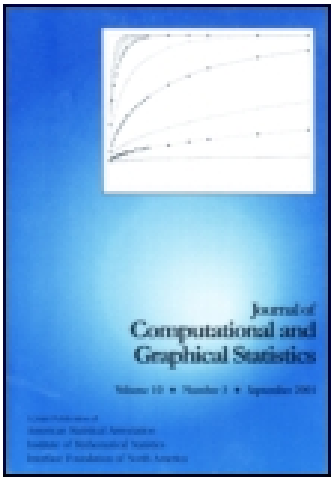

Journal of Computational and Graphical Statistics

Publication details, including instructions for authors and subscription information:
http://www.tandfonline.com/loi/ucgs20

A Computationally Efficient Oracle Estimator for Additive Nonparametric Regression with Bootstrap Confidence Intervals

Woocheol Kim [a] , Oliver B. Linton [a] & Niklaus W. Hengartner [b]

[a] Department of Economics , Yale University , 30 Hill House Avenue, New Haven , CT , 06520 , USA

[b] Department of Statistics , Yale University , USA

Published online: 21 Feb 2012.

To cite this article: Woocheol Kim , Oliver B. Linton & Niklaus W. Hengartner (1999) A Computationally Efficient Oracle Estimator for Additive Nonparametric Regression with Bootstrap Confidence Intervals, Journal of Computational and Graphical Statistics, 8:2, 278-297

To link to this article: http://dx.doi.org/10.1080/10618600.1999.10474814

A Computationally Efficient Oracle Estimator for Additive Nonparametric Regression with Bootstrap Confidence Intervals

Woocheol KIM, Oliver B. LINTON, and Niklaus W. HENGARTNER

This article makes three contributions. First, we introduce a computationally efficient estimator for the component functions in additive nonparametric regression exploiting a different motivation from the marginal integration estimator of Linton and Nielsen. Our method provides a reduction in computation of order n, which is highly significant in practice. Second, we define an efficient estimator of the additive components, by inserting the preliminary estimator into a backfitting algorithm but taking one step only, and establish that it is equivalent, in various senses, to the oracle estimator based on knowing the other components. Our two-step estimator is minimax superior to that considered in Opsomer and Ruppert, due to its better bias. Third, we define a bootstrap algorithm for computing pointwise confidence intervals and show that it achieves the correct coverage.

Key Words: Instrumental variables; Kernel estimation; Marginal integration.

1. INTRODUCTION

It is common practice to study the association between multivariate covariates and responses via regression analysis. Although nonparametric models for the conditional mean $m(x) = E(Y|X = x)$ are useful exploratory and diagnostic tools when X is one-dimensional, they suffer from the curse of dimensionality (Härdle 1990; Wand and Jones 1995) in that their best possible convergence rate is $n^{-q/(2q+d)}$, where d is the dimension of X and $m(\cdot)$ is q-times continuously differentiable. Additive regression models of the form

$$m(x) = c + m_1(x_1) + m_2(x_2) + \cdots + m_d(x_d), \tag{1.1}$$

with $x = (x_1, \ldots, x_d)^T \in \mathbb{R}^d$, offer a compromise between the flexibility of a full nonparametric regression and reasonable asymptotic behavior. In particular, in such additive regression models, the functions $m_j(\cdot)$ can be estimated with the one-dimensional rate of

Woocheol Kim is a Doctoral Candidate, and Oliver B. Linton is Professor, Department of Economics, Yale University, 30 Hill House Avenue, New Haven, CT 06520 (Email: woocheol@minerva.cis.yale.edu; linton@econ.yale.edu). Niklaus Hengartner is Assistant Professor, Department of Statistics, Yale University (Email: hengart@stat.yale.edu).

Journal of Computational and Graphical Statistics, Volume 8, Number 2, Pages 278–297

convergence—for example, $n^{q/(2q+1)}$ for q-times continuously differentiable functions—regardless of d; see Stone (1985, 1986). The backfitting algorithm of Breiman and Friedman (1985), Buja, Hastie, and Tibshirani (1989), and Hastie and Tibshirani (1990) is widely used to estimate the one-dimensional components $m_j(\cdot)$ and regression function $m(\cdot)$. More recently, Newey (1994), Tjøstheim and Auestad (1994), and Linton and Nielsen (1995) independently introduced the alternative method of *marginal integration* to estimate $m_\ell(x_\ell)$ [see also earlier work by Auestad and Tjøstheim (1991)]. One advantage of the integration method is that its statistical properties are easier to describe; specifically, one can easily prove central limit theorems and give explicit expressions for the asymptotic bias and variance of the estimators. Hengartner (1996) provided the weakest set of conditions, showing that the curse of dimensionality does not necessarily operate. Extensions to generalized additive models (Linton and Härdle 1996); to transformation models (Linton, Wang, Chen, and Härdle 1997); and to hazard models (Nielsen and Linton 1997) are also readily analyzed using this estimation method.

The marginal integration estimator relies on the following heuristic: if the regression function is additive and if Q is a $d-1$-dimensional probability measure, then $\int m(x)dQ(x_2, x_3, \ldots, x_d) = c' + m_1(x_1)$, where c' is a constant with respect to x_1. For identification, it is traditional to assume that $E\{m_\ell(X_\ell)\} = 0$ for $\ell = 1, \ldots d$, in which case the constant is $E(Y) = c$, if the integrating measure is exactly the probability distribution of $X_2, \ldots, X_d$. Applied to data, one replaces m by a d-dimensional pilot estimator $\widehat{m}$, and uses some known measure to integrate with. The empirical marginal integration method is the most convenient in practice. Partition $X_j = (X_{1j}, X_{2j})$ and $x = (x_1, x_2)$, where X_{1j} and x_1 are scalars, while X_{2j} and x_2 are $d-1$-dimensional, and consider the multidimensional Nadaraya–Watson nonparametric regression estimator

$$\widehat{m}(x) = \frac{1}{nh_o^d}\sum_{j=1}^{n} K\left(\frac{x - X_j}{h_o}\right) Y_j / \widehat{p}(x) \quad ; \quad \widehat{p}(x) = \frac{1}{nh_o^d}\sum_{l=1}^{n} K\left(\frac{x - X_l}{h_o}\right), \tag{1.2}$$

where $K(t_1, \ldots, t_d) = \prod_{j=1}^{d} k(t_j)$ with k a scalar kernel of order q, while h_o is a scalar bandwidth. Here, $\widehat{p}(x)$ is an estimator of the joint probability density $p(x)$. Then

$$\widehat{\gamma}_1(x_1) = n^{-1}\sum_{i=1}^{n} \widehat{m}(x_1, X_{2i}) \tag{1.3}$$

is the empirical marginal integration estimator as described by Linton and Nielsen (1995) and Chen, Härdle, Linton, and Severance-Lossin (1996). The latter paper showed that $\widehat{\gamma}_1(x_1)$ converges in probability to

$$\gamma_1(x_1) = \int m(x_1, x_2)p_2(x_2)dx_2 = m_1(x_1) + c,$$

when (1.1) holds, where $p_2(\cdot)$ is the density function of X_2. Assuming that $h_o = O(n^{-1/(2q+1)})$, the rate of convergence of $\widehat{\gamma}_1(x_1)$ is $n^{q/(2q+1)}$, and its asymptotic distribution follows from the stochastic expansion

$$\widehat{\gamma}_1(x_1) - \gamma_1(x_1) = \frac{1}{nh_o}\sum_{i=1}^{n} k\left(\frac{x_1 - X_{1i}}{h_o}\right)\frac{p_2(X_{2i})}{p(X_{1i}, X_{2i})}\varepsilon_i + b_n(x_1; h_o) + o_p(n^{-q/(2q+1)}), \tag{1.4}$$

where $\varepsilon_i = Y_i - E(Y_i|X_i)$, while the term $b_n(x_1; h_o)$ is a deterministic bias of order h_o^q.

There are two main disadvantages of the integration estimator. First, it is perhaps even more time consuming to compute than the backfitting estimator. To evaluate $\widehat{\gamma}_1(\cdot)$ at the observations X_{1i}, $i = 1, \ldots, n$, one must compute n^2 regression smooths evaluated at the pairs (X_{1i}, X_{2j}), $i, j = 1, \ldots, n$, each of which requires $O(nh_o)$ operations when the kernel $K(\cdot)$ has compact support. Thus, $\{\widehat{\gamma}_1(X_{1i})\}_{i=1}^n$ takes $O(n^3 h_o)$ operations to compute. The required number of operations for many other choices of pilot estimators, including the local linear regression smoother, is of the same order. For the backfitting estimator, computational time is $O(n^2 h_o r)$—where r is the number of iterations required to achieve convergence—which generally requires less operations. The second disadvantage of the integration estimator is that it is statistically inefficient. A reasonable standard against which to judge any estimate of the component $m_1(x_1)$ is that provided by the so-called *oracle* estimator, which is the corresponding one-dimensional smoother of the (unobserved) data $Y_i - c - \sum_{j=2}^d m_j(X_{ji})$ against X_{1i}. By this standard, the integration estimator is inefficient, and very much so for correlated data. Linton (1996) considered the hybrid procedure of applying one backfitting iteration to a pilot integration estimator of the additive components. If the pilot estimator undersmooths in all the variables, then this two-step estimator is itself asymptotically normal with the same asymptotic variance and bias as the *oracle* estimator (undersmoothing is needed to eliminate the bias in the pilot estimator, but plays no role in determining the variance). Recently, Opsomer and Ruppert (1997) provided conditional mean squared error expansions for one version—in fact, the exact solution of an nd by nd system of equations based on local linear smoothing—of the "backfitting" estimator. Although their procedure has optimal—that is, oracle—variance, its bias is not oracle, and is not even design adaptive, despite the fact that their procedure is based on local linear estimation throughout. Apart from anything else, this shows that the differences between alternative implementations of the "backfitting" idea are as significant as, say, the difference between Nadaraya–Watson kernel and local linear regression smoothing estimation.

