8-1-1994

Local Nonlinear Least Squares Estimation: Using Parametric Information Nonparametrically

Pedro Gozalo

Oliver B. Linton

COWLES FOUNDATION FOR RESEARCH IN ECONOMICS
AT YALE UNIVERSITY

Box 2125, Yale Station
New Haven, Connecticut 06520

COWLES FOUNDATION DISCUSSION PAPER NO. 1075

LOCAL NONLINEAR LEAST SQUARES: USING PARAMETRIC INFORMATION IN NONPARAMETRIC REGRESSION

Pedro Gozalo and Oliver Linton

August 1994
Revised: December 10, 1997

LOCAL NONLINEAR LEAST SQUARES: USING PARAMETRIC INFORMATION IN NONPARAMETRIC REGRESSION

Pedro Gozalo and Oliver Linton
Brown University and Yale University

December 10, 1997

Abstract

We introduce a new nonparametric regression estimator that uses prior information on regression shape in the form of a parametric model. In effect, we nonparametrically encompass the parametric model. We obtain estimates of the regression function and its derivatives along with local parameter estimates that can be interpreted from within the parametric model. We establish the uniform consistency and derive the asymptotic distribution of the local parameter estimates and of the corresponding regression and derivative estimates. For estimating the regression function our method has superior performance to the usual kernel estimators at or near the parametric model. It is particularly well motivated for binary data using the probit or logit parametric model as a base. We include an application to the Horowitz (1993) transport choice dataset.

Some key words: Binary Choice; Kernel; Local Regression; Nonparametric Regression; Parametric Regression.
Journal of Economic literature classification: C4, C5

1 Introduction

Methods for estimating nonlinear parametric models such as the generalized method of moments, nonlinear least squares or maximum likelihood are generally easy to apply, the results are easy to interpret, and the estimates converge at the rate $n^{1/2}$. However, parametric models can often impose too much structure on the data leading to inconsistent estimates and misleading inference. By contrast, methods based on nonparametric smoothing provide valid inference under a much broader class of structures. Unfortunately, the robustness of nonparametric methods is not free. Centered nonparametric smoothing estimators [such as kernels, nearest neighbors, splines, series] converge at rate $n^{1/2}s$, where $s \to 0$ is a smoothing parameter, which is slower than the $n^{1/2}$ rate for parametric estimators. The quantity s must shrink so that the 'smoothing' bias will vanish. Although this bias shrinks to zero as the sample size increases, it can be appreciable for moderate samples, and indeed is present in the limiting distribution of the standardized estimator when an optimal amount of smoothing is chosen. It is the presence of this limiting bias, as well as the slow rate of convergence, that distinguishes nonparametric from parametric asymptotic theory. In practice, the bias can seriously distort one's impression of the underlying relationship.[1] Therefore, it is important to have control of the bias and to try to minimize its influence on inference.

Some recent developments in the nonparametric smoothing literature have addressed these issues. Fan (1992) derived the statistical properties of the local linear estimator which involves fitting a line locally (as opposed to fitting a constant locally which is what the Nadaraya-Watson estimator does). Fan showed that local linear is design adaptive: its bias depends on the design to first order only through the second derivatives of the regression function. This is in contrast with the Nadaraya-Watson estimator whose bias also depends on the logarithmic derivative of the design density and can thus be quite large for some design densities even when the regression function is linear.[2]

There are special cases in which smoothing bias is small. For example, the Nadaraya-Watson regression smoother is exactly unbiased when the function being estimated is a constant. We say that Nadaraya-Watson estimators are centred at constants. Local polynomial estimators, see Fan and Gijbels (1992) and references therein, are centred at the corresponding polynomial. In fact, the local q'th order polynomial is centred at a reparameterization $p(z) = \alpha_0 + \alpha_1(z-x) + \ldots + \frac{\alpha_q}{q!}(z-x)^q$

[1] Confidence intervals are usually constructed without regard to this bias. This practice is justified by undersmoothing, see Bierens and Pott-Buter (1990). The bias effect can be worse in the boundary regions which are often of primary interest, see Müller (1988) and Anand et al. (1993) for some estimators, and in regions where the marginal density of the explanatory variables is changing rapidly.

[2] Fan (1993) formalized this further by showing that local linear is superior to Nadaraya-Watson according to a minimax criterion.

of the polynomial regression function $p(z) = \gamma_0 + \gamma_1 z + \ldots + \gamma_q z^q$ in which the α parameters are exactly the regression function and its derivatives. This reparameterization, however, changes with each evaluation point so the α parameters are not directly interpretable inside the global parametric model. It is of interest to estimate the γ parameters which do have an interpretation when the parametric model is true everywhere.

