ORACLE-EFFICIENT NONPARAMETRIC ESTIMATION OF AN ADDITIVE MODEL WITH AN UNKNOWN LINK FUNCTION

by

Joel L. Horowitz
Department of Economics
Northwestern University
Evanston, IL 60208-2600
U.S.A.
joel-horowitz@northwestern.edu

and

Enno Mammen
Department of Economics
University of Mannheim
L 7, 3 - 5
68131 Mannheim
Germany
emammen@rumms.uni-mannheim.de

November 2007

ABSTRACT

This paper describes an estimator of the additive components of a nonparametric additive model with an unknown link function. When the additive components and link function are twice differentiable with sufficiently smooth second derivatives, the estimator is asymptotically normally distributed with a rate of convergence in probability of $n^{-2/5}$. This is true regardless of the (finite) dimension of the explanatory variable. Thus, the estimator has no curse of dimensionality. Moreover, the asymptotic distribution of the estimator of each additive component is the same as it would be if the link function and the other components were known with certainty. Thus, asymptotically there is no penalty for not knowing the link function or the other components.

Key words: Dimension reduction, kernel estimation, orthogonal series estimation

AMS 2000 subject classifications: Primary 62G08; secondary 62G20

The research of Joel L. Horowitz was supported in part by NSF Grant SES-0352675 and the Alexander von Humboldt Foundation.

ORACLE-EFFICIENT NONPARAMETRIC ESTIMATION OF AN ADDITIVE MODEL WITH AN UNKNOWN LINK FUNCTION

1. Introduction

This paper is concerned with nonparametric estimation of the functions F and $m_1,...,m_d$ in the model

(1.1) $$Y = F[m_1(X^1) + ... + m_d(X^d)] + U,$$

where X^j (j = 1, …, d) is the j'th component of the random vector $X \in \mathbb{R}^d$ for some finite $d \geq 2$, F and $m_1,...,m_d$ are unknown functions, and U is an unobserved random variable satisfying $\boldsymbol{E}(U \mid X = x) = 0$ for almost every x. Estimation is based on an *iid* random sample $\{Y_i, X_i : i = 1,...,n\}$ of (Y, X). We describe estimators of F and $m_1,...,m_d$ that converge in probability pointwise at the rate $n^{-2/5}$ when F and the m_j's are twice differentiable and the second derivatives are sufficiently smooth. Only two derivatives are needed regardless of the dimension of X, so asymptotically there is no curse of dimensionality. Moreover, the estimators of the additive components m_j have an oracle property. Specifically, the centered, scaled estimator of each additive component is asymptotically normally distributed with the same mean and variance that it would have if F and the other components were known.

Model (1.1) is attractive for applied research because it nests nonparametric additive models and semiparametric single-index models. In a nonparametric additive model

(1.2) $$\boldsymbol{E}(Y \mid X) = m_1(X^1) + ... + m_d(X^d),$$

where the m_j's are unknown functions. In a single-index model

(1.3) $$\boldsymbol{E}(Y \mid X) = H(\theta' X),$$

where θ is a $d \times 1$ dimensional vector and H is an unknown function. Models (1.2) and (1.3) are non-nested. Each contains conditional mean functions that are not contained in the other, so an applied researcher must choose between the two specifications. If an incorrect choice is made, the resulting model is misspecified, and inferences based on it may be misleading. Model (1.1) nests (1.2) and (1.3), thereby avoiding the need to choose between additive and single-index specifications. Model (1.1) also nests the multiplicative specification

$$\boldsymbol{E}(Y \mid X) = H[\tilde{m}_1(X^1)\tilde{m}_2(X^2)...\tilde{m}_d(X^d)],$$

where H and the $\tilde{m}_j$'s are unknown functions. This model can be put into the form (1.1) by setting $F(v) = H(e^v)$ and $m_j(X^j) = \log \tilde{m}_j(X^j)$.

A further attraction of (1.1) is that it provides an informal, graphical method for checking additive and single-index specifications (1.2) and (1.3). One can plot the estimates of F and the m_j's. Approximate linearity of the estimate of F favors the additive specification (1.2), whereas approximate linearity of the m_j's favors the single-index specification (1.3). Linearity of F and the m_j's favors the linear model $\boldsymbol{E}(Y \mid X) = \theta' X$.

There is a large literature on estimating the m_j's in (1.1) nonparametrically when F is known to be the identity function. As is discussed by Carrasco, Florens, and Renault (2005), the identifying relation of an additive model is a Fredholm equation of the second kind, and estimating the model presents an ill-posed inverse problem. Stone (1985, 1986) showed that $n^{-2/5}$ is the optimal L_2 rate of convergence of an estimator of the m_j's when they are twice continuously differentiable. Stone (1994) and Newey (1997) describe spline estimators whose L_2 rate of convergence is $n^{-2/5}$. Breiman and Friedman (1985); Buja, Hastie, and Tibshirani (1989); Hastie and Tibshirani (1990); Opsomer and Ruppert (1997); Mammen, Linton, and Nielsen (1999); and Opsomer (2000) investigate the properties of backfitting estimators. Newey (1994); Tjøstheim and Auestad (1994); Linton and Nielsen (1995); Chen, Härdle, Linton, and Severance-Lossin (1996); and Fan, Härdle, and Mammen (1998) investigate the properties of marginal integration estimators. Horowitz, Klemelä, and Mammen (2006), hereinafter HKM, discuss optimality properties of a variety of estimators for nonparametric additive models with identity link functions. Estimators for the case in which F is not necessarily the identity function but is known have been developed by Linton and Härdle (1996), Linton (2000), and Horowitz and Mammen (2004). Using arguments like those of Carrasco, Florens, and Renault (2005), it can be shown that the identifying relation for an additive model with a link function can be written as a nonlinear integral equation. The linearization of this equation is a Fredholm equation of the second kind, and estimation of the model presents an ill-posed inverse problem. This argument carries over to the case of an unknown link function. The statistical properties of the nonlinear model (e.g., rates of convergence and oracle properties) are similar to those of the linear model, but the technical details of the nonlinear problem are different from those of the linear case and, consequently, require a separate treatment.

Estimators for the case of an unknown F have been developed by Horowitz (2001) and Horowitz and Mammen (2007). Horowitz's (2001) estimator is asymptotically normal, but its rate of convergence in probability is slower than $n^{-2/5}$. Moreover, it requires F and the m_j's to have an increasing number of derivatives as the dimension of X increases. Thus, it suffers from the curse of dimensionality. Horowitz and Mammen (2007) developed penalized least squares estimators of F and the m_j's that have L_2 rates of convergence of $n^{-2/5}$ and do not suffer from the curse of dimensionality. However, the asymptotic distributions of these estimators are unknown, and carrying out inference with them appears to be very difficult. The estimators presented in this paper avoid the curse of dimensionality and are pointwise asymptotically normal at an $n^{-2/5}$ rate when F and the m_j's are twice continuously differentiable. Therefore, (asymptotic) inference based on these estimators is straightforward. Moreover, the estimators of the m_j's are asymptotically equivalent to those of Horowitz and Mammen (2004) for the case of a known F. Therefore, asymptotically there is no penalty for not knowing F.

The estimators described in this paper are developed through a two stage procedure. In the first stage, a modified version of Ichimura's (1993) estimator for a semiparametric single-index model is used to obtain a series approximation to each m_j and a kernel estimator of F. The first-stage procedure imposes the additive structure of (1.1), thereby avoiding the curse of dimensionality. The first-stage estimates are inputs to the second stage. The second-stage estimator of, say, m_1 is obtained by taking one Newton step from the first-stage estimate toward a local nonlinear least-squares estimate. In large samples, the second-stage estimator has a structure similar to that of a kernel nonparametric regression estimator, so deriving its pointwise rate of convergence and asymptotic distribution is relatively easy.

The theoretical concept underlying our procedure is the same as that used by HKM for estimating a nonparametric additive model with a known, identity link function. HKM showed that each additive component can be estimated with the same asymptotic performance as a kernel smoothing estimator in a model in which the link function and all but one additive component are known. Here, we show that the same approach works in the considerably more complex case of an additive model with an unknown, possibly nonlinear link function. The procedure in HKM, like the one in this paper, has two stages. The first consists of obtaining an undersmoothed, nonparametric pilot estimator of each additive component. This estimator has an asymptotically negligible bias but a variance that converges to zero relatively slowly. The second-stage estimator is obtained by using a single backfitting step to update each component while setting

the others equal to their first-stage (pilot) estimates. The oracle property of the second-stage estimator follows because the bias of the first stage estimator is negligible and the rate of convergence of the variance is accelerated by the smoothing that takes place in the second stage. We conjecture that this method can be used to obtain oracle efficient estimators for a large class of other smoothing approaches. We also conjecture that if the m_j's and F are $r > 2$ times differentiable, then estimators of the m_j's can be obtained that are oracle-efficient and asymptotically normal with $n^{-r/(2r+1)}$ rates of convergence. However, we do not attempt to prove these conjectures in this paper.

The remainder of this paper is organized as follows. Section 2 provides an informal description of the two-stage estimator. The main results are presented in Section 3. Section 4 discusses the selection of bandwidths. Section 5 presents the results of a small simulation study, and Section 6 presents concluding comments. The proofs of theorems are in the appendix. Throughout the paper, subscripts index observations and superscripts denote components of vectors. Thus, X_i is the i'th observation of X, X^j is the j'th component of X, and X_i^j is the i'th observation of the j'th component.

2. Informal Description of the Estimators

Assume that the support of X is $\mathcal{X} \equiv [0,1]^d$. This can always be achieved by, if necessary, carrying out monotone transformations of the components of X. For any $x \in \mathcal{X}$ define $m(x) = m_1(x^1) + ... + m_d(x^d)$, where x^j is the j'th component of x. Observe that (1.1) remains unchanged if each m_j is replaced by $m_j + a_j$ for any constants a_j and $F(\nu)$ is replaced by $F^*(\nu) = F(\nu - a_1 - ... - a_d)$. Similarly, (1.1) is unchanged if each m_j is replaced by cm_j for any $c \neq 0$ and $F(\nu)$ is replaced by $F^*(\nu) = F(\nu / c)$. Therefore, location, sign, and scale normalizations are needed to make identification possible. Under the additional assumption that F is monotone, model (1.1) is identified if at least two additive components are not constant (Horowitz and Mammen 2007, Proposition 3.1). This assumption is also necessary for identification. To see why, suppose that only m_1 is not constant. Then the regression function is of the form $F[m_1(x^1) + \text{constant}]$. It is clear that this function does not identify F and m_1. In this paper we use a slightly stronger assumption for identification. We assume that the derivatives of two additive components are bounded away from 0. The indices j and k of these components do not need to be known for the implementation of our estimator. They are needed

only for the statement the result of the first estimation step (Theorem 1). We suppose for this purpose that $j = d$ and $k = d - 1$.

We achieve location normalization by setting

(2.1) $\int_0^1 m_j(v)dv = 0; \;\; j = 1,...,d$.

