m, see Tukey (1961). We chose the local parabolic fit since estimation of m'' is crucial. The local parabola fit by least suares gives a better curvature estimate than differencing the regressogram. We use the histogram to estimate f and f',vs. also using a quadratic fit, for reasons of simplicity. Note: efficency is not so important here as it is in the estimation of m''.

2.2 Blocks for a local smoothing parameter function

An additional advantage of the blockwise parabola methodology for pilot estimation is that it naturally allows local smoothing. In particular the estimates of variance and squared bias over each block provide an easy bandwidth choice for that block, given by

$$\hat{h}_0(c) = n^{-1/5}\left[\frac{\widehat{V}(c)}{4\widehat{B}^2(c)}\right]^{1/5}, \tag{2.15}$$

using the notation from (2.8) and (2.12). The logs (base 2) of these values are the heights of the dotted stepfunction in the upper left inset of Figure 1. Note that the second block in this picture has a much lower bandwidth which is sensible because of the large amount of curvature and relatively small variability in this region. This

results in the valley being deeper for the local smooth as compared to the global smooth. The two blocks on the right have a relatively larger bandwidth which helps smooth away the spurious wiggles which appear in the global smooth. The bandwidth $\hat{h}_1$ is represented by the dotted and dashed constant function in this inset bandwidth plot. This provides a useful reference for understanding the relative sizes of the local bandwidths. Note that the average of the $\hat{h}_0(c_j)$ does *not* give the global bandwidth $\hat{h}_1$ because the variance and bias terms need to be summed separately for the latter.

The local bandwidth estimates are best in the centers of the blocks. For the points away from the centers we use a smooth, represented by the solid curve, of the step function. This smooth is computed on a fixed grid of x's by the formula (2.1) with the X_i replaced by the *bin* centers and the Y_i replaced by the height of the step function. The kernel used is the quartic see (2.9) and the bandwidth is the block radius r_b. This choice of bandwidth guarantees that the smooth coincides with the stepfunction at the points where it is most accurate, i.e. at the *bin* centers.

To gain further insight into this method we constructed some simulated examples. We generated $n = 400$ X_i-values uniformly distributed on $[-1, 1]$. The regression function was taken to be

$$m(x) = x + \alpha_1 \exp(-\alpha_2 x^2)$$

for several values of the accentuation parameters α_1 and α_2. This family of regression functions represents a unimodel departure from the linear. The parameter α_1 controls the magnitude of the departure, whereas the parameter α_2 controls the width of the spike. The errors where chosen to be normal with variance function given by

$$\sigma^2(x) = \left(a_\sigma \left(|x| - \frac{1}{2} \right) + \sigma_0 \right)^2$$

for several values of a_σ and σ_0. Here the parameter a_σ controls the type and degree of heteroscedasticity, while σ_0 controls the overall variability. This parametrization of the variance function is convenient since $a_\sigma = 0$ corresponds to the homoscedastic case, $a_\sigma < 0$ means greater variability in the middle where m has more curvature, and $a_\sigma > 0$ moves the variability towards the boundaries where m has less curvature.

We considered this setup for a variety of choices of the parameters and present just two representative situations in Figure 3a,b. Both figures have $\alpha_1 = 1$, $\alpha_2 =$

60, $\sigma_0 = 0.4$ and $N = 5$. The inset pictures in the lower right for the raw data show a marked difference in heteroscedasticity, parametrized by $a_\sigma = -0.7$ for Figure 3a and $a_\sigma = 0.7$ for Figure 3b.

Figures 3a,b. Simulated data sets from the curve $m(x)$ (dotted). Raw data shown as circles together with blockwise parabolic fits in lower right insets. Local and global bandwidth shown in upper left inset. Resulting smooths shown in main picture.

In the setting of Figure 3a there is less need for local smoothing since the greater curvature of $m(x)$ near the center is balanced to some extent by the greater variability there. However the local smooth is still superior, most notably in terms of a smoother estimate in the regions where $m(x)$ is nearly linear. In Figure 3b the need for local smoothing is greater since there is less variablity in the central region where there is more curvature in $m(x)$. Our local method fulfills this need as illustrated by the upper left bandwidth plots, where the local bandwidth is smaller in the center and larger on the flanks. Note that the benefits of local smoothing are stronger here with marked improvement of the local over the global smooth in the estimation of the peak as well as the linear parts.

2.3 Speed considerations

Our scatterplot smoothing methodology consists of two essential steps. First we determine global and local bandwidths by simple and rapidly computable pilot estimates, see Appendix for details. The second step consists of smoothing the data using either the global bandwidth $\hat{h}_1$ or the location dependent $\hat{h}_0(x)$. This second step carries the larger share of the computational burden, especially when the data set is large and formula (2.1) is directly implemented.

A very effective practical means of reducing this computational burden is based on pre-binning the data, also called WARPing (Weighted Averaging of Rounded Points) in Härdle and Scott (1991). This technique consists of the following three steps

1. Discretize $\{(X_i, Y_i)\}_{i=1}^n$ into NB "small bins" of length δ by recording the frequencies of the X's, and the sum of the Y's in each bin.

2. Evaluate the kernel $K_h(u)$ at $u = 0, \delta, 2\delta, \ldots$

3. Convolute the discretized data with the discretized kernel.

A major source of the dramatic increase in speed provided by WARPing comes in Step 2. This is because only less then or equal to NB kernel evaluations are required, compared to n^2 (or $n^2 h$ for a compactly supported kernel) evaluations needed for a direct implementation of (2.1).

Another source of speed improvement comes from maintaining pointers to non empty bins. Empty bins that occur quite often for non uniform data need not be considered in the convolution step 3 in the algorithm above. This pointer structure and its application, is described in Härdle (1991, Section 5.1.5). Asymptotic analysis of the WARPed estimator is provided in Jones (1989).

3. Diagnostic Plots

The inset plots in each figure of this paper are intended to show visually how well our methods are performing. For example, the raw data-plots in Figures 2a and 2b clearly indicate that polynomials of degrees 2 and 5 respectively yield inappropriate final smooths. But the raw data-plots of Figure 2c shows the smooth there is better.

This raw data plot was also a useful diagnostic in the choice of $N = 4$ blocks for the motorcyle data example presented in Figure 1. Our initial choice of $N = 5$ blocks gave visually poor performance, because the "corner" near time 14 was in the interior of the second block. This gave a visually poor parabolic fit in that crucial time interval, which resulted in an oversmoothed global choice $\hat{h}_1$ and a less effective local bandwidth function, $\hat{h}_0(x)$. For these data correct estimation of this time location of the corner near time 14 is important for the resulting curve to properly reflect the physical phenomenon being studied, see Schmidt, Mattern and Schüler (1981).

The bandwidth plots enhance the understanding of the performance of the local smoothing method by showing the amount of smoothing done at each point. The effective bandwidth is shown on the log scale, because this parameter is multiplicative in character. Marron and Wand (1991) provide additional insight into why it is more practicable to consider adjusting bandwidth multiplicatively rather than additively.

A further diagnostic device, which we find useful, is to calculate the observed

significance level of the parabolic fit on each block. The numbers shown in the top part of the bandwidth plot are p-values for testing, within each block, the null hypothesis of linearity,

$$H_0 : \beta_{3,j} = 0$$

in the model (2.13). When these are small, there is strong evidence of curvature in the data, so our local bandwidth estimate should be reliable. Note that in most of our examples, the local method works well in many cases, even when the p-value is large see e.g. Figure 1,3a,b. However we have found this number to be quite useful for explaining phenomena such as that in Figure 4a.

Figure 4a. The food data set. Number of blocks $N = 3$. Quartic kernel. Display styles follow that of Figure 1.

This data stems from the Family Expediture Survey (1968-1983) as described in Härdle (1990, Chapter 3). In its original form it contains about 7000 observations for each year. For ease of computation we worked with a condensed version of this data set of $n = 727$. The interest of econometricians lies in estimating the mean expenditure for various goods as a function of $x =$ (total expenditure). The mean expenditure curve, also called the Engel curve, is the regression function $m(x)$ if Y denotes the expenditure for a specific good. In the above Figure we estimate the food Engel curve. Note the huge difference between the local (solid) and global (dotted-dashed) smooths is caused by the estimated bandwidth on the third block being larger than the X-range of the data. This results from being in a "low bias, high variance" situation which is reflected in the high p-value of 0.66.

One possibility to overcome the problem seen in Figure 4a is to combine the second and third block. In fact one could develop an automatic scheme for combining blocks, based on p-value, but this is not followed up here because it moves away from fast and simple methods. Another means of addressing the difficulty encountered in Figure 4a is to use a larger bandwidth in the smooth of the local bandwidth step function in the bandwidth plot, see Friedman (1984) for related ideas. This is not developed further here because this requires yet another decision on the amount of smoothing in the bandwidth plot.

Another application of our p-value diagnostic is given in Figure 4b where we show the fuel Engel curve.

Figure 4b. The fuel data set. Number of blocks $N = 3$. Quartic kernel. Display styles follow that of Figure 1.

No great importance should be attached to the bump in the local estimate near the center, because the large p-value of 0.63 on the second block indicates that the low bandwidth there could be due only to random fluctuations in the data. This effect in this case is due to the two very high data points visible in the raw data plot. Here again the two horizontal lines in the raw data plot are helpful. On the other hand the local bandwidth gives better performance on the first block, where the curvature estimate reflects actual structure, as indicated by the p-value of 0.00 in the bandwidth plot.

4. Number of blocks

Choice of the number of blocks is very important for application of our proposed methodology. This number is essentially another smoothing parameter, because one block usually gives an oversmoothed result, and many blocks result in undersmoothing.

Practical choice must depend on the particular situation. For exploratory analysis, where there is usually time for some interactive trial and error analysis, we recommend an initial attempt, with $N = 3$ for smaller data sets and $N = 5$ for larger ones. An important choice can be easily made, using some or all of the diagnostic ideas presented in the previous section. This is how we arrived at $N = 4$ in Figure 1. Since the process usually results in a satisfactory smooth after only one interation, we consider it to be an improvement over the several step trial and error, rather arbitrary, choice of bandwidth traditionally used, see section 3.4.1 of Silverman (1986). Another advantage is that less experienced data analysts should find it much easier to decide what is an effective smooth.

On the other hand, for situations requiring massive iterations of scatterplot smoothing such as application of dimensionality reduction methods, this interactive approach is not possible. However, the demands in terms of effective performance are usually much less stringent in these ceses. So we recommend simply using $N = 3$ for computational efficiency. Our experience indicates this will not always result in a visually pleasing smooth, but that is different goal from what is needed here. We

believe our methodology will do a better job of meeting the needs of massive iterations of scatterplot smoothing than either earlier fast and simple methods or else computationally intensive techniques (eg. cross validation, etc.)

5. Appendix

For straightforward application of our methods we now describe them in the easily implementable style of Kennedy and Gentle (1980). It is common to both the global and local bandwidth selection algorithm to partition the design interval into blocks. So we assume that the user provides the algorithm with an array `bgrid` indicating the block boundaries. Let `x` denote the vector of X variables, `y` the vector of Y response variables and `bgrid` the vector with values $l_1, l_2, \ldots, l_N, r_N$

Algorithm FASTH(x,y,bgrid)

Step	Description
1.	Set j=1, N=rows (bgrid)-1, n=rows(x). If j>N go to 13.
2.	Set l=bgrid[j], r=bgrid[j+1], c=(r+l)/2, rb=(r-l)/2.
3.	Set (xb,yb)=those (x,y)'s that lie in block j, xd= design matrix for model (2.13), xd1=design matrix with β_{3j}=0, degf=rows(xb)-3.
4.	Set temp1=xd1*inv(xd1'*xd1)*xd'*yb, bhat=inv(xd'*xd)*xd*yb, temp=xd*bhat.
5.	Set ehat=yb-temp temp1=(temp1-temp)'*(temp1-temp).
6.	Set fstat=degf*temp1/((ehat)'*ehat), pval[j]=F-distribution(fstat,1,degf), nl=number of xb's left of c, nr=number of xb's right of c.
7.	Set fhat=(nl+nr)/(2*n*rb), f1hat=(nr-nl)/(n*rb*rb); from (2.5) and (2.6).
8.	Set bmp=2*bhat[3]*c+bhat[2].
9.	Define blockwise squared bias: Set im2[j]=2*rb*(2*(1/7)*bhat[3] +bmp*(f1hat/fhat))∧2;

```
            from (2.12).

10.     Define blockwise variance:
            Set is2[j]=(5/7)*2*rb*ehat' *ehat/(rows(xb)*fhat)
            from (2.8).

11.     Determine local bandwidth:
            Set h0hat[j]=(is2[j]/(im2[j]*n))^0.2
            from(2.15).
            Store local parabolas in xpoly.

12.     Set j=j+1.  Goto step 1.

13.     Set h1hat=(sum(is2)/(sum(im2)*n))10.2.
            Return   (h1hat,h0hat,Xpoly,pval).
```

REFERENCES

Eubank, R.L. (1988). *Spline smoothing and nonparametric regression.* New York, Dekker.

Family Expenditure Survey, Annuals Base Tapes (1968-1983) Department of Employment, Statistics Division, Her Majesty's Stationery Office, London 1968-1983. The data utilized in this paper were made available by the ESRC Data Archive at the University of Essex.

Friedman, J. (1984). A variable span smoother. Department of Statistics Technical Report LCS5 , Stanford University, Stanford, CA.

Härdle, W. (1990). Applied nonparametric regression, *Econometrics Society Monograph Series*, N 19, Cambridge University Press, Cambridge.

Härdle, W. and Scott, D.W. (1992), Smoothing by weighted averaging of rounded points. *Computational Statistics,7,97-128.*

Härdle, W (1991). *Smoothing Techniques with implementation in S*, Springer-Verlag, New York.

Hastie, T.J. and Tibshirani, R.J. (1991). *Generalized additive models*, London: Chap-

man and Hall.

Jones, M.C. (1989), Discretized and interpolated kernel density estimates. *JASA*,84, 733-739.

Kennedy, W.J. and Gentle, J.E. (1980). *Statistical Computing*, New York: Marcel Dekker.

Marron, J.S. and Wand, M.P. (1991). Exact mean integrated squared error, *Annals of Statistics*,20,712-736.

Schmidt, G., Mattern, R. and Schüler, F. (1981). Biomechanical investigation to determine physical and traumatological differentiation criteria for the maximum load capacity of head and vertebral column with and without helmet under effects of impact. EEC Research Program on Biomechanics of Impacts. Final Report Phase III, Project 65, Institut für Rechtsmedizin, Universität Heidelberg, Germany.

Scott, D.W. (1985). Average shifted histogram: effective nonparametric density estimators in several dimensions. *Annals of Statistics*, 13, 1024-1040.

Shibata, R. (1981). An optimal selection of regression variables. *Biometrika*, 68, 45-54.

Silverman, B.W. (1982). Kernel density estimation using the Fast Fourier Transformation. *Applied Statistics*, 31, 93-7.

Silverman, B.W. (1984). Spline smoothing: the equivalent variable kernel method. *Annals of Statistics*, 12, 898-916.

Silverman, B.W. (1986). *Density estimation for statistics and data analysis*. London: Chapman and Hall.

Tukey, J.W. (1961). Curves as parameters and touch estimation. *Proc 4th Berkeley Symposium*, 681-94.

Wahba, G. (1991). *Spline Models for Observational Data*. CBMS-NSF Regional Conference Series in Applied Mathematics, SIAM Monograph Series 59.

ELSEVIER Computational Statistics & Data Analysis 19 (1995) 641–653

COMPUTATIONAL STATISTICS & DATA ANALYSIS

Testing increasing dispersion†

W. Härdle*
Humboldt-Universität zu Berlin, Germany

B.U. Park
Seoul National University, South Korea

Received February 1994

Abstract

Increasing dispersion in regression analysis means that with positive changes of the explanatory variable the residual variance increases. Motivated by theoretical questions in stability of demand systems and by bandwidth selection problem in kernel scatterplot smoothing, we consider the question of increasing dispersion in a nonparametric way. It amounts to testing the positive definiteness of differences of covariance matrices. The asymptotic distribution of the smallest eigenvalue of the estimator of this difference is rather complicated, and that is why we also apply bootstrapping. The proposed method is applied to family expenditure data from the United Kingdom.

1. Introduction

The residual pattern in regression analysis gives important information on the aptness of the model under investigation. Missing terms of a regression equation and heteroscedastic variance functions can be identified and estimated. Often one is interested in the variance function itself to understand the error structure. Estimation of variance functions in the context of linear models with unknown residual structure has been considered by Carroll (1982) and Robinson (1986) for two stage weighted least squares estimation of parameters. In these papers the residuals are computed from a parametric linear model. More generally the variance function may be computed from a nonparametric regression model, this is the situation we consider here. More precisely we are interested in increasing dispersion which is the property that the errors are more spread out as the value of the explanatory variable increases.

* Corresponding author.

† Research done while both authors were visiting CentER, Tilburg and Institut de Statistique and CORE at Université Catholique de Louvain.

SSDI 0167-9473(94)00021-A

Computational Statistics and Data Analysis, 19, 641-654

Suppose that we are given observations $\{X_i, \underline{Y}_i\}_{i=1}^n \in \mathbb{R}^{d+1}$, $\underline{Y} = (Y_1, \ldots, Y_d)$, independent and identically distributed with distribution function F and density f. If $d = 1$, then increasing dispersion describes a fan-shaped variance pattern of a point cloud. For a specific application we have in mind, we shall consider the general case $d \geq 1$. For a fixed x, define the conditional covariance $C(x) = \mathrm{cov}[\underline{Y} | X = x]$. The property of increasing dispersion that we want to consider is

$$C(x_2) - C(x_1) > 0 \text{ for given } x_1 \text{ and } x_2 \text{ with } x_2 > x_1. \tag{1.1}$$

Here "$A > 0$" means that the matrix A is positive definite. We shall also consider this property for the conditional second moment matrix $D(x) = E[\underline{Y}\underline{Y}^{\mathrm{T}} | X = x]$, i.e.

$$D(x_2) - D(x_1) > 0 \text{ for given } x_1 \text{ and } x_2 \text{ with } x_2 > x_1. \tag{1.2}$$

The motivation for consideration of these properties partly comes from theoretical work in economic demand theory. In connection with Hildenbrand's (1992) analysis of market demand, "increasing dispersion" of the form (1.1) and (1.2) is a key element for stability of demand systems, see also (Härdle et al., 1991) for details. Of course not only in this specific economic application, increasing dispersion is an interesting property to investigate in other contexts. For example, the mean squared errors of various scatterplot smoothers involve the residual variance function (see Jones, et al., 1994). Also the often used plug-in bandwidth selection rule requires estimation of this quantity, see (Härdle and Marron, 1993). At this stage, the estimation method would heavily depend on the variance pattern, a key question on which is whether changes in explanatory variable bring changes in variability. The problem of testing increasing dispersion also arises when one seeks better statistical analysis. In case of increasing dispersion, many standard statistical procedures are inappropriate. This difficulty may be eliminated by taking proper transformations on either the response or both the response and explanatory variable to promote stable dispersion. For detailed discussion on this issue, we refer to (Hoaglin et al., 1983).

How can we test the properties (1.1) and (1.2)? A matrix is positive definite if its smallest eigenvalue is greater than zero. So a natural way to approach this problem is to derive the distribution of the smallest eigenvalue of $\hat{C}(x_2) - \hat{C}(x_1)$ resp. $\hat{D}(x_2) - \hat{D}(x_1)$ where $\hat{C}(x)$ and $\hat{D}(x)$ are estimators of $C(x)$ and $D(x)$. A specific type of estimators investigated in this paper will be presented in the next section. Hypothesis of the form (1.1) or (1.2) can then be tested by plugging in the unknown parameters of the asymptotic distribution. An asymptotic analysis of the related problem on the average derivative of $D(x)$ has been carried out by Härdle and Hart (1992) where also the bootstrap technique is advocated to overcome the difficulties arising with the "plug-in technique". The complications arising in this approach come from the fact that the asymptotic variance of the desired distribution of the smallest eigenvalue involves complex unknowns that by itself have to be estimated. For this purpose one could employ in a two step approach once more a nonparametric estimation technique. Second-order properties then become questionable

(1995) Härdle, W. and Park, B. Testing Increasing Dispersion

Computational Statistics and Data Analysis, 19, 641-654

since a second smoothing parameter has to be estimated. That is why we also advocate the bootstrap technique which elegantly avoids the difficulties arising in plug-in technique.

In the next section we give the theoretical framework to test (1.1) and (1.2). In Section 3 we apply our bootstrap test to data[1] from the Family Expenditure Survey (1968–1983). The last section is devoted to proofs.

2. Tests on increasing dispersion

We begin by considering the testing problem (1.1). The theoretical framework to test (1.2) follows afterwards.

For estimation of the conditional variances

$$c_{\ell k}(x) = E[Y_\ell Y_k | X = x] - m_\ell(x) m_k(x) \tag{2.1}$$

with $m_\ell(x) = E[Y_\ell | X = x]$, the kernel method is applied. Let K denote a continuous symmetric kernel function integrating to one. An estimator for the marginal density $g(x)$ of X is then given by

$$\hat{g}_h(x) = n^{-1} \sum_{i=1}^{n} K_h(x - X_i) \tag{2.2}$$

where $K_h(\cdot) = h^{-1} K(\cdot / h)$ denotes the rescaled kernel function with bandwidth $h = h_n > 0$. For theoretical and practical properties of this kernel density estimator, we refer to (Silverman, 1986). The regression functions $m_\ell(x)$ in (2.1) can be estimated by the kernel method as well,

$$\hat{m}_{h,\ell}(x) = n^{-1} \sum_{i=1}^{n} K_h(x - X_i) Y_{i\ell} / \hat{g}_h(x). \tag{2.3}$$

For statistical details, in particular mean squared error expansions of this kernel regression smoother, see (Härdle, 1990, Ch. 3). Based on these estimators the conditional covariances can now be approximated by

$$\hat{c}_{\ell k}(x) = n^{-1} \sum_{i=1}^{n} K_h(x - X_i) Y_{i\ell} Y_{ik} / \hat{g}_h(x) - \hat{m}_{h,\ell}(x) \hat{m}_{h,k}(x). \tag{2.4}$$

Let $\hat{C}(x) = (\hat{c}_{\ell k}(x))$, let $\hat{\lambda}$ be the vector of eigenvalues (ordered in magnitude) of $C(x_2) - C(x_1)$, and define $\hat{\underline{\lambda}}$ correspondingly. Our testing procedure rejects the null hypothesis H_0: $\lambda_1 \leqslant 0$ in favor of H_1: $\lambda_1 > 0$ if $\hat{\lambda}_1$ is large enough. For this, first we derive the asymptotic distribution of $\hat{C}(x_2) - \hat{C}(x_1)$. From an analysis of variance of either $\hat{g}_h$ or $\hat{m}_{h,\ell}$ we see that, under smoothness assumptions on the moment

[1]The data utilized in this paper were made available by the ESRC Data Archive at the University of Essex.

(1995) Härdle, W. and Park, B. Testing Increasing Dispersion

Computational Statistics and Data Analysis, 19, 641-654

functions $E[Y_\ell Y_k | X = x]$, the variance is of order $(nh)^{-1}$. It is therefore reasonable to consider the distribution of

$$\sqrt{nh}[\text{uvec}(\hat{C}(x_2) - \hat{C}(x_1)) - \text{uvec}(C(x_2) - C(x_1))], \tag{2.5}$$

where, for $C = (c_{ij})$, the vector uvec(C) denotes the vectorized matrix of size $\frac{1}{2}d(d+1)$, i.e. $\text{uvec}(C) = (c_{11}c_{12}\cdots c_{1d}, c_{22}\cdots c_{2d}, \cdots, c_{dd})^T$. Even in the case of simple kernel regression smoothing, the bias is a complicated functional of the first and the second derivatives of the regression function and the marginal density. To avoid complication of this form we derive the limiting distribution in (2.5) without a bias term that would be reflected as a nonzero mean in the asymptotic normal distribution. This can be achieved by slightly "undersmoothing" the estimator, i.e., by letting $h = h_n$ tends to zero fast enough so that nh^5 tends to zero as well.

The writing of the asymptotic covariance involves some more notation. Let $\chi_{\ell k}(x) = (Y_\ell - m_\ell(x))(Y_k - m_k(x))$ and define

$$V_{\ell k,\ell' k'} = \int K^2(u)\,du[\text{cov}\{\chi_{\ell k}(x_1), \chi_{\ell' k'}(x_1) | X = x_1\}/g(x_1) + \text{cov}\{\chi_{\ell k}(x_2), \chi_{\ell' k'}(x_2) | X = x_2\}/g(x_2)]. \tag{2.6}$$

With these elements form the matrix $V = (V_{\ell k,\ell' k'})$ with $\frac{1}{2}d(d+1)$ columns and rows. The following set of assumptions is needed.

Assumptions

(A.1) The kernel $K(u)$ and $|uK(u)|$ are bounded.
(A.2) The first moment of the kernel is zero, $\int uK(u)\,du = 0$.
(A.3) The second kernel moment is finite, $\int u^2K(u)\,du < \infty$.
(A.4) Let $t_\ell(x) = m_\ell(x)g(x)$, $u_{\ell k}(x) = E[Y_\ell Y_k | X = x]g(x)$, $v_{\ell\ell' k}(x) = E[Y_\ell Y_{\ell'} Y_k | X = x]g(x)$, $w_{\ell\ell' kk'}(x) = E[Y_\ell Y_{\ell'} Y_k Y_{k'} | X = x]g(x)$.
The functions $g(x), t_\ell(x) u_{\ell k}(x), v_{\ell\ell' k}(x), w_{\ell\ell' kk'}(x)$ have bounded first derivatives.
(A.5) The density of X is positive in x_1, x_2: $g(x_1)g(x_2) > 0$
(A.6) $E[Y_\ell^{r_1} Y_k^{r_2}] < \infty$ for any ℓ, k and $r_1, r_2 = 0, 1, 2, 3$.
(A.7) $g''(x), t_\ell''(x), u_{\ell k}''(x)$ are bounded for any ℓ, k.
(A.8) The bandwidth h tends to zero in such a way that $nh^5 \to 0$ and $nh^3 \to \infty$.

Our first theorem states the asymptotic distribution of $\hat{C}(x_2) - \hat{C}(x_1)$.

Theorem 1. *Under the assumptions (A), as* $n \to \infty$,

$$\mathcal{L}\{\sqrt{nh}[\text{uvec}(\hat{C}(x_2) - \hat{C}(x_1)) - \text{uvec}(C(x_2) - C(x_1))] | F\} \to N(0, V).$$

In the statement of this theorem we have explicitly spelled out the distribution of F of the observations $\{(X_i, Y_i)\}_{i=1}^n$ since in a later theorem we employ the empirical distribution function $\hat{F}_n$ of the data to establish the same limit law for the bootstrap approximation.

Computational Statistics and Data Analysis, 19, 641-654

In order to formulate the limit result for the estimated eigenvalue vector we need some more notation. Let $D(\lambda) = \det(C(x_2) - C(x_1) - \lambda I)$ denote the characteristic polynomial of the difference matrix $C(x_2) - C(x_1)$ and define $\hat{D}(\lambda)$ accordingly. Let $B_{ij}(\lambda)$ be the cofactor of the (i,j)th element of $C(x_2) - C(x_1) - \lambda I$ and let

$$B(\lambda_k) = 2(B_{ij}(\lambda_k)) - \operatorname{diag}(B_{11}(\lambda_k), \ldots, B_{dd}(\lambda_k)). \tag{2.7}$$

With this matrix define

$$M = (\operatorname{uvec}(B(\lambda_1))/D'(\lambda_1), \ldots, \operatorname{uvec}(B(\lambda_d))/D'(\lambda_d)) \tag{2.8}$$

and $\Sigma = M^{\mathsf{T}} V M$. The asymptotic distribution of the eigenvalues is given in

Theorem 2. *Under the assumptions (A) and if all eigenvalues are distinct, then as* $n \to \infty$

$$\mathcal{L}\{\sqrt{nh}\,(\hat{\underset{\sim}{\lambda}} - \underset{\sim}{\lambda}) \mid F\} \to N(0, \Sigma).$$

The additional assumption that the elements $\underset{\sim}{\lambda}$ are all distinct is important here since otherwise the bootstrap will not work as has been pointed out by Hall et al., (1993) in rather general context. Beran and Srivastava (1985) also observed this for the covariance matrix of i.i.d. random vectors.

Bootstrapping the eigenvalue distribution

Theorems 1 and 2 give a flavor that, for testing the positive definiteness of $C(x_2) - C(x_1)$, additional difficulties arise. Estimating the covariance matrix Σ requires yet another smoothing method and in practice miscalculations might occur due to the complicated form of Σ. An easier approach is called for. The bootstrap method is a nonparametric resampling scheme that uses the observed sample again and helps in constructing approximations to the limiting distribution.

Let $X = \{(X_i, \underset{\sim}{Y}_i)\}_{i=1}^n$ denote the observations and $X^* = \{(X_i^*, \underset{\sim}{Y}_i^*)\}_{i=1}^n$ the bootstrap observations i.e.,

$$P[(X_i^*, \underset{\sim}{Y}_i^*)\} = (X_j, \underset{\sim}{Y}_j) \mid X] = n^{-1}, \quad 1 \le i, \quad j \le n.$$

With these new observations construct now $\hat{C}^*(x_2) - \hat{C}^*(x_1)$ from X^* as $\hat{C}(x_2) - \hat{C}(x_1)$ was constructed from X. We say that the bootstrap works if this new estimate $\hat{C}^*(x_2) - \hat{C}^*(x_1)$ centered around the already calculated $\hat{C}(x_2) - \hat{C}(x_1)$ tends to the same limit along almost all sample sequences. Since this last distribution can be repeatedly simulated on the computer, we can compute statistics like confidence limits without knowledge of the asymptotic normal distribution. Before we do this, we need to show that the bootstrap works. This is done in the following theorem.

Theorem 3. *Under the assumptions of Theorem* 1, *with probability equal to one, as* $n \to \infty$,

$$\mathcal{L}\{\sqrt{nh}[\operatorname{uvec}(\hat{C}^*(x_2) - \hat{C}^*(x_1)) - \operatorname{uvec}(\hat{C}(x_2) - \hat{C}(x_1))] \mid X\} \to N(0, V).$$

Computational Statistics and Data Analysis, 19, 641-654

Let $\hat{\underline{\lambda}}^*$ denote the eigenvalues of $\hat{C}^*(x_2) - \hat{C}^*(x_1)$. The analog of Theorem 2 is given in

Theorem 4. *Under the assumptions of Theorem 2, with probability equal to one, as* $n \to \infty$,

$$\mathcal{L}(\sqrt{nh}(\hat{\underline{\lambda}}^* - \hat{\underline{\lambda}}) | X) \to N(0, \Sigma).$$

We now describe the asymptotic distribution of $\hat{D}(x_2) - \hat{D}(x_1)$. For this, define

$$v_{\ell k, \ell' k'} = \int K^2(u)\,du[\operatorname{cov}\{Y_\ell Y_k, Y_{\ell'} Y_{k'} | X = x_1\}/g(x_1) + \operatorname{cov}\{Y_\ell Y_k, Y_{\ell'} Y_{k'} | X = x_2\}/g(x_2)]$$

and the $\frac{1}{2}d(d+1) \times \frac{1}{2}d(d+1)$ matrix $V_1 = (v_{\ell k, \ell' k'})$. The elements of $\hat{D}(x)$, the estimator of $D(x)$, are given by

$$\hat{d}_{\ell k}(x) = \hat{c}_{\ell k}(x) + \hat{m}_{h,\ell}(x) \cdot \hat{m}_{h,k}(x),$$

where $\hat{m}_{h,\ell}(x)$ is the regression kernel smoother defined in (2.3). The asymptotic distribution is given in

Theorem 5. *Under the assumptions of Theorem 1, as* $n \to \infty$

$$\mathcal{L}\{\sqrt{nh}[\operatorname{uvec}(\hat{D}(x_2) - \hat{D}(x_1)) - \operatorname{uvec}(D(x_2) - D(x_1))] | F\} \to N(0, V_1).$$

Again bootstrap versions for this and for the corresponding eigenvalues are possible. We do not state them since they are straightforward.

3. Application

The above resampling method has been applied to the data from the Family Expenditure Survey (1968–1983), also used in (Härdle et al., 1991). The data are cross section data, the i.i.d. assumption is thus justifiable. These data sets contain net family income (variable X) and expenditure on nine categories (variables Y's) from around 7000 households per year, a number fairly big enough to believe into the asymptotic statistical behavior described in the above theorems. For the specific names of categories, we refer to the aforementioned paper. To test the positive definiteness of $D(x_2) - D(x_1)$, the bootstrap distribution

$$\mathcal{L}\{\sqrt{nh}(\hat{\underline{\lambda}}^* - \hat{\underline{\lambda}}) | X\}$$

was simulated $N = 1000$ times. The 0.025 and the 0.975 quantiles were recorded to construct a confidence interval for $\lambda_{\min}$, the smallest eigenvalue. This confidence

Computational Statistics and Data Analysis, 19, 641-654

interval leads immediately to a test of increasing dispersion by checking whether 0 is in this interval. Table 1 shows the results of this bootstrap procedure. In the rows of Table 1 we see the results for the different years, in the columns the results for different x_1, x_2. $\lambda_{CLO}, \lambda_{CUP}$ denote the lower and upper 95%-confidence limits (computed from the bootstrap distribution) of λ_{min}.

The bandwidths h chosen for each location were on a logarithmic scale since the frequency of observations decreases rapidly as x increases over 1. (Note that the observations were scaled by mean budget, i.e. $x = 1$ denotes the households with average budget.) In particular the bandwidth for locations $x = 0.5, 1, 1.5, 2$ were $h = 0.1, 0.13, 0.169, 0.2197$ in a 30% increase of h as one moves to higher budget levels. These bandwidths have been often used in the estimation of mean function for this particular data set. As we mentioned in the previous section, the bootstrap method for testing on λ_{min} does not work when there are ties for the value of λ_{min}. Although we did not report the second smallest eigenvalues of $\hat{D}(x_2) - \hat{D}(x_1)$ here, we observed that the two smallest eigenvalues are distinct enough to assume that there are no ties for λ_{min}.

Note from Table 1 that the confidence interval for the smallest eigenvalue contains zero very often. Sometimes even the negative definiteness is evident since zero is even larger than the right bound of the confidence interval. This does not stand in contrast with the results of Härdle et al., (1991). In that later paper, an average derivative (a weighted average of differences) is calculated, and it is well possible that the average derivative matrix is positive definite although some difference $\Delta D = D(x_2) - D(x_1)$ is negative definite.

Table 1
Smallest eigenvalues of $\hat{D}(x_2) - \hat{D}(x_1)$ with confidence limits. All entries multiplied by 10^3

	$x_1 = 0.5, x_2 = 1$			$x_1 = 1, x_2 = 1.5$			$x_1 = 1.5, x_2 = 2$		
	λ_{CLO}	$\hat{\lambda}_1$	λ_{CUP}	λ_{CLO}	$\hat{\lambda}_1$	λ_{CUP}	λ_{CLO}	$\hat{\lambda}_1$	λ_{CUP}
1969	−0.4	0.05	0.8	−0.6	0.6	3.4	0.04	1.2	8.2
1970	−0.2	0.2	0.9	−0.4	0.6	2.2	−5.4	−1.4	7.5
1971	−0.2	0.1	0.8	−2.3	−0.6	2.5	−2.0	−0.07	7.5
1972	−0.8	−0.2	0.5	−2.6	−1.1	0.8	−2.6	−0.1	6.1
1973	−0.6	0.04	0.9	0.6	1.3	4.1	−6.5	−2.7	5.0
1974	−0.3	0.3	1.0	−2.7	−1.0	2.4	0.1	2.5	1.0
1975	−0.2	0.3	1.2	−0.8	0.7	2.9	2.3	3.4	13.8
1976	−0.07	0.6	1.3	−0.6	0.5	3.0	0.3	1.9	9.3
1977	−0.3	0.3	1.2	−2.2	−0.2	2.7	−0.8	2.1	8.9
1978	−0.01	0.6	1.5	−1.5	0.2	2.5	−1.4	−6.9	14.5
1979	−1.5	−1.0	−0.2	−0.2	1.1	3.9	−2.5	−0.9	4.8
1980	−0.9	−0.1	0.6	−0.6	0.9	3.1	−4.2	−2.1	1.9
1981	−1.0	−0.4	0.3	0.3	1.1	3.5	−1.0	−6.7	2.7
1982	−1.6	−1.1	−0.3	−0.02	0.9	3.2	−1.7	−0.04	6.8
1983	−1.0	−0.3	0.6	−2.3	−1.4	0.01	0.8	1.4	7.6

(1995) Härdle, W. and Park, B. Testing Increasing Dispersion

4. Proofs

For the proofs we need some more notation. Let $T_{n\ell k}$ denote the 8×1 vector

$$T_{n\ell k} = (\hat{g}_h(x_1), \hat{m}_{h,\ell}(x_1)\hat{g}_h(x_1), \hat{m}_{h,k}(x_1)\hat{g}_h(x_1), n^{-1}\textstyle\sum_{i=1}^n K_h(x_1 - X_i)\, Y_{i\ell}\, Y_{ik},$$
$$\hat{g}_h(x_2), \hat{m}_{h,\ell}(x_2)\hat{g}_h(x_2), \hat{m}_{h,k}(x_2)\hat{g}_h(x_2), n^{-1}\textstyle\sum_{i=1}^n K_h(x_2 - X_i)\, Y_{i\ell}\, Y_{ik})^{\mathrm{T}}. \tag{4.1}$$

Then define the $8 \times \frac{1}{2}d(d+1)$ matrix

$$T_n = (T_{n11}^{\mathrm{T}}, T_{n12}^{\mathrm{T}}, \ldots, T_{n1d}^{\mathrm{T}}, T_{n22}^{\mathrm{T}}, \ldots, T_{n2d}^{\mathrm{T}}, \ldots, T_{ndd}^{\mathrm{T}})^{\mathrm{T}}. \tag{4.2}$$

We shall consider the vector $U_{\ell k} = (1, Y_\ell, Y_k, Y_\ell Y_k)^{\mathrm{T}}$ and the moments

$$W_{\ell k,\ell' k'}(j) = g(x_j) E[U_{\ell k} U_{\ell' k'}^{\mathrm{T}} \mid X = x_j], \quad j = 1, 2.$$

Define also the matrices

$$W_{\ell k,\ell' k'} = \int K^2(u)\,\mathrm{d}u \begin{bmatrix} W_{\ell k,\ell' k'}(1) & 0 \\ 0 & W_{\ell k,\ell' k'}(2) \end{bmatrix} \tag{4.3}$$

and $W = (W_{\ell k,\ell' k'})$, $\ell \le k$, $\ell' \le k'$. The statistics $T_{n\ell k}$ will estimate

$$\theta_{\ell k} = (g(x_1), g(x_1)m_\ell(x_1), g(x_1)m_k(x_1), g(x_1)d_{\ell k}(x_1),$$
$$g(x_2), g(x_2)m_\ell(x_2), g(x_2)m_k(x_2), g(x_2)d_{\ell k}(x_2))^{\mathrm{T}},$$

and T_n will estimate $\theta = (\theta_{11}^{\mathrm{T}}, \theta_{12}^{\mathrm{T}}, \ldots, \theta_{1d}^{\mathrm{T}}, \theta_{22}^{\mathrm{T}}, \ldots, \theta_{2d}^{\mathrm{T}}, \ldots, \theta_{dd}^{\mathrm{T}})^{\mathrm{T}}$.

Proof of Theorem 1. The proof consists of the following three steps.

Step 1: $\mathcal{L}\{\sqrt{nh}\,[T_n - E(T_n)] \mid F\} \to N(0, W)$

Step 2: $\sqrt{nh}\,[E(T_n) - \theta] = o(1)$

Step 3: Application of the Mann-Wald theorem to derive the distribution of $\mathrm{uvec}[\hat{C}(x_2) - \hat{C}(x_1)]$

Proof of Step 1. Let

$$Z_{ni} = n^{-1/2} h^{1/2} (T_{ni} - E T_{ni}) \quad \text{and} \quad Z_n = \sum_{i=1}^{n} Z_{ni}$$

with

$$T_{ni} = (T_{n11i}^{\mathrm{T}}, \ldots, T_{n1di}^{\mathrm{T}}, T_{n22i}^{\mathrm{T}}, \ldots, T_{n2di}^{\mathrm{T}}, \ldots, T_{nddi}^{\mathrm{T}})^{\mathrm{T}}.$$

Note that $T_n = n^{-1}\Sigma_{i=1}^n T_{ni}$ and each $T_{n\ell k}$ from (4.1) can be written as a sample average of 8×1 vectors $T_{n\ell ki}$, i.e. $T_{n\ell k} = n^{-1}\Sigma_{i=1}^n T_{n\ell ki}$. Let now $c = (c_{11}^{\mathrm{T}}, \ldots, c_{1d}^{\mathrm{T}}, c_{22}^{\mathrm{T}}, \ldots, c_{2d}^{\mathrm{T}}, \ldots, c_{dd}^{\mathrm{T}})^{\mathrm{T}}$, then $8 \times \frac{1}{2}d(d+1)$ matrix composed from the 8×1 vectors $c_{\ell k} = (c_{1\ell k}, c_{2\ell k}, \ldots, c_{8\ell k})^{\mathrm{T}}$. Then we claim that

$$\mathrm{var}(c^{\mathrm{T}} Z_n) = c^{\mathrm{T}} W c + O(h). \tag{4.4}$$

Computational Statistics and Data Analysis, 19, 641-654

This can be seen as follows. The observations are i.i.d., therefore $\operatorname{var}(c^{\mathrm{T}} Z_n) = n \operatorname{var}(c^{\mathrm{T}} Z_{n1}) = h \operatorname{var}(c^{\mathrm{T}} T_{n1})$. Now

$$\operatorname{var}(c^{\mathrm{T}} T_{n1}) = \operatorname{var}\Big(\sum_{\ell \le k} \sum c_{\ell k}^{\mathrm{T}} T_{n\ell k1}\Big) = \sum_{\ell \le k} \sum \sum_{\ell' \le k'} \sum c_{\ell k}^{\mathrm{T}} \operatorname{cov}[T_{n\ell k1}, T_{n\ell' k'1}] c_{\ell' k'}.$$

And

$$\operatorname{cov}[T_{n\ell k1}, T_{n\ell' k'1}] = E T_{n\ell k1} T_{n\ell' k'1} - [E T_{n\ell k1}][E T_{n\ell' k'1}]^{\mathrm{T}}. \tag{4.5}$$

The first term in (4.5) is $h^{-1} W_{\ell k, \ell' k'} + \mathrm{O}(1)$. To see this, let us compute $E[Y_\ell Y_k Y_{\ell'} Y_{k'} K_h^2(x - X)]$ and $E[Y_\ell Y_k Y_{\ell'} Y_{k'} K_h(x_1 - X) K_h(x_2 - X)]$, for example. Other terms can be handled in a similar way.

$$\begin{aligned} E[Y_\ell Y_k Y_{\ell'} Y_{k'} K_h^2(x - X)] &= \int E[Y_\ell Y_k Y_{\ell'} Y_{k'} \mid X = t] K_h^2(x - t) g(t)\,\mathrm{d}t \\ &= \frac{1}{h} \int E[Y_\ell Y_k Y_{\ell'} Y_{k'} \mid X = x - hu] g(x - hu) K^2(u)\,\mathrm{d}u \\ &= \frac{1}{h} \{E(Y_\ell Y_k Y_{\ell'} Y_{k'} \mid X = x) g(x) \int K^2(u)\,\mathrm{d}u\} + \mathrm{O}(1). \end{aligned}$$

The last equation follows from the assumptions (A.1) and (A.6). Now using the same arguments as in (Schuster, 1972, p.86),

$$E[Y_\ell Y_k Y_{\ell'} Y_{k'} K_h(x_1 - X) K_h(x_2 - X)] = \mathrm{O}(1).$$

Since the second term in (4.5) is O(1), we have

$$\begin{aligned} \operatorname{var}(c^{\mathrm{T}} Z_n) &= n \cdot (n^{-1} h) \cdot \sum_{\ell \le k} \sum \sum_{\ell' \le k'} \sum c_{\ell k}^{\mathrm{T}} [h^{-1} W_{\ell k, \ell' k'} + \mathrm{O}(1)] c_{\ell' k'} \\ &= \sum_{\ell \le k} \sum \sum_{\ell' \le k'} \sum c_{\ell k}^{\mathrm{T}} W_{\ell k, \ell' k'} c_{\ell' k'} + \mathrm{O}(h). \\ &= c^{\mathrm{T}} W c + \mathrm{O}(h) \end{aligned}$$

which proves (4.4). Now

$$\begin{aligned} \sum_{i=1}^{n} E|c^{\mathrm{T}} Z_{ni}|^3 &= n^{-3/2} h^{3/2} \cdot n E|c^{\mathrm{T}} T_{n1} - E c^{\mathrm{T}} T_{n1}|^3 \\ &\le n^{-1/2} h^{3/2} |c|^3 E|T_{n1} - E T_{n1}|^3 \\ &\le N_1 n^{-1/2} h^{3/2} \max_{\ell,k} \max \{E|K_h(x - X_1) - E K_h(x - X_1)|^3, \ldots, \\ & E|Y_{1\ell} Y_{1k} K_h(x - X_1) - E Y_{1\ell} Y_{1k} K_h(x - X_1)|^3\}. \end{aligned}$$

All the terms in the curly brackets are of order $O(h^{-3})$ by (A.6). Thus

$$\sum_{i=1}^{n} E|c^T Z_{ni}|^3 \le N_2 (nh^3)^{-1/2} \to 0$$

if $nh^3 \to \infty$ as $n \to \infty$. Applying the Berry–Esséen theorem (Loéve, 1963, p. 288) to the triangular sequence $\{c^T Z_{ni}\}_{i=1}^{n}$, we obtain

$$\mathcal{L}\left\{c^T \sum_{i=1}^{n} Z_{ni}\right\} \to N(0, c^T W c)$$

for any c. By the Cramer–Wold device, this implies

$$\mathcal{L}\left\{\sum_{i=1}^{n} Z_{ni}\right\} \to N(0, W),$$

which proves Step 1. Note that W is not positive definite but rather positive semidefinite so that this limiting distribution is in fact improper. It is not hard to see why W is not positive definite. Recall the definition of $W_{\ell k, \ell' k'}$ from (4.3). We can permute rows and columns of W to obtain $\det(W) = \det(\tilde{W})$ for a $4d(d+1) \times 4d(d+1)$ matrix

$$\tilde{W} = \begin{bmatrix} \tilde{W}(1) & 0 \\ 0 & \tilde{W}(2) \end{bmatrix}, \qquad \tilde{W}(j) = (W_{\ell k, \ell' k'}(j)) = g(x_j) E[UU^T \mid X = x_j]$$

where $U = (U_{11}^T, U_{12}^T, \ldots, U_{1d}^T, U_{22}^T, \ldots, U_{2d}^T, \ldots, U_{dd}^T)^T$. Certainly there exists a $c \neq 0$: $c^T U = 0$, hence $\tilde{W}$ and thus W is not positive definite.

Proof of step 2. We claim

$$E_F(T_n) - \theta = O(h^2). \tag{4.6}$$

Note $E(T_n) = E(T_{n1})$ and $E(T_{n1})$ has typical elements, $EK_h(x - X)$, $EY_\ell K_h(x - X)$, $EY_\ell Y_k K_h(x - X)$. We show $|EY_\ell Y_k K_h(x - X) - g(x)E(Y_\ell Y_k \mid X = x)| = O(h^2)$. The other terms are very similar to handle.

$$EY_\ell Y_k K_h(x - X) = \int E(Y_\ell Y_k \mid X = t) K_h(x - t) g(t)\, dt$$

$$= \int E(Y_\ell Y_k \mid X = x - hu) g(x - hu) K(u)\, dy.$$

Hence

$$|EY_\ell Y_k K_h(x - X) - g(x)E(Y_\ell Y_k \mid X = x)|$$

$$= \left| \int [E(Y_\ell Y_k \mid (X = x - hu) g(x - hu) - E(Y_\ell Y_k \mid X = x) g(x)] K(u)\, du \right|$$

$$\le \sup_x \left| \frac{\partial^2}{\partial x^2} (E(Y_\ell Y_k \mid X = x) \cdot g(x)) \right| \cdot h^2 \int u^2 K(u)\, du/2 = O(h^2)$$

Computational Statistics and Data Analysis, 19, 641-654

by (A.7). Hence

$$\sqrt{nh}[E(T_n) - \theta] = O(n^{1/2}h^{5/2})$$
$$\to 0$$

since $nh^5 \to 0$.

Proof of step 3. Consider the transformation $H: R^{4d(d+1)} \to R^{d(d+1)/2}$ defined by

$$H(\underline{y}) = (H_{11}, H_{12}, \ldots, H_{1d}, H_{22}, \ldots, H_{2d}, \ldots, H_{dd})^{\mathrm{T}}(\underline{y})$$

where $H_{\ell k}(\underline{y}) = (y_{8\ell k}/y_{5\ell k}) - (y_{6\ell k}y_{7\ell k}/y_{5\ell k}^2) - (y_{4\ell k}/y_{1\ell k}) + (y_{2\ell k}y_{3\ell k}/y_{1\ell k}^2)$ and $\underline{y} = y_{111}, \ldots, y_{811}, y_{112}, \ldots, y_{812}, \ldots, y_{1dd}, \ldots, y_{8dd})$. Apply the Mann–Wald theorem (Rao, 1973, p. 387) to H. Then

$$\mathscr{L}\{(nh)^{1/2}[H(T_n) - H(\theta)]|F\} \to N\left(0, \left(\frac{\partial}{\partial \underline{\theta}} H(\underline{\theta})\right) W \left(\frac{\partial}{\partial \underline{\theta}} H(\underline{\theta})\right)^{\mathrm{T}}\right).$$

It is straightforward to show that $H(T_n) = \mathrm{uvec}(\hat{C}(x_2) - \hat{C}(x_1))$ and $H(\theta) = \mathrm{uvec}(C(x_2) - C(x_1))$. Also going through some tedious calculations, we can see

$$V = \left(\frac{\partial}{\partial \underline{\theta}} H(\underline{\theta})\right) W \left(\frac{\partial}{\partial \underline{\theta}} H(\underline{\theta})\right)^{\mathrm{T}}.$$

Proof of Theorem 2. Note that $\hat{D}(\lambda_i) = (\lambda_i - \hat{\lambda}_i)\hat{D}'(\tilde{\lambda}_i)$ where $\tilde{\lambda}_i$ lies in between λ_i and $\hat{\lambda}_i$. Using the same arguments as in (Härdle and Hart, 1992)

$$\lambda_i - \hat{\lambda}_i = \hat{D}'(\tilde{\lambda}_i)^{-1}\hat{D}(\lambda_i) = \hat{D}(\lambda_i)/D'(\lambda_i) + o_p(1).$$

This implies

$$\mathscr{L}[\sqrt{nh}(\hat{\underline{\lambda}} - \underline{\lambda})|F] \to M^{\mathrm{T}}N(0, V) \stackrel{\mathscr{L}}{=} N(0, \Sigma).$$

Proof of Theorem 3. We follow the triangular array approach described in (Beran, 1984). Define

$$J_n(F) = \mathscr{L}\{(nh)^{1/2}[T_n - E_F(T_n)]\,|\,F\}.$$

We say that a sequence F_n of distribution functions is in the class $\mathscr{C}(F)$ if and only if

(i) $hE_{F_n}[Y_\ell^{r_1}Y_k^{r_2}Y_{\ell'}^{r_3}Y_{k'}^{r_4}K_h^2(x - X)] \to E_F[Y_\ell^{r_1}Y_k^{r_2}Y_{\ell'}^{r_3}Y_{k'}^{r_4}\,|\,X = x]\cdot g(x)\int K^2(u)\,du,$

$$1 \le \ell, k, \ell', k' \le d \quad \text{and} \quad r_1, r_2, r_3, r_4 = 0, 1.$$

(ii) $hE_{F_n}[Y_\ell^{r_1}Y_k^{r_2}Y_{\ell'}^{r_3}Y_{k'}^{r_4}K_h(x_1 - X)K_h(x_2 - X)] \to 0, \quad 1 \le \ell, k, \ell', k' \le d$ and

$$r_1, r_2, r_3, r_4 = 0, 1$$

Computational Statistics and Data Analysis, 19, 641-654

(iii) $E_{F_n}(Y_\ell^{r_1} Y_k^{r_2} K_h(x - X)) \to E_F(Y_\ell^{r_1} Y_k^{r_2} | X = x) g(x)$,

$1 \le \ell, k \le d$ and $r_1, r_2 = 0, 1, 2$

(iv) $E_{F_n} Y_\ell^{r_1} Y_k^{r_2} \to E_F Y_\ell^{r_1} Y_k^{r_2}$, $1 \le \ell, k \le d$ and $r_1, r_2 = 0, 1, 2, 3$.

Let $\hat{F}_n$ denote the empirical distribution function, then $P_F[\hat{F}_n \in \mathscr{G}(F)] = 1$. Note that $E_{\hat{F}_n}(T_n^*) = T_n$. If we show that, for any $F_n \in \mathscr{G}(F)$,

$$J_n(F_n) \to N(0, W), \tag{4.7}$$

then this implies the theorem, going through the Step 3 described in the proof of Theorem 1. The proof of (4.7) can be done by the same arguments as in the Step 1 of the proof of Theorem 1.

Proof of Theorem 4. This proof is completely analogous to that Theorem 2.

Proof of Theorem 5. Consider the transformation $H^*: \mathbb{R}^{\frac{1}{2}d(d+1)} \to \mathbb{R}^{d(d+1)/2}$ defined by

$$H^*(\underline{y}) = (H_{11}^*, H_{12}^*, \ldots, H_{1d}^*, H_{22}^*, \ldots, H_{2d}^*, \ldots, H_{dd}^*)(\underline{y})$$

with $H_{\ell k}^*(\underline{y}) = (y_{8\ell k}/y_{5\ell k}) - (y_{4\ell k}/y_{1\ell k})$. Then by the Mann–Wald theorem again

$$\mathscr{L}\{\sqrt{nh}[H^*(T_n) - H^*(\theta)] | F\} \to N\left(0, \left(\frac{\partial}{\partial \underline{\theta}} H^*(\underline{\theta})\right) W \left(\frac{\partial}{\partial \underline{\theta}} H^*(\underline{\theta})\right)^\tau\right).$$

Note that $H^*(T_n) = \text{uvec}[\hat{D}(x_2) - \hat{D}(x_1)]$ and $H^*(\theta) = \text{uvec}[D(x_2) - D(x_1)]$. Straightforward but tedious calculations show that

$$V_1 = \left(\frac{\partial}{\partial \underline{\theta}} H^*(\underline{\theta})\right) W \left(\frac{\partial}{\partial \underline{\theta}} H^*(\underline{\theta})\right)^\tau.$$

References

Beran, R., Bootstrap methods in statistics, *Jber. der Dt. Math.-Verein.* **86** (1984) 14–30.

Beran, R. and M.S. Srivastava, Bootstrap tests and confidence regions for functions of a covariance matrix. *Ann. Statist.* **13** (1985) 95–115.

Carroll, R.J., Adapting for heteroscedasticity in linear models, *Ann. Statist.* **10** (1982) 1224–33.

Family Expenditure Survey, Annual Base Tapes (1968–1983). Department of Employment, Statistics Division, Her Majesty's Stationary Office, London 1968–1983.

Härdle, W., *Applied Nonparametric Regression.* Cambridge University Press, Econometric Society Monograph Series 19.

Härdle, W. and H. Hart, A Bootstrap test for positive definiteness of income effect matrices, *Econom. Theor.* **8** (1992) 276–290.

Härdle, W., W. Hildenbrand and M. Jerison, Empirical evidence on the law of demand, *Econometrica*, **59** (1991) 1525–1549.

Härdle, W. and J.S. Marron, Fast and simple scatterplot smoothing, submitted to *Comput. Statist. Data Anal.*

Hall, P., W. Härdle and L. Simar, On the inconsistency of bootstrap distribution estimators, *Comput. Statist. Data Anal.* **16** (1992) 11–18.

(1995) Härdle, W. and Park, B. Testing Increasing Dispersion

Hildenbrand, W., *An Essay on Market Demand* (University of Bonn, Bonn, 1992).
Hoaglin, D.C., F. Mosteller and J.W. Tukey, *Understanding Robust and explanatory data analysis* (Wiley, New York, 1983).
Jones, M.C., S.J. Davies and B.U. Park (1994). Versions of kernel-type regression estimators, *J. Amer. Statist. Assoc.*, in print.
Loéve, M., *Probability theory* (Van Nostrand, New York, 1963).
Rao, C.R., *Linear statistical inference and its applications* (Wiley, New York, 1973).
Robinson, P.M., Asymptotically efficient estimation in the presence of heteroscedasticity of unknown form, *Econometrica*, **55** (1986) 895–92.
Schuster, E.F., Joint asymptotic distribution of the estimated regression function at a finite number of distinct points, *Ann. Math. Statist.*, **43** (1972) 84–88.
Silverman, B.W., *Density estimation* (Chapman and Hall, London, 1986).

Statistics **26** (1995) 287–306
Reprints available directly from the publisher
Photocopying permitted by license only

BETTER BOOTSTRAP CONFIDENCE INTERVALS FOR REGRESSION CURVE ESTIMATION

W. HÄRDLE, S. HUET and E. JOLIVET

Humboldt Universität zu Berlin and INRA, Jouy-en-Josas

(Received April 1991; in final form 5 July 1994)

Summary. Bootstrap methods in curve estimation have been introduced for smoothing parameter selection and for construction of confidence intervals. Most of the papers on confidence intervals use explicit bias estimation or the technique of "undersmoothing" to deal with bias. Coverage accuracy has only been considered for curve estimates with constant variance function. In this paper we show that explicit bias estimation even with heteroscedastic variance structure leads to an improvement of coverage accuracy, when compared to undersmoothing. Bootstrapping with this bias correction using the so-called wild bootstrap leads to an improved coverage accuracy.

AMS 1991 subject classifications: Primary 62 G07; secondary 62 G09.

Key words: Bootstrap, nonparametric regression, wild bootstrap, kernel estimators.

1. INTRODUCTION

Bootstrap methods in nonparametric curve estimation have been considered recently by several authors. Basically, two types of results have been obtained. Resampling procedures have been shown to be consistent, i.e., *the bootstrap works*, the simulated distribution *approaches the true distribution*, as the distribution of the estmator. Also, errors of coverage probability have been evaluated: these have been calculated by use of Edgeworth expansion in the density estimation, Hall (1991) and regression function estimation setting, Hall (1992b).

The results obtained for regression function estimation are of somewhat restricted applicability in the sense that a root-n estimable conditional variance function is supposed to exist. This can be achieved on the assumption of either homoscedasticity of the errors or of the existence of some parametric modelisation of the variance structure. Under these hypothesis, resampling has been used by bootstrapping *all* the residuals. The stochastic structure is so that information about the errors can be drawn from all these residuals.

This resampling method has been generalized to the so-called *wild bootstrap* (Härdle and Mammen (1990); Härdle and Marron (1991)) which corrects locally and adapts for heteroscedasticity and local skewness. It is called wild bootstrap since one resamples from only one single residual instead of all residuals. A theoretical step of understanding the behaviour of this method has been done by Cao-Abad (1991). Using Berry–Essen inequalities, he obtained, for a confidence interval calculated with explicit bias estimation, and for a given fixed bandwidth, an error in coverage probability of order $n^{-2/9}$.

Statistics, 26, 287-306

In this note, we aim to refine this technique by using Edgeworth expansions, but also to compare explicit bias estimation with the technique of undersmoothing. We show that Studentized confidence intervals *with* explicit bias estimation achieve coverage accuracy of $n^{-3/8}$. This may be compared with the $n^{-2/5}$ rate of Hall (1991) who has used undersmoothing. Bootstrapping improves this accuracy. We obtain for our method $n^{-3/7}$. This is only a slight loss of $n^{3/35} = n^{-5/7}/n^{-4/5}$ when compared with the previously mentioned method of Hall (1991) who obtained the rate $n^{-4/5}$ when bootstrapping from constant variance residuals.

In summary we can say that explicit bias estimation pays off in confidence interval construction, although the improvement is in practical studies probably not visible unless the sample size is gigantic. The technique of undersmoothing requires the knowledge of what is an "undersmoothed curve", i.e., a fine tuning of the bandwidth. Our results need similar assumptions on the speed of bandwidths, more precisely, as we will see later, our method consists of undersmoothing in combination with oversmoothing for bias estimation.

Let $(X_1, Y_1), \ldots, (X_n, Y_n)$ be n observations in $\mathbb{R}^2$ such that $X_1, \ldots, X_n$ is a sample of i.i.d. random variables on an interval $I \subset \mathbb{R}$. Assume

$$Y_i = m(X_i) + \sigma(X_i)e_i, \quad i = 1, \ldots, n,$$

where $e_1, \ldots, e_n$ is a sample of independent random variables with 0 mean and unit variance, independent of $X_1, \ldots, X_n$. Let $\varepsilon_i = \sigma(X_i)e_i$. We make the following assumptions.

(A1) The functions m, σ, f are four times continuously differentiable on I. σ and f are strictly positive and bounded uniformly on I.

(A2) The distribution of ε has finite moments of all order: $\gamma_k(x) = \mathrm{E}(\varepsilon^k | X = x) < \infty$ for all $k \geqslant 3$ and all $x \in I$.

We are interested in confidence intervals for $m(x) = \mathrm{E}(Y|X = x)$ for a fixed $x \in \mathbb{R}$. For this purpose we use kernel estimation.

$$\hat{m}_h(x) = \frac{1}{nh} \sum_{i=1}^{n} \frac{Y_i K\left(\frac{x - X_i}{h}\right)}{\frac{1}{nh} \sum_{j=1}^{n} K\left(\frac{x - X_j}{h}\right)}$$

$$= \frac{1}{nh} \sum_{i=1}^{n} \frac{Y_i K\left(\frac{x - X_i}{h}\right)}{\hat{f}_h(x)}. \tag{1.1}$$

For the kernel we make the assumption.

(A3) The kernel K is a symmetric, twice differentiable probability density function with finite $C_k = \int K^k(x)\,\mathrm{d}x$, $k = 1, 2, 3$. K is piecewise strictly monotonous and has compact support.

It is well known that $\hat{m}_h(x)$ is a consistent estimate of $m(x)$ if $h \to 0, nh \to \infty$.

(1995) Härdle, W., Huet, S. and Jolivet, E.
Better Bootstrap Confidence Intervals for Regression Curve Estimation.

Statistics, 26, 287-306

2. CONFIDENCE INTERVALS

2.1. *Asymptotic Pivots*

Confidence intervals are constructed by using the asymptotic distribution of pivotal quantities. It is well known that $\hat{m}_h(x)$ is biased and that

$$S_n = \sqrt{nh}\frac{\hat{m}_h(x) - m(x)}{V_n^{1/2}} - \frac{B_n}{V_n^{1/2}} \tag{2.1}$$

with

$$B_n = \sqrt{nh}\left\{\frac{1}{n}\sum_{i=1}^{n} W_{hi}(x)m(X_i) - m(x)\right\} \tag{2.2}$$

$$V_n = \frac{h}{n}\sum_{i=1}^{n} W_{hi}^2(x)\sigma^2(X_i), \tag{2.3}$$

converges in distribution to a normalized Gaussian random variable, see, e.g., Hall (1992b). The weight sequence in (2.2) is $W_{hi}(x) = (1/h)K((x - X_i)/h)/\hat{f}_h(x) = K_h(x - X_i)/\hat{f}_h(x)$. Moreover, S_n can be rewritten as

$$S_n = \frac{\sum_{i=1}^{n} Z_i}{(\sum_{i=1}^{n} K_h^2(x - X_i)\sigma^2(X_i))^{1/2}},$$

where $Z_i = K_h(x - X_i)\varepsilon_i$, $i = 1, \ldots, n$ are independent random variables. The moments conditional on X_i of S_n are rather easy to evaluate (the two first being exactly 0 and 1) but the asymptotic pivot S_n cannot be used for practical computation of confidence intervals because B_n and V_n are unknown quantities. Consequently, we introduce two other asymptotic pivots, U_n and T_n.

U_n is obtained by estimating V_n from the residuals and neglecting the bias B_n:

$$\begin{aligned} U_n &= \sqrt{nh}\frac{\sum_{i=1}^{n} K_h(x - X_i)\varepsilon_i}{(\sum_{i=}^{n} K_h^2(x - X_i)\hat{\varepsilon}_i^2)^{1/2}} \\ &= \sqrt{nh}\frac{\hat{m}_h(x) - E(\hat{m}_h(x))}{\hat{V}_n^{1/2}} \\ &= S_n\left(\frac{V_n}{\hat{V}_n}\right)^{1/2}, \end{aligned} \tag{2.4}$$

with $\hat{\varepsilon}_i = Y_i - \hat{m}_h(X_i)$. As U_n can be shown to be asymptotically $N(0, 1)$, it can be used to compute asmptotic confidence intervals with prescribed asymptotic level for $E(\hat{m}_h(x))$. But, as pointed out before, by using a fine tuning of h, the bias can be made so small that the confidence interval for $E(\hat{m}_h(x))$ is also a confidence interval for $m(x)$. In order to achieve this goal, one uses undersmoothing, i.e., one selects an h faster than the optimal $n^{-1/5}$, see Hall (1992a).

T_n is obtained by estimating both B_n and V_n. Let $\check{m}(x)$, $x \in I$ be any estimator of $m(x)$. An estimation $\check{B}_n$ of B_n is given by

$$\check{B}_n = \sqrt{nh}\left(\frac{1}{n}\sum_{i=1}^{n} W_{hi}(x)\check{m}(X_i) - \check{m}(x)\right).$$

(1995) Härdle, W., Huet, S. and Jolivet, E.
Better Bootstrap Confidence Intervals for Regression Curve Estimation.

Statistics, 26, 287-306

In practice $\check{m}(x)$ could be a kernel estimator. Denote by $\hat{m}_g(.)$ a kernel estimator with a bandwidth g possibly distinct from h and using eventually a different kernel L. This gives as bias estimate $\hat{B}_n = \sqrt{nh}(1/n)\sum_{i=1}^n W_{hi}(x)\hat{m}_g(X_i) - \hat{m}_g(x))$. Finally, define T_n as

$$\begin{aligned} T_n &= \left[\sqrt{nh}\frac{\hat{m}_h(x) - m(x)}{\hat{V}_n^{1/2}} - \frac{\hat{B}_n}{\hat{V}_n^{1/2}} \right] \\ &= \left[S_n - \frac{(\hat{B}_n - B_n)}{V_n^{1/2}} \right] \left(\frac{V_n}{\hat{V}_n} \right)^{1/2} \\ &= U_n - \frac{\hat{B}_n - B_n}{\hat{V}_n^{1/2}}. \end{aligned} \tag{2.5}$$

T_n is also asymptotically $N(0, 1)$ and can be used directly to compute confidence intervals for $m(x)$.

In order to compare confidence intervals obtained from either U_n and T_n, we need to evaluate more precisely their respective levels. This will be done via Edgeworth expansion of their distributions. The same procedure will be repeated later for the bootstrap. We shall see that we then obtain more accurate coverage probability. In order to obtain the first terms of the Edgeworth expansion, we need to evaluate the first three moments of U_n and T_n.

2.2. *Asymptotic Expansion of* U_n

Let us introduce two intermediate quantities

$$\hat{U}_n = S_n \left[1 - \frac{1}{2}\frac{\hat{V}_n - V_n}{V_n} \right]. \tag{2.6}$$

and

$$\tilde{U}_n = S_n \left[1 - \frac{1}{2}\frac{\tilde{V}_n - V_n}{V_n} \right]. \tag{2.7}$$

where $\tilde{V}_n = (h/n)\sum_{i=1}^n W_{hi}^2(x)\varepsilon_i^2$. Obviously, $\hat{U}_n$ is the formal Taylor expansion of U_n, and $\tilde{V}_n$ approximates $\hat{V}_n$. Closeness between $\tilde{V}_n$ and $\hat{V}_n$ will be evaluated in lemma A.1. It can be shown (see the appendix) that the distribution of $\tilde{U}_n$ admits an Edgeworth expansion and that the distribution of U_n is sufficiently close to the one of $\tilde{U}_n$ to admit expansions with first term coinciding. More precisely, we have

Proposition 2.1. *Let* $nh^9 \to 0$, *then under* (*A1*)–(*A3*)

$$\limsup_{n\to\infty} \sqrt{nh} \left| P\{U_n \leqslant t\} - \Phi(t) - \frac{1}{\sqrt{nh}}\varphi(t)\frac{A_3(x)}{6}(2t^2 + 1) \right| = 0 \tag{2.8}$$

with

$$A_3(x) = \frac{1}{\sqrt{f(x)}} \frac{C_3}{C_2^{3/2}} \frac{\gamma_3(x)}{\sigma^3(x)}.$$

Here, $\Phi(.)$ is the distribution function of a standard Gaussian variable, $\phi(.)$ is its density. The proof of proposition 2.1 is given in the appendix.