We introduce a new kernel nonparametric regression estimator that can be centred at any parametric regression function. Our approach is to nonparametrically encompass the parametric model and to estimate the local parameters using information from a neighborhood of the point of interest. To facilitate the asymptotics we introduce a local reparameterization that separates out the parameters with common convergence rates, specifically setting one parameter equal to the regression function and others equal to partial derivatives.[3] Our regression function estimator is consistent for all possible functional forms, is design adaptive, but has especially good properties (i.e., essentially no bias) at or near the parametric model. This centering approach to nonparametric estimation has been recently considered as well in spline regression and series density estimation, see Ansley, Kohn and Wong (1993) and Fenton and Gallant (1996), but not as yet for kernels. The advantage of working with kernels is that we are able to obtain the asymptotic distribution of our procedure taking account of bias and variance. This important result is not available for these other smoothing methods.

We also derive the asymptotic properties of estimates of the identifiable original parameters of the parametric model by the delta method. These local parameter estimates are interpretable inside the parametric model and are of interest in their own right. They can be used as a device for evaluating the parametric model. See Staniswalis and Severini (1991) for a formal analysis of a likelihood ratio type test statistic in a similar setting, and Gourieroux, Monfort, and Tenreiro (1994) for an econometric time series application.

The paper is organized as follows. In the next section we introduce the model and procedure. In section 3 we establish uniform consistency and derive the asymptotic distribution of our estimators. Section 4 contains an application to the transport choice data set of Horowitz (1993) and includes some discussion of bandwidth choice. Section 5 concludes. Two appendices contain the proofs. Our consistency proof follows an approach similar to that found for example in the results of Pötscher and Prucha (1991), combined with the empirical process techniques of Pollard (1984) and Andrews (1994).

For any vectors $x = (x_1, \ldots, x_d)$ and $a = (a_1, \ldots a_d)$, define $|x| = \sum_{j=1}^{d} x_j$, $x! = x_1! \times \cdots \times x_d!$, and $x^a = (x_1^{a_1}, \ldots, x_d^{a_d})$ for any integer d, also let for any function g: $\mathbb{R}^{d+c} \to \mathbb{R}$,

[3] As is commonly done for the local linear estimator.

$$D_x^a g(x,y) = \frac{\partial^{|a|}}{\partial x_1^{a_1} \cdots \partial x_d^{a_d}} g(x,y), \quad x \in \mathbb{R}^d, y \in \mathbb{R}^c,$$

with $D_x^a = D^a$ when there is no ambiguity as to the variable being differentiated. We use $\|A\| = \left\{\text{tr}\left(A^T A\right)\right\}^{1/2}$ to denote the Euclidean norm of a vector or matrix A. Convergence of sets should be taken to mean relative to the Hausdorff metric $\rho_H(\cdot,\cdot)$, which is defined for compact subsets A, B of $\mathbb{R}^p$ by:

$$\rho_H(A,B) = \inf\left\{\delta\text{: } A \subset B^\delta \text{ and } B \subset A^\delta\right\},$$

where $A^\delta = \{x\text{: } \rho(x,A) < \delta\}$ is the δ-neighborhood of the set A with respect to the usual Euclidean distance $\rho(x,A) = \inf_{y\in A}\|x-y\|$, see Kelley (1955, p131). Finally, $\xrightarrow{P}$ denotes convergence in probability and $\Rightarrow$ signifies weak convergence.

2 The model and procedure

The data to be considered are given by the following assumption.

ASSUMPTION A0: *Our sample* $\{Z_i\}_{i=1}^n$, *where* $Z_i = (Y_i, X_i)$, *is* n *realizations drawn from an i.i.d. random vector* $Z = (Y, X)$ *defined on a probability space* $(\Omega, \mathcal{F}, P)$, *where the response variable* Y *takes values in* $\mathcal{Y} \subseteq \mathbb{R}$, *the covariates* X *take values on* $\mathcal{X} \subseteq \mathbb{R}^d$, *for some* d, *and* $E[Y^2] < \infty$.

Assumption 0 guarantees the existence of Borel measurable real valued functions g and σ^2 with domain $\mathcal{X}$ such that $E(Y|X=x) = g(x)$ and $\text{var}(Y|X=x) = \sigma^2(x)$. Let $u_i = Y_i - g(X_i)$, then

$$E(u_i|X_i) = 0 \quad a.s. \tag{1}$$

Our main objective is to estimate the unknown regression function g at an interior point x without making explicit assumptions about its functional form. We do, however, wish to introduce some potentially relevant information taking the form of a parametric function $m(X,\alpha)$. For concreteness, we carry this out using the (nonlinear) least squares loss function, although the idea can be extended

to a likelihood, (generalized) method of moments, or M-estimation setup. We therefore introduce the following local nonlinear least squares criterion

$$Q_n(x,\alpha) = n^{-1}\sum_{i=1}^{n}\{Y_i - m(X_i,\alpha)\}^2 \; K_H(X_i - x), \tag{2}$$