To describe our sign and scale normalizations, let $\{p_k : k = 1,2,...\}$ denote a basis for smooth functions on $[0,1]$ satisfying (2.1). A precise definition of "smooth" and conditions that the basis functions must satisfy are given in Section 3. The conditions include:

(2.2) $\int_0^1 p_k(v)dv = 0$;

(2.3) $\int_0^1 p_j(v)\,p_k(v)dv = \begin{cases} 1 \text{ if } j = k \\ 0 \text{ otherwise;} \end{cases}$

$p_1(v) = \sqrt{12}(v - 1/2)$, and

(2.4) $m_j(x^j) = \sum_{k=1}^{\infty} \theta_{jk}\, p_k(x^j)$

for each $j = 1,...,d$, each $x^j \in [0,1]$, and suitable coefficients $\{\theta_{jk}\}$. We achieve sign and scale normalization by setting $\theta_{d1} = 1$. This works because, to ensure identification, we assume that there is a finite constant $C_M > 0$ such that $m_d'(v) \geq C_M$ for all $v \in [0,1]$. Under this assumption, θ_{d1} is bounded away from 0 and can be normalized to equal 1. To see why θ_{d1} is bounded away from 0, use integration by parts to obtain

$$\int_0^1 m_d(v)\,p_1(v)dv = \sqrt{12}\int_0^1 m_d(v)(v - 1/2)dv$$

$$= \left(\sqrt{12}/2\right)\left[\, m_d(v)v(v-1)\big|_0^1 - \int_0^1 m_d'(v)(v^2 - v)dv \,\right]$$

$$= -\left(\sqrt{12}/2\right)\int_0^1 m_d'(v)(v^2 - v)dv$$

$$\geq C_M / \sqrt{12}.$$

Now, for any positive integer κ, define

$$P_\kappa(x) = [\,p_1(x^1),..., p_\kappa(x^1), p_1(x^2),..., p_\kappa(x^2),..., p_1(x^d),... p_\kappa(x^d)]'\,.$$

Then for $\theta_\kappa \in \mathbb{R}^{\kappa d}$, $P_\kappa(x)'\theta_\kappa$ is a series approximation to $m(x)$. Section 3 gives conditions that κ must satisfy. These require that $\kappa \to \infty$ at an appropriate rate as $n \to \infty$.

We now describe the two-stage procedure for estimating (say) m_1. We begin with the first-stage estimator. Let K be a kernel function (in the sense of nonparametric regression) on $[-1,1]$, and define $K_h(v) = K(v/h)$ for any real, positive constant h. Conditions that K and h must satisfy are given in Section 3. Let $\hat{F}_i(\theta_\kappa)$ be the following estimator of $F[P_\kappa(X_i)'\theta_\kappa]$:

$$\hat{F}_i(\theta_\kappa) = \frac{1}{nh\hat{g}_i(\theta_\kappa)} \sum_{\substack{j=1 \\ j \neq i}}^{n} Y_j K_h \left\{ \left[P_\kappa(X_i)' - P_\kappa(X_j)' \right] \theta_\kappa \right\},$$

where

$$\hat{g}_i(\theta_\kappa) = \frac{1}{nh} \sum_{\substack{j=1 \\ j \neq i}}^{n} K_h \left\{ \left[P_\kappa(X_i)' - P_\kappa(X_j)' \right] \theta_\kappa \right\}.$$

To obtain the first-stage estimators of the m_j's, let $\{Y_i, X_i : i = 1,...,n\}$ be a random sample of (Y, X). Let $\hat{\theta}_{n\kappa}$ be a solution to

$$\text{(2.5)} \qquad \underset{\theta \in \Theta_\kappa}{\text{minimize}} \quad S_{n\kappa}(\theta) \equiv n^{-1} \sum_{i=1}^{n} I(X_i \in \mathcal{A}_h)[Y_i - \hat{F}_i(\theta)]^2,$$

where Θ_κ and $\mathcal{A}_h$ are sets that are defined in the next paragraph. The series estimator of $m(x)$ for any $x \in \mathcal{A}_h$ is

$$\tilde{m}(x) = P_\kappa(x)'\hat{\theta}_{n\kappa}.$$

The estimator of $m_j(x^j)$ is the product of $[p_1(x^j),..., p_\kappa(x^j)]$ with the appropriate components of $\hat{\theta}_{n\kappa}$.

We now define the sets Θ_κ and $\mathcal{A}_h$. To ensure that the estimator obtained from (2.5) converges to m sufficiently rapidly as $n \to \infty$, $P_\kappa(X)'\theta_\kappa$ must have a bounded probability density whose value is sufficiently far from 0 at each $\theta_\kappa \in \Theta_\kappa$ and each $X \in \mathcal{A}_h$. We choose Θ_κ and $\mathcal{A}_h$ so that this requirement is satisfied. To do this, define the vectors $v = (v^1,...,v^d)$ and $v^{-d} = (v^1,...,v^{d-1})$. Define $m_j^*(v^j) = \sum_{k=1}^{\kappa} \theta_{jk} p_k(v^j)$ $(j = 1,...,d)$ and $m_{-d}^*(v^{-d}) = \sum_{j=1}^{d-1} \sum_{k=1}^{\kappa} \theta_{jk} p_k(v^j)$. Let f_X denote the probability density function of X. Then the density of $m^* \equiv m_{-d}^* + m_d^*$ is

(2.6) $$f_{m^*}(z) = \int \frac{f_X\{m_d^{*-1}[z - m_{-d}^*(x^{-d})], x^{-d}\}}{m_d^{*\prime}\{m_d^{*-1}[z - m_{-d}^*(x^{-d})]\}} dx^{-d} \ .$$

The function f_m has the desired properties if $m_d^{*\prime}$ is bounded and bounded away from 0, $m_d^{*\prime}$ is Lipschitz continuous, and the Lebesgue measure of the region of integration in (2.6) is sufficiently far from 0. To ensure that these properties hold, we assume that the basis functions p_k are continuously differentiable. We set

$$\mathcal{A}_h = [h, 1-h]^{d-1} \times [-(\log h)^{-1}, 1 + (\log h)^{-1}] .$$

The Lebesgue measure of the region of integration in (2.6) is bounded from below by a quantity that is proportional to $-(\log h)^{-1}$ whenever $z = m_{-d}^*(x^{-d}) + m_d^*(x^d)$, $x \in \mathcal{A}_h$, and $m_d^{*\prime}(v) \geq c_2$ for some constant $c_2 > 0$ and all $v \in [0,1]$.

To specify Θ_κ, let C_θ be a constant that satisfies assumption A5(iv) in Section 3. Let C_1, C_2, and C_3 be the constants defined in assumptions A3(i), A3(iii), and A3(vi). Let c_1, c_2, and c_3 be any finite constants satisfying $0 < c_2 < C_2$, $c_1 \geq C_1$ and $c_3 \geq C_3$. Set

$$\Theta_\kappa = \{\theta \in [-C_\theta, C_\theta]^{\kappa d} : |m_j^*(x^j)| \leq c_1, |m_d^{*\prime}(x^d)| \geq c_2,$$

$$|m_d^{*\prime}(x_2^d) - m_d^{*\prime}(x_1^d)| \leq c_3 |x_2^d - x_1^d| \text{ for all } x, x_1, x_2 \in \mathcal{A}_h\}.$$

To obtain the second-stage estimator of $m_1(x^1)$ at a point $x^1 \in (h, 1-h)$, let $\tilde{X}_i$ denote the i'th observation of $\tilde{X} \equiv (X^2, ..., X^d)$. Define $\tilde{m}_{-1}(\tilde{X}_i) = \tilde{m}_2(X_i^2) + ... + \tilde{m}_d(X_i^d)$, where $\tilde{m}_j$ is the first-stage estimator of m_j. For a bandwidth s define

$$[\hat{b}_{i0}(v), \hat{b}_{i1}(v)] = \underset{b_0, b_1}{\arg\min} \sum_{\substack{j=1 \\ j \neq i}}^{n} I(X_j \in \mathcal{A}_h)\{Y_j - b_0 - b_1[\tilde{m}(X_j) - v]\}^2 K_s[\tilde{m}(X_j) - v].$$

Define the local-linear estimators $\tilde{\tilde{F}}_i(v) = \hat{b}_{i0}(v)$ and $\tilde{\tilde{F}}_i'(v) = \hat{b}_{i1}(v)$. Let H be a symmetrical (about 0) probability density function on $[-1,1]$. For a bandwidth t, define $H_t(v) = H(v/t)$,

$$S_{n1}(x^1) = -2\sum_{i=1}^{n} I(X_i \in \mathcal{A}_h)\{Y_i - \tilde{\tilde{F}}_i[\tilde{m}_1(x^1) + \tilde{m}_{-1}(\tilde{X}_i)]\}\tilde{\tilde{F}}_i'[\tilde{m}_1(x^1) + \tilde{m}_{-1}(\tilde{X}_i)]H_t(x^1 - X_i^1),$$

and

$$S_{n2}(x^1) = 2\sum_{i=1}^{n} I(X_i \in \mathcal{A}_h)\tilde{\tilde{F}}_i'[\tilde{m}_1(x^1) + \tilde{m}_{-1}(\tilde{X}_i)]^2 H_t(x^1 - X_i^1) .$$

The second-stage estimator of $m_1(x^1)$ is

$$\text{(2.7)} \qquad \hat{m}_1(x^1) = \tilde{m}_1(x^1) - \frac{S_{n1}(x^1)}{S_{n2}(x^1)}.$$

The second stage estimators of $m_2(x^2),...,m_d(x^d)$ are obtained similarly.

The estimator (2.7) can be understood intuitively as follows. If $\tilde{\tilde{F}}_i$ and $\tilde{m}_{-1}$ were the true functions F and m_{-1}, then a consistent estimator of $m_1(x^1)$ could be obtained by minimizing

$$\text{(2.8)} \qquad S_{n1}(x^1, b) = \sum_{i=1}^{n} I(X_i \in \mathcal{A}_h)\{Y_i - \tilde{\tilde{F}}_i[b + \tilde{m}_{-1}(\tilde{X}_i)]\}^2 H_t(x^1 - X_i^1).$$

The estimator (2.7) is the result of taking one Newton step from the starting value $b = \tilde{m}_1(x^1)$ toward the minimum of the right-hand side of (2.8).

Section 3 gives conditions under which $\hat{m}_1(x^1) - m_1(x^1) = O_p(n^{-2/5})$ and $n^{2/5}[\hat{m}_1(x^1) - m_1(x^1)]$ is asymptotically normally distributed for any finite d when F and the m_j's are twice continuously differentiable. The second-stage estimator of F is the kernel nonparametric mean-regression of Y on $\hat{m} = \hat{m}_1 + ... + \hat{m}_d$. It is clear, though we do not prove here, that this estimator is $n^{-2/5}$-consistent and asymptotically normal.

3. Main Results

This section has two parts. Section 3.1 states the assumptions that are used to prove the main results. Section 3.2 states the results. The main results are the $n^{-2/5}$-consistency and asymptotic normality of the $\hat{m}_j$'s.

The following additional notation is used. For any matrix A, define the norm $\|A\| = [trace(A'A)]^{1/2}$. Also define $U = Y - F[m(X)]$, $V(x) = Var(U \mid X = x)$,

$$Q_\kappa = \boldsymbol{E}\{I(X \in \mathcal{A}_h)F'[m(X)]^2 P_\kappa(X) P_\kappa(X)'\},$$

and

$$\Psi_\kappa = Q_\kappa^{-1} \boldsymbol{E}\{I(X \in \mathcal{A}_h)F'[m(X)]^2 V(X) P_\kappa(X) P_\kappa(X)'\} Q_\kappa^{-1}$$

whenever the latter quantity exists. Q_κ and Ψ_κ are $d(\kappa) \times d(\kappa)$ positive semidefinite matrices, where $d(\kappa) = \kappa d$. Let $\lambda_{\kappa,\min}$ denote the smallest eigenvalue of Q_κ. Let $Q_{\kappa,ij}$ denote the (i, j)

element of Q_κ. Define $\zeta_\kappa = \sup_{x\in\mathcal{X}} \|P_\kappa(x)\|$. Let $\{\theta_{jk}\}$ be the coefficients of the series expansion (2.4). For each κ define $\theta_\kappa = (\theta_{11},...,\theta_{1\kappa},\theta_{21},...,\theta_{2\kappa},...,\theta_{d1},...,\theta_{d\kappa})'$.