Statistics, 26, 287-306

Using Cornish–Fisher expansion (see for instance Barndorff–Nielsen and Cox (1989)), that is inverting the above Edgeworth expansion, we obtain, uniformly for t in any compact subset of $\mathbb{R}$:

$$\mathrm{P}\left\{U_n \leqslant t - \frac{1}{\sqrt{nh}} \frac{A_3(x)}{6}(2t^2+1)\right\} = \Phi(t) + o\left(\frac{1}{\sqrt{nh}}\right), \tag{2.9}$$

which provides confidence intervals for $\mathrm{E}(\hat{m}_h(x))$ with accuracy of order $o(1/\sqrt{nh})$. Suppose now we parameterize $h = n^{-\alpha}$ with $\alpha > \frac{1}{9}$, to fulfill the assumption of Proposition 2.1. The remainder term in (2.9) is thus $o(\sqrt{n^{\alpha-1}})$. For the mean squared error optimal choice $\alpha = 1/5$, this is equal to $o(n^{-2/5})$. Thus from (2.9) a one-sided confidence interval for $\mathrm{E}(\hat{m}_h(x))$ with level $\Phi(t) + o(\sqrt{n^{\alpha-1}})$ is

$$\left[\hat{m}_h(x) - \frac{\hat{V}_n^{1/2}}{\sqrt{nh}}\left(t - \frac{1}{\sqrt{nh}} \frac{A_3(x)}{6}(2t^2+1)\right), +\infty\right[.$$

This confidence interval is also a confidence interval for $m(x)$ with accuracy of order $1/\sqrt{nh}$ if $h = o(n^{-1/5})$, as we can write

$$\mathrm{P}\left\{\sqrt{nh}\frac{\hat{m}_h(x) - m(x)}{\hat{V}_n^{1/2}} \leqslant u\right\} = \mathrm{P}\left\{U_n \leqslant u + \frac{B_n}{\hat{V}_n^{1/2}}\right\}$$

and under our conditions $B_n/\sqrt{nh}$ is $O(h^2)$ (Härdle, 1990).

Let z_a be the a-quantile of the standard Gaussian distribution: $\Phi(z_a) = a$. A confidence interval for $m(x)$ of asymptotic level a, without correction by expansion, is given by

$$I_{U_n} = \left[\hat{m}_h(x) - \frac{\hat{V}_n^{1/2}}{\sqrt{nh}} z_a, +\infty\right[$$

and has level equal to $a + O(1/\sqrt{nh}, \sqrt{nh^5})$. The second term, $O(\sqrt{nh^5})$, comes from the bias B_n. Obviously, the best achievable rate of convergence of this quantity to a is obtained for $1/\sqrt{nh} = \sqrt{nh^5}$, that is for $h = n^{-1/3}$. The error in coverage probability is then $n^{-1/3}$, that is to say

$$\mathrm{P}\{m(x) \in I_{U_n}\} = a + O(n^{-1/3}). \tag{2.10}$$

2.3. *Asymptotic Expansion of T_n*

The asymptotic expansion of T_n is established through a treatment analogous to the one used in the preceding section. We consider two approximations of T_n defined by

$$\hat{T}_n = S_n\left(1 - \frac{1}{2}\frac{\hat{V}_n - V_n}{V_n}\right) - \frac{\hat{B}_n - B_n}{V_n^{1/2}}\left(1 - \frac{1}{2}\frac{\hat{V}_n - V_n}{V_n}\right)$$

and

$$\tilde{T}_n = S_n\left(1 - \frac{1}{2}\frac{\tilde{V}_n - V_n}{V_n}\right) - \frac{\tilde{B}_n - B_n}{V_n^{1/2}}\left(1 - \frac{1}{2}\frac{\tilde{V}_n - V_n}{V_n}\right),$$

Statistics, 26, 287-306

where

$$\tilde{B}_n = \sqrt{nh}\left(\frac{\sum_{i=1}^n\sum_{j=1}^n \mathrm{E}(\Omega_{ij}/X_i) + \mathrm{E}(\Omega_{ij}/X_j) - \mathrm{E}(\Omega_{ij})}{\sum_{i=1}^n\sum_{j=1}^n \mathrm{E}(\Xi_{ij}/X_i) + \mathrm{E}(\Xi_{ij}/X_j) - \mathrm{E}(\Xi_{ij})} - \hat{m}_g(x)\right)$$

with

$$\Omega_{ij} = K_h(x - X_i)L_g(X_i - X_j)Y_j \tag{2.11}$$

and

$$\Xi_{ij} = K_h(x - X_i)L_g(X_i - X_j). \tag{2.12}$$

Recall that $L(.)$ is a kernel enjoying the same properties as K and g a bandwidth. As in the previous section, one can show that the distribution function of $\tilde{T}_n$ admits an Edgeworth expansion and that the distribution functions of T_n, $\hat{T}_n$ and $\tilde{T}_n$ are sufficiently close to each other to admit the same Edgeworth expansion up to the first order. Details are given in the appendix.

The moments of $\tilde{T}_n$ are given by

$$\begin{aligned}
\mathrm{E}(\tilde{T}_n) &= -\frac{1}{2}\frac{1}{\sqrt{nh}}\frac{1}{\sqrt{f(x)}}\frac{C_3}{C_2^{3/2}}\frac{\gamma_3(x)}{\sigma^3(x)} \\
&\quad -\frac{1}{4}\sqrt{nh}g^2h^2\frac{\sqrt{f(x)}}{\sigma(x)}\frac{1}{C_2^{1/2}}L^{(2)}K^{(2)}\left(\mu''(x) + 2\mu'(x)\frac{f'(x)}{f(x)}\right) \\
&\quad + O\left(g^2h^2, \frac{1}{\sqrt{nh}}\left(\frac{h}{g}\right)^{5/2}, \frac{1}{\sqrt{nh}}\frac{h^{3/2}}{g^{1/2}}, \frac{1}{nh}\right); \\
\mathrm{E}(\tilde{T}_n^2) &= 1 - \left(\frac{h}{g}\right)^3 K''(0)\frac{K^{(2)}}{C_2} \\
&\quad + O\left(\frac{1}{nh}, h^2g^2, \left(\frac{h}{g}\right)^{5/2}\frac{1}{\sqrt{nh}}, h^4, \frac{h^3}{g}, \frac{1}{\sqrt{nh}}\frac{h^{3/2}}{g^{1/2}}\right); \\
\mathrm{E}(\tilde{T}_n^3) &= -\frac{7}{2}\frac{1}{\sqrt{nh}}\frac{1}{\sqrt{f(x)}}\frac{C_3}{C_2^{3/2}}\frac{\gamma_3(x)}{\sigma^3(x)} \\
&\quad -\frac{3}{4}\sqrt{nh}g^2h^2\frac{L^{(2)}K^{(2)}}{C_2^{1/2}}\frac{\sqrt{f(x)}}{\sigma(x)}\left(\mu''(x) + 2\mu'(x)\frac{f'(x)}{f(x)}\right) \\
&\quad + O\left(\frac{1}{nh}, h^2g^2, \frac{h^5}{g^5}, \frac{h^3}{g}, nh^5g^4, \sqrt{nh}g^4h^2\right)
\end{aligned}$$

with

$$\mu(x) = m''(x) + 2m'(x)\frac{f'(x)}{f(x)}$$

$$K^{(2)} = \int_{-\infty}^{+\infty} u^2K(u)\,\mathrm{d}u \quad L^{(2)} = \int_{-\infty}^{+\infty} u^2L(u)\,\mathrm{d}u.$$

Our purpose now is to compare the precision of confidence intervals for $m(x)$ based on U_n or T_n. We have seen that the best approximation achievable for the level of the

(1995) Härdle, W., Huet, S. and Jolivet, E.
Better Bootstrap Confidence Intervals for Regression Curve Estimation.

Statistics, 26, 287-306

uncorrected confidence interval based on U_n is $n^{-1/3}$ obtained for $h = n^{-1/3}$. Can we do better with T_n? In other words, can be obtain a confidence interval whose coverage probability tends to its asymptotic value in such a way that the difference between both quantities is a $o(n^{-1/3})$ or even a $o(n^{-2/5})$?

Consider the moment expressions of $\tilde{T}_n$. Provided that all the terms contained in the remainder terms are negligible with respect to $1/\sqrt{nh}$, $(h/g)^3$, $\sqrt{nh}h^2g^2$, the coverage probability of a confidence intervals based on T_n will be equal to the nominal level up to a $o(n^{-1/3})$ if we find h and g such that each of these three quantities has itself an order of magnitude less than $n^{-1/3}$. For this purpose, choose $h = n^{-\alpha}$ and $g = n^{-\beta}$, and select α, β, such that $\alpha < \frac{1}{3}, \beta > \frac{5}{12}(1 - 3\alpha), \beta < \alpha - \frac{1}{9}$. The set of α, β fulfilling these inequalities is a nonempty set of $\mathbb{R}^2$. In particular, it contains $\alpha = \frac{1}{4}, \beta = \frac{1}{8}$ for which

$$\frac{1}{\sqrt{nh}} = \sqrt{nh}h^2g^2 = \left(\frac{h}{g}\right)^3 = n^{-3/8}.$$

One checks that the remaining terms in the expressions of the moments of $\tilde{T}_n$ are all $o(n^{-3/8})$. This choice for h and g is optimal in the sense that for $(\alpha_o, \beta_o) = (\frac{1}{4}, \frac{1}{8})$

$$\sup_t |\mathrm{P}\{T_n \leqslant t\} - \Phi(t)| = O(n^{-3/8})$$

whereas for $(\alpha, \beta) \neq (\alpha_o, \beta_o)$

$$\lim_{n\to\infty} \sup_t n^{3/8} |\mathrm{P}\{T_n \leqslant t\} - \Phi(t)| = \infty.$$

Both assertions are proved along the lines of proposition 2.1:

$$\sup_t |\mathrm{P}\{T_n \leqslant t\} - \Phi(t)| = O\left(\frac{1}{\sqrt{nh}}, \sqrt{nh}h^2g^2, \left(\frac{h}{g}\right)^3\right)$$

and $n^{3/8}O(1/\sqrt{nh}, \sqrt{nh}h^2g^2, (h/g)^3)$ is finite if and only if $\alpha = 1/4$ and $\beta = 1/8$.

Hence we have

Proposition 2.2. *For each $t \in \mathbb{R}$, the level of the confidence interval*

$$I_{T_n} = \left[\hat{m}_h(x) - \frac{\hat{B}_n}{\sqrt{nh}} - \frac{\hat{V}_n^{1/2}}{\sqrt{nh}} z_a, +\infty\right[\tag{2.13}$$

is $a + O(n^{-3/8})$ provided $h \sim n^{-1/4}$, $g \sim n^{-1/8}$, that is to say

$$\mathrm{P}\{m(x) \in I_{T_n}\} = a + O(n^{-3/8}). \tag{2.14}$$

Thus, there exist choices for h and g such that confidence intervals for $m(x)$ using explicit bias estimation – and both undersmoothing (for $\hat{m}_h(.)$!) and oversmoothing (for $\hat{m}_g(.)$!) – are better than using strict undersmoothing to obtain an asymptotic negligible bias. "Better" is to be understood here in terms of speed of convergence of the coverage probability to the asymptotic level. Note that the choice of Cao–Abad (1991), $h \sim n^{-1/5}$, $g = n^{-1/9}$ does not fulfill our conditions and also does not achieve the higher accuracy developed here.

(1995) Härdle, W., Huet, S. and Jolivet, E.
Better Bootstrap Confidence Intervals for Regression Curve Estimation.

Statistics, 26, 287-306

On first sight our results seem to be in contradiction with claims made in Hall (1991, 1992b) that undersmoothing is preferable to explicit bias estimation. In both papers, Hall assumes more regularity than we do and he proposes to estimate the r-th derivatives which appear in the main term of the bias. We use a different type of bias estimation, without assuming more regularity than fourth-order differentiability. In fact, the bias estimate turns out to be the wild bootstrap estimate of bias as we shall see in the next chapter.

3. COMPUTING CONFIDENCE INTERVALS FROM BOOTSTRAP PROCEDURES

3.1. The Wild Bootstrap

The technique of bootstrapping from heteroscedastic errors in linear models has been introduced by Wu (1986). In the context of regression smoothing this method has been called *wild bootstrapping* by Härdle and Mammen (1991). Let us recall briefly this method. Let $\hat{\varepsilon}_i = Y_i - \hat{m}_h(X_i)$ and define

$$Y_i^* = m^*(X_i) + \varepsilon_i^*,$$

where $m^*(.)$ is some estimator of $m(.)$ and ε_i^* is a random variable with mean 0, variance $\hat{\varepsilon}_i^2$ and third moment $\hat{\varepsilon}_i^3$. For instance, the distribution of ε_i^* could be

$$\gamma\delta_{a\hat{\varepsilon}_i} + (1-\gamma)\delta_{b\hat{\varepsilon}_i}$$

with $\gamma = (5+\sqrt{5})/10$, $a = (1-\sqrt{5})/2$, $b = (1+\sqrt{5})/2$, δ_x being the Dirac measure at x. Alternatively, one can choose ε_i^* distributed as $\hat{\varepsilon}_i(Z_1/\sqrt{2} + (Z_2^2 - 1)/2)$, Z_1 and Z_2 being two independent Gaussian random variables with 0 mean and unit variance, also independent of $\hat{\varepsilon}_i$. We can then construct bootstrap versions of U_n and T_n. Let

$$\hat{m}_h^*(x) = \frac{1}{nh}\left(\sum_{i=1}^{n} Y_i^* K\left(\frac{x - X_i}{h}\right)\right) \Big/ \hat{f}_h(x) \tag{3.1}$$

$$B_n^* = \sqrt{nh}\left[\frac{1}{n}\sum_{i=1}^{n} W_{hi}(x)m^*(X_i) - m^*(x)\right] \tag{3.2}$$

$$V_n^* = \frac{h}{n}\sum_{i=1}^{n} W_{hi}^2(x)\hat{\varepsilon}_i^2. \tag{3.3}$$

Clearly, $V_n^* = \hat{V}_n$ and, if we choose $m^*(.) = \hat{m}_g(.)$, $B_n^* = \hat{B}_n$, B_n^* and V_n^* are bootstrap versions of B_n and V_n respectively, and, as such, can be used to compute the bootstrap version of S_n

$$S_n^* = \sqrt{nh}\frac{\hat{m}_h^*(x) - m^*(x)}{V_n^{*1/2}} - \frac{B_n^*}{V_n^{*1/2}}. \tag{3.4}$$

But note that B_n^* and V_n^* are not good for mimicking $\hat{B}_n$ and $\hat{V}_n$ in the "bootstrap world". To define an analogue of U_n and T_n we need to consider a bias estimate entirely in the bootstrap world.

Statistics, 26, 287-306

Define

$$U_n^* = \sqrt{nh}\frac{\sum_{i=1}^n K_h(x-X_i)\hat{\varepsilon}_i^*}{(\sum_{i=1}^n K_h(x-X_i)\hat{\varepsilon}_i^{*2})^{1/2}}$$

$$= \sqrt{nh}\frac{\hat{m}_h^*(x) - \mathrm{E}^*(\hat{m}_h^*(x))}{\hat{V}_n^{*1/2}} \tag{3.5}$$

and

$$T_n^* = \frac{\sqrt{nh}(\hat{m}_h^*(x) - m^*(x)) - \hat{B}_n^*}{\hat{V}_n^{*1/2}} \tag{3.6}$$

with

$$\hat{V}_n^* = \frac{h}{n}\sum_{i=1}^n W_{hi}^2(X)\hat{\varepsilon}_i^{*2},$$

$$\hat{\varepsilon}_i^* = Y_i^* - \hat{m}_h^*(X_i),$$

and

$$\hat{B}_n^* = \sqrt{nh}\left[\frac{1}{n}\sum_{i=1}^n W_{hi}(x)\hat{m}_g^*(X_i) - \hat{m}_g^*(x)\right]$$

with

$$\hat{m}_g^*(x) = \frac{1}{n}\sum_{i=1}^n W_{gi}(x)Y_i^*.$$

E^* is the conditional expectation with respect to observations $\{(X_1, Y_1), \ldots, (X_n, Y_n)\}$. P^* is the corresponding conditional probability. Obviously, for both above-mentioned types of wild bootstrap, $E^*(\varepsilon_i^{*k}) = c_k\hat{\varepsilon}_i^k$ for any $k \geqslant 2$, where c_k is a finite constant.

By simulation, we can obtain approximations of the distribution functions of U_n^* and T_n^*, and use their quantiles in order to construct confidence intervals for $m(x)$. Let

$$u_a^* = \inf\{u : P^*\{U_n^* \leqslant u\} = a\} \tag{3.7}$$

$$t_a^* = \inf\{t : P^*\{T_n^* \leqslant t\} = a\}. \tag{3.8}$$

Then

$$I_{U_n^*} = \left[\hat{m}_h(x) - \frac{\hat{V}_n^{1/2}}{\sqrt{nh}}u_a^*, +\infty\right[$$

and

$$I_{T_n^*} = \left[\hat{m}_h(x) - \frac{\hat{B}_n}{\sqrt{nh}} - \frac{\hat{V}_n^{1/2}}{\sqrt{nh}}t_a^*, +\infty\right[$$

are two confidence intervals for $m(x)$. Let us look at their respective levels one more time through Edgeworth expansions. We choose $m^*(.) = \hat{m}_{g^*}(.)$, a kernel estimate with bandwidth g^*.

3.2. *Asymptotic Expansion of U_n^* and Level of $I_{U_n^*}$*

As for U_n in section 2.2, we use approximations for U_n^* by

$$\hat{U}_n^* = S_n^*\left[1 - \frac{1}{2}\frac{\hat{V}_n^* - V_n^*}{V_n^*}\right].$$

Statistics, 26, 287-306

and

$$\tilde{U}_n^* = S_n^* \left[1 - \frac{1}{2} \frac{\tilde{V}_n^* - V_n^*}{V_n^*} \right].$$

We aim to prove that the approximation of the distribution of U_n by the condtional distribution of U_n^* is *second order correct*. Hence, using the quantiles of the bootstrap distribution, we are able to construct one sided confidence intervals for $m(x)$ with better coverage probabilities than those obtained by Gaussian quantiles. The proof has the following steps: First we show that the conditional distribution of $\tilde{U}_n^*$, which is a *smooth function of the mean*, admits an Edgeworth expansion. To this end, the material provided by Hall (1992), section 5.5 is adapted. Second we show that the quantiles of the conditional distribution of $\tilde{U}_n^*$ approach those of the distribution of $\tilde{U}_n$ at the desired rate. Third we show that the difference between the conditional distribution functions of $\tilde{U}_n^*$ and U_n^* is uniformly small.

Combining these steps with the closeness of the distribution functions of U_n and $\tilde{U}_n$, we obtain the following statement.

Proposition 3.1. *Let* $nh^9 \to 0$, *then under* (*A1*)–(*A3*)

$$\mathrm{P}\{U_n \leqslant u_a^*\} = a + O\left(\sqrt{nh}h^2, \frac{1}{nh}, h^4, \frac{1}{ng^*}, h^2 g^{*4} \right). \tag{3.9}$$

Moreover,

$$\mathrm{P}\{m(x) \in I_{U_n^*}\} = \mathrm{P}\left\{ m(x) \geqslant \hat{m}_h(x) - \frac{\hat{V}_n^{1/2}}{\sqrt{nh}} u_a^* \right\} = a + O\left(\sqrt{nh}h^2, \frac{1}{nh}, h^4, \frac{1}{ng^*}, h^2 g^{*4} \right). \tag{3.10}$$

Restricting ourselves, as earlier, to bandwidths of the form $h = n^{-\alpha}, g^* = n^{-\gamma}$, we have

$$h^4 = o(\sqrt{nh}h^2) \quad \text{for any choice } 0 < \alpha < 1;$$

$$\frac{1}{ng^*} = o\left(\frac{1}{nh} \right) \quad \text{for any choice } 0 < \gamma < \alpha;$$

$$h^2 g^{*4} = o(\sqrt{nh}h^2) \quad \text{for any choice } 0 < \gamma < 1;$$

$$\sqrt{nh}h^2 = o\left(\frac{1}{\sqrt{nh}} \right) \quad \text{for any choice } 1/3 < \alpha < 1;$$

For such α and γ

$$\mathrm{P}\{m(x) \in I_{U_n^*}\} = a + O\left(\sqrt{nh}, \frac{1}{nh} \right).$$

The best choice is achieved for α such that $\sqrt{nh}h^2 = 1/nh$; that is $\alpha = 3/7$ and any γ such that $0 < \gamma < \alpha$. For such a choice, we have

$$\mathrm{P}\{m(x) \in I_{U_n^*}\} = a + O(n^{-4/7}). \tag{3.11}$$

Statistics, 26, 287-306

3.3. Asymptotic expansion of T_n^* and level of $I_{T_n^*}$

Again, using an intermediate approximation $\tilde{T}_n^*$ of T_n^* and evaluating its moments we have

Proposition 3.2. *Let $nh^9 \to 0$, then under (A1)–(A3)*

$$\mathrm{P}\{T_n \leqslant t_a^*\} = a + O\left(\frac{1}{nh}, h^2g^2, \frac{1}{\sqrt{nh}}\left(\frac{h}{g}\right)^{5/2}, \frac{h^5}{g^5}, \frac{h^3}{g}, nh^5g^4, \sqrt{nh}g^4h^2, h^2g^{*4},\right.$$

$$\left.\frac{1}{ng^*}, \sqrt{nh}h^2g^2g^{*2}, \frac{1}{\sqrt{ng}}\frac{h^3}{g^2}, \frac{h}{\sqrt{ng}}\right).$$

Moreover,

$$\mathrm{P}\{m(x)\in I_{T_n^*}\} = \mathrm{P}\left\{m(x) \geqslant \hat{m}_h(x) - \frac{\hat{B}_n}{\sqrt{nh}} - \frac{\hat{V}_n^{1/2}}{\sqrt{nh}}t_a^*\right\}$$

$$= a + O\left(\frac{1}{nh}, h^2g^2, \frac{1}{\sqrt{nh}}\left(\frac{h}{g}\right)^{5/2}, \frac{h^5}{g^5}, \frac{h^3}{g}, nh^5g^4, \sqrt{nh}g^4h^2, h^2g^{*4},\right.$$

$$\left.\frac{1}{ng^*}, \sqrt{nh}h^2g^2g^{*2}, \frac{1}{\sqrt{ng}}\frac{h^3}{g^2}, \frac{h}{\sqrt{ng}}\right).$$

As above, we take $h = n^{-\alpha}, g = n^{-\beta}, g^* = n^{-\gamma}$ with $\beta < \alpha, \gamma < \alpha$. The best choice, if any, is such that h and g^2 are of the same order of magnitude, that is $2\beta = \alpha$, as it balances terms like

$$\frac{1}{\sqrt{nh}}\left(\frac{h}{g}\right)^{5/2}, \quad \frac{1}{\sqrt{nh}}\frac{h^{3/2}}{g^{1/2}}.$$

In such a case, the remaining terms simplify to

$$O\left(\frac{1}{nh}, h^3, \frac{1}{\sqrt{n}}h^{3/4}, h^{5/2}, nh^7, \sqrt{nh}h^4, h^2g^{*4}, \sqrt{nh}h^3g^{*2}, \frac{1}{ng^*}\right).$$

The choice $\alpha = \frac{2}{7}, \beta = \frac{1}{7}$ balances the terms $1/nh, h^3/g, h^5/g^5, 1/\sqrt{nh}(h/g)^{5/2}, 1/\sqrt{nh}\,h^{3/2}/g^{1/2}$, $1/\sqrt{ng}h^2/g^2, h/\sqrt{ng}$, which are of order $n^{-5/7}$, the other ones being negligible provided $\gamma \in]\frac{3}{38}, \frac{2}{7}[$. No other choice (see T_n) can give better approximation of the actual level of $I_{T_n^*}$ by its asymptotic level and, for this choice

$$\mathrm{P}\{m(x)\in I_{T_n^*}\} = a + O(n^{-5/7}). \tag{3.12}$$

Summarizing our calculations from (2.10), (2.14), (3.11), (3.12), we obtain the following table.

In that table the speed is such that, if we denote it by s, $P\{m(x)\in I\} = a + O(n^{-s})$. N is the number of kernel smoothing necessary to achieve the computation, assuming B bootstrap simulations are performed. Let us, for instance, enumerate the smoothing

(1995) Härdle, W., Huet, S. and Jolivet, E.
Better Bootstrap Confidence Intervals for Regression Curve Estimation.

Statistics, 26, 287-306

Table 1 Comparison of accuracy of confidence intervals (C.I.)

C.I.	α	β	γ	Speed	N
I_{U_n}	$\frac{1}{3}$	—	—	$\frac{1}{3}$	3
I_{T_n}	$\frac{1}{4}$	$\frac{1}{8}$	—	$\frac{3}{8}$	6
$I_{U_n^*}$	$\frac{3}{7}$	—	$]0, \frac{3}{7}[$	$\frac{4}{7}$	$5+2B$
$I_{T_n^*}$	$\frac{2}{7}$	$\frac{1}{7}$	$]\frac{3}{28}, \frac{2}{7}[$	$\frac{5}{7}$	$8+4B$

involved in the calculation of $I_{T_n^*}$:

$$\hat{f}_h, \hat{m}_h, \hat{V}_n, \hat{f}_g, \hat{m}_g, \hat{B}_n, \hat{f}_{g^*}, \hat{m}_{g^*}, \hat{V}_n^*, \hat{m}_h^*, \hat{m}_g^* \text{ and } \hat{B}_n^*,$$

the four last being done for each bootstrap simulation. This results show that, if we aim to control as well as possible the level of confidence intervals for $m(x)$, it is better to estimate explicitly the bias, and it is better to use bootstrap confidence intervals, provided the bandwidths are chosen correctly. The bandwidths vary with the methods. Nevertheless, α is always greater than 1/5, the bandwidth order which minimizes the mean square error: we always try to lower the bias by undersmoothing. On the other hand, when we estimate explicitly B_n, We use an oversmoothed estimate of the regression function ($\beta < 1/5$): the problem is no more to have a small bias, but to obtain a not too much variable estimate of $m(x)$. The choice of the bootstrap bandwidth g^*, in turn, seems to be of little importance. Obviously, this discussion is meaningful only in an asymptotic framework. The problem of bandwidth choice for fixed n is a crucial one as always in nonparametric curve estimation. As our aim is to obtain confidence intervals with level as close as possible to a prescribed one, data-driven procedures have to be imagined to choose α, β and even γ. The effectiveness of such procedures remains to be examined.

APPENDIX

Proof of Proposition 2.1

Our purpose is to prove that the distribution of U_n is uniformly approximated by the distribution of $\tilde{U}_n$, which admits an Edgeworth expansion. Let $\hat{\Delta}_n = U_n - \hat{U}_n$ and $\tilde{\Delta}_n = \hat{U}_n - \tilde{U}_n$. Obviously, for any $\delta_1 > 0$, any $\delta_2 > 0$ and any t,

$$\mathrm{P}\{\hat{U}_n \leqslant t - \delta_1\} - \mathrm{P}\{|\hat{\Delta}_n| > \delta_1\} \leqslant \mathrm{P}\{U_n \leqslant t\} \leqslant \mathrm{P}\{\hat{U}_n \leqslant t + \delta_1\} + \mathrm{P}\{|\hat{\Delta}_n| > \delta_1\}$$

and

$$\begin{aligned} \mathrm{P}\{\tilde{U}_n \leqslant t - \delta_1 - \delta_2\} - \mathrm{P}\{|\hat{\Delta}_n| > \delta_1\} - \mathrm{P}\{|\tilde{\Delta}_n| > \delta_2\} &\leqslant \mathrm{P}\{U_n \leqslant t\} \\ \leqslant \mathrm{P}\{\tilde{U}_n \leqslant t + \delta_1 + \delta_2\} + \mathrm{P}\{|\hat{\Delta}_n| > \delta_1\} + \mathrm{P}\{|\tilde{\Delta}_n| > \delta_2\}. \end{aligned} \tag{A.1}$$

We have to control $\mathrm{P}\{|\hat{\Delta}_n| > \delta_1\}$ and $\mathrm{P}\{|\tilde{\Delta}_n| > \delta_2\}$. To this end, we first evaluate how close $\tilde{V}_n$ approaches $\hat{V}_n$.

(1995) Härdle, W., Huet, S. and Jolivet, E.
Better Bootstrap Confidence Intervals for Regression Curve Estimation.

Statistics, 26, 287-306

Lemma A.1. *Under conditions (A1), (A2), (A3), $(\hat{V}_n - \tilde{V}_n)/V_n$ is of the order $O_p(h^4, 1/nh)$.*

To prove this assertion, it will be shown that $\mathrm{E}[((\hat{V}_n - \tilde{V}_n)/V_n)^2]$ is an $O(h^8, 1/n^2h^2)$, and the Markov inequality will be applied. Let us recall that

$$V_n = \frac{h}{n}\sum_{i=1}^{n} W_{hi}^2(x)\sigma^2(X_i),$$

$$\hat{V}_n = \frac{h}{n}\sum_{i=1}^{n} W_{hi}^2(x)\hat{\varepsilon}_i^2,$$

$$\tilde{V}_n = \frac{h}{n}\sum_{i=1}^{n} W_{hi}^2(x)\varepsilon_i^2,$$

so that

$$\frac{(\hat{V}_n - \tilde{V}_n)^2}{V_n^2} = \frac{1}{(\sum_{i=1}^n K_h^2(x - X_i)\sigma^2(X_i))^2}\left[\sum_{i=1}^{n} K_h^4(x - X_i)(\hat{\varepsilon}_i^2 - \varepsilon_i^2)^2 + \sum_{i\neq l} K_h^2(x - X_i)K_h^2(x - X_l)(\hat{\varepsilon}_i^2 - \varepsilon_i^2)(\hat{\varepsilon}_l^2 - \varepsilon_l^2)\right]. \tag{A.2}$$

But $\hat{\varepsilon}_i = Y_i - \hat{m}_h(X_i)$ is also equal to $\varepsilon_i - \left\{1/n\sum_{j=1}^{n}(K_h(X_i - X_j)/\hat{f}_h(X_i))\varepsilon_j\right\} - B_{n,i}/\sqrt{nh}$. Here, $B_{n,i}$ is the bias calculated for $x = X_i$, or $B_{n,i} = \sqrt{nh}\{1/n\sum_{j=1}^n W_{hj}(X_i)m(X_j) - m(X_i)\}$.

Obviously,

$$(\hat{\varepsilon}_i^2 - \varepsilon_i^2)^2 = 4\varepsilon_i^2(\hat{\varepsilon}_i - \varepsilon_i)^2 + 4\varepsilon_i(\hat{\varepsilon}_i - \varepsilon_i)^3 + (\hat{\varepsilon}_i - \varepsilon_i)^4.$$

We use well known approximations with kernel weights (see for instance Härdle (1990), section 3.1). It is not difficult to check that the leading term of the conditional expectation with respect to X_i of the right-hand-side of the above expression is dominated by

$$\frac{8}{nh}\sigma^2(X_i)\left[\int_R K^2(u)\sigma^2(X_i - hu)f(X_i - hu)\,\mathrm{d}u + \frac{1}{nh}(E^{X_i}(B_{n,i}^2))\right].$$

Moreover, the same type of calculation shows that

$$\mathrm{E}\left[\frac{\sum_{i=1}^n K_h^4(x - X_i)\frac{1}{nh}\sigma^2(X_i)\int_R K^2(u)\sigma^2(X_i - hu)f(X_i - hu)\,\mathrm{d}u}{(\sum_{i=1}^n K_h^2(x - X_i)\sigma^2(X_i))^2}\right]$$

is of order $(nh)^{-2}$ when n trends to infinity.

On the other hand, the bias at X_i admits the following decomposition:

$$\frac{B_{n,i}}{\sqrt{nh}} = \frac{1}{nh}\sum_{j\neq i}\frac{K\left(\frac{X_i - X_j}{h}\right)m(X_j)}{\hat{f}_h(X_i)} - m(X_i)\left(1 - \frac{1}{nh}\frac{K(0)}{\hat{f}_h(X_i)}\right)$$

(1995) Härdle, W., Huet, S. and Jolivet, E.
Better Bootstrap Confidence Intervals for Regression Curve Estimation.

Statistics, 26, 287-306

and the leading term of $E((nh)^{-1}B_{n,i}^2)$ is given by

$$h^4\mu(X_i)^2 + \frac{1}{nh}C_2\frac{m^2(X_i)}{f(X_i)}.$$

Thus, the leading term of

$$E\left(\frac{\sum_{i=1}^n K_h^4(x-X_i)((nh)^{-1}B_{n,i}^2)}{(\sum_{i=1}^n K_h^2(x-X_i)\sigma^2(X_i))^2}\right)$$

is a $O(h^3/n, 1/(n^2h^2))$. The result of the lemma follows using equation (A.2) and the Schwartz inequality, as h^8 dominates $n^{-1}h^3$ and $n^{-2}h^{-2}$ is dominated by both h^8 and $n^{-1}h^3$. ■

In the following two lemmas, we derive upper bounds for $P\{|\hat{\Delta}_n| > \delta_1\}$ and $P\{|\tilde{\Delta}_n| > \delta_2\}$.

Lemma A.2. *Under conditions (A1), (A2), (A3),*

$$P\{|\hat{\Delta}_n| > \delta_1\} \leqslant \frac{1}{\delta_1}O\left(\frac{1}{n^2h^2}, h^8\right) + O\left(\frac{1}{nh}, h^4\right).$$

Actually,

$$P\{|\hat{\Delta}_n| > \delta_1\} \leqslant P\{|\hat{\Delta}_n| > \delta_1 \cap R_n < 1\} + P\{R_n \geqslant 1\}$$

where $R_n = ((\hat{V}_n - V_n)/b_n)$. But $P\{R_n \geqslant 1\} \leqslant E(R_n^2) \leqslant O(1/nh, h^4)$. On the other hand, the study of the function $x \mapsto 1/\sqrt{1+x} + x/2 - 1 + s$ for $|x| < 1$ shows the existence of two functions f_l and f_u such that

$$\left|\frac{1}{\sqrt{1+x}} + \frac{x}{2} - 1\right| \leqslant s \Rightarrow f_l(s) \leqslant x \leqslant f_u(s)$$

for any s, $0 \leqslant s < 1$. With the help of Taylor expansions of $f_l(.)$ and $f_u(.)$, one shows that

$$P\{|\hat{\Delta}_n| > \delta_1 \cap R_n < 1\} \leqslant P\{|S_nR_n^2| > c\delta_1 \cap R_n < 1\}$$

for some positive constant c. Morevoer, with computations similar to those used in the proof of lemma A.1, one shows that

$$P\{|S_nR_n^2| > c\delta_1 \cap R_n < 1\} \leqslant \frac{1}{\delta_1}O\left(\frac{1}{n^2h^2}, h^8\right). \quad ■$$

Lemma A.3. *Under conditions (A1), (A2), (A3),*

$$P\{|\tilde{\Delta}_n| > \delta_2\} \leqslant \frac{1}{\delta_2}O\left(\frac{1}{n^2h^2}, h^8\right).$$

The proof follows exactly the lines of the one of lemma A.1. ■

Now take $\delta_1 = \delta_2 = \delta$, and choose $\delta = 1/nh$: $(1/\delta)h^8 = o(1/nh)$, if $\alpha > 1/9$. A direct consequence of (A.1) and of the preceding two lemmas is that

$$\sup_{-\infty < t < \infty} |P\{\tilde{U}_n \leqslant t\} - P\{U_n \leqslant t\}| \leqslant O\left(\frac{1}{nh}\right). \tag{A.3}$$

Now let us give the Edgeworth expansion of $\tilde{U}_n$.

(1995) Härdle, W., Huet, S. and Jolivet, E.
Better Bootstrap Confidence Intervals for Regression Curve Estimation.

Statistics, 26, 287-306

Lemma A.4. *Under conditions (A1), (A2), (A3), the random variables U_n is such that*

(i) *its three first moments admit the following expansions*

$$\mathrm{E}(\tilde{U}_n) = -\frac{1}{2}\frac{1}{\sqrt{nh}}\frac{1}{\sqrt{f(x)}}\frac{C_3}{C_2^{3/2}}\frac{\gamma_3(x)}{\sigma^3(x)} + O\left(\frac{1}{nh}, h^4\right),$$

$$\mathrm{E}(\tilde{U}_n^2) = 1 + O\left(\frac{1}{nh}, h^4\right),$$

$$\mathrm{E}(\tilde{U}_n^3) = -\frac{7}{2}\frac{1}{\sqrt{nh}}\frac{1}{\sqrt{f(x)}}\frac{C_3}{C_2^{3/2}}\frac{\gamma_3(x)}{\sigma^3(x)} + O\left(\frac{1}{nh}, h^4\right),$$

(ii) *its distribution function of $\tilde{U}_n$ admits the following Edgeworth expansion*

$$\mathrm{P}\{\tilde{U}_n \leqslant t\} = \Phi(t) + \frac{1}{6\sqrt{nh}}\phi(t)\frac{1}{\sqrt{f(x)}}\frac{C_3}{C_2^{3/2}}\frac{\gamma_3}{\sigma^3(x)}(2t^2+1) + O\left(\frac{1}{nh}, h^4\right).$$

The proof of (*i*) is an evaluation of expressions of the same type than those considered in lemma A.1. The proof of (ii) parallels proofs of similar statements appearing in section 5.5 of Hall (1992a). Indeed, $\tilde{U}_n$ is a smooth function of means of i.i.d. random variables. Let

$$Z_{n1} = \frac{1}{\sqrt{nh}}\sum_{i=1}^n \left\{K\left(\frac{x-X_i}{h}\right)\varepsilon_i - \mathrm{E}\left[K\left(\frac{x-X_i}{h}\right)\varepsilon_i\right]\right\},$$

$$Z_{n2} = \frac{1}{\sqrt{nh}}\sum_{i=1}^n \left\{K^2\left(\frac{x-X_i}{h}\right)\varepsilon_i^2 - \mathrm{E}\left[K^2\left(\frac{x-X_i}{h}\right)\varepsilon_i^2\right]\right\},$$

$$Z_{n3} = \frac{1}{\sqrt{nh}}\sum_{i=1}^n \left\{K^2\left(\frac{x-X_i}{h}\right)\sigma^2(X_i) - \mathrm{E}\left[K^2\left(\frac{x-X_i}{h}\right)\sigma^2(X_i)\right]\right\},$$

$$v_2(x) = \frac{1}{h}\mathrm{E}\left[K^2\left(\frac{x-X_i}{h}\right)\sigma^2(X_i)\right] = \frac{1}{h}\mathrm{E}\left[K^2\left(\frac{x-X_i}{h}\right)\varepsilon_i^2\right].$$

Thus, $\tilde{U}_n = g(Z_{n1}, Z_{n2}, Z_{n3})$ with

$$g(u,v,w) = \frac{1}{2}\frac{u(2v_2(x) + (3w-v)(nh)^{-1/2})}{(v_2(x) + w(nh)^{-1/2})}.$$

From now on, we follow the lines of Hall (1992a), paragraph 5.5.2, with adaptations when needed. Providing a result analogous to the Lemma 5.6, p. 273, is the main modification. As a matter of fact, we need the following statement.

Lemma A.5. *Under conditions (A1), (A2), (A3), we have for any $\delta > 0$*

$$\sup_{|t_1|+|t_2|+|t_3|>\delta}\left|\int_{R^2} h\exp\{i(t_1K(u)z + t_2K^2(u)z^2 + t_3K^2(u)\sigma^2(x-hu))\}p_{X,\varepsilon}(x-hu,z)\,\mathrm{d}u\,\mathrm{d}z\right|$$

$$< 1 - C(x,\delta)h.$$

Statistics, 26, 287-306

Let us note that, as a consequence of condition (A2), the Cramer condition is fulfilled for the distribution of ε conditional to X. More precisely, $p_{\varepsilon/X}(./x)$ being the distribution density of ε given $\{X = x\}$, $\forall x \in I$, $\limsup_{|t| \to \infty} |\int_R \exp(itz) p_{\varepsilon/X}(z/x)\, dz| < 1$. ■

Remark. Assumption (A2) may certainly be somewhat relaxed. Nevertheless, assuming existence of moments of any order for ε makes the proofs easier, since they are always simply based on the Schwartz inequality.

The uniform bound in (A.3) together with lemma A.4 completes the proof of proposition 2.1.

Bias Estimation

The asymptotic pivot T_n depends on the bias. We only give here a few examples of the treatment of expressions including B_n and its approximations.

Lemma A.6. *$\hat{B}_n - B_n$ is of order $O_p(\sqrt{nh}g^2h^2)$ if $\alpha > \beta > 1/9$.*

One checks easily that

$$\hat{B}_n - B_n = \sqrt{nh}\left\{ A(\varepsilon) + n^{-1} \sum_{i=1}^{n} W_{hi}(x)[\varphi_g(X_i) - \varphi_g(x)] \right\}$$

with

$$A(\varepsilon) = n^{-1} \sum_{i=1}^{n} W_{hi}(x)\left[n^{-1} \sum_{j=1}^{n} (W_{gj}(X_i) - W_{gj}(x))\varepsilon_j \right]$$

$$\varphi_g(x) = n^{-1} W_{gj}(x) m(X_j) - m(x).$$

Using well known approximation techniques, the following asymptotic relations are obtained:

$$\mathrm{E}(A(\varepsilon)^2) \sim \frac{1}{4} \frac{1}{ng} \frac{h^4}{g^4} \frac{\sigma^2(x)}{f(x)} K^{(2)^2} \int L''^2(v)\, dv$$

$$\mathrm{E}\left(n^{-1} \sum_{i=1}^{n} W_{hi}(x) |\varphi_g(X_i) - \varphi_g(x)| \right) \sim \frac{g^2h^2}{4} L^{(2)} K^{(2)} \left(\rho''(x) + 2\rho'(x) \frac{f'(x)}{f(x)} \right)$$

where $\rho(x) = |\mu(x)|$. Using the Markov inequality, the conclusion of the lemma follows. ■

On the other hand, $\hat{B}_n$ and $\tilde{B}_n$ are close to each other.

Lemma A.7. $E(\hat{B}_n - \tilde{B}_n)^2 = O(1/(ng))$.

Recall that Ω_{ij} and Ξ_{ij} are given by equations (2.11) and (2.12) respectively. $\tilde{B}_n$ is an approximation of $\hat{B}_n$, using the Hausdorff decomposition $\mathrm{E}(\Omega_{ij}/X_i) + \mathrm{E}(\Omega_{ij}/X_j) - \mathrm{E}(\Omega_{ij})$ in place of Ω_{ij}, and similarly for Ξ_{ij}. This leads to the order of $O(1/(ng))$ for the mean square error. ■

Note that $1/(ng)$ tends to 0 at a higher rate than the remainder terms of the moments of $\tilde{T}_n$.

(1995) Härdle, W., Huet, S. and Jolivet, E.
Better Bootstrap Confidence Intervals for Regression Curve Estimation.

Statistics, 26, 287-306

Proof of Proposition 3.1

In order to prove proposition 3.1, we follow the indicated steps.

The first step is given in lemma A.8.. To prepare the following, note that as for U_n, an approximation of U_n^* depending on S_n^* and $\tilde{V}_n^*$ is needed:

$$\tilde{U}_n^* = S_n^* \left[1 - \frac{1}{2} \frac{\tilde{V}_n^* - \hat{V}_n}{\hat{V}_n} + O\left[\left(\frac{\tilde{V}_n^* - \hat{V}_n}{\hat{V}_n} \right)^2 \right] \right].$$

As for the proof of theorem 5.6 in Hall (1992a), the Edgeworth expansion of the conditional distribution of $\tilde{U}_n^*$ can be obtained by adapting the proof of lemma A.4 to the bootstrap case.

Lemma A.8. *Under conditions (A1), (A2), (A3), the conditional distribution of $\tilde{U}_n^*$ is such that:*

$$\sup_{-\infty<t<\infty} \left| P^*\{\tilde{U}_n^* \leqslant t\} - \Phi(t) - \frac{1}{6\sqrt{nh}} \frac{\sum_{i=1}^n K_h^3(x - X_i)\hat{\varepsilon}_i^3}{(\sum_{i=1}^n K_h^2(x - X_i)\hat{\varepsilon}_i^2)^{3/2}} (2t^2 + 1) \right|$$

$$= O\left(\frac{1}{nh}, h^4, \frac{1}{ng^*}, h^2 g^{*4} \right).$$

The second step is analogous to theorem 5.7 in Hall (1992a), p. 282. Define $\tilde{u}_a$ by $P^*\{\tilde{U}_n^* \leqslant \tilde{u}_a\} = a$.

Lemma A.9. *Under conditions (A1), (A2), (A3),*

$$\mathrm{P}\{\tilde{U}_n \leqslant \tilde{u}_a\} = a + O\left(\frac{1}{nh}, h^4, \frac{1}{ng^*}, h^2 g^{*4} \right)$$

For a rigorous treatment, we need to compute the following moments,

$$\mathrm{E}(\tilde{U}_n^*) = -\frac{1}{2} \frac{1}{\sqrt{nh}} \frac{1}{f(x)} \frac{C_3}{C_2^{3/2}} \frac{\gamma_3(x)}{\sigma^3(x)} + O\left(\frac{1}{nh}, h^4, \frac{1}{ng^*}, h^2 g^{*4} \right) \tag{A.4}$$

$$\mathrm{E}(\tilde{U}_n^{*2}) = 1 + O\left(\frac{1}{nh}, h^4, \frac{1}{ng^*}, h^2 g^{*4} \right) \tag{A.5}$$

$$\mathrm{E}(\tilde{U}_n^{*3}) = -\frac{7}{2} \frac{1}{\sqrt{nh}} \frac{1}{f(x)} \frac{C_3}{C_2^{3/2}} \frac{\gamma_3(x)}{\sigma^3(x)} + O\left(\frac{1}{nh}, h^4, \frac{1}{ng^*}, h^2 g^{*4} \right). \tag{A.6}$$

This can be seen by the evaluation of conditional moments of $\tilde{U}_n^*$. In Lemma A.10. we give the corresponding conditional moments of (A.4) as an example. The other moments are calculated in the same way.

(1995) Härdle, W., Huet, S. and Jolivet, E.
Better Bootstrap Confidence Intervals for Regression Curve Estimation.

Statistics, 26, 287-306

Lemma A.10. *Under conditions (A1), (A2), (A3),*

$$\mathrm{E}^*(\tilde{U}_n^*) = \mathrm{E}(\tilde{U}_n) + O_p\left(\frac{1}{nh}, h^4, \frac{1}{ng^*}, h^2 g^{*4}\right).$$

Actually

$$\mathrm{E}^*\left(S_n^* \frac{\tilde{V}_n^* - \hat{V}_n}{\hat{V}_n}\right) = \frac{\sum_{i=1}^n K_h^3(x - X_i)\hat{\varepsilon}_i^3}{(\sum_{i=1}^n K_h^2(x - X_i)\hat{\varepsilon}_i^2)^{3/2}}$$

$$= \left[\mathrm{E}\left(S_n \frac{\tilde{V}_n - V_n}{V_n}\right) + \frac{\sum_{i=1}^n K_h^3(x - X_i)(\hat{\varepsilon}_i^3 - \gamma_3(X_i))}{(\sum_{i=1}^n K_h^2(x - X_i)\sigma^2(X_i))^{3/2}}\right]$$

$$\times \left(1 - \frac{3}{2}\frac{\hat{V}_n - V_n}{V_n} + O\left(\left(\frac{\hat{V}_n - V_n}{V_n}\right)^2\right)\right).$$

It is not hard to see that the expectation of the second term inside the square brackets on the right hand side of the preceding equation is a $O_p(1/(nh), h^2/\sqrt{nh})$. For $h = n^{-\alpha}$, $0 < \alpha < 1$, $h^2/\sqrt{\mathrm{nh}} \leqslant \min\{1/(nh), h^4\}$. The proof of the lemma follows. ■

The third step is based on the following two lemmas.

Lemma A.11. *If h/g^* tends to 0 as n tends to infinity, $(\tilde{V}_n^* - \hat{V}_n^*)/\hat{V}_n$ is a $O_p(1/(nh), h^4, 1/(ng)^*, h^2 g^{*4})$.*

The proof is almost the same as the one of lemma A.1. Indeed, if $\hat{\varepsilon}_i^*, \varepsilon_i^*$ and $\hat{\varepsilon}_i$ are used instead of $\hat{\varepsilon}_i$, ε_i and $\sigma(X_i)$ respectively in formula (A.2), we obtain an expended expression of $((\tilde{V}_n^* - \hat{V}_n^*)/\hat{V}_n)^2$. It turns out that its expectation is of order $O(1/(n^2h^2), h^8, 1/(n^2g^{*2}), h^4 g^{*8})$. The Markov inequality yields the result. ■

Lemma A.12. *If h/g^* tends to 0 as n tends to infinity, $(\tilde{V}_n^* - \hat{V}_n)/\hat{V}_n$ is a $O_p(1/\sqrt{nh})$.*

$$\mathrm{E}^*\left(\frac{\tilde{V}_n^* - \hat{V}_n}{\hat{V}_n}\right)^2 = \frac{\sum_{i=1}^n K_h^4(x - X_i)\hat{\varepsilon}_i^4(c_4 - 1)}{(\sum_{i=1}^n K_h^2(x - X_i)\hat{\varepsilon}_i^2)^2},$$

thus its expectation is of order $1/(nh)$. The Markov inequality yields the result. ■

Obviously, we have a relation between the conditional distributions of $\tilde{U}_n^*$ and U_n^* analogous to relation (A.1). This leads to the counterpart of equation (A.3) in the bootstrap world

$$\sup_{-\infty < t < \infty} |P^*\{\tilde{U}_n^* \leqslant t\} - P^*\{U_n^* \leqslant t\}| \leqslant O_p\left(\frac{1}{nh}\right). \tag{A.7}$$

Hence, $nh|P^*\{\tilde{U}_n^* \leqslant u_a^*\} - a|$ is bounded in probability. The proof of proposition (3.1) is complete.

(1995) Härdle, W., Huet, S. and Jolivet, E.
Better Bootstrap Confidence Intervals for Regression Curve Estimation.

Statistics, 26, 287-306

Moments of $\tilde{T}_n^*$

Using methods of calculations along the lines of those used to prove lemmas A.10, A.11, A.12, the moments of T_n^* are obtained.

$$\mathrm{E}(\tilde{T}_n^*) = -\frac{1}{2}\frac{1}{\sqrt{nh}}\frac{1}{\sqrt{f(x)}}\frac{C_3}{C_2^{3/2}}\frac{\gamma_3(x)}{\sigma^3(x)}$$

$$-\frac{1}{4}\sqrt{nh}g^2h^2\frac{\sqrt{f(x)}}{\sigma(x)}\frac{1}{C_2^{1/2}}L^{(2)}K^{(2)}\left(\mu''(x)+2\mu'(x)\frac{f'(x)}{f(x)}\right)$$

$$+O\left(g^2h^2\frac{1}{\sqrt{nh}}\frac{h^{3/2}}{g^{1/2}},\frac{1}{\sqrt{nh}}\left(\frac{h}{g}\right)^{5/2},\frac{1}{nh},\frac{1}{ng^*},h^2g^{*4},\sqrt{nh}h^2g^2g^{*2},\frac{1}{\sqrt{ng}}\frac{h^2}{g^2},\frac{h}{\sqrt{ng}}\right).$$

$$\mathrm{E}(\tilde{T}_n^{*2}) = 1-\left(\frac{h}{g}\right)^3K''(0)\frac{K^{(2)}}{C^2}$$

$$+O\left(\frac{1}{nh},h^2g^2,\frac{1}{\sqrt{nh}}\left(\frac{h}{g}\right)^{5/2},\frac{h^3}{g},\frac{1}{\sqrt{nh}},\frac{h^{3/2}}{\sqrt{nh}},h^2g^{*4},\frac{1}{ng^*},\right.$$

$$\left.\sqrt{nh}h^2g^2g^{*2},\frac{1}{\sqrt{ng}}\frac{h^2}{g^2},\frac{h}{\sqrt{ng}}\right).$$

$$\mathrm{E}(\tilde{T}_n^{*3}) = -\frac{7}{2}\frac{1}{\sqrt{nh}}\frac{1}{\sqrt{f(x)}}\frac{C_3}{C_2^{3/2}}\frac{\gamma_3(x)}{\sigma^3(x)}$$

$$-\frac{3}{4}\sqrt{nh}g^2h^2\frac{L^{(2)}K^{(2)}}{C_2^{1/2}}\frac{\sqrt{f(x)}}{\sigma(x)}\left(\mu''(x)+2\mu'(x)\frac{f'(x)}{f(x)}\right)$$

$$+O\left(\frac{1}{nh}h^2g^2,\frac{1}{\sqrt{nh}}\left(\frac{h}{g}\right)^{5/2},\frac{h^5}{g^4},\frac{h^3}{g},nh^5g^4,\sqrt{nh}g^4h^2,h^2g^{*4},\frac{1}{ng^*},\sqrt{nh}h^2g^2g^{*2},\right.$$

$$\left.\frac{1}{\sqrt{ng}}\frac{h^2}{g^2},\frac{h}{\sqrt{ng}}\right).$$

If we assume that all the terms in the remainder terms are negligible with respect to $\max(1/\sqrt{nh}, \sqrt{nh}h^2g^2)$, proposition 3.2 is proved.

Acknowledgements

The research for this paper was partly carried out within Sonderforschungsbereich 373 at the Humboldt University Berlin and was printed using funds made available by the Deutsche Forschungsgemeinschaft. The authors express their gratefulness to one annonymous referee, who helped them to improve considerably a previous version of the manuscript.

Statistics, 26, 287-306

References

Barndorff-Nielsen, O. E. and Cox, D. R. (1989). *Asymptotic Technics for Use in Statistics.* Chapman and Hall, London.

Cao-Abad, R. (1991). Rate of convergence for the wild bootstrap in nonparametric regression. *Ann. Statist.* **19**, 2226–2231.

Franke, J. and Härdle, W. (1992). On bootstrapping kernel spectral estimates. *Ann. Statist.* **20**, 121–145.

Hall, P. (1991). Edgeworth expansions for nonparametric density estimators with applications. *Statistics* **22**, 215–232.

Hall. P. (1992a). *The Bootstrap and Edgeworth Expansion.* Springer Verlag, New York.

Hall, P. (1992b). On bootstrap confidence intervals in nonparametric regression. *Ann. Statist.* **20**, 695–711.

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

Härdle, W. and Bowman, A. (1988). Bootstrapping in nonparametric regression: local adaptive smoothing and confidence bands. *J. Amer. Statist. Assoc.* **83**, 102–110.

Härdle, W. and Mammen, E. (1990). Bootstrap methods in nonparametric regression. In *Non-parametric Functional Estimation and Related Topics*, G. Roussas ed., Kluwer Publishing Company, Series C: Mathematical and Physical Sciences **335**, 111–124.

Härdle, W. and Marron, J. S. (1991). Bootstrap simultaneous error bars for nonparametric regression. *Ann. Statist.* **19**, 778–796.

Petrov, V. V. (1975). *Sums of Independent Random Variables*, Springer Verlag, Berlin.

Wu, C. F. J. (1986). Jackknife, bootstrap and other resampling methods in regression analysis. *Journal of Statistics* **14**, 1261–343.

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

S. Huet, E. Jolivet
INRA
Laboratoire de Biométrie
F-78352 Jouy-en-Josas
France

(1995) Härdle, W., Huet, S. and Jolivet, E.
Better Bootstrap Confidence Intervals for Regression Curve Estimation.

Journal of Productivity Analysis, 6, 63-76.

Iterated Bootstrap with Applications to Frontier Models*

PETER HALL
Centre for Mathematics and its Applications, Australian National University, Canberra, A.C.T. 0200, Australia
halpstat@fac.anu.edu.au

WOLFGANG HÄRDLE
Institut für Statistik und Ökonometrie, Humboldt Universität zu Berlin, Spandauer Str. 1, 10178 Berlin, Germany
hardle@wiwi.hu-berlin.de

LÉOPOLD SIMAR
Institut de Statistique and CORE, Université Catholique de Louvain, 34 voie du Roman Pays, B-1348 Louvain-la-Neuve, Belgium
simar@stat.ucl.ac.be

Abstract

The iterated bootstrap may be used to estimate errors which arise from a single pass of the bootstrap and thereby to correct for them. Here the iteration is employed to correct for coverage probability of confidence intervals obtained by a percentile method in the context of production frontier estimation with panel data. The parameter of interest is the maximum of the intercepts in a fixed firm effect model. The bootstrap distribution estimators are consistent if and only if there are no ties for this maximum. In the regular case (no ties), poor distribution estimators can result when the second largest intercept is close to the maximum. The iterated bootstrap is thus suggested to improve the accuracy of the obtained distributions. The result is illustrated in the analysis of labor efficiency of railway companies.

Keywords. Bootstrap, iterated bootstrap, frontier models, railways efficiency, percentile method.

1. Introduction

The iterated bootstrap may be used to estimate errors which arise from a single pass of the bootstrap, and thereby to correct for them. Iteration may be employed in a variety of contexts where the bootstrap is inefficaceous, such as bias estimation, hypothesis testing and confidence interval construction. We shall consider only the latter case, since it is the one appropriate to the context of our work, and we shall focus specifically on the particular regression problem at hand, i.e., production frontier estimation with panel data.

Estimation of production frontiers provides a yardstick to measure the (technical) efficiencies of production units. The original work on this problem dates back to Farrell (1957): the efficiency of a production unit is characterized by the distance between the output (production) level attained by this unit and the level it should obtain if it were efficient. The latter is defined and the maximal output attainable for a given combination of inputs; the geometric locus of the optimal productions may be represented by a production frontier.

*This work was supported in part by grant No. 26 from the program "Pôle d'attraction interuniversitaire—Deuxième phase" to CORE and by the contract "Projet d'Actions de Recherche Concertées of the Belgian government" (PARC) to the Institute of Statistics, Université Catholique de Louvain. The first author was partly financed by the Institut de Mathématiques Appliquées, Université Catholique de Louvain.

From a statistical point of view, this frontier will be estimated from a set of observations of particular production units. Then the efficiencies have to be defined by the distances to the obtained frontier.

In this paper we concentrate our analysis on the parameter defining the level of the frontier and analyze its bootstrap distribution. We shall assume that at the first level of the bootstrap, the percentile bootstrap method is employed to construct confidence intervals. Thus, the purpose of the iterated bootstrap in our problem is to correct a percentile confidence interval for coverage error. In the context of frontier estimation there are few alternatives to the percentile method. For example, percentile-t is often not a viable option, owing to the difficulty of accurately estimating the variance of the estimates.

Section 2 describes the model and the first level of the bootstrap, and Section 3 discusses inaccuracies in that approach. Some of those inaccuracies may be remedied by iteration, which is introduced in Section 4. Implementation of the iterated bootstrap requires two layers of Monte Carlo simulation, which are described in Section 5. The method is illustrated through the analysis of labor efficiencies of railways in Section 6. Technical details, in particular for our discussion of properties of the single level bootstrap, are deferred to the Appendices.

The iterated bootstrap for confidence intervals was introduced by Hall (1986), Beran (1987) and Loh (1987). Unified accounts are given by Hall and Martin (1988) and Hall (1992, Section 1.4).

2. Percentile Method: The First Bootstrap Level

Assume that the data $\mathfrak{X} = \{(x_{ij}, y_{ij}), 1 \le i \le p, 1 \le j \le n_i\}$ are generated by the model

$$y_{ij} = \gamma_i + \beta x_{ij} + \epsilon_{ij}\sigma_i,$$

where γ_i, β and $\sigma_i > 0$ are unknown parameters and the ϵ_{ij}'s are independent and identically distributed random variables with zero mean and unit variance.

In the context of frontier models with panel data, this is a fixed effect model where i is the firm index and j may be the time index, y_{ij} is the output, x_{ij} are the inputs and γ_i represents the firm effect.

The production frontier corresponds to the maximal attainable output given the inputs and is generally determined by the highest level of the firm effects. Thus a parameter of particular interest will be $\max_{1 \le i \le p} \gamma_i$.

A measure of technical inefficiency of the firms is then simply given by the distance of each firm to the obtained frontier: $\alpha_i = \gamma_i - \max_i \gamma_i$. [For details see, e.g., Schmidt and Sickles (1984) or Simar (1992)]. In the sequel, we focus our attention on the parameter $\max_{1 \le i \le p} \gamma_i$.

Many possible methods are available for estimation of γ_i, β and σ_i. The technique used in the numerical work for this paper is weighted least-squares, although many other approaches, such as robust methods, could have been used. Let the method be $\mathfrak{M}$; its choice has no bearing on the veracity of our algorithm for either the first or the second bootstrap level. Let $\hat{\gamma}_i$, $\hat{\beta}$, $\hat{\sigma}_i$ denote the estimates computed from $\mathfrak{X}$ using method $\mathfrak{M}$, and define the residuals

$$\tilde{\epsilon}_{ij} = \frac{(y_{ij} - \hat{\gamma}_i - \hat{\beta}x_{ij})}{\hat{\sigma}_i},$$

and their centered, scaled counterparts

$$\hat{\epsilon}_{ij} = \hat{\tau}^{-1}(\epsilon_{ij} - \hat{\nu}),$$

where $\hat{\nu} = n^{-1}\Sigma\Sigma\tilde{\epsilon}_{ij}$, $\hat{\tau}^2 = n^{-1}\Sigma\Sigma(\tilde{\epsilon}_{ij} - \hat{\nu})^2$ and $n = \Sigma n_i$. (Depending on the method $\mathfrak{M}$, one or other of these standardizing operations may be unnecessary, since the raw residuals ϵ_{ij} may satisfy $\hat{\nu} = 0$ or $\hat{\tau}^2 = 1$).

We now introduce the bootstrap. Conditional on the data $\mathfrak{X}$, let $\mathcal{E}^* = \{\epsilon^*_{ij}, 1 \le i \le p, 1 \le j \le n_i\}$ denote a resample drawn randomly, with replacement, from the standardized residuals $\{\hat{\epsilon}_{ij}, 1 \le i \le p, 1 \le j \le n_i\}$. Define

$$y^*_{ij} = \hat{\gamma}_i + \hat{\beta}x_{ij} + \epsilon^*_{ij}\hat{\sigma}_i$$

and $\mathfrak{X}^* = \{(x_{ij}, y^*_{ij}), 1 \le i \le p, 1 \le j \le n_i\}$. Using method $\mathfrak{M}$ applied to the resample $\mathfrak{X}^*$, compute bootstrap versions $\hat{\gamma}^*_i$, $\hat{\beta}^*$, $\hat{\sigma}^*_i$ of $\hat{\gamma}_i$, $\hat{\beta}$, $\hat{\sigma}_i$. Put

$$m = \max_{1 \le i \le p} \gamma_i, \qquad \hat{m} = \max_{1 \le i \le p} \hat{\gamma}_i, \qquad \hat{m}^* = \max_{1 \le i \le p} \hat{\gamma}^*_i.$$

An equal-tailed, percentile method, nominal $(1 - \alpha)$-level confidence interval $\mathcal{I}_\alpha$ for m may be constructed as follows. Define $\hat{u}_t$ to be the solution of the equation

$$p(\hat{m}^* \le \hat{u}_t | \mathfrak{X}) = t, \quad 0 < t < 1,$$

and take $\mathcal{I}_\alpha = (\hat{u}_{\alpha/2}, \hat{u}_{1-\alpha/2})$.

3. Properties of the Percentile Method

As we show in Appendix (i), the bootstrap estimator of the distribution of $\hat{m} - m$ is consistent if and only if there are no ties for the max of the γ_i's. That is not a serious problem in itself, since it is unlikely that in practice two values of γ_i will be tied precisely for m. However, if the second largest γ_i is close to the largest γ_i, while not being precisely equal to it, then there can be problems, in small to moderate samples, with accurately estimating the distribution of $\hat{m} - m$. These will be reflected in coverage inaccuracies of the confidence interval $\mathcal{I}_\alpha$.

One approach to removing this difficulty is to use a method which produces intervals enjoying particularly accurate coverage properties in regular circumstances (i.e., no ties). It is then to be expected that the performance of the method will continue to produce good results as the largest and second largest γ_i move closer together; at least, it will give better accuracy than a more "average" method such as percentile. We suggest bootstrap iteration of the percentile method as a technique for producing such particularly accurate intervals.

This method does indeed give good results (compared with percentile) if the γ_i's are close, as we found in a simulation study [Hall, Härdle, Simar (1993)].

Hall (1991, section 1.4) gives a detailed account on the differences between percentile and percentile-t methods. In particular the bootstrap approximations of quantiles are in error by $O_p(n^{-1})$ for the percentile-t and of the order $n^{-1/2}$ for the percentile method. For good performance of the percentile-t method in practice it requires that the estimator of the variance of $n^{1/2}\hat{m}$ be accurate. This can be quite a problem in small to moderate samples when the largest and second-largest values of γ_i are close, although not tied. Let $\gamma_{i_0} = m$, the variance of $\hat{\gamma}_{i_0}$ is often not a good approximation to that of $\hat{m}$ in this case. That is the reason we have not pursued the percentile-t method in this paper and rather used the percentile method with the improvement by iteration which induces an error of magnitude n^{-1}.

4. Iterated Percentile Method: The Second Bootstrap Level

As noted in Section 3, the interval $\mathcal{I}_\alpha$ is asymptotically correct as a $(1 - \alpha)$-level confidence interval for m, provided there are no ties for $\max \gamma_i$. However, in finite samples the error $P(m \in \mathcal{I}_\alpha) - (1 - \alpha)$ may be significant. In the present section we show that a method based on the iterated bootstrap may be used to estimate the true coverage,

$$\pi(\alpha) = P(m \in \mathcal{I}_\alpha), \tag{4.1}$$

and thereby to estimate the value of α for which the true coverage is equal to a predetermined level $1 - \alpha$ such as 0.95. If $\hat{\pi}(\alpha)$ is our estimator of $\pi(\alpha)$ then we seek the solution $\hat{\alpha}$ of the equation

$$\hat{\pi}(\hat{\alpha}) = 1 - \alpha_0, \tag{4.2}$$

and take $\mathcal{I}_{\alpha_0} = \mathcal{I}_{\hat{\alpha}}$ as our final confidence interval.

We shall adopt the algorithm, and the notation, described in Section 2, and extend it as follows. Define the bootstrap residuals

$$\tilde{\epsilon}_{ij}^* = \frac{(y_{ij}^* - \hat{\gamma}_i^* - \hat{\beta}^* x_{ij})}{\hat{\sigma}_i^*},$$

and their centered, scaled counterparts

$$\hat{\epsilon}_{ij}^* = \hat{\tau}^{*-1}(\tilde{\epsilon}_{ij}^* - \hat{\nu}^*),$$

where $\hat{\nu}^* = n^{-1}\Sigma_i\Sigma_j\tilde{\epsilon}_{ij}^*$, $\hat{\tau}^{*2} = n^{-1}\Sigma_i\Sigma_j(\tilde{\epsilon}_{ij}^* - \hat{\nu}^*)^2$. Let $\{\epsilon_{ij}^{**}, 1 \le i \le p, 1 \le j \le n_i\}$, denote a re-resample drawn randomly, with replacement, from the collection $\{\hat{\epsilon}_{ij}^*, 1 \le i \le p, 1 \le j \le n_i\}$, and define

$$y_{ij}^{**} = \hat{\gamma}_i^* + \hat{\beta}^* x_{ij} + \epsilon_{ij}^{**}\hat{\sigma}_i^*$$

and $\mathfrak{X}^{**} = \{(x_{ij}, y_{ij}^{**}),\ 1 \le i \le p,\ 1 \le j \le n_i\}$. Using method $\mathfrak{M}$ applied to the resample $\mathfrak{X}^{**}$, compute the second-level bootstrap estimators $\hat{\gamma}_i^{**}$, and put $\hat{m}^{**} = \max \hat{\gamma}_i^{**}$. Define $\hat{u}_t^*$ to be the solution of the equation

$$P(\hat{m}^{**} \le \hat{u}_t^* | \mathfrak{X}^*) = t, \quad 0 < t < 1.$$

The resulting version of $\mathfrak{I}_\alpha = (\hat{u}_{\alpha/2}, \hat{u}_{1-\alpha/2})$ is $\mathfrak{I}_\alpha^* = (\hat{u}_{\alpha/2}^*, \hat{u}_{1-\alpha/2}^*)$.

Now repeat both levels of the bootstrap many times, and record $\hat{\pi}(\alpha)$, the proportion of times that $\mathfrak{I}_\alpha^*$ covers $\hat{m}$:

$$\hat{\pi}(\alpha) = P(\hat{m} \in \mathfrak{I}_\alpha^* | \mathfrak{X}).$$

This quantity is the derived estimator of the function $\pi(\alpha)$, defined at (4.1), and leads to the iterated bootstrap confidence interval $\mathfrak{I}_{\alpha_0}$ defined in that paragraph.

In practice we would only compute the value of $\hat{\pi}(\alpha)$ for a small number of values α (often five or ten) in the vicinity of the derived value of α_0. In other α's, the value of $\hat{\pi}(\alpha)$ would be approximated via linear interpolation, and equation (4.2) solved for this approximate form of the function.

5. Implementation

Here we describe in detail the implementation of both levels of the iterated bootstrap, using Monte Carlo simulation to approximate the quantities involved.

In this paragraph we develop the first bootstrap level. Take B_1 to be a large number, and let $\mathcal{E}_b^* = \{\epsilon_{bij}^*,\ 1 \le i \le p,\ 1 \le j \le n_i\}$, for $1 \le b \le B_1$, denote independent versions of $\mathcal{E}^*$. That is, the resamples $\mathcal{E}_b^*$ are drawn independently by sampling randomly, with replacement, from the collection of all standardized residuals $\hat{\epsilon}_{ij}$. Put

$$y_{bij}^* = \hat{\gamma}_i + \hat{\beta} x_{ij} + \epsilon_{bij}^* \hat{\sigma}_i$$

and $\mathfrak{X}_b^* = \{(x_{ij}, y_{ij}^*),\ 1 \le i \le p,\ 1 \le j \le n_i\}$. Let $\hat{\gamma}_{bi}^*$, $\hat{\beta}_b^*$, $\hat{\sigma}_{bi}^*$ denote the resulting versions of $\hat{\gamma}_i^*$, $\hat{\beta}^*$, $\hat{\sigma}_i^*$, and define $\hat{m}_b^* = \max_i \hat{\gamma}_{bi}^*$. Our Monte Carlo approximation to the probability $q(u) = P(\hat{m}^* \le u\ \mathfrak{X})$ is

$$\hat{q}_{B_1}(u) = B_1^{-1} \sum_{b=1}^{B_1} I(\hat{m}_b^* \le u),$$

where $I(\cdot)$ denotes the indicator function. We take the inverse of $\hat{q}_{B_1}$ to be

$$\hat{u}_{B_1,t} = \hat{q}_{B_1}^{-1}(t) = \inf\{u : \hat{q}_{B_1}(u) \ge t\}.$$

The resulting approximation to $\mathfrak{I}_\alpha$ is $\mathfrak{I}_{B_1\alpha} = (\hat{u}_{B_1,\alpha/2}, \hat{u}_{B_1,1-\alpha/2})$.