where $K_H(\cdot) = \det(H)^{-1}K(H^{-1}\cdot)$ with H a $d\times d$ nonsingular bandwidth matrix, while $K(\cdot)$ is a real-valued kernel function. We first minimize Q_n with respect to α over the parameter space $\mathcal{A}_x \subset \mathbb{R}^p$, letting $\mathcal{M}_n(x)$ be the set of minimizers. Then, for any $\widehat{\alpha}(x) \in \mathcal{M}_n(x)$, let

$$\widehat{g}(x) = m\{x, \widehat{\alpha}(x)\} \tag{3}$$

be our estimator of $g(x)$, and let

$$\widehat{D^a g}(x) = D_x^a m(x, \widehat{\alpha}(x)), \tag{4}$$

be an estimate of $D^a g(x)$ for any vector $a = (a_1, \ldots, a_d)$, where this is well-defined. In general, iterative methods are required to find $\widehat{\alpha}(x)$ and the procedure might be quite computationally expensive if one needs to estimate at a large number of points x. However, in our experience only a small number of iterations are required to find each maximum since the globally optimal parameter values provide good starting points.

Our procedure is related to the local likelihood estimator of Tibshirani (1984), although this work appears to focus on generalized linear models in which a single parameter θ can enter any "link" function [an example of a link function is the normal c.d.f. or its inverse in a probit model]. There are similarities with the nonparametric maximum likelihood estimator of Staniswalis (1989) in which there is again only one location parameter, but a more general criterion than least squares is allowed for. There has been much recent work in the nonparametric statistics literature, mostly focusing on density and hazard estimation. In particular see: Copas (1994), Hjort (1993), Hjort and Jones (1996) and Loader (1996). Work in other areas includes Robinson (1989) for a time series regression problem and the recent paper by Hastie and Tibshirani (1993) about random coefficient models. Work on regression has proceeded less quickly, and we appear to be among the first to fully exploit this idea in regression, although see the recent discussion of Hjort and Jones (1994). Our procedure nests the local polynomial regression that originated with Stone (1977) and was recently popularized by Fan (1992). These procedures are discussed in Härdle and Linton (1994, see p17 especially).

We now give some examples.

EXAMPLE 1. Suppose the parametric regression function is linear in the parameters

$$m(x, \alpha) = \alpha_1 r_1(x) + \ldots + \alpha_p r_p(x),$$

where $r_1, \ldots, r_p$ are known functions. Many demand and cost systems fall into this category, see Deaton (1986). In this case, the minimization problem (2) has, for each n, an explicit solution

$$\widehat{\alpha} = (\mathbf{R}^T \mathbf{K} \mathbf{R})^+ \mathbf{R}^T \mathbf{K} Y,$$

where $\mathbf{R}$ is an $n \times p$ matrix with (i, j) element $r_j(X_i)$, $Y = (Y_1, \ldots, Y_n)^T$, $\mathbf{K}$ =diag$\{K_H(X_i - x)\}$, and superscript + denotes a generalized inverse. For example, when $m(x, \alpha) = \alpha$, then $\widehat{g}(x)$ is the Nadaraya-Watson estimator. This special case also includes the local polynomial estimators of Fan (1992).

EXAMPLE 2. (Index structures) Let $m(x, \alpha) = F(\alpha_1 + \sum_{j=2}^{p} \alpha_j x_{j-1})$, for some function $F(\cdot)$. This generalization provides a convenient way to impose inequality restrictions like $0 \leq g(x) \leq 1$. For example, with binary data one might take the logit or probit c.d.f.'s; in a sample selection model taking F to be an inverse Mills ratio would be appropriate. This type of structure is used widely; for example in semiparametric models, where F would be of unknown functional form while α would be fixed parameters. See especially Stoker (1986). In our approach, F is taken to be known, while α is allowed to vary with x. Our method is fully nonparametric, while the semiparametric approach adopted in Stoker (1986) entails some restrictions on the modeled distribution. Fan, Heckman and Wand (1995) consider the case where x is scalar and $m(x, \alpha) = F(\alpha_1 + \alpha_2 x + \ldots + \alpha_p x^{p-1})$ for some known link function F.

EXAMPLE 3. (Production functions). A number of common micro production functions can be used in (2), including: the Cobb-Douglas $m(x, \alpha) = \alpha_1 x_1^{\alpha_2} \ldots x_d^{\alpha_p}$, the constant elasticity of substitution $m(x, \alpha) = \left\{\alpha_1 + \alpha_2 x_1^{\alpha_p} + \ldots + \alpha_{p-1} x_d^{\alpha_p}\right\}^{1/\alpha_p}$, and the Leontieff $m(x, \alpha) = \min\{\alpha_1, \alpha_2 x_1, \ldots, \alpha_p x_d\}$. The Leontieff function although continuous is not differentiable. Note that the Cobb-Douglas functional form implicitly imposes the requirement that $g(0) = 0$. One can also impose, for example, constant returns to scale by writing $\alpha_p = 1 - \sum_{j=2}^{p-1} \alpha_j$ and minimizing (2) with respect to $\alpha = (\alpha_1, \ldots, \alpha_{p-1})$. See McManus (1994) for an application of nonparametric production function estimation.