3.1 Assumptions

The results are obtained under the following assumptions. None of the constants appearing in these assumptions needs to be known for implementation of our estimator.

A1: The data, $\{(Y_i, X_i): i = 1,...,n\}$, are an *iid* random sample from the distribution of (Y, X), and $\boldsymbol{E}(Y \mid X = x) = F[m(x)]$ for almost every $x \in \mathcal{X}$.

A2: (i) The support of X is $\mathcal{X}$. (ii) The distribution of X is absolutely continuous with respect to Lebesgue measure. (iii) The probability density function of X is bounded away from zero and twice differentiable on $\mathcal{X}$ with a Lipschitz-continuous second derivative. (iv) There are constants $c_V > 0$ and $C_V < \infty$ such that $c_V \le Var(U \mid X = x) \le C_V$ for all $x \in \mathcal{X}$. (v) There is a constant $C_U < \infty$ such that $\boldsymbol{E}\,|U|^j \le C_U^{j-2}\, j!\, \boldsymbol{E}(U^2) < \infty$ for all $j \ge 2$.

A3: (i) Each function m_j is defined on $[0,1]$, and $|m_j(v)| \le C_1$ for each $j = 1,...,d$, all $v \in [0,1]$, and some constant $C_1 < \infty$. (ii) Each function m_j is twice continuously differentiable on $[0,1]$ with derivatives at 0 and 1 interpreted as one-sided. (iii) There is a finite constant C_2 such that $m_d'(v) \ge C_2$ and $|m_{d-1}'(v)| \ge C_2$ for all $v \in [0,1]$. (iv) There are constants $C_{F1} < \infty$, $c_{F2} > 0$, and $C_{F2} < \infty$ such that $F(v) \le C_{F1}$ and $c_{F2} \le F'(v) \le C_{F2}$ for all $v \in \{m(x): x \in [0,1]^d\}$. (v) F is twice continuously differentiable on $\{m(x): x \in [0,1]^d\}$. (vi) There is a constant $C_3 < \infty$ such that $|m_d''(v_2) - m_d''(v_1)| \le C_3 |v_2 - v_1|$ for all $v_1, v_2 \in [0,1]$ and $|F''(v_2) - F''(v_1)| \le C_3 |v_2 - v_1|$ for all $v_2, v_1 \in \{m(x): x \in [0,1]^d\}$.

A4: (i) There are constants $C_Q < \infty$ and $c_\lambda > 0$ such that $|Q_{\kappa,ij}| \le C_Q$ and $\lambda_{\kappa,\min} > c_\lambda$ for all κ and all $i, j = 1,...,d(\kappa)$. (ii) The largest eigenvalue of Ψ_κ is bounded for all κ.

A5: (i) The functions $\{p_k\}$ satisfy (2.2), (2.3), and $p_1(v) = \sqrt{12}(v - 1/2)$. (ii) There is a constant $c_\kappa > 0$ such that $\zeta_\kappa \ge c_\kappa$ for all sufficiently large κ. (iii) $\zeta_\kappa = O(\kappa^{1/2})$ as $\kappa \to \infty$. (iv) There are a constant $C_\theta < \infty$ and vectors $\theta_{\kappa 0} \in \Theta_\kappa$ such that $\sup_{x\in\mathcal{A}_h} |m(x) - P_\kappa(x)'\theta_{\kappa 0}| = O(\kappa^{-2})$ as $\kappa \to \infty$. (v) For each κ, θ_κ is an interior point of Θ_κ.

A6: (i) $\kappa = C_\kappa n^{4/15+\nu}$ for some constant C_κ satisfying $0 < C_\kappa < \infty$ and some ν satisfying $0 < \nu < 1/30$. (ii) $h = C_h n^{-1/5}$ and $t = C_t n^{-1/5}$ for some constants C_h and C_t satisfying $0 < C_h < C_t < \infty$. (iii) $s = C_s n^{-1/7}$ for some constant C_s satisfying $0 < C_h < \infty$.

A7: H and K are supported on $[-1,1]$, symmetrical about 0, and twice differentiable everywhere. H'' and K'' are Lipschitz continuous. In addition,

$$\int_{-1}^{1} v^j K(v)dv = \begin{cases} 1 \text{ if } j = 0 \\ 0 \text{ if } 1 \le j \le 3 \end{cases}$$

Differentiability of the density of X (Assumption A2(iii)) is used to ensure that the bias of our estimator converges to zero sufficiently rapidly. Assumption A2(v) restricts the thickness of the tails of the distribution of U and is used to prove consistency of the first-stage estimator. Assumption A3 defines the sense in which F and the m_j's must be smooth. A3(iii) and A3(iv) are used for identification. The need for higher-order smoothness of F depends on the metric one uses to measure the distance between the estimated and true additive components. Horowitz and Mammen (2007) show that only two derivatives of F are needed to enable the additive components to achieve an $n^{-2/5}$ rate of convergence in the L_2 metric. However, Juditsky, Lepski, and Tsybakov (2007) show that when $d = 2$, three derivatives are needed to achieve this rate with the uniform metric. This follows from the lower bound in the first part of the proof of their Theorem 1. This lower bound is stated for another model but it also applies to an additive model with unknown link and two unknown additive components. A4 ensures the existence and non-singularity of the covariance matrix of the asymptotic form of the first-stage estimator. This is analogous to assuming that the information matrix is positive definite in parametric maximum likelihood estimation. Assumption A4(i) implies A4(ii) if U is homoskedastic. Assumption A4(vi) requires higher-order smoothness of only one additive component. We conjecture that this condition can be weakened. Assumptions A5(iii) and A5(iv) bound the magnitudes of the basis functions and ensure that the errors in the series approximations to the m_j's converge to zero sufficiently rapidly as $\kappa \to \infty$. These assumptions are satisfied by spline and (for periodic functions) Fourier bases. We use B-splines in the Monte Carlo experiment reported in Section 5. Assumption A6 states the rates at which $\kappa \to \infty$ and the bandwidths converge to 0 as $n \to \infty$. The assumed rates of convergence of h and t are well known to be asymptotically optimal for one-dimensional kernel mean-regression when the conditional mean function is twice continuously differentiable. The required rate for κ ensures that the asymptotic bias and

variance of the first-stage estimator are sufficiently small to achieve an $n^{-2/5}$ rate of convergence in the second stage. The L_2 rate of convergence of a series estimator of m_j is maximized by setting $\kappa \propto n^{1/5}$, which is slower than the rates permitted by A6(i) (Newey (1997)). Thus, A6(i) requires the first-stage estimator to be undersmoothed. Undersmoothing is needed to ensure sufficiently rapid convergence of the bias of the first-stage estimator. We show that the first-order performance of our second-stage estimator does not depend on the choice of κ if A6(i) is satisfied. See Theorem 2. Optimizing the choice of κ would require a rather complicated higher-order theory and is beyond the scope of this paper, which is restricted to first-order asymptotics.

3.2 Theorems

This section states two theorems that give the main results of the paper. Theorem 1 gives the asymptotic behavior of the first-stage estimator. Theorem 2 gives the properties of the second-stage estimator. Define $U_i = Y_i - F[m(X_i)]$ $(i = 1,...,n)$ and $b_{\kappa 0}(x) = m(x) - P_\kappa(x)'\theta_{\kappa 0}$. Let $\|v\|$ denote the Euclidean norm of any finite-dimensional vector v.

Theorem 1: Let assumptions A1-A7 hold. Then

(a) $$\lim_{n\to\infty} \left\| \hat{\theta}_{n\kappa} - \theta_{\kappa 0} \right\| = 0$$

almost surely,

(b) $$\hat{\theta}_{n\kappa} - \theta_{\kappa 0} = O_p(\kappa^{1/2}/n^{1/2} + \kappa^{-2}),$$

and

(c) $$\sup_{x \in \mathcal{A}_h} |\tilde{m}(x) - m(x)| = O_p(\kappa/n^{1/2} + \kappa^{-3/2}).$$

In addition,

$$\hat{\theta}_{n\kappa} - \theta_{\kappa 0} = n^{-1} Q_\kappa^{-1} \sum_{i=1}^{n} I(X_i \in \mathcal{A}_h) F'[m(X_i)] P_\kappa(X_i) U_i$$

(d)

$$+ n^{-1} Q_\kappa^{-1} \sum_{i=1}^{n} I(X_i \in \mathcal{A}_h) F'[m(X_i)]^2 P_\kappa(X_i) b_{\kappa 0}(X_i) + R_n,$$

where $\|R_n\| = O_p(\kappa^{3/2}/n + n^{-1/2})$. ■

Now define

$$S_{n1}(x^1, m) =$$

$$-2\sum_{i=1}^{n} I(X_i \in \mathcal{X}_h)\{Y_i - F[m_1(x^1) + m_{-1}(\tilde{X}_i)]\}F'[m_1(x^1) + m_{-1}(\tilde{X}_i)]H_t(x^1 - X_i^1),$$

$$D(x^1) = 2\int F'[m_1(x^1) + m_{-1}(\tilde{x})]^2 f_X(x^1, \tilde{x})d\tilde{x},$$

$$A_K = \int_{-1}^{1} v^2 H(v)dv,$$

$$B_K = \int_{-1}^{1} H(v)^2 dv,$$

$$q(x^1, \tilde{x}) = (\partial^2/\partial\zeta^2)\{F[m_1(\zeta + x^1) + m_{-1}(\tilde{x})] - F[m_1(x^1) + m_{-1}(\tilde{x})]\} f_X(\zeta + x^1, \tilde{x})\Big|_{\zeta=0},$$

$$\beta_1(x^1) = 2C_t^2 A_K D(x^1)^{-1}\int q(x^1, \tilde{x})F'[m_1(x^1) + m_{-1}(\tilde{x})] f_X(x^1, \tilde{x})d\tilde{x},$$

and

$$V_1(x^1) = 4B_K C_t^{-1} D(x^1)^{-2}\int Var(U \mid x^1, \tilde{x})F'[m_1(x^1) + m_{-1}(\tilde{x})]^2 f_X(x^1, \tilde{x})d\tilde{x}.$$

The next theorem gives the asymptotic properties of the second-stage estimator. We state the result only for the estimator of m_1. Analogous results hold for the estimators of the other components.

Theorem 2: Let assumptions A1-A7 hold. Then

(a) $\hat{m}_1(x^1) - m_1(x^1) = -[ntD(x^1)]^{-1} S_{n1}(x^1, m) + o_p(n^{-2/5})$

for each $x^1 \in (0,1)$, and $\hat{m}_1(x^1) - m_1(x^1) = o_p(1)$ uniformly over $x^1 \in \mathcal{A}_h$.