In this article, we suggest a state-of-the-art implementation of an efficient estimator of $m_1(\cdot)$. We first introduce a computationally convenient and consistent, but inefficient, estimator of $m_j(\cdot)$, $j = 2, \ldots, d$. We give an alternative motivation of our initial estimator as being of the instrumental variable type. It takes only $O(n^2 h_o)$ operations to compute. This reduction of $O(n)$ operations is practically significant when implementing computer-intensive methods such as the bootstrap or cross-validation. We then define a two-step estimator by plugging in the initial estimator into a backfitting-like iteration. The efficient estimator takes just twice as many operations as the initial estimator. We establish the equivalence of the two-step efficient estimator and the oracle estimator in various senses: in probability, with probability one, and in distribution with rate, pointwise as well as uniformly. This allows us to address many of the practical questions concerning implementation of the additive estimator, because we argue that the host of results established for one-dimensional regression can now be applied with impunity to our feasible procedure. Finally, we define a bootstrap confidence interval and establish its large sample coverage. For presentation simplicity we shall maintain that the additive structure (1.1) holds. Furthermore, in contradistinction to previous work—for example, Linton and Härdle (1996) who used bias reduction in the direction not of interest (at

least when the dimensionality is large)—we use a common bandwidth h_o and kernel k of order q, where q is an even integer, for each direction. Therefore, our conditions in the following require that q is related to d: specifically, when $d > 4$ we must have $q > 2$. Of course, better performance can be achieved by using multiple kernels and multiple bandwidths, but this adds what we consider unnecessary complication.

This article is organized as follows. Section 2 introduces and motivates a computationally efficient pilot estimator for the additive component which is used in Section 3 to produce an oracle estimator for the additive components. Bootstrap of the oracle estimator for constructing confidence intervals for the individual components is studied in Section 4. Section 5 presents a simulation study of the finite sample performance of the bootstrap algorithm and a rule-of-thumb bandwidth selection method. The proofs are given in the Appendix.

2. A FAST INSTRUMENTAL VARIABLE PILOT ESTIMATOR

We give an alternative heuristic for estimating the additive model in place of the usual integration idea presented in the introduction. We would really like to apply just a one-dimensional smoother, so consider the nonparametric regression of Y on X_1,

$$E(Y|X_1 = x_1) = c + m_1(x_1) + E\left[m_2(X_2)|X_1 = x_1\right].$$

Unfortunately, $E(Y|X_1 = x_1)$ is a biased estimator of $c + m_1(x_1)$ due to the presence of the unknown function $E\left[m_2(X_2)|X_1 = x_1\right]$ on the right side. Now suppose we can find some "instrument" function $w(x_1, x_2)$ [see Angrist, Imbens, and Rubin (1996), and the accompanying discussion for an interesting recent discussion of the econometric concept of instrument and some background literature], for which

$$E\left[w(X_1, X_2)|X_1 = x_1\right] = 1 \;\; ; \;\; E\left[w(X_1, X_2)m_2(X_2)|X_1 = x_1\right] = 0 \tag{2.1}$$

with probability one, then

$$E\left[w(X_1, X_2)Y|X_1 = x_1\right] = c + m_1(x_1), \tag{2.2}$$

as required. In fact, $w(x_1, x_2) = p_1(x_1)p_2(x_2)/p(x_1, x_2)$ is such a function, because

$$\begin{aligned} \int w(x_1, x_2)\frac{p(x_1, x_2)}{p_1(x_1)}dx_2 &= \int p_2(x_2)dx_2 = 1 \\ \int w(x_1, x_2)m_2(x_2)\frac{p(x_1, x_2)}{p_1(x_1)}dx_2 &= \int m_2(x_2)p_2(x_2)dx_2 = 0. \end{aligned}$$

In practice, we must replace $w(x_1, x_2)$ by an estimate and replace population conditional expectation by a regression smoother. Thus, we can compute a variety of one-dimensional smooths of $\widehat{w}(X_{1j}, X_{2j})Y_j$ on X_{1j}, where $\widehat{w}(X_{1j}, X_{2j}) = \widehat{p}_1(X_{1j})\widehat{p}_2(X_{2j})/\widehat{p}(X_{1j}, X_{2j})$, including kernel or local polynomial. We propose the somewhat simpler version of the kernel estimate that eliminates the superfluous explicit estimation of p_1,

$$\widehat{\gamma}_1^{pi}(x_1) \equiv \widehat{\gamma}_1^{pi}(x_1; h_o) = \frac{1}{nh_o}\sum_{j=1}^{n} k\left(\frac{x_1 - X_{1j}}{h_o}\right)\frac{\widehat{p}_2(X_{2j})}{\widehat{p}(X_{1j}, X_{2j})}Y_j \tag{2.3}$$

as our estimate of $\gamma_1(x_1)$. It has several interpretations in addition to the above instrumental variable estimate. First, as a version of the one-dimensional regression smoother but adjusting internally by a conditional density estimate

$$\widehat{p}_{1|2}(X_{1j}|X_{2j}) = \frac{\widehat{p}(X_{1j}, X_{2j})}{\widehat{p}_2(X_{2j})},$$

instead of by a marginal density estimate. Second, one can think of (2.3) as a one-dimensional standard Nadaraya–Watson (externalized) regression smoother of the adjusted data $\widehat{Y}_j$ on X_{1j}, where $\widehat{Y}_j = \widehat{p}_1(x_1)\widehat{p}_2(X_{2j})Y_j / \widehat{p}(X_{1j}, X_{2j})$. Finally, note that $\widehat{\gamma}_1^{pi}(X_{1i})$ can be interpreted as a marginal integration estimator in which the pilot estimator is a fully internalized smoother [see Jones, Davies, and Park (1994), who considered only the version where $\widehat{p}$ is replaced by p and its relation to the local linear smoother.]

$$\tilde{m}(x) = \frac{1}{nh_o^d} \sum_{j=1}^{n} K\left(\frac{x - X_j}{h_o}\right) Y_j / \widehat{p}(X_{1j}, X_{2j}),$$

rather than the Nadaraya–Watson: by interchanging the orders of summation, we obtain

$$\begin{aligned}
\widehat{\gamma}_1^{pi}(X_{1i}; h_o) &= \frac{1}{nh_o} \sum_{j=1}^{n} k\left(\frac{X_{1i} - X_{1j}}{h_o}\right) \frac{\widehat{p}_2(X_{2j})}{\widehat{p}(X_{1j}, X_{2j})} Y_j \\
&= \frac{1}{nh_o} \sum_{j=1}^{n} \frac{k\left(\frac{X_{1i} - X_{1j}}{h_o}\right) Y_j}{\widehat{p}(X_{1j}, X_{2j})} \left\{ \frac{1}{nh_o^{d-1}} \sum_{k=1}^{n} K_2\left(\frac{X_{2k} - X_{2j}}{h_o}\right) \right\} \\
&= \frac{1}{n^2 h_o^d} \sum_{k=1}^{n} \sum_{j=1}^{n} \frac{k\left(\frac{X_{1i} - X_{1j}}{h_o}\right) K_2\left(\frac{X_{2k} - X_{2j}}{h_o}\right) Y_j}{\widehat{p}(X_{1j}, X_{2j})} \\
&= \frac{1}{n} \sum_{k=1}^{n} \tilde{m}(X_{1i}, X_{2k}),
\end{aligned}$$

provided the kernel k is symmetric about zero. Here, $K_2(s_1, \ldots, s_{d-1}) = \prod_{\ell=1}^{d-1} k(s_\ell)$ is a $(d-1)$-dimensional kernel.

Our procedure (2.3) is computational convenient—it involves just simple smoother matrices and takes only order n smoothing operations. It also does not require evaluation of either density or regression function outside of the joint support. It has very similar asymptotic properties to the empirical marginal integration estimator—in particular, it converges in distribution at rate $n^{q/(2q+1)}$ to a normal random variable. We will show later how to improve on the efficiency of this estimate. Define $D^q g(x_1, \ldots, x_d) = \sum_{\ell=1}^{d} \partial^q g(x_1, \ldots, x_d) / \partial x_\ell^q$ for any positive integer q.

Theorem 1. *Suppose the conditions given in the Appendix hold, and that* $h_o = \delta_o n^{-1/(2q+1)}$. *Then,*

$$n^{q/(2q+1)} \left[\widehat{\gamma}_1^{pi}(x_1) - \gamma_1(x_1)\right] \longrightarrow N\left[b_1(x_1), v_1^2(x_1)\right]$$

in distribution, where

$$b_1(x_1) = \frac{\delta_o^q}{q!}\left[\mu_q(k)D^q m_1(x_1) + \int\left\{\mu_q(K)\frac{m(x)p_2(x_2)}{p(x)}D^q p(x) - \mu_q(K_2)m(x)D^q p_2(x_2)\right\}dx_2\right]$$

$$v_1^2(x_1) = \delta_o^{-1}\|k\|_2^2\int\frac{p_2^2(x_2)}{p(x)}\left\{\sigma^2(x)+m^2(x)\right\}dx_2$$

with $\sigma^2(x) = var(Y|X=x)$, *while* $\mu_q(k) = \int k(t)t^q dt$ *and* $\|k\|_2^2 = \int k(t)^2 dt$.