3 Asymptotic properties

In this section we derive the uniform consistency and pointwise asymptotic normality of the regression and derivative estimators. We first outline the argument and then give the theorems in two subsections.

We shall show, under only continuity conditions on m and g, that the criterion function $Q_n(x,\alpha)$ converges uniformly [in x and α] as $n \to \infty$ to the non-random function

$$Q(x,\alpha) = \{g(x) - m(x,\alpha)\}^2 \; f_X(x) + \sigma^2(x) f_X(x), \tag{5}$$

where $f_X(x)$ is the marginal density of X, and let $\Phi_0(x)$ be the set of parameter values minimizing $Q(x,\alpha)$. Under additional smoothness, i.e., continuous derivatives of order $r > 0$, one can approximate $Q_n(x,\alpha)$ by a deterministic function $\overline{Q}^r(x,\alpha)$, obtained from a Taylor series expansion of $E\left[Q_n(x,\alpha)\right]$, with a smaller error. Let $\Phi_r(x)$ be the set of parameter values minimizing $\overline{Q}^r(x,\alpha)$ with respect to α; this set is defined inductively, i.e., $\Phi_j(x) = \{\alpha \in \Phi_{j-1}(x) : \alpha = \arg\min C_j(x,\alpha)\}$, for $j = 0, 1, \ldots, r$,[4] where

$$C_j(x,\alpha) = \int \left[\sum_{\{a:|a|=j\}} \frac{v^a}{a!} \left\{ D^a g(x) - D_x^a m(x,\alpha) \right\} \right]^2 K(v) dv. \tag{6}$$

Now similarly define $\Psi_0(x) = \{\alpha \in \mathcal{A}_x : g(x) = m(x,\alpha)\}$ and $\Psi_j(x) = \{\alpha \in \Psi_{j-1}(x) : D^a g(x) = D_x^a m(x,\alpha); |a| = j\}$ for $j = 1, \ldots, r$. Note that $\Psi_j(x) \subseteq \Phi_j(x)$ for $j = 0, 1, \ldots r$, while, under continuity and compactness assumptions, $\Phi_j(x) \neq \emptyset$. We shall assume that $\Phi_0(x) = \Psi_0(x)$ and that $\mathcal{A}_x$ is compact. A compact parameter space is frequently assumed in nonlinear problems for technical reasons; it guarantees that $\mathcal{M}_n(x) \neq \emptyset$. Nevertheless, it is not strictly necessary: it can be omitted when the criterion function is convex in the parameters, for example, see Pötscher and Prucha (1991). The requirement that $\Psi_0(x) \neq \emptyset$ rules out inappropriate functional forms: for example, if m were a cumulative distribution function, but the regression function itself were negative at the point of interest.

Let $\alpha^0(x)$ be a typical element of $\Phi_j(x)$ for whichever relevant $j = 0, \ldots, r$; a key issue here is identification of the parameter vector $\alpha^0(x)$ as well as $g(x)$ and its derivatives. The identification of the parameter $\alpha^0(x)$ will depend on the model chosen, and in particular on the number of parameters p relative to the number of derivatives, which we assume generically to be $r \geq 0$, of g and m at x. Letting $p_r = \sum_{j=0}^{r} t_j$, where $t_j = \binom{j+d-1}{d-1}$ is the number of distinct j'th order partial derivatives, we

[4] With the convention that $\Phi_{-1}(x) = \mathcal{A}_x$.

say that $\alpha^0(x)$ is underidentified when $p > p_r$, and identified when $p \leq p_r$.[5] The leading example of an underidentified case is when g is only a continuous function but $p > 1$. In the underidentified case, $\Phi_r(x)$ may have any cardinality and $\alpha^0(x)$ cannot be uniquely identified. Nevertheless, we establish that $\mathcal{M}_n(x)$ is uniformly strongly ρ_H-consistent for the set $\Phi_s(x)$, with $0 \leq s \leq r$, in the sense that

$$\sup_{x \in \mathcal{X}_0} \rho_H\{\mathcal{M}_n(x), \Phi_s(x)\} \to 0 \quad a.s., \tag{7}$$

for any compact set $\mathcal{X}_0 \subset \mathcal{X}$. In other words, every element of $\mathcal{M}_n(x)$ converges to an element of $\Phi_s(x)$. When $g(x)$ is only continuous but $p > 1$, even though $\alpha^0(x)$ is not identified, the regression function itself can be consistently estimated by the following argument. For any sequence of functions $\widehat{\alpha}(\cdot), \alpha^0(\cdot)$ with $\widehat{\alpha}(x) \in \mathcal{M}_n(x)$ and $\alpha^0(x) \in \Psi_0(x)$, we have that with probability one