(b) If $x^1 \in (0,1)$, then $n^{2/5}[\hat{m}_1(x^1) - m_1(x^1)] \to^d N[\beta_1(x^1), V_1(x^1)]$.

(c) If $j \neq 1$ and $x^1, x^j \in (0,1)$, then $n^{2/5}[\hat{m}_1(x^1) - m_1(x^1)]$ and $n^{2/5}[\hat{m}_j(x^j) - m_j(x^j)]$ are asymptotically independently normally distributed. ■

Theorem 2(b) implies that asymptotically, $n^{2/5}[\hat{m}_1(x^1) - m_1(x^1)]$ is not affected by random sampling errors in the first stage estimator or lack of knowledge of F. In fact, the second-stage estimator of $m_1(x^1)$ has the same asymptotic distribution that it would have if F and $m_2, ..., m_d$ were known and (2.8) were used to estimate $m_1(x^1)$ directly. In this sense, our estimator has an oracle property. Part (c) of Theorem 2 implies that the estimators of $m_1(x^1), ..., m_d(x^d)$ are asymptotically independently distributed. These results hold for any m_j's and F that satisfy assumptions A1-A7. No other information about these functions is needed.

$V_1(x^1)$ and $\beta_1(x^1)$ can be estimated consistently by replacing unknown population parameters with consistent estimators. Alternatively, one can eliminate the asymptotic bias, $\beta_1(x^1)$, by setting $t = C_t n^{-\gamma}$ for $1/5 < \gamma < 1$. Then $n^{(1-\gamma)/2}[\hat{m}_1(x^1) - m_1(x^1)] \to^d N[0, V_1(x^1)]$.

Theorems 1 and 2 are proved by showing that the estimators presented here are asymptotically equivalent to the estimators of Horowitz and Mammen (2004), which assume that F is known. Therefore, the estimators presented here have the same asymptotic properties as the estimators of Horowitz and Mammen (2004).

4. Bandwidth Selection

This section presents an informal method for choosing the values of κ and the bandwidth parameters h, s, and t in applications. The asymptotic distributions of the second-stage estimators of the m_j's do not depend on the first-stage tuning parameters κ, h, and s. As a practical consequence, the second-stage estimates are not highly sensitive to these parameters and it is not necessarily to choose them precisely. Our method for choosing them consists of fitting an approximate parametric model to the data and choosing bandwidths that would be optimal if the approximate model were, in fact, correct. The approximate parametric model is obtained by choosing θ_κ to minimize

$$\sum_{i=1}^{n} \{Y_i - \psi[P_\kappa(X_i)'\theta_\kappa]\}^2$$

for some fixed value of κ, where ψ is a function that is chosen by the analyst. The specific form of the function ψ and the value of κ are not crucial. They should be chosen to achieve a reasonable fit to the data as indicated by, say, a residuals plot, but it is neither necessary nor desirable to carry out elaborate specification searching or testing in this step. Let $\bar{\theta}_\kappa$ denote the resulting value of θ_κ. Next, consider the kernel nonparametric regression of Y_i on $P_\kappa(X_i)\bar{\theta}_\kappa$. Use a standard method for bandwidth choice, possibly cross-validation, to estimate the optimal bandwidth for this regression. This yields our proposed choice of h. Our choice of s is obtained by using the plug-in method to estimate the optimal bandwidth for local linear estimation of the derivative of the $\boldsymbol{E}[Y \mid P_\kappa(X)'\bar{\theta}_\kappa]$. Finally, the bandwidth t can be chosen by using the plug-in method of Horowitz and Mammen (2004, Sec. 4). Equation (4.1) of Horowitz and Mammen (2004) gives the formula for the optimal t. Theorem 5 of Horowitz and Mammen

(2004) shows how to estimate the required derivatives of the m_j's. Theorem 5 can also be used to estimate the required derivatives of F.

5. Monte Carlo Experiments

This section reports the results of a Monte Carlo experiment that illustrates the finite-sample performance of $\hat{m}_1$. The experiment consists of repeatedly estimating m_1 in the binary logit model

$$P(Y = 1 | X = x) = F[m_1(x^1) + m_2(x^2)],$$

where

$$F(v) = 1/[1 + \exp(-v)]$$

for any $v \in (-\infty, \infty)$ and

$$m_1(z) = m_2(z) = 0.25(z+1)^4$$

for any $z \in [-1,1]$. X^1 and X^2 are independent with the uniform distribution on $[-1,1]$. This model is not a single-index or additive model. The curvature of m_1 provides a test of the performance of the estimator.

We also carried out the experiment using the infeasible oracle estimator that assumes knowledge of F and m_2. This estimator cannot be used in applications, but it provides a benchmark against which the performance of the feasible estimator can be compared. The infeasible oracle estimator of $m_1(x^1)$, which we denote by $\hat{m}_{1,OR}(x^1)$, is

$$\hat{m}_{1,OR}(x^1) = \arg\min_{\mu} \sum_{i=1}^{n} \{Y_i - F[\mu + m_2(X_i^2)]\}^2 K_h(x^1 - X_i^1).$$

All local-linear or kernel estimation steps in obtaining $\hat{m}_1$ and $\hat{m}_{1,OR}$ used the kernel function

$$K(v) = \frac{15}{16}(1 - v^2)^2 I(|v| \leq 1).$$

The basis functions in the first estimation stage were B-splines. The procedures described in Section 4 for selecting κ and the bandwidths cannot be used in the Monte Carlo experiment because of the long computing times that are required. Therefore, κ and the bandwidths were chosen through Monte Carlo experimentation to approximately minimize the integrated mean-square error (IMSE's) of $\hat{m}_1$. This yielded $\kappa = 3$, $h = 0.25$, $s = 1$, and $t = 0.6$. The sample size used in the experiment was n = 500. There were 100 replications. The experiment was carried

out in GAUSS using GAUSS pseudo-random number generators. The function m_1 was estimated over the interval $[-.5,.5]$, rather than $[-1,1]$, to reduce edge effects.

The results of the experiments can be summarized by the IMSEs of $\hat{m}_1$, $\hat{m}_{1,OR}$, and the first-stage spline estimator, $\tilde{m}_1$. These are 0.137, 0.158, and 0.232, respectively. As predicted by the theory, the infeasible oracle and second-stage estimators have approximately the same IMSEs, and the IMSEs of both are smaller than the IMSE of the first-stage spline estimator.

Figure 1 illustrates the estimates of m_1. The solid line in the figure shows the true function. The dashed line shows the average of the 100 Monte Carlo estimates. The circles, squares, and triangles show individual estimates of m_1. The circles show the estimate whose IMSE is at the 25th percentile of the IMSEs of the 100 Monte Carlo replications. The squares and triangles show the estimates corresponding to the 50th and 75th percentiles of the IMSEs. The shape of the average of the Monte Carlo estimates is similar to the shape of the true m_1, though some bias is evident. The bias can be reduced at the expense of increased variance and IMSE by reducing t. As is to be expected, the shape of the individual estimate at the 25th percentile of the IMSE is close to the shape of the true m_1, whereas the shape of the estimate at the 75th percentile is further from the truth.

Figure 2 illustrates the performance of the infeasible oracle estimator. As in Figure 1, the dashed line shows the average of the 100 Monte Carlo estimates, the circles, squares, and triangles, respectively, show estimates of m_1 at the 25th, 50th and 75th percentiles of the IMSE; and the solid line shows the true m_1. Figures 1 and 2 have a similar appearance, which illustrates the ability of the two-stage estimator to perform essentially as well as the infeasible oracle estimator.

We have also done limited experimentation with a model in which there are five covariates. In this model, F, m_1, and m_2 are the same as in the two-covariate experiment, and $m_j(x) = x$ for $j = 3$, 4, and 5. Because of the long computing times required to obtain the first-stage spline estimates when there are 5 covariates, we have carried out only a few Monte Carlo replications. These suggest that as in the two-covariate case, the second-stage estimate has a shape that is similar to that of the true m_j and an IMSE that is slightly less than the IMSE of the infeasible oracle estimator that assumes knowledge of F and all but one of the m_j's.

6. Conclusions

This paper has described an estimator of the additive components of a nonparametric additive model with an unknown link function. When the additive components and link function are twice differentiable with sufficiently smooth second derivatives, the estimator is asymptotically normally distributed with a rate of convergence in probability of $n^{-2/5}$. This is true regardless of the (finite) dimension of the explanatory variable. Thus, the estimator has no curse of dimensionality. Moreover, the estimator has an oracle property. The asymptotic distribution of the estimator of each additive component is the same as it would be if the link function and the other components were known with certainty. Thus, asymptotically there is no penalty for not knowing the link function or the other components.

Computation of the first-stage estimator remains an important topic for further research. The optimization problem (2.5) is hard to solve, especially if θ is high-dimensional, because the objective function is not globally convex. Although the theory presented in this paper requires solving (2.5), in applications it may be possible to obtain good numerical results by using other methods. The numerical performance of the second-stage estimator tends to be satisfactory whenever the first-stage estimates are good approximations to the true additive components. Thus, in applications it may suffice to obtain the first-stage estimates by using methods that are relatively easy to compute and perform satisfactorily in numerical practice, even though their theoretical properties in our setting are not understood. The average derivative estimator of Hristache, Juditsky, and Spokoiny (2001) is an example of such a method. The penalized least squares method of Horowitz and Mammen (2007) is another. A further possibility is to use such a method to obtain an initial estimate of θ and then take several Newton or other steps toward the optimum of (2.5). Any first-stage estimator, including ours, must undersmooth the series estimator. Otherwise, the bias of the first-stage estimator will not be asymptotically negligible, and the second-stage estimator will not have the oracle property.

Appendix: Proofs of Theorems

Assumptions A1-A7 hold throughout this section.

a. Theorem 1

This section begins with lemmas that are used to prove Theorem 1. For any fixed κ, $\theta \in \Theta_\kappa$, and $X_i \in \mathcal{A}_h$, define $Z_{\kappa j} = P_\kappa(X_j)$,

$$\hat{G}_i(\theta) = \frac{1}{nh}\sum_{\substack{j=1 \\ j\neq i}}^{n} Y_j K_h[(Z_{\kappa i} - Z_{\kappa j})'\theta],$$

$$\bar{F}(z,\theta) = \boldsymbol{E}(Y_1 \mid Z_{\kappa 1}'\theta = z)$$

and

$$\bar{F}_i(\theta) = \bar{F}(Z_{\kappa i}'\theta, \theta).$$

Let $\bar{g}(z,\theta)$ denote the density of $Z_{\kappa 1}'\theta$ evaluated at $Z_{\kappa 1}'\theta = z$. Define $g_i(\theta) = \bar{g}(Z_{\kappa i}'\theta, \theta)$, $G_i(\theta) = \bar{F}_i(\theta) g_i(\theta)$, and $\mathcal{I} = \{i : X_i \in \mathcal{A}_h\}$.

Lemma 1: There is a function n_1 and constants $a > 0$, $C > 0$, and $\varepsilon_1 > 0$ such that

$$\boldsymbol{P}\left[\sup_{i\in\mathcal{I},\, \theta\in\Theta_\kappa} |\hat{F}_i(\theta) - \bar{F}_i(\theta)| > \varepsilon\right] \leq C\exp(-nha\varepsilon^2)$$

for $0 < \varepsilon < \varepsilon_1$ and $n > n_1(\varepsilon)$.