2.1 REMARKS

1. This estimator has an additional factor in the variance relative to the marginal integration estimator of Linton and Nielsen (1995), and is less efficient.
2. By using more bias reduction in the directions not of interest, as in Linton and Härdle (1996), one can eliminate some terms from the bias; specifically, those terms involving derivatives with respect to the direction not of interest are not present.
3. To estimate $m_1(x_1)$, we can use either of the following recentered estimates

$$\widehat{m}_1^{cd}(x_1) = \widehat{\gamma}_1^{pi}(x_1) - \overline{Y} \quad \text{or} \quad \widehat{m}_1^{ce}(x_1) = \widehat{\gamma}_1^{pi}(x_1) - \frac{1}{n}\sum_{i=1}^n \widehat{\gamma}_1^{pi}(X_{1i}).$$

   Since $\overline{Y} = \frac{1}{n}\sum_{i=1}^n Y_i = c + O_p(n^{-1/2})$, the bias and variance of $\widehat{m}_1^{cd}(x_1)$ are the same as those of $\widehat{\gamma}_1^{pi}(x_1)$. However, this is not the case for $\widehat{m}_1^{ce}(x_1)$; specifically, while the variance is the same as for $\widehat{\gamma}_1^{pi}(x_1)$, the bias of $\widehat{m}_1^{ce}(x_1)$ is $b_1(x_1) - \int b_1(x_1)p_1(x_1)dx_1$. Although $\widehat{m}_1^{ce}(\cdot)$ has (sample) mean zero by construction, $\widehat{m}_1^{cd}(\cdot)$ does not.
4. To estimate $m(x)$, we let

$$\widetilde{m}^{cd}(x) = \overline{Y} + \sum_{j=1}^d \widehat{m}_j^{cd}(x_j)] \;;\; \widetilde{m}^{ce}(x) = \frac{1}{nd}\sum_{j=1}^d\sum_{i=1}^n \widehat{\gamma}_j^{pi}(X_{ji}) + \sum_{j=1}^d \widehat{m}_j^{ce}(x_j).$$

   It is easy to see that the bias of $\widetilde{m}^{cd}(x)$ is the sum of the biases of $\widehat{\gamma}_j^{pi}(x_j)$, while the asymptotic bias of $\widetilde{m}^{ce}(x)$ is $\sum_{j=1}^d b_j(x_j) - \frac{d-1}{d}\int b_j(x_j)p_j(x_j)dx_j$. The estimates $\widehat{m}_j^{\ell}(x_j)$ and $\widehat{m}_k^{\ell}(x_k)$ are asymptotically uncorrelated for any j,k (this is true for both $\ell = cd$ and $\ell = ce$), so that the asymptotic variance of $\widetilde{m}^{\ell}(x)$ is just the sum of the asymptotic variances of $\widehat{m}_j^{\ell}(x_j)$.

3. ORACLE ESTIMATION OF ADDITIVE COMPONENTS

Suppose an oracle tells us all but one of the one-dimensional components $\{m_\ell(x_\ell) : \ell \neq j\}$. To estimate $m_j(x_j)$, construct $Y_{ji}^{\text{oracle}} = Y_i - \sum_{\ell\neq j} m_\ell(X_{\ell i}) - c$, and consider

$\widehat{m}_j^{\text{oracle}}(x_j)$, a univariate local polynomial smoother of the pairs $(X_{ji}, Y_{ji}^{\text{oracle}})$; that is, $\widehat{m}_j^{\text{oracle}}(x_j) = \widehat{a}_0(x_j)$, where $\widehat{a}_0(x_j), \widehat{a}_1(x_j), \ldots, \widehat{a}_{q-1}(x_j)$ solve

$$\min_{a_0,a_1,\ldots,a_{q-1}} \sum_{i=1}^{n} k\left(\frac{x_j - X_{ji}}{h}\right)\left[Y_{ji}^{\text{oracle}} - \sum_{\ell=0}^{q-1} a_\ell \left(X_{ji} - x_j\right)^\ell\right], \tag{3.1}$$

with $h = \delta n^{-1/(2q+1)}$ a bandwidth sequence. Under the conditions given in Fan (1992, 1993),

$$n^{q/(2q+1)}\left\{\widehat{m}_j^{\text{oracle}}(x_j) - m_j(x_j)\right\} \longrightarrow N\left[b_{ej}(x_j), v_{ej}^2(x_j)\right]$$

in distribution, where

$$b_{ej}(x_j) = \delta^q \frac{\overline{\mu}_q(k)}{q!} D^q m_j(x_j) \;\; ; \;\; v_{ej}^2(x_j) = \delta^{-1}\overline{\nu}_q(k)\frac{\sigma_j^2(x_j)}{p_j(x_j)},$$

in which $\sigma_j^2(x_j) = \text{var}(Y|X_j = x_j)$, while $\overline{\mu}_q(k)$ and $\overline{\nu}_q(k)$ are certain constants derived from k—see Fan and Gijbels (1992) for elucidation, and see Fan (1993) for a discussion of the minimax efficiency of this procedure. In fact, one can improve slightly on this method by using the fact that $\int m_j(x_j)p_j(x_j)dx_j = 0$. Let $\widehat{m}_j^{\text{coracle}}(x_j) = \widehat{m}_j^{\text{oracle}}(x_j) - n^{-1}\sum_{i=1}^n \widehat{m}_j^{\text{oracle}}(X_{ji})$, then

$$n^{q/(2q+1)}\left\{\widehat{m}_j^{\text{coracle}}(x_j) - m_j(x_j)\right\} \longrightarrow N\left[b_{cej}(x_j), v_{cej}^2(x_j)\right]$$

in distribution, where $v_{cej}^2(x_j) = v_{ej}^2(x_j)$, and

$$b_{cej}(x_j) = \delta^q \frac{\overline{\mu}_q(k)}{q!}\left\{D^q m_j(x_j) - \int D^q m_j(x_j)p_j(x_j)dx_j\right\}.$$

Note that $\widehat{m}_j^{\text{coracle}}(\cdot)$ is more efficient than $\widehat{m}_j^{\text{oracle}}(\cdot)$ according to integrated mean squared error, because

$$\text{var}_{X_j}\left\{D^q m_j(X_j)\right\} \leq E_{X_j}\left[\left\{D^q m_j(X_j)\right\}^2\right].$$

We call $\widehat{m}_j(x_j)$ an *oracle estimator* for $m_j(x_j)$ if it behaves (asymptotically) like $\widehat{m}_j^{\text{oracle}}(x_j)$. While in parametric regression oracle estimators for the slope parameter only exist when the covariate X_j is uncorrelated with $\{X_\ell : \ell \neq j\}$, we show that for additive regression, applying one step of the backfitting algorithm to an integration estimator produces an oracle estimator. Specifically, denote by $\widehat{\gamma}_\ell^{pi}(x_\ell; h_o)$ the integration estimator of the one-dimensional components described in Section 2 with bandwidth h_o, and set

$$\widetilde{Y}_{ji}^{2-\text{step}} = Y_i - \sum_{\ell \neq j} \widehat{\gamma}_\ell^{pi}(X_{\ell i}; h_o) + (d-1)\overline{Y}.$$

Theorem 2 states that with appropriate choices for the bandwidths h_o and h, $\widehat{m}_j^{2-\text{step}}(x_j; h, h_o) = \widehat{a}_0(x_j)$, in which $\widehat{a}_0(x_j), \widehat{a}_1(x_j), \ldots, \widehat{a}_{q-1}(x_j)$ solve

$$\min_{a_0,a_1,\ldots,a_{q-1}} \sum_{i=1}^{n} k\left(\frac{x_j - X_{ji}}{h}\right)\left[\widetilde{Y}_{ji}^{2-\text{step}} - \sum_{\ell=0}^{q-1} a_\ell \left(X_{ji} - x_j\right)^\ell\right] 1(X_i \in \mathcal{S}^{on}), \tag{3.2}$$

is an oracle estimator for $m_j(x_j)$. Here, $\mathcal{S}^{on}$ is the intersection of the trimmed one-dimensional supports; that is,

$$\mathcal{S}^{on} = \left\{x \in \mathbb{R}^d : \underline{b}_j + h_o \leq x_j \leq \bar{b}_j - h_o \quad j = 1, \ldots, d\right\},$$

and where $\underline{b}_j, \bar{b}_j$ are the lower and upper bounds of the support of X_{ji} (we have assumed a rectangular support for simplicity). This trimming is for technical reasons. Our initial estimator $\sum_{\ell \neq j} \widehat{\gamma}_\ell^{pi}(X_{\ell i}; h_o) - (d-1)\overline{Y}$ has poor boundary bias behavior, specifically boundary bias of order h_o compared with interior bias of order h_o^q. The trimming eliminates the observation at the boundary at no first-order cost, provided the bandwidth h_o converges to zero.

Theorem 2. *Suppose the conditions in the Appendix hold, that* $h_o = o(n^{-1/(2q+1)})$, *and that* $h = a_o n^{-1/(2q+1)}$ *for some* $a_o > 0$. *Then, for all* ϵ,

$$\Pr\left[n^{q/(2q+1)}\left\{\widehat{m}_j^{2-\text{step}}(x_j; h, h_o) - \widehat{m}_j^{\text{oracle}}(x_j; h)\right\} > \epsilon \middle| \mathcal{X}^n\right] \longrightarrow 0$$

with probability one, as $n \to \infty$. *Here,* $\mathcal{X}^n = \{X_1, \ldots, X_n\}$.

A similar theorem for a marginal integration estimator with an externally adjusted pilot can be found in Linton (1996). The immediate reason for this result is that $\widehat{m}_j^{2-\text{step}}$ is asymptotically independent of $\widehat{m}_k^{2-\text{step}}$ for all $k \neq j$, or rather the correlation between these two random variables is of order n^{-1} which is of smaller order than the variance of either. Thus, $\widehat{m}_j^{2-\text{step}}$ inherits, to first order, the asymptotic properties of $\widehat{m}_j^{\text{oracle}}(x_j)$. The reason for the low correlation can be understood from the simpler problem of comparing two marginal smooths on variables X_j and X_k. The marginal kernel windows contain $O(nh)$ observations, while the intersection of the marginal windows (which determines the correlation between the two estimators) contains $O(nh^2)$ observations.