$$\begin{aligned} Q(x, \widehat{\alpha}(x)) &= Q_n(x, \widehat{\alpha}(x)) + o(1) && (8) \\ &\leq Q_n(x, \alpha^0(x)) + o(1) && (9) \\ &= Q(x, \alpha^0(x)) + o(1), && (10) \end{aligned}$$

where the $o(1)$ errors are uniform in x [this is established in the appendix], while the inequality in (9) follows by definition of $\widehat{\alpha}(x)$. Since $Q(x, \widehat{\alpha}(x)) \geq Q(x, \alpha^0(x))$ by definition, we conclude that with probability one $Q(x, \widehat{\alpha}(x)) = Q(x, \alpha^0(x)) + o(1)$ and hence

$$m(x, \widehat{\alpha}(x)) = m(x, \alpha^0(x)) + o(1),$$

uniformly in x, where $m(x, \alpha^0(x)) = g(x)$ by assumption. In other words, under our conditions, we have that

$$\sup_{x \in \mathcal{X}_0} |\widehat{g}(x) - g(x)| \to 0 \quad a.s., \tag{11}$$

regardless of the cardinality of $\Psi_0(x)$. When $g(x)$ possesses r continuous partial derivatives and $\Psi_r(x) \neq \emptyset$, a similar argument allows us to conclude that for all vectors $a = (a_1, \ldots, a_d)$ with $|a| \leq r$,

$$\sup_{x \in \mathcal{X}_0} \left| \widehat{D^a g}(x) - D^a g(x) \right| \to 0 \quad a.s. \tag{12}$$

[5] Of course, these are only necessary conditions. The distinction between exactly identified, $p = p_r$, and overidentified, $p < p_r$ plays no role in our analysis.

In conclusion, the regression function and derivatives can typically be consistently estimated even when the parameters are not.

We now turn to the underparameterized scenario with $p \leq p_r$, in which case it may be possible to uniquely identify $\alpha^0(x)$, i.e., we can expect that $\Phi_r(x)$ will be a singleton. The 'identification' of $g(x)$ and its derivatives, however, is a different story. There are two subcases. First, when $p = p_s$ for some $s \leq r$, then we can expect $\Psi_j(x) \neq \emptyset$ and in fact $\Psi_j(x) = \Phi_j(x)$ for each $j \leq s$. In this case, $D^a g(x)$ for any vector a with $|a| \leq s$ can be uniquely identified; indeed, one can estimate these quantities via (4). The second case we call 'unsaturated' and corresponds to p satisfying $p_{s-1} < p < p_s$ for some $s \leq r$. In this case, the criterion function $C_s(x, \alpha)$ cannot be set to zero, and although we can identify and estimate $\alpha^0(x)$ as well as the partial derivatives through order $s-1$, we are not able to consistently estimated all of the s'th order partial derivatives by this method, i.e., $\Psi_j(x) = \Phi_j(x)$ for each $j \leq s-1$ but $\Psi_s(x) = \emptyset$.[6] The following examples clarify these points.

Example 4.

(a) Suppose that $m(x, \alpha) = \alpha_1 + \alpha_2 x$, i.e., $d = 1$ and $p = 2$. Then, $\Phi_0(x) = \{\alpha \in \mathcal{A}_x : g(x) = \alpha_1 + \alpha_2 x\} = \Psi_0(x)$ is a line segment in α-space. When $r \geq 1$,

$$\Phi_1(x) = \Psi_1(x) = \left\{ \alpha \in \mathcal{A}_x : g(x) = \alpha_1 + \alpha_2 x \text{and } \frac{dg}{dx}(x) = \alpha_2 \right\},$$

which is a singleton with $\alpha_1^0(x) = g(x) - x \cdot dg(x)/dx$ and $\alpha_2^0(x) = dg(x)/dx$.

(b) Now suppose that $m(x, \alpha) = \alpha_1 x_1 + \alpha_2 x_2$, i.e., $d = 2$ and $p = 2$. Then,

$$\Phi_0(x) = \Psi_0(x) = \{\alpha \in \mathcal{A}_x : g(x) = \alpha_1 x_1 + \alpha_2 x_2\}$$

is a line segment in α-space, while, when $r \geq 1$, $\Phi_1(x)$ is the orthogonal projection in $\mathbb{R}^2$ of the point $(\partial g(x)/\partial x_1 \, , \partial g(x)/\partial x_2)$ onto the affine subspace $\Phi_0(x)$; this gives a unique element of α-space both components of which are, in fact, linear combinations of $g(x)$, $\partial g(x)/\partial x_1$, and $\partial g(x)/\partial x_2$. However, these parameter values do not set (6) identically zero, i.e., $\Psi_1(x) = \emptyset$, so that although $g(x) = \alpha_1^0(x) \cdot x_1 + \alpha_2^0(x) \cdot x_2$, neither $\partial g(x)/\partial x_1 = \alpha_1^0(x)$ nor $\partial g(x)/\partial x_2 = \alpha_2^0(x)$ unless $x_1 \cdot \partial g(x)/\partial x_1 + x_2 \cdot \partial g(x)/\partial x_2 = g(x)$.