Proof: If suffices to show that there are positive constants a_1, a_2, C_1, and C_2 such that

(A1) $$\boldsymbol{P}\left[\sup_{\theta\in\Theta_\kappa} |\hat{G}_1(\theta) - G_1(\theta)| > \varepsilon\right] \leq C_1\exp(-nha_1\varepsilon^2)$$

and

(A2) $$\boldsymbol{P}\left[\sup_{\theta\in\Theta_\kappa} |\hat{g}_1(\theta) - g_1(\theta)| > \varepsilon\right] \leq C_2\exp(-nha_2\varepsilon^2).$$

Only (A1) is proved. The proof of (A2) is similar.

Divide Θ_κ into hypercubes (or fragments of hypercubes) of edge-length ℓ_θ. Let $\Theta_\kappa^{(1)},...,\Theta_\kappa^{(M)}$ denote the $L_\theta = (2C_\theta/\ell_\theta)^{d(\kappa)}$ cubes thus created. Let $\theta_{\kappa j}$ be the point the center of $\Theta_\kappa^{(j)}$. The maximum distance between $\theta_{\kappa j}$ and any other point in $\Theta_\kappa^{(j)}$ is $r = d(\kappa)^{1/2}\ell_\theta/2$, and $L_\theta = \exp\{d(\kappa)[\log(C_\theta/r) + (1/2)\log d(\kappa)]\}$. Define

$$\boldsymbol{E}_1\hat{G}_1(\theta) = \frac{1}{nh}\sum_{\substack{j=1 \\ j\neq 1}}^{n} \boldsymbol{E}\{Y_j K_h[(Z_{\kappa 1} - Z_{\kappa j})'\theta] \mid Z_{\kappa j}'\theta = Z_{\kappa 1}'\theta\}.$$

Define $\Delta\hat{G}_1(\theta) = \hat{G}_1(\theta) - \boldsymbol{E}_1\hat{G}_1(\theta)$ and

$$P_n = \boldsymbol{P}\left[\sup_{\theta\in\Theta_\kappa} |\Delta\hat{G}_1(\theta)| > 2\varepsilon/3\right].$$

Then

$$P_n \le \sum_{j=1}^{L_\theta} \boldsymbol{P}\left[\sup_{\theta \in \Theta_\kappa^{(j)}} |\Delta\hat{G}_1(\theta)| > 2\varepsilon/3 \right].$$

Now for $\theta \in \Theta_\kappa^{(j)}$

$$|\Delta\hat{G}_1(\theta)| \le |\Delta\hat{G}_1(\theta_j)| + |\Delta\hat{G}_1(\theta) - \Delta\hat{G}_1(\theta_j)|.$$

A Taylor series approximation gives

$$\hat{G}_1(\theta) - \hat{G}_1(\theta_j) = \frac{1}{nh}\sum_{i=2}^{n} Y_i K_h'(\xi_i) \frac{(Z_{\kappa 1} - Z_{\kappa i})'(\theta - \theta_j)}{h},$$

where ξ_i is between $(Z_{\kappa 1} - Z_{\kappa i})'\theta$ and $(Z_{\kappa 1} - Z_{\kappa i})'\theta_j$. But

$$[(Z_{\kappa 1} - Z_{\kappa i})'(\theta - \theta_j)]^2 \le \left\|Z_{\kappa 1} - Z_{\kappa i}\right\|^2 \left\|\theta - \theta_j\right\|^2$$

$$\le 4\zeta_\kappa^2 r^2,$$

and $|K'(\xi)| \le C_K$ for some constant $C_K < \infty$. Therefore,

$$|\hat{G}_1(\theta) - \hat{G}_1(\theta_j)| \le \frac{2C_K r\zeta_\kappa}{nh^2}\sum_{i=2}^{n} |Y_i|.$$

Moreover, $|\boldsymbol{E}_1\hat{G}_1(\theta) - \boldsymbol{E}_1\hat{G}_1(\theta_j)| \le C_F r\zeta_\kappa$. Therefore,

$$|\Delta\hat{G}_1(\theta) - \Delta\hat{G}_1(\theta_j)| \le \frac{2C_K r\zeta_\kappa}{nh^2}\sum_{i=2}^{n}(|Y_i| - \boldsymbol{E}|Y_i|)$$

$$+ \frac{C_K r\zeta_\kappa}{nh^2}\boldsymbol{E}|Y_i| + C_F r\zeta_\kappa.$$

Choose $r = h^2/\zeta_\kappa^2$. Then $\varepsilon/3 - [(C_K r\zeta_\kappa)/(nh^2)\boldsymbol{E}|Y_i| + C_F r\zeta_\kappa] > \varepsilon/6$ for all sufficiently large κ. Moreover,

(A3) $$\boldsymbol{P}\left[\frac{2C_K r\zeta_\kappa}{nh^2}\sum_{i=2}^{n}(|Y_i| - \boldsymbol{E}|Y_i|) > \varepsilon/6\right] \le 2\exp(-a_3 n\zeta_\kappa^2\varepsilon)$$

for some constant $a_3 > 0$ by Bernstein's inequality. Also by Bernstein's inequality, there is a constant $a_4 > 0$ such that

(A4) $$\boldsymbol{P}(|\Delta\hat{G}_1(\theta_j)| > \varepsilon/3) \le 2\exp(-a_4 nh^2\varepsilon^2).$$

Therefore,

$$P_n \le 2[L_\theta \exp(-a_4 n h^2 \varepsilon^2) + \exp(-a_3 n \zeta_\kappa^2 \varepsilon)]$$

$$\le C_1 \exp(-a_1 n h^2 \varepsilon^2)$$

for suitable constants $C_1 > 0$ and $a_1 > 0$. The proof is completed by observing that under assumption A3, $P_\kappa(x)'\theta$ has a continuous density. Therefore, by A3 and the definition of Θ_κ, $|\boldsymbol{E}_1\hat{G}_1(\theta) - G_1(\theta)| = o(1)$ uniformly over $\theta \in \Theta_\kappa$, so $|\boldsymbol{E}_1\hat{G}_1(\theta) - G_1(\theta)| < \varepsilon/3$ uniformly over $\theta \in \Theta_\kappa$ for all sufficiently large n. Q.E.D.

For $\theta \in \Theta_\kappa$, define

$$S_h(\theta) = \boldsymbol{E}(U^2) + \boldsymbol{E}n^{-1}\sum_{i=1}^{n} I(X_i \in \mathcal{A}_h)\{F[m(X_i)] - \bar{F}_i(\theta)\}^2$$

and $\tilde{\theta}_\kappa = \arg\min_{\theta \in \Theta_\kappa} S_h(\theta)$.

Lemma 2:

$$\lim_{n\to\infty} \sup_{\theta \in \Theta_\kappa} |S_{n\kappa}(\theta) - S_h(\theta)| = 0$$

almost surely.

Proof: It follows from Lemma 1 and Theorem 1.3.4 of Serfling (1980, p. 10) that as $n \to \infty$. $\hat{F}_i(\theta) \to \bar{F}_i(\theta)$ almost surely uniformly over $i \in \mathcal{I}$ and $\theta \in \Theta_\kappa$. The conclusion of the lemma follows from this result and Jennrich's (1969) uniform strong law of large numbers (Jennrich 1969). Q.E.D.

Define

$$S_{\kappa 0}(\theta) = \boldsymbol{E}I(X \in \mathcal{A}_h)\{Y - F[P_\kappa(X)'\theta + b_{\kappa 0}(X)]\}^2$$

and $b_k(x) = \mu + m(x) - P_\kappa(x)'\theta_\kappa$. Then

$$\theta_{\kappa 0} = \arg\min_{\theta \in \Theta_\kappa} S_{\kappa 0}(\theta).$$

Lemma 3: As $n \to \infty$, $|\tilde{\theta}_\kappa - \theta_{\kappa 0}| \to 0$.

Proof: Standard arguments for kernel estimators show that as $n \to \infty$, $\bar{F}_i(\tilde{\theta}_\kappa) \to G[\ell_1(X_i^1) + ... + \ell_d(X_i^d)]$ for almost any X_i and some functions G and $\ell_1, ..., \ell_d$ that satisfy

(A5) $\quad G[\ell_1(x^1) + ... + \ell_d(x^d)] = F[m_1(x^1) + ... + m_d(x^d)]$

for almost every $x \in \mathcal{X}$. Taking the ratio of the derivatives of each side of (A5) with respect to x^d and x^j for some $j > 1$ gives

$$\frac{\ell'_j(x^j)}{\ell_d(x^d)} = \frac{m'_j(x^j)}{m'_d(x^d)}$$

and

$$\ell'_j(x^j)\int_0^1 \frac{1}{\ell'_d(v)}dv = m'_j(x^j)\int_0^1 \frac{1}{m'_d(v)}dv .$$

Therefore,

(A6) $\quad \ell'_j(x^j) = \gamma m'_j(x^j)$

for some constant $\gamma > 0$ and almost every x^j. Integrating (A6) yields

$$\ell_j(x^j) = \gamma m_j(x^j) + \eta ,$$

where η is a constant. Imposing the location and scale normalizations of Section 2 gives $\gamma = 1$ and $\eta = 0$, so $\ell_j(x^j) = m_j(x^j)$ for almost every x^j. A similar argument with m_2 in place of m_d and m_d in place of m_j shows that $\ell_d(x^d) = m_d(x^d)$ for almost every x^d. The conclusion of the lemma follows from uniqueness of the Fourier representations of the m_j's. Q.E.D.

Lemma 4: $\left\|\hat{\theta}_{n\kappa} - \tilde{\theta}_\kappa\right\| \to 0$ almost surely.

Proof: For each κ, let $\mathcal{N}_\kappa \subset \mathbb{R}^{d(\kappa)}$ be an open set containing $\tilde{\theta}_\kappa$. Let $\bar{\mathcal{N}}_\kappa$ denote the complement of $\mathcal{N}_\kappa$ in Θ_κ. Define $T_\kappa = \bar{\mathcal{N}}_\kappa \bigcap \Theta_\kappa$. Then $T_\kappa \subset \mathbb{R}^{d(\kappa)}$ is compact. Define

$$\eta = \min_{\theta \in T_\kappa} S_h(\theta) - S_h(\tilde{\theta}_\kappa) .$$

Let A_n be the event $|S_{n\kappa}(\theta) - S_h(\theta)| < \eta/2$ for all $\theta \in \Theta_\kappa$. Then

$$A_n \Rightarrow S_h(\hat{\theta}_{n\kappa}) < S_{n\kappa}(\hat{\theta}_{n\kappa}) + \eta/2$$

and

$$A_n \Rightarrow S_{n\kappa}(\tilde{\theta}_\kappa) < S_h(\tilde{\theta}_\kappa) + \eta/2 .$$

But $S_{n\kappa}(\hat{\theta}_{n\kappa}) \le S_{n\kappa}(\tilde{\theta}_\kappa)$ by definition, so

$$A_n \Rightarrow S_h(\hat{\theta}_{n\kappa}) < S_{n\kappa}(\tilde{\theta}_\kappa) + \eta/2 .$$

Therefore,

$$A_n \Rightarrow S_h(\hat{\theta}_{n\kappa}) < S_h(\tilde{\theta}_\kappa) + \eta \Rightarrow S_h(\hat{\theta}_{n\kappa}) - S_h(\tilde{\theta}_\kappa) < \eta .$$

So $A_n \Rightarrow \hat{\theta}_{n\kappa} \in \mathcal{N}_\kappa$. Since $\mathcal{N}_\kappa$ is arbitrary, the result follows from Lemma 2 and Theorem 1.3.4 of Serfling (1980, p. 10). Q.E.D.