3.1 REMARKS

1. Under slightly stronger conditions—specifically $E\left(|Y|^s\right) < \infty$ for some $s > 2$—it is true that

$$n^{q/(2q+1)} \sup_{x_j} \left|\widehat{m}_j^{2-\text{step}}(x_j; h, h_o) - \widehat{m}_j^{\text{oracle}}(x_j; h)\right| \longrightarrow 0$$

with probability one (the supremum is taken over any compact set contained in the support of X_j); see Masry (1996). Combining this with the well-known optimal rate of uniform convergence for $\widehat{m}_j^{\text{oracle}}$; that is, $n^{q/(2q+1)}/\log n$, gives the rate for $\widehat{m}_j^{2-\text{step}}$.

2. If one uses the same bandwidth $h = h_o = \delta_o n^{-q/(2q+1)}$ in both (2.3) and (3.2), then

$$n^{q/(2q+1)}\left\{\widehat{m}_j^{2-\text{step}}(x_j) - \widehat{m}_j^{\text{oracle}}(x_j)\right\} \longrightarrow -\sum_{k \neq j}^{d} \int b_k(x_k) \frac{p_{j,k}(x_j, x_k)}{p_j(x_j)} dx_k$$

in probability, where $p_{j,k}$ is the joint density of X_j and X_k. Hence, $\widehat{m}_j^{2-\text{step}}(x_j)$

has the same variance as $\widehat{m}_j^{\text{oracle}}(x_j)$, but has bias

$$b_{e1}(x_1) - \sum_{j=2}^{d} \int b_j(x_j) \frac{p_{1,j}(x_1, x_j)}{p_1(x_1)} dx_j.$$

This procedure is evidently inferior, according to a minimax criterion, than when $h_o/h \longrightarrow 0$. We conjecture that by iterating to infinity, with $h = h_o$, one obtains the same bias as when $h_o/h \longrightarrow 0$.

Similar calculations can be performed for the centred estimates as well as for the estimates of $m(x)$.

4. BOOTSTRAP CONFIDENCE INTERVALS

The bootstrap distribution of studentized estimators usually improves upon the Normal approximation by capturing the first term of the Edgeworth expansion. In this section, we investigate the bootstrap for the two-stage estimator $\widehat{m}_j(x_j; h, h_o)$ introduced in the previous section, where h_o is the bandwidth of the pilot estimator $\widehat{\gamma}^{pi}(x_j)$ and h is the bandwidth of the local polynomial regression in the second step. Define the additive reconstruction

$$\widehat{m}_{\text{add}}(x; h, h_o) = \overline{Y} + \sum_{\ell=1}^{d} \widehat{m}_\ell(x_\ell; h, h_o),$$

and the residuals

$$e_i = Y_i - \widehat{m}_{\text{add}}(X_i; h, h_o), \qquad i = 1, \ldots, n.$$

The naive bootstrap, which samples with replacement from the centered residuals, leads to a bootstrap distribution with a different bias than that of the estimator. As a remedy, Härdle and Marron (1991) adapted an idea of Wu (1986) and proposed drawing $\varepsilon_1^\dagger, \varepsilon_2^\dagger, \ldots, \varepsilon_n^\dagger$ independently from the two point distributions $dG_i = \gamma_i \delta_{a_i} + (1-\gamma_i)\delta_{b_i}$, where δ is the Dirac delta function, $a_i = e_i^c(1+\sqrt{5})/2$, $b_i = e_i^c(1-\sqrt{5})/2$, where $e_i^c = e_i - \sum_j e_j/n$, and $\gamma_i = (5+\sqrt{5})/10$, to construct the bootstrap sample

$$Y_j^\dagger = \widehat{m}_{\text{add}}(x; g, h_o) + \varepsilon_j^\dagger.$$

To see why yet another bandwidth g is introduced for constructing the bootstrap sample, observe that $\widehat{m}_\ell(x_\ell; h, h_o)$ and its bootstrap counterpart $\widehat{m}_\ell^\dagger(x_\ell; h, h_o)$ have the same variances, but that their biases are

$$E\left[\widehat{m}_\ell(x_\ell; h, h_o) - m_\ell(x_\ell)\right] = h^q C(K) \cdot D^q m_\ell(x_\ell) + o(h^q),$$

and

$$E^\dagger\left[\widehat{m}_\ell^\dagger(x_\ell; h, h_o) - \widehat{m}_\ell(x_\ell; g, h_o)\right] = h^q C(K) \cdot D^q \widehat{m}_\ell(x_\ell; g, h_o) + o(h^q),$$

respectively, where $C(K)$ is a constant depending on the kernel. Here, the expectation $E^\dagger$ is understood to be conditionally on the data. Hence, the asymptotic distribution of

$$\sqrt{nh}\left\{\widehat{m}_\ell^\dagger(x_\ell; h, h_o) - \widehat{m}_\ell(x_\ell, g, h_o)\right\},$$

and

$$\sqrt{nh}\left\{\widehat{m}_\ell(x_\ell; h, h_o) - m_\ell(x_\ell)\right\}$$

will agree if $D^q\widehat{m}_\ell(x_\ell; g, h_o)$ converges to $D^q m_\ell(x_\ell)$. A bandwidth $g(n)$ of larger order than $n^{-1/(2q+1)}$ is needed to ensure this.

Our analysis relies on a strengthening of Theorem 2 to show that the bootstrap distributions of the two-step and the oracle estimators are asymptotically equivalent. Properties of the bootstrap of the two-step estimator follows from established theorems for one-dimensional nonparametric regression; see Härdle and Marron (1991).

Theorem 3. *In addition to the conditions in the Appendix, suppose that* $E\left(|Y|^s\right) < \infty$ *for some* $s > 2$, *and that* $\left(\frac{h_o}{h}\right)^q / \lambda_n \to 0$. *Then, there exists a random variable* $a_1(\mathcal{X}^n)$ *that is positive and finite with a probability tending to one, such that*

$$\Pr\left[n^{q/(2q+1)}|\widehat{m}_1^{2-\text{step}}(x_1) - \widehat{m}_1^{\text{oracle}}(x_1)| > \lambda_n|\mathcal{X}^n\right] \leq a_1\lambda_n^{-s}n^{-s/2(2q+1)}. \tag{4.1}$$

We have using standard arguments (see, e.g., Hall 1992, p. 292), that for any $\lambda_n \to 0$,

$$\begin{aligned} G(z; h, h_o) &= \Pr\left[n^{q/(2q+1)}\left\{\widehat{m}_j^{2-\text{step}}(x_j; h, h_o) - m_j(x_j)\right\} \leq z \middle| \mathcal{X}^n\right] \\ &= \Pr\left[n^{q/(2q+1)}\left\{\widehat{m}_j^{\text{oracle}}(x_j; h, h_o) - m_j(x_j)\right\} \leq z \middle| \mathcal{X}^n\right] + O(\lambda_n) \\ &\quad + O\left(\Pr\left[n^{q/(2q+1)}\left|\widehat{m}_j^{2-\text{step}}(x_j; h, h_o) - \widehat{m}_j^{\text{oracle}}(x_j; h)\right| > \lambda_n \middle| \mathcal{X}^n\right]\right). \end{aligned}$$

Applying the theorem with $\lambda_n = 1/\log n$, for example, the remainders are $o(1)$, for $s \geq 2$. A similar approximation holds for the bootstrap distribution

$$\begin{aligned} G^\dagger(z; h, h_o, g) &= \overset{\dagger}{\Pr}\left[n^{q/(2q+1)}\left\{\widehat{m}_j^\dagger(x_j; h, h_o) - \widehat{m}_j(x_j, g, h_o)\right\} \leq z\right] \\ &= \overset{\dagger}{\Pr}\left[n^{q/(2q+1)}\left\{\widehat{m}_j^{\text{oracle},*}(x_j; h, h_o) - \widehat{m}_j^{\text{oracle}}(x_j, g, h_o)\right\} < z\right] + o(1), \end{aligned}$$

where $\Pr^\dagger$ is computed conditionally on the data. Consistency of the bootstrap distribution for the two-stage estimator then follows from the consistency of the bootstrap of the oracle estimator.

Theorem 4. *Suppose that in addition to the assumptions set forth in the Appendix:* $h_o = o(n^{-1/(2q+1)})$, $H_n = [\underline{a}n^{-1/(2q+1)}, \overline{a}n^{-1/(2q+1)}]$, *and* $G_n = [n^{-1/(2q+1)+\delta}, n^{-\delta}]$, *where* $0 < \delta < 1/(2q+1)$, *while* $0 < \underline{a} \leq \overline{a} < \infty$. *Then,*

$$\sup_{g \in G_n}\sup_{h \in H_n}\left|G^\dagger(x; h_o, h, g) - G(x; h_o, h)\right| \longrightarrow 0$$

with probability one.

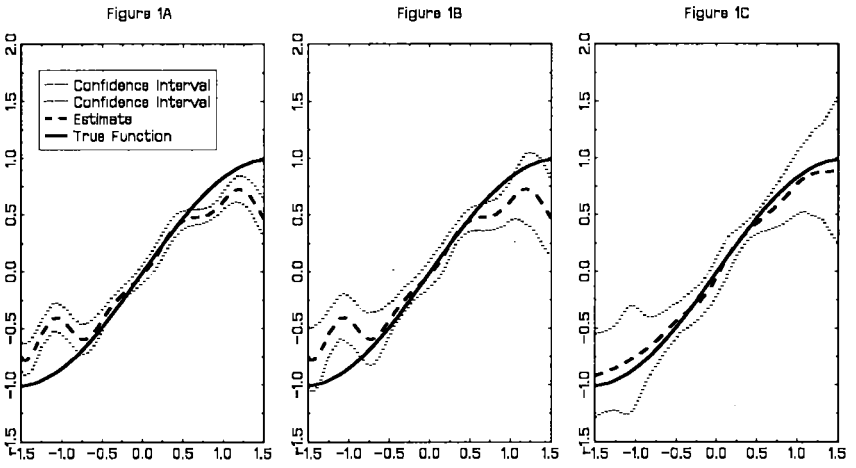

Figure 1. Figure 1(a) shows the (preliminary) fast additive smoother $\widehat{\gamma}_1$ along with 95% asymptotic confidence intervals. Figure 1(b) shows the same estimator with bootstrap confidence intervals. Figure 1(c) shows the efficient fast additive smoother $\widehat{\gamma}_1^e$ with bootstrap confidence intervals. In each figure, the true function is shown as a solid line.