Next we develop the second bootstrap level. Put

$$\tilde{\epsilon}^*_{bij} = \frac{y^*_{bij} - \hat{\gamma}^*_{bi} - \hat{\beta}^*_b x_{ij})}{\hat{\sigma}^*_{bi}}, \qquad \hat{\epsilon}^*_{bij} = \hat{\tau}_b^{*-1}(\tilde{\epsilon}^*_{bij} - \hat{\nu}^*_b)$$

where $\hat{\nu}^*_b = n^{-1}\Sigma_i\Sigma_j\tilde{\epsilon}^*_{bij}$, $\hat{\tau}^{*2}_b = n^{-1}\Sigma_i\Sigma_j(\tilde{\epsilon}^*_{bij} - \hat{\nu}^*_b)^2$. Take B_2 to be a large number, and let $\{\epsilon^{**}_{bcij}, 1 \le i \le p, 1 \le j \le n_i\}$, for $1 \le b \le B_1$, and $1 \le c \le B_2$, denote a re-resample drawn randomly, with replacement, from the collection $\{\hat{\epsilon}^*_{bij}, 1 \le i \le p, 1 \le j \le n_i\}$. Define

$$y^{**}_{bcij} = \hat{\gamma}^*_{bi} + \hat{\beta}^*_b x_{ij} + \epsilon^{**}_{bcij}\hat{\sigma}^*_{bi}$$

and $\mathfrak{X}^{**}_{bc} = \{(x_{ij}, y^{**}_{bcij}, 1 \le i \le p, 1 \le j \le n_i\}$. Let $\hat{\gamma}^{**}_{bci}$ denote the estimator of γ_i obtained by applying method $\mathfrak{M}$ to the re-resample $\mathfrak{X}^{**}_{bc}$, and put $\hat{m}^{**}_{bc} = \max_i \hat{\gamma}^{**}_{bci}$ and

$$\hat{q}^*_{bB_2}(u) = B_2^{-1}\sum_{i=1}^{B_2} I(\hat{m}^{**}_{bc} \le u).$$

The latter function estimates the probability $P(\hat{m}^*_{bc} \le u \,|\, \mathfrak{X}^*_b)$, for arbitrary c. From the inverse function,

$$\hat{u}^*_{bB_2t} = \hat{q}^{*-1}_{bB_2}(t) = \inf\{u : \hat{q}^*_{bB_2}(u) \ge t\},$$

our Monte Carlo approximation to u^*_t (when the resample $\mathfrak{X}^*$ is $\mathfrak{X}^*_b$). Let

$$\mathcal{I}^*_{bB_2\alpha} = (\hat{u}^*_{bB_2,\alpha/2}, \hat{u}^*_{bB_2,1-\alpha/2})$$

and

$$\hat{\pi}_{B_1B_2}(\alpha) = B_1^{-1}\sum_{b=1}^{B_1} I(\hat{m} \in \mathcal{I}^*_{bB_2\alpha}),$$

the Monte Carlo approximation to $\hat{\pi}(\alpha)$. Compute $\hat{\pi}_{B_1B_2}(\alpha)$ for a number of values α in the vicinity of α_0, where $1 - \alpha_0$ denotes the desired coverage of the final confidence interval; extend $\hat{\pi}_{B_1B_2}$ to other α's by interpolation; solve the equation $\hat{\pi}_{B_1B_2}(\alpha) = 1 - \alpha_0$ using the interpolated approximation to the left-hand side, obtaining a solution $\hat{\alpha}_{B_1B_2}$ say; and take the final confidence interval to be $\mathcal{I}_{B_1,\hat{\alpha}_{B_1B_2}}$, our Monte Carlo approximation to $\mathcal{I}_{\alpha_0} = \mathcal{I}_{\hat{\alpha}}$.

The values of B_1 and B_2 should each be taken to be at least several hundred, see Hall (1991, section 1.4). Efficiency of the simulation can be enhanced by using balanced bootstrap resampling, rather than uniform bootstrap resampling, at both levels. The balanced bootstrap was proposed by Davison, Hinkley and Schechtman (1986). An algorithm for its implementation has been given by Gleason (1988). In this paper, though, we didn't apply the balanced bootstrap since the statistic of interest (a maximum of least squares parameters) is easy to compute (see remark (ii) of section 5 in Davison, Hinkley and Schechtman (1986)).

6. Applications to Railways

As pointed out in Section 2, linear models have been extensively used in the literature for describing the production frontier of a sector of activity. In particular, with panel data, a firm effect can be introduced allowing identification of a measure of technical efficiency for each firm.

We consider here a fixed effect model which produces consistent estimators of the parameters even if the efficiency levels are correlated with the exogenous variables. Note, however, that if one of the latter is time-invariant, the model is not identified, as Schmidt and Sickles (1984) remark.

In order to illustrate the above results, and point out the sensitivity of the coverage probability of the bootstrap confidence interval to the possible presence of ties for the most efficient firms, we consider the analysis of labor efficiencies of 19 railway companies observed over a period of 14 years. Data on the activity of the main international railway companies can be found in the annual reports of the *Union Internationale des Chemins de Fer* (U.I.C.). An analysis of labor efficiencies of railways using a more complete set of data may be found in Gathon and Perelman (1990), where the use of a labor function to analyze the production process is largely motivated.

The railway companies retained for the illustration are the following:

Network	Country	Network	Country
BR	Great Britain	NS	Netherlands
CFF	Switzerland	NSB	Norway
CFL	Luxemburg	OBB	Austria
CH	Greece	RENFE	Spain
CIE	Ireland	SJ	Sweden
CP	Portugal	SNCB	Belgium
DB	Germany	SNCF	France
DSB	Denmark	TCDD	Turkey
FS	Italy	VR	Finland
JNR	Japan		

The data are available for each network on an annual basis. In this study we used the period from 1970 to 1983. This provides 266 observations on the whole set of variables.

For the labor function, we use the following linear model:

$$y_{ij} = \gamma_i + \beta' x_{ij} + \beta_6 z_j + \sigma_j \epsilon_{ij}, \quad i = 1, \ldots, 19; \quad j = 1, \ldots, 14; \tag{6.1}$$

where

y_{ij} = labor (total number of employees)/total length; of the network (in kms);
$x_{1,ij}$ = total distance covered by trains (in 10^3 kms);
$x_{2,ij}$ = ratio of passenger trains in $x_{1,ij}$ (in %);
$x_{3,ij}$ = density of the network (kms of lines by 100 km^2);
$x_{4,ij}$ = mean distance covered by a passenger (kms);
$x_{5,ij}$ = mean distance covered by a ton of freight; transported (kms);
z_j = trend (year).

All these variables, except trend, are in logarithms.

Note that the total length of a network is equal to $x_{3,ij} * S_i$ where S_i is the size of the country (in 100 km^2). Its introduction in the input (division of the labor) allows us to eliminate a size effect. Thus the total length of a network is implicitly considered as exogenous. The introduction of S_i, as such, as an exogenous variable would not allow us to identify the model (6.1). The elasticity of labor w.r.t. the density of the network x_3, is thus given by $(1 + \beta_3)$. Note that x_1 represents a rough measure of the output of a railway whereas x_2, x_4, and x_5 characterize some aspects of the demand; x_3 is a physical measure of the density of a network. The trend variable, z, is included in order to capture the technical progress.

Since it is an input function, the technical efficiencies will be given by $\alpha_i = \gamma_i - \min_{1 \le i \le p} \gamma_i$. The parameters are estimated by weighted least squares where the weight matrix is simply given by

$$W = \begin{pmatrix} \sigma_1 I_{n_1} & & 0 \\ & \ddots & \\ 0 & & \sigma_p I_{n_p} \end{pmatrix}$$

and the variances σ_i^2 are estimated by the residuals e_i of the ith railways obtained by a first step global OLS on (6.1): $\hat{\sigma}_i^2 = (e_i' e_i)/(n_i - 1)$, $i = 1, \ldots, p$. The results of the estimation may be found in Table A.1 of Appendix (ii). Note that all the elasticities have the right sign (except for x_1 which is not significant).

Our interest here is the bootstrap distribution of $m = \min_i \gamma_i$ and the double bootstrap estimation of the coverage probabilities of the confidence intervals $\mathcal{I}_\alpha$. In this illustration, we choose $B_1 = B_2 = 200$. Table 1 shows these intervals for some values of $(1 - \alpha)$, and gives, for each α, the estimation $\hat{\pi}(\alpha)$ of the coverage probabilities obtained by the double bootstrap algorithms presented in Section 5.

Figure 1 shows that the coverage probabilities of $\mathcal{I}_\alpha$ are slightly overestimated; for instance if the predetermined level $(1 - \alpha_0)$ is 0.90 then $\hat{\alpha}$, the solution of equation (4.2),

Table 1. Bootstrap confidence intervals $\mathcal{I}_\alpha$ for m and estimated coverage probabilities $\hat{\pi}(\alpha)$ (complete set of data).

$1 - \alpha$	$\mathcal{I}_\alpha$		$\hat{\pi}(\alpha)$
0.9900	6.6512	12.8570	0.9800
0.9800	6.8325	12.5279	0.9800
0.9700	6.8889	10.3168	0.9650
0.9600	6.9523	10.1067	0.9400
0.9500	7.0461	10.0639	0.9100
0.9400	7.1449	11.9086	0.9150
0.9300	7.2542	11.9074	0.9050
0.9200	7.4162	11.7978	0.8900
0.9100	7.4296	11.7718	0.8800
0.9000	7.5136	11.7680	0.8800
0.8500	7.8714	11.2509	0.8300
0.8000	8.0002	11.1807	0.7600

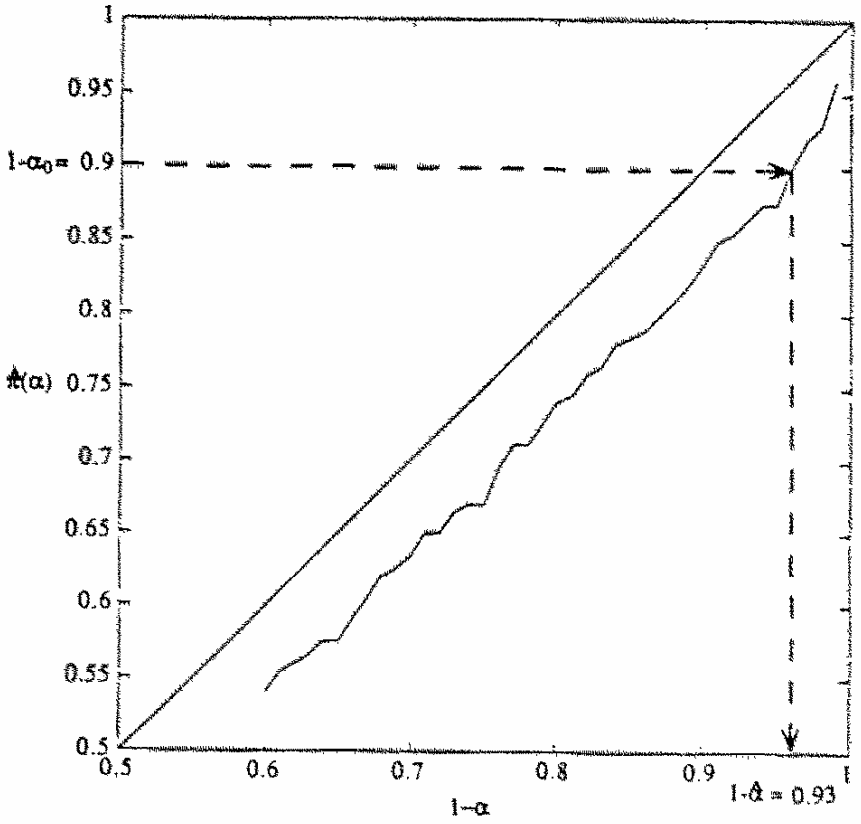


Figure 1. Estimated coverage probabilities $\hat{\pi}(\alpha)$ (complete set of data).

is given by 0.07. Our final estimate of the 0.90 confidence interval for m is $\mathcal{I}_{0.10} = \mathcal{I}_{0.07} =$ [7.2542, 11.9074] in place of $\mathcal{I}_{0.10} =$ [7.5136, 11.7680]. This is a minor correction, since it appears from the results in Appendix (ii) that there is a very small probability of a tie for $\min_i \gamma_i$. The network NSB (Norway) is clearly the most efficient (it appeared to be almost always the most efficient in the 200 replications of the bootstrap).

In order to analyze the effect of the possibility of ties for $\min_i \gamma_i$, we did the same computations without the railway NSB. The results of the estimation are given in Table A.2 of Appendix (ii). The results are very similar to the preceding case but note the higher possibility of ties for $\min_i \gamma_i$. In particular, the railways SJ (Sweden) with VR (Finland) compete now to be the most efficient.

Table 2 gives the percentile confidence intervals $\mathcal{I}_\alpha$ for different values of α and allows, with Figure 2, correction for coverage errors: here those errors are more important.

For instance, our final estimate of the 0.90 confidence intervals for m is $\mathcal{I}_{0.10} = \mathcal{I}_{0.04} =$ [6.3409, 12.0119] in place of $\mathcal{I}_{0.10} =$ [6.8427, 11.5698], a much wider interval. Here the corrections of the confidence intervals are more important due to the augmented risk of ties for $\min_i \gamma_i$ obtained by eliminating the clearly most efficient railway (NSB) from the data set.

So the illustration shows clearly the improvements due to the iterated bootstrap for estimating the coverage probabilities obtained by the percentile method. As expected, the corrections may become important when the risk of a tie for the most efficient railways has increased (compare Figure 1 and Figure 2).

Table 2. Bootstrap confidence intervals ϑ_α for m and estimated coverage probabilities $\hat{\pi}(\alpha)$ (without NSB Railways).

$1-\alpha$	ϑ_α		$\hat{\pi}(\alpha)$
0.9900	5.8023	13.8666	0.9600
0.9800	5.8528	12.4493	0.9300
0.9700	6.1365	12.2157	0.9200
0.9600	6.3409	12.0119	0.9000
0.9500	6.5145	11.8649	0.8750
0.9400	6.6158	11.8229	0.8750
0.9300	6.6256	11.7256	0.8650
0.9200	6.7393	11.7181	0.8550
0.9100	6.8414	11.5960	0.8500
0.9000	6.8427	11.5698	0.8350
0.8500	7.1469	11.2449	0.7850
0.8000	7.2736	10.8900	0.7400

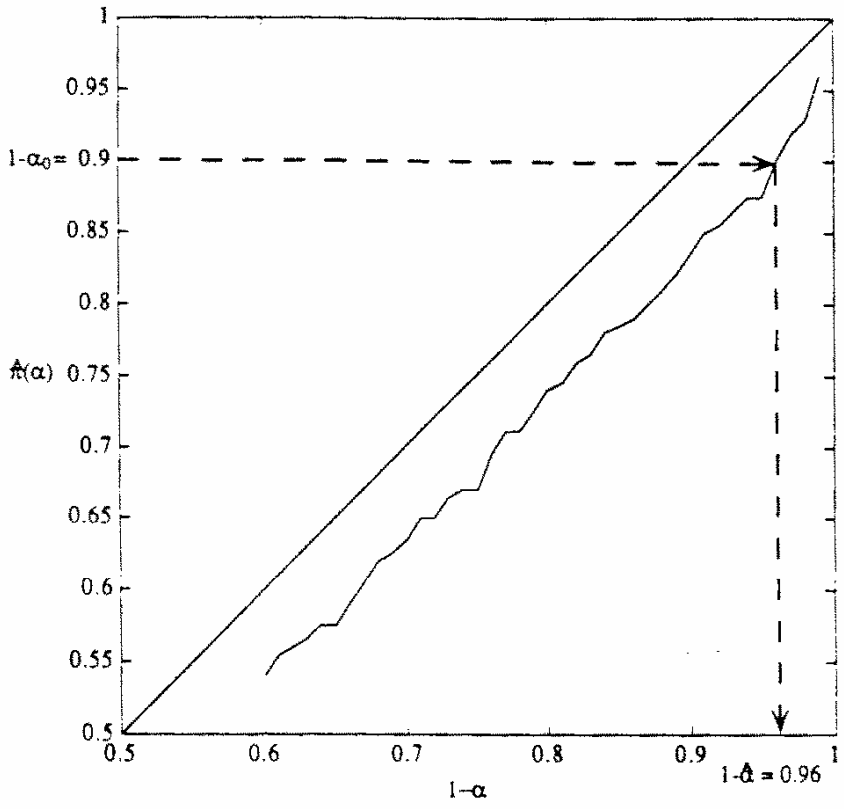


Figure 2. Estimated coverage probabilities $\hat{\pi}(\alpha)$ (without NSB Railways).

7. Conclusions

In this paper, we analyze a situation (estimation of a production frontier with panel data) where the bootstrap distribution of the parameter of interest (the level of the frontier $m = \max_i \gamma_i$, i.e. the maximum of least-squares parameters) may be inconsistent. It is shown

that consistency is achieved if and only if there are no ties for this maximum. In the regular situation however, poor distribution estimation can result because the percentile-t method may not be a viable option, due to the absence of a good estimator of the variance of m.

This is particularly true when the largest and second-largest values of γ_i are close. The iterated bootstrap proposed in this situation allows improvement of the estimation of the coverages probabilities of confidence intervals obtained by the percentile method.

The application to the analysis of labor efficiency of railways illustrates the improvements that can be expected by using the double bootstrap. In particular, in the case where the risk of a tie for the most efficient railway has been increased by eliminating the clearly most efficient firm from the data set, the corrections for coverage probabilities appeared to be important.

Appendix (i)

Here we show that the following three conditions are equivalent: (a) the asymptotic distribution of $\max \hat{\gamma}_i$ is normal, (b) the bootstrap estimate of the distribution of $\max \hat{\gamma}_i$ is consistent, (c) there are no ties for $\max \gamma_i$. We also note the effect which ties have in the large-sample properties of the iterated bootstrap.

Recall that the model is

$$y_{ij} = \gamma_i + \beta x_{ij} + \epsilon_{ij}\sigma_i,$$

where the ϵ_{ij}'s are independent and identically distributed with zero mean and unit variance. Estimators $\hat{\gamma}_i$, $\hat{\beta}$, $\hat{\sigma}_i$ of the unknown parameter may be derived by a variety of different methods, including weighted least-squares and robust regression. In these techniques, and a variety of others, the estimators $\hat{\gamma}_i$ are root-n consistent and asymptotically normally distributed. That is, under mild regularity conditions (e.g., finite variance of the ϵ_{ij}'s, finite variance of the distributions generating the design points x_{ij} and the assumption that $(\min n_i)/n$ is bounded away from zero), the p-vectors $\gamma = (\gamma_i)$ and $\hat{\gamma} = (\hat{\gamma}_i)$ satisfy

$$n^{1/2}(\hat{\gamma} - \gamma) \to N(0, V) \tag{A.1}$$

in distribution, for a $p \times p$ positive definite matrix V.

Put $m = \max \gamma_i$ and $\hat{m} = \max \hat{\gamma}_i$. We derive the asymptotic distribution of $\hat{m} - m$. Suppose there are precisely k distinct values of i such that $\gamma_i = m$. Let these be $i_1, \ldots, i_k$, let V_1 denote the $k \times k$ sub-matrix of V formed from the intersection of rows and columns with indices $i_1, \ldots, i_k$, and let the random k-vector $Z = (Z_j)$ be normal $N(0, V_1)$. Now, $P(\hat{m} = \max_j \hat{\gamma}_{i_j}) \to 1$, and so

$$\begin{aligned} P\{n^{1/2}(\hat{m} - m) \le u\} &= P\{n^{1/2}(\max_j \hat{\gamma}_{i_j} - m) \le u\} + o(1) \\ &= P\{n^{1/2}(\max_j \hat{\gamma}_{i_j} - \gamma_{i_j}) \le u\} + o(1) \\ &\to \psi(u) = P(\max Z_j \le u). \end{aligned} \tag{A.2}$$

The distribution of max Z_j is normal if and only if $k = 1$, i.e., if and only if there are no ties for max γ_i. Should there be a tie, the distribution of max Z_j depends on V_1 through quantities other than scale, and cannot be readily transformed to a distribution which does not depend on unknowns.

Next we consider the bootstrap version of the limit theorem at (A.2). Put

$$\psi(u \mid z_1, \ldots, z_k) = P\{\max(Z_j + z_j) - \max z_j \le u\},$$

for constants $z_1, \ldots, z_k$. Note particularly that $\psi(u \mid \cdot, \ldots, \cdot) \equiv \psi(u)$ if and only if $k = 1$. Let $\hat{\gamma}_i^*$ denote the estimator computed from the resample $\mathfrak{X}^*$, as outlined in Section 2, and define $\hat{m}^* = \max \hat{\gamma}_i^*$. Under the regularity conditions introduced two paragraphs earlier, the bootstrap distribution of $\hat{\gamma}^* = (\hat{\gamma}_i^*)$ enjoys the same asymptotic distribution as $\hat{\gamma}$, in the sense that for any Borel set $\mathcal{S} \subseteq \mathbb{R}^p$,

$$P\{n^{1/2}(\hat{\gamma}^* - \hat{\gamma}) \in \mathcal{S} \mid \mathfrak{X}\} \to P\{N(0, V) \in \mathcal{S}\}$$

in probability. Compare (A.1). Furthermore, $P(\hat{m}^* = \max_j \hat{\gamma}_{i_j} \mid \mathfrak{X}) = 1 + o_p(1)$. Therefore we may obtain the following analogue of (A.2) for the bootstrap estimator $\hat{\psi}(u)$ of $\psi(u)$:

$$\begin{aligned}
\hat{\psi}(u) &= P\{n^{1/2}(\hat{m}^* - m^*) \le u \mid \mathfrak{X}\} \\
&= P\{n^{1/2}(\max_j \hat{\gamma}_{i_j}^* - \max_j \hat{\gamma}_{i_j}^*) \le u \mid \mathfrak{X}\} + o_p(1) \\
&= P[n^{1/2} \max_j \{(\hat{\gamma}_{i_j}^* - \hat{\gamma}_{i_j}) + (\hat{\gamma}_{i_j}^* - m)\} - n^{1/2} \max_j (\hat{\gamma}_{i_j} - m) \le u \mid \mathfrak{X}\}] + o_p(1) \\
&= \psi\{u \mid n^{1/2}(\hat{\gamma}_{i_1} - m), \ldots, n^{1/2}(\hat{\gamma}_{i_k} - m)\} + o_p(1). \qquad \text{(A.3)}
\end{aligned}$$

Since the variables $n^{1/2}(\hat{\gamma}_{i_1} - m), \ldots, n^{1/2}(\hat{\gamma}_{i_k} - m)$ do not converge in probability then it follows from (A.3) that $\hat{\psi}$ converges in probability to ψ if and only if $k = 1$—that is, if and only if there are no ties for m. In the event that $k \ge 2$, $\hat{\psi}$ does not coverge in probability at all, although it does converge weakly:

$$\hat{\psi}(u) \to \psi(u \mid Z_1, \ldots, Z_k)$$

in distribution.

It is readily checked that an application of the iterated bootstrap produces a consistent estimator of the distribution function $\psi_1(u) = E\{\psi(u \mid Z_1, \ldots, Z_k)\}$, not of $\psi(u)$. However, $\psi_1(u)$ is a nonrandom, smoothed version of the weak limit of the conditional distribution of $n^{1/2}(\hat{m}^* - \hat{m})$, and as such it represents a better approximation to the nonrandom distribution function $\psi(u)$ than does the random, heavily sample-dependent distribution function $\psi\{u \mid n^{1/2}(\hat{\gamma}_{i_1} - m), \ldots, n^{1/2}(\hat{\gamma}_{i_q} - m)\}$ which $\hat{\psi}(u)$ actually approximates.

Appendix (ii)

Table A.1. Estimation of model (6.1) by W.L.S. (complete set of observations).

	Estimation	Standard Deviation
γ for BR	11.8822	1.2983
γ for CFF	11.9229	1.3121
γ for CFL	12.0668	1.3800
γ for CH	10.2421	1.2948
γ for CIE	10.0865	1.3220
γ for CP	10.9430	1.3104
γ for DB	12.2039	1.3047
γ for DSB	11.3744	1.3066
γ for FS	11.9430	1.2744
γ for JNR	12.2009	1.2870
γ for NS	11.8023	1.3040
γ for NSB	9.5726	1.3075
γ for OBB	11.8311	1.3123
γ for RENFE	10.4857	1.2818
γ for SJ	9.8583	1.2902
γ for SNCB	12.3886	1.3239
γ for SNCF	11.3373	1.2892
γ for TCDD	10.0856	1.3056
γ for VR	9.8497	1.2984
Total distance	−0.0056	0.0534
Ratio passengers	−0.3388	0.0811
Density	−0.6644	0.0929
Mean distance passengers	−0.0812	0.0482
Mean distance freight	−0.0869	0.0392
Trend	−0.0029	0.0009
$R^2 = 0.9884$		

Table A.2. Estimation of model (6.1) by W.L.S. (without NSB railways).

	Estimation	Standard Deviation
γ for BR	10.9933	1.3709
γ for CFF	10.9751	1.3875
γ for CFL	11.0974	1.4533
γ for CH	9.2584	1.3701
γ for CIE	9.1001	1.3967
γ for CP	9.9905	1.3836
γ for DB	11.2744	1.3818
γ for DSB	10.4171	1.3811
γ for FS	10.9874	1.3527
γ for JNR	11.2914	1.3610
γ for NS	10.8658	1.3790
γ for OBB	10.8611	1.3898
γ for RENFE	9.5293	1.3578
γ for SJ	8.8888	1.3668

Hall, P., Hardle, W. and Simar, L. (1995) Iterated Bootstrap with Applications to Frontier Models.

Table A.2. Continued.

	Estimation	Standard Deviation
γ for SNCB	11.4361	1.4009
γ for SNCF	10.3731	1.3691
γ for TCDD	9.1016	1.3804
γ for VR	8.8683	1.3743
Total distance	−0.0323	0.0549
Ratio passengers	−0.4185	0.0880
Density	−0.6297	0.0976
Mean distance passengers	−0.0577	0.0510
Mean distance freight	−0.0400	0.0494
Trend	−0.0023	0.0010
$R^2 = 0.9872$		

References

Beran, R. (1987). "Prepivoting to reduce level error of confidence sets." *Biometrika* 74, 457–468.

Davison, A.C., Hinkley, D.V., and Schechtman. (1986). "Efficient bootstrap simulation." *Biometrika* 73, 555–566.

Farrell, M.J. (1957). "The measurement of production efficiency." *Journal of the Royal Statistical Society* A 120, 253–281.

Gathon, H.J. and S. Perelman. (1990). "Measuring technical efficiency in national railways: A panel data approach." Mimeo, Université de Liège, Belgium.

Gleason, J.R. (1988). "Algorithms for balanced bootstrap simulations." *American Statisticians* 42, 263–266.

Hall, P. (1986). "On the bootstrap and confidence intervals." *Annals of Statistics* 14, 1431–1452.

Hall, P. (1992). *The Bootstrap and Edgeworth Expansion*. New York/Berlin: Springer-Verlag.

Hall, P., Härdle, W., and L. Simar. (1993). "On the inconsistency of bootstrap distribution estimators." *Computational Statistics and Data Analysis* 16, 11–18.

Hall, P. and M.A. Martin. (1988). "On bootstrap resampling and iteration." *Biometrika* 75, 661–671.

Loh, W.-Y. (1987). "Calibrating confidence coefficients." *Journal of the American Statistical Association* 82, 155–162.

Schmidt, P. and R.E. Sickles. (1984). "Production frontiers and panel data." *Journal of Business and Economic Statistics* 2, 367–374.

Simar, L. (1992). "Estimating efficiencies from frontier models with panel data: A comparison of parametric, nonparametric and semiparametric methods with bootstrapping." *The Journal of Productivity Analysis* 3, 171–203.

U.I.C. (1970–1983). *Statistiques Internationales des Chemins de Fer*, Union Internationale des Chemins de Fer, Paris.

Journal of Nonparametric Statistics, 5, 157-184.

Additive Nonparametric Regression on Principal Components

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

A.B. Tsybakov
Laboratoire de Statistique Theorique et Appliquée
Université Paris VI
4, pl. Jussieu, B.P. 158, 75252 Paris, France.

January 1994

Abstract

Nonparametric regression smoothing in high dimensions faces the problem of data sparseness. Additive regression models alleviate this problem by fitting a sum of one-dimensional smooth functions. A common approach for dimension reduction in multivariate statistics is to replace the original high dimensional predictor variables by their dominant principal components. In this paper consider an additive nonparametric regession model on principal components. A three-stage procedure is proposed to decide how many and which components should be included into such an additive model. In a first step the predictor variables are made orthogonal by the principal component transformation. After the second step, determining the number and sequence of components, the additive regession model is fit by the kernel method. The asymptotic distribution of this regression estimate is given. The practical performance is investigated via a simulation study.

Keywords: nonparametric regression in high dimensions, additive regression models, principal components kernel estimation, dimensionality reduction, model choice.

AMS Subject Classification: 62G05, 62G20.

1. Introduction

It is well known that nonparametric estimation methods for regression analysis in high dimensions face the problem of dimensionality. Theoretically this can be seen through the analysis of rates of convergence which depend exponentially on the dimension. Practically speaking this means that the "window" over which we average the response variables contains almost no observations.

One way of avoiding this problem of dimensionality is to impose more structure on the regression function. Additivity is such an option of which Stone (1986) has shown that it results in better (essentially one-dimensional) rates of convergence. An additive regession model is one where the multidimensional regression function is represented as a sum of smooth one-dimensional functions operating on each predictor variable separately.

In practice though it might well be that some of these regression components are actually zero which means that these particular variables have no influcence on the response. It is therefore interesting to determine *how many* and *which* components should be included into such an additive regression model.

Let (X, Y) be a random variable with $X \in I\!R^d$ and $Y \in I\!R$ and let $\{(X_i, Y_i)\}_{i=1}^n$ be an i.i.d. sample of this random variable. Consider the estimation of the regression function $m(x) = E(Y \mid X = x)$. Stone (1980, 1982), Ibragimov and Hasminskii (1980) have shown that the optimal rate for estimating m is $n^{-\frac{\ell}{2\ell+d}}$ with ℓ an index of smoothness of m. An additive structure for m is a regression function of the form $m(x) = \sum_{j=1}^d m_j(x_j)$ where $x = (x_1, \ldots, x_d)$ are the d-dimensionl predictor variable and m_j are one-dimensional nonparametric functions operating on each element of the vector or predictor variables. Stone (1986) has shown that for such regression curves one obtains a one-dimensional rate of convergence with $n^{-\frac{\ell}{2\ell+1}}$. Thus one speaks of dimensionality reduction through additive modelling.

A classical dimension reduction technique is the computation of principal components with dominant eigenvalues, see Mardia, Kent and Bibby (1979) or Flury and Riedwyl (1988). The number of variables d is typically reduced to $p^* = 2$ or 3 principal components that resolve the main proportion of variation in the predictor variable. In this paper we combine these two techniqus by considering additive nonparametric regression on principal components.

Suppose that X can be represented as

$$X = AU + B$$

where $A = \Sigma^{1/2}$ is the unknown positive definite symmetric root of a covariance matrix of full rank d, $B = \mu$ is the unknown mean vector, and $U = (U_1, \ldots, U_d)$ is a random vector with independent components $U_1, \ldots, U_d$ such that $E(U) = 0$ and $cov(U)$ is the $d \times d$ identity matrix.

Consider now regression functions m that follow an additive model in U-coordinates, i.e.,

$$m(x) = \sum_{j=1}^{d} g_j \left[\{A^{-1}(x - B)\}_j\right] \tag{1.1}$$

where $\{A^{-1}(x - B)\}_j$ is the jth coordinate of the vector $A^{-1}(x - B)$. The problem in practice is to choose the components g_j, that give significant contributions to m. Suppose that only components with index $j \in J = \{j, \ldots, j_{p^*}\}$ are different from zero. Then (1.1) can be written as:

$$m(x) = \sum_{k=1}^{p^*} g_{j_k}\left(\{A^{-1}(x - B)\}_{j_k}\right) \tag{1.2}$$

The problem is now to estimate the *significant directions* $J = \{j_1, \ldots, j_{p^*}\}$, the *significant functions* $\{g_{j_k}\}_{k=1}^{p^*}$, and the principal component transformation A, B.

In principal component analysis the U-variables are uncorrelated, and if they are Gaussian, also independent. It is supposed here that the U components are independent for technical reasons. The additional technical difficulty for estimating (1.2) is that not only the number p^* of significant directions is unknown but also the individual indices $j_k \in J$. It is assumed throughout the paper that the additive nonparametric regression on principal components model (1.2) is true.

The model (1.2) has also an interpretation in terms of projection pursuit additive modelling (Friedman and Tukey (1974), Friedman and Stuetzle (1980), Hastie and Tibshirani (1990)).

A projection pursuit regression model is defined as $m(x) = \sum_{k=1}^{p^*} g_{j_k}\left(\beta_{j_k}^T x\right)$ where $\beta_{j_k} \in R^d$ are unknown projection vectors. If we denote the rows of A^{-1} by $\beta_1, \ldots, \beta_d$; then $(A^{-1}x)_j = \beta_j^T x$, and thus the projection pursuit model reduces to (1.2) with $B = 0$.The significant directions are the projections along which the smooth functions g_{j_k} are applied. The fitting technique described in the above references is different from ours though.

We deal with the problem of estimation of $m(x)$ in three separate steps:

(i) Estimation of the matrix A and vector B using only the X-sample;

(ii) Estimation of significant directions;

(iii) Estimation of significant functions.

The projection pursuit techique is to fit the projections β_j and the functions simultaneously. The steps (i), (ii) are parametric problems. Indeed, we show that A, B are estimated $\sqrt{n}$-consistently and the estimators of significant directions coincide with true significant directions with probability that goes to 1 as $n \to \infty$. Using these results we apply the sum of one-dimensional kernel smoothers along the significant directions to obtain the final estimator of m. The convegence rate to m is the optimal one (Stone, 1986) and the asymptotic distribution of the additive nonparametric regression estimator is normal.

2. The Estimators

(i) Estimators of the tranformation matrix A and mean vector B are defined in a usual way. Set

$$\begin{aligned} \widehat{B}_n &= \frac{1}{n}\sum_{i=1}^{n} X_i, \\ \widehat{A}_n &= \left\{\frac{1}{n}\sum_{i=1}^{n}(X_i - \widehat{B}_n)(X_i - \widehat{B}_n)^T\right\}^{1/2}. \end{aligned}$$

(ii) The significant directions are picked as follows. Define for each index j the kernel estimators

$$\widehat{g}_j(t) = \frac{\widehat{r}_j(t)}{\widehat{f}_j(t)} = \frac{\frac{1}{n}\sum_{i=1}^{n} K_h\{t - \{\widehat{A}_n^{-1}(X_i - \widehat{B}_n)\}_j\}Y_i}{\frac{1}{n}\sum_{i=1}^{n} K_h\{t - \{\widehat{A}_n^{-1}(X_i - \widehat{B}_n)\}_j\}}. \tag{2.1}$$

Here $K_h(\bullet) = h^{-1}K(\bullet/h)$ and $K(\bullet)$ is a kernel. To estimate the set J we calculate the quantities

$$\widehat{S}_j = n^{-1}\sum_{i=1}^{n} \widehat{r}_j^2\left(\{\widehat{A}_n^{-1}(X_i - \widehat{B}_n)\}_j\right), \qquad j = 1, \ldots, d,$$

and select those directions j that statisfy $\widehat{S}_j \geq b_n$, where $b_n > 0$ is some prescribed level. Denote

$$\widehat{J} = \{j : \widehat{S}_j \geq b_n\}.$$

(iii) The regression function estimator we finally consider is

$$\widehat{m}(x) = \sum_{k \in \widehat{J}} \widehat{g}_{j_k}\left(\{\widehat{A}_n^{-1}(x - \widehat{B}_n)\}_{j_k}\right) \tag{2.2}$$

where $\widehat{J}$ is the random subset defined above. Note that the bandwidth h for the steps (ii) and (iii) need not be the same. To mark this difference we denote $h = h_n$ the bandwidth for (ii), and $h = h_{0n}$ the bandwidth for (iii).

Remark. Other approaches to selection of significant directions are possible.

(1)The choice of $\widehat{S}_j$ can be different. Note that, under the appropriate assumptions, stated below in Section 3, $\widehat{S}_j$ is a consistent estimator of the functional $\widehat{S}_n = \int g_j^2(u) f_j^3(u) du$, where $f_j(u)$ is the marginal distribution of $\{A^{-1}(x_i - B)\}_j$. The choice of the functional S_j is done here for mathematical and computational convenience. Other functionals satisfying $S_j > 0, j \in J, S_j = 0, j \notin J$, can be implemented as well. Our motivation for this particular one came from considering the partial residual sum of squares

$$n^{-1} \sum_{i=1}^{n} \{Y_i - \widehat{g}_j(U_i)\}^2.$$

This sum, under proper conditions, behaves approximately as $n^{-1} \sum_{i=1}^n Y_i^2 - n^{-1} \sum_{i=1}^n \{\widehat{g}_j(U_i)\}^2$. For a significant direction the second term will give a significant negative contribution. Discarding the random denominator of $\widehat{g}_j$ for technical reasons and the constant $n^{-1} \sum_{i=1}^n Y_i^2$ we come to $\widehat{S}_j$ as a reasonable statistic to detect significant directions.

(2) One can propose another procedure, similar to those model selectors in regression analysis that incorporate penalty term which is linear in the "complexity" p of the model, see e.g., Akaike (1977, 1974), Mallows (1973). A basic difference though is that we need to order the directions, whereas in the classical nested model situation this is not necessary. This procedure works as follows.

Arrange $\widehat{S}_j$ in the decreasing order:

$$\widehat{S}^{(1)} \geq \widehat{S}^{(2)} \geq \cdots \geq \widehat{S}^{(d)}.$$

Let $(1)_n$ be the integer that equals to j with maximal value $\widehat{S}_j = \widehat{S}^{(1)}$. Let then $(2)_n$ be the integer that equals to j with $\widehat{S}_j = \widehat{S}^{(2)}$ etc. Thus

$$(k)_n = j \in \{1, \ldots, d\} : \widehat{S}_j = \widehat{S}^{(k)}.$$

Without loss of generality assume that all $\widehat{S}^{(k)}$ are different (thus $(k)_n$ is uniquely defined). In particular we have

$$\widehat{S}^{(k)} = \frac{1}{n} \sum_{i=1}^{n} \widehat{r}^2_{(k)_n} \left(\{\widehat{A}_n^{-1}(X_i - \widehat{B}_n)\}_{(k)_n} \right).$$

Choose $\widehat{p}$ as the minimizer of the following statistic

$$\widehat{p} = \left[\operatorname*{argmin}_{p \leq d} (\widehat{S}^{(p)} + p b_n) \right] - 1,$$

where b_n is a sequence such that $n b_n^2 \to \infty$. The estimate of the set J is defined as $\widehat{J} = \{(1)_n, \ldots, (\widehat{p})_n\}$.

It is clear that this procedure works if $p^* < d$, and it is equivalent to the comparison of differences of consecutive $\widehat{S}(j)$ values with a level b_n.

In Section 3 we give details of the procedure (i) – (iii). In particular we prove that $\widehat{J}$ coincides with J with probability tending to one. In Section 4 we prove the asymptotic normality of the additive estimator (2.2). An example is given in Section 5. The last section is devoted to proofs.

3. Estimators of A, B, and of Significant Directions: Asymptotic Properties

Assume that the kernel and the bandwidth satisfy the following conditions.

(A1) K is bounded, nonnegative, compactly supported, Lipschitz continuous, and $\int K(u) du = 1$.

(A2) The sequence of bandwidths h_n is such that $\lim_n nh_n^2 = \infty$.

We also need some assumptions on the functions $g_j(\bullet)$, the marginal densities $f_j(\bullet)$ of the random variables $U_{ij} = \{A^{-1}(X_i - B)\}_j$ and on the distribution of errors $\varepsilon_i(X_i) = Y_i - m(X_i)$.

(A3) The functions $g_j(\bullet)$, $j \in J$, are bounded and Lipschitz continuous.

(A4) The random variables $U_{ij} = \{A^{-1}(X_i - B)\}_j$ are mutually independent for different j.

(A5) $E[g_j(U_{1j})] = 0$.

(A6) The functions $f_j(\bullet)$ are bounded, Lipschitz continuous and bounded away from zero. The support of $f_j(\bullet)$ is compact.

(A7) $\sup_x E(\varepsilon_1^4(X_1) \mid X_1 = x) \le C_1 < \infty$.

(A8) $S_j = \int g_j^2(u)\, f_j^3(u)\, du \neq 0$, $j \in J$.

To simplify the notation, suppose that $\{j_1, \ldots, j_{p^*}\} = \{1, \ldots, p^*\}$, i.e., the significant directions are the first p^* directions. This can be done without loss of generality since the statistician does not know this information.

The next lemma gives $\sqrt{n}$-consistency of $\widehat{A}_n$ and $\widehat{B}_n$.

Lemma 1. *Assume (A4), (A6) and*

(A9) $E(U_{1j}) = 0$, $E(U_{1j}^2) = 1$, $j = 1, \ldots, d$.

Then

$$E(|\widehat{B}_n - B|^2) = O\left(\frac{1}{n}\right), \quad n \to \infty \tag{3.1}$$

$$E(\|\widehat{A}_n - A\|^2) = O\left(\frac{1}{n}\right), \quad n \to \infty \tag{3.2}$$

and there exists $C^* > 0$ such that

$$P\left(\|\widehat{A}_n^{-1} - A^{-1}\| \ge \frac{\tau}{\sqrt{n}}\right) \le \frac{1}{C^*}\left(\exp(-C^*\tau^2) + \exp(-C^* n)\right), \tau > 0, \tag{3.3}$$

where $|\bullet|$ is the Euclidean norm of a vector in $\mathbb{R}^d$, and $\|\bullet\|$ is the maximum-of-elements norm for $(d \times d)$-matrices.

Proof of Lemma 1 is standard, and we omit it. The exponential estimate (3.3) comes from the Bernstein's inequality since the $|X_i|$'s are uniformly bounded (in fact, the support of $f_j(\bullet)$ is compact by (A6)). Clearly, (3.3) may be extended to the case of noncompactly supported $f_j(\bullet)$ under the Cramer condition.

The estimation of significant directions is based on the fact that $\widehat{S}_j$ is close to S_j for n large enough. However, we do not need that $\widehat{S}^{(j)}$ were a consistent estimate of S_j. In particular, we prefer to choose a constant $h_n = h$ (small enough) rather that $h_n \to 0$.

The next lemma gives the conditions under which $\widehat{S}_j$ is close to S_j. Denote

$$r_j^*(t) = \frac{1}{n}\sum_{i=1}^n K_h(t - U_{ij})Y_i$$

$$S_j^* = \frac{1}{n}\sum_{m=1}^n r_j^{*2}(U_{mj}).$$

Let $D, L, g_{\max}$ be some positive numbers. Define the class of pairs *(f,g)* as follows:

$$\begin{aligned}\mathcal{T} \quad = \quad & \{(f,g) \mid f=[-D,D] \to R^1, g:[-D,D] \to R^1, \\ & f,g \text{ are Lipschitz continuous with constant } L, \\ & \max_u \mid g(u) \mid \leq g_{\max}, f \text{ is a probability density}\}\end{aligned}$$

Clearly,

$$\mathcal{T} = \mathcal{T}(D, L, g_{max}).$$

Consider the following class of vector-functions $\mathbf{f} = (f_1, \ldots, f_d), \mathbf{g} = (g_1, \ldots, g_d)$:

$$\mathcal{P}_{p^*} = \{(\mathbf{f},\mathbf{g}) \mid (f_j, g_j) \in \mathcal{T},\ j = 1, \ldots, d, \int g_j^2(u) f_j^3(u) du \geq s,\ j = 1, \ldots, p^*; g_j(u) \equiv 0,\ j = p^*+1, \ldots, d\},$$

where $s > 0$ is a constant. We will refer to this class as simply $\mathcal{P}$, assuming p^* and s fixed.

Lemma 2. *Assume (A1)–(A9). Then*

(a) there exists a constant $C_2 > 0$ such that

$$\sup_{(\mathbf{f},\mathbf{g})\in\mathcal{P}} E[E(S_j^*) - S_j^*]^2 \leq \frac{C_2}{n}, j = 1, \ldots, d, \tag{3.4}$$

$$\sup_{(\mathbf{f},\mathbf{g})\in\mathcal{P}} P\left(\sqrt{nh_n^2}|\widehat{S}_j - S_j^*| \geq a\right) \leq C_2 \frac{1+\log^2 a}{a^2} + \frac{1}{C^*}\exp(-C^* n), \tag{3.5}$$
$$j = 1, \ldots, d, \quad a > 0,$$

(b) let $D_1 = L\int |t|K(t)dt$, then there exists $C_3 > 0$ such that

$$\max_{1\leq j\leq p^*} \sup_{(\mathbf{f},\mathbf{g})\in\mathcal{P}} |E(S_j^*) - S_j| \leq D_1 h_n + \frac{C_3}{nh_n}, \tag{3.6}$$

and

$$\sup_{(\mathbf{f},\mathbf{g})\in\mathcal{P}} |E(S_j^*)| \leq \frac{C_3}{nh_n}, \quad j = p^*+1, \ldots, d. \tag{3.7}$$

Proof of Lemma 2 is deferred to Section 6.

Note that for any $a > 0$, and any sequence $\{h_n\}$ such that $\limsup_n h_n < \infty$ (3.4), (3.5) entail

$$\sup_{(\mathbf{f},\mathbf{g})\in\mathcal{P}} P(|\widehat{S}_j - E(S_j^*)| \geq a) \leq C_4 \frac{\log^2 n}{a^2 n h_n^2}, \quad n \to \infty, \tag{3.8}$$

for some $C_4 > 0$.

Now we are able to state the main result of this section.

Theorem 1. *Assume* (A1) *to* (A9). *Let b_n be such that $C_3/nh_n < b_n < s - D_1 h_n - C_3/nh_n$, and let $\limsup_{n\to\infty} h_n < \infty$. Then*

$$\sup_{(\mathbf{f},\mathbf{g})\in\mathcal{P}} P(\widehat{J} \neq J) \leq C_4 d \frac{\log^2 n}{nh_n^2}\left[\frac{1}{(b_n - C_3/nh_n)^2} + \frac{1}{(s - b_n - D_1 h_n - \frac{C_3}{nh_n})^2}\right]. \tag{3.9}$$

The minimum of the right-hand side of (3.9) *is attained for $h_n \equiv s/(2D_1), b_n = \frac{s}{4} + 2C_3 D_1/(s_n)$, and under this choice of h_n and b_n we have*

$$\sup_{(\mathbf{f},\mathbf{g})\in\mathcal{P}} P(\widehat{J} \neq J) \leq \frac{128 C_4 d D_1^2}{s^4} \cdot \frac{\log^2 n}{n}. \tag{3.10}$$

Proof. Note that

$$P(\widehat{J} \neq J) \leq P(\exists\, j \in J : \widehat{S}_j < b_n) + P(\exists\, j \notin J : \widehat{S}_j \geq b_n). \tag{3.11}$$

By (3.6), (3.8) the first probability in the right-hand side of (3.11) is bounded as follows

$$\begin{aligned} P(\exists\, j \in J : \widehat{S}_j < b_n) &\leq \operatorname{card}(J) \max_{j \in J} P\left(\widehat{S}_j - E(S_j^*) < b_n - E(S_j^*)\right) \\ &\leq d \max_{j \in J} P\left(|\widehat{S}_j - E(S_j^*)| > s - D_1 hn - \frac{C_3}{nh_n} - b_n\right) \\ &\leq \frac{dC_4 \log^2 n}{nh_n^2 (s - D_1 h_n - b_n - C_3/nh_n)^2}. \end{aligned} \tag{3.12}$$

For the second probability in the right-hand side of (3.11) we obtain by (3.7), (3.8)

$$\begin{aligned} &P(\exists\, j \notin J : \widehat{S}_j \geq b_n) \leq \\ &\leq d \max_{j \notin J} P\left(|\widehat{S}_j - E(S_j^*)| \geq b_n - \frac{C_3}{nh}\right) \leq \frac{dC_4 \log^2 n}{nh_n^2 (b_n - C_3/nh_n)^2}. \end{aligned} \tag{3.13}$$

Inequalities (3.12) and (3.13) entail (3.9).

To prove (3.10), denote $b_{n1} = b_n - C_3/nh_n$, and note that for any fixed h_n the RHS of (3.9) is minimised by $b_{n1} = (s - D_1 h_n)/2$. Substituting this into the RHS of (3.9) we obtain the expression

$$C_4 d \frac{\log^2 n}{nh_n^2} \quad \frac{8}{(s - D_1 h_n)^2}$$

which is minimised by $h_n = s/(2D_1)$. ∎

4. Asymptotic Distribution of the Additive Regression Estimate

In this section, we establish the asymptotic distribution of $\widehat{m}(x)$ at a fixed point x. We assume that $x = Au + B$ where $u = (u_1, \ldots, u_d)$ belongs to the interior of the support of $f(x) = \prod_{j=1}^d f_j(x_j)$.

In addition to (A1)–(A8) the following assumptions are needed.

(A10) The functions $g_j(\bullet)$ and $f_j(\bullet)$ are $\ell - 1$ times continuously differentiable, and there exist one-sided ℓth derivatives $g_j^{(\ell)}(x \pm 0)$, $f_j^{(\ell)}(x \pm 0)$, $\ell \geq 2$ is an integer, $j = 1, \ldots, d$.

(A11) $h_{0n} = \beta n^{-1/(2\ell+1)}, \quad \beta > 0.$

(A12) $\int u^j K(u) du = 0, \quad j = 1, \ldots, \ell - 1$, and K has the Lipschitzian first derivative.

(A13) The variance $V(z) = E((Y_1 - m(X_1))^2 \mid X_1 = Az + B)$, $z \in \mathbb{R}^d$, is a continuous function on the support of f.

Assumption (A11) guarantees in combination with (A12) that the bandwidth $h = h_n$ has the speed necessary for an optimal rate of convergence, see, e.g., Müller (1988). The following result allows construction of pointwise asymptotic confidence intervals.

Theorem 2. *Under the assumptions (A1)–(A13) we have*

$$n^{\ell/(2\ell+1)}(\widehat{m}(x) - m(x)) \xrightarrow{\mathcal{L}} \mathcal{N}(b(x), \sigma^2(x))$$

where

$$
\begin{aligned}
b(x) &= \frac{\beta^\ell}{\ell!} \sum_{k=1}^{p^*} f_k^{-1}(u_k) \int K(w) w^\ell H^{(\ell)}(u_k; w) dw \\
\sigma^2(x) &= \frac{1}{\beta} \int K^2(w) dw \sum_{k=1}^{p^*} \left\{ f_k^{-1}(u_k) \left[\sum_{\substack{j=1 \\ j \neq k}}^{p^*} E\left(g_j^2(U_{1j})\right) + \right. \right. \\
&\quad \left. \left. + \int V(z_1, \ldots, z_{k-1}, u_k, z_{k+1}, \ldots, z_d) \prod_{\substack{j=1 \\ j \neq k}}^{d} f_j(z_j) dz_j \right] \right\}, \\
H^{(\ell)}(u_k; w) &= \begin{cases} H_+^{(\ell)}(u_k), & w \geq 0, \\ H_-^{(\ell)}(u_k), & w < 0, \end{cases}
\end{aligned}
$$

and $H_+^{(\ell)}$, $H_-^{(\ell)}$ *are one-sided derivatives of* $H^{(\ell-1)}$, *with* $H(v) = (g_k(v) - g_k(u_k)) f_k(v)$, $u = A^{-1}(x - B)$, $u = (u_1, \ldots, u_d)$.

Remark. It follows from Theorem 2 that the rate of convergence of $\widehat{m}(x)$ is free from the "curse of dimensionality". This is the rate $n^{\ell/(2\ell+1)}$ that is achieved in estimation of scalar regression functions (cf. Stone (1986), Hall (1989)). Although, the constant factor in the rate of convergence deteriorates. In particular, the additional variance term $\sum_{\substack{j=1 \\ j \neq k}}^{p^*} E(g_j^2(U_{1j}))$ appears.

Example. Assume that the variance $V(z) \equiv V$ where V is a constant. Assume also that $\ell = 2$ and that the second derivatives g_j'', f_j'' are continuous. Then the bias of $\widehat{m}(x)$ is

$$
b(x) = \beta^2 \sum_{k=1}^{p^*} \left(\frac{f_k'(u_k) g_k'(u_k)}{f_k(u_k)} + \frac{g_k''(u_k)}{2} \right) \int w^2 K(w) dw,
$$

and the variance is

$$
\sigma^2(x) = \frac{1}{\beta} \int K^2(w) dw \sum_{k=1}^{p^*} \left(f_k^{-1}(u_k) \left[V + \sum_{\substack{j=1 \\ j \neq k}}^{p^*} E(g_j^2(U_{1j})) \right] \right).
$$

The construction of asymptotic confidence intervals with coverage probability $1 - \alpha$ unfortunately involves estimating the terms $b(x)$ and $\sigma^2(x)$. The bootstrap provides one alternative for approximating desired p-values.

To find such a sample based approximation to the distribution of $n^{\ell/(2\ell+1)}(\widehat{m}(x) - m(x))$ one can use, for instance, the wild bootstrap (Härdle and Marron (1991)).

5. An Example

We tested the proposed procedure in a simulated situation. We have chosen a $d = 10$ dimensional normal X variable with covariance $A = I$ and mean $B = 0$. Define $\rho(u) = \sin(u)I(|u| \leq \pi)$. The significant functions were defined as

$$\begin{aligned} g_1(u_1) &= 2\rho(u_1) \\ g_3(u_3) &= 1.75\rho(u_3) \\ g_5(u_5) &= 1.5\rho(u_5) \\ g_7(u_7) &= 1.25\rho(u_7) \\ g_9(u_9) &= \rho(u_9) \end{aligned}$$

Thus the set of significant directions is $J = \{1, 3, 5, 7, 9\}$. We have chosen a set of significant functions g_{j_k} with decreasing values of S_j in order to make the problem harder for increasing j. The values of s_j are

$$\begin{aligned} s_1 &= 0.09, s_3 = 0.07, \\ s_5 &= 0.05, s_7 = 0.04, s_9 = 0.02, \\ s_2 &= s_4 = s_6 = s_8 = s_{10} = 0. \end{aligned}$$

The response variables Y_i were then generated by adding to the sum of the significant functions a normal error variable with standard deviation $\sigma = 0.1$. This experiment was repeated $N = 100$ times for sample size $n = 100$.

Figure 1 gives an impression of how difficult it is to detect significant directions.

Figure 1

Figure 1. A simulated data set for n=100. The points are a scatterplot of (X_{i1}, Y_i), the line denotes the function $\hat{r}_1(u)$ for h=0.6.

In this figure we show a scatterplot of the first variable $\widehat{U}_1$ (a significant one) versus the response variable and the kernel estimator $\widehat{r}_1$. One sees that the equation "data = curve + noise" that we carry in our heads is not fulfilled here since the curve $\widehat{r}$,is almost buried in noise. Although σ seems to be small relative to g_1, note that in the above scatterplot the Y variable contains the sum of the other 9 variables. The nonsignificant directions though have only zero mean contribution. One such zero mean contribution is shown as a dotted line in Figure 2.

Figure 2

Figure 2. The functions $\hat{r}_1, \hat{r}_2$ for n=100. The solid line denotes $\hat{r}_1$, the dotted line $\hat{r}_2$.

The solid line shows the curve $\widehat{r}_1(\widehat{U}_{i1})$ as in Figure 1,the dotted curve $\widehat{r}_2(\widehat{U}_{i2})$ with $\widehat{U}_{ij} = \{\widehat{A}^{-1}(X_i - \widehat{B})\}_j$. Note that the dotted curve corresponds to a nonsignificant direction with index $j = 2$.

How well is the true order of the model estimated? This is shown in the following Table 1.

These two plots seem to display a rather pessimistic situation. What counts though in selecting the significant directions is the score $S_j = \int g_j^2 f_j^3$. This quantity is rather well estimated as Figure 3 shows.

Figure 3

Figure 3. The mean score function with standard error bands, $h = 1.2$.

It shows the averaged scores over $N = 100$ simulations together with the observed standard errors. The significant directions are on the average well pronounced and show clearly the decreasing score $S_j, j = 1, 3, 5, 7, 9$. The standard errors in this picture make it evident that in practice a rule for selecting b_n, the cutoff point, is needed. The choice of b_n presented in section 3 corresponds to drawing a horizontal line in a plot of $\widehat{S}_j$ versus j and to declare those j's corresponding to $\widehat{S}_j \geq b_n$ as elements of the desired index set J. But where to draw exactly this line?

For practical implementation we propose a technique used in principal component analysis: order first the $\widehat{S}_j$ in decreasing value. Let the sequence of ordered scores be denoted as $\widehat{S}_{(j)}$ and let

$$\psi_k = \frac{\widehat{S}_{(1)} + \ldots + \widehat{S}_{(k)}}{\sum_{j=1}^{d} \widehat{S}_j}$$

be the proportion of variation explained by the first k highest scores. As in principal component analysis keep now those j's with ψ_k greater than some prescribed percentage level. We have varied this level $1 - p$ for $p = 0.05, 0.1, 0.2$. We have also studied how sensitive the results are with respect to the choice of bandwidths. The results are shown in Table 1.

	1	2	3	4	5	6	7	8	9	10
$h = 0.3$ $perc = 0.1$	100	32	100	44	97	52	93	40	89	48
$h = 0.6$ $perc = 0.1$	100	16	100	31	99	31	94	29	77	33
$h = 1.2$ $perc = 0.1$	100	10	100	13	97	17	93	13	73	14
$h = 1.8$ $perc = 0.1$	100	12	99	13	98	18	89	16	57	17
$h = 1.2$ $perc = 0.05$	100	25	100	30	99	43	99	35	92	37
$h = 1.2$ $perc = 0.2$	100	05	94	01	87	08	46	06	24	01

Table 1. Number of times an index was selected out of $N = 100$ simulations.

The values in table 1 indicate that the frequency distribution of hitting the right model is quite robust with respect to choice of h. The first 4 lines of table 1 show for $p = 0.1$ the number of times an index

was selected to be significant out of 100 simulations. The bandwith $h = 0.3$ is apparently too small: the wrong indices (j even) were selected up to 52 percent. The range of $h = 0.6, 1.2, 1.8$ does not affect drastically the hitting frequencies except for $j = 9$ and $h = 1.8$. This observation of insensitiveness of the frequency distribution is in accordance with the theoretical result that h can in fact be chosen constant for the problem of estimating J.

The next two lines in Table 1 show the distribution for $p = 0.05$ and $p = 0.2$. As to be expected, the theoretical threshold $p = 0.05$ allows for too many false indices whereas the threshold $p = 0.2$ rather excludes correct indices. We conclude thus from this table that a 90 percent resolution of variation (i.e. $\psi \geq 0.9$) shows the best performance and could be recommended for practice.

6. Proofs

Proof of Lemma 2. Prove (3.6), (3.7) first. Note that

$$E(S_j^*) = E(E(S_j^* \mid X_1, \ldots, X_n)) = \sum_{i,\ell,t=1}^{n} A_{i\ell t} + \frac{1}{n^3 h_n^2} \sum_{i,\ell=1}^{n} E\left(E(\varepsilon_\ell^2 \mid X_1, \ldots, X_n) K^2\left(\frac{U_{\ell j} - U_{ij}}{h_n}\right)\right) \tag{6.1}$$

where $\varepsilon_\ell = \varepsilon_\ell(X_\ell)$ for brevity, and

$$A_{i\ell t} = E\left(\frac{1}{n^3 h_n^2} g_j(U_{\ell j}) g_j(U_{tj}) K\left(\frac{U_{\ell j} - U_{ij}}{h_n}\right) K\left(\frac{U_{tj} - U_{ij}}{h_n}\right)\right).$$

Assume that $j \in \{1, \ldots, p^*\}$. For $\ell \neq t$ we have

$$\begin{aligned} A_{i\ell t} &= E\left(\frac{1}{n^3 h_n^2} g_j(U_{1j}) g_j(U_{2j}) K\left(\frac{U_{1j} - U_{3j}}{h_n}\right) K\left(\frac{U_{2j} - U_{3j}}{h_n}\right)\right) \\ &= \frac{1}{n^3} \int g_j(w + t h_n) g_j(w + t' h_n) K(t) K(t') f_j(w) f_j(w + t h_n) f_j(w + t' h_n) dt dt' dw. \end{aligned}$$

Since g_j and f_j are Lipschitz continuous we have

$$\left| \int g_j(w + t h_n) f_j(w + t h_n) K(t) dt - g_j(w) f_j(w) \right| \leq \left(L \int |t| K(t) dt \right) h_n = D_1 h_n.$$

Thus,

$$\left| A_{i\ell t} - S_j n^{-3} \right| \leq D_1 h_n n^{-3}, \quad \ell \neq t, \quad j = 1, \ldots, p^*. \tag{6.2}$$

It is easy to prove also that

$$A_{i\ell\ell} = O\left(\frac{1}{n^3 h_n}\right), \qquad j = 1, \ldots, p^*. \tag{6.3}$$

where $O(\bullet)$ depends only on L and $g_{\max}$. If $j \in \{p^* + 1, \ldots, d\}$ then $g_j(\bullet) \equiv 0$, $S_j = 0$, $A_{i\ell t} = 0$, $\forall\, i, \ell, t$, and $E(S_j^*)$ equals to the second sum in the right-hand side of (6.1). Let us estimate this sum. Assumption

(A7) entails that $E(\varepsilon_\ell^2 \mid X_1, \ldots, X_n) \le C_1^{1/2}$, and hence

$$\frac{1}{n^3 h_n^2} \sum_{i,\ell=1}^n E\left(E(\varepsilon_\ell^2 \mid X_1, \ldots, X_n) K^2\left(\frac{U_{\ell j} - U_{ij}}{h_n}\right)\right) \le \tag{6.4}$$

$$\le C_1^{1/2} K^2(0) n^{-2} h_n^{-2} + \frac{1}{n h_n^2} \int K^2\left(\frac{w-u}{h_n}\right) f_j(w) f_j(u) dw du = O\left(\frac{1}{n h_n}\right), \quad n \to \infty.$$

where $O(\bullet)$ depends only on L. Clearly, (3.6), (3.7) follow from (6.1)–(6.4).

Prove (3.4) now. We have

$$E\left(E(S_j^*) - S_j^*\right)^2) \le 2E\left((S_j^* - E(S_j^* \mid X_1, \ldots, X_n))^2\right) \tag{6.5}$$

$$+2E\left((E(S_j^* \mid X_1, \ldots, X_n) - E(S_j^*))^2\right).$$

Define

$$B^{(1)} = S_j^* - E(S_j^* \mid X_1, \ldots, X_n) = \frac{1}{n^3 h_n^2} \sum_{i,\ell,t=1}^n \varepsilon_\ell \varepsilon_t K\left(\frac{U_{\ell j} - U_{ij}}{h_n}\right) K\left(\frac{U_{tj} - U_{ij}}{h_n}\right).$$

Note that

$$E((B^{(1)})^2 \mid X_1, \ldots, X_n) = \frac{1}{n^6 h_n^4} \sum_{\substack{i,\ell,t,\\ q,k,s}} E(Z_{\ell t i} Z_{ksq} \mid X_1, \ldots, X_n), \tag{6.6}$$

where

$$E(Z_{\ell t i} Z_{ksq} \mid X_1, \ldots, X_n) = \varphi(X_1, \ldots, X_n) E(\varepsilon_\ell \varepsilon_t \varepsilon_k \varepsilon_s \mid X_1, \ldots, X_n),$$

$\varphi(X_1, \ldots, X_n)$ being some function of $X_1, \ldots, X_n$. Note that $E(\varepsilon_\ell \varepsilon_t \varepsilon_k \varepsilon_s \mid X_1, \ldots, X_n) \ne 0$ only if there are either two pairs of coinciding indices in the quadruple (ℓ, t, k, s), or all the indices are the same $(\ell = t = k = s)$. In the first case $\varphi(X_1, \ldots X_n)$ is of the form

$$K^2\left(\frac{U_{\ell j} - U_{ij}}{h_n}\right) K^2\left(\frac{U_{sj} - U_{ij}}{h_n}\right),$$

or

$$K\left(\frac{U_{\ell j} - U_{ij}}{h_n}\right) K\left(\frac{U_{tj} - U_{ij}}{h_n}\right) K\left(\frac{U_{\ell j} - U_{qj}}{h_n}\right) K\left(\frac{U_{tj} - U_{qj}}{h_n}\right),$$

and in the second case it is of the form

$$K^2\left(\frac{U_{\ell j} - U_{ij}}{h_n}\right) K^2\left(\frac{U_{\ell j} - U_{qj}}{h_n}\right).$$

Moreover, from (A7) it follows that

$$|E(\varepsilon_\ell \varepsilon_t \varepsilon_k \varepsilon_s \mid X_1, \ldots, X_n)| \le C_1 < \infty.$$

Now, for the first case (two pairs of coinciding indices) the number of nonzero terms in the corresponding part of the RHS of (6.6) is $O(n^4)$, and the expected value of φ (i.e., of (6.6)) is $O(h_n^2)$. Thus the expected value of this part is $n^{-6} h_n^{-4} O(n^4 h_n^2) = O\left(\frac{1}{n^2 h_n^2}\right)$. For the second case of $\ell = t = k = s$ the number of terms is $O(n^3)$, and the expected value is $O(h_n)$. Thus the expected value of this part is $n^{-6} h_n^{-4}\, O(n^3 h_n) = O(n^{-3} h_n^{-3})$. Hence we get

$$E((B^{(1)})^2) = O\left(\frac{1}{n^2 h_n^2}\right), \quad n \to \infty. \tag{6.7}$$

Now we prove that $E((B^{(2)})^2) = O\left(\frac{1}{n}\right)$ where

$$B^{(2)} = E(S_j^* \mid X_1, \ldots, X_n) - E(S_j^*).$$

This will complete the proof of lemma. Denote

$$V_{i\ell t} = m(X_\ell)m(X_t)\frac{1}{h_n^2}K\left(\frac{U_{\ell j}-U_{ij}}{h_n}\right)K\left(\frac{U_{tj}-U_{ij}}{h_n}\right).$$

Then

$$\begin{aligned} E\left((B^{(2)})^2\right) &= \frac{1}{n^6}\sum_{\substack{i,\ell,t,\\ q,k,s}} E\Big((V_{i\ell t}-E(V_{i\ell t}))(V_{qks}-E(V_{qks}))\Big) \\ &= \frac{1}{n^6}\sum_{\substack{i,\ell,t,\\ q,k,s}} E(V_{i\ell t}V_{qks}) - \left(\frac{1}{n^3}\sum_{i,\ell,t}E(V_{i\ell t})\right)^2. \end{aligned}$$

Note that if $\{i,\ell,t\}\cap\{q,k,s\}=\emptyset$ then $E(V_{i\ell t}V_{qks})=E(V_{i\ell t})E(V_{qks})$, and

$$\frac{1}{n^6}\sum_{\{i,\ell,t\}\cap\{q,k,s\}=\emptyset} E(V_{i\ell t}V_{qks}) = \left(\frac{1}{n^3}\sum_{i,\ell,t}E(V_{i\ell t})\right)^2 - \frac{1}{n^6}\sum_{\{i,\ell,t\}\cap\{q,k,s\}\neq\emptyset} E(V_{i\ell t})E(V_{qks}).$$

Hence

$$E\left[(B^{(2)})^2\right] = \frac{1}{n^6}\sum_{\{i,\ell,t\}\cap\{q,k,s\}\neq\emptyset} E(V_{i\ell t}V_{qks}) - \frac{1}{n^6}\sum_{\{i,\ell,t\}\cap\{q,k,s\}\neq\emptyset} E(V_{i\ell t})E(V_{qks}). \tag{6.8}$$

The last sum contains $O(n^5)$ terms, as follows from (A1), (A6), (A7),

$$\begin{aligned} E(V_{qks}) &= E(V_{123}) = \frac{1}{h_n^2}\int m(U_2)m(U_3)K\left(\frac{U_{2j}-U_{1j}}{h_n}\right)K\left(\frac{U_{3j}-U_{1j}}{h_n}\right)f_j(U_{1j})dU_{1j} \\ &\quad\times\prod_{s=1}^{d} f_s(U_{2s})f_s(U_{3s})dU_{2s}dU_{3s} = O(1), \quad n\to\infty. \end{aligned}$$

Thus, the last sum is $O\left(\frac{1}{n}\right)$.