[6] When K is sufficiently smooth, it is possible to estimate the remaining partial derivatives by direct differentiation of $\widehat{\alpha}(x)$ with respect to x.

In the identified case, $p \leq p_r$, one can use a standard Taylor series expansion about the single point $\alpha^0(x)$ to establish asymptotic normality for $\widehat{\alpha}(x)$. In the underidentified case consistency of the function and derivative estimates is established; however, the fact that the smallest set $\Phi_r(x)$ need not be a singleton makes proving the asymptotic distribution theory more complicated and is not attempted here.

For notational convenience in the sequel we restrict our attention to scalar bandwidths $H = hI$ and product kernels $K_h(v) = \prod_{j=1}^{d} k_h(v_j)$ where $k_h(\cdot) = h^{-1}k(\cdot/h)$.

3.1 Uniform Consistency

Let $\mathcal{X}_0$ be a compact set with $\mathcal{X}_0^\delta \subset \mathcal{X}$ for some $\delta > 0$, and let $\mathcal{A}$ be the closure of $\cup_{x\in\mathcal{X}_0}\mathcal{A}_x$. Define the class $\mathcal{F}_{r;s,\lambda}(\mathcal{X}_0 \times \mathcal{A})$ of functions q: $\mathcal{X}_0 \times \mathcal{A} \to \mathbb{R}$ for which $D^{(a,b)}_{(x,\alpha)}q(x,\alpha)$ exists and is continuous in x and α for all vectors a, b with $|a| \leq r$ and $|b| \leq s$, and, furthermore, there exists a non-negative bounded function ϕ and a positive constant λ such that for all $x \in \mathcal{X}_0$ and $\alpha, \alpha' \in \mathcal{A}_x$,

$$\left| D^{(a,b)}_{(x,\alpha)}q(x,\alpha) - D^{(a,b)}_{(x,\alpha)}q(x,\alpha') \right| \leq \phi(x) \left\|\alpha - \alpha'\right\|^\lambda$$

for all vectors a, b with $|a| \leq r$ and $|b| \leq s$. Functions smooth in x that do not explicitly depend on α can be embedded in a suitable $\mathcal{F}_{j;0,0}(\mathcal{X}_0 \times \mathcal{A})$ in an obvious manner.

Assumption A

(1) *The marginal density of X, f_X, is bounded away from zero for all $x \in \mathcal{X}_0$. Furthermore, $f_X, \sigma^2 \in \mathcal{F}_{0;0,0}(\mathcal{X}_0 \times \mathcal{A})$, while $g \in \mathcal{F}_{r;0,0}(\mathcal{X}_0 \times \mathcal{A})$ and $m \in \mathcal{F}_{r;r,\lambda}(\mathcal{X}_0 \times \mathcal{A})$ for some $\lambda > 0$.*

(2) *For an $s \leq r$ specified in the theorem, we suppose that for all $t = 0, 1, \ldots, s$ we have the following condition. For any δ-neighborhood $\Phi_t^\delta(x)$ of $\Phi_t(x)$, $\delta > 0$, there is an $\varepsilon > 0$ such that for all subsets $A(x) \subseteq \Phi_{t-1}(x) \setminus \Phi_t^\delta(x)$, we have*

$$\inf_{x\in\mathcal{X}_0} \inf_{\alpha\in A(x)} C_t(x,\alpha) \geq \varepsilon.$$

(3) *The kernel weighting function k is symmetric about zero, continuous, of bounded variation, and satisfies $\int k(t)dt = 1$, $\int |k(t)|dt \leq \hat{K} < \infty$, and for some $s \leq r$ specified in the theorem $\int k(t)t^{2s}dt < \infty$.*

(4) *$\{h(n) : \ n \geq 1\}$ is a sequence of nonrandom bounded positive constants satisfying $h \to 0$ and $nh^{d+2s}/\log n \to \infty$ for some $s \leq r$ specified in the theorem.*

Assumption A2 is a form of identifiable uniqueness necessary for the identification of $\Phi_s(x)$. It is a generalization of the typical identification condition in parametric models where $\Phi_t(x)$ is a singleton to the case where it can consist of a continuum of elements. Assumption A2 is plausible from the form of the criterion function in the leading cases $s = 0, 1, 2$. For example, under our assumptions, $C_1(x,\alpha) = \int t^2 k(t)dt \times \sum_{\{a:|a|=1\}} \{D^a g(x) - D^a m(x,\alpha)\}^2$, while

$$C_2(x,\alpha) = \text{vec}\{G'(x) - M'(x,\alpha)\}' \left\{ \int (vv' \otimes vv')K(v)dv \right\} \text{vec}\{G(x) - M(x,\alpha)\},$$

where $G(x) = \left(\frac{\partial^2 g}{\partial x_j \partial x_k}(x)\right)_{j,k}$ and $M(x,\alpha) = \left(\frac{\partial^2 m}{\partial x_j \partial x_k}(x,\alpha)\right)_{j,k}$ are matrices of second derivatives. In both these cases, we should have $\Psi_s(x) = \Phi_s(x) = \{\alpha^0(x)\}$, for $s = 1, 2$, provided there are sufficiently many parameters [i.e., $p = p_1$ or $p = p_2$]. This is clearly true in the case of $C_1(x,\alpha)$. It also holds for $C_2(x,\alpha)$ provided the matrix $\int (vv' \otimes vv')K(v)dv$ is finite and positive definite.