4.1 REMARK

The uniformity over H_n and G_n implies that the bootstrap of estimators with data-dependent bandwidth is consistent.

5. NUMERICAL RESULTS

We investigate the bootstrap confidence intervals on simulated data. We took the following setup

$$Y_i = \sin X_{1i} + \sin X_{2i} + \sigma\varepsilon_i, \quad i = 1, \ldots, n,$$

where X_i is uniformly distributed on $[-\pi, \pi]^2$ and ε_i is standard normal independent of X_i, while $\sigma = .1$. We computed the estimator using the following matrix commands, which makes for speedy processing in programs like Gauss or Matlab. Defining the $n \times n$ smoother matrices

$$S_1 = \left[\frac{1}{nh} k\left(\frac{X_{1i} - X_{1j}}{h}\right)\right]_{i,j} ; \; S_2 = \left[\frac{1}{nh} k\left(\frac{X_{2i} - X_{2j}}{h}\right)\right]_{i,j},$$

we can write the $n \times 1$ vector of estimates $\widehat{\gamma}_1 = (\widehat{\gamma}_1(X_{11}), \ldots, \widehat{\gamma}_1(X_{1n}))^T$ as

$$\widehat{\gamma}_1 = S_1 \left\{ y \odot (S_2 i)./(n(S_1 \odot S_2)i) \right\},$$

in which $\odot$ and $./$ denote matrix Hadamard product and division, respectively, while $i = (1, \ldots, 1)^T$ and $y = (Y_1, \ldots, Y_n)^T$. By comparison, the marginal integration estimate involves n smoother matrices. We also computed the efficient smoothers $\widehat{\gamma}_1^e$ and $\widehat{\gamma}_2^e$ using $\widehat{\gamma}_1$ and $\widehat{\gamma}_2$ as starting values; that is,

$$\widehat{\gamma}_1^e = S_1 y_1^e./(S_1 i) \; ; \; y_1^e = y - \widehat{\gamma}_2 - \overline{Y} i.$$

Table 1. Rejection Frequency of Bootstrap Intervals ($n = 50$)

	c_{FB}			c_{EB}		
c	1%	5%	10%	1%	5%	10%
.10	.5144	.5882	.6258	.0010	.0026	.0050
.30	.3960	.4550	.4960	.0046	.0096	.0168
.50	.2062	.2598	.3072	.0180	.0324	.0470
.55	.1754	.2252	.2712	.0212	.0400	.0602
.60	.1584	.2008	.2402	.0286	.0508	.0694
.65	.1394	.1798	.2180	.0330	.0588	.0822
.75	.1262	.1656	.1944	.0538	.0796	.1004
1.00	.1308	.1722	.2108	.0862	.1088	.1282
2.00	.4346	.5208	.5800	.1584	.1664	.1712

NOTE: c_{FB} denotes the rejection frequency of the confidence interval of the preliminary fast additive smoother is denoted, while c_{EB} denotes this quantity for the efficient fast additive smoother. c in the first column denotes the bootstrap bandwidth constant; that is, the actual bandwidth is $g = c \times \sigma\, n^{-1/9}$.

We estimated the regression function vector $m = (m(X_1), \ldots, m(X_n))^T$ by $\widetilde{m} = \widehat{\gamma}_1 + \widehat{\gamma}_2 - \overline{Y}i$ and by the efficient version $\widetilde{m}^e = \widehat{\gamma}_1^e + \widehat{\gamma}_2^e - \overline{Y}i$.

We took $h_j = .5\widehat{\sigma}_{X_j} n^{-1/5}$, $j = 1, 2$, as the bandwidths for estimation, where $\widehat{\sigma}^2_{X_j} = n^{-1}\sum_{i=1}^n (X_{ji} - \overline{X}_j)^2$, and we examined a grid of bootstrap bandwidths $g = c\widehat{\sigma}_{X_j} n^{-1/9}$, for c in the range $.1 - 2.0$ using $nb = 200$ bootstrap samples. Figure 1 shows the estimation results from a typical sample along with bootstrap confidence intervals. The preliminary estimator is badly biased, and the asymptotic confidence intervals seem way too narrow. The bootstrap bands seem to be wide enough in both cases. Finally, the efficient estimator has much smaller bias.

Tables 1 and 2 give rejection frequencies for the 1%, 5%, and 10% level bootstrap confidence bands for the fast preliminary estimator, c_{FB}, and the efficient estimator, c_{EB}. The confidence interval is centered at a single random point (which changes from sample to sample) and there were 5,000 replications.

For c_{FA}—that is, the asymptotic confidence interval based on the fast additive smoother—the rejection frequencies are .5764, .6938, and .7498, respectively, when $n = 50$ and .5446, .6740, and .7368, respectively, when $n = 100$.

We should point out that no attempt has been made to optimize the estimation bandwidths h_j—their specific value was chosen by their visual performance for the

Table 2. Rejection Frequency of Bootstrap Intervals ($n = 100$)

	c_{FB}			c_{EB}		
c	1%	5%	10%	1%	5%	10%
.10	.4838	.5488	.5898	.0002	.0006	.0008
.30	.3006	.3662	.4078	.0026	.0080	.0118
.50	.1278	.1682	.2012	.0134	.0274	.0410
.55	.1088	.1476	.1790	.0170	.0336	.0484
.60	.0998	.1334	.1614	.0230	.0420	.0582
.65	.0914	.1216	.1482	.0288	.0518	.0700
.75	.0848	.1140	.1394	.0442	.0668	.0830
1.00	.0940	.1204	.1464	.0652	.0842	.1002
2.00	.3658	.4686	.5362	.0934	.1004	.1038

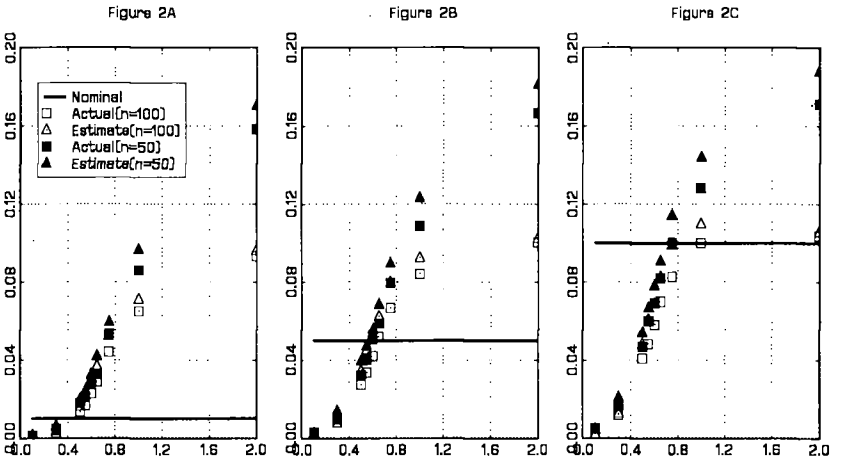

Figure 2. Figure 2(a) shows the rejection frequency for the 1% confidence bands constructed with the bandwidth $g = c \times \sigma n^{-1/9}$ *as c varies. This is the same information given in Table 1, and we use the term "actual" to denote it. We also give the "estimated" trade-off which was obtained using the rule-of-thumb method described in the text. Figure 2(b) shows the same information for the 5% confidence bands, and Figure 2(c) shows the same information for the 10% confidence bands. We show both the* $n = 50$ *(solid symbols) and* $n = 100$ *(hollow symbols).*

efficient estimator. No doubt, better performance could be achieved for the fast estimator by careful choice of bandwidth.

The main conclusion from these simulations is that the efficient estimator in conjunction with the bootstrap method performs quite well, provided a good resampling bandwidth is chosen. We now evaluate a practical method for selecting the bootstrap bandwidth based on an elaboration of Silverman's rule-of-thumb idea. More sophisticated bandwidth selection rules, such as those reviewed by Härdle (1990) and Jones, Marron, and Sheather (1996), can also be applied here, but our purpose here is just to investigate one very simple method. The approach is to specify a parametric model for the conditional distribution $\mathcal{L}(Y_i|X_i,\theta)$ for the purpose of selecting g. One estimates $\widehat{\theta}$ by maximum likelihood and then generates samples $\left\{\widetilde{Y}_i\right\}_{i=1}^{n}$ from $\mathcal{L}(Y_i|X_i,\widehat{\theta})$, which are used to perform a simulation experiment like we did above to determine an optimal bandwidth. Figure 2 reports the results of such a method taking a Gaussian quartic regression model as the conditional model and compares it with the infeasible bandwidth selection method based on the true design. Everything else is as before. The "estimated" trade-off quite closely mimics the "actual" trade-off, although there is a tendency towards slight over-coverage of the interval by this method. This would presumably be reduced by taking a richer parametric model for the pilot.

6. CONCLUDING REMARKS

Our work might be interpreted as supporting the use of the standard backfitting method of estimating additive models. However, we point out that there are some substantial differences between the asymptotic properties of different versions of the "back-

fitting" method. The one such method that has been successfully analyzed from a statistical point of view (see Opsomer and Ruppert 1997) differs from the oracle estimator in its bias, and in fact is not design adaptive—except when the covariates are mutually independent—even if local linear estimation is used throughout. By contrast, our two-step estimator has the same asymptotic variance and bias as the oracle estimator—that is, it is design adaptive. It is thus minimax superior to the Opsomer and Ruppert (1997) procedure. Furthermore, our two-step estimator is much less computationally demanding than their method. Finally, while we require stronger smoothness/dimensionality tradeoff than they did, we have not restricted the dependence among the covariates as they did.