Consider the sum

$$Q_n = \frac{1}{n^6}\sum_{\mathrm{Card}(\{i,\ell,t\}\cap\{q,k,S\})=1} E(V_{i\ell t}V_{qks}).$$

This is the main term in the RHS of (6.8). Other terms are $o(Q_n)$, $n\to\infty$. For example, if $\mathrm{Card}(\{i,\ell,t\}\cap\{q,k,s\})=2$ then the corresponding sum contains $O(n^4)$ terms, and each term has an additional h_n^{-1} in the denominator. Thus it is $\sim\frac{1}{nh_n}Q_n = o(Q_n)$. We prove that $Q_n = O\left(\frac{1}{n}\right)$. Consider only the case $i=q$. Other cases are handled in a similar way. Direct calculations show that $E(V_{i\ell t}V_{iks}) = O(1)$, $n\to\infty$.

Thus the contribution of the sum with $i=q$ is $\frac{1}{n^6}O(n^5)\cdot O(1) = O\left(\frac{1}{n}\right)$, and we have proved that $E((B^{(2)})^2) = O\left(\frac{1}{n}\right)$. This together with (6.5) and (6.7) gives (3.4). (Inspection of the proof shows that all $O(\bullet)$ depend only on g_{max}, C_1 and L).

Finally, we prove (3.5). Introduce the random event

$$\mathcal{A} = \{\|\widehat{A}_n^{-1} - A^{-1}\| < \tau/\sqrt{n}\}, \quad \tau > 0.$$

It follows from Lemma 1 that

$$\begin{aligned} P\left(\sqrt{nh_n^2}|\widehat{S}_j - S_j^*| \geq a\right) &\leq P\left(\{\sqrt{nh_n^2}|\widehat{S}_j - S_j^*| \geq a\}\cap\mathcal{A}\right) \\ &\quad + \frac{1}{C^*}\left(\exp(-C^*\tau^2) + \exp(-C^*n)\right). \end{aligned} \tag{6.9}$$

Now,

$$|S_j^* - \widehat{S}_j| \le | E\left(S_j^* - \widehat{S}_j \mid X_1, \ldots, X_n\right)| + \tag{6.10}$$
$$+ \left|\left(S_j^* - E(S_j^* \mid X_1, \ldots, X_n)\right) - \left(\widehat{S}_j - E(\widehat{S}_j \mid X_1, \ldots, X_n)\right)\right|.$$

Here (cf. (6.1))

$$E\left(S_j^* - \widehat{S}_j \mid X_1, \ldots, X_n\right) = \tag{6.11}$$
$$= \tfrac{1}{n^3 h_n^2}\left[\textstyle\sum_{i,\ell,t=1}^n g_j(U_{\ell j}) g_j(U_{tj}) \alpha_{i\ell t} + \sum_{i,\ell=1}^n E(\varepsilon_\ell^2 \mid X_1, \ldots, X_n)\alpha_{i\ell\ell}\right]$$

where

$$\alpha_{i\ell t} = K\left(\frac{U_{\ell j} - U_{ij}}{h_n}\right) K\left(\frac{U_{tj} - U_{ij}}{h_n}\right) - K\left(\frac{\{\widehat{A}_n^{-1}(X_\ell - X_i)\}_j}{h_n}\right) K\left(\frac{\{\widehat{A}_n^{-1}(X_t - X_i)\}_j}{h_n}\right).$$

Since K is bounded and compactly supported there exists a constant $k_1 > 0$ such that

$$K(w) \le k_1 I\{|w| \le k_1\}. \tag{6.12}$$

Also we have

$$\left|\{\widehat{A}_n^{-1}(X_t - X_i)\}_j - (U_{tj} - U_{ij})\right| = \left|\{(\widehat{A}_n^{-1} - A^{-1})(X_t - X_i)\}_j\right| \le \tag{6.13}$$
$$\le \|\widehat{A}_n^{-1} - A^{-1}\| \mid X_t - X_i \mid \le \tfrac{C_5 \tau}{\sqrt{n}}$$

if $\mathcal{A}$ holds. Here $C_5 > 0$ is a constant depending only on the size of support D of $f_j(\bullet)$.

Lemma 4. *If* $\|\widehat{A}_n^{-1} - A^{-1}\| \le \tau/\sqrt{n}$ *then*

$$|\alpha_{i\ell t}| \le C_6 \tfrac{\tau}{h_n\sqrt{n}} K_{1n}\left(\tfrac{U_{tj} - U_{ij}}{h_n}\right) K_{1n}\left(\tfrac{U_{\ell j} - U_{ij}}{h_n}\right), \tag{6.14}$$

where

$$K_{1n}(w) = I\{|w| \le k_1 + \tfrac{C_5\tau}{\sqrt{n}h_n}\}.$$

and $C_6 > 0$ *depends only on* g_{max}, C_1, L *and* D.

Proof of Lemma 4. It follows from (6.12), (6.13) that

$$K\left(\frac{\{\widehat{A}_n^{-1}(X_t - X_i)\}_j}{h_n}\right) \le k_1 I\left(-\frac{C_5\tau}{\sqrt{n}h_n} + \frac{|U_{tj} - U_{ij}|}{h_n} \le k_1\right) = k_1 K_{1n}\left(\frac{U_{tj} - U_{ij}}{h_n}\right). \tag{6.15}$$

Now, with $w_1 = \frac{U_{tj} - U_{ij}}{h_n}$, $w_2 = \frac{\{\widehat{A}_n^{-1}(X_t - X_i)\}_j}{h_n}$, we get

$$|K(w_1) - K(w_2)| = I\{\max(|w_1|, |w_2|) \le k_1\}|K(w_1) - K(w_2)| \le \tag{6.16}$$
$$\le K_{1n}\left(\tfrac{U_{tj} - U_{ij}}{h_n}\right) L_K C_5 \tfrac{\tau}{\sqrt{n}h_n}.$$

where we used (6.13) and the Lipschitz condition on K. Here L_K is the Lipschitz constant for K. Now, (6.14) follows directly from (6.15), (6.16). ∎

Applying (6.11), (6.14), (A3), (A7) we find that if $X_1, \ldots, X_n$ are such that $\|\widehat{A}_n^{-1} - A^{-1}\| \le \tau/\sqrt{n}$, then

$$|E\left(S_j^* - \widehat{S}_j \mid X_1, \ldots, X_n\right)| \le \tag{6.17}$$
$$\le \tfrac{C_7\tau}{h_n\sqrt{n}}\left\{\tfrac{1}{n^3 h_n^2}\left[\textstyle\sum_{i,\ell,t=1}^n K_{1n}\left(\tfrac{U_{tj} - U_{ij}}{h_n}\right) K_{1n}\left(\tfrac{U_{\ell j} - U_{ij}}{h_n}\right) + \sum_{i,\ell=1}^n K_{1n}\left(\tfrac{U_{\ell j} - U_{ij}}{h_n}\right)\right]\right\}.$$

Here we used the fact that $K_{1n}^2(w) = K_{1n}(w)$.

Hence, by the Chebyshev's inequality

$$P\left(\left\{\sqrt{n}h_n^2|E\left(S_j^* - \widehat{S}_j \mid X_1,\dots,X_n\right)| \geq \frac{a}{2}\right\}\cap\mathcal{A}\right) \leq \tag{6.18}$$

$$\leq \frac{16C_7^2\tau^2}{a^2n^6h_n^4}\cdot 2\left[\sum_{\substack{i,\ell,t\\ i_1,\ell_1,t_1=1}}^{n} E\left(K_{1n}\left(\frac{U_{tj}-U_{ij}}{h_n}\right)K_{1n}\left(\frac{U_{\ell j}-U_{ij}}{h_n}\right)\times\right.\right.$$

$$\left.\times K_{1n}\left(\frac{U_{t_1j}-U_{i_1j}}{h_n}\right)K_{1n}\left(\frac{U_{\ell_1j}-U_{i_1j}}{h_n}\right)\right) +$$

$$\left. + \sum_{\substack{i,\ell\\ i_1,\ell_1=1}}^{n} E\left(K_{1n}\left(\frac{U_{\ell j}-U_{ij}}{h_n}\right)K_{1n}\left(\frac{U_{\ell_1j}-U_{i_1j}}{h_n}\right)\right)\right].$$

By the same argument as in the profof of (6.7) we get that the first sum in the RHS of (6.18) is $O(n^6h_n^4(k_1+C_5\tau/(\sqrt{n}h_n))^4)$, and the second sum is $O(n^4h_n^2(k_1+C_5\tau/(\sqrt{n}h_n))^2)$, as $n\to\infty$. Hence

$$P\left(\left\{\sqrt{n}h_n^2|E\left(S_j^* - \widehat{S}_j \mid X_1,\dots,X_n\right)| \geq \frac{a}{2}\right\}\cap\mathcal{A}\right) \leq C_8\frac{\tau^2}{a^2}\left(1+\frac{\tau}{\sqrt{n}h_n}\right)^4. \tag{6.19}$$

Next,

$$|S_j^* - E(S_j^* \mid X_1,\dots,X_n)) - (\widehat{S}_j - E(\widehat{S}_j \mid X_1,\dots,X_n))| = \left|\frac{1}{n^3h_n^2}\sum_{i,\ell,t=1}^{n}\varepsilon_\ell\varepsilon_t\alpha_{i\ell t}\right|.$$

Lemma 4 and (A7) imply

$$E\left(\left|\tfrac{1}{n^3h_n^2}\textstyle\sum_{i,\ell,t=1}^{n}\varepsilon_\ell\varepsilon_t\alpha_{i\ell t}\right|^2 I\{\mathcal{A}\}\right) \leq$$

$$\leq C_9\frac{\tau^2}{nh_n^2}\left[\frac{1}{n^6h_n^4}\sum_{\substack{i,\ell,t\\ i_1,\ell_1,t_1=1}}^{n} E\left(K_{1n}\left(\frac{U_{tj}-U_{ij}}{h_n}\right)K_{1n}\left(\frac{U_{\ell j}-U_{ij}}{h_n}\right)\times\right.\right.$$

$$\left.\left.\times K_{1n}\left(\frac{U_{t_1j}-U_{i_1j}}{h_n}\right)K_{1n}\left(\frac{U_{\ell_1j}-U_{i_1j}}{h_n}\right)\right)\right] \leq C_{10}\frac{\tau^2}{nh_n^2}\left(1+\frac{\tau}{\sqrt{n}h_n}\right)^4.$$

Thus

$$P\left(\{\sqrt{n}h_n^2|\left(S_j^* - E(S_j^* \mid X_1,\dots,X_n) - \right.\right. \tag{6.20}$$

$$\left.\left. - \left(\widehat{S}_j - E(\widehat{S}_j \mid X_1,\dots,X_n\right)| \geq \tfrac{a}{2}\}\cap\mathcal{A}\right) \leq C_{11}\frac{\tau^2}{a^2}\left(1+\frac{\tau}{\sqrt{n}h_n}\right)^4.$$

Put $\tau = \sqrt{(2/C^*)\log a}$ for $a > 1$ and $\tau = 1$ for $a \leq 1$. Then (6.9), (6.10), (6.19), (6.20) give (3.5). ∎

Proof of Theorem 2. Denote $\Omega_n = \{\widehat{J} = J\}$. By Theorem 1 we have

$$P(\overline{\Omega}_n) \to 0, \quad n\to\infty.$$

On Ω_n we have

$$\widehat{m}(x) = \sum_{k=1}^{p^*}\widehat{g}_k(\{\widehat{A}_n^{-1}(x-\widehat{B}_n)\}_k). \tag{6.21}$$

Thus for any $t \in I\!R^1$

$$\left| P\left(n^{\ell/(2\ell+1)}(\widehat{m}(x) - m(x)) \le t\right) \right. \tag{6.22}$$
$$\left. - P\left(n^{\ell/(2\ell+1)}\left[\sum_{k=1}^{p^*}\{\widehat{g}_k(\{\widehat{A}_n^{-1}(x-\widehat{B}_n)\}_k) - g_k(\{A^{-1}(x-B)\}_k)\}\right] \le t\right) \right| \le P\left(\overline{\Omega}_n\right).$$

Moreover, the following lemma is true.

Lemma 5.

$$n^{\ell/(2\ell+1)} \sum_{k=1}^{p^*} \left\{\widehat{g}_k(\{\widehat{A}_n^{-1}(x-\widehat{B}_n)\}_k) - g_k^*(\{A^{-1}(x-B)\}_k)\right\} \xrightarrow{p} 0,$$

as $n \to \infty$, *where*

$$g_k^*(t) = \frac{r_k^*(t)}{f_k^*(t)} = \frac{\frac{1}{n}\sum_{i=1}^n K_{h_{0n}}(t - U_{ik})Y_i}{\frac{1}{n}\sum_{i=1}^n K_{h_{0n}}(t - U_{ik})}.$$

Proof of Lemma 5 will be given later.

It follows from (6.21), (6.22) and from Lemma 5 that the asymptotic distribution of

$$n^{\ell/(2\ell+1)}(\widehat{m}(x) - m(x))$$

is the same as the asymptotic distribution of

$$n^{\ell/(2\ell+1)}\left(\sum_{k=1}^{p^*}(g_k^*(u_k) - g_k(u_k))\right) = \sum_{k=1}^{p^*} s_{kn}\eta_{kn} = (\mathbf{a}_n, \boldsymbol{\eta}_n)$$

where $u_k = (A^{-1}(x - B))_k$,

$$\begin{aligned} s_{kn} &= \frac{1}{f_k^*(u_k)}, \\ \eta_{kn} &= n^{\ell/(2\ell+1)}(r_k^*(x_k) - g_k(u_k)f_k^*(x_k)), \\ \mathbf{a}_n &= (s_{1n}, \ldots, s_{p^*n}) \\ \boldsymbol{\eta}_n &= (\eta_{1n}, \ldots, \eta_{p^*n}), \end{aligned}$$

and $(\cdot, \cdot)$ is the scalar product.

It follows from Parzen (1962) that under the assumptions of the theorem

$$f_k^*(u_k) \xrightarrow{p} f_k(u_k), \quad k = 1, \ldots, p^*, \tag{6.23}$$

as $n \to \infty$. Define

$$\mathbf{a} = \left(\frac{1}{f_1(u_1)}, \ldots, \frac{1}{f_{p^*}(u_{p^*})}\right).$$

If we prove that

$$\boldsymbol{\eta}_n \xrightarrow{\mathcal{D}} \mathcal{N}(\mathbf{b}, \Sigma_0), \quad n \to \infty, \tag{6.24}$$

where $\mathbf{b}$ is a vector and $\Sigma_0 > 0$ is some matrix then

$$(\mathbf{a}_n, \boldsymbol{\eta}_n) \xrightarrow{\mathcal{D}} \mathcal{N}((\mathbf{a}, \mathbf{b}), \mathbf{a}^T\, \Sigma_0\, \mathbf{a}), \quad n \to \infty,$$

since $\mathbf{a}_n \xrightarrow{p} \mathbf{a}$. Thus it remains to prove (6.24) with appropriate $\mathbf{b}$ and Σ_0.

The expected value of the kth component η_{kn} is

$$E(\eta_{kn}) = \frac{1}{\beta n^{\ell/(2\ell+1)}} E\left(\sum_{i=1}^{n}\left(\sum_{j=1}^{p^*} g_j(U_{ij}) - g_k(u_k)\right) K\left(\frac{U_{ik}-u_k}{h_{0n}}\right)\right).$$

Since U_{ij} and U_{ik} are independent for $j \neq k$, and (A5) holds then

$$\begin{aligned}(6.25)\qquad E(\eta_{kn}) &= \frac{1}{\beta}\frac{n}{n^{\ell/(2\ell+1)}}\int (g_k(z)-g_k(u_k))\,K\left(\frac{z-u_k}{h_{0n}}\right) f_k(z)dz \\ &= n^{\ell/(2\ell+1)}\frac{h_{0n}^{\ell-1}}{(\ell-2)!}\int_0^1 (1-\theta)^{\ell-2}\left[\int K(w)w^{\ell-1}H^{(\ell-1)}(x_k+wh_n\theta)dw\right]d\theta,\end{aligned}$$

where $H(v) = (g_k(v) - g_k(u_k))f_k(v)$, $v \in I\!R^1$, and we used the Taylor expansion and (A12) with $j = 1, \ldots, \ell-2$.

Now

$$(6.26)\qquad \int K(w)w^{\ell-1}H^{\ell-1}(u_k+wh_{0n}\theta)dw = \theta h_{0n}\int K(w)w^{\ell}H^{(\ell)}(u_k;w)dw + o(h_{0n}),$$

where again we used (A12) with $j = \ell - 1$.

From (6.25) and (6.26) we get

$$(6.27)\qquad E(\eta_{kn}) = \beta^{\ell}\frac{1}{\ell!}\int K(w)w^{\ell}H^{(\ell)}(u_k;w)dw + o(1), \quad n \to \infty.$$

Now calculate the components of the covariance matrix of η_n.

$$E(\eta_{kn}\eta_{tn}) - E(\eta_{kn})E(\eta_{tn}) = \frac{1}{\beta^2 n^{2\ell/(2\ell+1)}}\sum_{i,s=1}^{n} E(W_{ik}\,W_{st}) - E(\eta_{kn})E(\eta_{tn}),$$

where

$$W_{ik} = \left[\varepsilon_i(X_i) + \sum_{j=1}^{p^*} g_j(U_{ij}) - g_k(u_k)\right] K\left(\frac{U_{ik}-u_k}{h_{0n}}\right).$$

If $i \neq s$ then

$$E(W_{ik}\,W_{st}) = E(W_{ik})E(W_{st})$$

by independence of (ε_i, X_i) and (ε_s, X_s). Thus

$$\begin{aligned}(6.28)\qquad &E(\eta_{kn}\eta_{tn}) - E(\eta_{kn})E(\eta_{tn}) \\ &= \tfrac{1}{\beta^2 n^{2\ell/(2\ell+1)}}\left[\textstyle\sum_{i=1}^{n} E(W_{ik}W_{it}) - \sum_{i=1}^{n} E(W_{ik})E(W_{it})\right].\end{aligned}$$

Note that

$$E(W_{ik}) = \frac{\beta n^{\ell/(2\ell+1)}}{n}E(\eta_{kn}) = O(n^{-1/(2\ell+1)})$$

by (6.27). Thus

$$(6.29)\qquad \frac{1}{\beta^2 n^{2\ell/(2\ell+1)}}\sum_{i=1}^{n} E(W_{ik})E(W_{it}) = o(1), \quad n \to \infty.$$

From (6.28) and (6.29) we obtain

$$(6.30)\qquad E(\eta_{kn}\eta_{tn}) - E(\eta_{kn})E(\eta_{tn}) = \frac{1}{\beta^2 n^{2\ell/(2\ell+1)}}\sum_{i=1}^{n} E(W_{ik}W_{it}) + o(1).$$

Now

$$E(W_{ik}W_{it}) = E\left[\varepsilon_i^2(X_i)K\left(\frac{U_{ik}-u_k}{h_{0n}}\right)K\left(\frac{U_{it}-u_t}{h_{0n}}\right)\right] \tag{6.31}$$
$$+E\left[K\left(\frac{U_{ik}-u_k}{h_{0n}}\right)K\left(\frac{U_{it}-u_t}{h_{0n}}\right)\left(\sum_{j=1}^{p^*} g_j(U_{ij}) - g_k(u_k)\right)\left(\sum_{j=1}^{p^*} g_j(U_{ij}) - g_t(u_t)\right)\right].$$

Here for $k \neq t$

$$E\left[\varepsilon_i^2(X_i)K\left(\frac{U_{ik}-u_k}{h_{0n}}\right)K\left(\frac{U_{it}-u_t}{h_{0n}}\right)\right] \tag{6.32}$$
$$= \int E\left((Y_1 - m(X_1))^2 \mid X_1 = Az + B\right)K\left(\frac{z_k-u_k}{h_{0n}}\right)K\left(\frac{z_t-u_t}{h_{0n}}\right)\prod_{j=1}^d f_j(z_j)dz_j = O(h_n^2),$$
$$n \to \infty, z = (z_1, \ldots, z_d).$$

(The last equality follows from (A8) and from the boundedness of $f_j(\bullet)$, $j = 1, \ldots, d$). For $k = t$ by continuity of $V(z)$ and f_j we get

$$E\left(\varepsilon_i^2(X_i)K^2\left(\frac{U_{ik}-u_k}{h_{0n}}\right)\right) \tag{6.33}$$
$$= h_{0n}\left(f_k(u_k)\left[\int V(z_1, \ldots, z_{k-1}, u_k, z_{k+1}, \ldots, z_d)\prod_{\substack{j=1\\ j\neq k}}^d f_j(z_j)dz_j\right]\int K^2(w)dw + o(1)\right), n \to \infty.$$

Consider now the last summand in (6.31). Since all $g_j(\bullet)$ are bounded, then for $k \neq t$

$$E\left[K\left(\frac{U_{ik}-u_k}{h_{0n}}\right)K\left(\frac{U_{it}-u_t}{h_{0n}}\right)\left(\sum_{j=1}^{p^*} g_j(U_{ij}) - g_k(u_k)\right)\left(\sum_{j=1}^{p^*} g_j(U_{ij}) - g_t(u_t)\right)\right] \tag{6.34}$$
$$\leq C\,E\left(K\left(\frac{U_{ik}-u_k}{h_{0n}}\right)K\left(\frac{U_{it}-u_t}{h_{0n}}\right)\right) = O(h_n^2), \quad n \to \infty,$$

where $C > 0$ is a constant.

For $k = t$

$$E\left(K^2\left(\frac{U_{ik}-u_k}{h_{0n}}\right)\left(\sum_{j=1}^{p^*} g_j(U_{ij}) - g_k(u_k)\right)^2\right) \tag{6.35}$$
$$= E\left(K^2\left(\frac{U_{ik}-u_k}{h_{0n}}\right)\left(\sum_{j\neq k} g_j(U_{ij})\right)^2\right)$$
$$+2E\left(K^2\left(\frac{U_{ik}-u_k}{h_{0n}}\right)\left(\sum_{j\neq k} g_j(U_{ij})\right)(g_k(U_{ik}) - g_k(u_k))\right)$$
$$+E\left(K^2\left(\frac{U_{ik}-u_k}{h_{0n}}\right)(g_k(U_{ik}) - g_k(u_k))^2\right) = \overline{Q}_1 + \overline{Q}_2 + \overline{Q}_3.$$

Since U_{ij} is independent of U_{ik}, $j \neq k$, and $E(g_j(U_{1j})) = 0$ we obtain

$$\overline{Q}_1 = \int K^2\left(\frac{z-u_k}{h_{0n}}\right)f_k(z)dz E\Big(\sum_{j\neq k} g_j(U_{1j})\Big)^2 \tag{6.36}$$
$$= h_{0n}\left(\int K^2(w)dw\; f_k u(u_k) + o(1)\right)\sum_{j\neq k} E(g_j^2(U_{1j})),$$

$$\overline{Q}_2 = 0. \tag{6.37}$$

By continuity of $g_k(\bullet)$, $f_k(\bullet)$

$$\overline{Q}_3 = E\left(K^2\left(\frac{U_{ik}-u_k}{h_{0n}}\right)(g_k(U_{ik}) - g_k(u_k))^2\right) \tag{6.38}$$
$$= \int K^2\left(\frac{z-u_k}{h_{0n}}\right)f_k(z)(g_k(z) - g_k(u_k))^2 dz = o(h_{0n}), \quad n \to \infty.$$

Putting (6.31)-(6.38) together we get

$$E(W_{ik}W_{it}) = \begin{cases} o(h_{0n}), & n \to \infty,\ k \neq t, \\ h_{0n}\Delta_k + o(h_{0n}), & n \to \infty,\ k = t, \end{cases}$$

$$\text{where} \Delta_k = f_k(u_k) \int K^2(w)dw \left[\sum_{j \neq k} E(g_j^2(U_{1j})) + \int V(z_1, \ldots, z_{k-1}, u_k, z_{k+1}, \ldots, z_d) \prod_{\substack{j=1 \\ j \neq k}}^{d} f_j(z_j)dz_j \right].$$

This together with (6.30) entails

$$\lim_{n \to \infty} (E(\eta_{kn}\eta_{tn}) - E(\eta_{kn})E(\eta_{tn})) = \begin{cases} 0, & k \neq t, \\ \frac{1}{\beta}\Delta_k, & k = t. \end{cases} \tag{6.39}$$

To prove Theorem 2 it suffices to show that in addition to (6.27) and (6.39) the Lyapunov condition of the CLT is satisfied. We can express η_{kn} in the form

$$\eta_{kn} = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \xi_{ikn}$$

where

$$\xi_{ikn} = \frac{1}{\beta} n^{1/2(2\ell+1)} \left(\varepsilon_i(X_i) + \sum_{j=1}^{p^*} g_j(U_{ij}) - g_k(u_k) \right) K\left(\frac{U_{ik} - u_k}{h_{0n}} \right).$$

The Lyapunov condition is satisfied since, by (A1), (A3), (A8),

$$E(\xi_{1kn}^4) \leq \tfrac{8}{\beta^4} n^{2/(2\ell+1)} \left[E\left(\varepsilon_i^4(X_i) K^4\left(\tfrac{U_{ik}-u_k}{h_{0n}} \right) \right) \right.$$
$$\left. + E\left(\left(\textstyle\sum_{j=1}^{p^*} g_j(U_{ij}) - g_k(u_k) \right)^4 K^4\left(\tfrac{U_{ik}-u_k}{h_{0n}} \right) \right) \right]$$
$$= O(n^{1/(2\ell+1)}), \quad n \to \infty.$$

This concludes the proof of Theorem 2. ∎

Proof of Lemma 5. Let $u_k = \{A^{-1}(x-B)\}_k$. In view of (6.23) and (A6) $f_k^*(u_k)$ converges in probability to a positive limit for every $k = 1, \ldots, p^*$. The standard argument of nonparametric regression (see, e.g., Härdle (1990)) shows that under the assumptions of Theorem 2

$$r_k^*(u_k) \xrightarrow{p} f_k(u_k)g_k(u_k), \quad \text{as} \quad n \to \infty.$$

Hence to prove the lemma it suffices to show that

$$R_k = n^{\ell/(2\ell+1)}(\widehat{r}_k(\{A_n^{-1}(x - \widehat{B}_n)\}_k) - r_k^*(u_k)) \xrightarrow{p} 0, \tag{6.40}$$

$$\Phi_n = n^{\ell/(2\ell+1)}(\widehat{f}_k(\{A_n^{-1}(x - \widehat{B}_n)\}_k) - f_k^*(u_k)) \xrightarrow{p} 0, \tag{6.41}$$

as $n \to \infty$. We prove only (6.40) since the proof of (6.41) is quite similar.

Denote

$$\widehat{A}_{n,i,m} = \left(\frac{1}{n-2} \sum_{\substack{\ell=1 \\ \ell \neq i,m}}^{n} (X_\ell - \widehat{B}_{n,i,m})(X_\ell - \widehat{B}_{n,i,m})^T \right)^{1/2},$$

$$\widehat{B}_{n,i,m} = \frac{1}{n-2} \sum_{\substack{\ell=1 \\ \ell \neq i,m}}^{n} X_\ell,$$

$$\nu_i = Y_i\left[K\left(\frac{\{\widehat{A}_n^{-1}(X_i-x)\}_k}{h_{0n}}\right)-K\left(\frac{\{A^{-1}(X_i-x)\}_k}{h_{0n}}\right)\right],$$

$$\nu_{im} = Y_i\left[K\left(\frac{\{\widehat{A}_{n,i,m}^{-1}(X_i-x)\}_k}{h_{0n}}\right)-K\left(\frac{\{A_n^{-1}(X_i-x)\}_k}{h_{0n}}\right)\right].$$

Introduce the random events

$$\mathcal{B}_{n0} = \left\{\|\widehat{A}_n^{-1}-A^{-1}\|\le n^{-\delta}\right\}$$

where $\dfrac{\ell+2}{2(2\ell+1)}<\delta<\dfrac{1}{2}$, and

$$\mathcal{B}_n = \mathcal{B}_{n0}\cap\left\{\max_{1\le i,m\le n}\|\widehat{A}_{n,i,m}^{-1}-A_n^{-1}\tilde{}\|\le\frac{C_0}{n}\right\}$$

for some fixed $C_0>0$. By (3.3)

(6.42) $$P\left(\overline{\mathcal{B}}_{n0}\right)\le 2\exp(-C^*n^{1-2\delta})$$

for n large enough. Also, $|X_i|$ are bounded uniformly in i since the supports of $f_j(\bullet)$ are compact by (A6). This implies that there exists some $C_0>0$ such that $\max_{1\le i,m\le n}\|\widehat{A}_{n,i,m}^{-1}-A_n^{-1}\|\le C_0/n$ (a.s.) for n large. Hence, under this choice of C_0

(6.43) $$P\left(\overline{\mathcal{B}}_n\right)\le 2\exp(-C^*n^{1-2\delta}),$$

for n large.

Since $h_{0n}=\beta n^{-1/(2\ell+1)}$ we have

(6.44) $$\begin{aligned}E(R_n^2) &= \beta^{-2}n^{2\ell/(2\ell+1)}\sum_{i,m=1}^{n}E(\nu_1\nu_m)\\ &= \beta^{-2}n^{2\ell/(2\ell+1)}[nE(\nu_1^2)+n(n-1)E(\nu_1\nu_2)].\end{aligned}$$

To prove (6.40) we show that $E(R_n^2)\to 0$ as $n\to\infty$. Using the same argument as in (6.15), (6.16) we find that if $\mathcal{B}_{n0}$ holds then

(6.45) $$|\nu_1|\le|Y_i|\frac{C_{12}}{n^\delta h_{0n}}K_{2n}\left(\frac{U_{ik}-u_k}{h_{0n}}\right)$$

where $C_{12}>0$ does not depend on i, and

$$K_{2n}(w)=I\{|w|\le k_1+C_5n^{-\delta}h_{0n}^{-1}\}.$$

Note that $n^{-\delta}h_{0n}^{-1}=o(1)$ as $n\to\infty$, and thus supp K_{2n} is bounded. Therefore

(6.46) $$E\left(\nu_1^2I\{\mathcal{B}_{n0}\}\right)\le C_{12}^2n^{-2\delta}h_{0n}^{-2}E\left(Y_1^2K_{2n}^2\left(\tfrac{U_{1k}-u_k}{h_{0n}}\right)\right)=O\left(n^{-2\delta}h_{0n}^{-1}\right),\quad n\to\infty.$$

By Cauchy–Schwartz inequality and (6.42)

(6.47) $$E\left(\nu_1^2I\{\overline{\mathcal{B}}_{n0}\}\right)\le\left(E(\nu_1^4)\right)^{1/2}\left(P(\overline{\mathcal{B}}_{n0})\right)^{1/2}\le C_{13}\exp(-C^*n^{-(1-2\delta)}/2),$$

where we used the boundedness of $E(\nu_1^4)$ which follows from (A3) and (A6), (A7).

The substitution of (6.46), (6.47) into (6.44) gives

(6.48) $$\begin{aligned}E(R_n^2) &= O\left(n^{(2/(2\ell+1))-2\delta}\right)+E(\nu_1\nu_2)O\left(n^{(2\ell+2)/(2\ell+1)}\right)\\ &= E(\nu_1\nu_2)O\left(n^{(2\ell+2)/(2\ell+1)}\right)+o(1).\end{aligned}$$

Now

$$|E(\nu_1\nu_2) - E(\nu_{12}\nu_{21})| \le E(|\nu_1 - \nu_{12}||\nu_2|) + |\nu_2 - \nu_{21}||\nu_{12}|) \tag{6.49}$$
$$\le E(|\nu_1 - \nu_{12}||\nu_2|I\{\mathcal{B}_n\}) + E(|\nu_{12}||\nu_2 - \nu_{21}|I\{\mathcal{B}_n\}) + C_{14}\exp(-C^*n^{1-2\delta}/2)$$

by the same argument as in (6.47).

If $\mathcal{B}_n$ holds, then, as in (6.45),

$$|\nu_{im}| \le |Y_i|\frac{C_{15}}{n^\delta h_{0n}}K_{3n}\left(\frac{U_{ik} - u_k}{h_{0n}}\right) \tag{6.50}$$

where $C_{15} > 0$ does not depend on i, m, and

$$K_{3n}(w) = I\{|w| \le k_1 + C_5(n^{-\delta}h_{0n}^{-1} + C_0 n^{-1}h_{0n}^{-1})\}.$$

Also, by the Lipschitz condition for K we get

$$|\nu_i - \nu_{im}| \le |Y_i|\frac{C_{15}}{nh_{0n}}K_{3n}\left(\frac{U_{ik} - u_k}{h_{0n}}\right), \tag{6.51}$$

uniformly on i, m if $\mathcal{B}_n$ holds. From (6.45), (6.49)–(6.51), we deduce the following:

$$\begin{aligned} |E(\nu_1\nu_2) &- E(\nu_{12}\nu_{21})| \\ &= O\left(E\left(|Y_1Y_2|\frac{1}{n^{1+\delta}h_{0n}^2}K_{3n}^2\left(\frac{U_{1k} - u_k}{h_{0n}}\right)\right)\right) + C_{14}\exp(-C^*n^{1-2\delta}/2) \\ &= O\left(\frac{1}{n^{1+\delta}h_{0n}^2}\right) + C_{14}\exp(-C^*n^{1-2\delta}/2) = o\left(n^{-(2\ell+2)/(2\ell+1)}\right). \end{aligned}$$

This, together with (6.48), yields

$$E(R_n^2) = E(\nu_{12}\nu_{21})O\left(n^{(2\ell+2)/(2\ell+1)}\right) + o(1). \tag{6.52}$$

Consider the event

$$\mathcal{B}_{n,1,2} = \{\|\widehat{A}_{n,1,2}^{-1}A - I_{d\times d}\| \le n^{-\delta}\}$$

where $I_{d\times d}$ is $d \times d$ identity matrix. It follows from (3.3) that

$$P\left(\overline{\mathcal{B}}_{n,1,2}\right) \le \exp(-C_{17}n^{1-2\delta})$$

for n large enough. Hence, as in (6.47),

$$E\left(\nu_{12}\nu_{21}I\{\overline{\mathcal{B}}_{n,1,2}\}\right) \le \left(E(\nu_{12}^4)\right)^{1/2}\left(P(\overline{\mathcal{B}}_{n,1,2})\right)^{1/2} \le C_{18}\exp(-C_{17}n^{-(1-2\delta)}/2), \tag{6.53}$$

It remains to evaluate $E(\nu_{12}\nu_{21}I\{\mathcal{B}_{n,1,2}\})$. Note that ν_{12} and ν_{21} are conditionally independent for fixed $X_3, X_4, \ldots, X_n$ since the matrix $\widehat{A}_{n,1,2} = \widehat{A}_{n,2,1}$ depends on $X_3, X_4, \ldots, X_n$ only. Hence

$$E\left(\nu_{12}\nu_{21}I\{\overline{\mathcal{B}}_{n,1,2}\}\right) = E\left(E(\nu_{12} \mid X_3, X_4, \ldots, X_n)E(\nu_{12} \mid X_3, X_4, \ldots, X_n)I\{\mathcal{B}_{n,1,2}\}\right) \tag{6.54}$$
$$= E\left(E^2\nu_{12} \mid X_3, X_4, \ldots, X_n)I\{\mathcal{B}_{n,1,2}\}\right).$$

Also note that, by definition of U_{1j}, u_j, $j = 1, \ldots, d$, we have

$$\frac{\left\{\widehat{A}_{n,1,2}^{-1}(X_1 - x)\right\}_k}{h_{0n}} = \sum_{j=1}^{d}\alpha_j\frac{U_{1j} - u_j}{h_{0n}}$$

where $(\alpha_1, \ldots, \alpha_d)$ is the kth row of the matrix $\widehat{A}^{-1}_{n,1,2}A$. Thus,

$$E(\nu_{12} \mid X_3, X_4, \ldots, X_n) =$$
$$= \int \sum_{m=1}^d g_m(w_m) \left[K\left(\sum_{j=1}^d \alpha_j \tfrac{w_j-u_j}{h_{0n}}\right) - K\left(\tfrac{w_k-u_k}{h_{0n}}\right)\right] \prod_{j=1}^d f_j(w_j)dw_j.$$

Now, by (A12) we have that for some $C_{19} > 0$

$$\left| K\left(\sum_{j=1}^d \alpha_j \tfrac{w_j-u_j}{h_{0n}}\right) - K\left(\tfrac{w_k-u_k}{h_{0n}}\right) - K'\left(\tfrac{w_k-u_k}{h_{0n}}\right)\left(\sum_{j=1}^d \alpha'_j \tfrac{w_j-u_j}{h_{0n}}\right)\right| \le$$
$$\le C_{19}\left(\sum_{j=1}^n \alpha'_j \tfrac{w_j-u_j}{h_{0n}}\right)^2 K_{4n}\left(\tfrac{w_k-u_k}{h_{0n}}\right)$$

where $K_{4n}(w) = I\{|w| \le k_1 + C_5 \sum_{j=1}^d |\alpha'_j| h_{0n}^{-1}\}$, and

$$\alpha'_j = \begin{array}{ll} \alpha_j, & j \ne k, \\ \alpha_j - 1, & j = k. \end{array}$$

This gives

$$|E(\nu_{12} \mid X_3, X_4, \ldots, X_n)| \le R_{1n} + R_{2n}$$

where

$$R_{1n} = \left| \int \sum_{j,m=1}^d g_m(w_m) K'\left(\frac{w_k-u_k}{h_{0n}}\right) \alpha'_j \left(\frac{w_j-u_j}{h_{0n}}\right) \prod_{j=1}^d f_j(w_j)dw_j \right|,$$

$$R_{2n} = C_{19} \int \sum_{m=1}^d |g_m(w_m)| \left(\sum_{j=1}^d \alpha'_j \frac{w_j-u_j}{h_{0n}}\right)^2 K_{4n}\left(\frac{w_k-u_k}{h_{0n}}\right) \prod_{j=1}^d f_j(w_j)dw_j.$$

In view of (A12),

$$\int K'(w)dw = 0,$$

and

$$\int K'\left(\tfrac{w_k-u_k}{h_{0n}}\right) f_k(w_k)dw_k = O(h_{0n}^2),$$
$$\int K'\left(\tfrac{w_k-u_k}{h_{0n}}\right) g_m(u_m) \tfrac{w_m-u_m}{h_{0n}} f_k(w_k) f_m(w_m)dw_k dw_m = O(h_{0n}), \qquad m \ne k.$$

Using (A5) we find

$$\int g_m(w_m) \frac{w_j-u_j}{h_{0n}} K'\left(\frac{w_k-u_k}{h_{0n}}\right) f_m(w_m) f_j(w_j) f_k(w_k)dw_m\, dw_j\, dw_k = 0, \qquad m \ne j, \quad m \ne k.$$

Quite similarly we show that the integrals of summands in R_{1n} for the cases of $m = j \ne k$, $m = j = k$ are of order $O(h_{0n})$. Thus $R_{1n} = O(\alpha'_* h_{0n})$, where $\alpha'_* = \max_{j=1,\ldots,d} |\alpha'_j|$.

With R_{2n} we use the rough estimate

$$R_{2n} = O\left((\alpha'_*)^2 h_{0n}^{-2} \int K_{4n}\left(\tfrac{w_k-u_k}{h_{0n}}\right) dw_k\right) = O\left((\alpha'_*)^2 h_{0n}^{-1}(k_1 + C_5 d\alpha'_* h_{0n}^{-1})\right).$$

Note that $\alpha'_* \le \|\widehat{A}^{-1}_{n,1,2}A - I_{d\times d}\|$, and thus on $\mathcal{B}_{n,1,2}$ we have $\alpha'_* \le n^{-\delta}$. Hence, on $\mathcal{B}_{n,1,2}$

$$R_{1n} = O(n^{-\delta - 1/(2\ell+1)}), \qquad R_{2n} = O(n^{-2\delta + 1/(2\ell+1)}).$$

This gives with (6.54), (6.55):

$$E(\nu_{12}\nu_{21} I\{\mathcal{B}_{n,1,2}\}) = o(n^{-(2\ell+2)/(2\ell+1)}).$$

Finally, (6.52), (6.53), (6.57) prove that $E(R_{1n}^2) \to 0$. ∎

REFERENCES

Akaike, H. (1974). A new look at the statistical identification model. *IEEE transaction* AC, 19, 6.

Akaike, H. (1977). On entropy maximization principle. *Applications in Statistics*, ed. by P.R. Krishaiah, Amsterdam.

Friedman, J. and Tukey, J. (1974).A projection pursuit algorithm for exploratory data analysis. *IEEE Transactions on Computers* C-23, 881-90.

Friedman, J. and Stuetzle, W. (1981). Projection pursuit regression, *JASA*, 76, 817-823

Flury, B. and Riedwyl, H. (1988). Multivariate Statistics. A practical Approach. *Chapman and Hall, London.*

Hall, P. (1989) On Projection pursuit regression. *Annals of Statistics*, 17, 573-588.

Hastie, T. and Tibshirani, R. (1991). Generalized Additive Models. *Chapman and Hall, London.*

Härdle, W. and Scott, D.W. (1992). Smoothing in high and low dimension weighted averaging of rounded points. *Computational Statistics*, 7,97 –128.

Härdle, W. and Marron, J.S. (1991). Bootstrap simultaneous errors bars for nonparametric regression. *Annals of Statistics*, 19, 778-796.

Härdle, W. (1990). Applied Nonparametric Regression. Econometric Monograph Series 19. Cambridge University Press.

Ibgragimov, I.A. and Hasminskii, R.Z. (1980). On nonparametric estimation of regression, *Soviet Math. Dokl.*, 21, 810-4.

Mallows, C.L. (1973). Some comments on C_p. *Technometrics*, 15, 661-675.

Mardia, K.V., Kent, J.T. and Bibby, J.M. (1979). Multivariate Analysis, *Academic Press, London.*

Müller, H.G. (1988). Adaptive nonparametric peak estimation. *Annals of Statistics*, 17, 1053-1069.

Nadaraya, E.A. (1964). On estimating regression. *Theory of Probability and its Applications* 10, 186-190.

Parzen, E. (1962). On estimation of a probability density and mode. *Annals of Mathematical Statistics*, 35, 1065-1076.

Stone, C.J. (1980). Optimal rates of convergence for nonparametric estimators. *Annals of Statistics*, 8, 1348-1360.

Stone, C.J. (1982). Optimal global rates of convergence for nonparametric regression. *Annals of Statistics*, 8, 1040-1053.

Stone, C.J. (1985). Additive regression and other nonparametric models. *Annals of Statistics*, 13, 685-705.

Stone, C.J. (1986). The dimensionality reduction principle for generalized additive models. *Annals of Statistics*, 14, 592-606.

Watson, G.S. (1964). Smooth regression analysis. *Sankhyā*, Series A, 26, 359-372.

Journal of Econometrics, 67, 227-257.

ELSEVIER Journal of Econometrics 67 (1995) 227–257

JOURNAL OF Econometrics

Nonclassical demand
A model-free examination of price–quantity relations in the Marseille fish market

Wolfgang Härdle[a], Alan Kirman*[,b]

[a] *Institut für Statistik und Ökonometrie, Humboldt-Universität, Berlin, Germany*
[b] *European University Institute, 50016 San Domenico di Fiesole, Italy*

Abstract

We suggest a model for a market, the Marseille wholesale fish market, in which purchases do not correspond to standard competitive demand. We use nonparametric methods to detect the properties of price–quantity relations which reveal 'strategic demand'. Our data over three months include price quantity details of each transaction for each fish, and the identities of the buyers and sellers. The observed distributions of prices are stable over time, thus the market can be treated as a repeated game. Strategic demand curves are obtained by local fitting. They are downward-sloping at the aggregate level but not in general at the individual level. Thus regularities are generated by aggregation rather than derived from individual behaviour.

Key words: Aggregation; Demand; Markets; Nonparametric; Monotonicity
JEL classification: D4; C4; C7

1. Introduction

In economics, aggregate behaviour is often tested to see if it meets restrictions that can be derived from individual maximising behaviour. Thus it is common practice to treat data arising from aggregate purchases of some commodity over

*Corresponding author.

We would like to thank Dorothea Herreiner, Jan Magnus, Steve Marrou, and Mary Morgan for helpful comments and suggestions, and Peter Møllgaard and Gilles Teyssière for effective research assistance.

SSDI 0304-4076(94)01634-C

Härdle, W. and Kirman, A. (1995) Nonclassical demand: A model-free examination of price-quantity relations in the Marseille fish market.

time as if these were the expression of the competitive demand of some representative individual. This approach involves a number of implicit assumptions, in particular that the underlying micro-data observed can be thought of as corresponding to individual Walrasian demand, and furthermore that aggregation considerations do not invalidate the use of restrictions derived from individual behaviour.

In this paper we find, empirically, that, for the particular market that we study, properties which we show to hold at the aggregate level and which might be thought of as 'downward-sloping demand curves' are not derived from similar properties at the individual level. Furthermore when this market is appropriately modelled we show that there is no theoretical reason to expect any such simple relation between individual and aggregate behaviour. In this we are following directly in the line of work by Becker (1962)[1] who showed that downward-sloping demand curves at the market level could be derived from random individual choice behaviour subject only to a budget constraint. Whereas he summarised his result as saying that 'households may be irrational and yet markets quite rational', a better summary of our results would be that 'sophisticated and complicated individual behaviour may lead to simple aggregate properties'.

In particular, we will be concerned with the properties of the purchases of particularly perishable goods, different types of fish, for which we have data at the individual level, from the Marseille fish market. Although fish markets have been widely used as an example in the economic literature, we will argue that, for several reasons, it is inappropriate to think of purchases on such markets as corresponding to Walrasian demand. Having briefly described the type of market involved, we then propose a simple theoretical model, which describes how individuals could be thought of as behaving when determining the quantities that they purchase and sell. This model shows that the quantities 'demanded' or 'supplied' by the individuals in question cannot be expected to correspond to those that would be 'demanded' or 'supplied' by the standard individual in a competitive market. Instead we develop a more appropriate theoretical notion, that of 'strategic demand and supply'.

Using this, it is clear that the appropriate equilibrium concept can be defined and corresponds to a nondegenerate price distribution (see Diamond, 1987; Butters, 1977). For the use of our model to be justified and if we are to be able to use data from the whole period that we study, it should be the case that the observed price distributions remain fairly constant over time. If this is the case, then one can think of the market as repeating itself from day to day, or at least from week to week. We test for the intertemporal stability of the price

[1] This idea has been developed recently by Gode and Sunder (1993).

distributions for some representative types of fish. In particular, it should be noted that we do this without imposing any a priori restrictions on the form of the distributions. Then, having established that these do indeed remain constant, we use nonparametric methods to fit two different aggregate price–quantity relations and find that these relations, in contrast to those at the individual level, exhibit monotonicity once a certain class of observations is identified and removed.

Since we are arguing that Walrasian demand is not the appropriate concept in our context, nor, indeed in many of the contexts in which it is used, it is worth looking at the way in which, historically, it has come to occupy such a central role in empirical studies. This is of particular interest since the market for fish, which we examine here, has been frequently used as an example.

In the nineteenth century there was a very active debate over the nature of demand and little concern about its estimation. An extensive debate took place between John Stuart Mill (1869, 1871, 1972) and Thornton (1869, 1870) over the meaning and nature of the equilibrium price in the fish market, and the interpretation that could be given to demand in such a market. More precisely, the particular structure considered in the examples they discussed, was that of an auction, and there was a suggestion that either the standard continuity property of demand was violated in the examples given or that transactions reflected disequilibrium behaviour.[2] Until Marshall, there was considerable discussion as to the correct definition of demand for a single commodity. However, in the more formal literature there was convergence on the rather abstract Walrasian notion that demand simply represented the quantity that an individual would purchase at given prices which he was unable to influence. The subsequent theoretical literature concentrated largely on the extension of the analysis to interdependent markets and the problem of demand systems rather than single demand equations still maintaining the abstract Walrasian approach. Until recently, the idea that demand should be treated in this way has not really been challenged, neither in the economic nor in the econometric literature.

Once the twentieth century literature had converged on this precise theoretical definition, econometricians concentrated on more sophisticated techniques for the estimation and identification of demand systems. The agreed definition, that of competitive demand, concerning the quantities of goods an individual would buy at given prices, were he only constrained by his income, was retained. In Working's (1927) paper the conceptual nature of demand and supply are not questioned. The only real problem for him was that of which of the two was

[2] This debate produced echoes recently, when Negishi (1985, 1986, 1989) (and Ekelund and Thommeson, 1989) discussed the precise nature of the difficulties involved in the Mill and Thornton examples.

fluctuating over time. However for many markets, and this is the subject of this paper, this conceptual framework is not satisfactory. For example, in our particular case, the wholesale fish market in Marseille, all transactions are bilateral and no prices are posted. When we look at the relation between the prices charged and the quantities purchased on this sort of market, a number of questions which were very present in the earlier debate as to the appropriate notion of 'demand' recur.

Let us, therefore, return to the implicit assumptions underlying the usual empirical analysis based on Walrasian demand theory and see whether they are appropriate in our context.

The first question that arises is whether the purchaser of a good is, in fact, the final consumer. If this is not the case, then one would have to show that properties of individual demand carry over to properties of quantities purchased by an intermediary at different prices. If one considers the simple case of a purchaser who is a retailer and has a monopoly locally of the product that he buys and resells, then it is easy to construct examples in which this will not be the case. This question was raised by Working (1927) and mentioned again in the classical studies by Schultz (1938), who although using individual properties of demand made his estimations using data for farm prices and not shop prices. More recently, in a specific study of the Belgian fish market, Barten and Bettendorf (1989) refer to this question.

The second problem arises even if one accepts that the final consumers are present on the market in question and that it does function 'competitively'. The problem is that of identification, in this case, separating out supply changes from demand changes. In a truly Walrasian, or Arrow–Debreu world such a distinction could, of course, not be made, since all transactions over time represent one supply and one demand decision taken in some initial period. However, this problem is usually circumvented in the empirical literature by making an implicit assumption of stationarity and separability, i.e., that the market is somehow repeated over time, and that decisions are taken in the same way at each point in time. This should, of course, be tested, but does mean that one can talk of successive observations. However, in this case the appropriate theory is that referred to as temporary general equilibrium theory. The problem with this is that, without full knowledge of future prices, expectations have to be taken into account. Without unreasonable assumptions about these, short-run demand loses many of the properties of its Walrasian counterpart. It does not satisfy homogeneity or the Weak Axiom of Revealed Preference, for example (see, e.g., Grandmont, 1983). Trying to fit a demand system based on the usual theoretical restrictions makes little sense therefore.

Nevertheless, if we are prepared to accept the idea that changes in the prices of fish do not result in a large amount of intertemporal substitution, then thinking of a sequence of equilibria in a market which repeats itself is more acceptable. This explains why, when considering particular markets, fish has been so widely

used as an example (e.g., by Marshall, Pareto, Hicks) since with no stocks, successive markets can be thought of as independent. In our case, when fitting our price–quantity relations we are implicitly treating price changes as resulting from random shocks to the supply of fish, although the amount available is, at least in part, a result of strategic choice.

The next problem is that of aggregation. If we fit a demand system in the usual way, we are assuming that market behaviour corresponds to that of an individual. Examination of individual data reveals none of the properties that one would expect from standard individual demand. Thus, even if such properties are found at the aggregate level, they cannot be attributed to individual behaviour. This is one side of the problem of aggregation. The other is that, even if individuals satisfy certain properties, it is by no means necessary that these properties carry over to the aggregate level (see, e.g., Sonnenschein, 1972; Debreu, 1974). The two taken together mean that there is no direct connection between micro and macro behaviour. This basic difficulty in the testing of aggregate models has recently been insisted upon (see Kirman, 1992; Summers, 1991; Lewbel, 1989) when discussing representative individual macro models, but as Lewbel observes, this has not stopped, and is unlikely to stop, the profession from testing individually derived hypotheses at the aggregate level. Hence when we establish some empirical properties of the aggregate relationships between prices charged and quantities purchased, we suggest that these should be viewed as independent of standard maximising individual behaviour.

The next point is that the organisation of the market for the product in question may not be competitive. In this case, it is not possible to talk of a single market price. If different lots of the same good are auctioned off successively, for example, the average price will not necessarily correspond to the price which would have solved the Walrasian problem for that market. The problem here is that techniques for the econometric analysis of data arising from differently organized markets such as auctions, for example, have been little developed and there is always a temptation to return to standard and sophisticated techniques, even if these should not really be applied to the type of market in question. Barten and Bettendorf (1989) are well aware of this difficulty, and suggest that the aggregate behaviour in the fish market can be reduced to that of a Walrasian mechanism by looking at an inverse demand system. They reason as follows:

> 'Price taking producers and price taking consumers are linked by traders who select a price which they expect clears the market. In practice, this means that at the auction the wholesale traders offer prices for the fixed quantities which, after being augmented with a suitable margin, are suitably low to induce consumers to buy the available quantities. The traders set the prices as a function of the quantities. The causality goes from quantity to price.'

Although the authors are only making explicit what is commonly done, it is clear that one should *prove* that, even if the auction price is well defined, it is indeed related to prices charged to consumers through a simple mark-up. Necessarily, if different purchasers pay different prices and the mark-up principle does apply, then a distribution of prices will be observed on the retail market.

This bring us to a further point. Since our market does not function as a standard auction, and individual traders strike bargains amongst themselves and are well aware of each others' identities, different prices can be, and are, charged to different purchasers for the same product. This discrimination is an important feature of the market, and there are significant variations in the average prices paid by different buyers (see Kirman and McCarthy, 1990). This means that reducing prices to averages may well lose a significant feature of the data. Furthermore, it means that the average price cannot be regarded as a reasonable sufficient statistic and that other properties of the price distribution must be taken into account. This reduces the plausibility of the argument advanced by Barten and Bettendorf.

Lastly we emphasise that, when fitting the price–quantity relations, we use nonparametric estimation techniques, since these are less likely to lead to mistakenly accepting a monotone relation, and furthermore reveal interesting features of the data that standard techniques, using predefined functional forms, would have been unlikely to detect.

Now we turn to the development of our formal model of the market.

2. A simple strategic model

As we have already suggested and as has been emphasised by many authors, the structure and organisation of the market are of particular importance in determining the nature of the equilibrium realised. We therefore give a simple model of our type of market, restricting ourselves, for simplicity, to the case of one type of fish.

We thus consider the market for one perishable product with m sellers and n buyers.[3] The market evolves in a fixed number T of rounds. Each seller i has strategies which at each round t specify a vector $x_{it} \in R_+^n$ of the prices which he will charge to each of the buyers. A strategy for each buyer j specifies at each round t a demand function $q_{jt}(p)$: $R_+ \to R_+$. In both cases the choice of the prices set and the demand functions will depend on two things: firstly, the strategies of the other players and, secondly, on who has met whom in the

[3] In Kirman and Vignes (1991) we considered a continuum of buyers and sellers, but this was to facilitate the solution of the technical problem of establishing the continuity of strategies.

market. The model is then completed by specifying a matching process which, in keeping with the literature, will be assumed to be random. Thus a matching at time t, a realisation of the random variable, will be a mapping g from the integers $J = \{1, \ldots, n\}$ to the integers $I = \{1, \ldots, m\}$. A probability distribution must then be specified over the outcomes of the matching process for every time t. One might think, as an example, of each buyer as choosing a seller with uniform probability $1/m$, independently at each time t. However, many other matching processes could be considered, including those in which some particular buyers and sellers are always matched together. A best strategy for a buyer i then will consist for each realisation of the matching process and for the associated price vectors of each seller i and demand functions of the other buyers $h \neq i$ of a demand function for each period t. Similarly, for a seller it will consist of specifying the best price vectors for each matching and each period.

Thus to sum up, the model consists of:

(1) *A basic strategy set* B for buyers which is a subset of the product space of Q where Q is the set of functions q: $R_+ \to R_+$ satisfying
   (i) every q is continuous and monotone decreasing,
   (ii) there exists a $\bar{p} > 0$ such that for every q in Q, $q(p) = 0$ for all $p > \bar{p}$.
   Thus $B \subset \prod_T Q$, i.e., $Q \times Q \times \ldots \times Q$. Furthermore, we assume B is compact and convex.

(2) *A basic strategy set* S for sellers which is a compact convex subset of R^{nT}.

(3) *A matching process* M, i.e.,
   (i) the mappings $f = (f_1, f_2, \ldots, f_n)$ where f_i: $I \to J$,
   (ii) a probability distribution over the finite set F of f.

(4) *A full strategy* $\tilde{q}$ for a buyer then associates with each f an element of Q, i.e., q: $F \to B$, and $\tilde{s}$ for a seller associates with each element of F, an element of S, i.e., s: $F \to S$. The set of full strategies for buyers is denoted $\tilde{Q}$ and for sellers $\tilde{S}$.

(5) There is *a continuous payoff function* for each player i, $\prod_i(q_1 \ldots q_n, s_{n+1} \ldots s_{n+m})$, $i = 1, \ldots, n+m$.

(6) *The response function* Γ for each player is given for buyers by Γ: $\tilde{Q}^n \times \tilde{S}^m \to \tilde{Q}$ where $\Gamma_i = \max_{q_i} \prod_i(q_1 \ldots q_{i-1}, q_i, q_{i+1} \ldots q_n, s_{n+1}, s_{n+m})$ and for sellers by Γ: $\tilde{Q}^n \times \tilde{S}^m \to \tilde{S}$ where $\Gamma_j = \max_{s_j} \prod_j(q_1 \ldots q_n, s_{n+1} \ldots s_{j-1}, s_j, s_{j+1} \ldots s_{n+m})$. We assume that for every i, $i = 1, \ldots, n$, and j, $j = n+1, \ldots, n+m$, Γ_i is continuous.

Denoting by $\tilde{Q}^n$ the n product of $\tilde{Q}$ and by $\tilde{S}^m$ the m product of $\tilde{S}$, then each Γ $(\Gamma_1 \ldots \Gamma_{n+m})$ defines a mapping from $\tilde{Q}^n \times \tilde{S}^m$ into itself. Given our assumptions, a standard fixed point argument can be used to show the existence of an equilibrium.

The market can be envisaged as follows:

Period 0: Initial stocks become available.

Period 1: Sellers specify prices, buyers specify demands. Matching takes place. Transactions occur.

Period 2: Given their information about what happened in period 1, sellers respecify prices, buyers respecify demands. Matching occurs. Exchanges follow.

Period T: Last specifications by sellers and buyers, last matching and exchanges.

As it stands, we have done no more than give a formal framework which enables us to define the concept of an equilibrium. To characterise precisely the nature of an equilibrium requires that the Γ_t be derived from maximising behaviour. For example, Γ_t for a buyer would maximise his expected utility at each round t given the known strategies of the other players and the matching up until t. The real difficulty here is *proving* the continuity of Γ_t for the players. This difficulty is illustrated by Kormendi (1979) and Benabou (1988).

Whether or not we give a complete specification of the maximising behaviour of the individuals, our model would allow for extensive price dispersion (particularly since we have assumed that price discrimination is possible as each seller knows the buyers' characteristics), there will be no necessary tendency for prices to decline during the day as is commonly supposed and, as we have mentioned, there is no a priori reason to assume that individual buyers will or will not search.

One important point to emphasise is that any strategy must be such that if the information set up to time t is the same in two realisations, the next component of the strategy at time $t-1$ should be the same. Thus it is important to specify what is known at each time. If, for instance, the individuals know only their own initial stocks and only observe their own transaction outcomes, they will be much more limited than if they observe everything that has occurred. Furthermore, it may well be the case that individuals actually choose to condition strategies on a limited part of the information they have available.

Although it is difficult to prove the continuity of strategies in a fully optimising context, it is possible that agents develop simple rules which are continuous. An interesting problem is how such rules are developed.

Having given an outline of the structure of the sort of process we examine, it is not surprising that the outcomes do not necessarily satisfy standard demand properties at the individual level, since observed transactions are the results of the interaction between buyers' and sellers' strategies. We shall refer to the observed purchases as reflecting '*strategic demand*' since they reflect the outcome of the process which clears the market at the end of the day. The difference from day to day on the market is the amount of fish available. This is due in part to exogenous factors such as weather, but is also due to the choice of sellers when anticipating demand changes. This latter factor should be incorporated into a complete model.

Table 1
Characteristics of the data for the Marseille fish market

Period	July–September 1988
Organization	Pairwise trading, no posted prices
Number of buyers	574
Number of sellers	37
Number of types of fish	129
Level of disaggregation	Every transaction recorded
Data for each transaction	a) Name of seller b) Name of buyer c) Type of fish d) Quantity sold (weight) e) Price per kilo
Time	Transactions listed in chronological order during the day for each seller
Salient features of the market	
Concentration	
Sellers	One seller accounts for 15.5% of all transactions; the six biggest sellers account for 50% of all transactions
Buyers	One buyer accounts for 14% of all transactions; all others account for less than 1%
Diversity	Average number of fish types traded by a seller during a day varies from 1 to 32, 50% of traders trade in less than 10 fish types
Fish types selected	Sole, sardine, whiting, and trout

In Table 1 a description of the data set and some of its characteristics are given, and we now turn to an analysis of some of its features.

3. Price–quantity relations

What we have learned from our theoretical analysis is that there is no a priori reason to expect any particular structure of the relationship between prices (or average prices) and quantities sold. Testing standard properties to verify the theory underlying demand functions or demand systems would make little sense in this context for the reasons we have indicated. What we are observing does not reflect consumer demand, discriminatory pricing is taking place and prices evolve strategically over the day.

However, determining whether or not our data do satisfy certain properties is of interest. The one feature that we do observe is that over the day markets do more or less clear in the sense that the surplus left unsold never exceeds 4%. Since sellers become aware, from the reactions of buyers to their offers, of the

amount available on the market and vice versa, it would not be unreasonable to expect average strategic equilibrium prices to be lower on those days where the quantity is higher, but some buyers transact early, before such information becomes available, and others only make one transaction for a given fish on a given day. Thus to deduce such a property formally would require much stronger assumptions than we have made. If we can establish such a property, i.e., of 'downward-sloping demand', it certainly could not be attributed to the normal utility-maximising model as is frequently done, but is rather a property that emerges from a rather complicated noncooperative game.

To look at this we now proceed to an empirical examination of the behaviour of the market. We might like to find out whether, for example, when we consider the four fish that we have taken as examples, the quantities purchased at each price $D(p)$ for those fish display the monotonicity property, i.e., for $p \neq p'$ and $p > 0$, $p' > 0$, p in R^4_+,

$$(D(p) - D(p'))\cdot(p - p') \leqslant 0.$$

Such a property, when $D(p)$ is interpreted as a standard demand system, is described as the 'Law of Demand' by Hildenbrand (1983) following Hicks. In particular, it implies that each partial 'own demand curve' for the fish is downward-sloping.[4] One approach would be to estimate in a standard parametric way the whole 'demand system', but since we have no a priori reason to impose any sort of functional form on the system, we have chosen to look at the weaker property, negatively sloped price quantity relations for each individual fish. In doing so we are open to the criticism that we are not taking into account substitution effects between fish. Thus, it could be argued that what we gain in using more flexible estimation methods is offset by what we lose in overlooking these effects. There are three responses to this. Firstly, many buyers such as restaurant owners have a pre-determined vector of fish quantities which they do not vary much in response to relative price changes. Secondly, there are other buyers who only buy one type of fish and therefore do not substitute. Lastly, some of the exogenous factors influencing the amount of fish available, such as weather, are common to many fish, thus limiting the amount of substitution possible. For all of these reasons we have analysed each of our four fish separately.

In undertaking our analysis of the 'demand' for each fish, we do an exercise designed to elicit some of the basic characteristics of the data. Basically, we take the data for a given fish and aggregate it by taking the quantity of that fish sold on a particular day and the weighted average price for that day. There is

[4] Of course to take observed quantities purchased as representing a marginal curve is not correct since the ceteris paribus condition is violated. However, this makes the resultant monotonicity more rather than less convincing.

a problem of separation of strategies here. There are not only variations in the supply of fish due to weather, etc., but more fish is landed on active market days by choice. The variations over the week are due in part to obvious institutional factors (fish-shops are closed on Sundays), but also to more indirect ones. As Robbins (1935) observed before his discussion of the market for herring in England:

> 'The influence of the Reformation made no change in the forces of gravity. But it certainly must have changed the demand for fish on Fridays.'

We then fit the resulting data by nonparametric smoothing methods. Sufficient details to give a basic understanding of the techniques used are given in Appendix A (for a full account see Härdle, 1990). We use nonparametric methods since they enable us to pick up any lack of monotonicity of the fitted curve over some particular price range. Nevertheless in all four cases the fitted curves are indeed monotone decreasing and two examples are given in Figs. 1a and 1b.

Simple inspection of the graphs is, of course, not sufficient, and since we have no explicit functional form for the fitted curves, we have actually to *show* that

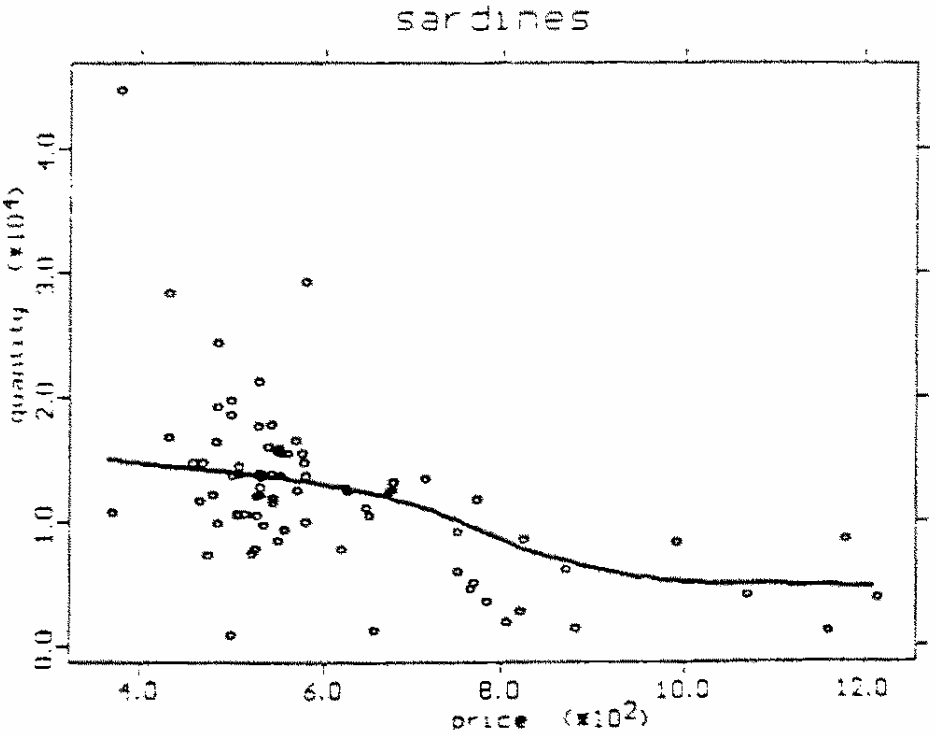


Fig. 1a

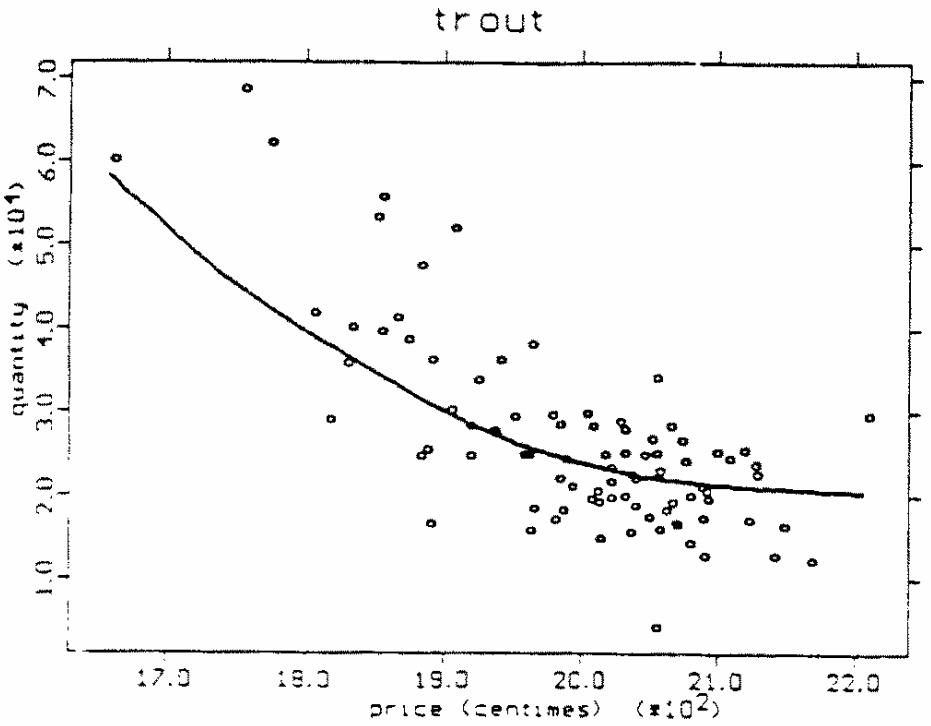


Fig. 1b

they are monotonic. This is easily done since we can check successive differences in the y values for each of the x values corresponding to the grid imposed by the original observations. If the maximum of these is negative, then it can easily be shown that the continuous fitted curve is monotone decreasing. This was the case for all of our curves. As explained in Appendix A, the local smoothing procedure used is 'optimal', and this monotonicity property is not vulnerable to changes about the optimum. Although we have established the monotonicity of the fitted relation, what we are really interested in is the monotonicity of the 'true' relation. This requires establishing an upper confidence band on the *derivative* of the function and showing that this is negative. In our case, it is enough to establish that the upper confidence bound on the maximum of the derivative is negative. However, since no theoretical results are available for this, we had to make this calculation at each discrete point on the x axis corresponding to one of the bin means. In every case the negativity property was satisfied. In the light of this evidence, that the monotonicity property is apparently robust, an economist might naively have suggested that these curves represented aggregate demand for the fish in question and that their monotonicity was derived from the underlying classical individual demands.