The bandwidth condition A4 is the same as in Silverman (1978), when $s = 0$, and is essentially the weakest possible. Specializing the assumptions and the theorem to $\mathcal{X}_0 = \{x\}$ we obtain a pointwise consistency result. In this case, the bandwidth condition can be weakened to requiring only $h \to 0$ and $nh^{d+2s} \to \infty$, see Härdle (1990) for similar conditions and discussion of the literature. Schuster and Yakowitz (1979) show uniform consistency for the Nadaraya-Watson regression estimator by direct methods. See Andrews (1994) for a review of the theoretical literature and results for more general sampling schemes using empirical process techniques.

THEOREM 1. *Suppose that assumptions A0-A4 hold for some $0 \le s \le r$, and let $\mathcal{M}_n(x)$, $x \in \mathcal{X}_0$ be the sequence of local nonlinear least squares estimators. (i) Then, (7) holds; (ii) Also, taking $s = 0$, (11) holds; and (iii) Taking s such that $p \ge p_s$, (12) holds for all vectors a with $|a| \le s \le r$, provided that for all $x \in \mathcal{X}_0$, $\Psi_s(x) \ne \emptyset$; (iv) Suppose in addition: that $p \le p_r$, that s in A2-A4 is such that $p_{s-1} < p \le p_s$, and that for this s, $\Phi_s(x) = \Psi_s(x) = \{\alpha^0(x)\}$ for all $x \in \mathcal{X}_0$. Then,*

$$\sup_{x \in \mathcal{X}_0} \left|\widehat{\alpha}(x) - \alpha^0(x)\right| \to 0 \quad a.s. \tag{13}$$

Note how in the underidentified case where $p > p_r$, $\Phi_r(x)$ is not a singleton, and although (7), (11) and (12) will hold, (13) will not.

In the next section we find the pointwise rate of convergence.

3.2 Asymptotic Normality

In this section we derive the pointwise asymptotic distributions of the local parameter estimates $\widehat{\alpha}(x)$, the regression function estimate $\widehat{g}(x)$, and the derivative estimates $\widehat{D^a g}(x)$. We restrict our attention to the identified case, and in particular take $p = p_1 = d + 1$ and $r \geq 2$ so that $\alpha^0(x)$ and $g(x)$ and its first partial derivatives can be uniquely identified. Our results are thus comparable with those of Ruppert and Wand (1994) for multivariate local linear estimators. This specialization is primarily for notational convenience, and we remark on the general case following the theorem.

Our approach is as follows. We first work with a convenient reparameterization of m, for which we use the original notation $\alpha(x)$, and derive the asymptotic properties of the local parameter estimates $\widehat{\alpha}(x)$ in this case. In fact, we work with a parameterization for which $\alpha_1^0(x) = g(x)$ and $\alpha_{j+1}^0(x) = \partial g(x)/\partial x_j$, $j = 1, \ldots, d$, so that $\widehat{\alpha}(x)$ actually estimates the regression function and its derivatives. The original parameters, which we now denote by γ, are smooth functions of $\alpha(x)$ and x. We then derive the properties of $\widehat{\gamma}(x) = \gamma(\widehat{\alpha}(x), x)$ by the delta method. The reason for our approach is as follows.

The asymptotic properties of extremum estimators depend crucially on the behaviour of the Hessian matrix and its appropriate normalization. We can generally find a sequence of $p \times p$ scaling matrices $H(n)$ with the property that

$$H^{-1} \frac{\partial^2 Q_n}{\partial \alpha \partial \alpha^T}(x, \alpha^0(x)) H^{-1} \xrightarrow{P} \mathcal{I}_{\alpha\alpha}(x, \alpha^0(x)),$$

where the limit matrix is positive definite. The ideal case is when the scaling matrices and the local information matrix $\mathcal{I}_{\alpha\alpha}(x, \alpha^0(x))$ are diagonal (or block diagonal), which we call an orthogonal parameterization, following Cox and Reid (1987). In this case the diagonal elements of $H(n)$ measure the relative rates of convergence of the asymptotically mutually orthogonal parameter estimates. For any given parametric function $m(x, \gamma)$, the components of γ each generally contains information about both the regression function $g(x)$ and its partial derivatives $\partial g(x)/\partial x_j$, $j = 1, 2, \ldots, d$, which are themselves known to be estimable at different rates, see Stone (1982). A consequence of this is that the corresponding information matrix may be singular or the scaling matrices may have a complicated structure that can even depend on g itself.