APPENDIX: REGULARITY CONDITIONS AND PROOFS

Let $\mathcal{X}^n = \{X_1, \ldots, X_n\}$. We use the following regularity conditions:

A1. The kernel k is symmetric about zero and of order q; that is, $\int k(u)u^j du = 0$, $j = 1, \ldots, q-1$. Furthermore, k is supported on $[-1, 1]$, bounded, and Lipschitz continuous; that is, there exists a finite constant c such that $|k(u) - k(v)| \leq c|u - v|$ for all u, v.

A2. The functions $m(\cdot)$ and $p(\cdot)$are q-times continuously differentiable in each direction, where $q \geq (d-1)/2$.

A3. The joint density $p(\cdot)$ is bounded away from zero and infinity on its compact support, which is $\times_{j=1}^{d}[\underline{b}_j, \overline{b}_j]$.

A4. The conditional variance $\sigma^2(x) = \text{var}(Y|X = x)$ is Lipschitz continuous, and is bounded away from zero and infinity.

Proof of Theorem 1: We start with decomposing the estimation error as,

$$\widehat{\gamma}_1^{pi}(x_1) - \gamma_1(x_1) = \frac{1}{n}\sum_{k=1}^{n}\frac{1}{h_0}k\left(\frac{X_{1k} - x_1}{h_0}\right)\frac{\widehat{p}_2(X_{2k})}{\widehat{p}(X_{1k}, X_{2k})}Y_k - \gamma_1(x_1) = S_{1n} + S_{2n} + S_{3n},$$

where

$$\begin{aligned}
S_{1n} &\equiv \frac{1}{n}\sum_{k=1}^{n}\frac{1}{h_0}k\left(\frac{X_{1k} - x_1}{h_0}\right)\frac{\widehat{p}_2(X_{2k})}{\widehat{p}(X_{1k}, X_{2k})}\varepsilon_k, \\
S_{2n} &\equiv \frac{1}{n}\sum_{k=1}^{n}\frac{1}{h_0}k\left(\frac{X_{1k} - x_1}{h_0}\right)\frac{\widehat{p}_2(X_{2k})}{\widehat{p}(X_{1k}, X_{2k})}\left[m_1(X_{1k}) - m_1(x_1)\right], \\
S_{3n} &\equiv \frac{1}{n}\sum_{k=1}^{n}\left\{\frac{1}{h_0}k\left(\frac{X_{1k} - x_1}{h_0}\right)\frac{\widehat{p}_2(X_{2k})}{\widehat{p}(X_{1k}, X_{2k})}m(x_1, X_{2k}) - \gamma_1(x_1)\right\}.
\end{aligned}$$

1. We first examine the term S_{1n}. Note that $E\left(S_{1n}|\chi^n\right) = 0$, while

$$\text{var}\left(S_{1n}|\chi^n\right) = \frac{1}{nh_0}\left\{\frac{1}{nh_0}\sum_{k=1}^{n}k^2\left(\frac{X_{1k} - x_1}{h_0}\right)\frac{\widehat{p}_2^2(X_{2k})}{\widehat{p}^2(X_{1k}, X_{2k})}\sigma^2(X_k)\right\}$$

$$= \frac{1}{nh_0}\left\{\frac{1}{nh_0}\sum_{k=1}^{n} k^2\left(\frac{X_{1k}-x_1}{h_0}\right)\frac{\widehat{p}_2^2(X_{2k})}{\widehat{p}^2(X_{1k},X_{2k})}\sigma^2(X_k)\right\}\left[1+o_p(1)\right],$$

by the uniform convergence of $\widehat{p}$ to p and $\widehat{p}_2$ to p_2 [which follows under our conditions, see Masry (1996)]. Then, applying a law of large numbers and changing variables, we get that

$$v_1'(x_1) \equiv \operatorname{var}\left(\sqrt{nh_0}S_{1n}\right) = \|k\|_2^2 \int \frac{p_2^2(z_2)}{p(x_1,z_2)}\sigma^2(x_1,z_2)dz_2\left[1+o(1)\right].$$

In fact, applying the Lindeberg central limit theorem,

$$\sqrt{nh_0}S_{1n} \longrightarrow N\left[0, v_1'(x_1)\right],$$

in distribution—the Lindeberg condition is satisfied due to $k(\cdot)$ having bounded support.

2. S_{2n} : We approximate S_{2n} by $\frac{1}{n}\sum_{k=1}^{n}\frac{1}{h_0}k(\frac{X_{1k}-x_1}{h_0})\frac{\overline{p}_2(X_{2k})}{\overline{p}(X_{1k},X_{2k})}\left[m_1(X_{1k})-m_1(x_1)\right]$ with an error of $o_p(\frac{1}{\sqrt{nh_0}})$, where $\overline{p}(x_1,x_2)=E(\widehat{p}(x_1,x_2))=\int\frac{1}{h_0^q}K(\frac{z-x}{h_0})(z)dz$, and $\overline{p}_2(x_2)=E(\widehat{p}_2(x_2))=\int\frac{1}{h_0^{d-1}}K_2(\frac{z_2-x_2}{h_0})p_2(z_2)dz_2$. By a Taylor expansion,

$$\begin{aligned}
B_{2n} &\equiv E(S_{2n}) = \int \frac{1}{h_0}k\left(\frac{z_1-x_1}{h_0}\right)\frac{\overline{p}_2(z_2)}{\overline{p}(z_1,z_2)}\left[m_1(z_1)-m_1(x_1)\right]p(z_1,z_2)dz_1dz_2 \\
&= \int \frac{1}{h_0}k\left(\frac{z_1-x_1}{h_0}\right)\overline{p}_2(z_2)\left[m_1(z_1)-m_1(x_1)\right]dz_1dz_2 \\
&\quad -\int \frac{1}{h_0}k\left(\frac{z_1-x_1}{h_0}\right)\frac{\overline{p}_2(z_2)}{\overline{p}(z_1,z_2)}\left[m_1(z_1)-m_1(x_1)\right] \\
&\quad \times\left[\overline{p}(z_1,z_2)-p(z_1,z_2)\right]dz_1dz_2 \\
&= h_0^q\frac{\mu_q(k)}{q!}D^q m_1(x_1)+o(h_0^q).
\end{aligned}$$

The variance of S_{2n} is negligible, having the order of $O\left(\frac{h_0}{n}\right)$. To understand this, consider

$$\begin{aligned}
\operatorname{var}(S_{2n}) &= \frac{1}{n}\int\frac{1}{h_0^2}k^2\left(\frac{z_1-x_1}{h_0}\right)\frac{\overline{p}_2^2(z_2)}{\overline{p}^2(z_1,z_2)} \\
&\quad\times\left[m_1(z_1)-m_1(x_1)\right]^2 p(z_1,z_2)dz_1dz_2-\frac{1}{n}E^2(B_n) \\
&= \frac{h_0}{n}\int k^2(u)\frac{\overline{p}_2^2(z_2)m_1'^2(\widetilde{z}_1)u_1^2}{\overline{p}^2(x_1-hu_1,z_2)}p(x_1-h_0u_1,z_2)du_1dz_2 \\
&\quad +O\left(\frac{h_0^{2q}}{n}\right)=O\left(\frac{h_0}{n}\right),
\end{aligned}$$

where $\widetilde{z}_1$ is a point between x_1 and $x_1-h_0u_1$.

3. S_{3n}: Using the fact that $\frac{1}{n}\sum_{k=1}^{n}\frac{1}{h_0}k\left(\frac{X_{1k}-x_1}{h_0}\right)\left[\frac{\widehat{p}_2(X_{2k})}{\widehat{p}(X_{1k},X_{2k})}-\frac{\overline{p}_2(X_{2k})}{\overline{p}(X_{1k},X_{2k})}\right]m(x_1,X_{2k})=o_p(\frac{1}{\sqrt{nh_0}})$, we get an approximation of S_{3n} :

$$S_{3n}=S_{3n}'+\frac{1}{n}\sum_{k=1}^{n}m_2(X_{2k})+o_p\left(\frac{1}{\sqrt{nh_0}}\right)=S_{3n}'+O_p\left(\frac{1}{\sqrt{n}}\right)+o_p\left(\frac{1}{\sqrt{nh_0}}\right),$$

where $S'_{3n} = \frac{1}{n}\sum_{k=1}^{n}\left\{\frac{1}{h_0}k(\frac{X_{1k}-x_1}{h_0})\frac{\overline{p}_2(X_{2k})}{\overline{p}(X_{1k},X_{2k})} - 1\right\}m(x_1, X_{2k})$. The approximation in the second equation is a direct result from the Lindeberg central limit theorem, which holds under our assumptions A2, A3. Now, let

$$S'_{3n} = S''_{3n} + B_{3n} \equiv \frac{1}{n}\sum_{k=1}^{n}\widetilde{\zeta}_k + \overline{\zeta},$$

where $\zeta_k \equiv \left\{\frac{1}{h_0}k(\frac{X_{1k}-x_1}{h_0})\frac{\overline{p}_2(X_{2k})}{\overline{p}(X_{1k},X_{2k})} - 1\right\}m(x_1, X_{2k})$, $\overline{\zeta} \equiv E(\zeta_k)$, and $\widetilde{\zeta}_k \equiv \zeta_k - \overline{\zeta}$. A simple calculation gives the additional bias term as

$$\begin{aligned} B_{3n} &= \int \frac{1}{h_0}k(\frac{z_1-x_1}{h_0})\frac{\overline{p}_2(z_2)}{\overline{p}(z_1,z_2)}m(x_1,z_2)p(z_1,z_2)dz_1dz_2 - \gamma_1(x_1) \\ &= h_0^q\frac{\mu_q(K_2)}{q!}\int D^q p_2(z_2)m(x_1,z_2)dz_2 \\ &\quad -h_0^q\frac{\mu_q(K)}{q!}\int \frac{p_2(z_2)}{p(x_1,z_1)}m(x_1,z_2)D^q p(x_1,z_2)dz_2 + o(h_0^q) \end{aligned}$$

