The Annals of Statistics
1993, Vol. 21, No. 1, 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

Received March 1991; revised February 1992.
[1]Research supported by NSF Grant SES-88-0993.
AMS 1991 *subject classifications*. Primary 62H99; secondary 62HO5.
Key words and phrases. Bandwidth, heteroscedastic, kernel estimator, projection pursuit, regression, single index model, smoothing.

Institute of Mathematical Statistics is collaborating with JSTOR to digitize, preserve, and extend access to *The Annals of Statistics*.
www.jstor.org

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 $A^{2h} = \{x \in \mathbb{R}^p\colon \|x - y\| \le 2h \text{ for some } y \in A\}$.

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

$$\hat{g}(u|\theta) = \left\{ \sum_{j=1}^{n} Y_j K_h\big(u - \theta^T x_j\big) \right\} \Big/ \left\{ \sum_{j=1}^{n} K_h\big(u - \theta^T x_j\big) \right\},$$

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

$$\hat{g}_i(u|\theta) = \left\{ \sum_{j \neq i} Y_j K_h\big(u - \theta^T x_j\big) \right\} \Big/ \left\{ \sum_{j \neq i} K_h\big(u - \theta^T x_j\big) \right\}.$$

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

$$\hat{S}(\theta, h) = \sum_i{}' \big\{ Y_i - \hat{g}_i\big(\theta^T x_i|\theta\big) \big\}^2,$$

where Σ_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

$$\tilde{S}(\theta) = \sum_i{}' \big\{ Y_i - g\big(\theta^T x_i|\theta\big) \big\}^2 \tag{2.2}$$

is the distance measure we would employ instead of $\hat{S}$ if we knew $g(\cdot|\theta)$, and

$$T(h) = \sum_i{}' \big\{ \hat{g}_i\big(\theta_0^T x_i|\theta_0\big) - g\big(\theta_0^T x_i\big) \big\}^2 \tag{2.3}$$

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

$$\begin{aligned} \hat{S}(\theta, h) = \tilde{S}(\theta) + T(h) + \{&\text{terms of smaller order than } T(h) \\ &\text{and not depending on } \theta\} \\ + \big\{&\text{terms of smaller order than either } \tilde{S} \text{ or } T(h)\big\}. \end{aligned} \tag{2.4}$$

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 $0 < C_1 < C_2 < \infty$, $\Theta_n = \{\theta \in \Theta\colon \|\theta - \theta_0\| \le Cn^{-1/2}\}$, $\mathscr{H}_n = \{h\colon C_1 n^{-1/5} \le h \le C_2 n^{-1/5}\}$. 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

$$\mu(x|\theta) = E\big(X_A|\theta^T X_A = \theta^T x\big), \qquad K_1 = \int z^2 K(z)\,dz,$$

$$K_2 = \int K^2(z)\,dz, \tag{2.11}$$

$$V = \sum_i{}'\{x_i - \mu(x_i|\theta_0)\}g'\big(\theta_0^T x_i\big)\varepsilon_i,$$

$$A_1 = K_2 \int_A \gamma\big(\theta_0^T x\big)^{-1}\sigma(x)^2 f(x)\,dx, \qquad A_2 = \tfrac{1}{4}K_1^2 \int_A g''\big(\theta_0^T x\big)^2 f(x)\,dx. \tag{2.12}$$

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,

$$\tilde{S}(\theta) = n\big\{W_0^{1/2}(\theta - \theta_0) - n^{-1/2}\sigma Z\big\}^T\big\{W_0^{1/2}(\theta - \theta_0) - n^{-1/2}\sigma Z\big\} + R_3 + R_4(\theta), \tag{2.15}$$

$$T(h) = A_1 h^{-1} + A_2 n h^4 + R_5(h), \tag{2.16}$$

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

$$G\left\{\hat{g}(\hat{\theta}_1^T x_i|\hat{\theta}_1)\right\}^{-1},$$

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 $\hat{h}_2 h_0 \to 1$ 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