Now define $Z_{\kappa i} = I(X_i \in \mathcal{A}_h)F'[m(X_i)]P_\kappa(X_i)$. Define $\hat{Q}_\kappa = n^{-1}\sum_{i=1}^n Z_{\kappa i}Z_{\kappa i}'$.

Lemma 5: $\left\|\hat{Q}_\kappa - Q_\kappa\right\|^2 = O_p(\kappa^2/n)$.

Proof: See Horowitz and Mammen (2004, Lemma 4). Q.E.D.

Define $\gamma_n = I(\lambda_{\kappa,\min} \geq c_\lambda/2)$, where I is the indicator function. Let $\bar{U} = (U_1,...,U_n)'$.

Lemma 6: $\gamma_n\left\|\hat{Q}_\kappa^{-1}Z_\kappa'\bar{U}/n\right\| = O_p(\kappa^{1/2}/n^{1/2})$ as $n \to \infty$.

Proof: See Horowitz and Mammen (2004, Lemma 5). Q.E.D.

Define $B_n = \hat{Q}_\kappa^{-1}n^{-1}\sum_{i=1}^n F'[m(X_i)]Z_{\kappa i}b_{\kappa 0}(X_i)$.

Lemma 7: $\|B_n\| = O(\kappa^{-2})$ with probability approaching 1 as $n \to \infty$.

Proof: See Horowitz and Mammen (2004, Lemma 6). Q.E.D.

Proof of Theorem 1: Part (a) follows by combining Lemmas 3 and 4. To prove the remaining parts, observe that $\hat{\theta}_{n\kappa}$ satisfies the first-order condition $\partial S_{n\kappa}(\hat{\theta}_{n\kappa})/\partial\theta = 0$ almost surely for all sufficiently large n. Define $M_i = m(X_i)$ and $\Delta M_i = P_\kappa(X_i)'\hat{\theta}_{n\kappa} - M_i$ $= P_\kappa(X_i)'(\hat{\theta}_{n\kappa} - \theta_{\kappa 0}) - b_{\kappa 0}(X_i)$. For $\theta \in \Theta_\kappa$ define $\Delta\hat{F}_i(\theta) = \hat{F}_i(\theta) - F[P_\kappa(X_i)'\theta]$. Then a Taylor series expansion yields

$$n^{-1}\sum_{i=1}^n Z_{\kappa i}U_i - (\hat{Q}_\kappa + R_{n1})(\hat{\theta}_{n\kappa} - \theta_{\kappa 0}) + n^{-1}\sum_{i=1}^n F'(M_i)Z_{\kappa i}b_{\kappa 0}(X_i) + R_{n2} = 0,$$

almost surely for all sufficiently large n. R_{n1} is defined by

$$R_{n1} = n^{-1}\sum_{i=1}^n I(X_i \in \mathcal{A}_h)\{-U_iF''(M_i) - U_i[F''(\tilde{\tilde{M}}_i) - F''(M_i)] + [(3/2)F''(\tilde{\tilde{M}}_i)F'(M_i)$$

$$+(1/2)F''(\tilde{M}_i)F''(\tilde{\tilde{M}}_i)\Delta M_i + (1/2)F''(\tilde{M}_i)F''(\tilde{\tilde{M}}_i)(\Delta M_i)^2]\Delta M_i -$$

$$[F''(\tilde{\tilde{M}}_i)F'(M_i) - (1/2)F''(\tilde{M}_i)F'(\tilde{M}_i) + F''(\tilde{\tilde{M}}_i)F''(\tilde{M}_i)b_{\kappa 0}(X_i)]b_{\kappa 0}(X_i)\}P_\kappa(X_i)P_\kappa(X_i)'$$

$$+R_{n3},$$

where

$$R_{n3} = \frac{2}{n}\sum_{i=1}^{n} I(X_i \in \mathcal{A}_h)\left\{2[Y_i - F_i(\tilde{\theta})]\frac{\partial^2 \Delta\hat{F}_i(\tilde{\theta})}{\partial\theta\partial\theta'} - 2\frac{\partial\Delta\hat{F}_i(\tilde{\theta})}{\partial\theta}\frac{\partial F_i(\tilde{\theta})}{\partial\theta'} + \Delta\hat{F}_i(\tilde{\theta})\frac{\partial^2 F_i(\tilde{\theta})}{\partial\theta\partial\theta'}\right.$$

$$\left. + \frac{\partial F_i(\tilde{\theta})}{\partial\theta}\frac{\partial\Delta\hat{F}_i(\tilde{\theta})}{\partial\theta'}\right\},$$

$\tilde{M}_i$ and $\tilde{\tilde{M}}_i$ are points between $P_\kappa(X_i)'\hat{\theta}_{n\kappa}$ and M_i, and $\tilde{\theta}$ is between $\hat{\theta}_{n\kappa}$ and $\theta_{\kappa 0}$. R_{n2} is defined by

$$R_{n2} = -n^{-1}\sum_{i=1}^{n} I(X_i \in \mathcal{A}_h)\{U_i F''(\tilde{\tilde{M}}_i) + U_i[F''(\tilde{\tilde{M}}_i) - F''(M_i)]$$

$$+ [F''(\tilde{\tilde{M}}_i)F'(M_i) - (1/2)F''(\tilde{M}_i)F'(M_i)]b_{\kappa 0}(X_i) - (1/2)F''(\tilde{\tilde{M}}_i)F''(\tilde{M}_i)b_{\kappa 0}(X_i)^2\}b_{\kappa 0}(X_i)$$

$$+ R_{n4},$$

where

$$R_{n4} = \frac{2}{n}\sum_{i=1}^{n} I(X_i \in \mathcal{A}_h)\left[U_i\frac{\partial\Delta\hat{F}_i(\theta_{\kappa 0})}{\partial\theta} + \Delta\hat{F}_i(\theta_{\kappa 0})\frac{\partial F_i(\theta_{\kappa 0})}{\partial\theta} + \Delta\hat{F}_i(\hat{\theta}_{n\kappa})\frac{\partial\Delta\hat{F}_i(\hat{\theta}_{n\kappa})}{\partial\theta}\right].$$

Lengthy arguments similar to those used to prove Lemma 1 show that

$$\|R_{n3}\|^2 = O\left[\kappa^2(\log n)^2/(nh^3) + \int[P_\kappa(x)'(\hat{\theta}_{n\kappa} - \theta_{\kappa 0})]^2 dx + \kappa^{-3}\right]$$

almost surely and

$$\|R_{n4}\|^2 = O_p[\kappa^2(\log n)^4/(n^2h^3) + \kappa^{-3}].$$

Now let either

$$\xi = I(X \in \mathcal{A}_h)\hat{Q}_\kappa^{-1}Z_\kappa'\bar{U}/n$$

or

$$\xi = \hat{Q}_\kappa^{-1}\left[n^{-1}\sum\nolimits_{i=1}^{n} I(X_i \in \mathcal{A}_h)F'(M_i)^2 P_\kappa(X_i)b_{\kappa 0}(X_i) + R_{n2}\right].$$

Note that

$$\left\|n^{-1}\sum_{i=1}^{n} I(X_i \in \mathcal{A}_h)U_i F''(M_i)P_\kappa(X_i)P_\kappa(X_i)'\right\|^2 = O_p(\kappa^2/n).$$

Then

$$\gamma_n\left\|[(\hat{Q}_\kappa + R_{n1})^{-1} - \hat{Q}_\kappa^{-1}]\hat{Q}_\kappa\xi\right\|^2 = \gamma_n\left\|(\hat{Q}_\kappa + R_{n1})^{-1}R_{n1}\xi\right\|^2$$

$$= Trace\{\gamma_n[\xi' R_{n1}(\hat{Q}_\kappa + R_{n1})^{-2} R_{n1}\xi]\}$$

$$= O_p\left(\left\|\xi' R_{n1}\right\|^2\right)$$

$$= O_p(\xi'\xi) O_p\left\{\kappa^2/n + \frac{\kappa^2(\log n)^2}{nh^3}\left\|\hat{\theta}_{n\kappa} - \theta_{\kappa 0}\right\|^2 + \int [P_\kappa(x)'(\hat{\theta}_{n\kappa} - \theta_{\kappa 0})]^2 dx + \sup_{x\in\mathcal{X}} |b_{\kappa 0}(x)|^2\right\}$$

$$= O_p(\xi'\xi) O_p\left\{\kappa^2/n + \left[\kappa + \frac{\kappa^2(\log n)^2}{nh^3}\right]\left\|\hat{\theta}_\kappa - \theta_{\kappa 0}\right\|^2 + \kappa^{-3}\right\}.$$

Setting $\xi = I(X \in \mathcal{A}_h)\hat{Q}_\kappa^{-1} Z_\kappa' \bar{U}/n$ and applying Lemma 6 yields

$$\left\|[(\hat{Q}_\kappa + R_{n1})^{-1} - \hat{Q}_\kappa^{-1}] Z_\kappa' \bar{U}/n\right\|^2 = O_p\left\{\kappa^3/n^2 + \left[\frac{\kappa^2}{n} + \frac{\kappa^3(\log n)^2}{n^2 h^3}\right]\left\|\hat{\theta}_{n\kappa} - \theta_{\kappa 0}\right\|^2 + 1/(n\kappa^2)\right\}.$$

If $\xi = \hat{Q}_\kappa^{-1}\left[n^{-1}\sum_{i=1}^n I(X_i \in \mathcal{A}_h) F'(M_i)^2 P_\kappa(X_i) b_{\kappa 0}(X_i) + R_{n2}\right]$, then applying Lemma 7, and using the result $\left\|\hat{Q}_\kappa^{-1} R_{n2}\right\|^2 = O_p[\kappa^2(\log n)^4/(n^2 h^3)]$ yields

$$\left\|[(\hat{Q}_\kappa + R_{n1})^{-1} - \hat{Q}_\kappa^{-1}]\left[n^{-1}\sum_{i=1}^n I(X_i \in \mathcal{A}_h) F'(M_i) Z_{\kappa i} b_{\kappa 0}(X_i) + R_{n2}\right]\right\|^2 =$$

$$O_p\left[\frac{\kappa^5(\log n)^6}{n^4 h^6}\left\|\hat{\theta}_{n\kappa} - \theta_{\kappa 0}\right\|^2 + \frac{(\kappa \log n)^4}{n^3 h^3}\right].$$

It follows from these results that

$$\hat{\theta}_{n\kappa} - \theta_{\kappa 0} = n^{-1}\hat{Q}_\kappa^{-1}\sum_{i=1}^n I(X_i \in \mathcal{A}_h) F'[m(X_i)] P_\kappa(X_i) U_i$$

$$+ n^{-1}\hat{Q}_\kappa^{-1}\sum_{i=1}^n I(X_i \in \mathcal{A}_h) F'[m(X_i)]^2 P_\kappa(X_i) b_\kappa(X_i) + R_n,$$

where $\|R_n\| = O_p(\kappa^{3/2}/n + n^{-1/2})$. Part (d) of the theorem now follows from Lemma 5. Part (b) follows by applying Lemmas 6 and 7 to Part (d). Part (c) follows from Part (b) and Assumption A5(iii). Q.E.D.