Next, we turn to $S''_{3n} = \frac{1}{n}\sum_{k=1}^{n}\widetilde{\zeta}_k$. It is mean zero by construction. Its variance is simply equal to $E\left[\frac{1}{n^2}\sum_{k=1}^{n}\widetilde{\zeta}_k^2\right]$, since the covariance terms disappear due to the independence X_k and X_l. That is,

$$\text{var}\left(S''_{3n}\right) = \frac{1}{n}E\left(\zeta_k^2\right) - \frac{1}{n}\overline{\zeta}^2 = \frac{1}{nh_0}\|k\|_2^2\int\frac{p_2^2(z_2)}{p(x_1,z_2)}m^2(x_1,z_2)dz_2 + O\left(\frac{1}{n}\right).$$

Hence, applying the Lindeberg CLT theorem for $\overline{S}''_{3n}$, it follows that

$$\sqrt{nh_0}S''_{3n} \longrightarrow N\left(0, \|k\|_2^2\int\frac{m^2(x_1,z_2)p_2^2(z_2)}{p(x_1,z_2)}dz_2\right)$$

in distribution. Finally, combining all the results above for S_{1n}, B_{2n}, B_{3n}, and S''_{3n}, together with $E\left(S_{1n}\overline{S}''_{3n}\right) = E\left(\overline{S}''_{3n}E\left(S_{1n}|\mathcal{X}^n\right)\right) = 0$, we get the asymptotic distribution of $\sqrt{nh_0}\left[\widehat{\gamma}_1^{pi}(x_1) - \gamma_1(x_1)\right]$.

Proof of Theorem 2: To prove Theorem 2, we need to determine the equivalent kernel of the local polynomial regression. Such a representation is stated in Lemma 1 whose proof can be found in Wand and Jones (1995).

Lemma 1. *The weights $W_{n,i}$ of the univariate local polynomial regression satisfy*

$$W_{n,i}(h) = \frac{1}{nh}\frac{1}{p_1(x_1)}k_\lambda^*\left(\frac{x_1 - X_{1i}}{h}\right) + o_p(1),$$

where

$$k_\lambda^*(u) = \sum_{t=0}^{q-1}s_{\lambda t}u^t k(u), \quad s_{st} = \left[S^{-1}\right]_{st}, \quad \textit{with} \quad [S]_{st} = \int u^{s+t-2}K(u)du,$$

and satisfy

$$\int k_0^*(\cdot)\,du = 1,\ p = \lambda = 0, \quad \int u^r k_0^*(\cdot)\,du = 0, \quad \textit{for} \quad 0 < r \le q-1.$$

We continue with the proof of Theorem 2. Define $\widehat{m}_1^*(x_1) = \widehat{a}_0(x_1)$ that minimizes

$$\min_{a_0,a_1,\ldots,a_{q-1}} \sum_{i=1}^{n} k\left(\frac{x_1 - X_{1i}}{h}\right)\left[Y_{1i}^{\text{oracle}} - \sum_{\ell=0}^{q-1} a_\ell \left(X_{1i} - x_{1j}\right)^\ell\right] 1(X_i \in \mathcal{S}^{on}).$$

Because the set $\mathcal{S}^{on}$ contains order $n(1-h_o)$ observations,

$$\widehat{m}_1^{\text{oracle}}(x_1) - \widehat{m}_1^*(x_1) = O_p\left(\frac{h_o}{1-h_o}(nh)^{-1}\right) = O_p\left(n^{-1}\right).$$

Whence it suffices to compare $\widehat{m}_1^{2-\text{step}}(x_1)$ to $\widehat{m}^*(x_1)$. We write

$$\widehat{m}_1^{2-\text{step}}(x_1) - \widehat{m}_1^*(x_1) = \sum_{i=1}^{n} W_{n,i}(h)\,(\mu - \overline{Y}) - \sum_{j=2}^{d}\sum_{i=1}^{n} W_{n,i}(h)\left[\widehat{\gamma}_j^{pi}(X_{ji}; h_o) - \gamma_j(X_{ji})\right],$$

where $W_{n,i}(h)$ are the weights of the local polynomial smoother. The first term $\sum_{i=1}^{n} W_{n,i}(h)\,(\mu - \overline{Y}) = O_p(n^{-1/2})$. From the result of Theorem 1 and Lemma 1, we get

$$\begin{aligned}\sum_{i=1}^{n} W_{n,i}(h)\left[\widehat{\gamma}_j^{pi}(X_{ji}; h_o) - \gamma_j(X_{ji})\right] &= \frac{1}{p_1(x_1)}\frac{1}{n}\sum_{i=1}^{n}\frac{1}{h}k_0^*\left(\frac{X_{1i}-x_1}{h}\right) \\ &\quad \times \left[B_{2n}(X_{ji}) + B_{3n}(X_{ji})\right] \\ &\quad + \frac{1}{p_1(x_1)}\frac{1}{n}\sum_{i=1}^{n}\frac{1}{h}k_0^*\left(\frac{X_{1i}-x_1}{h}\right) \\ &\quad \times \left[S_{1n}(X_{ji}) + S_{3n}''(X_{ji})\right] + o_p\left(\frac{1}{\sqrt{nh}}\right) \\ &\equiv B_{n,j}^*(x_1) + V_{n,j}^*(x_1) + o_p(\frac{1}{\sqrt{nh}}).\end{aligned}$$

Since $W_{n,i} = 0$ for $X_i \notin \mathcal{S}^{on}$, only points away from the boundary contribute to the bias, which, by Theorem 1, is

$$B_{n,j}^*(x_1) = h_o^q \frac{1}{p_1(x_1)}\frac{1}{n}\sum_{i=1}^{n}\frac{1}{h}k_0^*\left(\frac{X_{1i}-x_1}{h}\right)B_n(X_{ji}) + o_p(h_o^q) = O(h_o^q),$$

where

$$\begin{aligned}B_n(X_{ji}) &= \frac{\mu_q(k)}{q!}D^q m_j(X_{ji}) + \frac{\mu_q(K_2)}{q!}\int D^q p_{-j}(z_{-j}) m(X_{ji}, z_{-j})\,dz_{-j} \\ &\quad - \frac{\mu_q(K)}{q!}\int \frac{p_{-j}(z_{-j})}{p(x_1, z_{-j})} m(X_{ji}, z_{-j}) D^q p(X_{ji}, z_{-j})\,dz_{-j} = O(1).\end{aligned}$$

To analyze the variance term, recall that

$$S_{1n}(X_{ji}) + S''_{3n}(X_{ji}) = \frac{1}{nh_o}\sum_{k=1}^{n} k\left(\frac{X_{jk}-X_{ji}}{h_o}\right)\frac{\widehat{p}_{-j}(X_{-jk})}{\widehat{p}(X_k)}\varepsilon_k + \frac{1}{n}\sum_{k=1}^{n}\widetilde{\zeta}_{ik}^{j},$$

where $\widetilde{\zeta}_{ik}^{j} = \zeta_{ik}^{j} - \overline{\zeta}_{i}^{j}$ with $\zeta_{ik}^{j} \equiv \left[\frac{1}{h_0}k(\frac{X_{jk}-X_{ji}}{h_0})\frac{\overline{p}_{-j}(X_{-jk})}{\overline{p}(X_{jk},X_{-jk})} - 1\right] m(X_k)$ and $\overline{\zeta}_{i}^{j} \equiv E_i\left(\zeta_{ik}^{j}\right)$. That is,

$$V^*_{n,j}(x_1) = \frac{1}{p_1(x_1)}\frac{1}{n}\sum_{k=1}^{n}\widehat{p}^*_{1,,j}(x_1, X_{jk})\frac{\widehat{p}_{-j}(X_{-jk})}{\widehat{p}(X_k)}\varepsilon_k$$
$$+\frac{1}{p_1(x_1)}\frac{1}{n^2}\sum_{i=1}^{n}\sum_{k=1}^{n}\frac{1}{h}k_0^*\left(\frac{X_{1i}-x_1}{h}\right)\widetilde{\zeta}_{ik}^{j} \equiv V^*_{1n,j}(x_1) + V^*_{2n,j}(x_1),$$

where $\widehat{p}^*_{1,j}(x_1, X_{jk}) = \frac{1}{n}\sum_{i=1}^{n}\frac{1}{h}k_0^*\left(\frac{X_{1i}-x_1}{h}\right)\frac{1}{h_0}k(\frac{X_{jk}-X_{ji}}{h_0})$.