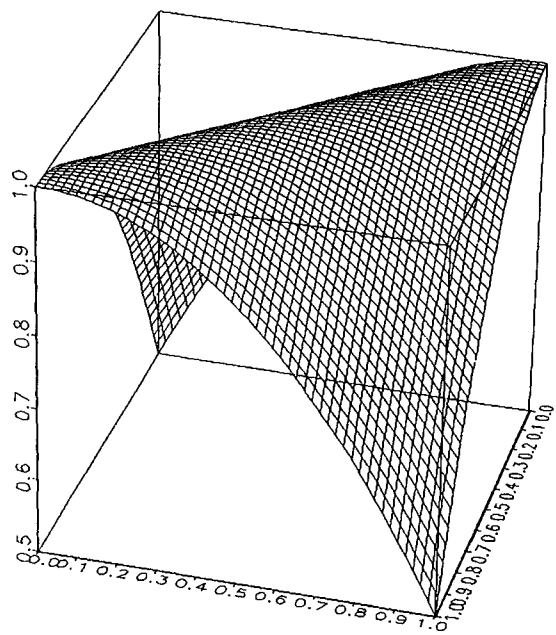

FIG. 1. *The function* $g(\theta_0^T X; 1)$ *on the unit square.*

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

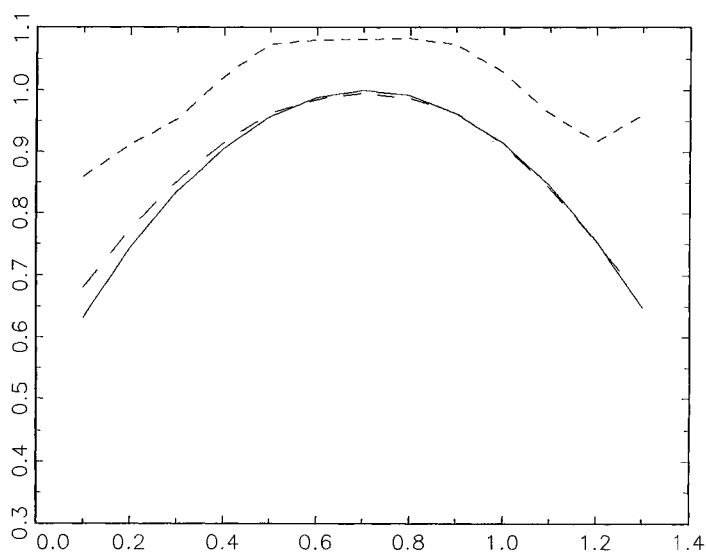

FIG. 2. *The true curve $g(u; 1)$ (solid line), the mean of $\hat{g}_{\hat{h}}(u|\hat{\theta})$ over 200 simulations (long dashes), the upper 95% confidence intervals (short dashes) and the lower 95% confidence intervals (dotted line).*