The important thing to re-emphasise here is that the 'nice' monotonicity property of the aggregate price quantity curves does *not* reflect and is *not* derived

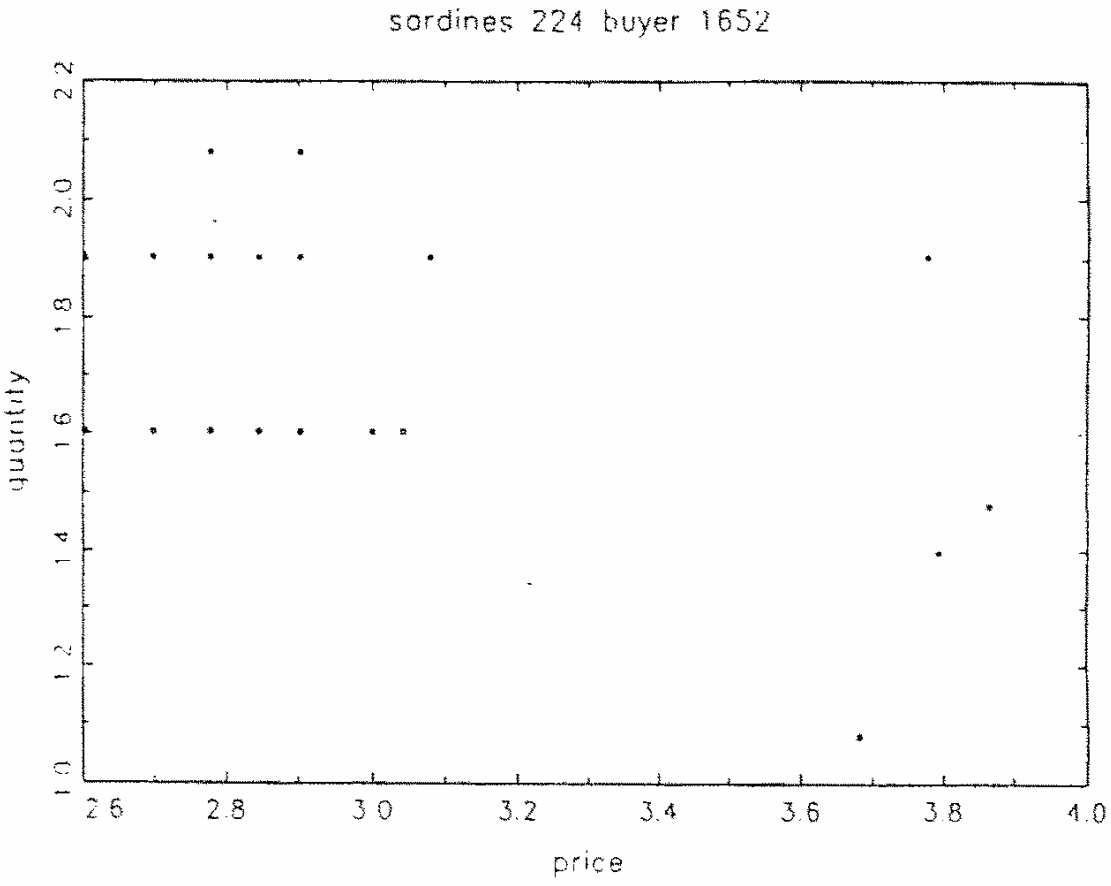


Fig. 2a

from the corresponding characteristics of individual behaviour. Nor indeed, given our discussion, should we expect it to be.

Transactions at the individual level

To illustrate the lack of 'good behaviour' at the micro level, it is therefore worth looking at the plots for the quantities of sardines purchased at different prices by three individuals. These are illustrated in Figs. 2a, 2b, and 2c. The observed price–quantity pairs of the first two buyers are far from corresponding to what classical demand theory might lead us to expect, whereas the third might conceivably meet those conditions. The sardine was chosen to illustrate this, not because it has any particular significance but since the price–quality relations for all four fish are, as we have seen, well-behaved on the aggregate level. Once again it is important to recall the nature and organization of the transactions on this market, both over the day and as it varies between matching

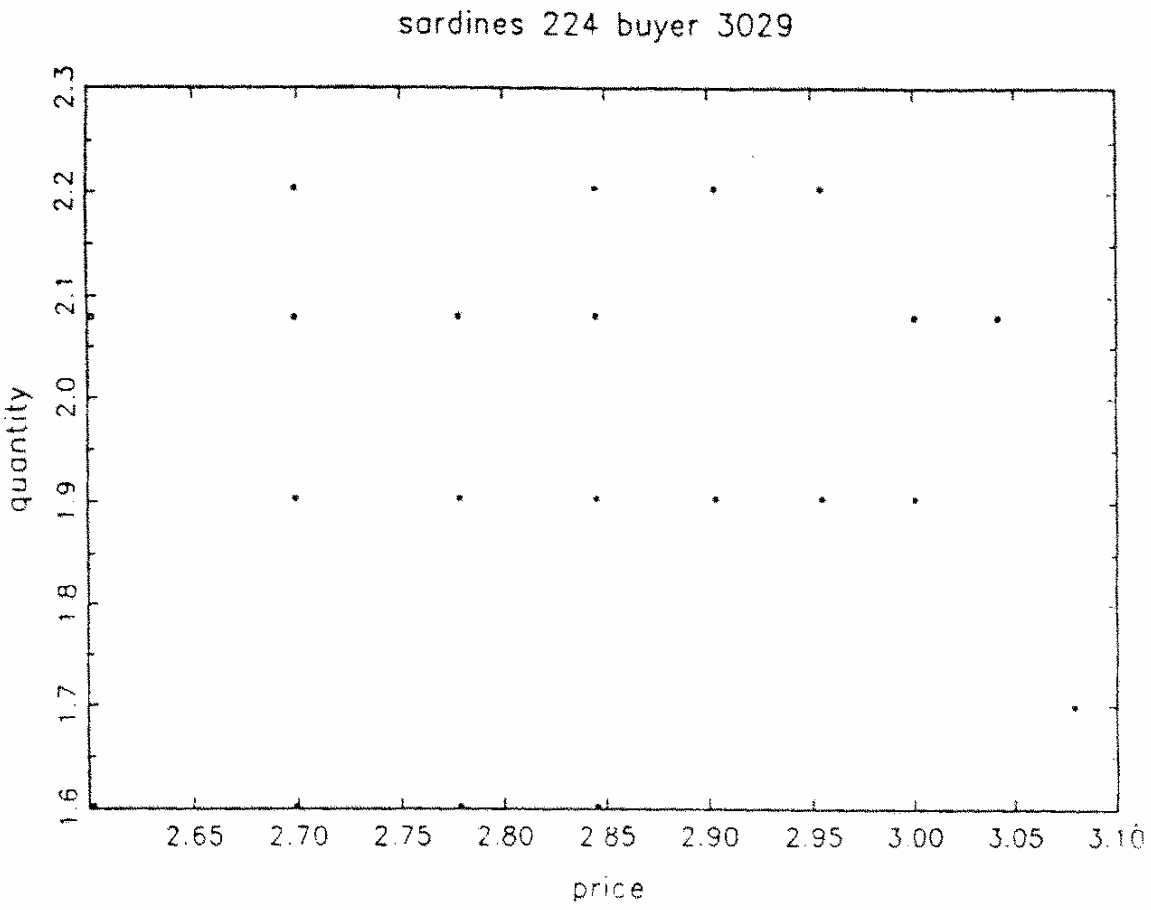


Fig. 2b

of different buyers and sellers, to understand the apparent eccentricity of the individual demands.

4. Stability of price distributions

If, after this initial examination of the data, we are to consider deriving something corresponding to short-term 'strategic demand' from observations over time in our particular market, then we have to be sure that market conditions remained essentially the same over the whole period. Since, as we have already observed, we would not expect an equilibrium price, but rather an equilibrium price distribution for the game that we have described, we should therefore check that the distributions observed in successive periods remain the same. To do this we test the hypothesis that for each individual fish the daily observed price distribution is stable over time. It is important to understand

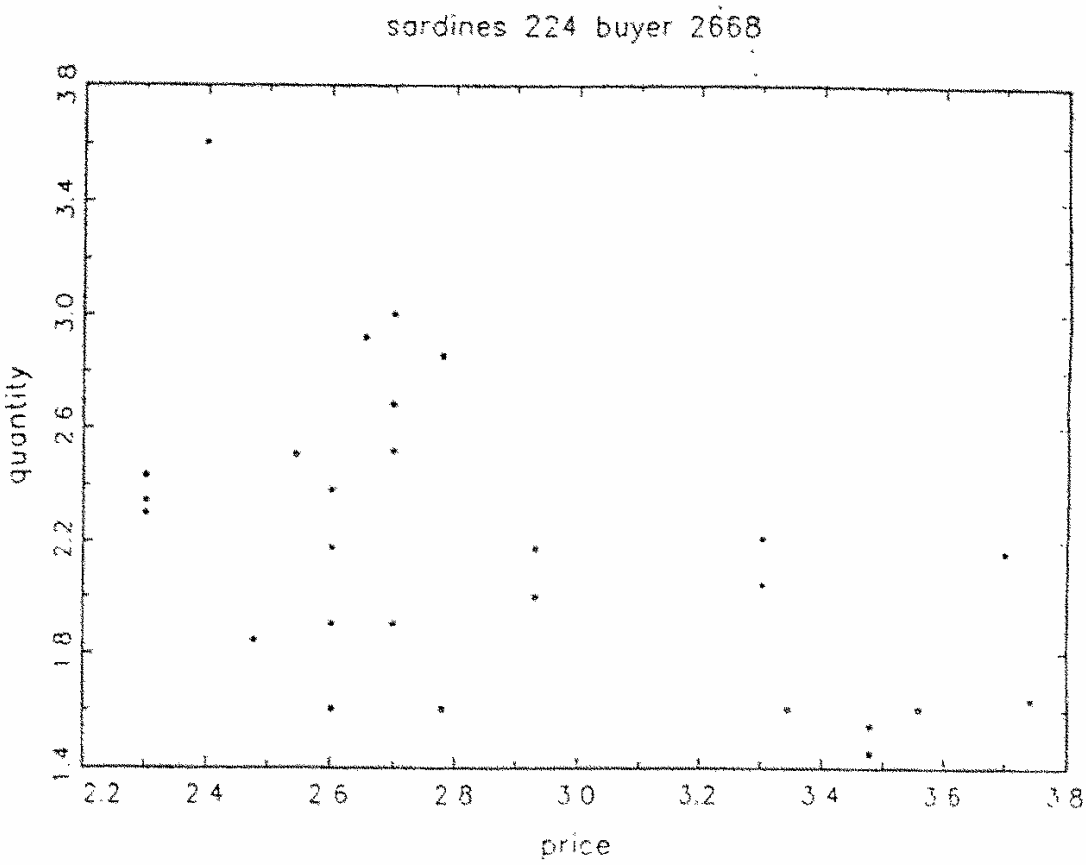


Fig. 2c

what we mean here by distribution. We count the total number of kilos transacted in each price interval. The alternative would be to count the *number of transactions* at each price level, but, in effect, we consider each kilo as a separate transaction. This distinction is usually avoided in the literature on price dispersion where individuals demand one unit of an invisible good (see Rothschild, 1973, and Diamond, 1987, for example). Thus the distribution h of prices is given by

$$h(p_j) = \frac{\sum \textit{quantities sold at prices in the jth interval}}{\textit{Total quantities sold}}.$$

We proceed by fitting a function to each of the distributions and then seeing by how much the distance of each of these functions from the others varies. In Figs. 3a and 3b the results for the three months for sardines on a general and on a focused scale are shown. Fig. 3c shows the same analysis of trout on a focused scale.

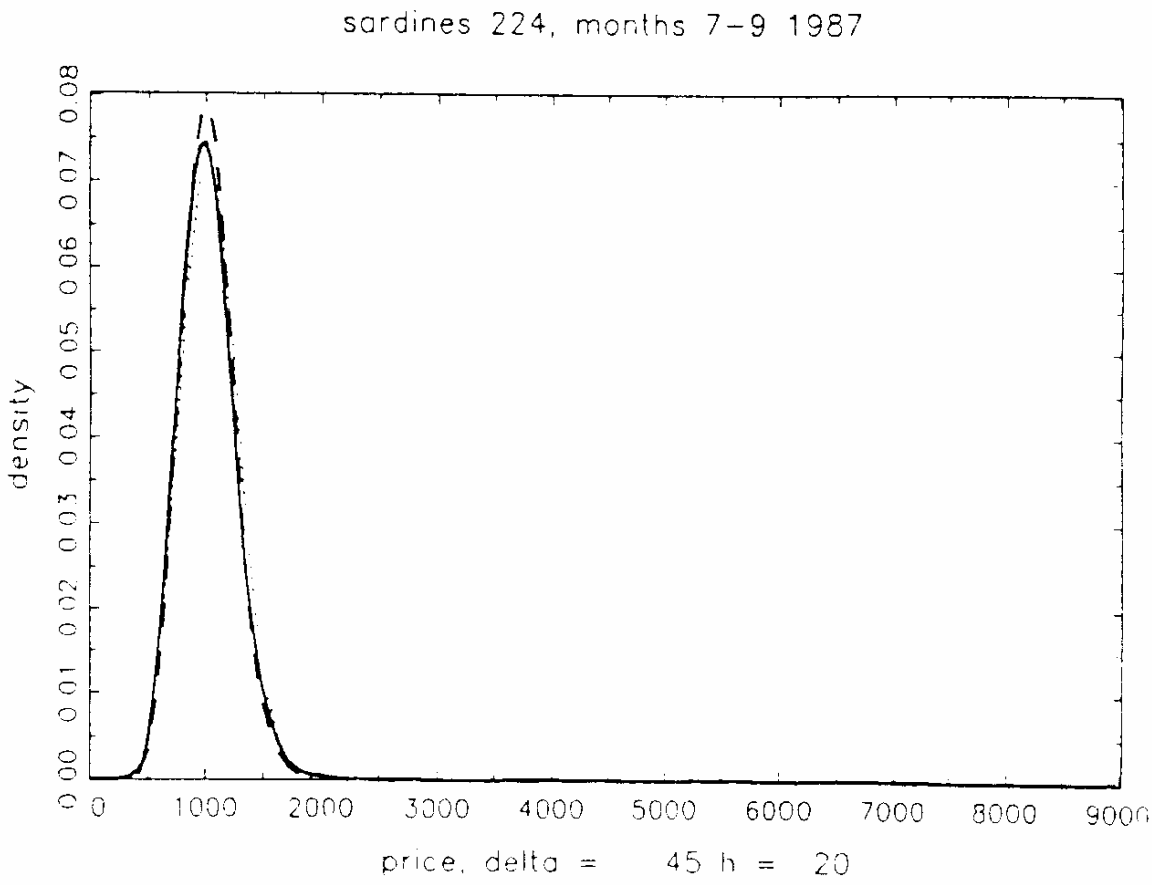


Fig. 3a

Full details of the tests for the stability of the distribution are given in Appendix B. We could not reject the hypothesis that the distributions were constant over time. That is, when we considered the following hypotheses:

$$H_0: \quad f_i = f_j, \qquad i \neq j,$$

$$H_1: \quad f_i \neq f_j, \qquad i \neq j,$$

for each of our four fish over the three months in question, in none of the cases could we reject H_0.

Since the numerical values of the confidence bounds constitute curves we have not reproduced the graphs here, but similar illustrations may be found in Härdle (1990). However, the relative stability of the smoothed fits over the three months in question can be seen in Figs. 3a to 3c.[5] The importance of the discretisation

[5] In each case, July is given by the solid line, August by the dashed line, and September by the dotted line.

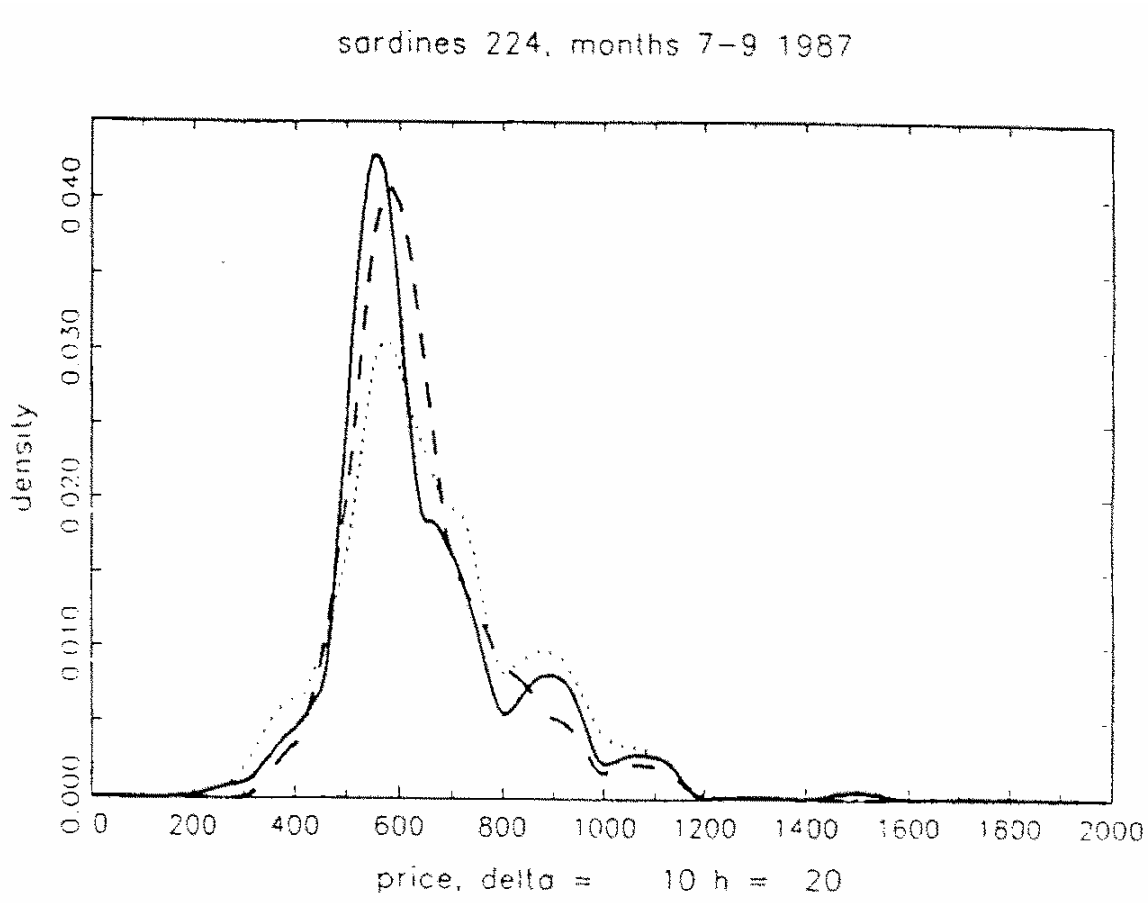


Fig. 3b

parameter d is seen by comparing the two figures for sardines, Figs. 3a and 3b. The smoothing parameter is h.

As our statistical analysis shows, once we accept the idea that we are dealing with a distribution of prices which reflects the equilibrium strategies of the different players, we cannot reject the hypothesis that these distributions are stable over time.

An important point which merits further discussion is to what extent is it legitimate to use stability tests developed for independently drawn observations on the sort of data we examine here? Two remarks can be made. Firstly, it is by no means infrequent to apply a stochastic model to data which is not derived from such a model. Stochastic models of deterministic processes are often very useful. (See Erdös and Spencer, 1974, for example.)

Thus, even if buyers met sellers in a predetermined way, without the appropriate information the modeller may well have to treat the data as if generated by some stochastic matching process.

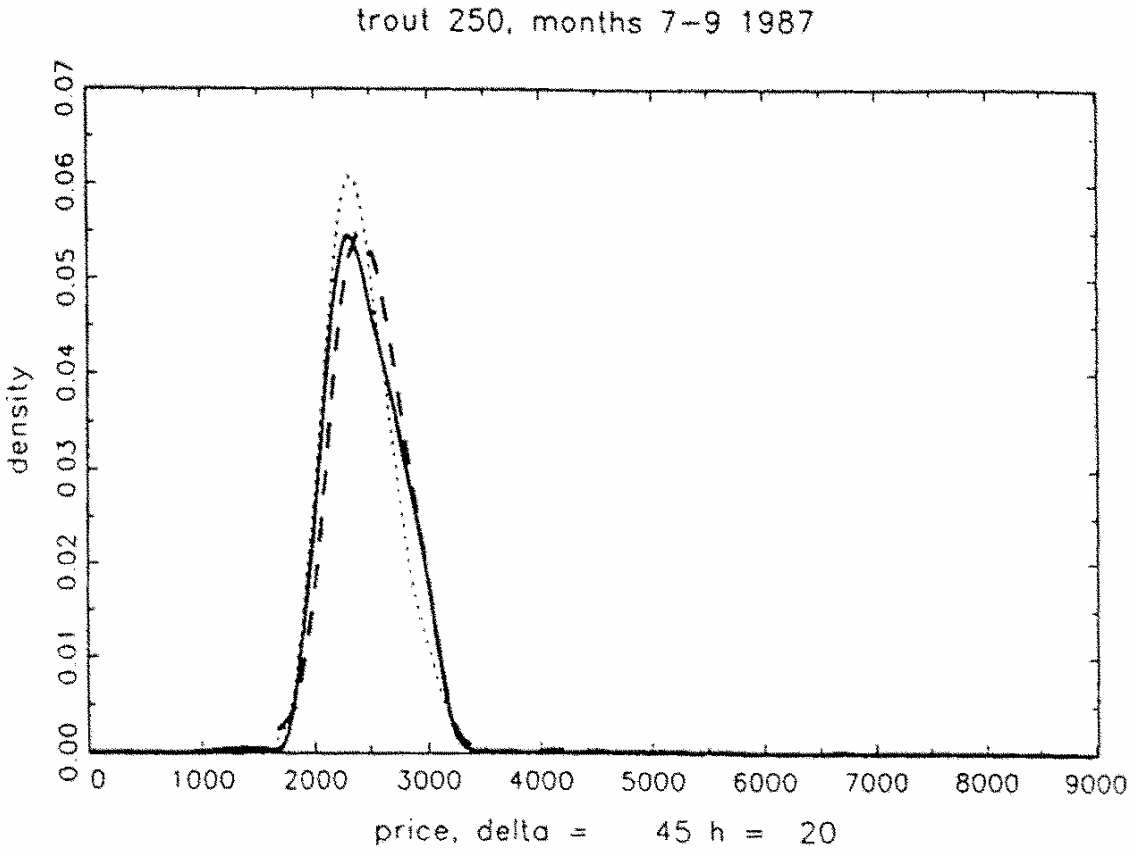


Fig. 3c

In fact, there is a certain amount of stochastic behaviour in the market in that searching for low prices does take place. The problem is that, although the evidence from the fitted densities seems to be clear, for the *statistical* tests for stability to be valid, the observations should be independently identically distributed. This cannot be strictly true, since certain buyers pay higher prices for example. Although these buyers are probably of particular types, restaurants, etc., they are only identified by code. We therefore do not have prior information on which to condition and cannot treat them as different.[6]

[6] In treating our observations as drawn from the same population in this way we are following Theil (1971), for example, who in his 'convergence' approach thought of N consumers as independent elements of an infinite consumer population and the parameters of their utility functions as identically distributed.

5. Price–quantity relations: A second analysis

When we estimated our smoothed 'demand' curves in the previous section we did so by averaging prices. However, this loses a lot of the information contained in our data. We might therefore ask a simpler question for which generally data are not available. Are more kilos of fish transacted at lower prices? By this we mean taking as one observation the total quantity transacted over the whole period at a specific price. Fitting these observations with a smoothed curve corresponds to finding the best fit for the average quantity sold for prices in some appropriately chosen price interval.

We thus obtained smoothed curves by summing over quantities at each given price and by 'binning' data in price intervals. The results are shown in Figs. 4a to 4c. Aggregation over transactions produces nicely behaved, that is, essentially monotone demand curves for three out of the four fish. Sardines are illustrated in Figs. 4a and 4b. Whiting in Fig. 4c, however, displays a monotone increasing

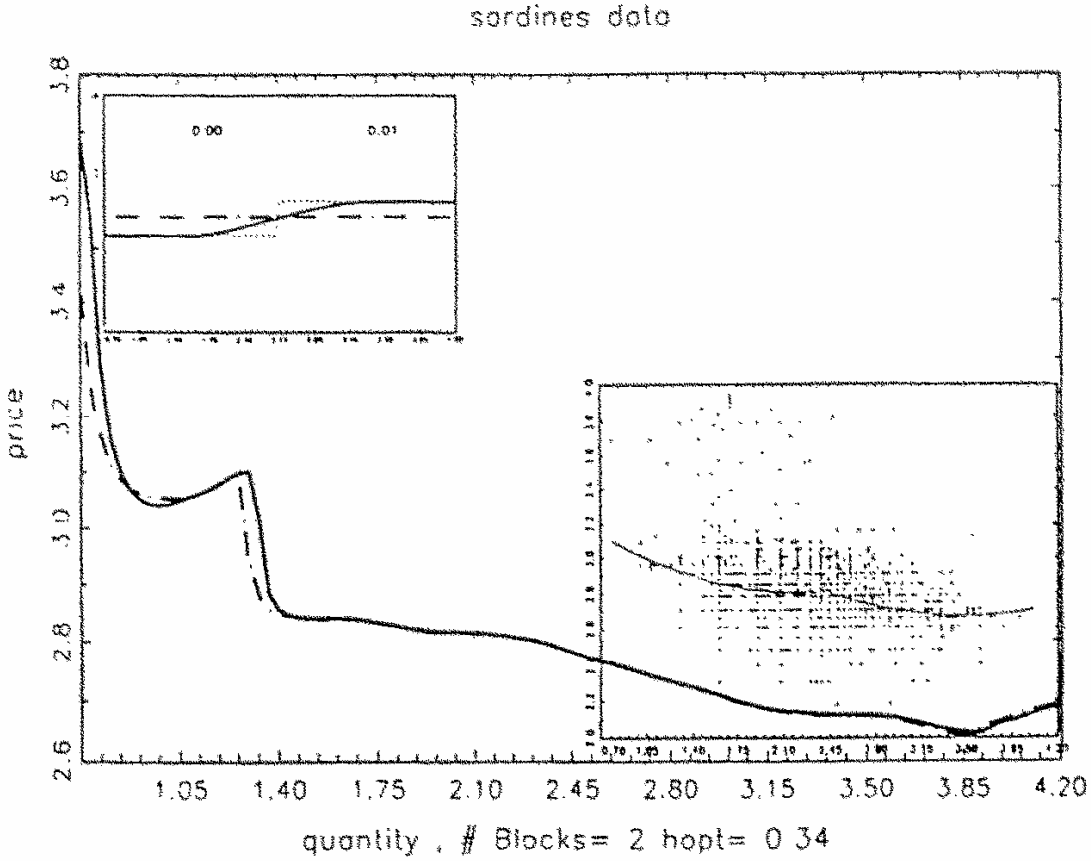


Fig. 4a

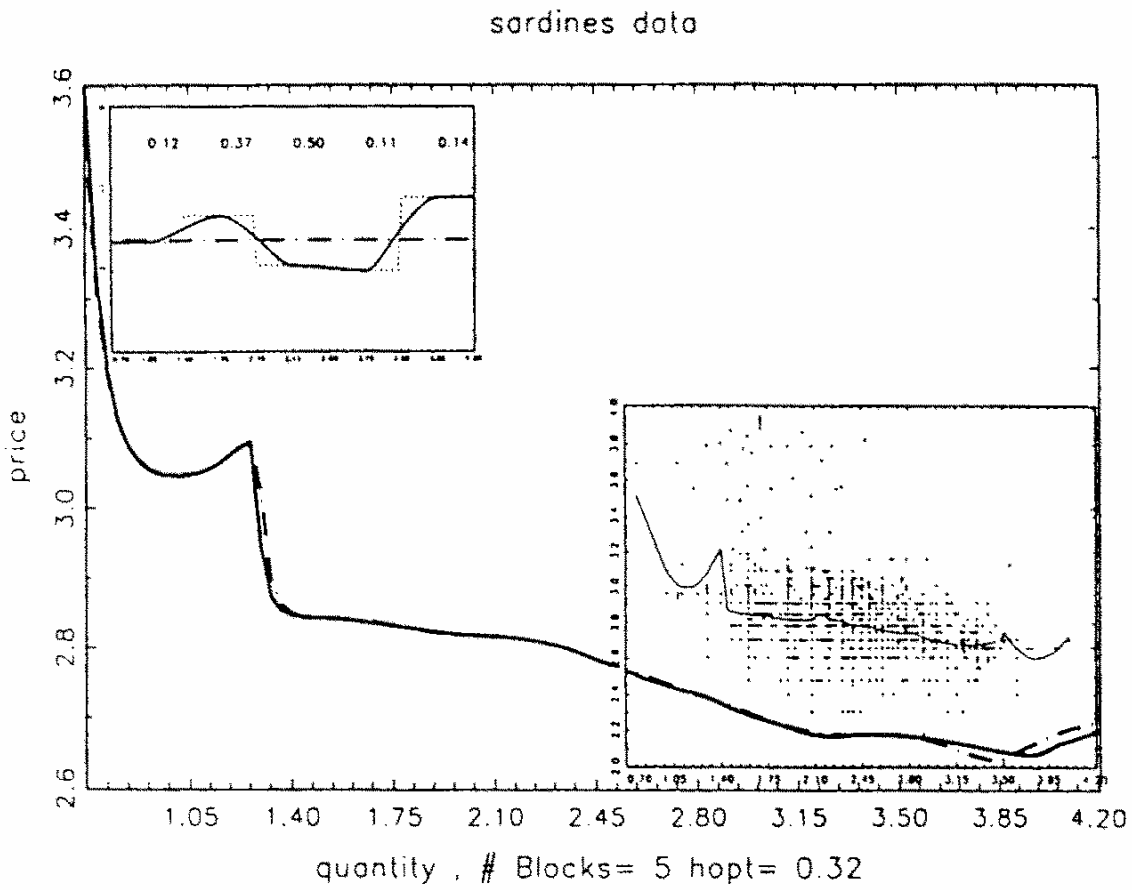


Fig. 4b

characteristic over higher prices. The reason for this is the cluster of points near the origin. One explanation that could be advanced for this is the existence of many transactions at low prices corresponding to the clearing of stocks at the end of the day. To check for this we isolated data for late transactions, and as Table 2 indicates, we found examples in which the prices of late transactions fall sharply.

This result must be treated with some caution since we also find examples where prices fall, only to rise again at the end of the day – a situation corresponding to that when a buyer with inelastic demand wished to add to his purchases before the end of the market.

Nevertheless the important thing to realise here is that the use of nonparametric methods enabled us to identify this particular feature. Trying various nonlinear but *parametric* forms gave excellent fits with monotone decreasing functions. Thus we would not have detected the presence of this group of 'less well-behaved' observations. Once the offending observations were removed we obtained a monotone decreasing nonparametric curve.

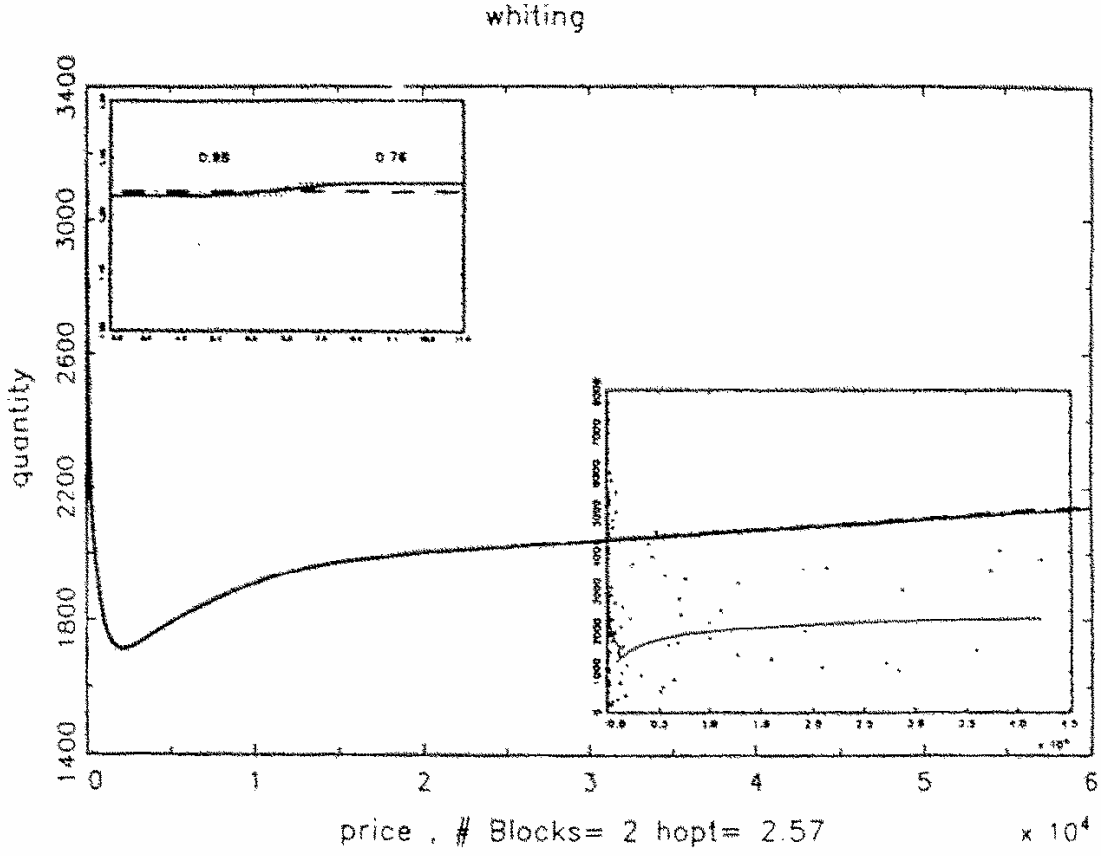


Fig. 4c

Table 2
Late transactions, in chronological order (price dropped if transaction is in last 10% of those made by seller in the day)

Date	Seller no.	Price	Quantity
87/07/04	40		
Normal prices		82	60
		85	60
		75	100
		75	50
		85	60
		75	50
		85	60
		85	70
Dropped price		25	85
87/07/08			
Normal prices	28	82	50
		82	20
Dropped price		20	64

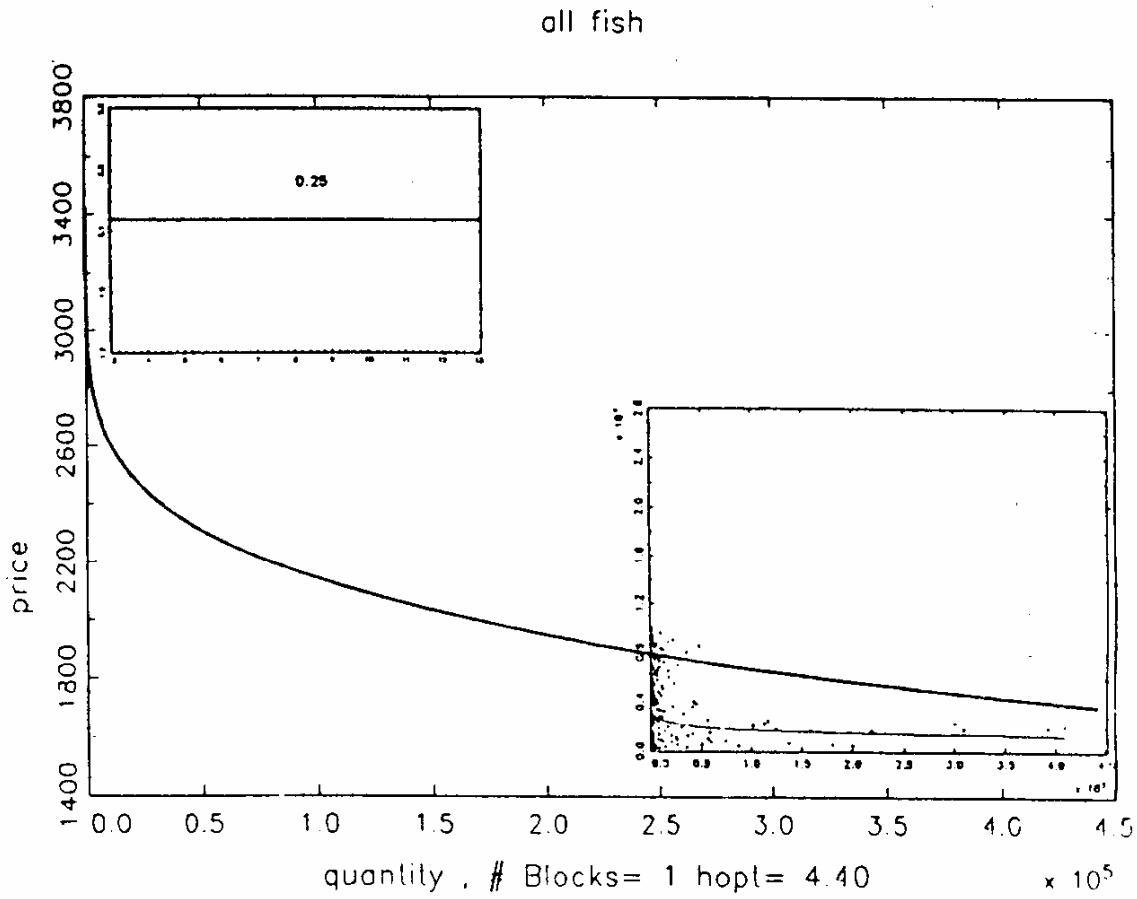


Fig. 5a

Finally we look at all fish with quantities added as if they were one commodity, and the data and resultant curves are shown in Figs. 5a and 5b. Although, of course, this simple addition is extremely primitive as a procedure, the resultant smoothed demand curves are monotone. Weighting the quantities more appropriately would not have modified this. Thus aggregating, even in an arbitrary way, reinforces a characteristic which is weaker at a less aggregated level and even absent for many observations at the micro level. The market is, therefore, 'well-behaved', even though many of the individuals participating in it are not.

6. Conclusion

In this paper we have examined detailed data from the fish market for Marseille. We built a theoretical model for the behaviour in this market. The perishable nature of the product enables us to think of successive markets as

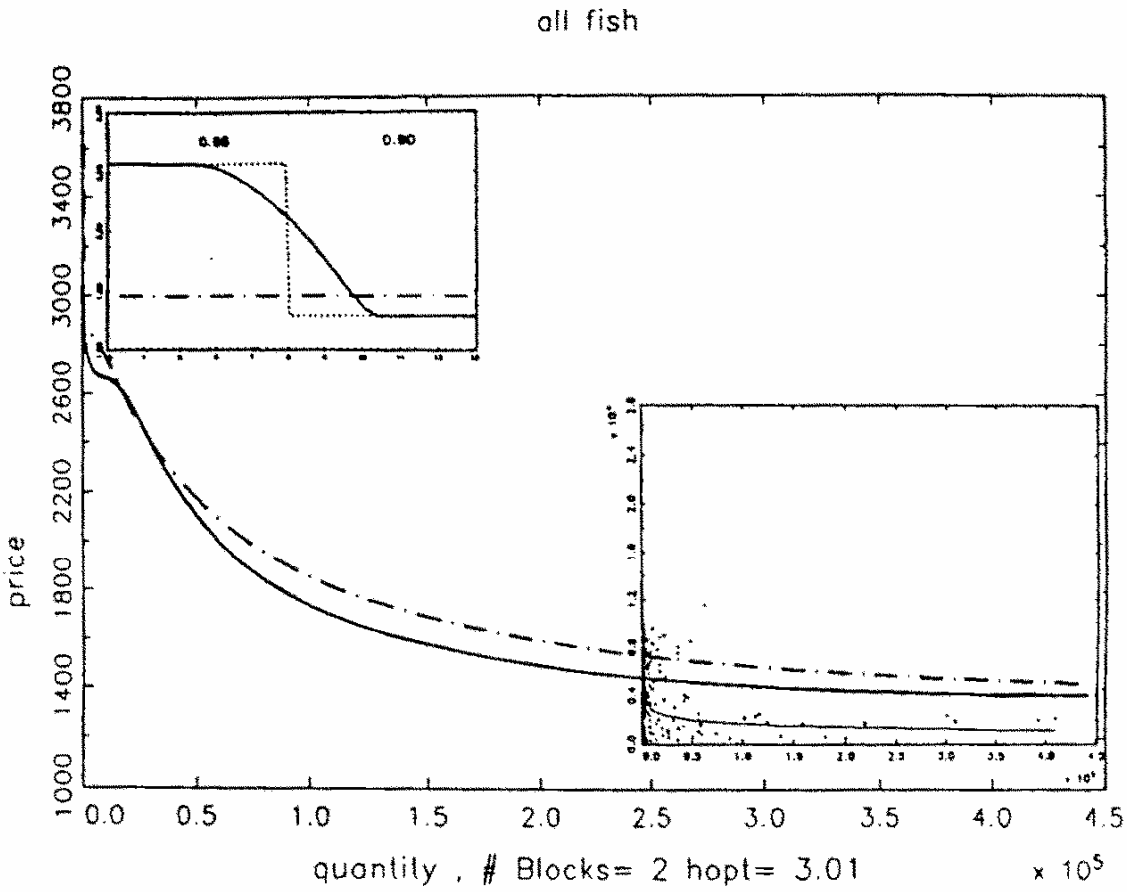


Fig. 5b

being separated and to analyse the short-run behaviour of the relation between prices and quantities in such a market. Although it is tempting to look at these data as resulting from the interaction between competitive supply and demand, the organization of the market and the identity of the participants makes this unreasonable. Individual transactions show none of the characteristics that standard demand analysis would lead one to expect. Fitting seasonalised average price over a day to quantities sold on that day did at the *aggregate level* give rise to a monotone decreasing relation. Thus, this property does not reflect individual behaviour but rather results from aggregation. The use of nonparametric methods makes this finding particularly striking.

We then turned to looking at the data in the light of our models. The price distributions, the appropriate equilibrium notion were stable over time, allowing us to think of the market game as being repeated over time. We then analysed nonparametrically the transactions at each price and were able to identify a particular feature of the data, many small transactions at low prices at the end of trading which destroyed the monotone character of the relation for

one fish. However, once this problem was identified, trading for the earlier part of the day did have a monotone price–quantity relation.

Although this analysis now needs to be generalised to all types of fish over longer periods, two aspects are important. Firstly, the underlying behaviour should not just be taken to reflect a standard competitive market. Doing so may lead to false inferences on the aggregate level about individual behaviour. Aggregate phenomena in this sort of market are not simply the magnified reflection of micro phenomena. Testing aggregate data for properties derived from the theory of individual behaviour is not an appropriate procedure. Secondly, once this is taken into account, it is interesting to observe that using nonparametric methods to fit a different price–quantity relation did allow us to identify features of the special behavioural structure of this market, at the micro level.

Appendix A

A.1. The local smoothing method

Our smoothing method assumes that the response variables $\{Y_i\}_{i=1}^n$ are of the form:

$$Y_i = m(X_i) + \varepsilon_i, \qquad i = 1, \ldots, n,$$

with explanatory variables X_i, independent errors $\{\varepsilon_i\}_{i=1}^n$, and the smooth regression function $m(x)$. We are interested in estimating the function m. The kernel smoother is defined by

$$\hat{m}_h(x) = n^{-1} \sum_{i=1}^{n} K_h(x - X_i) Y_i \Big/ n^{-1} \sum_{i=1}^{n} K_h(x - X_i), \tag{A.1}$$

where $K_h(\cdot) = h^{-1}K(\cdot/h)$ is the rescaled kernel function K with bandwidth h. Behaviour of this smoother is crucially dependent on the choice of h. A simple and useful quantification of the influence of h is the analysis of the asymptotic mean integrated squared error. The variance of the kernel smoother $\hat{m}_h(x)$ is approximated by

$$n^{-1}h^{-1}V(x) = n^{-1}h^{-1}\int K^2(u)\,du\,\frac{\sigma^2(x)}{f(x)}, \tag{A.2}$$

where $\sigma^2(x)$ denotes the variance function $E(Y^2|x) - m^2(x)$ and $f(x)$ is the marginal density of the X variables. The bias is approximated by

$$h^2B(x) = \frac{h^2}{2}\int u^2K(u)\left[m''(x) + 2\,\frac{m'(x)f'(x)}{f(x)}\right]. \tag{A.3}$$

This indicates that the best bandwidth at each location x is well represented by

$$h_0(x) = n^{-1/5}\left[\frac{V(x)}{4B^2(x)}\right]^{1/5}.$$

A good global bandwidth, i.e., one suitable in an average sense, is given by

$$h_1(x) = n^{-1/5}\left[\frac{\int V(x)}{4\int B^2(x)}\right]^{1/5}. \tag{A.4}$$

This bandwidth is obtained by minimising the approximate integrated mean squared error, $n^{-1}h^{-1}\int V(x)\mathrm{d}x + h^4\int B^2(x)\mathrm{d}x$.

A.2. A global smoothing parameter

Practical use of the above representation for h_1 in formula (A.4) requires estimates of $\int V(x)$ and $\int B^2(x)$, which in turn can be built up, using formulas (A.2) and (A.3), from estimates of $m(x)$ and $f(x)$. We shall use simple estimates like histograms for this.

Histograms are constructed by first partitioning the design interval $[a, b]$ into blocks B_j, $j = 1, \ldots, N$. For simplicity, we work explicitly here with equal length intervals

$$B_j = \left[a + \frac{(j-1)(b-a)}{N}, a + \frac{j(b-a)}{N}\right].$$

Let B denote a generic block B_j, and r and l denote right and left boundaries of this block. The proportion of X_i falling in each interval reflects the height of the density near the centre of the block (bin). Let $c = (r + l)/2$ denote the block centre and $r_b = (r - l)/2$ denote the block radius. The histogram density estimate is

$$\hat{f}(c) = \frac{1}{2nr_b}\sum_{i=1}^{n} I(|c - X_i| \leqslant r_b). \tag{A.5}$$

To estimate the derivative of f on B we use a simple differencing method. Define

$$n_l = \sum_{i=1}^{n} I(l \leqslant X_i < c), \qquad n_r = \sum_{i=1}^{n} I(c \leqslant X_i < r).$$

If these frequency estimates are combined we obtain the score function estimate

$$(\hat{f}'/\hat{f})(c) = 2(n_r - n_l)/r_b(n_r + n_l). \tag{A.6}$$

The estimation of $V(x)$ in (A.4) is constructed from a sum of squared residuals (RSS) about an estimate $\hat{m}(x)$ of $m(x)$, normalised by an estimate of f. In particular, in the generic block B define $RSS = \sum_{X_i \in B}(Y_i - \hat{m}(X_i))^2$. An

estimate of (σ^2/f) is then given by

$$\widehat{(\hat{\sigma}^2/f)}(c) = \frac{2nr_b}{n_l + n_r} RSS,$$

which leads to

$$\hat{V}(c) = \tfrac{5}{4}\widehat{(\sigma^2/f)}(c), \tag{A.7}$$

for the quartic kernel,

$$K(u) = \tfrac{15}{16}(1 - u^2)^2 I(|u| \leqslant 1),$$

which we used throughout.

For estimation of m', m'' we use blockwise parabolic fitting. As an estimate of $V = \int V(x)$, we obtain

$$\hat{V} = \sum_{j=1}^{N} 2r_b \hat{V}(c_j), \tag{A.8}$$

using notation from (A.7) and the quartic kernel given before. The bias $B(c_j)$ for the quartic kernel in each block B is estimated via

$$\hat{B}(c_j) = \tfrac{1}{2}\tfrac{1}{4} 2\hat{\beta}_{3j} + 2[2\beta_{3j}(c_j) + \hat{\beta}_{2j}](\hat{f}'/f)(c_j), \tag{A.9}$$

where $\hat{\beta}_{2j}$ and β_{3j} are least squares estimates in the model

$$\beta_{1j} + \beta_{2j}x + \beta_{3j}x^2 \tag{A.10}$$

over block β_j. The final estimate of $B_2 = \int B^2(x)\,dx$ is obtained by summing up the squares of these quantities,

$$\hat{B}_2 = \sum_{j=1}^{N} 2r_b \hat{B}^2(c_j). \tag{A.11}$$

This leads finally to

$$\hat{h}_1 = n^{-1/5}[\hat{V}/4\hat{B}_2]^{1/5}.$$

A.3. Blocks for a local smoothing parameter function

The estimates of variance and squared bias over each block provide an easy bandwidth choice for each block, given by

$$\hat{h}_0(c) = n^{-1/5}[\hat{V}(c)/4\hat{B}^2(c)]^{1/5}, \tag{A.12}$$

using the notation from (A.7) and (A.10). The logs (base 2) of these values are the heights of the dotted step function in the upper left inset of Fig. 4a. The bandwidth $\hat{h}_1$ is represented by the dotted and dashed constant function in this inset bandwidth plot. This provides a useful reference for understanding the relative sizes of the local bandwidths. The average of the $\hat{h}_0(c_j)$ does *not* give the

global bandwidth $\hat{h}_1$ because the variance and bias terms need to be summed separately for the latter!

The local bandwidth estimates are best in the centre of the blocks. For the points away from the centres we use a smooth, represented by the solid curve, of the step function. This smooth is computed on a fixed grid of x's by the formula (A.1) with the X_i replaced by the *bin* centres and the Y_i replaced by the height of the step function. The kernel used is the quartic, and the bandwidth is the block radius r_b. This choice of bandwidth guarantees that the smooth coincides with the step function at the points where it is most accurate, i.e., at the *bin* centres.

A.4. Diagnostic plots

The inset plots in Fig. 4a are intended to show visually how well our methods are performing. For example, Fig. 4a shows in the left upper corner the fact that a bigger bandwidth has to be chosen for larger quantities.

The raw data plot in the lower right was also a useful diagnostic in the choice of $N = 2$ blocks for the sardine data example presented in Fig. 4a. Our initial choice of $N = 5$ blocks gave visually poor performance because too many 'corners' appeared, as can be seen in Fig. 4b. This gave a visually poor parabolic fit, which resulted in an oversmoothed global choice $\hat{h}_1$ and a less effective local bandwidth function $\hat{h}_0(x)$.

The bandwidth plots enhance the understanding of the performance of the local smoothing method by showing the amount of smoothing done at each point. The effective bandwidth is shown on the log scale, because this parameter is multiplicative in character.

A further diagnostic device, which we find useful, is to calculate the observed significance level of the parabolic fit on each block. The numbers shown in the top part of the bandwidth plot are p-values for testing, within each block, the null hypothesis of linearity,

$$H_0: \quad \beta_{3,j} = 0,$$

in the local parabolic model. When these are small, there is strong evidence of curvature in the data, so our local bandwidth estimate should be reliable. Note that in most of our examples the local method works well in many cases, even when the p-value is large.

Appendix B

Stability of price distributions

Recall that the distribution of prices is given by

$$f(p) = \frac{\sum \textit{ quantities sold at prices in the jth interval}}{\textit{Total quantities sold}}.$$

Let us denote the pairs $\{X_i, Y_i\}_{i=1}$ as observations of X_i = price of the transactions and Y_i = quantity of ith transactions. We shall assume that the data $\{X_i, Y_i\}_{i=1}$ are independent and identically distributed observations.

This can be justified in several ways. Firstly, not all players are present all the time (randomness). Second, buyers do not stick to 'their' seller (independence) and buy different quantities on different occasions (identical distribution). Third, sellers do not stick to 'their' prices, as empirical analysis of late transactions shows. We are aware of the fact that this is an important assumption, but nevertheless pose it in order to be able to analyse the stability of the distribution.

The price distribution at $p = x$ can be re-written as

$$n^{-1}h^{-1}\tfrac{1}{2}\sum_{i=1}^{n} Y_i I(|x - X_i| \leqslant h) \Big/ n^{-1}\sum_{i=1}^{n} Y_i,$$

when we have essentially rescaled by $2h$, the length of the interval over which we are computing the price distribution. This information is essentially a kernel estimator.

$$\hat{r}_h(x)/\bar{y} = n^{-1}\sum_{i=1}^{n} K_h(x - X_i)Y_i/\bar{y},$$

for kernel $K(u) = \frac{1}{2}I(|u| \leqslant 1)$, $K_h(\cdot) = h^{-1}K(\cdot/h)$.

A kernel is a symmetric probability density. Examples are

$K(u) = \frac{3}{4}(1 - u^2)I(|u| \leqslant 1)$ Epanechenkov,

$K(u) = \frac{15}{16}(1 - u^2)I(|u| \leqslant 1)$ Quartic,

$K(u) = (1/\sqrt{2u})\exp(-u^2, 2)$ Gaussian.

The parameter h controlling the 'window' over which we are averaging is called bandwidth (see Härdle, 1990, Ch. 3). What is $\hat{r}_h(x)$ estimating? Denote by $m(x) = \mathrm{E}(y \mid Y = x)$ and (in contrast to earlier notation) $f(x)$ the density of the prices of all transactions.

Thus, following well-known arguments of Härdle (1990),

$$\mathrm{E}\hat{r}_h(x) = \int K_h(x - u)m(u)f(u)\,\mathrm{d}u$$

$$= m(x)f(x) + \frac{h^2}{2}\int u^2K(u)\mathrm{d}u(mf)''(x) + o(h^2) \quad \text{as} \quad h \to 0.$$

The bias is thus of order $O(h^2)$. The variance of $\hat{r}_h(x)$ is given by

$$\mathrm{var}[\hat{r}_h(x)] = n^{-1}\mathrm{var}[K_h(x - X)Y]$$

$$= n^{-1}h^{-1}\mathrm{E}(Y^2 \mid X = x)f(x)\int K^2(u)\,\mathrm{d}u + O(n^{-1}h^{-1})$$

$$\text{as} \quad nh \to \infty.$$

Hence the mean squared error can be written with constraints C_1 and C_2 as

$$MSE(x) = u^{-1}h^{-1}C_1 + h^4C_4.$$

The optimum for h is reached when $h \sim n^{-1/5}$, yielding a spread of $MSE(x) \sim n^{-4/5}$. Therefore positive intervals have length (and 'shrinking rate') $n^{-2/5}$. More precisely

$$\sqrt{nh}\left(\frac{\hat{r}_h(x)}{\bar{y}} - \frac{r(x)}{\mu}\right) \xrightarrow{L} N(0, M_2(x) f(x) \int K^2),$$

for $nh5 \to 0$, $\mu = E(Y)$, $r = mf$.

Of course, pointwise confirmation intervals do not help us in testing the stability of the price distribution, we need uniform confirmation bounds. They can be developed as follows. With

$$\hat{r}_h(x) - r(x) = \iint K_h(x - u)v \, d(F_n - F)(u, v),$$

when F_n diverts the empirical distribution function of the data $\{(X_i, Y_i)\}_{i=1}^n$ and F their theoretical distribution. Suppose the data has been rescaled so that $X \in [0, 1]$. The process $\sqrt{n}(F_n - F)$ can be approximated by Brownian bridges. so that asymptotically as the number of transactions becomes very large,

$$\sqrt{nh}(\hat{r}_h(x) - r(x)) \approx [M_2(x) f(x)]^{1/2} h^{-1/2} \int K\left(\frac{x - u}{h}\right) dW(u),$$

for a Wiener process W and $M_2(x)$: $E(Y^2 | X = x)$.

The desired uniform confidence bound can be constructed from the following statement:

$$P\{(2\delta \log n)^{1/2}[(nh/\int K^2)^{1/2} \sup(\hat{M}_2(x) \hat{f}_h(x))^{-1/2} | - d_n] < z\}$$
$$\to \exp(-2\exp(-z)) \quad \text{as} \quad n \to \infty.$$

Here,

$$d_n = (2\delta \log n)^{1/2} + (1/2\delta \log n)^{1/2}\{\log(C_3/2\pi)\},$$
$$h = n^{-\delta}, \qquad \delta > \tfrac{1}{2},$$
$$C_3 = \int [K'(u)]^2 du / 2\int [K(u)]^2 du.$$

For an algorithm for these confidence bounds see Härdle (1990).

References

Barten, A.P. and L.J. Bettendorf, 1989, Price formation of fish: An application of an inverse demand system, European Economic Review 33, 1509–1525.

Becker, G.S., 1962, Irrational behavior and economic theory, Journal of Political Economy 70, 1–13.

Benabou, R., 1988, Search, price setting and inflation, Review of Economic Studies 55, 353–376.

Butters, G.R., 1977, Price distributions of sales and advertising prices, Review of Economic Studies 44, 465–491.

Debreu, G., 1974, Excess demand functions, Journal of Mathematical Economics 1, 15–23.
Diamond, P., 1987, Consumer differences and prices in a search model, Quarterly Journal of Economics 102, 429–436.
Ekelund, R.B. Jr. and S. Thommesen, 1989, Disequilibrium theory and Thornton's assault on the laws of supply and demand, History of Political Economy 21, 567–592.
Erdös, P. and J. Spencer, 1974, Probabilistic methods in combinatorics (Academic Press, New York, NY).
Gode, D.K. and S. Sunder, 1993, Allocative efficiency of markets with zero-intelligence traders: Markets as a partial substitute for individual rationality, Journal of Political Economy 101, 119– 137.
Gorman, W.M., 1959, The demand for fish: An application of factor analysis, Research paper no. 6, Series A (Faculty of Commerce and Social Science, University of Birmingham); Abstracted in: Econometrica 28, 649–650.
Grandmont, J.-M., 1983, Money and value, Econometric Society monographs in pure theory (Cambridge University Press, Cambridge, and Editions de la Maison des Sciences de l'Homme, Paris).
Härdle, W., 1990, Applied nonparametric regression, Econometric Society monograph series 19 (Cambridge University Press, Cambridge).
Hildenbrand, W., 1983, On the law of demand, Econometrica 51, 997–1019.
Kirman, A.P., 1992, What or whom does the representative individual represent?, Journal of Economic Perspectives 6, 117–136.
Kirman, A.P. and M. McCarthy, 1990, Equilibrium prices and market structure: The Marseille fish market, Paper presented at the 1990 congress of the Royal Economic Society.
Kirman, A.P. and A. Vignes, 1991, Price dispersion: Theoretical considerations and empirical evidence from the Marseille fish market, in: K.J. Arrow, ed., Issues in contemporary economics (Macmillan, London).
Kormendi, R.C., 1979, Dispersed transactions prices in a model of decentralised pure exchange, in: S.A. Lippman and J. McCall, eds., Studies in the economics of search (North-Holland, Amsterdam).
Lewbel, A., 1989, Exact aggregation and a representative consumer, Quarterly Journal of Economics 104, 622–633.
Mill, J.S., 1869, Thornton on labour and its claims, in: Collected works, 1967, Essays on economics and society (Toronto University Press, Toronto) 631–668.
Mill, J.S., 1871, Principles of political economy, edited by W.J. Ashley (New York, NY).
Mill, J.S., 1972, Later letters of John Stuart Mill, 1849–1873, in: Collected works, Vols. 14–17, (Toronto).
Negishi, T., 1985, Non-Walrasian foundations of macroeconomics, in: G.R. Feiwel, ed., Issues in contemporary macroeconomics and distribution (London).
Negishi, T., 1986, Thornton's criticism of equilibrium theory and Mill, History of Political Economy 18, 567–577.
Negishi, T., 1989, On equilibrium and disequilibrium – A reply to Ekelund and Thommesen, History of Political Economy 21, 593–600.
Phlips, L., 1988, The economics of imperfect information (Cambridge University Press, New York, NY).
Robbins, L., 1935, An essay on the nature and significance of economic science (Macmillan, London).
Roth, A.K., J.K. Murnighan, and F. Schoumaker, 1988, The deadline effect in bargaining: Some experimental evidence, American Economic Review 78, 806–823.
Rothschild, M., 1973, Models of market organisation with imperfect information: A survey, Journal of Political Economy 81, 1283–1301.

Salop, S.C. and J.E. Stiglitz, 1982, The theory of sales: A simple model of equilibrium price dispersion with identical agents, American Economic Review 72, 1121–1130.

Sonnenschein, H., 1972, Market excess demand functions, Econometrica 40, 549–563.

Summers, L.H., 1991, The scientific illusion in empirical macroeconomics, Scandinavian Journal of Economics 93, 129–148.

Theil, H., 1971, Principles of econometrics (Wiley, New York, NY).

Thornton, W.H., 1870, On labour: Its wrongful claims and rightful dues, its actual present and possible future, 2nd ed. (London).

Varian, H.R., 1980, A model of sales, American Economic Review 76, 651–659.

Working, E.J., 1927, What do statistical 'demand curves' show?, Quarterly Journal of Economics, 212–235.

Alg 665 Applied Statistics, 44, No.2, 258-268.

OPTIMAL MEDIAN SMOOTHING*

W. Härdle
Humboldt-Universität Berlin, Wirtschaftswissenschaftliche Fakultät
Institut für Statistik und Ökonometrie
Spandauer Str. 1, D - 10178 Berlin

W. Steiger
Department of Computer Science
Rutgers University
New Brunswick, N.J. 08903, U.S.A.

May 1994

Abstract

Median smoothing of a series of data values is considered. Naive programming of such an algorithm would result in large amount of computation, especially when the series of data values is long. By maintaining a heap structure that we update when moving along the data we obtain an optimal median smoothing algorithm.

* This research was supported by the Deutschen Forschungsgemeinschaft. Sonderforschungsbereich 303 and 373 and by C.O.R.E. and Institute de Statistique, Université Catholique de Louvain, Belgium.

Language: ISO Pascal

Keywords: Smoothing, Running Medians, Heaps.

Description and Purpose

High variability in a given series $X_1, \ldots, X_N$ may obscure important structural features which would become evident if the data were replaced by a smoother, less variable version, $X_1^*, \ldots, X_N^*$. One smoothing strategy consists of replacing X_i by

$$X_i^* = median(X_{i-k}, \ldots, X_{i+k}), i = k+1, \ldots, N-k \quad (1)$$

which, due to the robustness of the median, will often give a superior result to

$$average(X_{i-k}, \ldots, X_{i+k}).$$

It is common terminology to call the series X_i^* the *running median* of order $K = 2k+1$ and we refer to $X_{i-k}, \ldots, X_{i+k}$ as the *window of size* K *around* X_i. In this note we describe an efficient running median algorithm using the HEAP data structure and we mention an interesting recent lower bound which shows that the algorithm has, up to constants, optimal running time.

An obvious approach (METHOD 1) to the computation of X_i^* could use the fast median algorithm of Schönhage, Patterson and Pippenger (1976) with which each X_i^* could be obtained in at most $3K$ steps. As is usual, we count each pairwise comparison as a STEP and argue that the running time of any "reasonable" implementation would be proportional to this complexity measure. The smoothed series would then be obtained in at most $3K(N-K)$ steps. By this method we obtained $N-K$ running medians at an average cost of $3K$.

Another possibility (METHOD 2) would be to maintain the window

$$X_{i-k}, \ldots, X_{i+k}$$

in sorted order as

$$Z_1 \leq \ldots \leq Z_K.$$

In this case $X_i^* = Z_{k+1}$ and the Z's may be prepared for the determination of X_{i+1}^* by locating, then removing X_{i-k}, and then correctly inserting X_{i+k+1} so the Z's are once again sorted. Each operation may be done via a binary search requiring at most $log\ K$ steps (all logarithms are base 2). Although it is possible to implement this strategy so it actually runs in time $O(log\ K)$ (this means $\leq b\ log\ K$ for all K, b an absolute constant), it is quite difficult to do so and in any case, the constant b is quite large. Easily implemented algorithms to maintain a sorted

window use pointers into the Z's. We expect $K/2$ pointer values to be changed per update, on the average, and this would certainly contribute to the running time. While an update really takes only $2\,log\,K$ comparison steps, the hidden cost of pointer updates would force the actual running time of this algorithm to be

$$(2\,log\,K + cK/2)(N - K)$$

c a constant reflecting the time for a pointer manipulation in relation to that of a comparison.

Our proposal (METHOD 3) maintains $X_{i-k}, \ldots, X_{i+k}$ in a priority queue that supports easy and quick updates. We use a data structure based on two heaps which we call partitioning heaps. It may be updated in time proportional to $log\,K$. The array $Z_1, \ldots, Z_m$ is called a (max) heap if it is partially ordered so as to satisfy

$$Z_i \geq max(Z_{2i}, Z_{2i+1}); \tag{2}$$

a min (heap) reverses the inequality. A convenient reference for heaps and other data structures is the text of Sara Baase (1988) or the one of Mehlhorn (1984). There it is shown that a heap of size m may be constructed in at most $4m$ steps and that a new item may be inserted so as to preserve the heap property in at most $log\,m$ steps. It is clear that after an item is deleted, the heap property can be restored in at most $2\,log\,m$ steps.

The complexity of the algorithm that we describe in the next section is no more than

$$4K + (3\,log(K/2))(N - K), \tag{3}$$

steps. More importantly, the implementation uses at most $4\,log(K/2)$ pointer manipulations per update, each using one multiplication or division. This gives a total running time bounded by

$$4K + (3 + 4d)\,log(K/2)(N - K), \tag{4}$$

d a constant reflecting the time for a division relative to that of a comparison. If, as seems reasonable, the window size increases with N, this method becomes infinitely more efficient than the two that were mentioned previously.

It is interesting to consider the behaviour of an optimal algorithm for running medians. This would give a standard against which all algorithms should

be compared and has practical as well as theoretical significance. Though it seems likely that the average cost of determining X_i^* would grow with K, the number of X's determining each median, the only obvious statement is the trivial lower bound of $2(N-K)$ steps for smoothing by running medians which is established as follows: If $X_1, \ldots, X_N$ is partitioned into N/K non-overlapping segments of length K each, it is necessary to perform $2K$ steps to obtain each of $X^*_{k+1}, X^*_{K+k+1}, \ldots, X^*_{(N/K-1)K+k+1}$, via a recent result of Bent and John (1985). This gives a total cost of $2K(N/K)$, or an average update cost of 2 steps per median. By (3), our method has an average cost of

$$3\,log(K/2) + (K-1)/(N-K)$$

steps per update. A recent lower bound of Gill, Steiger and Wigderson (1988) shows that at least $(a\,log\,K)(N-K)$ steps must be performed by any algorithm that correctly computes $X^*_{k+1}, \ldots, X^*_{N-K}$, $a > 0$ an absolute constant. This means that our algorithm is optimal at least up to the constant of proportionality in (4). It is hard to imagine a realizable algorithm that does less work than $2\,log(K/2)$ steps per update.

Finally, it seems worth mentioning some recent active interest in data structures aimed at implementing priority queues for specialized types of insertions and deletions (see Aktinson, Sack, Santoro and Strothotte (1986) or Carlsson (1987)). Neither of these techniques seems as well suited to compute the running median as the partioning heaps we will describe in the next section.

Numerical Method

We use two heaps of size k to represent the current window

$$X_{i-k}, \ldots X_{i+k}, \quad i = k+1, \ldots, N-k.$$

The window is stored in the array

$$H_{-k}, \ldots, H_{-1}, H_0, H_1, \ldots, H_k$$

that preserves the following structure:

	(i)	H_0 is X_i^*
(*)	(ii)	$max(H_{-2i}, H_{-2i-1}) \leq H_{-i} \leq X_i^*$

(iii) $$min(H_{2i+1}, H_{2i}) \geq H_i \geq X_i^*.$$

These conditions mean that X_i^* partitions H into a (max) heap of those window elements $H_{-1}, \ldots, H_{-k}$ not larger than X_i^* and a (min) heap

$$H_1, \ldots, H_k$$

of those window elements not smaller than X_i^*. The data structure

$$H_{-k}, \ldots, H_{-1}, H_0, H_1, \ldots, H_k$$

may be visualized as follows:

Figure 1.

The ordering between levels obeys (ii) and (iii). No order is imposed on values within the same level. The structure has a depth at most $2(1 + log\, k)$.

To update this structure so it will represent the next window, we need to find X_{i-k} and remove it, correctly insert X_{i+k+1}, and restore the structure so (*) still holds. Our method involves pointers $i_{-k}, \ldots, i_k$ and $j_{-k}, \ldots, j_k$, both permutations of $-k, \ldots -1, 0, 1, \ldots k$. They have the property that

$$H_{i_m} = X_{i+m}, m = -k, \ldots, k; \tag{5}$$
$$X_{i+j_m} = H_m, m = -k, \ldots, k.$$

The pointers i_m locate a particular window element in the heap while the j_m locate a particular heap element in the window. Clearly

$$X_{i+j_{i_m}} = X_{i+m}, \tag{6}$$
$$H_{i_{j_m}} = H_m.$$

When X_{i-k} is removed from the window the data structure has an empty place at i_{-k}. To update we propagate the "hole" to the apex of the relevant heap. For example if $i_{-k} < -1$ the steps

$$j_{i_{-k}} \leftarrow j_{i_{-k}}/2, \tag{7}$$
$$i_{-k} \leftarrow i_{-k}/2$$

swap the hole at i_{-k} with the value of the parent node, $H_{i_{-k}/2}$ and correctly adjust the relevant pointers. The case $i_{-k} > 1$ is analogous. At most $log(k)$ steps are performed, each involving a division of the pointer value by 2.

Inserting the new value X_{i+k+1} into the data structure is analogous: A comparison of X_{i+k+1} with X_i^* determines if the new value goes in the top or bottom heap. In the former case, for example, assuming X_{i-k} came from the bottom heap, X_{i+k+1} it is initially placed in the hole just created at position -1; $i_k \leftarrow -1$ and $j_{-1} \leftarrow k$ reflect this operation. H_{-1} is compared with H_{-2} and H_{-3} and if (ii) in (*) holds, the insert is finished. Otherwise the larger of H_{-2}, H_{-3} is swapped with H_{-1}, ets., until the new item trickles down to its correct place and (*) is once again satisfied. At most $2\,log(k)$ comparisons are required along with $2log(k)$ pointer updates. Thus, the total cost of maintaining the structure in H while moving from the current window to the next one is at most $2log(k)$ comparisons and $4log(k)$ pointer updates.