The same argument used to show $S_{1n} = O_p\left(\frac{1}{\sqrt{nh_0}}\right)$ in Theorem 1 gives

$$V^*_{1n,j}(x_1) = \frac{1}{p_1(x_1)}\frac{1}{n}\sum_{k=1}^{n}\widehat{p}^*_{1,,j}(x_1, X_{jk})\frac{\widehat{p}_{-j}(X_{-jk})}{\widehat{p}(X_k)}\varepsilon_k = O_p\left(\frac{1}{\sqrt{n}}\right).$$

It remains to show

$$V^*_{2n,j}(x_1) = \frac{1}{p_1(x_1)}\frac{1}{n^2}\sum_{i=1}^{n}\sum_{k=1}^{n}\frac{1}{h}k_0^*\left(\frac{X_{1i}-x_1}{h}\right)\widetilde{\zeta}_{ik}^{j} = O_p\left(\frac{1}{\sqrt{n}}\right).$$

Note that $\frac{1}{p_1(x_1)}\frac{1}{n^2}\sum_{k=1}^{n}\frac{1}{h}k_0^*\left(\frac{X_{1k}-x_1}{h}\right)\widetilde{\zeta}_{kk}^{j} = O_p\left(\frac{1}{n\sqrt{nh}}\right)$, for the same reason in the proof of $S''_{3n} = O_p\left(\frac{1}{\sqrt{nh_0}}\right)$ in Theorem 1. Thus, we only need to check that the variance of $\frac{1}{p_1(x_1)}\frac{1}{n^2}\sum\sum_{i\neq k}\widetilde{\tau}_{ik}$ is smaller order than $O\left(\frac{1}{nh}\right)$, where $\widetilde{\tau}_{ik} \equiv \frac{1}{h}k_0^*\left(\frac{X_{1i}-x_1}{h}\right)\widetilde{\zeta}_{ik}^{j}$, $\overline{\tau}_i \equiv E_i(\tau_{ik})$, $\tau_{ik} \equiv \widetilde{\tau}_{ik} + \overline{\tau}_i$, and $\overline{\tau} \equiv E(\overline{\tau}_i)$. The variance of this double sum consists of $\sum\sum_{i\neq k} E(\widetilde{\tau}_{ik}^2)$, $\sum\sum_{i\neq k} E(\widetilde{\tau}_{ik}\widetilde{\tau}_{ki})$, $\sum\sum\sum_{i\neq k\neq h} E(\widetilde{\tau}_{ih}\widetilde{\tau}_{kh})$, $\sum\sum\sum_{i\neq k\neq h} E(\widetilde{\tau}_{ik}\widetilde{\tau}_{ih})$, and $\sum\sum\sum\sum_{i\neq k\neq h\neq m} E(\widetilde{\tau}_{ik}\widetilde{\tau}_{hm})$. The following direct computations complete the proof.

$$\frac{1}{n^2}E(\widetilde{\tau}_{ik}^2) = \frac{1}{n^2}E(\tau_{ik}^2) + O\left(\frac{h_0^{2q}}{n^2}\right) = O\left(\frac{1}{n^2h_0h}\right)$$

$$\frac{1}{n^2}E(\widetilde{\tau}_{ik}\widetilde{\tau}_{ki}) = \frac{1}{n^2}E(\tau_{ik}\tau_{ki}) - \frac{2}{n^2}E(\overline{\tau}_i\tau_{ki}) + O\left(\frac{h_0^{2q}}{n^2}\right) = O\left(\frac{1}{n^2h_0}\right)$$

$$\frac{1}{n}E(\widetilde{\tau}_{im}\widetilde{\tau}_{km}) = \frac{1}{n^2}E(\tau_{im}\tau_{km}) + O\left(\frac{h_0^{2q}}{n}\right) = O\left(\frac{1}{n}\right).$$

Finally, $E(\widetilde{\tau}_{ik}\widetilde{\tau}_{im}) = E\left[E_i(\widetilde{\tau}_{ik}\widetilde{\tau}_{im})\right] = E\left[E_i(\widetilde{\tau}_{ik})E_i(\widetilde{\tau}_{im})\right] = 0$, and $E(\widetilde{\tau}_{ik}\widetilde{\tau}_{lm}) = E(\widetilde{\tau}_{ik})E(\widetilde{\tau}_{lm}) = 0$, which is immediate from the conditioning argument.

Proof of Theorem 3: First, note that by the triangle inequality,

$$\begin{aligned} p &\equiv \Pr\left[n^{q/(2q+1)}|\widehat{m}_1^{2-\text{step}}(x_1) - \widehat{m}_1^{\text{oracle}}(x_1)| > \lambda_n|\mathcal{X}^n\right] \\ &\leq \Pr\Bigg[n^{q/(2q+1)}|\widehat{m}_1^{2-\text{step}}(x_1) - \widehat{m}_1^{\text{oracle}}(x_1) \\ &\qquad - \sum_{j=2}^{d} B_{n,j}^*(x_1)| > \lambda_n - n^{q/(2q+1)}|\sum_{j=2}^{d} B_{n,j}^*(x_1)||\mathcal{X}^n\Bigg]. \end{aligned}$$

Applying inequalities of Markov and Esseen, we get, for $s > 2$,

$$\begin{aligned} p &\leq \frac{E\left[n^{sq/(2q+1)}|\widehat{m}_1^{2-\text{step}}(x_1) - \widehat{m}_1^{\text{oracle}}(x_1) - \sum_{j=2}^{d} B_{n,j}^*(x_1)|^s|\mathcal{X}^n\right]}{\left|\lambda_n - n^{q/(2q+1)}|\sum_{j=2}^{d} B_{n,j}^*(x_1)|\right|^s} \\ &= \frac{\lambda_n^{-s}E\left[n^{sq/(2q+1)}|\widehat{m}_1^{2-\text{step}}(x_1) - \widehat{m}_1^{\text{oracle}}(x_1) - \sum_{j=2}^{d} B_{n,j}^*(x_1)|^s|\mathcal{X}^n\right]}{\left|1 - \lambda_n^{-1}O\left(\left(\frac{h_o}{h}\right)^q\right)\right|^s} \\ &= a_1\lambda_n^{-s}n^{-s/2(2q+1)}, \end{aligned}$$

where we used $\sum_{j=2}^{d} B_{n,j}^*(x_1) = O\left(h_o^q\right)$ and $\widehat{m}_1^{2-\text{step}}(x_1) - \widehat{m}_1^{\text{oracle}}(x_1) - \sum_{j=2}^{d} B_{n,j}^*(x_1) = O_p\left(n^{-1/2}\right)$, from Theorem 2, and the condition, $\left(\frac{h_o}{h}\right)^q/\lambda_n \to 0$. □

ACKNOWLEDGMENTS

We would like to thank the National Science Foundation for financial support. We would also like to thank Jens Perch Nielsen and Arthur Lewbel for helpful comments. GAUSS programs are available at http:\\www.econ.yale.edu\~linton.

[Received February 1998. Revised August 1998.]

REFERENCES

Angrist, J. D., Imbens, G. W., and Rubin, D. B. (1996), "Identification of Causal Effects Using Instrumental Variables" (with discussion), *Journal of the American Statistical Association*, 91, 444–473.

Auestad, B., and Tjøstheim, D. (1991), "Functional Identification in Nonlinear Time Series," in *Nonparametric Functional Estimation and Related Topics*, ed. G. Roussas, Kluwer Academic: Amsterdam, pp. 493–507.

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., and Tibshirani, R. (1989), "Linear Smoothers and Additive Models" (with discussion), *The Annals of Statistics*, 17, 453–555.

Chen, R., Härdle, W., Linton, O. B., and Severance-Lossin, E. (1996), "Estimation in Additive Nonparametric Regression," in *Proceedings of the COMPSTAT Conference Semmering*, eds. W. Härdle and M. Schimek, Heidelberg: Physika Verlag, pp. 247–265.

Fan, J. (1992), "Design-Adaptive Nonparametric Regression," *Journal of the American Statistical Association*, 82, 998–1004.

——— (1993), "Local Linear Regression Smoothers and Their Minimax Efficiencies," *The Annals of Statistics*, 21, 196–216.

Fan, J., and Gijbels, I. (1992), "Variable Bandwidth and Local Linear Regression Smoothers," *The Annals of Statistics*, 20, 2008–2036.

Hall, P. (1992), *The Bootstrap and Edgeworth Expansion*, Berlin: Springer-Verlag.

Härdle, W. (1990), *Applied Nonparametric Regression. Econometric Monograph Series 19*, Cambridge: Cambridge University Press.

Härdle, W., and Marron, E. (1991), "Bootstrap Simultaneous Error Bars for Nonparametric Regression," *The Annals of Statistics*, 19, 778–796.

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

Hengartner, N. W. (1996), "Rate Optimal Estimation of Additive Regression via the Integration Method in the Presence of Many Covariates," Preprint, Department of Statistics, Yale University, http:\\www.stat.yale.edu.

Jones, M. C., Davies, S. J., and Park, B. U. (1994), "Versions of Kernel-Type Regression Estimators," *Journal of the American Statistical Association*, 89, 825–832.

Jones, M. C., Marron, J. S., and Sheather, S. J. (1996), "A Brief Survey of Bandwidth Selection for Density Estimation," *Journal of the American Statistical Association*, 91, 401–407.

Linton, O. B. (1996), "Efficient Estimation of Additive Nonparametric Regression Models," *Biometrika*, 84, 469–474.

Linton, O. B., and Härdle, W. (1996), "Estimating Additive Regression Models With Known Links," *Biometrika*, 83, 529–540.

Linton, O. B., and Nielsen, J. P. (1995), "A Kernel Method of Estimating Structured Nonparametric Regression Based on Marginal Integration," *Biometrika*, 82, 93–100.

Linton, O. B., Wang, N., Chen, R., and Härdle, W. (1997), "An Analysis of Transformation for Additive Nonparametric Regression," *Journal of the American Statistical Association*, 92, 1512–1521.

Masry, E. (1996), "Multivariate Local Polynomial Regression for Time Series: Uniform Strong Consistency and Rates," *Journal of Time Series Analysis*, 17, 571–599.

Newey, W. K. (1994), "Kernel Estimation of Partial Means," *Econometric Theory*, 10, 233–253.

Nielsen, J. P., and Linton, O. B. (1997), "Multiplicative Marker Dependent Hazard Estimation," unpublished manuscript.

Opsomer, J. D., and Ruppert, D. (1997), "On the Existence and Asymptotic Properties of Backfitting Estimators," *The Annals of Statistics*, 25, 186–211.

Stone, C. J. (1985), "Additive Regression and Other Nonparametric Models," *The Annals of Statistics*, 13, 685–705.

——— (1986), "The Dimensionality Reduction Principle for Generalized Additive Models," *The Annals of Statistics*, 14, 592–606.

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

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

Wu, C. J. F. (1986), "Jacknife, Bootstrap and Other Resampling Methods in Regression Analysis," *The Annals of Statistics*, 14, 1261–1343.