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

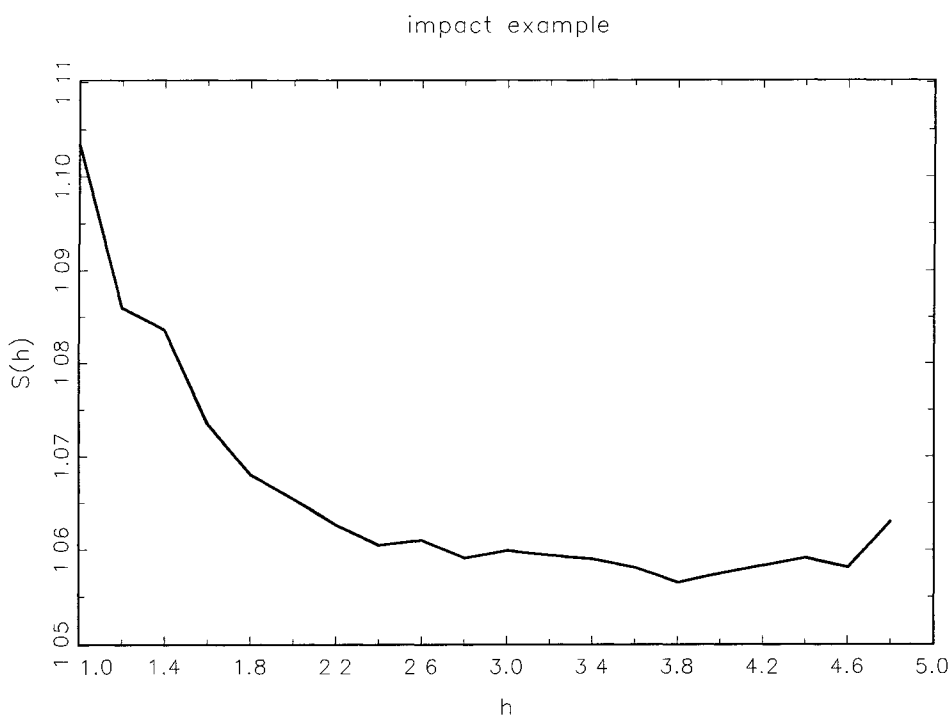

FIG. 3. *The objective function $S(\hat{\theta}, h)$ as a function of h.*

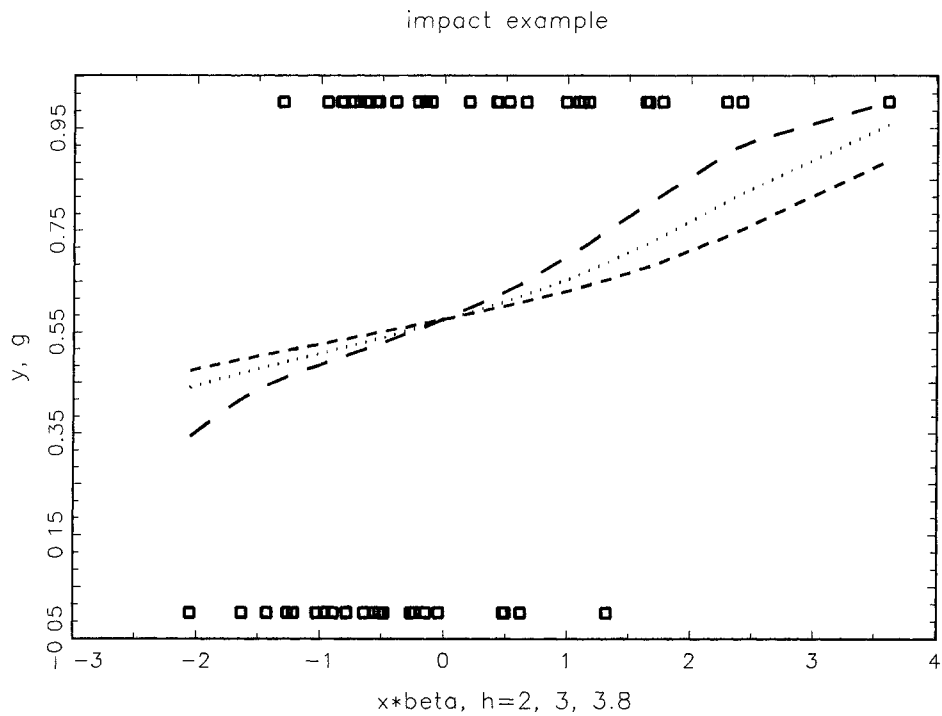

FIG. 4. *The projected data* $\{X_i^T\hat{\theta}, Y_i\}_{i=1}^n$ *and the optimal kernel regression estimate* $\hat{g}(u|\hat{\theta})$ *with* $\hat{\theta} = (0.3, 0.3, 0.9)$ *and* $h = 2$ (*solid line*), $h = 3$ (*dashed line*), $h = 2$ (*dotted line*).

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 $\hat{g}_h(u|\hat{\theta})$. 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 $\tilde{S}(\theta) = \Sigma_i'\{Y_i - g(\theta^T x_i|\theta)\}^2$,

$$D_i = \hat{g}_i(\theta_0^T x_i|\theta_0) - g(\theta_0^T x_i), \qquad \delta_i = g(\theta^T x_i|\theta) - g(\theta_0^T x_i),$$
$$\Delta_i = \hat{g}_i(\theta^T x_i|\theta) - g(\theta^T x_i|\theta) - \{\hat{g}_i(\theta_0^T x_i|\theta_0) - g(\theta_0^T x_i)\}.$$

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 $\Sigma_i' E(D_i^2) = O(n^{1/5})$, whence $T(h) = \Sigma_i' D_i^2 = O_p(n^{1/5})$. 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, $\Sigma_i'\delta_i^2 = O(1)$, 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 $|\Sigma_i' D_i\delta_i| = O_p(n^{-3/10+\xi})$, and step (viii) that $|\Sigma_i'\Delta_i\varepsilon_i| = O_p(n^{-1/5+\xi})$, for all $\xi > 0$. Therefore, the right-hand side of (4.1) equals $o_p(1)$. We prove in step (vii) that the term $\Sigma_i' D_i\varepsilon_i$, 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)] $= A_1 h^{-1} + A_2 nh^4 + O(n^{1/5})$, 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 $\varphi_n = \Sigma_i' \Delta_i^2$ [from steps (iii) and (iv); call this Example 1] and $\varphi_n = \Sigma_i' D_i \varepsilon_i$ [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 $\alpha > 0$,

$$\begin{aligned} P\Big\{ \sup_{\theta \in \Theta_n',\, h \in \mathscr{H}_n'} |\varphi_n(\theta, h)| > \alpha n^a \Big\} \\ \leq 2n^c \sup_{\theta \in \Theta_n',\, h \in \mathscr{H}_n'} P\{|\varphi_n(\theta, h)| > \alpha n^a\} \\ \leq 2n^c \big(n^b / \alpha n^a\big)^{2l} \sup_{\theta \in \Theta_n',\, h \in \mathscr{H}_n'} E\{\varphi_n(\theta, h)/n^b\}^{2l} = O(1), \end{aligned}$$