The running times reported in the next section show that the average cost to compute for each X_i^* is proportional to $6\,log(k)$. Maintaining pointer values as permutations of $-k, \ldots, k$ facilitates quick updates because it is very well suited to the data structure we use: The parent node of node j is $j/2$ and the children are $2j$ and $2j + 1$. Finally we observe that roundoff error and stability is not an issue because the algorithm does not perform floating point arithmetic.

Structure

Required global declarations

```
Constant
  maxk                                          maximum window size (e.g.60)
  maxn                                          maximum number of
                                                observations (e.g.1023)

Type
  keytype     real                              type of value
  elem        RECORD
              key :  keytype
              END
  nelemarray  ARRAY [1..maxn] of elem           array of all observations
  kelemarray  ARRAY [-maxk..maxk] of elem       array of observations in window
  outlistype  ARRAY [1..maxn] of integer        array of pointer indices
  nrlistype   ARRAY [-maxk..maxk] of integer    array of pointer indices

  medianarray ARRAY [1..maxn] of elem.          array of medians

PROCEDURE runmed (n, k :  integer;
VAR unsorted :  nelemarray; VAR median :  medianarray);
```

Formal parameters

n	integer	Value:	the real size of the array <=maxn
k	integerthe	Value:	real size of the window <= maxk
unsorted	nelemarray	Value:	only var because of storage
median	medianarray	Value:	returns the running median

Time

To check the approximations of the expression in (4) we smoothed series of $n = 16000$ using windows of $K = 7, 15, 31, 63, 127, 255, 511, 1023, 2047, 4095$. The X_i were generated to be uniformly distributed (0,1) random numbers using the internal random mechanism of an ATARI home computer. For each value of K the smoothing experiment was repeated ten times with independently generated series $X_1, ..., X_N$. The running times were then averaged over the ten replications. As a basis of comparison to the proposed optimal median smoothing algorithm we also smoothed using a simplification of METHOD 2 where $X_i^* = W_{k+1}$ is simply extracted from the sorted window $W_1, ..., W_k$ and then X_{i+k+1} is correctly inserted by a sequential search costing $O(K)$ steps on the average.

Table 1 shows the timings for the optimal running median algorithm in three situations. In the first column of Table 1 we report the timings for smoothing a series of values that were initially in descending order. The second column gives the timing on a series that was initially increasing. The last column of Table 1 shows the timings averaged over ten repetitions of smoothing pure random sequences.

k	*Descending*	*Ascending*	*Random*
7	11.2	10.8	6.8
15	15.6	15.3	9.2
31	20.3	19.8	11.4
63	25.1	24.2	13.8
127	29.1	28.6	16.2
255	33.3	32.8	18.4
511	37.1	36.6	20.4
1023	41.2	39.7	22.0
2047	41.3	40.9	22.4
4095	38.6	38.2	21.2

Table 1. Timings (in sec) for Optimal Median Smoothing

k	*Descending*	*Ascending*	*Random*
7	0 : 11.2	0 : 08.4	0 : 08.3
15	0 : 22.7	0 : 16.3	0 : 11.8
31	0 : 45.6	0 : 33.2	0 : 16.7
63	1 : 31.4	1 : 06.1	0 : 36.9
127	3 : 02.4	2 : 11.6	1 : 14.9
255	6 : 02.1	4 : 21.0	2 : 21.5

Table 2. Timings (in min:sec) for the straight insertion method.

It is interesting to note that the optimal running median algorithm is worst on an initially sorted series, whether ascending or descending. The reason is that the new element entering the smoothing window will trickle all the way downwards (or upwards) through a heap.

Table 2 presents the timings for the straight insertion method. The columns refer to the same experiments as in Table 1. Since the algorithm was extremely slow for large K we present results only for K up to 255. Comparison of Table 1 and 2 reveals that for the above experiment the optimal median smoothing algorithm is about ten times faster than the insertion method. Inserting the new value X_{i+k+1} into the window would not improve the insertion algorithm appreciably due to the linear cost of data movement on each update. The timings for smoothing random series by the new method (Table 1) grow proportionally to $logK$; those for the insertion method increase proportional to K.

Acknowledgement. We would like to thank a referee for his careful reading of the code.

PROGRAM LIST

```
PROCEDURE runmed (n,
                  k            : integer;
                  VAR unsorted : nelemarray;
                  VAR median   : medianarray);
(* n          the real size of the array   <= maxn
   k          the real size of the window  <= maxk
   unsorted   only var because of storage
   median     returns the running median

   cause there is no unconditional loop in vs pascal rel. 1.0 we use
   repeat
   until false;
   to leave this loop, we jump to it"s end with some goto - statements

   if you translate this units into modula-2, you can replace
   repeat ...  until true; with loop ...  end;  and the goto - statement
   with exit.

   the structure of the "double heap" for example :

                       101  102  103  104
                          \ /      \ /
                           105     106
                              \ /
                              107
                               !
                              108
                               !
                              109
                              /  \
                            110    111
                           /  \     /  \
                         112  113  114  115                    *)

(*=============================================================*)

VAR
  window        : kelemarray;
  outlist       : outlistype;
  nrlist        : nrlistype;
```

```
k2,
shiftindex    : INTEGER;

(******** subroutines common to inittree and runmedint *******)

  PROCEDURE switch (l, r : INTEGER);
  VAR
    x   : elem;
    nrl,
    nrr : INTEGER;
  BEGIN
    x               := window [l];
    nrl             := nrlist [l];
    nrr             := nrlist [r];

    window [l]    := window [r];
    outlist [nrr] := l;
    nrlist [l]    := nrr;

    window [r]    := x;
    outlist [nrl] := r;
    nrlist [r]    := nrl;
  END; (* switch *)



(* now follow the first main procedure "inittree" with all the
   neccessary subroutines *)

PROCEDURE inittree ( k : INTEGER; VAR unsorted : nelemarray );

(* we built the first " double heap " with sorting the the first k
   elements of the array with a heapsort. *)

VAR
  ih1           : INTEGER;
  big           : REAL;

(************ all neccessary subroutines for inittree *******)

  PROCEDURE siftup (l, r : INTEGER);
  LABEL 7;
  VAR
```

```
      i,
      j,
      nrold : INTEGER;
      x     : elem;
    BEGIN
      i     := l;
      j     := 2*i;
      x     := window [i];
      nrold := nrlist [i];
      WHILE j <= r DO
      BEGIN
        IF j < r THEN
  IF window [j] key < window [j+1]. key THEN
          j := j+1;
IF x. key >= window [j]. key THEN goto 7;
window [i]             := window [j];
outlist [nrlist [j]] := i;
nrlist  [i]            := nrlist [j];
i                        := j;
j                        := 2*i;
      END;
7:    window [i]      := x;
      outlist [nrold] := i;
      nrlist [i]      := nrold;
    END; (* siftup *)

    PROCEDURE heapsort(ug,og : INTEGER);
    VAR
      l,
      r : INTEGER;
    BEGIN
      l := (og DIV 2)+1;
      r := og;
      WHILE l > ug DO
      BEGIN
        l := l-1;
siftup (l,r);
      END;

      WHILE r > ug DO
      BEGIN
switch (l,r);
        r := r-1;
siftup (l,r);
```

```
    END;
  END; (* heapsort *)

(* procedure-body of inittree *)
BEGIN
  k2 := k DIV 2;                    (* (2 * k2)+1  is the window *)
  FOR ih1 := 1 TO k DO
  BEGIN
    window [ih1]  := unsorted [ih1];
    nrlist [ih1]  := ih1 ;
    outlist [ih1] := ih1;
  END;

  heapsort (1,k);

  FOR ih1 := -k2 TO k2 DO
  BEGIN
    window [ih1]       :=  window [ih1+1+k2];
    nrlist [ih1]       :=  nrlist [ih1+1+k2];
    outlist [ih1+1+k2] :=  outlist [ih1+1+k2]-1-k2;
  END;

  big:=abs( unsorted[1].key );
  FOR ih1 := 2 TO n DO
    IF big < abs( unsorted[ih1 ].key ) THEN
   big:=abs( unsorted[ih1].key );

  FOR ih1 := k2+1 TO k DO
  BEGIN
    window [-ih1]. key   := -big ;
    window [ih1]. key    :=  big ;
  END;
END; (* inittree *)

(* now follow the second procedure runmedint with all neccessary
   subroutines *)

PROCEDURE runmedint (n,
                     k            : INTEGER;
                     VAR unsorted : nelemarray;
                     VAR median   : medianarray);
VAR
  nrnew,
  nrold,
```

```
    outold,
    childl,
    childr,
    father,
    nk,
    nk1        : INTEGER;

  (************ all neccessary subroutines for runmedint ******)

    PROCEDURE toroot;
    BEGIN
      REPEAT
father                          := outold DIV 2 ;
window  [outold]              := window [father];
        outlist [nrlist [father]] := outold;
nrlist  [outold]              := nrlist [father];
outold                          := father;
      UNTIL father = 0;
      window  [0]     := unsorted [nrnew];
      outlist [nrnew] := 0;
      nrlist  [0]     := nrnew;
    END; (* toroot *)

    PROCEDURE downtoleave;
    LABEL 1;
    BEGIN
      REPEAT
        childl := 2*outold;
        childr := childl-1;     (* because negative part of array *)
        IF window [childl]. key < window [childr]. key THEN
          childl := childr;
        IF window [outold]. key >= window [childl]. key THEN
          goto 1;
        switch (outold, childl);
        outold := childl;
      UNTIL false; (* loop *)
1:  END; (* downtoleave *)

    PROCEDURE uptoleave;
    LABEL 2;
    BEGIN
      REPEAT
        childl  := 2*outold;
        childr  := childl+1 ;
```

```
        IF window [childl]. key > window [childr]. key THEN
          childl := childr;
        IF window [outold]. key <= window [childl]. key THEN
          goto 2;
        switch (outold, childl);
        outold := childl;
      UNTIL false;  (* loop *)
2:  END; (* uptoleave *)

    PROCEDURE upperoutupperin;
    BEGIN                          (* upper out, upper in *)
      uptoleave;

      father := outold DIV 2 ;
      WHILE window [outold]. key < window [father]. key DO
      BEGIN
        switch (outold, father);
        outold := father;
        father := outold DIV 2 ;
      END
    END; (* upperoutupperin *)

    PROCEDURE upperoutdownin;
    BEGIN                         (* upper out, down in *)
      toroot;

      IF window [0]. key < window [-1]. key THEN BEGIN
        switch (0, -1);
outold := -1;
downtoleave;
      END;
    END; (* upperoutdownin *)

    PROCEDURE downoutdownin;
    BEGIN                          (* down out, down in *)
      downtoleave;

      father := outold DIV 2 ;
      WHILE window [outold]. key > window [father]. key DO
      BEGIN
        switch (outold, father);
        outold := father;
        father := outold DIV 2 ;
      END
```

```
      END; (* downoutdownin *)

      PROCEDURE downoutupperin;
      BEGIN                          (* down out, upper in *)
        toroot;

        IF window [0]. key > window [1]. key THEN BEGIN
          switch (1, 0);
outold := 1;
uptoleave;
        END;

      END; (* downoutupperin *)

      PROCEDURE wentoutone;
      BEGIN
        switch (1, 0);
        outold := 1;
        uptoleave;
      END; (* wentoutone *)

      PROCEDURE wentouttwo;
      BEGIN
        switch (-1, 0) ;
        outold := -1;
        downtoleave;
      END; (* wentouttwo *)

    (* procedure body of runmedint *)
    BEGIN
      nk  := n-k;
      nk1 := n-k+1;
      FOR nrold :=1 TO nk DO
        BEGIN
        median  [nrold]  := window [0];
        nrnew              := nrold+k;
        outold             := outlist [nrold];
        outlist [nrnew]  := outold ;
        nrlist  [outold] := nrnew;
        window  [outold] := unsorted [nrnew];

        IF outold > 0 THEN
          IF unsorted [nrnew] .key >= window [0] .key THEN
            upperoutupperin
```

```
        ELSE
          upperoutdownin
      ELSE IF outold < 0 THEN
        IF unsorted [nrnew]. key < window [0]. key THEN
          downoutdownin
        ELSE
          downoutupperin
      ELSE                              (* root (window [0] went out*)
        IF window [0]. key > window [1]. key THEN
          wentoutone
        ELSE IF window [0]. key < window [-1]. key THEN
          wentouttwo;
    END; (* for *)
    median [nk1] := window [0];
  END; (* runmedint *)

(* procedure-body of runmed *)

BEGIN
  k2 := k DIV 2;
  k  := 2*k2 +1;                  (* to be sure that k is odd         *)
  inittree (k, unsorted);         (* we built the first "double heap" *)
  runmedint (n, k, unsorted, median);
  (* shifting the results on the correct position *)
  FOR shiftindex := n DOWNTO n-k2+1 DO BEGIN
    median [shiftindex]. key := median [n-k+1].key
  END;
  FOR shiftindex := n-k2 DOWNTO k2+1 DO BEGIN
    median [shiftindex]. key := median [shiftindex-k2]. key
  END;
  FOR shiftindex := k2 DOWNTO 1 DO BEGIN
    median [shiftindex]. key := median[1].key
  END;
END;  (* rummed *)
```

REFERENCES

Atkinson, M., Sack, J.R., Santoro, N. and Strothotte, T. (1986). Min-Max Heaps and Generalized Priority Queues. *Comm.A.C.M.*, 29, 996–1000.

Baase, S. (1988). *Computer Algorithms. Introduction to Design and Analysis.* Second Edition, Addison-Wesley, Reading, Mass.

Bent, S.W. and John, J. (1985). Finding the Median Requires 2n Comparisons. *Proc*, 17^{th} *ACM Symposium on the Theory of Computing*, 213–216.

Carlsson, S. (1987). The Deap — A Double Ended Heap to Implement Double Ended Priority Queues. *Inf.Proc.Letters*, 26, 33–36.

Gill, J., Steiger, W. and Wigderson, A. (1988). The Complexity of Weighted Median Smoothing is $O(logK)$ Steps per Median. Technical Report, Dept. of Computer Science, Rutgers University.

Gonnet, G. and Munro, I. (1982). Heaps on Heaps. *Proc.* 9^{th} *ICLAP, Aarhus, Lecture Notes in Computer Science*, # 140, 282–287.

Mehlhorn, K. (1984). *Data Structures and Algorithms.* Vol.1: *Sorting and Searching.* Springer-Verlag, Berlin.

Schönhage, A., Patterson, M. and Pippenger, N. (1976). Finding The Median. *J.Comp.Syst.Sci.*, 13, 184–199.

Journal of Multivariate Analysis, 55, 205-218.

Estimation of Non-sharp Support Boundaries

W. HÄRDLE

Humboldt Universität zu Berlin, Berlin, Germany

B. U. PARK

Seoul National University, Seoul, Korea

AND

A. B. TSYBAKOV

Université Paris VI, Paris, France

Let $X_1, ..., X_n$ be independent identically distributed observations from an unknown probability density $f(\cdot)$, such that its support $G = \text{supp } f$ is a subset of the unit square in $\mathbb{R}^2$. We consider the problem of estimating G from the sample $X_1, ..., X_n$, under the assumption that the boundary of G is a function of smoothness γ and that the values of density f decrease to 0 as the power α of the distance from the boundary. We show that a certain piecewise-polynomial estimator of G has optimal rate of convergence (namely, the rate $n^{-\gamma/((\alpha+1)\gamma+1)}$) within this class of densities.

1. INTRODUCTION

Let $X_1, ..., X_n$ be independent identically distributed observations from an unknown probability density $f(\cdot)$ defined on the unit square $K = [0, 1] \times [0, 1]$ in $\mathbb{R}^2$. Consider the problem of estimating the supporty of f, i.e., the closed set $G = \text{supp } f = \text{Cl}\{x \in K : f(x) > 0\}$, given the sample $X_1, ..., X_n$. Here and later $\text{Cl}\{D\}$ means the closure of a set D.

We make some more specific assumptions on G, assuming that it is the set under a smooth curve, and on the density f, assuming that it is sufficiently regular near the boundary of G. Let the points of the square K be

denoted by $x=(x_1, x_2)$ and let $g(x_1)$, $0 \leqslant x_1 \leqslant 1$, be a function such that, for given γ, $L>0$ and $0<h<\frac{1}{2}$, we have $h \leqslant g(x_1) \leqslant 1-h$, $x_1 \in [0, 1]$, and

$$|g^{(k)}(t)-g^{(k)}(t')| \leqslant L\,|t-t'|^{\gamma-k} \qquad \forall t, t' \in [0, 1],$$

where $k=\lfloor \gamma \rfloor$. To every function g satisfying these conditions we assign the set

$$G=\{x=(x_1, x_2): 0 \leqslant x_1 \leqslant 1,\ 0 \leqslant x_2 \leqslant g(x_1)\}.$$

Denote $\mathscr{G}(\gamma, L, h)$ the set of all such sets G. We assume that the true density support G belongs to the class $\mathscr{G}(\gamma, L, h)$.

The fact that we consider the density support of a rather particular form prescribed by the class $\mathscr{G}(\gamma, L, h)$ is due to several reasons. First, as shown by Korostelev and Tsybakov [10], the reconstruction of such sets G is a building block for reconstruction of more general sets. Second, the class $\mathscr{G}(\gamma, L, h)$ itself is interesting in some applications, for example, in the problem of measuring economic efficiency [3, 8].

Here we impose one of the following assumptions on the density f.

ASSUMPTION A. *There exists a constant $C>0$ such that*

$$f(x) \geqslant C, \qquad x \in G, \qquad \text{and} \qquad f(x)=0, \qquad x \notin G,$$

where $G \in \mathscr{G}(\gamma, L, h)$.

ASSUMPTION B. *There exist constants α, C_1, $C_2>0$, such that the conditional density $f(x_2 \mid x_1)$ of x_2 given x_1 satisfies*

$$f(x_2 \mid x_1)=0 \qquad \text{for} \quad 0 \leqslant x_1 \leqslant 1, \quad g(x_1) \leqslant x_2 \leqslant 1, \tag{1}$$

and

$$f(x_2 \mid x_1) \geqslant C_1\,|g(x_1)-x_2|^{\alpha} \qquad \text{for} \quad 0 \leqslant x_1 \leqslant 1, \quad 0 \leqslant x_2 \leqslant g(x_1), \tag{2}$$

where g is such that $G \in \mathscr{G}(\gamma, L, h)$, and the marginal density $\mu(\cdot)$ of x_1 satisfies $\mu(x_1) \geqslant C_2$ for $x_1 \in [0, 1]$.

Note that the constant C_1 cannot be chosen arbitrarily large. The requirement that $G \in \mathscr{G}(\gamma, L, h)$ and that $f(x_2 \mid x_1)\mu(x_1)$ integrates to 1 implies that one needs at least $C_1<(\alpha+1)/(1-h)^{\alpha+1}$ to get a non-degenerate class of probability densities satisfying Assumption B. In fact, for our result, we need only a local version of (2), satisfied in some neighbourhood of the curve g, together with a strict positivity of $f(x_2 \mid x_1)$ elsewhere. We use the global inequality (2) to avoid additional notation. For the same reason we will assume where necessary that C_1 is as small as we need (but fixed).

Assumption A defines the class of probability densities f on K which we denote by $\mathcal{F}(\gamma, 0, L, h, C)$, or, briefly $\mathcal{F}(\gamma, 0)$. Quite similarly, the class of densities f satisfying Assumption B will be denoted by $\mathcal{F}(\gamma, \alpha, L, h, C_1, C_2)$, or, briefly $\mathcal{F}(\gamma, \alpha)$. Under Assumption A we set $\alpha = 0$ and we are in the case of a *sharp boundary* of density support, which means that the density f has a jump at the boundary. If Assumption B is satisfied, the density f may decrease to 0 continuously, therefore we call this the case of *non-sharp boundary*. Note, however, that under Assumption B we may also have $f(x) \geq C$, $x \in G$, for some $f \in \mathcal{F}(\gamma, \alpha)$, so that the class $\mathcal{F}(\gamma, \alpha)$ contains $\mathcal{F}(\gamma, 0)$ if $\alpha > 0$.

The problem of density support estimation was considered earlier by several authors (see, e.g., [5; 12; 13; 2; 14; 4; 11; 8; 16; 9; 10, Chap. 7]. In most of these papers the density f was supposed to be either uniform on G, or separated from 0 on G by a positive constant. (Except for the paper of Devroye and Wise [4] which provides a consistency result in a rather general setting). The estimation of G is equivalent to the estimation of the function g. The simplest histogram-type estimator of g was proposed by Geffroy [5] who proved its pointwise convergence with the rate $n^{-\gamma/(\gamma+1)}$ if $0 < \gamma \leq 1$ under Assumption A. Korostelev and Tsybakov [9, 10, Chap. 7] considered the case of the uniform density f and showed that the rate of convergence $n^{-\gamma/(\gamma+1)}$ is optimal in asymptotically minimax sense for all γ and that this rate is attained on piecewise-polynomial estimators, containing the Geffroy's estimator as a special case.

In this paper we prove that the optimality of convergence rate $n^{-\gamma/(\gamma+1)}$ holds not only for uniform densities f, but also under Assumption A. Next, we show that under Assumption B the optimal rate of convergence depends on α, and it equals $n^{-\gamma(\alpha+1)/(\gamma+1)}$, $\alpha > 0$. In the case $\alpha = 0$ we obtain "by continuity" the result under Assumption A. If $\alpha = 1$, which allows the Lipschitz continuous densities, we get the usual rate of convergence for nonparametric regression [6, 15]. This is not surprising, since the estimation of g can be viewed as nonparametric estimation of a functional of conditional distribution of x_2 given x_1 (namely, of the extremal point of support of conditional density). It is quite clear that under the appropriate assumptions on the conditional density (in fact, the Lipschitz condition is sufficient) this problem is characterized by the same optimal rates of convergence as the classical nonparametric regression. As α grows to infinity the rates of convergence become worse, since the average number of observations in a small neighbourhood of the curve $g(\cdot)$ decreases. We show that for every $\alpha > 0$ the estimator which attains the optimal rate can be chosen as the piecewise-polynomial estimator of Korostelev and Tsybakov [9, 10], with the suitable bin size depending on α. This, however, does not exclude other possible choices of optimal estimators. In particular, the analogy with non-regular parametric problems (cf. [7, Chaps. 5, 6]) suggests

that to improve the asymptotical constant factors multiplying the rates of convergence it could be better to use piecewise-polynomial Bayesian-type estimators rather than the piecewise-polynomial maximum-likelihood-type estimators of Korostelev and Tsybakov [9, 10].

Finally, we note that our estimator requires knowledge of α and γ, and it is suboptimal if the wrong parameters are used. It would be interesting to find an adaptive estimator which attains the asymptotically optimal behavior, but which does not depend on α and γ. This problem remains open.

2. The Results

Let us give some definitions. We consider estimators $\hat{G}_n$ of G. By an estimator $\hat{G}_n$ of G we mean an arbitrary closed set in K measurable with respect to $X_1, \ldots, X_n$. To measure the closedness of $\hat{G}_n$ to G we use the distance $d_1(G, \hat{G}_n)$, where, for any two closed sets G_1 and G_2 the value $d_1(G_1, G_2)$ is the Lebesgue measure of their symmetric difference:

$$d_1(G_1, G_2) = \mathrm{mes}(G_1 \Delta G_2).$$

(We use the subscript 1 in the notation d_1 to emphasize the fact that this is the L_1-type distance).

Let E_f be the expectation with respect to the distribution P_f of observations $X_1, \ldots, X_n$ when the underlying density is f. The error of an estimator $\hat{G}_n$ is measured by the risk function

$$\mathscr{R}(f, \hat{G}_n, q) = E_f[d_1^q(G, \hat{G}_n)],$$

where $q > 0$.

Following Ibragimov and Khasminskii [6, 7] and Stone [15] we say that $\psi_n \to 0$ is the correct normalizing factor (or optimal rate of convergence) for estimation of G in the class of densities $\mathscr{F}(\gamma, \alpha)$ if

$$\lambda_1(q) \leqslant \inf_{\hat{G}_n} \sup_{f \in \mathscr{F}(\gamma, \alpha)} \psi_n^{-q} \mathscr{R}(f, \hat{G}_n, q) \leqslant \lambda_2(q) \qquad \forall q > 0, \tag{3}$$

as $n \to \infty$, where λ_1 and λ_2 are positive constants, possibly depending on q, and $\inf_{\hat{G}_n}$ denotes the infimum over all estimators.

We say that the estimator G_n^* of G has the optimal rate of convergence in the class of densities $\mathscr{F}(\gamma, \alpha)$ if

$$\sup_{f \in \mathscr{F}(\gamma, \alpha)} \psi_n^{-q} \mathscr{R}(f, G_n^*, q) \leqslant \lambda_2(q) \qquad \forall q > 0, \tag{4}$$

where ψ_n is the correct normalizing factor, and $\lambda_2(q)$ is a constant.

In this paper we prove that a piecewise-polynomial estimator of maximum likelihood type has the optimal rate of convergence in the class $\mathcal{F}(\gamma, \alpha)$. To define this estimator we need some notation.

Let $\delta_n = n^{-1/((\alpha+1)\gamma+1)}$. Denote $M = M_n = 1/\delta_n$ and assume w.l.o.g. that M is an integer. Define $u_l = l\delta_n$, $l = 0, 1, ..., M$, and $U_j = [u_{j-1}, u_j)$, $j = 1, ..., M-1$, $U_M = [u_{M-1}, 1]$. Divide the square K into M slices of the form $A_j = U_j \times [0, 1]$ and in each slice consider the parametric family of sets

$$B_j(\theta) = \{x = (x_1, x_2) : x_1 \in U_j,\ 0 \leqslant x_2 \leqslant \theta_0 + \theta_1(x_1 - u_{j-1}) + \cdots + \theta_k(x_1 - u_{j-1})^k\},$$

where $k = \lfloor \gamma \rfloor$ is the maximal integer such that $k < \gamma$, and $\theta = (\theta_1, ..., \theta_k)$ is a vector of real parameters. Let

$$\Theta^{(K)} = \{\theta : 0 \leqslant \theta_0 + \theta_1(x_1 - u_{j-1}) + \cdots + \theta_k(x_1 - u_{j-1})^k \leqslant 1,\ \forall x_1 \in U_j\}.$$

(Note that $\Theta^{(K)}$ does not depend on j.) If $\theta \in \Theta^{(K)}$, the set $B_j(\theta)$ is located within the square K.

For each slice A_j we define a discretized estimator $\hat{\theta}_j = (\hat{\theta}_{0j}, ..., \hat{\theta}_{kj})$ of θ based on the observations in this slice,

$$\hat{\theta}_j = \underset{\theta \in \Theta_n :\, B_j(\theta) \supseteq \{X_i :\, X_i \in A_j\}}{\arg\min} \operatorname{mes}(B_j(\theta)),$$

where $\Theta_n = \{\theta = (m_0 \delta_n^{\gamma}, m_1 \delta_n^{\gamma-1}, ..., m_k \delta_n^{\gamma-k})\} \cap \Theta^{(K)}$, where $m = (m_0, m_1, ..., m_k)$ is a $(k+1)$-tuple of integers.

Finally, we define the piecewise-polynomial estoimator G_n^* of G as the closure of the set

$$\{x = (x_1, x_2) : 0 \leqslant x_1 \leqslant 1,\ 0 \leqslant x_2 \leqslant \hat{g}(x_1)\},$$

where

$$\hat{g}(x_1) = \sum_{j=1}^{M} (\hat{\theta}_{0j} + \hat{\theta}_{1j}(x_1 - u_{j-1}) + \cdots + \hat{\theta}_{kj}(x_1 - u_{j-1})^k)\, I\{x_1 \in U_j\}.$$

The result of this paper is the following.

THEOREM. *Let Assumption A or Assumption B be satisfied. Then the piecewise-polynomial estimator G_n^* defined above has the optimal rate of convergence in the class of densities $\mathcal{F}(\gamma, \alpha)$ and this rate equals $\psi_n = n^{-\gamma((\alpha+1)\gamma+1)}$. (Here $\alpha = 0$ under Assumption A and $\alpha > 0$ under Assumption B.)*

3. Proofs

To prove the theorem it suffices to show the inequalities

$$\limsup_{n\to\infty} \sup_{f\in\mathcal{F}(\gamma,\alpha)} \psi_n^{-q} E_f\,[d_1^q(G, G_n^*)] < \infty, \qquad q>0, \tag{5}$$

$$\liminf_{n\to\infty} \inf_{\hat G_n} \sup_{f\in\mathcal{F}(\gamma,\alpha)} \psi_n^{-q} E_f\,[d_1^q(G, \hat G_n)] > 0, \qquad q>0, \tag{6}$$

where $\psi_n = n^{-\gamma(\alpha+1)/(\gamma(\alpha+1)+1)}$.

Proof of Inequality (5) *under Assumption B.* We use the notation $\theta=\theta(m)$ to emphasize the dependence of θ on the vector m of integers. Also, if $\theta=\theta(m)$, we write for brevity

$$g_{lm}(x_1) = \theta_0 + \theta_1(x_1 - u_{l-1}) + \cdots + \theta_k(x_1 - u_{l-1})^k.$$

Let $B_l(\theta_l^{(1)})$ be the set from the family $\{B_l(\theta),\ \theta\in\Theta_n\}$ defined on p. 190 of Korostelev and Tsybakov [10] and satisfying

$$\begin{aligned} &B_l(\theta_l^{(1)}) \supseteq D_l = A_l \cap G, \\ &\sup_{u_{l-1}\le x_1<u_l} |g(x_1) - g_l^{(1)}(x_1)| \le C_3\delta_n^\gamma, \end{aligned} \tag{7}$$

where $C_3>0$ is a constant and $g_l^{(1)}(x_1) = g_{lm_l^{(1)}}(x_1)$, where the vector of integers $m_l^{(1)}$ is such that $\theta(m_l^{(1)}) = \theta_l^{(1)}$.

Fix $R>2C_3$. For $r=0, 1, \ldots$ define

$$L(r) = \{m : Rr < |m - m_l^{(1)}| \le R(r+1)\} \cap \{m : \theta(m)\in\Theta^{(K)}\},$$

where $|m|$ denotes the maximum of components norm of m. Using (7) we find

$$\begin{aligned} &P_f\Big(\sup_{u_{l-1}\le x_1<u_l} |\hat g(x_1) - g(x_1)| \ge R\delta_n^\gamma\Big) \\ &\le P_f\Big(\sup_{u_{l-1}\le x_1<u_l} |\hat g(x_1) - g_l^{(1)}(x_1)| + \sup_{u_{l-1}\le x_1<u_l} |g(x_1) - g_l^{(1)}(x_1)| \ge R\delta_n^\gamma\Big) \\ &\le P_f\Big(\sup_{u_{l-1}\le x_1<u_l} |\hat g(x_1) - g_l^{(1)}(x_1)| \ge R\delta_n^\gamma/2\Big) \\ &= P_f\Big(\sup_{u_{l-1}\le x_1<u_l} |\hat g(x_1) - g_l^{(1)}(x_1)| \ge R\delta_n^\gamma/2 \\ &\qquad \text{and } X_i\notin D_l\backslash B_l(\hat\theta_l)),\ i=1,\ldots,n\Big), \end{aligned}$$

where the last equality is due to the definition of $\hat{\theta}_l$. Let

$$L'(r) = L(r) \cap \{m : \sup_{u_{l-1} \leqslant x_1 < u_l} |g_{lm}(x_1) - g_l^{(1)}(x_1)| \geqslant R\delta_n^{\gamma}/2\}.$$

Then, using elementary argument, we obtain

$$P_f(\sup_{u_{l-1} \leqslant x_1 < u_l} |\hat{g}(x_1) - g(x_1)| \geqslant R\delta_n^{\gamma})$$

$$\leqslant \sum_{r=0}^{\infty} \text{card } L'(r) \max_{m \in L'(r)} P_f(X_i \notin D_l \backslash B_l(\theta(m)),\ i = 1, \ldots, n). \tag{8}$$

Let us estimate the probabilities in the last sum. Using Assumption B and denoting $\tilde{g}_{lm}(x_1) = \max\{g_{lm}(x_1), g(x_1)\}$, we have

$$P_f(X_1 \in D_l \backslash B_l(\theta(m)))$$

$$= \int_{u_{l-1}}^{u_l} \mu(x_1)\, dx_1 \int_{g_{lm}(x_1)}^{\tilde{g}_{lm}(x_1)} f(x_2 \mid x_1)\, dx_2$$

$$\geqslant C_1 C_2 \int_{u_{l-1}}^{u_l} dx_1 \int_{g_{lm}(x_1)}^{\tilde{g}_{lm}(x_1)} |g(x_1) - x_2|^{\alpha}\, dx_2$$

$$= C_1 C_2 (\alpha + 1)^{-1} \int_{u_{l-1}}^{u_l} |g(x_1) - g_{lm}(x_1)|^{\alpha+1} I\{g_{lm}(x_1) \leqslant g(x_1)\}\, dx_1. \tag{9}$$

It follows from (7) that

$$d_s^s(B_l(\theta_l^{(1)}), D_l) \equiv \int_{u_{l-1}}^{u_l} |g(x_1) - g_l^{(1)}(x_1)|^s\, dx_1 \leqslant C_4 \delta_n^{s\gamma+1} \qquad \forall s > 0, \tag{10}$$

here $C_4 > 0$ is a constant.

Note that

$$\int_{u_{l-1}}^{u_l} |g(x_1) - g_{lm}(x_1)|^{\alpha+1} I\{g_{lm}(x_1) \leqslant g(x_1)\}\, dx_1$$

$$\geqslant 2^{-\alpha-1} \int_{u_{l-1}}^{u_l} |g_l^{(1)}(x_1) - g_{lm}(x_1)|^{\alpha+1}$$

$$\cdot I\{g_{lm}(x_1) \leqslant g_l^{(1)}(x_1)\}\, dx_1 - C_5 \delta_n^{(\alpha+1)\gamma+1}, \tag{11}$$

where $C_5 > 0$ is a constant. To prove (11) observe that (with an obvious abbreviation of notation)

$$\int |g_l^{(1)} - g_{lm}|^{\alpha+1} I\{g_{lm} \leqslant g_l^{(1)}\}\, dx_1$$

$$= \int |g_l^{(1)} - g_{lm}|^{\alpha+1} I\{g_{lm} \leqslant g\}\, dx_1$$

$$+ \int |g_l^{(1)} - g_{lm}|^{\alpha+1} I\{g < g_{lm} \leqslant g_l^{(1)}\}\, dx_1$$

$$\leqslant 2^{\alpha+1} \int (|g_l^{(1)} - g|^{\alpha+1} + |g - g_{lm}|^{\alpha+1}) I\{g_{lm} \leqslant g\}\, dx_1$$

$$+ \int |g_l^{(1)} - g|^{\alpha+1}\, dx_1$$

$$\leqslant 2^{\alpha+1} \int |g - g_{lm}|^{\alpha+1} I\{g_{lm} \leqslant g\}\, dx_1 + (2^{\alpha+1} + 1) \int |g_l^{(1)} - g|^{\alpha+1}\, dx_1$$

$$\leqslant 2^{\alpha+1} \int |g - g_{lm}|^{\alpha+1} I\{g_{lm} \leqslant g\}\, dx_1 + (2^{\alpha+1} + 1)\, C_4 \delta_n^{(\alpha+1)\gamma+1},$$

where in the last inequality (10) was used.

An important step in our argument is the following. By definition of $\hat{\theta}_l$, the Lebesgue measure $\mathrm{mes}(B_l(\hat{\theta}_l))$ is smaller than $\mathrm{mes}(B_l(\theta))$ for any set $B_l(\theta)$, $\theta \in \Theta_n$, such that $B_l(\theta)$ contains all the sample points in the slice A_l. But $B_l(\theta_l^{(1)})$ is one of such sets. Hence $\mathrm{mes}(B_l(\hat{\theta}_l)) \leqslant \mathrm{mes}(B_l(\theta_l^{(1)}))$, and $\mathrm{mes}(B_l(\hat{\theta}_l) \setminus B_l(\theta_l^{(1)})) \leqslant \mathrm{mes}(B_l(\theta_l^{(1)}) \setminus B_l(\hat{\theta}_l))$. It means that in the sum in (8) it suffices to consider only such $m \in L'(r)$ that

$$\mathrm{mes}(B_l(\theta(m)) \setminus B_l(\theta_l^{(1)})) \leqslant \mathrm{mas}(B_l(\theta_l^{(1)}) \setminus B_l(\theta(m))). \tag{12}$$

Denote such a subset of $L'(r)$ by $\bar{L}(r)$. Thus, we replace $L'(r)$ by $\bar{L}(r)$ in (8). It will be convenient for us to write (12) in the form

$$\int_{u_{l-1}}^{u_l} |g_l^{(1)}(x_1) - g_{lm}(x_1)|\, I\{g_{lm}(x_1) > g_l^{(1)}(x_1)\}\, dx_1$$

$$\leqslant \int_{u_{l-1}}^{u_l} |g_l^{(1)}(x_1) - g_{lm}(x_1)|\, I\{g_{lm}(x_1) \leqslant g_l^{(1)}(x_1)\}\, dx_1. \tag{13}$$

In the following we use (13) together with the next lemma.

LEMMA. *Let π be a polynomial of order k on $[0, 1]$ and let*

$$\int |\pi| I\{\pi > 0\} \leqslant \int |\pi| I\{\pi \leqslant 0\} \tag{14}$$

(the integrals are over $[0, 1]$). Then, for every $s > 1$, there exists a constant $C^* > 0$ which depends on k and s only, such that

$$\int |\pi|^s I\{\pi > 0\} \leqslant C^* \int |\pi|^s I\{\pi \leqslant 0\}.$$

Proof of the lemma. Owing to the equivalence of L_1 and L_s norms for polynomials, there exists a constant $C' > 0$ depending on k and s only, such that

$$\left(\int |\pi|^s\right)^{1/s} \leqslant C' \int |\pi|. \tag{15}$$

Combining (14) and (15) we find

$$\int |\pi|^s I\{\pi > 0\} \leqslant \int |\pi|^s \leqslant \left(C' \int |\pi|\right)^s \leqslant \left(2C' \int |\pi| I\{\pi \leqslant 0\}\right)^s$$

$$\leqslant (2C')^s \int |\pi|^s I\{\pi \leqslant 0\}.$$

This proves the lemma. ∎

It is clear that, by change of the scale of x_1, (13) reduces to (14), with a polynomial π which does not depend on n. Hence, if $m \in \bar{L}(r)$ then, applying the lemma, we obtain

$$\int_{u_{l-1}}^{u_l} |g_l^{(1)}(x_1) - g_{lm}(x_1)|^{\alpha+1} I\{g_{lm}(x_1) > g_l^{(1)}(x_1)\}\, dx_1$$

$$\leqslant C^* \int_{u_{l-1}}^{u_l} |g_l^{(1)}(x_1) - g_{lm}(x_1)|^{\alpha+1} I\{g_{lm}(x_1) \leqslant g_l^{(1)}(x_1)\}\, dx_1$$

and, consequently,

$$\int_{u_{l-1}}^{u_l} |g_l^{(1)}(x_1) - g_{lm}(x_1)|^{\alpha+1} I\{g_{lm}(x_1) \leqslant g_l^{(1)}(x_1)\}\, dx_1$$

$$\geqslant (C^* + 1)^{-1} \int_{u_{l-1}}^{u_l} |g_l^{(1)}(x_1) - g_{lm}(x_1)|^{\alpha+1}\, dx_1$$

$$= (C^* + 1)^{-1} d_{\alpha+1}^{\alpha+1}(B_l(\theta(m)), B_l(\theta_l^{(1)})). \tag{16}$$

From (9), (11), and (16) we get

$$P_f(X_1 \in D_l \setminus B_l(\theta(m))) \geqslant C_1 C_2(\alpha+1)^{-1}[2^{-\alpha-1}(C^*+1)^{-1} d_{\alpha+1}^{\alpha+1}(B_l(\theta(m)), B_l(\theta_l^{(1)})) - C_5 \delta_n^{(\alpha+1)\gamma+1}]. \tag{17}$$

On the other hand, the same argument as in Lemma 4.2.1 of Korostelev and Tsybakov [10] shows that for any $s>0$

$$\lambda_1 \delta_n^{s\gamma+1} \|m-m_l^{(1)}\|^s \leqslant d_s^s(B_l(\theta(m)), B_l(\theta_l^{(1)})) \leqslant \lambda_2 \delta_n^{s\gamma+1} \|m-m_l^{(1)}\|^s, \tag{18}$$

where λ_1 and λ_2 are positive constants. Hence, for sufficiently large R we have

$$P_f(X_1 \in D_l \setminus B_l(\theta(m))) \geqslant C_6(rR)^{\alpha+1} \delta_n^{(\alpha+1)\gamma+1} = C_6(rR)^{\alpha+1} n^{-1}, \quad \text{if} \quad m \in \bar{L}(r), \tag{19}$$

where $C_6>0$ is a constant. Now, replacing in (8) $L'(r)$ by $\bar{L}(r)$, as agreed, and using (19) we get

$$\begin{aligned} P_f\left(\sup_{u_{l-1} \leqslant x_1 < u_l} |\hat{g}(x_1)-g(x_1)| \geqslant R\delta_n^{\gamma}\right) &\leqslant \sum_{r=0}^{\infty} \operatorname{card} \bar{L}(r) \max_{m \in \bar{L}(r)} P_f(X_i \notin D_l \setminus B_l(\theta(m)),\ i=1, \ldots, n) \\ &\leqslant a_0 + \sum_{r=1}^{\infty} \operatorname{card} \bar{L}(r)(1-C_6(rR)^{\alpha+1} n^{-1})^n \\ &\leqslant a_0 + \sum_{r=1}^{\infty} \operatorname{card} L(r) \exp(-C_6(rR)^{\alpha+1}) \\ &\leqslant a_0 + C_7 \exp(-R^{\alpha+1}/C_7), \end{aligned} \tag{20}$$

where $a_0 = \operatorname{card} L(0) \max_{m \in L'(0)} P_f(X_i \notin D_l \setminus B_l(\theta(m)),\ i=1, \ldots, n)$ and $C_7>0$ is a constant. (Here we used the fact that $\operatorname{card} L(r) = \mathcal{O}(r^k R^{k+1})$).

It remains to evaluate a_0. Note that if $m \in L'(0)$, then, using the definitions of the polynomials g_{lm} and $g_l^{(1)}$, we find

$$\|m-m_l^{(1)}\| \geqslant (k+1)^{-1} \delta_n^{-\gamma} \sup_{u_{l-1} \leqslant x_1 < u_l} |g_{lm}(x_1)-g_l^{(1)}(x_1)| \geqslant \frac{R}{2(k+1)}.$$

This inequality, together with (17) and (18), implies that for R large enough

$$P_f(X_1 \in D_l \backslash B_l(\theta(m))) \geqslant C_6' R^{\alpha+1} \delta_n^{(\alpha+1)\gamma+1} = C_6' R^{\alpha+1} n^{-1}, \qquad \text{if} \quad m \in L'(0),$$

and thus $a_0 \leqslant C_7' \exp(-R^{\alpha+1}/C_7')$, where C_6' and C_7' are positive constants. After the substitution of this estimate for a_0 in (20), we can finish the proof in the same way as in Theorem 7.4.1 of Korostelev and Tsybakov [10].

Proof of Inequality (5) *under Assumption A.* If Assumption A holds ($\alpha = 0$), the proof is obtained as an easy modification of Theorem 7.4.1 in Korostelev and Tsybakov [10]. In fact, it suffices to change P_G, E_G into P_f, E_f, respectively, and to replace the formula (7.14) there by (here we use some notation defined in Korostelev and Tsybakov [10])

$$\begin{aligned} P_f(X_i \notin D_l \backslash B_l(\theta(m)),\ i = 1, \ldots, n) &= (1 - P_f(X_1 \in D_l \backslash B_l(\theta(m))))^n \\ &\leqslant (1 - C \operatorname{mes}(D_l \backslash B_l(\theta(m))))^n \\ &\leqslant (1 - rc' R \delta_n^{\gamma+1})^n, \qquad m \in L'(r), \end{aligned}$$

where $c' > 0$. The rest of the argument is exactly as in the proof of Theorem 7.4.1 in Korostelev and Tsybakov [10].

Proof of Inequality (6). Consider the family of functions

$$g(x_1, \bar{\omega}) = \tfrac{1}{2} + (\lambda/M^\gamma) \sum_{j=1}^{M} \omega_j \varphi(M(x_1 - b_j)),$$

where φ is non-negative infinitely many times continuously differentiable function with support $(-\frac{1}{2}, \frac{1}{2})$, b_j is the center of the interval U_j,

$$M = 1/\delta_n = n^{1/((\alpha+1)\gamma+1)}, \qquad \bar{\omega} = (\omega_1, \ldots, \omega_M), \qquad \omega_j \in \{0, 1\},$$

and we suppose w.l.o.g. that M is an integer. Define the domains

$$G(\bar{\omega}) = \{x \in K : 0 \leqslant x_2 \leqslant g(x_1, \bar{\omega})\}.$$

It is easy to see that $G(\bar{\omega}) \in \mathcal{G}(\gamma, L, h)$ for λ small enough.

Define the family of bivariate density functions

$$f(x, \bar{\omega}) = \begin{cases} C_0 (g(x_1, \bar{\omega}) - x_2)_+^\alpha, & \text{if} \quad x_2 \geqslant \frac{1}{2}, \\ A_{\bar{\omega}}, & \text{if} \quad 0 \leqslant x_2 \leqslant \frac{1}{2}, \end{cases}$$

where $x = (x_1, x_2) \in K$ and C_0, $A_{\bar{\omega}}$ are positive constants. We choose $C_0 > C_1$, and define $A_{\bar{\omega}}$ so that $f(x, \bar{\omega})$ is a probability density, i.e.,

$$\tfrac{1}{2} A_{\bar{\omega}} + C_0 \int_0^1 \int_{1/2}^{g(x_1, \bar{\omega})} |g(x_1, \bar{\omega}) - x_2|^\alpha \, dx_2 \, dx_1 = 1.$$

This entails, in view of the chosen form of $g(x_1, \bar{\omega})$,

$$A_\omega = 2\left(1 - C_0[\lambda^{\alpha+1}/(\alpha+1)]\, M^{-\gamma(\alpha+1)-1} \sum_{j=1}^{M} \omega_j \int \varphi^{\alpha+1}(u)\, du\right)$$

$$= 2 + o(1). \tag{21}$$

Clearly, $f(x, \bar{\omega}) \in \mathscr{F}(\gamma, \alpha)$ if C_1 is small enough (we admit the smallness of C_1 for simplicity: see the remark after Assumption B). Denote by $P_{i,\omega}$ the distribution of X_i when the underlying density is $f(x, \bar{\omega})$.

For a fixed collection $(\omega_1, ..., \omega_M)$ of M values ω_j we introduce the vectors

$$\bar{\omega}_{j0} = (\omega_1, ..., \omega_{j-1}, 0, \omega_{j+1}, ..., \omega_M),$$

$$\bar{\omega}_{j1} = (\omega_1, ..., \omega_{j-1}, 1, \omega_{j+1}, ..., \omega_M).$$

Now we use the Assouad [1] lemma to prove the lower bound. In our particular case of support estimation we refer to the proof of Theorem 7.3.1 in Korostelev and Tsybakov [10] based on the Assouad technique. As noted there, it suffices to consider the case $q = 1$. Moreover, it suffices to consider the case $\alpha > 0$, since for $\alpha = 0$ the inequality (6) is a direct consequence of Theorem 7.3.1 in Korostelev and Tsybakov [10]. Following the lines of this theorem we find that, in the case $\alpha > 0$,

$$\sup_{f \in \mathscr{F}(\gamma, \alpha)} E_f[d_1(G, \hat{G}_n)]$$

$$\geqslant 2^{-1} \sum_{j=1}^{M} d_1(G(\bar{\omega}_{j0}), G(\bar{\omega}_{j1})) \int \min\left(\prod_{i=1}^{n} dP_{i,\omega_{j0}}, \prod_{i=1}^{n} dP_{i,\omega_{j1}}\right)$$

$$\geqslant 4^{-1} \sum_{j=1}^{M} d_1(G(\bar{\omega}_{j0}), G(\bar{\omega}_{j1})) \prod_{i=1}^{n} (1 - H^2(P_{i,\omega_{j0}}, P_{i,\omega_{j1}})/2)^2, \tag{22}$$

where $H^2(P_{i,\omega_{j0}}, P_{i,\bar{\omega}_{j1}})$ denotes the Hellinger distance between the probability measures $P_{i,\omega_{j0}}$ and $P_{i,\omega_{j1}}$.

Now, $d_1(G(\bar{\omega}_{j0}), G(\bar{\omega}_{j1}))$ does not depend on $(\omega_1, ..., \omega_M)$ and it equals

$$d_1(G(\bar{\omega}_{j0}), G(\bar{\omega}_{j1})) = (\lambda/M^\gamma) \int \varphi(M(x_1 - b_j))\, dx_1$$

$$= \left(\lambda \int \varphi(u)\, du\right) M^{-\gamma-1}. \tag{23}$$

Next,

$$H^2(P_{i,\omega_{j0}}, P_{i,\omega_{j1}}) = C_0 \int_0^1 \int_{g(x_1,\omega_{j0})}^{g(x_1,\omega_{j1})} |g(x_1,\bar{\omega}_{j1}) - x_2|^\alpha \, dx_2 \, dx_1$$
$$+ \int_0^1 \left(\int_0^{1/2} (A_{\omega_{j1}}^{1/2} - A_{\omega_{j0}}^{1/2})^2 \, dx_2 \right) dx_1$$
$$= C_0[\lambda^{\alpha+1}/(\alpha+1)] M^{-\gamma(\alpha+1)-1} \int \varphi^{\alpha+1}(u) \, du$$
$$+ \int_0^1 \int_0^{1/2} (A_{\omega_{j1}}^{1/2} - A_{\omega_{j0}}^{1/2})^2 \, dx_2 \, dx_1. \tag{24}$$

Using (21) and (24) we get

$$H^2(P_{i,\omega_{j0}}, P_{i,\omega_{j1}}) \leqslant C_8(M^{-\gamma(\alpha+1)-1} + |A_{\omega_{j1}} - A_{\omega_{j0}}|^2)$$
$$\leqslant C_9 M^{-\gamma(\alpha+1)-1} = C_9 n^{-1}, \tag{25}$$

where C_8 and C_9 are positive constants. By substitution of (23) and (25) into (22) we find

$$\sup_{f \in \mathcal{F}(\gamma,\alpha)} E_f[d_1(G, \hat{G}_n)] \geqslant 4^{-1} \left(\lambda \int \varphi(u) \, du \right) M^{-\gamma}(1 - C_9 n^{-1})^n.$$

This, in view of the definition of M, proves (6) for $q = 1$.

Acknowledgment

The research of the second author was supported by a grant from Korea Research Foundation 1993-1995.

References

[1] Assouad, P. (1983). Deux remarques sur l'estimation. *C.R. Acad. Sci. Paris* **296** 1021-1024.

[2] Chevalier, J. (1976). Estimation du support et du contenu du support d'une loi de probabilité. *Ann. Inst. H. Poincaré Sect. B* **12** (4) 339–364.

[3] Deprins, D., Simar, L., and Tulkens, H. (1984). Measuring labor efficiency in post offices. In *The Performance of Public Enterprises: Concepts and Measurements* (M. Marchand, P. Pestieau, and H. Tulkens, Eds.), pp. 243–267. North-Holland, Amsterdam.

[4] Devroye, L., and Wise, G. L. (1980). Detection of abnormal behavior via nonparametric estimation of the support. *SIAM J. Appl. Math.* **38** 480–488.

[5] Geffroy, J. (1964). Sur un problème d'estimation géométrique. *Publ. Inst. Statist. Univ. Paris* **13** 191–120.

[6] Ibragimov, I. A., and Khasminskii, R. Z. (1980). On nonparametric estimation of regression, *Dokl. Acad. Nauk SSSR* **252** 780-784.

[7] IBRAGIMOV, I. A., AND KHASMINSKII, R. Z. (1981). *Statistical Estimation: Asymptotic Theory*. Springer-Verlag, New York.

[8] KOROSTELEV, A. P., SIMAR, L., AND TSYBAKOV, A. B. (1995). Efficient estimation of monotone boundaries. *Ann. Statist.* **23**, in press.

[9] KOROSTELEV, A. P., AND TSYBAKOV, A. B. (1993). Estimation of support of a probability density and estimation of support functionals. *Problems Inform. Transmission* **29** 3–18.

[10] KOROSTELEV, A. P., AND TSYBAKOV, A. B. (1993). *Minimax Theory of Image Reconstruction*. Lecture Notes in Statistics, Vol. 82. Springer-Verlag, New York.

[11] MAMMEN, E., AND TSYBAKOV, A. B. (1995). Asymptotical minimax recovery of sets with smooth boundaries. *Ann. Statist.* **23**, in press.

[12] RÉNYI, A., AND SULANKE, R. (1963). Über die konvexe Hülle von n zufällig gewählten Punkten. *Z. Wahrsch. Verw. Gebiete* **2** 75–84.

[13] RÉNYI, A., AND SULANKE, R. (1964). Über die konvexe Hülle von n zufällig gewählten Punkten, II. *Z. Wahrsch. Verw. Gebiete* **3** 138–147.

[14] RIPLEY, B. D., AND RASSON, J. P. (1977). Finding the edge of a Poisson forest. *J. Appl. Probab.* **14** 483–491.

[15] STONE, C. J. (1980). Optimal rates of convergence for nonparametric estimators. *Ann. Statist.* **8** 1348–1360.

[16] TSYBAKOV, A. B. (1994). Multidimensional change-point problems and boundary estimation. In *Change-Point Problems*, IMS Lecture Notes Monograph Series, Vol. 23, pp. 317–329. Inst. Math. Statist., Hayward, CA.

Biometrika, 83, 529-540

ESTIMATION OF ADDITIVE REGRESSION MODELS WITH KNOWN LINKS

O.B. LINTON

Cowles Foundation for Research in Economics, Yale University

New Haven, CT 06520, USA.

W. HÄRDLE

Institut für Statistik und Ökonometrie, Humboldt-Universität zu Berlin

Spandauer Str. 1, D-10178 Berlin, Germany.

SUMMARY

We consider estimation of an additive nonparametric regression model with known link function. The asymptotic distribution of this regression estimate is given. The practical performance is investigated via an application to the study of migration between East and West Germany.

Some key words: Additive regression models; Dimensionality reduction; Kernel estimation; Nonparametric regression.

Biometrika, 83, 529-540

1. INTRODUCTION

Additive models are useful in a wide range of data analyses. They embody a key simplifying assumption that in some scale covariate effects are separable. This simplifying structure is present in many models of economic behavior, see Leontief (1947) for example. Combining separability with an unrestricted functional form for the covariate effects provides a very general class of models, that are collectively called additive nonparametric regression models.

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 & Hasminskii (1980) showed that the optimal rate for estimating m is $n^{-\frac{\ell}{2\ell+d}}$ with ℓ an index of smoothness of m. An additive structure for m is a regression function of the form $m(x) = c + \sum_{\alpha=1}^d m_\alpha(x_\alpha)$, where $x = (x_1, \ldots, x_d)'$ 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 (1986) showed that for such regression curves the optimal rate for estimating m is the one-dimensional rate of convergence with $n^{-\frac{\ell}{2\ell+1}}$. Thus one speaks of dimensionality reduction through additive modelling.

In practice, the backfitting procedures proposed in Breiman & Friedman (1985) and Buja, Hastie & Tibshirani (1989) are widely used to estimate the additive components. These methods have been evaluated on numerous datasets and been refined quite considerably since their introduction. They are based on iteratively calculating one dimensional smoothers until some convergence criterion is satisfied. The resulting procedure is linear in Y, provided the one dimensional smoothers are linear; this is the basis for confidence intervals and degrees of freedom calculation. However, neither the bias nor the variance have been obtained explicitly. Recently, Linton & Nielsen (1995) and Tjøstheim & Auestad (1994) have proposed an alternative procedure for estimating the components of additive regression. These new procedures are based on direct integration of the initial multidimensional smoothers exploiting the following idea. Suppose that $m(x, z)$ is any bivariate function, and consider the quantities $\mu_1(x) = \int m(x, z) dQ(z)$ and $\mu_2(z) = \int m(x, z) dQ(x)$, where Q is a probability measure. If $m(x, z) = m_1(x) + m_2(z)$, then $\mu_1(.)$ and $\mu_2(.)$ are

$m_1(.)$ and $m_2(.)$, respectively, up to a constant. In practice one replaces m by an estimate. The statistical properties of the integration method are straightforward to derive.

For many situations, especially binary and survival time data, a more appropriate framework for modelling is, in the parametric case, provided by Generalised Linear Models, see McCullagh & Nelder (1990). Hastie & Tibshirani (1991) extend these ideas to nonparametric modelling. We consider two different versions of this: the partial and full model specifications. In the full model specification, the conditional distribution of Y given X belongs to an exponential family with known link function G and mean m, where

$$G\{m(x)\} = c + \sum_{\alpha=1}^{d} f_\alpha(x_\alpha), \tag{1.1}$$

where $E\{f_\alpha(X_\alpha)\} = 0$, $\alpha = 1, \ldots, d$. This full model specification is what is usually called Generalised Additive Model. It implies for example that the variance is functionally related to the mean. In some respects we prefer the partial model specification in which we keep (1.1), but do not restrict ourselves to the exponential family. In this case, the variance function is unrestricted. This flexibility is a relevant consideration for many datasets where there is overdispersion, see Cox (1983). When G is the identity function we have the additive regression model examined in Linton & Nielsen (1995). Other examples include the logit and probit link functions for binary data, and the logarithm transform for Poisson count data, see McCullagh & Nelder (1990), or more generally for when the regression function is multiplicative.

The Backfitting procedure in conjunction with Fisher Scoring is widely used to estimate Generalised Additive Models, see Hastie & Tibshirani (1991). These methods exploit the likelihood structure, but are even less tractable from a statistical point of view when G is not the identity, since the estimate is not linear in Y.

We propose an integration based method of estimating the components f_α based on (1.1). The main advantage of our method is that one can obtain its asymptotic properties. It is asymptotically normal at the optimal one dimensional convergence rate. One version of our procedure involves merely exploiting (1.1), but we also suggest how to take account of the additional information provided by the exponential family structure. We discuss the merits of imposing this additional structure.

(1996) Linton, O.B. and Härdle, W.
Estimation of Additive Regression Models with known Links.

In § 2 we define the estimation procedures. In §3 we give the asymptotic properties of the procedures. In §5 we give the results of an application to German migration data.

2. Estimation Procedures

2.1 *Estimating the Additive Components*

Partition $X = (X_1, X_2)$, where X_1 is the one dimensional direction of interest and X_2 is a $d-1$-dimensional nuisance direction, and let $x = (x_1, x_2)$. For any regression function m and transformation G, define the functional

$$\varphi_1(x_1) = \int G\{m(x_1, x_2)\}\, p_2(x_2) dx_2, \tag{2.1}$$

where $p_2(x_2)$ is the joint density of x_2. We here consider estimation of this functional motivated by the fact that under the additive structure (1.1), φ_1 is f_1 up to the additive constant c. The general strategy is to replace both m and p_2 in (2.1) by estimates. We use the multidimensional Nadaraya-Watson kernel estimator

$$\widehat{m}(x_1, x_2) = \frac{n^{-1}\sum_{k=1}^n K_h(x_1 - X_{1k}) L_g(x_2 - X_{2k}) Y_k}{n^{-1}\sum_{k=1}^n K_h(x_1 - X_{1k}) L_g(x_2 - X_{2k})},$$

where K and L are compactly supported Lipschitz continuous kernels integrating to one. Here, $K_h(.) = h^{-1}K(h^{-1}.)$ and $L_g(.) = g^{-(d-1)}L(g^{-1}.)$. We take K to be a second order kernel and L to be a product of univariate kernels of order q, i.e. $\int L(u)u^j du = 0$ for $j = 1, \ldots, q-1$. For large dimensions d it will be necessary to use bias reduction on the nuisance directions to achieve the optimal one-dimensional rate of convergence for the direction of interest. Note that $\widehat{m}(x_1, x_2) = \sum_{k=1}^n w_k(x_1, x_2) Y_k$ for weights $\{w_k\}_{k=1}^n$ depending only on the design. In principle any respectable smoother can be used in place of $\widehat{m}(x_1, x_2)$, although we have only proven results for the Nadaraya-Watson estimator.

We estimate $\varphi_1(x_1)$ by the sample version of (2.1):

$$\widetilde{\varphi}_1(x_1) = n^{-1}\sum_{i=1}^n G\{\widehat{m}(x_1, X_{2i})\}. \tag{2.2}$$

When G is the identity function, $\widetilde{\varphi}_1(x_1)$ is linear, i.e. $\widetilde{\varphi}_1(x_1) = \sum_{k=1}^n \overline{w}_k(x_1) Y_k$, where $\overline{w}_k(x_1) = n^{-1}\sum_{i=1}^n w_k(x_1, X_{2i})$. In general however, $\widetilde{\varphi}_1(x_1)$ is a nonlinear function of Y_i. We are using the empirical weighting version of the Linton & Nielsen (1995) procedure.

(1996) Linton, O.B. and Härdle, W.
Estimation of Additive Regression Models with known Links.

2.2 *Estimation of the Regression Surface, Residuals, and Model Selection*

The above procedure is carried out on each direction by in each case redefining the j'th coordinate to be X_1 and the rest to be X_2. We obtain estimates of each φ_α at each sample point. Let $\tilde{c} = d^{-1}n^{-1}\sum_{\alpha=1}^{d}\sum_{i=1}^{n}\tilde{\varphi}_\alpha(X_{\alpha i})$ and $\tilde{f}_1(x_1) = \tilde{\varphi}_1(x_1) - \tilde{c}$. We reestimate $m(x)$ by

$$\widetilde{m}(x) = F\left\{\sum_{\alpha=1}^{d}\tilde{f}_\alpha(x_\alpha) + \tilde{c}\right\}, \tag{2.3}$$

where $F = G^{-1}$. Let $\tilde{\varepsilon}_i = Y_i - \widetilde{m}(X_i)$ be the additive regression residuals which estimate the errors $\varepsilon_i = Y_i - m(X_i)$. These residuals, or alternatively the deviance residuals if a likelihood framework is adopted, can be used to test the additive structure, i.e. to look for interactions. When (1.1) is true, $\tilde{\varepsilon}_i$ should be approximately uncorrelated with any function of X_i.

In some cases, especially with many candidate variables, one may wish to select an even more restricted model that excludes some variables from (1.1), depending on their importance in resolving variance. Let $S_\alpha = \int f_\alpha^2(x)p_\alpha(x)dx$, then S_α captures the magnitude of the effect of X_α in an obvious way, not restricted to the likelihood framework. Let $\tilde{S}_\alpha = n^{-1}\sum_{i=1}^{n}\tilde{f}_\alpha^2(X_{\alpha i})$. We might choose directions α for which $\tilde{S}_\alpha$ is larger than some preselected threshold, as suggested by Härdle & Korostelev (1995), or at least report the magnitudes of these quantities. This is more informatively done in the ratio scale $\tilde{S}_\alpha / \sum_{\gamma=1}^{d}\tilde{S}_\gamma$. This is reminiscent of principal component or factor analysis.

3. Asymptotic Properties

3.1 *Estimators of the additive components*

We first establish the asymptotic properties of $\tilde{\varphi}_1$ at an interior point x_1. The theorem is proved for the most general case where m is the regression function of Y on X; no other structure such as additivity is being maintained. Only the interpretation of what is being estimated requires this structure. We work throughout with a design that has bounded support, and densities p_1, p_2, and p that are bounded away from zero. As far

as smoothness is concerned, we assume that all the densities and component functions are continuously differentiable of order q. These assumptions are fairly standard to the nonparametric regression literature. We also suppose that the link function G is twice continuously differentiable, and the variance function $\sigma^2(x) = \text{var}\,(\mathrm{Y}|\mathrm{X}_1 = \mathrm{x}_1, \mathrm{X}_2 = \mathrm{x}_2)$ is Lipschitz continuous. The logit and probit link functions satisfy these latter restrictions. Let $\|K\|_2^2 = \int K^2(u)du$ and $\mu_2(K) = \int u^2 K(u)du$.

THEOREM 1. *Let the order q of L satisfy $q > d-1$. Let $h = \beta n^{-1/5}$, and assume that $n^{2/5}g^q \to 0$ and $n^{2/5}g^{d-1} \to \infty$. Then*

$$n^{2/5}\left\{\widetilde{\varphi}_1(x_1) - \varphi_1(x_1)\right\} \to N\left\{b_1(x_1), v_1(x_1)\right\} \tag{3.1}$$

in distribution, where

$$b_1(x_1) = \beta^2\mu_2(K)\left[\tfrac{1}{2}f_1''(x_1)\int G'\left\{m(x)\right\}p_2(x_2)dx_2 + f_1'(x_1)\int G'\left\{m(x)\right\}\tfrac{\partial \ln p}{\partial x_1}(x)p_2(x_2)dx_2\right]$$

$$v_1(x_1) = \beta^{-1}\|K\|_2^2\int G'\left\{m(x)\right\}^2\sigma^2(x)\tfrac{p_2^2(x_2)}{p(x)}dx_2.$$

Finally, $v_1(x_1)$ can be consistently estimated by

$$\widetilde{v}_1(x_1) = \sum_{k=1}^{n}\widetilde{\delta}_k^2\widetilde{\varepsilon}_k^2, \tag{3.2}$$

where $\widetilde{\delta}_k = n^{-1}\sum_{j=1}^{n}G'\left\{\widetilde{m}(x_1, X_{2j})\right\}w_k(x_1, X_{2j})$.

PROOF. By a Taylor expansion

$$\frac{1}{n}\sum_{i=1}^{n}\left[G\left\{\widehat{m}(x_1, X_{2i})\right\} - G\left\{m(x_1, X_{2i})\right\}\right] = \frac{1}{n}\sum_{i=1}^{n}G'\left\{m(x_1, X_{2i})\right\}\left\{\widehat{m}(x_1, X_{2i}) - m(x_1, X_{2i})\right\} + R,$$

where $R = \frac{1}{2n}\sum_{i=1}^{n}G''(m_i^*)\left\{\widehat{m}(x_1, X_{2i}) - m(x_1, X_{2i})\right\}^2$, with m_i^* intermediate between $\widehat{m}(x_1, X_{2i})$ and $m(x_1, X_{2i})$. By the Cauchy-Schwarz inequality,

(1996) Linton, O.B. and Härdle, W.
Estimation of Additive Regression Models with known Links.

$$|R| \leq \frac{1}{2}\frac{1}{n}\sum_{i=1}^{n}\left\{G''(m_i^*)^2\right\}^{1/2}\left\{\sup_{x_2}|\widehat{m}(x_1,x_2)-m(x_1,x_2)|\right\}^2, \tag{3.3}$$

and the first term on the right hand side of (3.3) is of order one, while the second term is of order $\max\left\{h^4, g^{2q}, n^{-1}h^{-1}g^{-(d-1)}\right\}$ by standard theory for regression smoothers, see Härdle (1990). Provided $n^{2/5}g^{d-1} \to \infty$ and $n^{2/5}g^{2q} \to 0$, we have that $n^{2/5}\{\tilde{\varphi}_1(x_1) - \varphi_1(x_1)\}$ is asymptotically equivalent to

$$T_n = n^{2/5}\frac{1}{n}\sum_{i=1}^{n} G'\{m(x_1, X_{2i})\}\{\widehat{m}(x_1, X_{2i}) - m(x_1, X_{2i})\}$$

which can be handled as in Linton & Nielsen (1995). This is done in the technical appendix. □

When $d = 2$, a second order kernel may be used in the nuisance direction for some choice of bandwidths, but in higher dimensions at least some bias reduction must be used in order to achieve the natural rate $n^{2/5}$. This bias reduction strategy is similar to what is needed in certain semiparametric problems: Robinson (1988) showed that to obtain $n^{1/2}$ consistent estimates of the slope coefficients in the partially linear model, one must make the bias in the nonparametric estimator smaller than $n^{-1/4}$ which for large dimensions requires higher order kernels and undersmoothing. The variance of the nonparametric estimator is taken care of by averaging. In our case, a weighted average of the variance of the d-dimensional pilot smoother affects $\tilde{\varphi}_1(x_1)$ through R: this contribution is necessarily of order $n^{-1}h^{-1}g^{-(d-1)}$. Making this term smaller than $n^{-2/5}$ requires the above restrictions on g.

If the local linear smoother, see Fan (1992), were used as a pilot in place of the Nadaraya-Watson estimator, the asymptotic variance of $\tilde{\varphi}_1(x_1)$ would be the same but the bias would take the simpler form

$$b_1(x_1) = \beta^2\mu_2(K)\left[\frac{1}{2}f_1''(x_1)\int G'\{m(x)\}\, p_2(x_2)dx_2\right].$$

The construction of asymptotic confidence intervals with coverage probability $1-\alpha$ unfortunately involves estimating the terms $b_1(x_1)$ and $v_1(x_1)$, although with undersmoothing, i.e. $h = o(n^{-1/5})$, it suffices to approximate the variance, see Härdle & Linton

(1996) Linton, O.B. and Härdle, W.
Estimation of Additive Regression Models with known Links.

(1995). The bootstrap provides one alternative for approximating desired p-values. To find such a sample based approximation to the distribution of $n^{2/5}\{\widetilde{\varphi}_1(x_1) - \widetilde{\varphi}_1(x_1)\}$ one can use, for instance, the wild bootstrap, see Härdle & Marron (1991). Finally, bandwidth choice can be based on a rule-of-thumb method as in Linton & Nielsen (1995).