b. Theorem 2

This section begins with a lemma that is used to prove Theorem 2. For any $\tilde{x} \equiv (x^2,...,x^d) \in [-1,1]^{d-1}$, set $m_{-1}(\tilde{x}) = m_2(x^2) + ... + m_d(x^d)$. Define

$$S_{n10}(x^1) =$$

$$-2\sum_{i=1}^{n} I(X_i \in \mathcal{A}_h)\{Y_i - F[\tilde{m}_1(x^1) + \tilde{m}_{-1}(\tilde{X}_i)]\}F'[\tilde{m}_1(x^1) + \tilde{m}_{-1}(\tilde{X}_i)]H_t(x^1 - X_i^1)$$

and

$$S_{n20}(x^1) = 2\sum_{i=1}^{n} I(X_i \in \mathcal{A}_h)F'[\tilde{m}_1(x^1) + \tilde{m}_{-1}(\tilde{X}_i)]^2 H_t(x^1 - X_i^1).$$

Let $[\bar{F}(v), \bar{F}'(v)] = \text{plim}_{n\to\infty}[\tilde{\bar{F}}_i(v), \tilde{\bar{F}}_i'(v)]$, $\tilde{\Delta}F_i = \tilde{\bar{F}}_i - \bar{F}$, $\bar{\Delta}F = \bar{F} - F$, $\tilde{\Delta}F_i' = \tilde{\bar{F}}_i' - \bar{F}'$, $\bar{\Delta}F' = \bar{F}' - F'$, $\tilde{m}(x^1, \tilde{X}_i) = \tilde{m}_1(x^1) + \tilde{m}_{-1}(\tilde{X}_i)$, $m(x^1, \tilde{X}_i) = m_1(x^1) + m_{-1}(\tilde{X}_i)$ and $\Delta m_i = \tilde{m}(x^1, \tilde{X}_i) - m(x^1, \tilde{X}_i)$. By arguments like those used to prove Lemma 1, $\tilde{\Delta}F_i[\tilde{m}(x)], \tilde{\Delta}F_i[m(x)] = o[(ns)^{-1/2}\log n]$ and $\tilde{\Delta}F_i'[\tilde{m}(x)], \tilde{\Delta}F_i'[m(x)] = o[(ns^3)^{-1/2}\log n]$ almost surely uniformly over $x \in \mathcal{A}_h$.

Lemma 8: (a) $(nt)^{-1/2}S_{n1}(x^1) = (nt)^{-1/2}S_{n10}(x^1) + o_p(1)$ for each $x^1 \in [t, 1-t]$.

In addition, the following hold uniformly over $x^1 \in [t, 1-t]$:

(b) $(nt)^{-1}S_{n2}(x^1) = (nt)^{-1}S_{n20}(x^1) + o_p(1)$

Proof: Only part (a) is proved. The proof of part (b) is similar. Write $(nt)^{-1/2}S_{n1}(x^1) = (nt)^{-1/2}S_{n10}(x^1) + \sum_{j=1}^{6} L_j(x^1)$, where

$$L_1(x^1) =$$

$$2(nt)^{-1/2}\sum_{i=1}^{n} I(X_i \in \mathcal{A}_h)\tilde{\Delta}F_i[\tilde{m}(x^1, \tilde{X}_i)]\bar{F}'[\tilde{m}(x^1, \tilde{X}_i)]H_t(x^1 - X_i^1),$$

$$L_2(x^1) =$$

$$-2(nt)^{-1/2}\sum_{i=1}^{n} I(X_i \in \mathcal{A}_h)\{Y_i - F[\tilde{m}(x^1, \tilde{X}_i)]\}\tilde{\Delta}F_i'[\tilde{m}(x^1, \tilde{X}_i)]H_t(x^1 - X_i^1),$$

$$L_3(x^1) =$$

$$2(nt)^{-1/2}\sum_{i=1}^{n} I(X_i \in \mathcal{A}_h)\bar{\Delta}F[\tilde{m}(x^1, \tilde{X}_i)]\tilde{\Delta}F_i'[\tilde{m}(x^1, \tilde{X}_i)]H_t(x^1 - X_i^1),$$

$$L_4(x^1) =$$

$$2(nt)^{-1/2}\sum_{i=1}^{n} I(X_i \in \mathcal{A}_h)\tilde{\Delta}F_i[\tilde{m}(x^1, \tilde{X}_i)]\tilde{\Delta}F_i'[\tilde{m}(x^1, \tilde{X}_i)]H_t(x^1 - X_i^1),$$

$$L_5(x^1) =$$

$$-2(nt)^{-1/2}\sum_{i=1}^{n} I(X_i \in \mathcal{A}_h)\{Y_i - F[\tilde{m}(x^1, \tilde{X}_i)]\}\bar{\Delta}F'[\tilde{m}_1(x^1, \tilde{X}_i)]H_t(x^1 - X_i^1),$$

and

$$L_6(x^1) =$$

$$2(nt)^{-1/2}\sum_{i=1}^{n} I(X_i \in \mathcal{A}_h)\bar{\Delta}F[\tilde{m}(x^1, \tilde{X}_i)]\bar{F}'[\tilde{m}(x^1, \tilde{X}_i)]H_t(x^1 - X_i^1).$$

Standard properties of kernel estimators yield the results that

$$L_1(x^1) = O[(nt)^{1/2}(ns)^{-1/2}\log n] = o(1)$$

and

$$L_4(x^1) = O[(nt)^{1/2}(ns)^{-1/2}(ns^3)^{-1/2}(\log n)^2] = o(1)$$

almost surely uniformly over $x^1 \in [t, 1-t]$. In addition, it follows from Theorem 1(c) and the properties of kernel estimators that

$$L_3(x^1) = O_p[(nt)^{1/2}(ns^3)^{-1/2}]\sup_{x\in\mathcal{X}} |\tilde{m}(x) - m(x)|$$

$$= o_p(1).$$

$L_5(x^1)$ and $L_6(x^1)$ can be written

$$L_5(x^1) =$$

$$-2(nt)^{-1/2}\sum_{i=1}^{n} I(X_i \in \mathcal{A}_h)\{Y_i - F[\tilde{m}(x^1, \tilde{X}_i)]\}\bar{F}''[m(x^1, \tilde{X}_i)]\Delta m_i H_t(x^1 - X_i^1) + o_p(1)$$

and

$$L_6(x^1) =$$

$$2(nt)^{-1/2}\sum_{i=1}^{n} I(X_i \in \mathcal{A}_h)\bar{F}'[m(x^1,\tilde{X}_i)]^2[\tilde{m}(x^1,\tilde{X}_i) - m(x^1,\tilde{X}_i)]H_t(x^1 - X_i^1) + o_p(1).$$

$L_5(x^1) = o_p(1)$ and $L_6(x^1) = o_p(1)$ uniformly over $x^1 \in [t,1-t]$ now follow by the arguments used to prove Lemma 10 of Horowitz and Mammen (2004).

Now consider $L_2(x^1)$. For $x^1 \in [t,1-t]$, a Taylor series expansion gives

$$L_2(x^1) = L_{2a}(x^1) + L_{2b}(x^1) + L_{2c}(x^1),$$

where

$$L_{2a}(x^1) = -2(nt)^{-1/2}\sum_{i=1}^{n} I(X_i \in \mathcal{A}_h)\{Y_i - F[m(x^1,\tilde{X}_i)]\}\tilde{\Delta}F_i'[m(x^1,\tilde{X}_i)]H_t(x^1 - X_i^1),$$

$$L_{2b}(x^1) = -2(nt)^{-1/2}\sum_{i=1}^{n} I(X_i \in \mathcal{A}_h)\{Y_i - F[m(x^1,\tilde{X}_i)]\}\tilde{\Delta}F_i''(\tilde{\tilde{m}}_i)\Delta m_i H_t(x^1 - X_i^1),$$

$$L_{2c}(x^1) = 2(nt)^{-1/2}\sum_{i=1}^{n} I(X_i \in \mathcal{A}_h)F'(\tilde{\tilde{m}}_i)\tilde{\Delta}F_i'[\tilde{m}(x^1,\tilde{X}_i)]\Delta m_i H_t(x^1 - X_i^1)$$

and $\tilde{\tilde{m}}_i$ is between $\tilde{m}(x^1,\tilde{X}_i)$ and $m(x^1,\tilde{X}_i)$. Since $\Delta m_i = O_p(\kappa/n^{1/2})$ uniformly over $x^1 \in [t,1-t]$, we have $L_{2c}(x^1) = O_p[(nt)^{1/2}(ns^3)^{-1/2}\kappa n^{-1/2}\log n] = o_p(1)$ uniformly over $x^1 \in [t,1-t]$. Now consider L_{2a}. Divide [0,1] into $J_n = O(n^{1/2})$ subintervals of length $1/J_n$. Denote the j'th subinterval by I_j ($j = 1,...,J_n$), and let x_j^1 denote the midpoint of I_j. Then for any $\varepsilon > 0$,

$$\boldsymbol{P}\left[\sup_{x^1\in[t,1-t]} |L_{2a}(x^1)| > \varepsilon\right] = \boldsymbol{P}\left\{\bigcup_j \left[\sup_{x^1\in I_j} |L_{2a}(x^1)| > \varepsilon\right]\right\}$$

$$\leq \boldsymbol{P}\left\{\bigcup_j \left[|L_{2a}(x_j^1)| > \varepsilon/2\right]\right\} + \boldsymbol{P}\left\{\bigcup_j \left[\sup_{x^1\in I_j} |L_{2a}(x^1) - L_{2a}(x_j^1)| > \varepsilon/2\right]\right\}$$

$$\equiv P_{n1} + P_{n2}.$$

We have $\boldsymbol{E}L_{2a}(x_j^1) = O[(nt)^{-1/2}nt^3s^3] = O(n^{-3/7})$ for each $j = 1,...,J_n$, and a straightforward though lengthy calculation shows that $Var[L_{2a}(x_j^1)] = O[t/(ns^4)] = O(n^{-22/35})$. Therefore, it follows from Markov's inequality that $P_{n1} = O(J_n n^{-22/35}) = o(1)$ as $n \to \infty$. Now if $x^1 \in I_j$,

$$L_{2a}(x^1) - L_{2a}(x_j^1) = 2(nt)^{-1/2}\sum_{i=1}^{n}\Big\{U_{i1}\tilde{\Delta}F_i''[m(\tilde{x}^1, \tilde{X}_i)]H_t(\tilde{x}^1 - \tilde{X}_i)$$

$$+ t^{-1}U_{i1}\tilde{\Delta}F_i'[m(\tilde{x}^1, \tilde{X}_i)]H_t'(\tilde{x}^1 - \tilde{X}_i) - F'[m(\tilde{x}^1, \tilde{X}_i)]\tilde{\Delta}F_i'[m(\tilde{x}^1, \tilde{X}_i)]H_t(\tilde{x}^1 - \tilde{X}_i)\Big\}(x^1 - x_j^1),$$

where $U_{i1} = Y_i - F[m(x^1, \tilde{X}_i)]$ and $\tilde{x}^1$ is between x_j and x^1. But $\tilde{\Delta}F_i''[m(x^1, \tilde{X}_i)] = O[(\log n)/(ns^5)^{1/2}] = O(n^{-1/7}\log n)$ and $\tilde{\Delta}F_i'[m(x^1, \tilde{X}_i)] = O[(\log n)/(ns^3)^{1/2}] = O(n^{-2/7}\log n)$ uniformly over $x^1 \in [t, 1-t]$. Therefore,

$$\sup_{j,\, x \in I_j} |L_{2a}(x) - L_{2a}(x_j^1)| = O(J_n^{-1}n^{11/35}\log n)$$

almost surely. It follows that $P_{n2} = o(1)$.