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

$$\sup_{\theta \in \Theta'_n,\, h \in \mathscr{H}'_n} \left| \varphi_n(\theta, h) - \varphi_n(\theta', h') \right| = o_p(n^a). \tag{4.7}$$

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,

$$\begin{aligned} \mathrm{Var}\{B_i(\theta, x) - B_i(\theta_0, x)\} &= O\big\{(nh)^{-2}n(n^{-1/2}h^{-1})^2 h\big\} \\ &= O(n^{-2}h^{-3}) = O\big\{(n^{-1/2}h)^2\big\}, \end{aligned}$$

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,

$$E(T^2) = O\Bigg[\Bigg|\sum_j \sum_l \sum_{i \neq j, l} E\{a(X_i, X_j) a(X_i, X_l) I(X_i \in A)\}\Bigg|$$
$$+ \Bigg|\sum \sum_{i \neq j} E\{a(X_i, X_j) a(X_j, X_i) I(X_i, X_j \in A)\}\Bigg|\Bigg]$$
$$= O\{(nh)^{-2} n (nh)^2 h^4 + (nh)^{-2} n^2 h^3\} = O(nh^4).$$

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,

$$\operatorname{Var}\Big(\sum{}'D_i\delta_i\Big) \le 2(nh)^{-2}\eta^2\sum{}'v_j^2\sigma(x_j)^2 + O(n^{\xi}h^4).$$

Now, v_j equals the observed value of $V_j = \Sigma'_{i\neq j}b(X_i, X_j)$, 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,

$$\operatorname{Var}\Big(\sum{}'O_i\delta_i\Big) = O\{(nh)^{-2}\eta^2 n^{2+\xi}h + n^{\xi}h^4\} = O(n^{\xi}h^4),$$

as required.

Step (vii). We show that with X-probability 1 and for all $\xi > 0$,

$$E\Big(\sum{}'D_i\varepsilon_i\Big)^2 = O(n^{1/5+\xi}). \tag{4.16}$$

Note that $E(D_i) = O(h^2n^{\xi})$ uniformly in i, for all $\xi > 0$. Hence,

$$E\Big\{\sum{}'E(D_i)\varepsilon_i\Big\}^2 = \sum{}'(ED_i)^2\sigma(x_i)^2 = O(n^{1+\xi}h^4) = O(n^{1/5+\xi}) \tag{4.17}$$

for all $\xi > 0$. Define $s_{ij} = E\{(D_i - ED_i)\varepsilon_i(D_j - ED_j)\varepsilon_j\}$. Then $s_{ii} = \operatorname{Var}(D_i)\sigma(x_i)^2$, and for $i \neq j$,

$$s_{ij} = K_h\{\theta_0^T(x_i - x_j)\}K_h\{\theta_0^T(x_j - x_i)\}\Big[\sum_{k\neq i}K_h\{\theta_0^T(x_i - x_k)\}\Big]^{-1}$$

$$\times\Big[\sum_{k\neq j}K_h\{\theta_0^T(x_j - x_k)\}\Big]^{-1}\sigma(x_i)^2\sigma(x_j)^2.$$

Therefore,

$$\begin{aligned} E\Big\{\sum{}'(D_i - ED_i)\varepsilon_i\Big\}^2 &= \sum{}'s_{ii} + \sum_i{}'\sum_{i\neq j}{}'s_{ij} \\ &= O\{nh^4 + n\cdot nh(nh)^{-2}\} = O(n^{1/5}). \end{aligned} \tag{4.18}$$

The desired result (4.16) follows from (4.17) and (4.18).

Step (viii). We show that with X-probability 1 and for all $\xi > 0$,

$$E\Big(\sum{}'\Delta_i\varepsilon_i\Big)^2 = O(n^{-2/5+\xi}). \tag{4.19}$$

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

$$\begin{aligned} \tilde{S} &= \sum_i{}'\big\{\varepsilon_i + g(\theta_0^T x_i) - g(\theta^T x_i|\theta)\big\}^2 \\ &= \sum_i{}'\varepsilon_i^2 - 2\eta\theta_{00}^T V + \eta^2\theta_{00}^T W\theta_{00} + o_p(1) \\ &= \sum_i{}'\varepsilon_i^2 - nZ'Z + n\big(W_0^{1/2}\eta\theta_{00} - n^{-1/2}\sigma Z\big)^T\big(W_0^{1/2}\eta\theta_{00} - n^{-1/2}\sigma Z\big) \\ &\quad + o_p(1), \end{aligned}$$

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