3.2 *Estimators of the Regression Surface*

Here, we work with the additive structure given in (1.1). Our theorem is not limited to the exponential family structure discussed in the introduction. The properties of $\widetilde{m}(x)$ follow from Theorem 1, the delta method, and from the fact that $\widetilde{\varphi}_j(x)$ and $\widetilde{\varphi}_k(x)$ are asymptotically uncorrelated for $j \neq k$.

THEOREM 2. *Suppose in addition to the assumptions of the previous theorem that F is twice continuously differentiable and that (1.1) holds. Then*

$$n^{2/5}\{\widetilde{m}(x) - m(x)\} \to N\{b(x), v(x)\},$$

in distribution, where

$$b(x) = F'\left\{\sum_{\alpha=1}^{d} f_\alpha(x_\alpha) + c\right\} \sum_{\alpha=1}^{d} b_\alpha(x_\alpha)$$
$$v(x) = F'\left\{\sum_{\alpha=1}^{d} f_\alpha(x_\alpha) + c\right\}^2 \sum_{\alpha=1}^{d} v_\alpha(x_\alpha).$$

It follows from Theorem 2 that the rate of convergence of $\widetilde{m}$ is free from the "curse of dimensionality". We obtain the rate $n^{2/5}$ that is achieved in estimation of scalar regression functions, see Stone (1986) and Hall (1989). It is possible also to exploit the additional smoothness in the direction of interest by using a higher order kernel in place of K. Under appropriate restraints on the smoothness and dimensionality, one can obtain the optimal rate of convergence $n^{-\ell/2\ell+1}$.

4. EXPLOITING ADDITIONAL STRUCTURE

Now suppose that the exponential family structure discussed in the introduction is present. This restriction on the conditional distribution can be exploited in estimation. We consider two different approaches. The first approach is to use a pilot smoother that exploits the additional structure; for example the local likelihood procedure considered in the unpublished Standford dissertation of Tibshirani (1984) and in Fan, Heckman & Wand (1995). Let

$$Q_n(x;\beta) = \sum_{i=1}^{n} \ell_i(\beta) K_h(X_{1i} - x_1) L_g(X_{2i} - x_2)$$

where $\ell_i(\beta)$ is a likelihood function for observation i. In the binary data case

$$\ell_i(\beta) = y_i \ln\left[F\left\{\beta_0 + \beta_1'(X_i - x)\right\}\right] + (1 - y_i)\ln\left[1 - F\left\{\beta_0 + \beta_1'(X_i - x)\right\}\right],$$

where $\beta = (\beta_0, \beta_1)$ is a $d+1$ by 1 vector. Now choose $\widehat{\beta}(x)$ to minimize $Q_n(x;\beta)$, and let $\widehat{m}(x)$ be what is implied by $\widehat{\beta}(x)$, in the binary case $F\left\{\widehat{\beta}_0(x)\right\}$. We now integrate this new pilot estimator $\widehat{m}(x)$ as before.

An alternative method of exploiting additional structure works by producing several different estimates of the additive component and then combining them optimally through the minimum distance method. This method is widely used in econometrics; it allows one to be somewhat selective about the information that one uses, see Rothenberg (1972). In the binomial example, $\sigma^2 = m(1-m)$, thus $m = \frac{1}{2} - \frac{1}{2}(1-4\sigma^2)^{1/2}$ for $\sigma^2 \leq 1/4$. Therefore, an alternative procedure is to integrate $G_2(\widehat{\sigma}^2)$, where $G_2(t) = G\left\{\frac{1}{2} - \frac{1}{2}(1-4t)^{1/2}\right\}$ and $\widehat{\sigma}^2(.)$ is a nonparametric estimate of the conditional variance function $\sigma^2(.)$. One may then want to combine the estimate obtained from the mean with the estimate obtained from the variance. The optimal way of doing this is through minimum distance. More generally, let $\widetilde{\Phi}_1$ be a J by 1 vector of estimates of the additive components at a point x_1, and define the minimum distance estimate of φ_1,

$$\widetilde{\varphi}_1^* = (e'\widehat{A}^{-1}e)^{-1}e'\widehat{A}^{-1}\widetilde{\Phi}_1,$$

where e is a J by 1 vector of ones, while $\widehat{A}$ is an estimate of the covariance matrix of $\widetilde{\Phi}_1$.

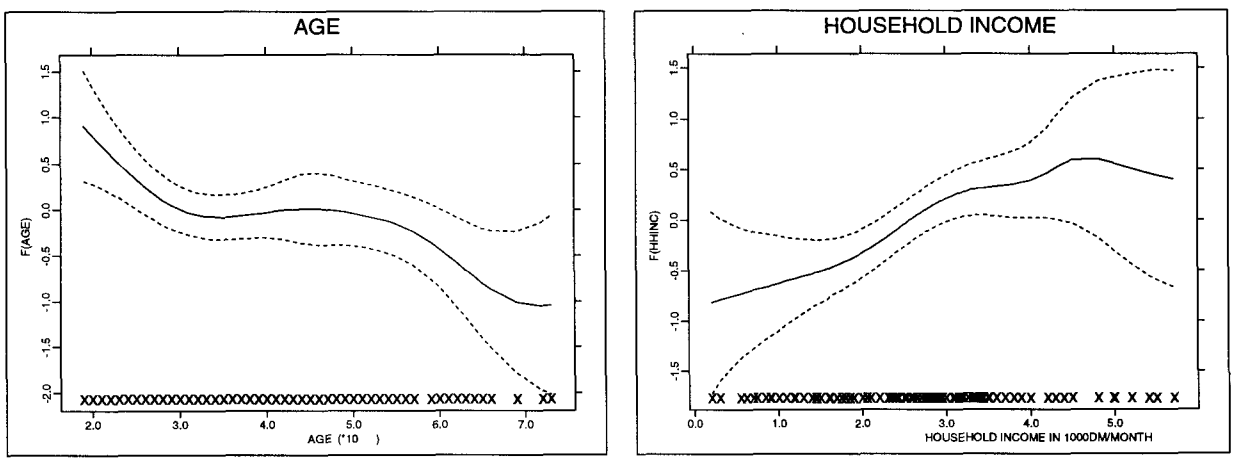
The question is, what do we gain from exploiting this additional information? Although the variance should improve, the bias may actually worsen. A similar point is

addressed by Jones (1993); in particular whether one should take account of heteroskedasticity in estimating a nonparametric regression. He argues for the unadjusted smoother in this case.

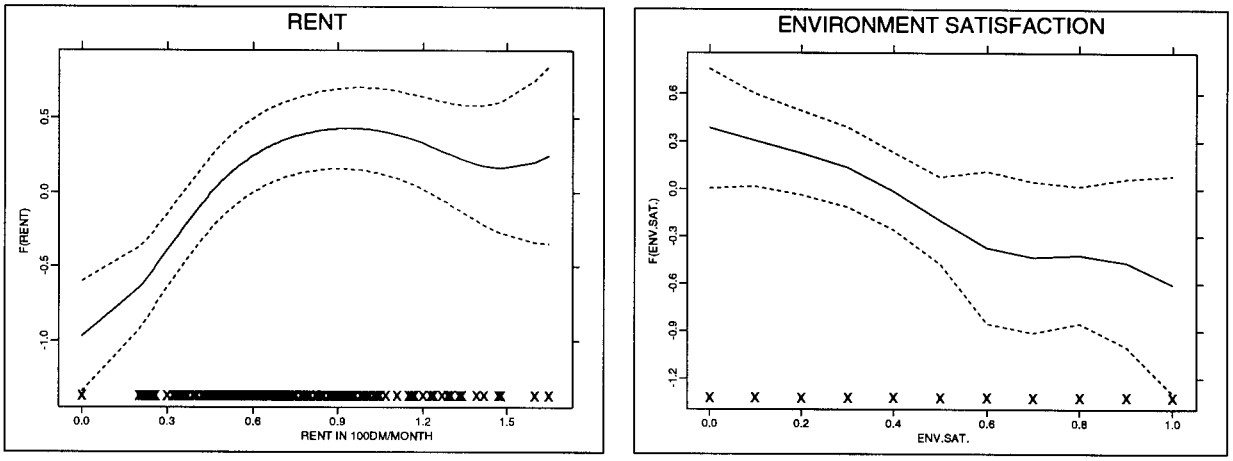
5. Empirical Illustration

We applied our procedures to the study of migration between East and West Germany using data from the 1991 Social and Economic Panel survey conducted by the Deutsche Institut Wirtschaftsforschung. The dependent variable is whether the individual intended to migrate from East to West Germany at this time. To explain this, we have four continuous variables: age, household income, rent, and a subjective measure of personal satisfaction on a scale of 1 to 10. Our sample consists of 315 individuals who all had at least Abitur education and had some friends in the West. Of these, 172 had the intention of leaving the East.

We used the logit link function. In all procedures we took the same Gaussian kernel with bandwidths $h = 0.5$ and $g = 1.0$ (relative to the studentized design). We experimented with the choice of bandwidths but found that this choice worked well for our dataset. Below we present our estimates of the additive components along with symmetric 90% pointwise confidence intervals calculated as suggested by Theorem 1.



(1996) Linton, O.B. and Härdle, W.
Estimation of Additive Regression Models with known Links.



FIGURES 1-4

The interesting curves are those for age and rent. The effect of rent is nonmonotonic: initially the probability of migration increases with rent, but this levels off and even declines after a certain point. This effect is not due to boundary issues, as can be seen from the design density. The probability of migration generally decreases with age but is effectively constant between 30 and 50. The relative importance ($S_\alpha / \sum S_\beta$) of the variables is as follows: rent (64.7%), income (14.5%), satisfaction (10.9%), and age (9.9%).

Finally, for comparison we give the results of a parametric logit fit.

TABLE 1

Variable	Coeff.	Std.Err.	$P > \|z\|$
constant	0.21515	0.120979	0.07632
age	-0.02867	0.010997	0.00957
hhinc	0.00022	0.000121	0.06827
rent	0.01188	0.002765	2.3158e-5
envsat	-0.11613	0.052526	0.02777

The coefficient on rent, 0.012, is quite close to the average slope of the rent component reported in Figure 3 above. Thus the two methods produce similar results in this average sense, but clearly the parametric method obliterates the structure apparent in Figure 1-4. Computer programs are available from the authors upon request.

6. FINAL REMARKS

(1996) Linton, O.B. and Härdle, W.
Estimation of Additive Regression Models with known Links.

The backfitting procedure produces the closest additive approximation to the regression function whether or not the true model is additive. The integration procedure does not have this property except under (1.1). However, the functional φ_1 that we estimate always has a sensible interpretation: it is the effect of X_1 in the transformed scale after averaging with respect to the other variables.

Finally, we would like to mention some semiparametric extensions of these models that are currently under investigation. One model of interest is

$$G\{m(x)\} = c + \sum_{\alpha=1}^{a} f_\alpha(x_\alpha) + \sum_{\delta=1}^{d} \theta_\delta x_\delta,$$

i.e. partially linear inside the link, which generalizes the semiparametric regression model of Cuzick (1992). Using the integration method we can estimate the parameters θ and the component functions fully exploiting the additive structure.

ACKNOWLEDGEMENTS

We thank two anonymous referees, Ray Carroll, Enno Mammen, Jens Perch Nielsen, and Eric Severance-Lossin for helpful comments. We thank Stefan Sperlich for help with the computations. We also thank for financial support: the Sonderforschungsbereich 373 "Quantifikation und Simulation Ökonomischer Prozesse", the National Science Foundation, and NATO through a Collaborative Research Grant.

(1996) Linton, O.B. and Härdle, W.
Estimation of Additive Regression Models with known Links.

Biometrika, 83, 529-540

APPENDIX

We use the following assumptions

A1. *The function G has bounded second derivative over any compact interval.*

A2. $\sigma^2(x_1, x_2) = \mathrm{var}\,(\mathrm{Y}|\mathrm{X}_1 = \mathrm{x}_1, \mathrm{X}_2 = \mathrm{x}_2)$ *is bounded and Lipschitz continuous.*

A3. *The functions f_1 and f_2 are bounded and Lipschitz continuous.*

A4. *The densities $p_1(\cdot)$, $p_2(\cdot)$, and $p(\cdot)$ are bounded, Lipschitz continuous and bounded away from zero by a constant p_0.*

A5. *The functions f_1, f_2 and all the joint densities are continuously q times differentiable.*

A6. *The kernel function $K(\cdot)$ is bounded, nonnegative, compactly supported, Lipschitz continuous and $\int K(u)du = 1$. $\|K\|_2^2 = \int K^2(u)du < \infty$, $\mu_2(K) = \int u^2 K(u)du < \infty$.*

A7. *The kernel function $L(\cdot)$ is bounded, compactly supported, Lipschitz continuous and $\int L(u)du = 1$, $\int u^i L(u)du = 0$, $i = 1, \ldots, q-1$.*

A8. $h = \beta n^{-1/5}$.

A9. *The sequence of bandwidths g are such that $g^q n^{2/5} \to 0$ and $n^{2/5} g^{d-1} \to \infty$.*

A10. *The function F has bounded second derivative over any compact interval.*

Let E_* and E_i denote expectation conditional on the design and on the $i'th$ design observation respectively.

PROOF OF THEOREM 1. Let $\widehat{m}(x) = \frac{\widehat{r}(x)}{\widehat{p}(x)}$. We first write

$$\widehat{m}(x_1, X_{2i}) - m(x_1, X_{2i}) = \frac{\widehat{r}(x_1, X_{2i}) - m(x_1, X_{2i})\widehat{p}(x_1, X_{2i})}{\widehat{p}(x_1, X_{2i})} \equiv \frac{\widehat{a}_i}{\widehat{p}_i},$$

then the leading term T_n can be decomposed into "bias" and "variance" terms

$$T_n = \frac{1}{n}\sum_{i=1}^{n} G'\{m(x_1, X_{2i})\}\frac{E_i(\widehat{a}_i)}{\widehat{p}(x_1, X_{2i})} + \frac{1}{n}\sum_{i=1}^{n} G'\{m(x_1, X_{2i})\}\frac{\widehat{a}_i - E_i(\widehat{a}_i)}{\widehat{p}(x_1, X_{2i})} = T_{1n} + T_{2n}.$$

The terms T_{1n} and T_{2n} can be further approximated by

(1996) Linton, O.B. and Härdle, W.
Estimation of Additive Regression Models with known Links.

$$T_{1n} = \frac{1}{n}\sum_{i=1}^{n} G'\{m(x_1, X_{2i})\}\frac{E_i(\widehat{a}_i)}{p(x_1,X_{2i})}\{1+o_p(1)\} = \tilde{T}_{1n}\{1+o_p(1)\}$$

$$T_{2n} = \frac{1}{n}\sum_{i=1}^{n} G'\{m(x_1, X_{2i})\}\frac{\widehat{a}_i - E_i(\widehat{a}_i)}{p(x_1,X_{2i})}\{1+o_p(1)\} = \tilde{T}_{2n}\{1+o_p(1)\},$$

since $\max_{i\leq n}|\frac{\widehat{p}_i - p_i}{p_i p_i}| = o_p(1)$, where $p_i = p(x_1, X_{2i})$, by standard uniform convergence arguments, such as Silverman (1978). This follows from assumptions A4 and A6-A9; in particular, we require $n^{4/5}g^{d-1} \to \infty$.

It remains to work with the approximations $\tilde{T}_{1n}$ and $\tilde{T}_{2n}$. We deal first with the bias term

$$\tilde{T}_{1n} = \frac{1}{n}\sum_{i=1}^{n} G'\{m(x_1, X_{2i})\}\frac{E_i(\widehat{a}_i)}{p_i} = O_p(h^2) + O_p(g^q),$$

where $p(x_1, X_{2i})^{-1}E_i(\widehat{a}_i)$ is an approximation to the conditional bias of the Nadaraya-Watson estimator at (x_1, X_{2i}). Its behaviour is well known from regression analysis. Therefore, the single sum $\tilde{T}_{1n}$ converges, on standardisation, to its population mean by Chebychev's Law of Large Numbers that holds by A1,A3-A9. In order for the $O_p(g^q)$ bias term not to show up in the asymptotic approximation, it is necessary that $n^{2/5}g^q \to 0$.

We now turn to the stochastic term

$$\tilde{T}_{2n} = \frac{1}{n}\sum_{i=1}^{n} G'\{m(x_1, X_{2i})\}\frac{\widehat{a}_i - E_i(\widehat{a}_i)}{p(x_1,X_{2i})}$$

We further write

$$\widehat{a}_i - E_i(\widehat{a}_i) = \widehat{a}_i - E_*(\widehat{a}_i) + E_*(\widehat{a}_i) - E_i(\widehat{a}_i),$$

with E_* denoting expectation conditional on the design, and

$$\widehat{a}_i - E_*(\widehat{a}_i) = n^{-1}\sum_{j=1}^{n} K_h(x_1 - X_{1j})L_g(X_{2i} - X_{2j})\varepsilon_j$$

and hence

$$\frac{1}{n}\sum_{i=1}^{n} G'\{m(x_1,X_{2i})\}\frac{\widehat{a}_i - E_*(\widehat{a}_i)}{p(x_1,X_{2i})}$$

$$= \frac{1}{n}\sum_{i=1}^{n} G'\{m(x_1,X_{2i})\}\frac{1}{p(x_1,X_{2i})} n^{-1}\sum_{j=1}^{n} K_h(x_1-X_{1j})L_g(X_{2i}-X_{2j})\varepsilon_j$$

$$= n^{-1}\sum_{j=1}^{n} K_h(x_1-X_{1j})\varepsilon_j\left[\frac{1}{n}\sum_{i=1}^{n} G'\{m(x_1,X_{2i})\}\frac{1}{p(x_1,X_{2i})}L_g(X_{2i}-X_{2j})\right]$$

$$= n^{-1}\sum_{j=1}^{n} K_h(x_1-X_{1j})\varepsilon_j G'\{m(x_1,X_{2j})\}\frac{p_2(X_{2j})}{p(x_1,X_{2j})}\{1+o_p(1)\},$$

by the same uniform convergence arguments as above. The term

$$\widetilde{T}_{3n} = n^{-1}\sum_{j=1}^{n} K_h(x_1-X_{1j})\varepsilon_j G'\{m(x_1,X_{2j})\}\frac{p_2(X_{2j})}{p(x_1,X_{2j})}$$

provides the asymptotic variance of the estimator; it is $O_p(n^{-1/2}h^{-1/2})$, since only smoothing with respect to X_1 is present. Moreover, $\widetilde{T}^*_{2n}$ satisfies the Lindberg-Feller Central Limit Theorem by virtue of A1-A9, see Härdle (1990).

We now turn to the term involving $E_*(\widehat{a}_i) - E_i(\widehat{a}_i)$. This double sum is

$$\frac{1}{n}\sum_{i=1}^{n} G'\{m(x_1,X_{2i})\}\frac{E_*(\widehat{a}_i)-E_i(\widehat{a}_i)}{p(x_1,X_{2i})} = \sum_{i=1}^{n}\sum_{j=1}^{n}\widetilde{\zeta}_{ij} = \sum_i \widetilde{\zeta}_{ii} + \sum\sum_{i\neq j}\widetilde{\zeta}_{ij},$$

where $\widetilde{\zeta}_{ij} = \zeta_{ij} - \overline{\zeta}_i$ and $\overline{\zeta}_i = E_i(\zeta_{ij})$, with

$$\zeta_{ij} = \frac{1}{n^2}\frac{G'\{m(x_1,X_{2i})\}}{p(x_1,X_{2i})}K_h(x_1-X_{1j})L_g(X_{2i}-X_{2j})\{m(X_{1i},X_{2i})-m(x_1,X_{2j})\}.$$

This double sum has mean zero. When $i=j$, we have

$$\zeta_{ii} = \frac{1}{n^2 g^{d-1}}L(0)\frac{G'\{m(x_1,X_{2i})\}}{p(x_1,X_{2i})}K_h(x_1-X_{1i})\{m(X_{1i},X_{2i})-m(x_1,X_{2i})\}.$$

and the single sum $\sum_i \widetilde{\zeta}_{ii}$ is $O_p\left\{(nh)^{-1/2}(ng^{d-1})^{-1/2}\right\}$. We now calculate the variance of the double sum $\sum\sum_{i\neq j}\widetilde{\zeta}_{ij}$, which involves the following calculations

(1996) Linton, O.B. and Härdle, W.
Estimation of Additive Regression Models with known Links.

$$\sum_{i\neq j}\sum \text{var}(\zeta_{ij}),\ \sum_{i\neq j}\sum E(\tilde{\zeta}_{ij}\tilde{\zeta}_{ji}),\ \sum_{i\neq j,i\neq k,j\neq k}\sum\sum E(\tilde{\zeta}_{ij}\tilde{\zeta}_{ik}),$$

since all other terms are mean zero by a conditioning argument. Now

$$E(\zeta_{ij}\zeta_{ik}) = E[E_j^2(\zeta_{ij})],$$

for $i \neq j, i \neq k, j \neq k$ using conditional independence. But

$$\begin{aligned} E_j(\zeta_{ij}) &= \tfrac{1}{n^2}K_h(x_1 - X_{1j})E_j\left[\tfrac{G'\{m(x_1,X_{2i})\}}{p(x_1,X_{2i})}L_g(X_{2i}-X_{2j})\left\{m(X_{1i},X_{2i}) - m(x_1,X_{2j})\right\}\right] \\ &= \tfrac{1}{n^2}K_h(x_1 - X_{1j})O(h^2+g^q), \end{aligned}$$

and so $\sum_{i\neq j,i\neq k,j\neq k}\sum\sum E(\zeta_{ij}\zeta_{ik}) = O(n^{-1}h^{-1})O(h^4+g^{2q})$. The other calculations follow similarly.

Finally, the uniform consistency of $\widehat{m}$ and the conditions on G'' ensure that $n^{-1}\sum G''\{\widehat{m}^*\} = O_p(1)$. □

PROOF OF THEOREM 2. By Taylor expansion,

$$\widetilde{m}(x) - m(x) = F'[G\{m(x)\}]\left[\sum_{\alpha=1}^{d}\left\{\tilde{f}_\alpha(x_\alpha) - f_\alpha(x_\alpha)\right\} + \tilde{c} - c\right] + R',$$

where the remainder R' depends on the second derivatives of F. The same arguments used above then apply. □

REFERENCES

[1] BREIMAN, L. AND J.H. FRIEDMAN (1985). Estimating optimal transformations for multiple regression and correlation, (with discussion). *J. Am. Statist. Assoc.* **80**, 580-619.

[2] BUJA, A., HASTIE, T. & R. TIBSHIRANI. (1989). Linear smoothers and additive models (with discussion). *Ann. Stat.* **17**, 453-555.

[3] COX, D.R. (1983). Some remarks on overdispersion. *Biometrika* **70**, 269-74.

[4] CUZICK, J. (1992). Efficient estimates in semiparametric additive regression models with unknown error distribution. *Ann. Stat.* **20**, 1129-1136.

[5] FAN, J. (1992). Design adaptive nonparametric regression. *J. Am. Stat. Assoc.* **87**, 998-1004.

[6] FAN, J. N. HECKMAN, AND M. WAND (1995). Local polynomial kernel regression for Generalised linear models and quasi-likelihood functions. *J. Am. Stat. Assoc.* **90**, 141-150.

[7] HALL, P. (1989) On Projection pursuit regression. *Ann. Stat.* **17**, 573-588.

[8] HÄRDLE, W. (1990). *Applied Nonparametric Regression.* Econometric Monograph Series 19. Cambridge University Press.

[9] HÄRDLE, W. AND KOROSTELEV (1995). Search of significant directions in additive nonparametric regression. *Biometrika.* To appear.

[10] HÄRDLE, W. AND O.B. LINTON (1995) Nonparametric Regression. Vol. 12 of *The Encyclopedia of Statistical Science.* Eds S. Kotz, C. Read, and . John Wiley.

[11] HÄRDLE, W. AND MARRON, J.S. (1991). Bootstrap simultaneous errors bars for nonparametric regression. *Ann. Stat.* **19**, 778-796.

[12] HASTIE, T. AND TIBSHIRANI, R. (1991). Generalised Additive Models. *Chapman and Hall, London.*

[13] IBRAGIMOV, I.A. AND HASMINSKII, R.Z. (1980). On nonparametric estimation of regression, *Soviet Math. Dokl.*, **21**, 810-4.

[14] JONES, M.C. (1993) Do not weight for heteroskedasticity in nonparametric regression. *Aus. J. Stat.* **35**, 89-92.

[15] LEONTIEFF, W. (1947). Introduction to a theory of an internal structure of functional relationships. *Econometrica* **15**, 361-373.

[16] LINTON, O.B. AND J.P. NIELSEN. (1995). Estimating structured nonparametric regression by the kernel method. *Biometrika* **82**, 93-101.

[17] MCCULLAGH, P., AND J.A. NELDER (1989) *Generalized Linear Models* 2nd edition. Chapman and Hall.

[18] ROBINSON, P.M. (1988). Root-n consistent semiparametric regression. *Econometrica* **56**, 931-954.

[19] ROTHENBERG, T.J. (1972) *Efficient estimation with a priori information.* Cowles Foundation Monograph. New Haven, CT.

[20] SILVERMAN, B.W. (1978). Weak and strong uniform consistency of the kernel estimate of a density function and its derivatives. *Ann. Statist.* **6**, 177–184.

[21] STONE, C.J. (1980). Optimal rates of convergence for nonparametric estimators. *Ann. Stat.* **8**, 1348-1360.

[22] STONE, C.J. (1982). Optimal global rates of convergence for nonparametric regression. *Ann. Stat.* **8**, 1040-1053.

[23] STONE, C.J. (1985). Additive regression and other nonparametric models. *Ann. Stat.* **13**, 685-705.

[24] STONE, C.J. (1986). The dimensionality reduction principle for Generalised additive models. *Ann. Stat.* **14**, 592-606.

[25] TJØSTHEIM, D., AND AUESTAD (1994). Nonparametric identification of nonlinear time series: projections. *J. Am. Stat. Assoc.* **89**, 1398-1409.

Biometrika, 83, 541-549

Search of Significant Variables in Nonparametric Additive Regression

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

A. Korostelev*
Institute for System Analysis
Prospekt 60-Let Oktjabrja, 9
Moscow 117312, Russia

Abstract

Nonparametric additive regression is studied under the assumption that only a subset of nonparametric components is separated away from zero. Each of these non-zero components depends on its own particular explanatory variable called a significant variable. The search problem for significant variables is considered and the algorithm is proposed which guarantees the exponentially fast decreasing error probabilities as the sample size grows. We show that it is reasonable to use a rough estimator rather than to estimate the nonparametric components with the fastest possible rate.

*The research described in this paper was made possible in part by Grant MCF000 from the International Scientific Foundation. The research for this paper was carried out within Sonderforschungsbereich 373 at the Humboldt-University Berlin and was printed using funds made available by the Deutsche Forschungsgemeinschaft. Support of INRA, INSEE, Paris, France is gratefully acknowledged.

Biometrika, 83, 541-549

1 Introduction

Consider a nonparametric regression model with a d-dimensional explanatory variable $X = (X^{(1)}, \ldots, X^{(d)})$ and a one-dimensional response function Y. Assume that the regression function

$$m(X) = M(X^{(1)}, \ldots, X^{(d)}) = E[Y|X] = \sum_{j=1}^{d} g_j(X^{(j)})$$

is a sum of nonparametric components g_j each depending on one particular explanatory variable $X^{(j)}$.

Let $\{(X_i, Y_i)\}_{i=1}^n$ be a sample of i.i.d. observations such that

$$Y_i = \sum_{j=1}^{d} g_j(X_i^{(j)}) + \epsilon_i, \qquad i = 1, \ldots, n, \tag{1.1}$$

where the errors ϵ_i's are conditionally i.i.d. given the design points X_i's and

$$E[\epsilon_i | X_1, \ldots, X_n] = 0, \quad i = 1, \ldots, n,$$

The additive regression model (1.1) is quite useful in the nonparametric regression theory since Stone (1985, 1986) proved that it can be estimated with the typical one-dimensional rate of convergence in estimating each function g_j. This rate is free of the dimension d and the "curse of dimensionality" is, thus, avoided. Further extensions of the additive regression were studied in Buja, Hastie and Tibshirani (1989), Huber (1985), Hall (1989), Hastie and Tibshirani (1990), Härdle et al. (1992), Härdle and Tsybakov (1993). In this paper we investigate the problem under an additional assumption on the structure of model (1.1).

We consider the situation where some of the predictor variables have no effect on the response. That is we assume that there is a set $J \in \{1, \ldots, d\}$ of indices such that $g_j(X^{(j)}) \not\equiv 0$ iff $j \in J$. The variables $X^{(j)}$ with $j \in J$ are called significant while the corresponding g_j's are significant functions (components). The number of the significant variables is usually much less than the total number of variables, i.e. $card(J) \ll d$. The case $J = \emptyset$ is not excluded from our consideration. Our goal is to detect these significant variables with the least error probabilities. The search is based on the sample

(1996) Härdle, W. and Korostelev, A.
Search of Significant Variables in Nonparametric Additive Regression.

Biometrika, 83, 541-549

$\{(X_i, Y_i)\}_{i=1}^n$, and an algorithm is proposed which guarantees exponentially fast decreasing error probabilities as $n \longrightarrow \infty$. The problem of choosing significant functions is important in practical situations where we often are confronted with too many predictor variables and are, thus, interested in reducing the statistical model (1.1) to a managable size. We will not suppose that the elements of X are independent. The selection problems for the significant variables in the situation of independent $X^{(j)}$ were studied in Härdle and Tsybakov (1994), Chen, Härdle, Linton and Serverance-Lossin (1995), Maljutov and Wynn (1994). In the latter paper exponentially fast decreasing errors were obtained by means of a sequential search algorithm.

Now we specify the model (1.1) more precisely.

(A1) The design points X_i's are i.i.d. with the continuous density $\varphi = \varphi(t^{(1)}, \ldots, t^{(d)})$ in the cube $K = [0,1]^d$; the density φ is separated away from zero.

(A2) The functions $g_j, j \in J$, are continuous, bounded, i.e.

$$||g_j||_\infty = \max_{0 \le t \le 1} |g_j(t)| \le g_0,$$

and

$$\int_0^1 g_j(t) f_j(t)\, dt = 0, \qquad j \in J,$$

where f_j is the marginal density of $X_i^{(j)}$, and g_0 is a known constant.

(A3)

$$E[\varepsilon_i^2 | X_1, \ldots, X_n] \le \sigma_0^2 < +\infty$$

where σ_0^2 is a given constant, $\sigma_0^2 > 0$.

To make the problem of search possible, the significant functions must be separated away from zero in some way. Let $\Phi(g)$ be a smooth functional of g such that

$$\Phi(g) \ge 0 \text{ and } \Phi(g) = 0 \quad \text{iff} \quad g \equiv 0.$$

Then we assume that the significant functions are defined by the restriction

$$\Phi(g_j) \ge c_0 \quad \text{if} \quad j \in J.$$

(1996) Härdle, W. and Korostelev, A.
Search of Significant Variables in Nonparametric Additive Regression.

Biometrika, 83, 541-549

In this paper we restrict ourselves to the two functional $\Phi(g)$ only: $\Phi(g)$ is either the sup–norm of g or the weighted L_2-norm. For this reason we do not specify the assumptions on the smoothness of this functional.

(B1) For any $j \in J$ the inequality holds

$$||g_j||_\infty \geq c_0$$

with a known positive constant c_0.

In Section 2 we give a tutorial example of testing hypotheses in the case of a one-dimensional location parameter. In Section 3 the search problem is considered under Assumption (B1). Another possibility comes from the restrictions in L_2-norm.

(B2) For any $j \in J$ the inequality holds

$$||g_j\sqrt{f_j}||_2^2 = \int_0^1 g_j^2(t)f_j(t)\,dt \geq c_1^2 > 0$$

with a known constant c_1.

The search under Assumption (B2) is also discussed briefly in Section 3.

Note that the smooth functional $\Phi(g_j)$ usually can be estimated consistently as $n \longrightarrow \infty$, e.g., this is the case under Assumptions (B1) or (B2). Thus, the corresponding estimate obtained from the observations can be used to test the null hypothesis: $\Phi(g_j) = 0$ versus the alternative: $\Phi(g_j) \geq c_0$ for each $j = 1, \ldots, d$. This would lead us to a consistent search of significant functions. Let $\hat{J}_n$ be the chosen set of significant variables. There is a temptation to estimate the functional $\Phi(g_j)$ with the fastest possible rate of convergence. Indeed, this was done in Härdle and Tsybakov (1994). Unfortunately, this does not always help to minimize the probability $Pr\{\hat{J}_n \neq J\}$. Our goal here is to minimize the error probabilities $Pr\{\hat{J}_n \neq J\}$ uniformly over the class of regressions satisfying (A1)–(A3) with a prescribed modulus of continuity.

As an illustrative example consider, $\Phi(g) = ||g||_2^2$. This smooth functional can be estimated root–n consistently, i.e. there is an estimator $\hat{\Phi} = \hat{\Phi}\{(X_1, Y_1), \ldots, (X_n, Y_n)\}$ such that uniformly in g

$$\lim_{n\to\infty} E[\sqrt{n}\{\hat{\Phi} - \Phi(g)\}] = 0,$$

Biometrika, 83, 541-549

$$\limsup_{n\to\infty} E[n\{\hat{\Phi} - \Phi(g)\}^2] \leq \sigma_\Phi^2$$

with some finite limiting variance σ_Φ^2.

Now, if our decision rule (based on estimation of $\Phi(g)$) is:

$$j \in \hat{J}_n \quad \text{iff} \quad \hat{\Phi} \geq c_0/2,$$

then the Chebyshev inequality guarantees that

$$\limsup_{n\to\infty} nPr\{\hat{J}_n \neq J\} \leq d\frac{4\sigma_\Phi^2}{c_0^2}$$

(cf. Härdle and Tsybakov (1994)). The error probability, thus, decreases reciprocally to the sample size. Our objective is not to estimate $\Phi(g)$, but rather to study tests based on rough estimators of $\Phi(g)$ whose error probabilities decrease exponentially fast in the sample size n. We will prove the following

THEOREM 1.1 *Let Assumptions (A1)–(A3), and (B1) hold. Then, there exists a decision rule $\hat{J}_n$ for significant variables and positive constants A_0 and A_1 independent of n such that*

$$Pr(\hat{J}_n \neq J) \leq dA_0 \exp(-nA_1), \qquad n \geq 1. \tag{1.2}$$

Biometrika, 83, 541-549

2 Basic idea

The idea of the inequality (1.2) is, in fact, very simple and can be seen from the following tutorial example. Let y_i be i.i.d. observations of a one-dimensional location parameter θ in the model

$$y_i = \theta + \epsilon_i, \quad E[\epsilon_i] = 0, \quad E[\epsilon_i^2] = \sigma_0^2 > 0, \quad i = 1, \ldots, n. \tag{2.1}$$

The null hypothesis: $\theta = 0$ versus the alternative: $g_0 \geq |\theta| \geq c_0 > 0$ can be tested based on the mean value $\bar{y}_n = n^{-1} \sum_{i=1}^n y_i$. The null hypothesis is rejected iff $\bar{y}_n \geq c_0/2$ in which case the Chebyshev inequality guarantees that the error probabilities are of the order $O(n^{-1})$ as $n \longrightarrow \infty$. Take a positive value z the final choice of which is made below. Define the truncation of observations y_i by

$$y_i^z = \begin{cases} z, & \text{if} \quad y_i > z \\ y_i, & \text{if} \quad |y_i| \leq z \\ -z, & \text{if} \quad y_i < -z. \end{cases}$$

Consider another test which rejects the null hypothesis iff

$$\bar{y}_n^z = n^{-1} \sum_{i=1}^n y_i^z \geq c_0/2.$$

LEMMA 2.1 *Let η_i be i.i.d. random variables and $|\eta_i| \leq \gamma$ almost surely, $\gamma > 0$. Then for any $c > \mu = |E\eta_i|$ there exists a positive constant $H = H(c) > 0$ such that*

$$Pr\{|n^{-1} \sum_{i=1}^n \eta_i| > c\} \leq 2 \exp\{-nH\}, \quad n \geq 1. \tag{2.2}$$

Proof:

For any $\alpha \geq 0$

$$Pr\{n^{-1} \sum_{i=1}^n \eta_i > c\} \leq Pr\{\sum_{i=1}^n \eta_i' > n(c-\mu)\} \leq \exp\{-\alpha(c-\mu)n\}(E\, e^{\alpha\eta_i'})^n \tag{2.3}$$

where $\eta_i' = \eta_i - E\eta_i$. Now we show that for any random variable η' such that $E\eta' = 0, \quad |\eta'| \leq \gamma + \mu$, the inequality holds

$$E[e^{\alpha\eta'}] \leq \cosh(\alpha(\gamma + \mu)). \tag{2.4}$$

(1996) Härdle, W. and Korostelev, A.
Search of Significant Variables in Nonparametric Additive Regression.

Biometrika, 83, 541-549

Indeed, let η' be a discrete variable and $Pr\{\eta' = x\} = p_1 > 0$ at a point $x \in [0, \gamma + \mu)$. Consider some other random variable η'' which distribution coincides with η', but the mass p_1 is divided in half between the points $\gamma + \mu$ and $2x - (\gamma + \mu)$. Then,

$$E[e^{\alpha\eta''}] - E[e^{\alpha\eta'}] \geq \frac{p_1}{2} exp\{\alpha(\gamma + \mu)\} + \frac{p_1}{2} exp[\alpha\{2x - (\gamma + \mu)\}] - p_1 exp(\alpha x) > 0,$$

since $e^{\alpha x}$ is convex in x. The same is true if $x \in (-(\gamma + \mu), 0]$. Thus, the random variable η'' has a smaller mass inside the interval $(-(\gamma+\mu), \gamma+\mu)$ and a higher value of the expectation: $E[e^{\alpha\eta''}] > E[e^{\alpha\eta'}]$. Continuing this process, we come to the distribution concentrated at points $\pm(\gamma+\mu)$ with the masses which are to be equal, since the expected value of all the variables is zero. Finally, any continuous distribution can be approximated by a discrete one. This proves (2.4). Substituting (2.4) into (2.3), we come to

$$Pr\{n^{-1}\sum_{i=1}^{n}\eta_i > c\} \leq exp(-n[\alpha(c-\mu) - \log\cosh\{\alpha(\gamma+\mu)\}]).$$

Note that $\frac{\partial}{\partial\alpha}[\log\cosh(\alpha(\gamma+\mu))]\Big|_{\alpha=0} = 0$. Hence

$$H = \max_{\alpha\geq 0}[\alpha(c-\mu) - \log\cosh\{\alpha(\gamma+\mu)\}] > 0$$

and (2.2) follows. □

COROLLARY 2.1 *There exist a sufficiently large truncation level z and $H > 0$ such that*

$$Pr\left(\bar{y}_n^z \geq \frac{c_0}{2}|\theta = 0\right) \leq 2\,exp(-nH); \quad Pr\left(\bar{y}_n^z < \frac{c_0}{2}|\theta \geq c_0\right) \leq 2\,exp(-nH). \qquad (2.5)$$

Proof:

Note that $|E\,y_i^z| \leq \frac{\sigma_0^2}{z-g_0}$ for any θ, $|\theta| < g_0$, and (2.5) follows from (2.2) if we take $\eta_i = y_i^z$ and choose $z : \frac{\sigma_0^2}{z-g_0} < \frac{c_0}{2}$. □

REMARK 2.1 *The best choice of z is the maximizer of*

$$H = H(z) = \alpha\left(\frac{c_0}{2} - \frac{\sigma_0^2}{z-g_0}\right) - \log\cosh\left(z + \frac{\sigma_0^2}{z-g_0}\right)$$

over z : $z > g_0$, $\frac{\sigma_0^2}{z-g_0} < \frac{c_0}{2}$. Note that there is a trade-off in the choice of the truncation level z. Larger values of z reduce the bias term $\frac{\sigma_0^2}{z-g_0}$ but increase the bound $z + \frac{\sigma_0^2}{z-g_0}$ and vice versa.

(1996) Härdle, W. and Korostelev, A.
Search of Significant Variables in Nonparametric Additive Regression.

Biometrika, 83, 541-549

3 Rough bin estimators and decision rules

Take an integer M and divide the interval (0,1) into intervals

$$\triangle_k = \left(\frac{k-1}{M}, \quad \frac{k}{M}\right), \qquad k = 1, \ldots, M.$$

We estimate g_j by a bin piecewise constant estimator (see Härdle (1990))

$$\hat{g}_j(t^{(j)}): \quad \hat{g}_j(t^{(j)}) = \hat{g}_j(k/M) \text{ if } \quad t^{(j)} \in \triangle_k, \quad k = 1, \ldots, M,$$

and

$$\hat{g}_j(k/M) = \frac{M}{n} \sum_{i=1}^{n} Y_i^z \prod_{m \neq j} f_m(X_i^{(m)}) I(X_i^{(j)} \in \triangle_k)/\varphi(X_i), \quad j = 1, \ldots, d, \tag{3.1}$$

where Y_i^z's are truncations of observations (1.1). Define the decision rule by

$$\hat{J}_n = \{j : ||\hat{g}_j||_\infty \geq \frac{c_0}{2}\}. \tag{3.2}$$

As shown below, the rule (3.2) satisfies (1.2). This decision rule still depends on the joint density φ as well as on the marginal densities f_m. As we will see from the proof, estimates of φ and f_j can be plugged in.

LEMMA 3.1 *For any arbitrary small $\mu_0 > 0$ and for any k, $k = 1, \ldots, M$, there exist z and M such that*

$$|E\,\hat{g}_j(k/M) - g_j(k/M)| < \mu_0, \qquad j = 1, \ldots, d. \tag{3.3}$$

Proof:

We may consider $z = \infty$ in the right-hand side of (3.1) since the expected value of this expression tends with $z \longrightarrow \infty$ to

$$E\{M\, Y_j \prod_{m \neq j} f_m(X_i^{(m)})\, I(X_i^{(j)} \in \triangle_k)/\varphi(X_i)\} =$$

$$= M \int_K \{g_1(t^{(1)}) + \ldots + g_j(t^{(j)}) + \ldots + g_d(t^{(d)})\} \prod_{m \neq j} f_m(t^{(m)})\, I(t^{(j)} \in \triangle_k) dt^{(1)} \ldots dt^{(d)} \tag{3.4}$$

Biometrika, 83, 541-549

where the cube K is defined in (A1).

As follows from Assumption (A2),

$$\int_K g_l(t^{(l)}) \prod_{m \neq j} f_m(t^{(m)})\, I(t^{(j)} \in \triangle_k)\, dt^{(1)} \dots dt^{(d)} = 0$$

if $l \neq j, \quad l = 1, \dots, d$, and

$$\int_K g_j(t^{(j)}) \prod_{m \neq j} f_m(t^{(m)})\, I(t^{(j)} \in \triangle_k)\, dt^{(1)} \dots dt^{(d)} = \int_{\triangle_k} g_j(t^{(j)})\, dt^{(j)}.$$

Thus, the right-hand side of (3.4) equals $M \int_{\triangle_k} g_j(t^{(j)})\, dt^{(j)}$ and

$$|M \int_{\triangle_k} g_j(t^{(j)})\, dt^{(j)} - g_j(k/M)| \leq \max_{t \in \triangle_k} |g_j(t) - g_j(k/M)|.$$

Since the modulus of continuity of g_j is vanishing at zero, the lemma follows. □

Proof of Theorem 1.1:

For the decision rule (3.2) we have

$$Pr(\hat{J}_n \neq J) \leq \sum_{j=1}^{d} Pr(||\hat{g}_j - g_j||_\infty \geq c_0/2) \leq \sum_{j=1}^{d} \sum_{k=1}^{M} Pr(|\hat{g}_j(k/M) - g_j(k/M)| \geq c_0/2).$$

Applying Lemma 2.1 to the bounded random variables

$$\eta_i = M\, Y_i^z \prod_{m \neq j} f_m(X_i^{(m)})\, I(X_i^{(j)} \in \triangle_k)/\varphi(X_i)$$

and choosing $\mu_0 < c_0/2$ in Lemma 3.1, we come to the inequality

$$Pr(\hat{J}_n \neq J) \leq 2dM\ exp\{-nH\}.$$

This proves the theorem. □

REMARK 3.1 *A priori knowledge of the density φ is not crucial in (3.1) for it may be substituted by a proper estimate as well as the marginal densities f_j.*

Now we turn to the case of Assumption (B2) on the L_2-norms of significant variables. Note that

$$||g_j||_\infty^2 \geq \int_0^1 g_j^2(t) \left\{ \frac{f_j(t)}{||f_j||_\infty} \right\} dt \geq \frac{c_1^2}{||f_j||_\infty} = c_0^2 > 0,$$

(1996) Härdle, W. and Korostelev, A.
Search of Significant Variables in Nonparametric Additive Regression.

Biometrika, 83, 541-549

and the decision rule (3.2) is applicable to this case.

In some applications it is interesting to derive the decision rules from a proper estimator of the functional $\Phi_j(g_j) - ||g_j\sqrt{f_j}||_2^2$. Let n be even. Split the whole sample $\{(X_i, Y_i)\}_{i=1}^n$ into the parts with odd and even indices. Let $\hat{g}_j^{odd}$ be the estimator of g_j obtained similarly to (3.1) from the odd subsample. Consider the following estimator:

$$\hat{\Phi}_j = \frac{4}{n}\sum_{i=1}^{n/2} \hat{g}_j^{odd}(X_{2i}^{(j)}) \prod_{m=1}^{d} f_m(X_{2i}^{(m)})\, Y_{2i}^z/\varphi(X_{2i}) - \int_0^1 \{\hat{g}_j^{odd}(t)\}^2 f_j(t)\, dt. \qquad (3.5)$$

The idea of (3.5) is that the expected value of the sum in the right-hand side asymptotically equals

$$2\int_0^1 \hat{g}_j^{odd}(t)\, g_j(t)\, f_j(t)\, dt, \qquad \text{as } z \longrightarrow \infty.$$

The bias term

$$E[\hat{\Phi}_j] - \Phi_j(g_j) \approx -||(\hat{g}_j^{odd} - g_j)\sqrt{f_j}||_2^2$$

in practical implementations may be smaller than that in the uniform norm. The exponential bounds on the error probabilities in this case are, of course, the same as above.

References

Buja, A.; Hastie, T.J. and Tibshirani, R.J. (1989) *Linear smoothers and additive models (with discussion).* The Annals of Statistics, 17, 453-555.

Cheng, R. and Härdle, W. (1994) *Estimation and Variable Selection in Generalized Additive Models*, Manuscript.

Hall, P. (1989) *On projection pursuit regression.* The Annals of Statistics, 17, 573-588.

Härdle, W. (1990) *Applied Nonparametric Regression. Econometric Monograph Series 19*, Cambridge University Press.

Härdle, W.; Hart, J.; Marron, J.S. and Tsybakov, A.B. (1992) *Bandwidth Choice for Average derivative estimation.* Journal of the American Statistical Association, 87, 817-823.

Biometrika, 83, 541-549

Härdle, W. and Tsybakov, A.B.; (1993) *How sensitive are average derivatives?* Journal of Econometrics, 58, 31-48.

Härdle, W. and Tsybakov, A.B. (1994) *Additive Nonparametric Regression on Principal Components*, Journal of Nonparametric Statistics, accepted.

Hastie, T.J. and Tibshirani, R.J. (1990) *Generalized Additive Models*, Chapman and Hall, London.

Maljutov, M.B. and Wynn, H.P. (1994) *Sequential Screening of Significant Variables of an Additive Model. In: Markov Process and Applications.* Dynkin's Festschrift. Birkhäuser Progress, 253-265.

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.

(1996) Härdle, W. and Korostelev, A.
Search of Significant Variables in Nonparametric Additive Regression.

Journal of the American Statistical Association, 91, 1632-1640

Direct Semiparametric Estimation of Single-Index Models With Discrete Covariates

Joel L. HOROWITZ and Wolfgang HÄRDLE

Others have developed average derivative estimators of the parameter β in the model $E(Y|\mathbf{X}=\mathbf{x}) = G(\mathbf{x}\beta)$, where G is an unknown function and $\mathbf{X}$ is a random vector. These estimators are noniterative and easy to compute but require that $\mathbf{X}$ be continuously distributed. This article develops a noniterative, easily computed estimator of β for models in which some components of $\mathbf{X}$ are discrete. The estimator is $n^{1/2}$ consistent and asymptotically normal. An application to data on product innovation by German manufacturers illustrates the estimator's usefulness.

KEY WORDS: Average derivative estimation; Index model.

1. INTRODUCTION

A single-index mean-regression model has the form

$$E(Y|\mathbf{X}=\mathbf{x}) = G[v(\mathbf{x},\beta)], \tag{1}$$

where Y is a scalar-dependent variable, $\mathbf{X}$ is a vector of explanatory variables, β is a vector of parameters whose values are unknown, $v(\cdot,\cdot)$ is a known function, and G is a function that may or may not be known. Many widely used parametric models have this form; examples include linear regression, binary logit and probit, and tobit models. These models assume that G is known up to a finite-dimensional parameter. When G is unknown, (1) provides a specification that is more flexible than a parametric model while avoiding the loss of precision that occurs in fully nonparametric estimation with a multidimensional $\mathbf{x}$. In most applications $v(\mathbf{x},\beta) = \mathbf{x}\beta$, where $\mathbf{x}'$ and β are $k \times 1$ vectors. Thus

$$E(Y|\mathbf{X}=\mathbf{x}) = G(\mathbf{x}\beta). \tag{2}$$

This article is concerned with estimating β in (2) when G is unknown.

Several estimators of β that do not require a parametric specification of G already exist. Ichimura (1993) developed a semiparametric least squares estimator of β. This estimator is closely related to projection pursuit regression (Friedman and Stuetzle 1981). Han (1987) and Sherman (1993) described a maximum rank correlation estimator. Klein and Spady (1993) developed a quasi–maximum likelihood estimator for the case in which Y is binary. This estimator achieves the asymptotic efficiency bound of Cosslett (1987) if G is a distribution function. The estimators of Ichimura, Han, Klein and Spady, and Sherman are $n^{1/2}$ consistent and asymptotically normal under regularity conditions.

The foregoing estimators have the disadvantage of being difficult to compute because they require solving nonlinear optimization problems whose objective functions are not necessarily concave (convex, in the case of semiparametric least squares) or unimodal. If $\mathbf{X}$ is a continuous random variable, then the computational difficulty of estimating β can be greatly reduced through the use of average-derivative estimators. These estimators rely on the fact that for any weight function $w(\cdot)$, $E[w(\mathbf{X})\partial G(\mathbf{X}\beta)/\partial \mathbf{X}] \propto \beta$. Average-derivative estimation does not require solving an optimization problem, and computation of average-derivative estimates is noniterative and fast. The estimators are $n^{1/2}$ consistent and asymptotically normal under regularity conditions (Härdle and Stoker 1989; Powell, Stock, and Stoker 1989; Stoker 1991). Ai (1991) discussed the case in which $v(\mathbf{x},\beta)$ is nonlinear in $\mathbf{x}$.

Average-derivative methods cannot be used to estimate components of β that multiply discrete components of $\mathbf{X}$. This is because derivatives of $\mathbf{G}(\mathbf{X}\beta)$ with respect to discrete components of $\mathbf{X}$ are not identified. Because $\mathbf{X}$ has discrete components in many applications, a direct (noniterative) method for estimating β when $\mathbf{X}$ has such components is needed. This article develops such a method. The resulting noniterative estimator is much easier to compute than estimators that require solving nonlinear optimization problems.

Section 2 describes the estimator and its properties. Section 3 presents the results of a Monte Carlo investigation of the estimator's finite-sample properties, and Section 4 illustrates the use of the estimator by applying it to data on product innovation by German manufacturers. Section 5 presents concluding comments. The Appendix gives all proofs.

2. THE ESTIMATOR

To distinguish between continuous and discrete covariates, we rewrite (2) in the form

$$E(Y|\mathbf{X}=\mathbf{x},\mathbf{Z}=\mathbf{z}) = G(\mathbf{x}\beta + \mathbf{z}\alpha), \tag{3}$$

where $\mathbf{X}$ denotes a $1 \times k$ vector of continuous random variables, $\mathbf{Z}$ denotes a $1 \times l$ vector of discrete random variables, and β and α are conformable vectors of parameters that must be estimated from data. Identification of β and α requires that (3) have at least one continuous explanatory

Joel L. Horowitz is Professor, Department of Economics, University of Iowa, Iowa City, IA 52242. Wolfgang Härdle is Professor, Institute for Statistics and Econometrics, Humboldt University, 10178 Berlin, Germany. This research was supported in part by Deutsche Forschungsgemeinschaft, Sonderforschungsbereich 373, "Quantifikation und Simulation Öknonomischer Prozesse." The research of Joel L. Horowitz was supported in part by National Science Foundation grants DMS-9208820 and SBR-9307677. The authors thank N. E. Savin for comments on this research.

Journal of the American Statistical Association
December 1996, Vol. 91, No. 436, Theory and Methods

Journal of the American Statistical Association, 91, 1632-1640

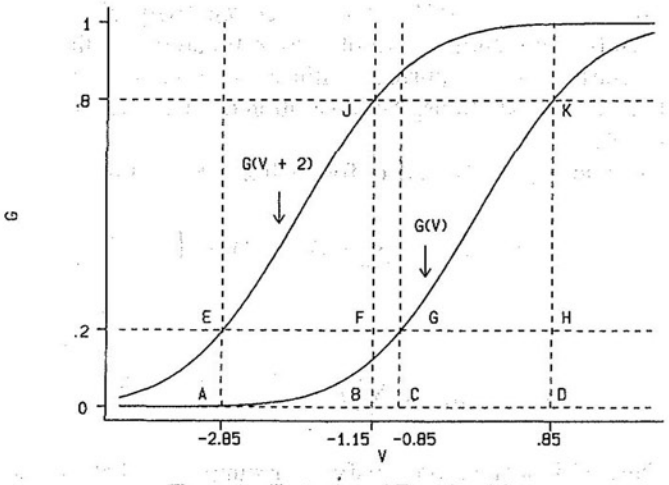


Figure 1. Illustration of Equation (5).

variable (Ichimura 1993; Klein and Spady 1993; Manski 1988), so $k \geq 1$. Bierens and Hartog (1988) discussed the case $k = 0$ in detail. There do not have to be any discrete explanatory variables in (3), but we assume that there is at least one, because the focus of this article is on estimating $\boldsymbol{\alpha}$. Thus $l \geq 1$. Because $\boldsymbol{\beta}$ and $\boldsymbol{\alpha}$ are identified only up to sign and scale, sign and scale normalizations are needed. We use $\beta_1 = 1$, where β_1 is the coefficient of the first component of $\mathbf{X}$. Let $\tilde{\mathbf{X}}$ and $\tilde{\boldsymbol{\beta}}$ denote components 2 through k of $\mathbf{X}$ and $\boldsymbol{\beta}$, if $k > 1$.

The main problem to be solved here is estimating $\boldsymbol{\alpha}$. The parameter $\boldsymbol{\beta}$ can be estimated using existing methods. For example, one can use average-derivative methods to estimate $\boldsymbol{\beta}$ for each $\mathbf{z}$ in the support of $\mathbf{Z}$ and then form a (possibly weighted) average of these estimators. Accordingly, in the remainder of this section we concentrate on estimating $\boldsymbol{\alpha}$.

2.1 Informal Description of the Estimator

The essential idea of our estimator of $\boldsymbol{\alpha}$ can be understood most easily by assuming for the moment that G is a known function. Define $S_{\mathbf{Z}} \equiv \{\mathbf{z}^{(i)}: i = 1, \ldots, M\}$ to be the support of the discrete random vector $\mathbf{Z}$. Our estimator works by deducing the horizontal distance between $G(v + \mathbf{z}^{(i)}\boldsymbol{\alpha})$ and $G(v + \mathbf{z}^{(1)}\boldsymbol{\alpha})$ $(i = 2, \ldots, M)$ on a set of v values on which $G(v + \mathbf{z}\boldsymbol{\alpha})$ is assumed to satisfy a weak monotonicity condition. Specifically, we assume that there are finite numbers v_0, v_1, c_0, and c_1 such that $v_0 < v_1, c_0 < c_1$, and $G(v + \mathbf{z}\boldsymbol{\alpha}) < c_0$ for each $\mathbf{z} \in S_{\mathbf{Z}}$ if $v < v_0$, and $G(v + \mathbf{z}\boldsymbol{\alpha}) > c_1$ for each $\mathbf{z} \in S_{\mathbf{Z}}$ if $v > v_1$. To ensure that $G(v + \mathbf{z}\boldsymbol{\alpha})$ is identified on $v_0 \leq v \leq v_1$, we also assume that for each $\mathbf{z} \in S_{\mathbf{Z}}$, the density of $\mathbf{X}\boldsymbol{\beta}$ conditional on $\mathbf{Z} = \mathbf{z}$ exceeds zero everywhere on $[v_0, v_1]$. To see the implications of these assumptions for estimation of $\boldsymbol{\alpha}$, let $I(\cdot)$ denote the indicator function. For $\mathbf{z} \in S_{\mathbf{Z}}$, define

$$J(\mathbf{z}) = \int_{v_0}^{v_1} \{c_0 I[G(v + \mathbf{z}\boldsymbol{\alpha}) < c_0] + c_1 I[G(v + \mathbf{z}\boldsymbol{\alpha}) > c_1] + G(v + \mathbf{z}\boldsymbol{\alpha}) I[c_0 \leq G(v + \mathbf{z}\boldsymbol{\alpha}) \leq c_1]\}\, dv. \quad (4)$$

The key fact that leads to our estimator is stated in the following equation, which is proved in Lemma 1 of the Appendix:

$$J[\mathbf{z}^{(i)}] - J[\mathbf{z}^{(1)}] = (c_1 - c_0)[\mathbf{z}^{(i)} - \mathbf{z}^{(1)}]\boldsymbol{\alpha}; \quad i = 2, \ldots M. \quad (5)$$

Figure 1 gives a graphical explanation of (5) for a model in which $\mathbf{z}$ is a scalar whose two possible values are $[\mathbf{z}^{(2)}, \mathbf{z}^{(1)}] = (1, 0)$, and $\boldsymbol{\alpha} = 2$. Let $(c_0, c_1) = (.2, .8)$, and $(v_0, v_1) = (-2.85, .85)$. The integrands of $J[\mathbf{z}^{(1)}]$ and $J[\mathbf{z}^{(2)}]$ are EFGK and EJK. $J[\mathbf{z}^{(2)}]$ is the area EFJ + ABFE + BDKJ = EFJ + $1.7c_0 + 2c_1$. $J[\mathbf{z}^{(1)}]$ is the area ACGE + CDHG + GHK = $2c_0 + 1.7c_0$ + GHK. But EFJ = GHK, so $J[\mathbf{z}^{(2)}] - J[\mathbf{z}^{(1)}] = 2(c_1 - c_0) = (c_1 - c_0)[\mathbf{z}^{(2)} - \mathbf{z}^{(1)}]\boldsymbol{\alpha}$.

Equation (5) constitutes $M - 1$ linear equations in the l unknown components of $\boldsymbol{\alpha}$. These equations may be solved for $\boldsymbol{\alpha}$ if a unique solution exists. To do this, define the $(M - 1) \times 1$ vector $\boldsymbol{\Delta}\mathbf{J}$ by

$$\boldsymbol{\Delta}\mathbf{J} = \begin{bmatrix} J[\mathbf{z}^{(2)}] - J[\mathbf{z}^{(1)}] \\ \cdots \\ J[\mathbf{z}^{(M)}] - J[\mathbf{z}^{(1)}] \end{bmatrix}. \quad (6)$$

Also, define the $(M - 1) \times l$ matrix $\mathbf{W}$ by

$$\mathbf{W} = \begin{bmatrix} \mathbf{z}^{(2)} - \mathbf{z}^{(1)} \\ \cdots \\ \mathbf{z}^{(M)} - \mathbf{z}^{(1)} \end{bmatrix}. \quad (7)$$

Then it can be proved (see Lemma 1 of the Appendix) that if $\mathbf{W}'\mathbf{W}$ is a nonsingular matrix, then

$$\boldsymbol{\alpha} = (c_1 - c_0)^{-1}(\mathbf{W}'\mathbf{W})^{-1}\mathbf{W}'\boldsymbol{\Delta}\mathbf{J}. \quad (8)$$

Equation (8) forms the basis for our estimator of $\boldsymbol{\alpha}$. Of course, (8) cannot be used directly in estimation, because $G(v + \mathbf{z}\boldsymbol{\alpha})$, and thus $\boldsymbol{\Delta}\mathbf{J}$, are not known in applications. We solve this problem by replacing $G(v + \mathbf{z}\boldsymbol{\alpha})$ in (4) with a nonparametric regression estimator of $E(Y|\mathbf{X}\mathbf{b}_n = v, \mathbf{Z} = \mathbf{z})$, where $\mathbf{b}_n$ is the estimator of $\boldsymbol{\beta}$. We use a kernel estimator because it is relatively easy to analyze and implement, but other estimators could be used. Denote the estimator of $G(v + \mathbf{z}\boldsymbol{\alpha})$ by $G_{nz}(v)$. The estimator of $\boldsymbol{\alpha}$ is

$$\boldsymbol{\alpha}_n = (c_1 - c_0)^{-1}(\mathbf{W}'\mathbf{W})^{-1}\mathbf{W}'\boldsymbol{\Delta}\mathbf{J}_n, \quad (9)$$

where

$$\boldsymbol{\Delta}\mathbf{J}_n = \begin{bmatrix} J_n[\mathbf{z}^{(2)}] - J_n[\mathbf{z}^{(1)}] \\ \cdots \\ J_n[\mathbf{z}^{(M)}] - J_n[\mathbf{z}^{(1)}] \end{bmatrix}, \quad (10)$$

and for each $\mathbf{z} \in S_{\mathbf{Z}}$,

$$J_n(\mathbf{z}) = \int_{v_0}^{v_1} \{c_0 I[G_{nz}(v) < c_0] + c_1 I[G_{nz}(v) > c_1] + G_{nz}(v) I[c_0 \leq G_{nz}(v) \leq c_1]\}\, dv. \quad (11)$$

These ideas are formalized in Section 2.2, where we give conditions under which $\boldsymbol{\alpha}_n$ is consistent for $\boldsymbol{\alpha}$ and $n^{1/2}(\boldsymbol{\alpha}_n - \boldsymbol{\alpha})$ is asymptotically normal.

(1996) Horowitz, J. and Härdle, W.
Direct semiparametric estimation of single-index models with discrete covariates.

Journal of the American Statistical Association, 91, 1632-1640

2.2 Assumptions and Results

We begin this section by presenting our assumptions. Let S_v denote the support of the distribution of $V \equiv \mathbf{X}\beta$. Let $f(v|\mathbf{z})$ be the probability density of V conditional on $\mathbf{Z} = \mathbf{z}$, let $p(v, \tilde{\mathbf{x}}|\mathbf{z})$ be the joint density of $(V, \tilde{\mathbf{X}})$ conditional on $\mathbf{Z} = \mathbf{z}$, let $p(\mathbf{z})$ be the probability that $\mathbf{Z} = \mathbf{z}$ ($\mathbf{z} \in S_{\mathbf{Z}}$), and let $f(v, \mathbf{z}) = f(v|\mathbf{z})p(\mathbf{z})$. Let $\{Y_i, \mathbf{X}_i, \mathbf{Z}_i: i = 1, \ldots, n\}$ be a random sample of size n of $\{Y, \mathbf{X}, \mathbf{Z}\}$, and set $V_i = \mathbf{X}_i\beta$. Let $r \geq 4$ be an integer and $\|\cdot\|$ denote the Euclidean norm.

Assumption 1.

a. $S_{\mathbf{Z}}$ is a finite set.
b. $E(\|\tilde{\mathbf{X}}\|^2|\mathbf{Z} = \mathbf{z}) < \infty$ and $E(|Y|\|\tilde{\mathbf{X}}\|^2|\mathbf{Z} = \mathbf{z}) < \infty$ for each $\mathbf{z} \in S_{\mathbf{Z}}$.
c. $E(|Y|^2\|\tilde{\mathbf{X}}\|^2|V = v, \mathbf{Z} = \mathbf{z}), E(|Y|^2|V = v, \mathbf{Z} = \mathbf{z})$, and $f(v, \mathbf{z})$ are bounded uniformly over $v \in [v_0 - \varepsilon, v_1 + \varepsilon]$ for some $\varepsilon > 0$ and all $\mathbf{z} \in S_{\mathbf{Z}}$.
d. For each $\mathbf{z} \in S_{\mathbf{Z}}, p(v, \tilde{\mathbf{x}}|\mathbf{z})$ is everywhere three times continuously differentiable with respect to v; the third derivative is bounded uniformly over $(v, \tilde{\mathbf{x}})$.
e. $\text{var}(Y|V = v, \mathbf{Z} = \mathbf{z}) > 0$ for all $\mathbf{z} \in S_{\mathbf{Z}}$ and almost every v.

The requirement that $S_{\mathbf{Z}}$ be finite can always be satisfied by truncating the distribution of $\mathbf{Z}$.

Assumption 2. Define $\mathbf{W}$ as in (7). $\mathbf{W}'\mathbf{W}$ is nonsingular.

Assumption 3. $E(Y|\mathbf{X} = \mathbf{x}, \mathbf{Z} = \mathbf{z}) = G(\mathbf{x}\beta + \mathbf{z}\alpha)$. $G(\cdot)$ is r times continuously differentiable ($r \geq 4$). $G(\cdot)$ and its first r derivatives are bounded on all bounded intervals.

Assumption 3 makes $E(Y|\mathbf{X} = x, \mathbf{Z} = z)$ a single-index model. The smoothness requirements are standard in nonparametric estimation.

Assumption 4. There are finite numbers v_0, v_1, c_0, and c_1 such that $v_0 < v_1, c_0 < c_1$, and $G(v) = c_0$ or c_1 at only finitely many values of v, and for each $\mathbf{z} \in S_{\mathbf{Z}}$:

a. $G(v + \mathbf{z}\alpha) < c_0$ if $v < v_0$
b. $G(v + \mathbf{z}\alpha) > c_1$ if $v > v_1$
c. $f(v|\mathbf{z})$ is bounded away from zero on an open interval containing $[v_0, v_1]$.

The purpose of this assumption is explained in the discussion of Equation (4). The results presented here hold with obvious modifications if $c_0 > c_1, G(v + \mathbf{z}\alpha) > c_0$ for all $\mathbf{z} \in S_{\mathbf{Z}}$ if $v < v_0$, and $G(v + \mathbf{z}\alpha) < c_1$ for all $\mathbf{z} \in S_{\mathbf{Z}}$ if $v > v_1$.

Assumption 5. If $k > 1$, there are (a) a $n^{1/2}$-consistent estimator of $\tilde{\beta}$, denoted by $\tilde{\mathbf{b}}_n$, and (b) a $(k-1) \times 1$ vector-valued function $\boldsymbol{\Omega}(y, \mathbf{x}, \mathbf{z})$ satisfying

$$n^{1/2}(\tilde{\mathbf{b}}_n - \tilde{\beta}) = n^{-1/2}\sum_{i=1}^{n} \boldsymbol{\Omega}(Y_i, \mathbf{X}_i, \mathbf{Z}_i) + o_p(1)$$

as $n \to \infty$, where $E\boldsymbol{\Omega}(Y, \mathbf{X}, \mathbf{Z}) = 0$ and $n^{-1/2}\sum_{i=1}^{n} \boldsymbol{\Omega}(Y_i, \mathbf{X}_i, \mathbf{Z}_i) \xrightarrow{d} N(0, \mathbf{V}_{\boldsymbol{\Omega}})$ for some finite matrix $\mathbf{V}_{\boldsymbol{\Omega}}$.

The estimators of Härdle and Stoker (1989), Powell et al. (1989), and Stoker (1991) satisfy this assumption, after sign and scale normalization, under regularity conditions given by these authors. All of these estimators are direct in the sense of not requiring nonlinear optimization or other iterative computations. An illustration of $\boldsymbol{\Omega}$ is given in Section 2.3.

Define $\hat{V}_i = \mathbf{X}_i\mathbf{b}_n$. Also, for each $\mathbf{z} \in S_{\mathbf{Z}}$, define

$$A_{nz}(v) = (nh_n)^{-1}\sum_{i=1}^{n} I(\mathbf{Z}_i = \mathbf{z})Y_iK\left(\frac{v - \hat{V}_i}{h_n}\right)$$

and

$$f_{nz}(v) = (nh_n)^{-1}\sum_{i=1}^{n} I(\mathbf{Z}_i = \mathbf{z})K\left(\frac{v - \hat{V}_i}{h_n}\right).$$

where K is a function satisfying assumption 6 below, and $\{h_n\}$ is a sequence of positive real numbers satisfying assumption 7. Set

$$G_{nz}(v) = A_{nz}(v)/f_{nz}(v),$$

Assumption 6. K is a bounded, symmetrical, differentiable function that is nonzero only on $[-1, 1]$. $K'(\cdot)$, the derivative of K, is Lipschitz continuous. For each integer i between 0 and $r (r \geq 4)$:

$$\int_{-1}^{1} v^iK(v)\,dv = \begin{cases} 1 & \text{if } i = 0, \\ 0 & \text{if } 1 < i < r, \\ \text{nonzero} & \text{if } i = r. \end{cases}$$

Assumption 7. As $n \to \infty, nh_n^{r+3} \to \infty$ and $nh_n^{2r} \to 0$.

A higher order kernel ($r \geq 4$) with undersmoothing is needed to prevent $n^{1/2}(\boldsymbol{\alpha}_n - \boldsymbol{\alpha})$ from being asymptotically biased. Higher order kernels with undersmoothing are used for similar reasons in average derivative estimation (Härdle and Stoker 1989; Powell et al. 1989) and estimation of semilinear regression models (Robinson 1988).