Now write $L_{2b}(x^1)$ in the form $L_{2b}(x^1) = L_{2b1}(x^1) + L_{2b2}(x^1)$, where

$L_{2b1}(x^1) =$

$$-2(nt)^{-1/2}\sum_{i=1}^{n} I(X_i \in \mathcal{A}_h)\{F[m(x^1, \tilde{X}_i)] - F[m(X_i^1, \tilde{X}_i)]\}\tilde{\Delta}F_i'[\tilde{m}(x^1, \tilde{X}_i)]H_t(x^1 - X_i^1).$$

and

$$L_{2b2}(x^1) = -2(nt)^{-1/2}\sum_{i=1}^{n} I(X_i \in \mathcal{A}_h)U_i\tilde{\Delta}F_i'[\tilde{m}(x^1, \tilde{X}_i)]H_t(x^1 - X_i^1).$$

$L_{2b1} = O[(nt)^{1/2}\kappa n^{-1/2}(ns^3)^{-1/2}] = o(1)$ almost surely uniformly over $x^1 \in [t, 1-t]$. Now let $\tilde{m}^i(x^1, X^i)$ be the version of $\tilde{m}(x^1, X^i)$ that is obtained by leaving observation i out of the estimation of θ in the first stage. Then

$$\text{(A7)} \qquad L_{2b2}(x^1) = -2(nt)^{-1/2}\sum_{i=1}^{n} I(X_i \in \mathcal{A}_h)U_i\tilde{\Delta}F_i'[\tilde{m}^{-i}(x^1, \tilde{X}_i)]H_t(x^1 - X_i^1) + o_p(1)$$

uniformly over $x \in \mathcal{X}$. The first term on the right-hand side of (A7) has mean 0 and variance $O[(ns^3)^{-1}]$ for each $x^1 \in [t, 1-t]$, so $L_{2b2}(x^1) = o_p(1)$. In addition, $(nt)^{-1/2}L_{2b2}(x^1) = o_p(1)$ uniformly over $x^1 \in [t, 1-t]$. Q.E.D.

Proof of Theorem 2: It follows from lemma 8 that

$$(nt)^{1/2}\hat{m}_1(x^1) = (nt)^{1/2}\tilde{m}_1(x^1) - \frac{(nt)^{-1/2}S_{n10}(x^1)}{(nt)^{-1}S_{n20}(x^1)} + o_p(1)$$

for each $x^1 \in (0,1)$. Moreover,

$$\hat{m}_1(x^1) = \tilde{m}_1(x^1) - \frac{(nt)^{-1}S_{n10}(x^1)}{(nt)^{-1}S_{n20}(x^1)} + o_p(1)$$

uniformly over $x^1 \in [t, 1-t]$. Now proceed as in the proof of Theorem 2 of Horowitz and Mammen (2004). Q.E.D.

REFERENCES

Breiman, L. and Friedman, J.H. (1985). Estimating optimal transformations for multiple regression and correlation. *Journal of the American Statistical Association*, 80, 580-598.

Buja, A., Hastie, T. and Tibshirani, R.J. (1989). Linear smoothers and additive models. *Annals of Statistics*,17, 453-510.

Carrasco, M., J.-P. Florens, and E. Renault (2005). Linear inverse problems in structural econometrics: estimation based on spectral decomposition and regularization. In *Handbook of Econometrics*, Vol. 6, E.E. Leamer and J.J. Heckman, eds, Amsterdam: North-Holland, forthcoming.

Chen, R., Härdle, W., Linton, O.B., and Severance-Lossin, E. (1996). Estimation in additive nonparametric regression, in *Proceedings of the COMPSTAT Conference Semmering*, ed. by W. Härdle and M. Schimek, Heidelberg: Phyisika Varlag.

Fan, J. and Chen, J. (1999). One-step local quasi-likelihood estimation. *Journal of the Royal Statistical Society* B, 61, 927-943.

Fan, J., Härdle, W., and Mammen, E. (1998). Direct estimation of low-dimensional components in additive models. *Annals of Statistics*, 26, 943-971.

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

Horowitz, J.L. (2001). Nonparametric estimation of a generalized additive model with an unknown link function, *Econometrica*, 69, 499-513.

Horowitz, J.L. and Mammen, E. (2004). Nonparametric estimation of an additive model with a link function, *Annals of Statistics*, 32, 2412-2443.

Horowitz, J.L. and Mammen, E. (2007). Rate-optimal estimation for a general class of nonparametric regression models with unknown link functions, *Annals of Statistics*, forthcoming..

Horowitz, J.L., Klemelä, J. and Mammen, E. (2006). Optimal estimation in additive regression models, *Bernoulli*, 12, 271-298.

Hristache, M., Juditsky, A., and Spokoiny, V. (2001). Structure Adaptive Approach for Dimension Reduction, *Annals of Statistics*, 29, 1-32.

Ichimura, H (1993). Semiparametric least squares (SLS) and weighted SLS estimation of single-index models. *Journal of Econometrics* 58: 71-120.

Jennrich, R.I. (1969). Asymptotic properties of non-linear least squares estimators, *Annals of Mathematical Statistics*, 40, 633-643.

Juditsky, A.B., O.V. Lepski, and A.B. Tsybakov (2007). Nonparametric estimation of composite functions, working paper, University of Paris VI.

Linton, O.B (2000). Efficient estimation of generalized additive nonparametric regression models, *Econometric Theory*, 16, 502-523.

Linton, O.B. and Härdle, W. (1996). Estimating additive regression with known links. *Biometrika*, 83, 529-540.

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

Mammen, E., Linton, O.B., and Nielsen, J.P. (1999). The existence and asymptotic properties of backfitting projection algorithm under weak conditions, *Annals of Statistics*, 27, 1443-1490.

Newey, W.K. (1994). Kernel estimation of partial means and a general variance estimator, *Econometric Theory*, 10, 233-253.

Newey, W.K. (1997). Convergence rates and asymptotic normality for series estimators. *Journal of Econometrics*, 79, 147-168.

Opsomer, J.D. (2000). Asymptotic properties of backfitting estimators. *Journal of Multivariate Analysis*, 73, 166-179.

Opsomer, J.D. and Ruppert, D. (1997). Fitting a bivariate additive model by local polynomial regression. *Annals of Statistics*, 25, 186-211.

Serfling, R.J. (1980). *Approximation Theorems of Mathematical Statistics*, New York: Wiley.

Stone, C.J. (1985). Additive regression and other nonparametric models, *Annals of Statistics*, 13, 689-705.

Stone, C.J. (1986). The dimensionality reduction principle for generalized additive models. *Annals of Statistics*, 14, 590-606.

Stone, C.J. (1994). The use of polynomial splines and their tensor products in multivariate function estimation, *Annals of Statistics*, 2, 118-171

Tjøstheim, D. and Auestad, B.H. (1994). Nonparametric identification of nonlinear time series: projections, *Journal of the American Statistical Association*, 89, 1398-1409.

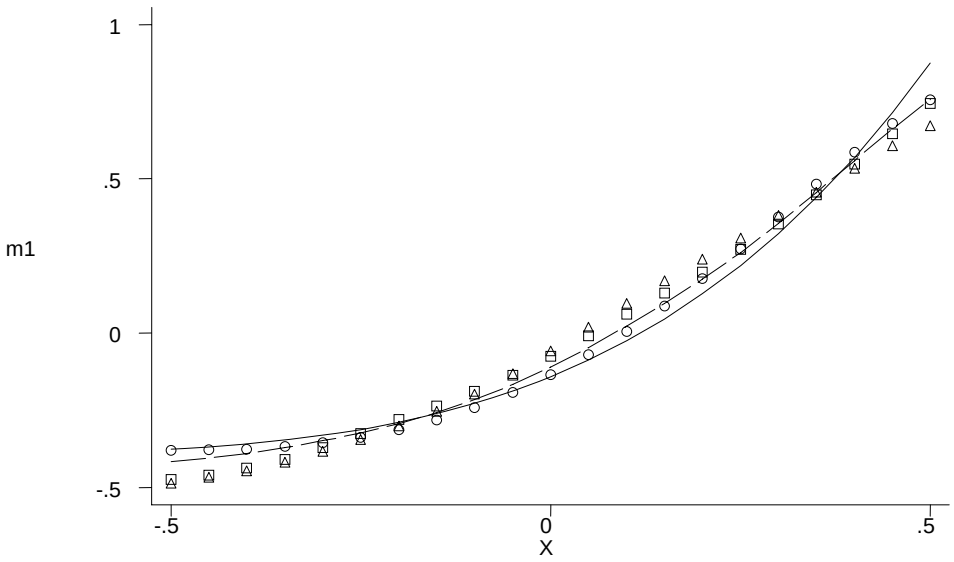

Figure 1: Performance of Second-Stage Estimator. Solid line is true m_1, dashed line is average of 100 estimates of m_1, small circles denote the estimate at the 25th percentile of the IMSE, squares denote the estimate at the 50th percentile of the IMSE, and diamonds denote the estimate at the 75th percentile of the IMSE.

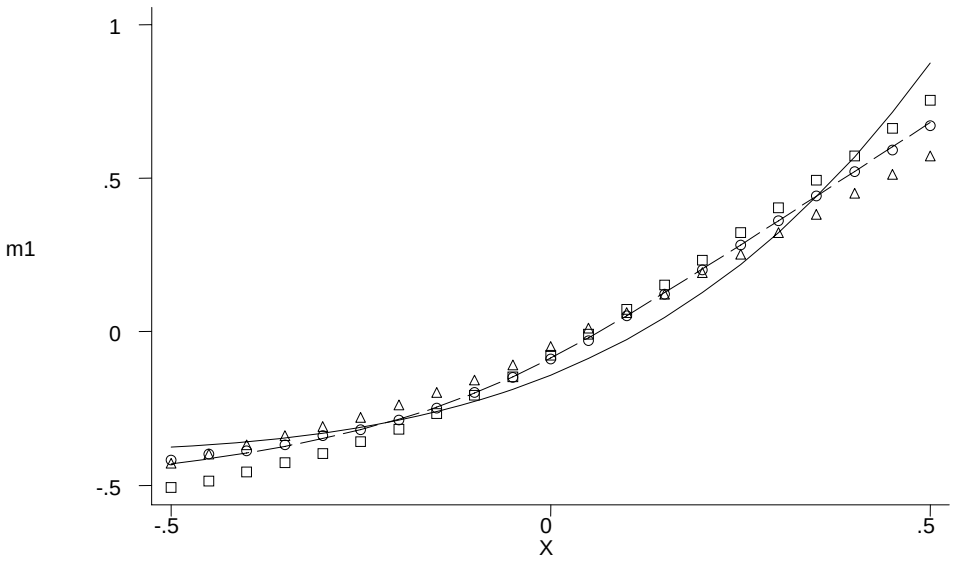

Figure 2. Performance of Infeasible Oracle Estimator. Solid line is true m_1, dashed line is average of 100 estimates of m_1, small circles denote the estimate at the 25th percentile of the IMSE, squares denote the estimate at the 50th percentile of the IMSE, and diamonds denote the estimate at the 75th percentile of the IMSE.