For each $\mathbf{z} \in S_{\mathbf{Z}}$, define $G_{\mathbf{z}}(v) = G(v + \mathbf{z}\alpha), G'_{\mathbf{z}}(v) = dG_{\mathbf{z}}(v)/dv$, and

$$\boldsymbol{\Gamma}_{\mathbf{z}} = -\int_{v_0}^{v_1} G'_{\mathbf{z}}(v)E(\tilde{\mathbf{X}}|v, \mathbf{z})I[c_0 \leq G(v + \mathbf{z}\alpha) \leq c_1]\,dv.$$

The following theorem shows that $\boldsymbol{\alpha}_n$ is consistent and $n^{1/2}(\boldsymbol{\alpha}_n - \boldsymbol{\alpha})$ is asymptotically normal under Assumptions 1–7.

Theorem 1. Let Assumptions 1–7 hold. As $n \to \infty$, (a) $\boldsymbol{\alpha}_n \xrightarrow{p} \boldsymbol{\alpha}$, and (b) $n^{1/2}(\boldsymbol{\alpha}_n - \boldsymbol{\alpha}) \xrightarrow{d} N(0, \boldsymbol{\Sigma}_{\alpha})$, where $\boldsymbol{\Sigma}_{\alpha}$ is the covariance matrix of the $\ell \times 1$ random vector $\boldsymbol{\Lambda}_n$ whose mth component ($m = 1, \ldots, \ell$) is

$$\sum_{j=2}^{M}[(\mathbf{W}'\mathbf{W})^{-1}\mathbf{W}']_{mj}n^{-1/2}\sum_{i=1}^{n}\{I(\mathbf{Z}_i = \mathbf{z}^{(j)})f(V_i, \mathbf{z}^{(j)})^{-1}$$
$$\times [Y_i - G_{\mathbf{z}^{(j)}}(V_i)]I[c_0 \leq G_{\mathbf{z}^{(j)}}(V_i) \leq c_1]$$
$$- I(\mathbf{Z}_i = \mathbf{z}^{(1)})f(V_i, \mathbf{z}^{(1)})^{-1}[Y_i - G_{\mathbf{z}^{(1)}}(V_i)]$$
$$\times I[c_0 \leq G_{\mathbf{z}^{(1)}}(V_i) \leq c_1] + (\boldsymbol{\Gamma}_{\mathbf{z}^{(j)}} - \boldsymbol{\Gamma}_{\mathbf{z}^{(1)}})\boldsymbol{\Omega}(Y_i, \mathbf{X}_i, \mathbf{Z}_i)\}.$$

(1996) Horowitz, J. and Härdle, W.
Direct semiparametric estimation of single-index models with discrete covariates.

Journal of the American Statistical Association, 91, 1632-1640

Table 1. Monte Carlo Parameter Estimates

		$n = 250$			$n = 500$		
		Mean (std. deviation)			Mean (std. deviation)		
α_1	α_2	$\tilde{b}_n$	α_{n1}	α_{n2}	$\tilde{b}_n$	α_{n1}	α_{n2}
			Direct semiparametric estimator				
0	0	2.031	.014	−.009	2.011	−.009	−.014
		(.394)	(.373)	(.245)	(.262)	(.227)	(.157)
0	.5	2.053	−.002	.484	2.006	0	.506
		(.423)	(.406)	(.306)	(.276)	(.257)	(.176)
0	1.0	2.050	.041	1.048	2.025	.001	1.053
		(.419)	(1.669)	(1.146)	(.302)	(.294)	(.234)
.5	0	2.051	.497	−.010	2.004	.509	−.010
		(.404)	(.387)	(.252)	(.277)	(.252)	(.151)
.5	.5	2.017	.493	.511	2.022	.523	.515
		(.402)	(.603)	(.414)	(.282)	(.266)	(.189)
.5	1.0	2.020	.420	.859	2.040	.489	1.034
		(.421)	(1.630)	(1.040)	(.272)	(.311)	(.235)
			Parametric maximum likelihood estimator				
0	0	2.018	.012	−.005	2.004	−.005	−.0005
		(.273)	(.246)	(.183)	(.195)	(.170)	(.118)
0	.5	2.053	.003	.505	2.018	.005	.497
		(.306)	(.235)	(.194)	(.214)	(.165)	(.132)
0	1.0	2.026	−.006	1.008	2.025	.005	1.005
		(.306)	(.258)	(.219)	(.209)	(.172)	(.157)
.5	0	2.055	.510	.004	2.017	.498	.005
		(.293)	(.255)	(.178)	(.204)	(.173)	(.119)
.5	.5	2.013	.501	.507	2.022	.504	.501
		(.287)	(.251)	(.184)	(.200)	(.181)	(.129)
.5	1.0	2.018	.523	.991	2.016	.504	1.018
		(.290)	(.285)	(.215)	(.216)	(.198)	(.160)

NOTE: Based on 500 replications. $\tilde{b}_n$ is the estimate of the second component of β; $\alpha_{ni} (i = 1, 2)$ estimates the ith component of α.

2.3 Estimating Σ_α

Σ_α can be estimated consistently by replacing unknown quantities with consistent estimators. It is not difficult to show that under Assumptions 1 and 3–7, Γ_z is estimated consistently by

$$\Gamma_{nz} = -n^{-1}\sum_{i=1}^{n}\tilde{\mathbf{X}}_i I(\mathbf{Z}_i = \mathbf{z})I(v_0 \le \hat{V}_i \le v_1) \times I[c_0 \le G_{nz}(\hat{V}_i) \le c_1]G'_{nz}(\hat{V}_i)/f_{nz}(\hat{V}_i)],$$

where $G'_{nz}(v) = dG_{nz}(v)/dv$. Define $\lambda(y, v, \mathbf{z})$ to be the $(M-1)\times 1$ vector whose $(j-1)$ component $(j = 2, \ldots, M)$ is

$$\begin{aligned}\lambda_j(y, v, \mathbf{z}) = {} & I(\mathbf{z} = \mathbf{z}^{(j)})f_{nz^{(j)}}(v)^{-1}[y - G_{nz^{(j)}}(v)] \\ & \times I[c_0 \le G_{nz^{(j)}}(v) \le c_1] \\ & - I(\mathbf{z} = \mathbf{z}^{(1)})f_{nz^{(1)}}(v)^{-1}[y - G_{nz^{(1)}}(v)] \\ & \times I[c_0 \le G_{nz^{(1)}}(v) \le c_1].\end{aligned}$$

Let Ω_n be a consistent estimator of Ω. Then under Assumptions 1–7, Σ_α is estimated consistently by the sample covariance of the $\ell \times 1$ vector whose mth component $(m = 1, \ldots, \ell)$ is

$$\sum_{j=2}^{M}[(\mathbf{W}'\mathbf{W})^{-1}\mathbf{W}']_{mj}[\lambda_j(Y_i, \hat{V}_i, Z_i) + (\Gamma_{nz^{(j)}} - \Gamma_{nz^{(1)}})\Omega_n(Y_i, \mathbf{X}_i, \mathbf{Z}_i)].$$

The details of Ω_n depend on the estimator of β that is used. To illustrate, let $p(\mathbf{x}|\mathbf{z})$ $(\mathbf{z} \in S_\mathbf{Z})$ be the probability density function of $\mathbf{X}$ conditional on $\mathbf{Z} = \mathbf{z}$, and let $p_n(\mathbf{z})$ be the empirical probability that $\mathbf{Z} = \mathbf{z}$. In Sections 3–4 we estimate β by (a) using the method of Powell et al. (1989) to estimate the density-weighted average derivative $[Ep(\mathbf{X}|\mathbf{z})\partial G(\mathbf{X}\beta + \mathbf{z}\alpha)/\partial \mathbf{x}]$ for each $\mathbf{z} \in S_\mathbf{Z}$, (b) forming a weighted average of the resulting estimates with weights $p_n(\mathbf{z})$, and (c) imposing the normalization $\beta_1 = 1$. Let $\delta = E[p(\mathbf{X}|\mathbf{Z})\partial G(\mathbf{X}\beta + \mathbf{Z}\alpha)/\partial x^{(1)}]$, where $x^{(1)}$ is the first component of $\mathbf{x}$. It follows by applying the delta method to equations (3.14) and (3.16) of Powell et al. (1989) that

$$\Omega(y, \mathbf{x}, \mathbf{z}) = -2\delta^{-1}p(\mathbf{z})[y - G(\mathbf{x}\beta + \mathbf{z}\alpha)] \times [\partial p(\mathbf{x}|\mathbf{z})/\partial \tilde{\mathbf{x}}' - \tilde{\beta}\partial p(\mathbf{x}|\mathbf{z})/\partial x^{(1)}]. \quad (12)$$

Let δ_n be the weighted average of the estimates of $[Ep(\mathbf{X}|\mathbf{z})\partial G(\mathbf{X}\beta + \mathbf{z}\alpha)/\partial x^{(1)}]$ using weights $p_n(\mathbf{z})$. Ω_n can be obtained from (12) by replacing δ with δ_n, p with p_n, $G(\mathbf{x}\beta + \mathbf{z}\alpha)$ with $G_{nz}(\mathbf{x}b_\mathbf{n})$, and $p(\mathbf{x}|\mathbf{z})$ with a kernel density estimator.

(1996) Horowitz, J. and Härdle, W.
Direct semiparametric estimation of single-index models with discrete covariates.

Journal of the American Statistical Association, 91, 1632-1640

Table 2. Monte Carlo Estimates of Variances and Levels of Nominal .05-Level t Tests

α_1	α_2		α_{n1}	α_{n2}
		Variances[a]		
0	0	Asymptotic	.0924	.0471
		Empirical	.0514	.0246
0	.5	Asymptotic	.0815	.0419
		Empirical	.0662	.0311
0	1.0	Asymptotic	.0607	.0332
		Empirical	.0864	.0548
.5	0	Asymptotic	.0985	.0503
		Empirical	.0633	.0229
.5	.5	Asymptotic	.0762	.0385
		Empirical	.0708	.0357
.5	1.0	Asymptotic	.0499	.0269
		Empirical	.0970	.0553
		Empirical Level of t Test		
0	0		.028	.016[b]
0	.5		.052	.048
0	1.0		.164[b]	.184[b]
.5	0		.044	.024[b]
.5	.5		.066	.082[b]
.5	1.0		.242[b]	.236[b]

[a] Based on 500 replications. "Asymptotic" is the mean of the estimates obtained from the asymptotic formula given in Section 2.3.
[b] Significantly different from the nominal level based on an asymptotic .01-level t test.

3. MONTE CARLO EXPERIMENTS

This section reports the results of a small-scale Monte Carlo investigation of the finite-sample behavior of α_n for model (2). In the experiments $k = l = 2$ and $n = 250$ or 500. G is the cumulative standard normal distribution function. The components of $\mathbf{X}$ are independently distributed as $N(0,1)$. The first component of $\mathbf{Z}$ is 0 with probability .5 and 1 with probability .5. The second component of $\mathbf{Z}$ takes the values 0, 1, and 2 with probabilities .25, .5, and .25. The components of $\mathbf{Z}$ are independent of one another and of $\mathbf{X}$. The first component of β is 1 by scale normalization. The second component, whose true value is 2, was estimated by forming a weighted average of density-weighted average-derivative estimates (Powell et al. 1989) that were computed for each point in $S_{\mathbf{Z}}$. The weights in the weighted average of estimates were the empirical probabilities $p_n(\mathbf{z})$. K is the fourth-order kernel $K(v) = (105/64)(1 - 5v^2 + 7v^4 - 3v^6)I(|v| \le 1)$.

Our theory does not indicate how to choose h_n, the bandwidth required to estimate $\tilde{\beta}, v_0, v_1, c_0$, and c_1 in applications. Härdle et al. (1992) and Härdle and Tsybakov (1992) derived the asymptotically optimal bandwidths for certain average-derivative estimators, but the resulting bandwidths are not asymptotically optimal for the estimation problem considered here.

In the absence of theoretical guidance, we have used a simple bandwidth selection procedure that can be implemented easily in Monte Carlo experiments, satisfies Assumption 7 and the regularity conditions of density-weighted average-derivative estimation, and performs well in the experiments. For density-weighted average-derivative estimation of β at a given $\mathbf{z} \in S_{\mathbf{Z}}$, we used the bandwidth $5n_{\mathbf{z}}^{-1/6}$, where $n_{\mathbf{z}}$ is the number of sampled observations for which $\mathbf{Z} = \mathbf{z}$. To compute $G_{n\mathbf{z}}$, we used the bandwidth $h_{n\mathbf{z}} = s_{v\mathbf{z}} n_{\mathbf{z}}^{-1/7.5}$, where $s_{v\mathbf{z}}$ is the sample standard deviation of $\mathbf{X}b_n$ conditional on $\mathbf{Z} = \mathbf{z} \in S_{\mathbf{Z}}$.

The integral in (11) was computed by Gauss-Legendre quadrature. To avoid edge effects in estimating $G_{\mathbf{z}}$, the limits of integration were set at

$$v_{n1} = \min_{\mathbf{z} \in S_{\mathbf{Z}}} \max_{1 \le i \le n} \{\mathbf{X}_i \mathbf{b}_n - h_{n\mathbf{z}} : \mathbf{Z}_i = \mathbf{z}\}$$

and

$$v_{n0} = \max_{\mathbf{z} \in S_{\mathbf{Z}}} \min_{1 \le i \le n} \{\mathbf{X}_i \mathbf{b}_n + h_{n\mathbf{z}} : \mathbf{Z}_i = \mathbf{z}\}.$$

To compute c_0 and c_1, we reestimated $G_{\mathbf{z}}$ for each $\mathbf{z} \in S_{\mathbf{Z}}$ using a second-order kernel (the standard normal density). Call the resulting estimate $G^*_{n\mathbf{z}}$. We then set

$$c_0 = \max_{\mathbf{z} \in S_{\mathbf{Z}}} \max_{\mathbf{X}_i \mathbf{b}_n \le v_{n0}} G^*_{n\mathbf{z}}(\mathbf{X}_i \mathbf{b}_n) \tag{13}$$

and

$$c_1 = \min_{\mathbf{z} \in S_{\mathbf{Z}}} \min_{\mathbf{X}_i \mathbf{b}_n \ge v_{n1}} G^*_{n\mathbf{z}}(\mathbf{X}_i \mathbf{b}_n). \tag{14}$$

Using a second-order kernel in (13) and (14) produces values of c_0 and c_1 that are more stable than those obtained with a fourth-order kernel.

There were 500 replications in each experiment. The computations were carried out in GAUSS using GAUSS pseudo–random number generators.

Table 1 shows the empirical means and standard deviations of $\tilde{\mathbf{b}}_n, \alpha_{n1}$, and α_{n2}. We also computed the empirical medians and interquartile ranges of the estimates. These lead to the same conclusions as the means and standard deviations, so they are not shown. To provide a basis for judging the performance of the semiparametric estimator, Table 1 also shows the means and standard deviations of the parametric maximum likelihood estimates of $\tilde{\beta}$ and α. The asymptotic efficiency bound for semiparametric estimation of β and α exceeds the Cramer–Rao bound (Cosslett 1987), so no semiparametric estimator can achieve the precision of the parametric maximum likelihood estimator. The estimator of Klein and Spady (1993) achieves the semiparametric efficiency bound, but its computational complexity

Table 3. Estimated Coefficients (Standard Errors) for a Model of Product Innovation

EMPLP	EMPLF	CAP	DEM
	Semiparametric model		
1	.032	.346	1.732
	(.023)	(.078)	(.509)
	Probit model		
1	.516	.520	1.895
	(.242)	(.163)	(.387)
	Monte Carlo experiment		
1	.032	.485	1.297
	(.073)	(.421)	(1.407)

NOTE: The coefficient of EMPLP is 1 by sign-scale normalization.

(1996) Horowitz, J. and Härdle, W.
Direct semiparametric estimation of single-index models with discrete covariates.

Journal of the American Statistical Association, 91, 1632-1640

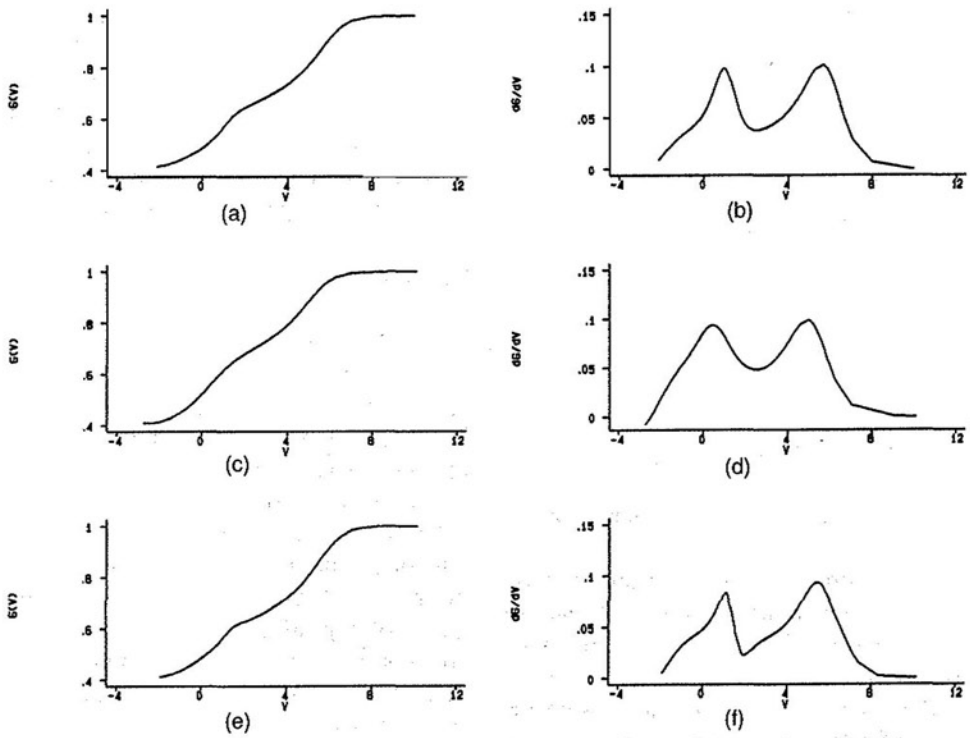


Figure 2. Estimates of G and dG/dv for the IFO Data. (a) Estimate of G with baseline bandwidths; (b) estimate of dG/dv with baseline bandwidths; (c) estimate of G with bandwidths increased by 20%; (d) estimate of dG/dv with bandwidths increased by 20%; (e) estimate of G with bandwidths decreased by 20%; (f) estimate of dG/dv with bandwidths decreased by 20%.

precludes carrying out Monte Carlo experiments to compare its finite-sample performance with that of the direct estimator.

The differences between the true values of α and the means of the semiparametric estimates are small except in the experiment with $(\alpha_1, \alpha_2) = (.5, 1)$ and $n = 250$. When $n = 500$, the root mean squared errors (RMSE's) of the semiparametric estimates of α_1 and α_2 exceed those of the maximum likelihood estimates by factors of 1.3 to 1.7. When $n = 250$, the semiparametric RMSE's exceed the maximum likelihood RMSE's by factors of 1.3 to 2.4, except in the experiments with $\alpha_2 = 1$, where the semiparametric RMSE's exceed the maximum likelihood RMSE's by factors of 5–6.

The bias and large RMSE's in the experiments with $\alpha_2 = 1$ and $n = 250$ can be understood by observing that α is estimated from the horizontal difference between functions $G_{n\mathbf{z}}$ corresponding to different values of $\mathbf{z}$. If the shifts caused by variations in $\mathbf{Z}\alpha$ are large and $n_{\mathbf{z}}$ is small (depending on $\mathbf{z}$, its average value is either 31 or 62 in the experiments with $n = 250$), then there may be few values of $\mathbf{X}\mathbf{b}_n$ in the interval on which the ranges of the functions $G_{n\mathbf{z}}$ overlap. This causes the estimates of $\Delta\mathbf{J}$ and α to be imprecise. The problem decreases with increasing n, as can be seen from the results of the experiments with $n = 500$.

Table 2 shows the empirical variances of α_{n1} and α_{n2} for the experiments with $n = 500$, together with the average variances estimated using the asymptotic formula of Section 2.3. The table also shows the empirical levels (rejection probabilities) of nominal .05-level t tests of the hypotheses that α_1 and α_2 have their true values. The denominators of the t statistics are computed using the asymptotic formula of Section 2.3. Except for the experiments with $\alpha_2 = 1$, the asymptotic formula overestimates the variances of α_{n1} and α_{n2}, and the empirical levels of the t test range from .016 to .082. The asymptotic formula underestimates the variances of α_{n1} and α_{n2} in the experiments with $\alpha_2 = 1$, and the empirical levels of the t tests are far above the nominal levels.

4. AN APPLICATION

This section illustrates the semiparametric estimator by applying it to data on product innovation by German manufacturers of investment goods. The data, assembled in 1989 by the IFO Institute in Munich, consist of observations on 1,100 manufacturers. The dependent variable is $Y = 1$ if a manufacturer realized an innovation during 1989 in a specific product category (defined by a four-digit code assigned by IFO) and zero otherwise. The continuous independent variables are the number of employees in the product category (EMPLP), the number of employees in the entire firm (EMPLF), and an indicator of the firm's production capacity utilization (CAP). There is one discrete independent variable, DEM, which is 1 if a firm expected increasing demand in the product category and 0 otherwise. We standardized the continuous variables, so they have units of standard deviations from their means. Scale/sign normalization was achieved by setting $\beta_{\text{EMPLP}} = 1$. The kernel and methods for choosing c_0, c_1, v_0, v_1, and the bandwidths are as described in Section 3.

The semiparametric estimates of β and α are shown in the top panel of Table 3. The middle panel shows estimates obtained from a parametric probit model. Figure 2 shows

(1996) Horowitz, J. and Härdle, W.
Direct semiparametric estimation of single-index models with discrete covariates.

Journal of the American Statistical Association, 91, 1632-1640

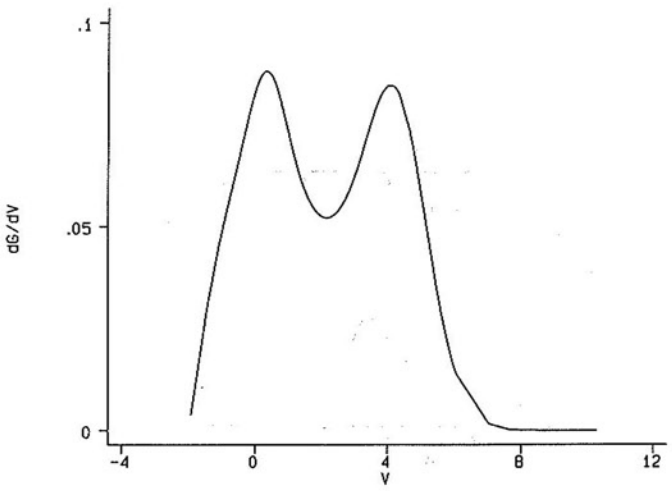


Figure 3. Estimate of dG/dv Using Probit Estimate of $\mathbf{X}\beta + \mathbf{Z}\alpha$.

estimates of G and dG/dv obtained from kernel nonparametric regression of Y on the semiparametric estimate of $\mathbf{X}\beta + \mathbf{Z}\alpha$. A second-order kernel (the normal density function) was used for this purpose. Figures 2a and 2b show the estimates obtained when α and β are estimated using the bandwidths described in Section 3. Figures 2c and 2d show the estimates obtained when the bandwidths are increased by 20%, and 2e and 2f show the estimates obtained when the bandwidths are decreased by 20%.

There are two important differences between the semiparametric and probit estimates. First, the semiparametric estimate of β_{EMPLF} is small and statistically nonsignificant, whereas the probit estimate is significant at the .05 level and similar in size to β_{CAP}. Second, Figure 2 reveals that dG/dv is bimodal. This contradicts the probit model, which assumes that dG/dv is a unimodal (normal) pdf. Bimodality is also present if the nonparametric regression is carried out using the probit estimate of $\mathbf{X}\beta + \mathbf{Z}\alpha$ (Fig. 3), so bimodality is not an artifact of the semiparametric estimation procedure. The bimodality of dG/dv suggests that the data may be a mixture of two populations. Although further investigation of this possibility is beyond the scope of this article, an obvious next step would be to search for variables that characterize these populations.

To gain additional insight into whether the semiparametric estimates reflect genuine features of the sampled population or are artifacts of our choices of bandwidths and other tuning parameters, we carried out a Monte Carlo experiment in which simulated data sets of size 250 were generated by sampling $(Y, \mathbf{X}, \mathbf{Z})$ randomly without replacement from the IFO data. Each simulated data set is a random sample from the distribution that generated the IFO data, rather than from an assumed model that may not capture essential features of this distribution. The centered, normalized parameter estimates have the same asymptotic distribution as the estimates obtained from the full sample (Politis and Romano 1994).

The bottom panel of Table 3 shows the means and standard errors of the parameter estimates obtained in 100 Monte Carlo replications. The standard errors indicate the variability of the Monte Carlo estimates. They are not estimates of the finite-sample standard deviations that would be obtained from independent random sampling of the true population distribution. The Monte Carlo parameter estimates are close to those obtained from the full data set. The Monte Carlo estimate of β_{EMPLF} is much closer to the full-data semiparametric estimate than to the probit estimate. Thus it appears that the main features of the semiparametric estimates are not artifacts of the choices of tuning parameters.

5. CONCLUSIONS

This paper has described a direct (noniterative) method for estimating the parameters of a semiparametric single-index model when some of the explanatory variables are discrete. The resulting estimator is $n^{1/2}$ consistent and asymptotically normal. The method described here is considerably less demanding computationally than other methods for estimating semiparametric single-index models with discrete explanatory variables, because other methods require solving difficult nonlinear optimization problems. An application to data on product innovation by German manufacturers has illustrated the usefulness of the semiparametric estimator.

APPENDIX: PROOF OF THEOREM 1

The proof is based on four lemmas. Assumptions 1–7 hold throughout.

Lemma 1.

a. For each $i = 2, \ldots, M$,

$$J[\mathbf{z}^{(i)}] - J[\mathbf{z}^{(1)}] = (c_1 - c_0)[\mathbf{z}^{(i)} - \mathbf{z}^{(1)}]\alpha.$$

b. $\alpha = (c_1 - c_0)^{-1}(\mathbf{W}'\mathbf{W})^{-1}\mathbf{W}'\Delta\mathbf{J}$.

Proof. To prove part a, define $v_a = \max\{v_0 + \mathbf{z}\alpha\colon \mathbf{z} \in S_{\mathbf{Z}}\}$ and $v_b = \min\{v_1 + \mathbf{z}\alpha\colon \mathbf{z} \in S_{\mathbf{Z}}\}$. Let $\mathbf{z} = \mathbf{z}^{(1)}$ or $\mathbf{z}^{(i)}$. Make the change of variables $v = u - \mathbf{z}\alpha$ on the right-hand side of (4). By Assumption 4, $I[G(u) < c_0] = 0$ if $u > v_b$, $I[G(u) > c_1] = 0$ if $u < v_a$, and $I[c_0 \le G(u) \le c_1] = 0$ if $u < v_a$ or $u > v_b$. Therefore,

$$\begin{aligned} J(\mathbf{z}) &= c_0 \int_{v_0+\mathbf{z}\alpha}^{v_a} I[G(u) < c_0]\,du + c_0 \int_{v_a}^{v_b} I[G(u) < c_0]\,du \\ &\quad + \int_{v_a}^{v_b} G(u) I[c_0 \le G(u) \le c_1]\,du \\ &\quad + c_1 \int_{v_a}^{v_b} I[G(u) > c_1]\,du \\ &\quad + c_1 \int_{v_b}^{v_1+\mathbf{z}\alpha} I[G(u) > c_1]\,du \\ &= c_0(v_a - v_0 - \mathbf{z}\alpha) + c_0 \int_{v_a}^{v_b} I[G(u) < c_0]\,du \\ &\quad + \int_{v_a}^{v_b} G(u) I[c_0 \le G(u) \le c_1]\,du \\ &\quad + c_1 \int_{v_a}^{v_b} I[G(u) > c_1]\,du + c_1(v_1 - v_b + \mathbf{z}\alpha). \end{aligned}$$

(1996) Horowitz, J. and Härdle, W.
Direct semiparametric estimation of single-index models with discrete covariates.

Journal of the American Statistical Association, 91, 1632-1640

It follows that

$$J[\mathbf{z}^{(i)}] - J[\mathbf{z}^{(1)}] = (c_1 - c_0)[\mathbf{z}^{(i)} - \mathbf{z}^{(1)}]\boldsymbol{\alpha}, \tag{A.1}$$

which proves part a. Part b follows from nonsingularity of $\mathbf{W}'\mathbf{W}$ and the observation that by part a, $\mathbf{W}'\boldsymbol{\Delta}\mathbf{J} = (c_1 - c_0)\mathbf{W}'\mathbf{W}\boldsymbol{\alpha}$.

Define

$$\boldsymbol{\Gamma}_1(v, \mathbf{z}) = p(\mathbf{z}) \int \tilde{\mathbf{x}}(\partial/\partial v)p(v, \tilde{\mathbf{x}}|\mathbf{z})\, d\tilde{\mathbf{x}},$$

$$\boldsymbol{\Gamma}_2(v, \mathbf{z}) = p(\mathbf{z}) \int \tilde{\mathbf{x}}(\partial/\partial v)]G_{\mathbf{z}}(v)p(v, \tilde{\mathbf{x}}|\mathbf{z})]\, d\tilde{\mathbf{x}},$$

$$\boldsymbol{\Gamma}_3(v, \mathbf{z}) = -G'_{\mathbf{z}}(v)E(\tilde{\mathbf{X}}|v, \mathbf{z}),$$

and

$$\boldsymbol{\Psi}_n = n^{-1}\sum_{i=1}^{n} \boldsymbol{\Omega}(Y_i, \mathbf{X}_i, \mathbf{Z}_i).$$

Lemma 2. Define $G_{\mathbf{z}}(v) = G(v + \mathbf{z}\boldsymbol{\alpha})(\mathbf{z} \in S_{\mathbf{Z}})$. For each $\mathbf{z} \in S_{\mathbf{Z}}$, the following hold uniformly over $v \in [v_0, v_1]$:

a. $A_{nz}(v) = (nh_n)^{-1}\sum_{i=1}^{n} I(\mathbf{Z}_i = \mathbf{z})Y_i K\left[\frac{v - V_i}{h_n}\right] - \boldsymbol{\Gamma}_2(v, \mathbf{z})\boldsymbol{\Psi}_n + O_p[(nh_n^3)^{-1}]$

b. $A_{nz}(v) - \mathbf{G}_{\mathbf{z}}(v)f(v, \mathbf{z}) = O_p[(nh_n)^{-1/2}(\log n)]$

c. $f_{nz}(v) = (nh_n)^{-1}\sum_{i=1}^{n} I(\mathbf{Z}_i = \mathbf{z})K\left[\frac{v - V_i}{h_n}\right] - \boldsymbol{\Gamma}_1(v, \mathbf{z})\boldsymbol{\Psi}_n + O_p[(nh_n^3)^{-1}]$

d. $f_{nz}(v) - f(v, \mathbf{z}) = O_p[(nh_n)^{-1/2}(\log n)]]$.

Proof. Only parts a and b are proven here. The proofs of parts c and d are similar. To begin, use a Taylor series expansion to obtain

$$\begin{aligned} &A_{nz}(v) \\ &= (nh_n)^{-1}\sum_{i=1}^{n} I(\mathbf{Z}_i = \mathbf{z})Y_i K\left[\frac{v - V_i}{h_n}\right] \qquad \text{(A.2)} \\ &\quad - (nh_n^2)^{-1}\sum_{i=1}^{n} I(\mathbf{Z}_i = \mathbf{z})Y_i K'\left[\frac{v - V_i}{h_n}\right]\tilde{\mathbf{X}}_i(\tilde{\mathbf{b}}_n - \tilde{\boldsymbol{\beta}}) + R_n, \\ &\equiv A_{nz1}(v) - \mathbf{A}_{nz2}(v)(\tilde{\mathbf{b}}_n - \tilde{\boldsymbol{\beta}}) + R_n, \qquad \text{(A.3)} \end{aligned}$$

where

$$R_n = -(nh_n^2)^{-1}\sum_{i=1}^{n} I(\mathbf{Z}_i = \mathbf{z}) \times Y_i\tilde{\mathbf{X}}_i\left\{K'\left[\frac{v - V_i^*}{h_n}\right] - K'\left[\frac{v - V_i}{h_n}\right]\right\}(\tilde{\mathbf{b}}_n - \tilde{\boldsymbol{\beta}})$$

and V_i^* is between $\hat{V}_i$ and V_i. Write $A_{nz1}(v) - G_{\mathbf{z}}(v)f(v, \mathbf{z}) = A_{nz}^*(v) - EA_{nz}^*(v) + [EA_{nz1}(v) - G_{\mathbf{z}}(v)f(v, \mathbf{z})] + \{A_{nz1}(v) - EA_{nz1}(v) - [A_{nz}^*(v) - EA_{nz}^*(v)]\}$, where

$$A_{nz}^*(v) = n(nh_n)^{-1}\sum_{i=1}^{n} I(\mathbf{Z}_i = \mathbf{z})(Y_i/n)K\left[\frac{v - V_i}{h_n}\right] I(Y_i \le n).$$

It follows from theorem (2.37) of Pollard (1984) that $|A_{nz}^*(v) - EA_{nz}^*(v)| = O[(nh_n)^{-1/2}(\log n)]$ almost surely uniformly over $v \in [v_0, v_1]$. Standard methods for kernel estimation show that $EA_{nz1}(v) = G_{\mathbf{z}}(v)f(v|\mathbf{z})p(\mathbf{z}) + O(h_n^r)$ uniformly over $v \in [v_0, v_1]$. An argument identical to that used to prove proposition 1 of Mack and Silverman (1982), except with their x restricted to $[v_0, v_1]$ instead of unbounded, shows that $\{A_{nz1}(v) - EA_{nz1}(v) - [A_{nz}^*(v) - EA_{nz}^*(v)]\} = O(n^{-1})$ almost surely uniformly over $v \in [v_0, v_1]$. Combining these results gives

$$A_{nz1}(v) = G_{\mathbf{z}}(v)f(v|\mathbf{z})p(\mathbf{z}) + O_p(h_n^r) + O_p[(nh_n)^{-1/2}(\log n)] \tag{A.4}$$

uniformly over $v \in [v_0, v_1]$. Similar arguments applied to $\mathbf{A}_{nz2}$ yield

$$\mathbf{A}_{nz2}(v) = \boldsymbol{\Gamma}_2(v, \mathbf{z}) + O_p(h_n^r) + O_p[(nh_n^3)^{-1/2}(\log n)] \tag{A.5}$$

uniformly over $v \in [v_0, v_1]$. Under Assumption 1, Lipschitz continuity of K' and $n^{1/2}$-consistency of $\tilde{\mathbf{b}}_n$ imply that

$$R_n = O_p[(nh_n^3)^{-1}] \tag{A.6}$$

uniformly over $v \in [v_0, v_1]$. Part a of the lemma follows by substituting (A.5) and (A.6) into (A.3) and using assumption 5. Part b follows by substituting (A.4)–(A.6) into (A.3).

Lemma 3. For each $\mathbf{z} \in S_{\mathbf{Z}}$,

$$\begin{aligned} &G_{nz}(v) - G_{\mathbf{z}}(v) \\ &\quad = [nh_n f(v|\mathbf{z})]^{-1}\sum_{i=1}^{n} I(\mathbf{Z}_i = \mathbf{z})[Y_i - G_{\mathbf{z}}(v)]K\left[\frac{v - V_i}{h_n}\right] \\ &\qquad + \boldsymbol{\Gamma}_3(v, \mathbf{z})\boldsymbol{\Psi}_n + O_p(h_n^r) + O_p[(nh_n^3)^{-1}] \end{aligned}$$

uniformly over $v \in |v_0, v_1]$.

Proof. A Taylor series expansion yields

$$\begin{aligned} G_{nz}(v) - G_{\mathbf{z}}(v) &= f(v, \mathbf{z})^{-1}[A_{nz}(v) - G_{\mathbf{z}}(v)f_{nz}(v)] \\ &\quad + O\{[A_{nz}(v) - G_{\mathbf{z}}(v)f(v, \mathbf{z})] \\ &\qquad \times [f_{nz}(v) - f(v, \mathbf{z})]/f(v, \mathbf{z})^2\} \\ &\quad + O\{[f_{nz}(v) - f(v, \mathbf{z})]^2/f(v, \mathbf{z})^2\}. \qquad \text{(A.7)} \end{aligned}$$

The lemma follows by applying Lemma 2 and Assumption 7 to (A.7).

Lemma 4. For each $\mathbf{z} \in S_{\mathbf{Z}}$,

$$J_n(\mathbf{z}) - J(\mathbf{z}) = n^{-1}\sum_{i=1}^{n} I(\mathbf{Z}_i = \mathbf{z})I[c_0 \le G_{\mathbf{z}}(V_i) \le c_1]f(V_i, \mathbf{z})^{-1} \times [Y_i - G_{\mathbf{z}}(V_i)] + \boldsymbol{\Gamma}_{\mathbf{z}}\boldsymbol{\Psi}_n + o_p(n^{-1/2}).$$

Proof. Define

$$H_{ni} = (1/h_n)\int_{v_0}^{v_1} I[c_0 \le G_{\mathbf{z}}(v) \le c_1] \times [Y_i - G_{\mathbf{z}}(v)]f(v, \mathbf{z})^{-1}K\left[\frac{v - V_i}{h_n}\right] dv$$

and

$$H_i^* = I(\mathbf{Z}_i = \mathbf{z})I[c_0 \le G_{\mathbf{z}}(V_i) \le c_1]f(V_i, \mathbf{z})^{-1}[Y_i - G_{\mathbf{z}}(V_i)].$$

It follows from Assumption 4 and Lemma 2 that $\int_{-\infty}^{\infty} |I[G_{nz}(v) < c_0] - I[G_{\mathbf{z}}(v) < c_0]|\, dv = o_p(n^{-1/2})$ uniformly over $v \in [v_0, v_1]$. The same result holds if c_0 is replaced by c_1 and/or the directions of the inequalities are reversed. Therefore, it follows from Lemma 3 that

$$J_n(\mathbf{z}) - J(\mathbf{z}) = n^{-1}\sum_{i=1}^{n} I(\mathbf{Z}_i = \mathbf{z})H_{ni} + \boldsymbol{\Gamma}_{\mathbf{z}}\boldsymbol{\Psi}_n + o_p(n^{-1/2}).$$

A straightforward but lengthy calculation based on Taylor series expansions shows that for each $\mathbf{z} \in S_{\mathbf{Z}}, E(H_{ni} - H_i^*) = O(h_n^r)$ and $\text{var}(H_{ni} - H_i^*) = O(h_n/n)$. The lemma now follows from Chebyshev's inequality.

(1996) Horowitz, J. and Härdle, W.
Direct semiparametric estimation of single-index models with discrete covariates.

Journal of the American Statistical Association, 91, 1632-1640

Proof of Theorem 1

By Lemma 4 and the definition of α_n,

$$\alpha_n - \alpha = (\mathbf{W}'\mathbf{W})^{-1}\mathbf{W}'\Lambda_n + o_p(n^{-1/2}).$$

Part a of the theorem follows by applying the weak law of large numbers to Λ_n, and part b follows by applying the Lindeberg–Levy theorem.

[Received October 1994. Revised February 1996.]

REFERENCES

Ai, C. (1991), "The Regression-Based Estimation Method of Index Model," unpublished manuscript, State University of New York at Stony Brook, Dept. of Economics.

Bierens, H. J., and Hartog, J. (1988), "Non-Linear Regression With Discrete Explanatory Variables, With an Application to the Earnings Function," *Journal of Econometrics,* 38, 269–299.

Cosslett, S. R. (1987), "Efficiency Bounds for Distribution-Free Estimators of the Binary Choice and the Censored Regression Models," *Econometrica,* 55, 559–585.

Friedman, J. H., and Stuetzle, W. (1981), "Projection Pursuit Regression," *Journal of the American Statistical Association,* 76, 817–823.

Han, A. K. (1987), "Non-Parametric Analysis of a Generalized Regression Model," *Journal of Econometrics,* 35, 303–316.

Härdle, W., Hart, J., Marron, J. S., and Tsybakov, A. B. (1992), "Bandwidth Choice for Average Derivative Estimation," *Journal of the American Statistical Association,* 87, 218–226.

———, and Stoker, T. M. (1989), "Investigating Smooth Multiple Regression by the Method of Average Derivatives," *Journal of the American Statistical Association,* 84, 986–995.

———, and Tsybakov, A. B. (1993), "How Sensitive are Average Derivatives?" *Journal of Econometrics,* 58, 31–48.

Ichimura, H. (1993), "Semiparametric Least Squares (SLS) and Weighted SLS Estimation of Single-Index Models," *Journal of Econometrics,* 58, 71–120.

Klein, R. L., and Spady, R. H. (1993), "An Efficient Semiparametric Estimator for Discrete Choice Models," *Econometrica,* 61, 387–422.

Mack, Y. P., and Silverman, B. W. (1982), "Weak and Strong Uniform Consistency of Kernel Regression Estimates," *Zeitschrift fur Wahrscheinlichkeitstheorie und verwandte Gebiete,* 61, 405–415.

Manski, C. F. (1988), "Identification of Binary Response Models," *Journal of the American Statistical Association,* 83, 729–738.

Politis, D. N., and Romano, J. P. (1994), "Large-Sample Confidence Regions Based on Subsamples Under Minimal Assumptions," *The Annals of Statistics,* 22, 2031–2050.

Pollard, D. (1984), *Convergence of Stochastic Processes,* New York: Springer-Verlag.

Powell, J. L., Stock, J. H., and Stoker, T. M. (1989), "Semiparametric Estimation of Index Coefficients," *Econometrica,* 51, 1403–1430.

Robinson, P. M. (1989), "Root-N Consistent Semiparametric Regression," *Econometrica,* 56, 931–954.

Sherman, R. P. (1993), "The Limiting Distribution of the Maximum Rank Correlation Estimator," *Econometrica,* 61, 123–138.

Stoker, T. M. (1991), "Equivalence of Direct, Indirect and Slope Estimators of Average Derivatives," in *Nonparametric and Semiparametric Methods in Econometrics and Statistics,* eds. W. A. Barnett, J. Powell, and G. Tauchen, New York: Cambridge University Press, pp. 99–118.

(1996) Horowitz, J. and Härdle, W.
Direct semiparametric estimation of single-index models with discrete covariates.

Allg. Statistisches Archiv 80, 433 – 441, ISSN 0002-6018

B. Rundschau

Sonderforschungsbereich 373 „Quantifikation und Simulation ökonomischer Prozesse“ an der Humboldt-Universität zu Berlin

Von WOLFGANG HÄRDLE und SIBYLLE SCHMERBACH, Berlin

Der Sonderforschungsbereich 373 wurde als erster geisteswissenschaftlicher SFB in den neuen Bundesländern an der Humboldt-Universität zu Berlin unter Beteiligung der Freien Universität Berlin, der Universität Potsdam und des Weierstraß-Institutes für Angewandte Analysis und Stochastik, Berlin, eingerichtet. Die Wissenschaftler des SFB 373 nahmen am 1. Januar 1994 ihre Arbeit auf. Zum Zeitpunkt der Niederschrift dieses Artikels befindet sich der Sonderforschungsbereich am Ende seiner ersten Förderperiode. Wir geben daher einen Überblick sowohl über die bisher geleistete Arbeit als auch über die geplante Forschungsarbeit in einer zweiten Förderperiode.

I. Förderperiode 1994 bis 1996

A. Wissenschaftliches Programm und wissenschaftliche Kooperation

Im Mittelpunkt der Arbeit der Wissenschaftler in der Antragsperiode 1994 bis 1996 stand die Zusammenführung ökonomischer Konzepte, mathematischer Methoden und statistischer Verfahren zur quantitativen Beurteilung der Dynamik ökonomischer Prozesse. Dabei bildeten solche ökonomischen Phänomene wie Lage und Entwicklung des Arbeitsmarktes (insbesondere die Migration von Arbeitskräften), die Geldnachfrage, die Dynamik des Preiswettbewerbes und die Stochastik von Finanzmärkten den Gegenstand der Forschung. Die Untersuchung neuer Klassen von mathematischen und statistischen Problemen wurde hierzu notwendig. Hauptaugenmerk wurde auf die Entwicklung und Implementation modellflexibler und nichtparametrischer Verfahren gelegt.

Dieses langfristig angelegte Forschungsprogramm des SFB 373 ist ohne einen kontinuierlichen, konstruktiven und kritischen Dialog zwischen den Vertretern der ökonomischen Theoriebildung, der empirisch orientierten Datenanalyse, der ökonometrisch-statistischen Modellierung sowie der angewandten Mathematik undenkbar. Die konzeptionelle Untergliederung des Sonderforschungsbereiches in die drei Projektbereiche

A *Quantitative Verfahren,*
B *Mathematische Methoden,*
C *Ökonomische Modellierung,*

Allg. Statistisches Archiv, 80, 433-441

hat sich für diese Aufgabenstellung als außerordentlich zweckmäßig erwiesen. Der Projektbereich *A* wurde von den quantitativen Lehrstühlen der Institute Statistik/Ökonometrie und Wirtschaftsinformatik der Wirtschaftswissenschaftlichen Fakultät der Humboldt-Universität getragen. Die entwickelten Methoden dieses Projektbereiches waren für die empirische Arbeit der benachbarten Projektbereiche von zentraler Bedeutung. Dies trifft sowohl auf die Bereitstellung nichtparametrischer datenanalytischer Verfahren wie auf die Entwicklung des Methodenservers MMM und des Software Environments XploRe zu.

Im Rahmen des *Projektbereich A* arbeiten folgende Teilprojekte:

A1 Semiparametrische Modelle (Härdle)
A2 Nichtparametrische Zeitreihenanalyse (Lütkepohl)
A3 MMM – Eine Methodenbank zur interoperablen Modellierung ökonomischer Prozesse (Günther)

Hauptziel der Arbeit des Teilprojektes *A1* war die Entwicklung und Erprobung von Verfahren und Techniken, die eine Beurteilung der Relevanz semiparametrischer Methoden in der ökonometrischen Praxis erlauben. Dazu wurden Schätzmethoden für flexiblere Regressionsmodelle (additive, semiparametrische und nichtparametrische Modelle) weiterentwickelt sowie statistische Testverfahren untersucht und konzipiert, die es ermöglichen, zwischen parametrischen, semiparametrischen und nichtparametrischen Modellen zu unterscheiden. Darüber hinaus war ein wesentlicher Teil der Projektarbeit *A1* der computergestützten Statistik gewidmet, d.h. der Algorithmenoptimierung und Entwicklung von XploRe als interaktive statistische Programmierumgebung.

Das Teilprojekt *A2* widmete seine Arbeit dem Ziel, die Anwendung der multivariaten Zeitreihenanalyse in der empirischen Wirtschaftsforschung durch geeignete Modellierung und aussagekräftige Modellinterpretation zum Erfolg zuführen. Diese Arbeit wurde in vier Teilbereichen geleistet:

a) Schätzen von VAR(∞)-Prozessen,
b) Interpretation von VAR-Prozessen,
c) Konkurrierende Modellansätze,
d) Nichtlineare Prozesse.

Ziel des Teilprojektes *A3* war die Erforschung und Entwicklung von sogenannten Methodenbanken. Die Kernidee einer Methodenbank besteht darin, Nutzern einen möglichst komfortablen Zugriff auf Methoden (d.h. Softwarmodule) zu ermöglichen, die im Rechnernetz verteilt abgelegt sind. Unabhängig von den verwendeten (heterogenen) Hard- und Softwareplattformen sollen die Nutzer — Mathematiker und Ökonomen — ihre unterschiedlichen Datensätze und Methoden beliebig miteinander kombinieren können.

Hauptziel des **Projektbereiches B** mit den Teilprojekten

B1 Kurvenschätzung und Resamplingverfahren (Bunke)
B2 Stochastische Modelle für Finanzmärkte und Statistik von Prozessen (Küchler)
B3 Nichtlineare Modell und Verfahren (Läuter)

(1996) Härdle, W. and Schmerbach, S. Sonderforschungsbereich 373 Quantifikation und Simulation ökonomischer Prozesse.

Allg. Statistisches Archiv, 80, 433-441

war die Entwicklung theoretischer Grundlagen für die Analysen der Projektbereiche *A* und *C*. Diese Teilprojekte werden von den mathematischen Instituten an der Mathematisch–Naturwissenschaftlichen Fakultät II der Humboldt–Universität zu Berlin, dem Weierstraß–Institut für Angewandte Analysis und Stochastik, Berlin, und dem Mathematischen Institut an der Wirtschaftswissenschaftlichen Fakultät der Universität Potsdam getragen.

Schwerpunkt der Forschungsarbeit des Teilprojektes *B1* war die Weiterentwicklung der statistischen Methodologie zur Modellierung von ökonomischen Zusammenhängen, zur Konstruktion von Vorhersageverfahren und zur Auffindung von Einflußvariablen. Hierzu wurden die nachstehenden Teilbereiche bearbeitet:

a) Modellierung und Resampling,
b) Waveletmethoden und nichtparametrische Glättungsverfahren,
c) Kurvenschätzung und qualitative Gestaltsmerkmale der Kurve,
d) Statistische Modellierung von Finanzmärkten.

Das Teilprojekt *B2* sah seine Hauptaufgabe darin, Beiträge zur stochastischen Modellierung von Finanzmärkten, insbesondere von derivativen Finanzprodukten, und zur Statistik stochastischer Prozesse zu leisten. Im Mittelpunkt standen dabei die Bestimmung von Arbitragegrenzen auf dem deutschen Markt für festverzinsliche Wertpapiere, Untersuchungen betreffs Einflußfaktoren für die Stochastik von Zinssätzen, Untersuchungen zu hyperbolischen Verteilungen von Aktienrenditen, mikroökonomische Modelle von Finanzmärkten, asymptotische Eigenschaften von Maximum–Likelihood–Schätzern bei gewissen stochastischen Differentialgleichungen und Exponentialfamilien stochastischer Prozesse.

Im Teilprojekt *B3* wurde an der Entwicklung von Methoden und Techniken gearbeitet, die auf statistischen Prinzipien beruhen und geeignet sind, Zusammenhänge qualitativer und quantitativer Grössen zu finden und zu beschreiben. Diese wurden an konkreten ökonomischen Problemen erprobt. Folgende Gliederung der Arbeit wurde vorgenommen:

a) Nichtparametrische Schätzung und Tests,
b) Schätzung in parametrischen Modellen,
c) Schlecht gestellte Probleme und numerische Stabilität.

Der *Projektbereich C* mit den Teilprojekten

C1 Bestimmungsfaktoren von erwarteten Aktienrenditen (Stehle),
C2 Wandel auf dem Arbeitsmarkt (Burda),
C3 Stabilität der Geldnachfrage in der Bundesrepublik Deutschland (Wolters/Lütkepohl),
C4 Dynamik des Wettbewerbs (Wolfstetter/Schwalbach),

widmete sich der Untersuchung makroökonomischer Aggregate, der Mobilität von Arbeitskräften, den Bestimmungsfaktoren von Aktienrenditen und der Dynamik des Preiswettbewerbs. Diese Teilprojekte wurden von den ökono-

(1996) Härdle, W. and Schmerbach, S. Sonderforschungsbereich 373 Quantifikation und Simulation ökonomischer Prozesse.

Allg. Statistisches Archiv, 80, 433-441

mischen Instituten der Wirtschaftswissenschaftlichen Fakultät der Humboldt-Universität und der Freien Universität Berlin getragen.

Das Teilprojekt *C1* beschäftigte sich mit Bestimmungsfaktoren von erwarteten Aktienrenditen, in Zusammenarbeit mit *B2* mit Zinsstrukturkurven und arbeitete gemeinsam mit *A3* an der projektübergreifenden Installation der Methodenbank MMM.

Das Teilprojekt *C2* konzentrierte sich auf die Erforschung der Dynamik neu entstandener Arbeitsmärkte und der Beweggründe der Mobilität der Bevölkerung sowie der Interaktion zwischen Lohnfindung, Gewerkschaftsverhalten und Mobilität. Ersteres erweitert als theoretische Grundlage das Konzept der Matching-Funktion um räumliche Interaktionsprozesse sowie den Einfluß von Maßnahmen der aktiven Arbeitsmarktpolitik.

Im Rahmen des Teilprojektes *C3* wurde die Stabilität der Geldnachfrage in der Bundesrepublik Deutschland als wichtige Voraussetzung für eine Bewertung und Beinflussung des Transmissionsprozesses geldpolitischer Maßnahmen vom monetären zum realen Sektor und damit für die Inflationsrate untersucht. Ziel dieses Projektes war es daher, die deutsche Geldnachfrage zu modellieren und ihre Stabilität zu beobachten.

Das Teilprojekt *C4* behandelte die strategische Interaktion zwischen managergeleiteten Unternehmen. Mit der vorrangigen Behandlung von Einzelfragen des Preiswettbewerbs, der Theorie und Empirie von Unternehmenszielen sowie der Eintrittsdynamik bei vollständiger und unvollständiger Information wurde das Ziel verfolgt, die Grundlagen für eine Zusammenführung der Analyse des unvollkommenen Wettbewerbs und der Agency Theorie zu schaffen.

Die wissenschaftliche Kooperation der Projektbereiche im Sonderforschungsbereich 373 und ihrer Teilprojekte war in der Förderperiode Grundprinzip der Forschungstätigkeit aller beteiligten Wissenschaftler. Dies fand Ausdruck in verschiedenen Formen.

Wichtige Stationen der Zusammenarbeit waren der planmäßig in jedem Semester mehrmals stattfindende „JOUR FIX" aller wissenschaftlichen Mitglieder des SFB sowie die regelmäßig in jedem Sommersemester stattfindende dreitägige Klausurtagung. Zu diesen Anlässen stellten jeweils ein bzw. mehrere Teilprojekte ihre aktuellen Forschungsergebnisse vor, die von benachbarten Projektbereichen oder Teilprojekten kommentiert und anschließend allgemein von der Versammlung der anwesenden Wissenschaftler diskutiert wurden. Auch die laufenden projekt- und fakultätsübergreifenden Seminarveranstaltungen zu den Schwerpunktthemen des Sonderforschungsbereiches trugen wesentlich dazu bei, die erzielten Forschungsresultate der wissenschaftlichen Öffentlichkeit vorzulegen und die kritische und konstruktive Diskussion anzufachen.

Das Begutachtungsverfahren für neu erarbeitete Discussion Papers war ebenfalls ein wirkungsvolles Mittel zur Förderung der wissenschaftlichen Zusammenarbeit. Es legt fest, daß vor jeder Veröffentlichung in der Discussion-Paper-Reihe des SFB eine Begutachtung durch einen benachbarten Projektbereich erfolgen muß.

(1996) Härdle, W. and Schmerbach, S. Sonderforschungsbereich 373 Quantifikation und Simulation ökonomischer Prozesse.

Allg. Statistisches Archiv, 80, 433-441

B. Gästeprogramm und Tagungsprogramm

Das Gästeprogramm des Sonderforschungsbereiches war ein wichtiges Element seiner Forschungstätigkeit und bildete ein attraktives Element des Wissenschaftsstandortes Berlin/Brandenburg. Wesentliche Forschungsarbeiten gingen aus der Kooperation mit Gästen hervor. Es entstanden durch internationale Kooperation mehrere Bücher bzw. Buchprojekte wie zum Beispiel:

Härdle, W., Klinke, S. & Turlach, B. (1995): *XploRe — An Interactive Statistical Computing Environment*, Springer Verlag, New York;

Härdle, W. & Schimek, M. (eds.) (1996): *Statistical Theory and Computational Aspects of Smoothing*, Physika Verlag;

Wolfstetter, E. (1996), *Topics in Microeconomics*, Buchmanuskript (bisher 400 Seiten) — wird demnächst abgeschlossen;

Küchler, U. & Sorensen (1997), *Exponential Families of Stochastic Processes* — wird bei Springer eingereicht;

Lütkepohl, H. (ed.) (1997), *Nonparametric Dynamic Modelling*, — erscheint als Sonderheft des Journal of Econometrics.

Das Gästeprogramm des SFB ist für die interessierte Fachwelt jederzeit abrufbar. So wurden alle eingeladenen Wissenschaftler mit ihren aktuell bearbeiteten Forschungsthemen und Kontaktpunkten im Newsletter des Sonderforschungsbereiches angekündigt, der auch über World Wide Web gelesen werden kann.

Wichtige Impulse gingen von diesem internationalen wissenschaftlichen Austausch auch auf die vom Sonderforschungsbereich durchgeführten nationalen und internationalen Tagungen aus. So wurden im Förderzeitraum 1994 bis 1996 durch den SFB die folgenden Tagungen veranstaltet bzw. unterstützt:

1994

Küchler — "Statistik von Prozessen, Numerik stochastischer Differentialgleichungen und Anwendungen auf Finanzmärkte", 5. – 10. September 1994 in Berlin (30 Teilnehmer)

Schwalbach — „Internationale Konferenz zur Managerkompensation", Juni 1994 in Berlin (30 Teilnehmer)

Härdle/Jolivet — „Seminaire Paris – Berlin" September 1994 in Garchy (24 Teilnehmer). Dieses Seminar dient der deutsch–französischen Nachwuchsförderung in Statistik/Ökonometrie.

Lütkepohl — „Fifth Meeting of the European Conference Series in Quantitative Economics and Econometrics (EC)" (120 Teilnehmer)

1995

Burda — „Entry and Exit of Firms and Workers and the Interactions Between Labour and Product Markets", Juni 1995 in Berlin (30 Teilnehmer)

(1996) Härdle, W. and Schmerbach, S. Sonderforschungsbereich 373 Quantifikation und Simulation ökonomischer Prozesse.

Allg. Statistisches Archiv, 80, 433-441

Sonderforschungsbereich 373 — „Smoothing and Resampling in Economics", Oktober 1995 in Berlin (120 Teilnehmer)

Mammen/Nussbaum — „Seminar Berlin - Paris", September 1995 in Schmerwitz (30 Teilnehmer)

Schmerbach — „Empirische Wirtschaftsstatistik", November 1995 in Berlin (35 Teilnehmer)

1996

Sonderforschungsbereich 373 — „Stochastics, Information and Markets", Oktober 1996 in Berlin (voraus. 50 Teilnehmer)

Härdle/Müller/Hüet/Nussbaum — „Seminaire Paris - Berlin", September 1996 in Garchy (voraus. 35 Teilnehmer)

Stehle — „Der Gang an die Börse", September 1996 in Berlin (voraus. 20 Teilnehmer)

Mammen/Nussbaum/Härdle — „Asymptotic Methods in Stochastics Dynamics and Nonparametric Statistics", September 1996 in Berlin (voraus. 90 Teilnehmer)

Küchler/Föllmer — „Mathematical Finance and Applications„, Oktober 1996 in Berlin (voraus. 50 Teilnehmer)

II. Förderperiode 1997 bis 1999

A. Forschungsprogramm (Fortsetzung und Weiterentwicklung)

In natürlicher Fortsetzung des Forschungsprogramms der ersten Förderperiode von 1994 bis 1996 wurden durch den Sonderforschungsbereich 373 als Forschungsschwerpunkte die nichtparametrische computergestützte Datenanalyse, die Dynamik auf den Finanzmärkten sowie die experimentelle Wirtschaftsforschung formuliert.

Der *Projektbereich A* wird sich nunmehr hauptsächlich der quantitativen Datenanalyse, der parameterflexiblen Modellierung sich zeitlich entwickelnder ökonomischer Systeme, dem netzbasierten wissenschaftlichen Rechnen und der Analyse von auf Scannerdaten basierenden Kaufentscheidungen widmen. Im Mittelpunkt der Forschungsarbeiten des *Projektbereiches B* wird die Entwicklung theoretischer und methodologischer Grundlagen für die Fragestellungen der benachbarten Projektbereiche stehen. Es werden Resamplingverfahren, Waveletmethoden, nichtparametrische Schätzer in der Zinsstrukturanalyse, nichtlineare Regression und die stochastische Analysis von Finanzderivaten untersucht. Der ökonomische *Projektbereich C* wird sich mit der Analyse von Märkten, der Preisbildung und der simulationsbasierten experimentellen Analyse ökonomischer Prozesse beschäftigen.

Dazu konnte der SFB für die neue Förderperiode drei weitere Teilprojekte hinzugewinnen:

(1996) Härdle, W. and Schmerbach, S. Sonderforschungsbereich 373 Quantifikation und Simulation ökonomischer Prozesse.

Allg. Statistisches Archiv, 80, 433-441

Projektbereich A:

A4 Marketing–Mix–Modelle und die Analyse von Wettbewerbsbeziehungen (Hildebrandt)

Kern dieses Forschungsvorhabens ist die Entwicklung eines integrierten Modellansatzes, der auf der Grundlage eines vollständigen Attraktionsmodells und Scannerpaneldaten(des CCHM–Modelltyps) die Wettbewerbseffekte bei unterschiedlichen Marketing–Mix–Kombinationen (z.B. Preis, Promotion, Display) schätzt und die Wettbewerbseffekte über drei–modale Analysen auf grundlegende Wirkungsgrößen zurückführt. Alternativ sollen die Modelle des MNL–Typs für individuelle Datenanalysen weiterentwickelt werden.

Projektbereich B:

B4 Stochastische Analyse von Derivaten (Föllmer/Schweizer)

In diesem Projekt sollen einige der mathematischen Probleme untersucht werden, die an den Schnittstellen zwischen der Theorie der Finanzmärkte und der mathematischen Stochastik auftreten. Schwerpunkte sind dabei

- Optimalitätskriterien, Restriktionen und Kostenanalyse für Absicherungsstrategien für Derivate in unvollständigen Finanzmarktmodellen,
- Nichtlineare Effekte, die sich aus der Rückwirkung von Absicherungsstrategien auf die zugrunde liegende stochastische Dynamik ergeben,
- Stochastische Modellierung von mikroökonomischen Interaktionen zwischen verschiedenen Gruppen von Marktteilnehmern (noise trading/information trading), Analyse der zugehörigen temporären Preisgleichgewichte und Übergang zu Diffusionsmodellen über ein Invarianzprinzip.

Projektbereich C:

C5 Experimentelle Wirtschaftsforschung (Güth)

Das Teilprojekt C5 sieht vor, die experimentelle Forschung innerhalb des SFB zu etablieren. Mit erprobten Methoden der experimentellen Forschung sollen hier vielfältige Untersuchungen zu den verschiedenen Forschungsschwerpunkten des Sonderforschungsbereiches vorgenommen werden. Dazu ist die Einrichtung eines eigens für diese Zwecke optimal augestalteten Computerlabors geplant und bereits genehmigt. Ein Schwerpunkt soll dabei die Entwicklung von interaktiven Computerexperimenten sein.

Darüber hinaus konnte innerhalb des *Projektbereiches C* das Teilprojekt *C1* (Stehle) durch das Forschungsgebiet Marktmikrostruktur (Müller) erweitert werden. Bei der Bearbeitung dieses Forschungsthemas stehen die folgenden Fragen im Mittelpunkt:

- Wie wirken sich unterschiedliche Handelsregeln auf die Preisbildung von Finanztiteln aus?
- Wie sollten Primärmärkte und Sekundärmärkte gestaltet werden?

Alle beteiligten Wissenschaftler des Sonderforschungsbereiches 373 sehen in der weiteren Zusammenführung ökonomischer Konzepte, mathematisch–stochastischer Methoden und statistischer Verfahren zur quantitativen Beurteilung

(1996) Härdle, W. and Schmerbach, S. Sonderforschungsbereich 373 Quantifikation und Simulation ökonomischer Prozesse.

Allg. Statistisches Archiv, 80, 433-441

der Dynamik ökonomischer Prozesse den einzig gangbaren Weg zu einer erfolgreichen Zusammenarbeit von Ökonomen, Statistikern und Mathematikern. Das Kooperieren und Interagieren der von diesen Wissenschaftlern vertretenen Forschungsgebiete ist typischerweise sequentieller Natur. In einem kontinuierlichen wissenschaftlichen Dialog werden die Projekte diskutiert und dann Schritt für Schritt entwickelt.

Der inhaltliche Zusammenhang der einzelnen Teilprojekte sowie geeignete Formen der Kooperation untereinander lassen sich durch folgende Kohärenzmatrix anschaulich zeigen:

	A1	A2	A3	A4	B1	B2	B3	B4	C1	C2	C3	C4	C5
A1	•	1⊕	2	3	4⊕⋆	5⋆	⊕			6⊕	⊕		7
A2		•	8⋆	⊕	9⋆				10	⊕	11⊕		12
A3			•						12⋆				
A4				•			18			⊕			
B1					•	13	14⊕		15		16		
B2						•		17⊕	⊕		13		
B3							•					18	
B4								•	17⊕				19
C1									•			20	
C2										•			
C3											•		
C4												•	
C5													•

⊕: regelmäßiges gemeinsames Seminar

⋆: gemeinsame Mitarbeiter

1. Neuronale Netze in der nichtparametrischen Zeitreihenanalyse
2. Networked Computing
3. Semiparametric discrete choice modelling
4. Wavelets, Bootstrapverfahren für additive Modelle
5. Volatilitätsuntersuchungen
6. Mobilitätsanalysen, Option-Value of Waiting
7. Savingsprojekt
8. MMM für nichtparametrische Zeitreihenanalysen
9. Bootstrapverfahren für Zeitreihen
10. Zeitreihenanalysen von Aktienrenditen
11. Kointegration
12. Methodenbank
13. Zinsstrukturkurven
14. bedingte U-Statistiken
15. Modellierung von Aktienrenditen
16. Local Polynomial Regression
17. Optionsbewertungen
18. Nichtlineare Regression
19. Interaktionssysteme
20. Matching Modelle

(1996) Härdle, W. and Schmerbach, S. Sonderforschungsbereich 373 Quantifikation und Simulation ökonomischer Prozesse.

C. Literatur

Zusendung von Beiträgen und Besprechungsexemplaren, Anfragen usw. werden erbeten an: Prof. Dr. Horst Rinne, Professur für Statistik und Ökonometrie, Justus-Liebig-Universität Gießen, Licher Str. 64, 35394 Gießen. Für angenommene Beiträge hat der Autor eine ASCII-Datei oder eine LaTeX-Datei zu liefern. Unaufgefordert eingesandte Bücher können leider nicht zurückgeschickt werden.

Heftpreis DM/sFr 44,– / öS 326; Jahresband DM/sFr 128,– / öS 947 zzgl. Porto.

Abbestellungen können nur berücksichtigt werden, wenn sie innerhalb 8 Wochen nach Ausgabe des Schlußheftes eines Bandes beim Verlag vorliegen.

Verlag Vandenhoeck & Ruprecht, Theaterstraße 13, 37073 Göttingen

ISSN 0002-6018

Allg. Statistisches Archiv, 80, 433-441

B. Gäste- und Tagungsprogramm

Ein internationales Gästeprogramm mit hochqualifizierten und erfahrenen Wissenschaftlern soll auch in der zweiten Förderperiode die Arbeit der Wissenschaftler des SFB 373 bereichern und unterstützen. Dazu ist wiederum vorgesehen, vor allem längerfristige Gastaufenthalte zu planen und zu realisieren. Der Dialog zwischen Datenanalyse und Modellentwicklung z.B erfordert ein sorgfältiges Umgehen mit den vorhandenen Datenquellen einerseits und schrittweises Herantasten an eine geeignete Modellierungsform andererseits.

Gäste tragen vor allem auch durch ihre internationale Erfahrung dazu bei, Promovenden und Habilitanden, also dem wissenschaftlichen Nachwuchs, bei der Formulierung ihrer Forschungsergebnisse zu helfen. Dieser zunächst lokale Effekt führt aber zu einer sehr schnellen Verbreitung der erzielten wissenschaftlichen Ergebnisse und liefert gleichzeitig einen globalen Effekt: Berlin als Wissenschafts- und Forschungsstandort wird international sichtbarer. Bereits in der ersten Förderperiode konnte der SFB durch ein anspruchsvolles Gästeprogramm eine Reihe von Tagungen (siehe Abs. I.B) mit hochrangigen Gastwissenschaftlern durchführen. Der Sonderforschungsbereich 373 plant auch für die zweite Förderperiode und zwar 1998 und 1999, seine Forschungsergebnisse anläßlich zweier internationaler Tagungen vorzustellen. Es ist vorgesehen, diese Veranstaltungen den folgenden Schwerpunktthemen zu widmen:

- „Labour market",
- „Smoothing and Resampling for time series",
- „Financial markets".

C. Informationen

Weitere Informationen zum Sonderforschungsbereich 373 sind über die Geschäftsführung des SFB über World Wide Web unter der Adresse

http://www.wiwi.hu-berlin.de/institute/sfb

erhältlich. Alle Discussion Paper des Sonderforschungsbereiches sind ebenfalls unter dieser Adresse einsehbar.

Die Softwareschnittstelle MMM (Teilprojekt A3) ist über

http://mmm.wiwi.hu-berlin.de

aufrufbar. Die interaktive statistische Umgebung XploRe (Teilprojekt A1) hat eine MMM Schnittstelle, einen WWW Browser und eine JAVA Schnittstelle unter

http://wotan.wiwi.hu-berlin.de.

Prof. Dr. Wolfgang Härdle
Dr. Sibylle Schmerbach
Humboldt-Universität zu Berlin
Institut für Statistik und Ökonometrie
Spandauer Straße 1
D–10178 Berlin

(1996) Härdle, W. and Schmerbach, S. Sonderforschungsbereich 373 Quantifikation und Simulation ökonomischer Prozesse.