The parameterization itself is not unique, and can be changed without affecting the resulting estimate of $g(x)$. We therefore reparameterize to an orthogonal parameterization for which H is diagonal and $\mathcal{I}_{\alpha\alpha}$ is block diagonal. In fact, we work with the particular orthogonal parameterization for which $\alpha_1^0(x) = g(x)$ and $\alpha_{j+1}^0(x) = \partial g(x)/\partial x_j$, $j = 1, \ldots, d$, which we call canonical, in which the score for the parameter $\alpha_1^0(x)$ is orthogonal to the scores for each $\alpha_{j+1}^0(x)$, $j = 1, \ldots, d$. This

separates out the parameters with different convergence rates. Given $m(z,\gamma)$, a general method for finding its reparameterization is to solve the system of partial differential equations:

$$m(x,\gamma) = \alpha_1(x) \quad ; \quad \frac{\partial m}{\partial x_j}(x,\gamma) = \alpha_{j+1}(x), \quad j = 1,\ldots,d,$$

for $\gamma(\alpha(x),x)$. Then write $m(z,\alpha^0(x)) = m(z,\gamma(\alpha^0(x),x))$. That is, $m(z,\alpha^0(x))$ and its partial derivatives will equal $\alpha_1^0(x) = g(x)$ and $\alpha_{j+1}^0(x) = \partial g(x)/\partial x_j$, $j = 1,\ldots,d$, respectively, when evaluated at $z = x$. Frequently, the canonical orthogonal parameterization is given by replacing $m(z,\gamma)$ by $m(z-x,\alpha(x))$, as in the local linear estimator of Fan (1992).[7] To illustrate the general method consider the Cobb-Douglas model $m(z,\gamma) = \gamma_1 z_1^{\gamma_2}\cdots z_d^{\gamma_p}$. We have to solve

$$\gamma_1 x_1^{\gamma_2}\cdots x_d^{\gamma_p} = \alpha_1(x) \quad ; \quad \gamma_{j+1}x_j^{-1}\gamma_1 x_1^{\gamma_2}\cdots x_d^{\gamma_p} = \alpha_{j+1}(x), \quad j = 1,\ldots,d,$$

which yields

$$\gamma_{j+1}(\alpha(x),x) = x_j\frac{\alpha_{j+1}(x)}{\alpha_1(x)}\,, \quad j = 1,\ldots,d\,; \quad \gamma_1(\alpha(x),x) = \frac{\alpha_1(x)}{x_1^{x_1\alpha_2(x)/\alpha_1(x)}\cdots x_d^{x_d\alpha_p(x)/\alpha_1(x)}}\,,$$

and

$$m(z,\alpha(x)) = \alpha_1(x)\left(\frac{z_1}{x_1}\right)^{x_1\alpha_2(x)/\alpha_1(x)}\cdots\left(\frac{z_d}{x_d}\right)^{x_d\alpha_p(x)/\alpha_1(x)}.$$

We use some convenient notation. Derivatives are denoted by subscripts so that $m_\alpha(x,\alpha)$ is the $p\times 1$ vector of partial derivatives of m with respect to α, while $m_x(x,\alpha)$ and $m_{xx}(x,\alpha)$ are $d\times 1$ and $d\times d$ first and second derivative arrays of m, and $m_{\alpha x}(x,\alpha)$ is the $p\times d$ matrix of cross partials. Finally, let $\Delta_{xx} = g_{xx}(x) - m_{xx}(x,\alpha^0(x))$, and note that $\Delta_x = g_x(x) - m_x(x,\alpha^0(x)) = 0$. We make the following additional assumptions:

ASSUMPTION B

(1) *Assumption A2 is satisfied for* $\Phi_1(x) = \Psi_1(x) = \{\alpha^0(x)\}$ *with* $\alpha^0(x) = (g(x), g_x^T(x))^T$ *in the interior of* $\mathcal{A}_x$ *and* x *in the interior of* $\mathcal{X}$.

(2) *There is some neighborhood* $\mathcal{U}$ *of* $(x,\alpha^0(x))$ *in* $\mathcal{X}\times\mathcal{A}_x$ *for which* $m\in\mathcal{F}_{3;2,0}(\mathcal{U})$ *and* $g\in\mathcal{F}_{r;0,0}(\mathcal{U})$ *for some integer* r, *while* $f\in\mathcal{F}_{1;0,0}(\mathcal{U})$.

[7] A similar transformation works for index models, thus an orthogonal reparameterization of $F(\gamma_1+\sum_{j=1}^d\gamma_{j+1}z_j)$ is provided by $F[\alpha_1+\sum_{j=1}^d\alpha_{j+1}(z_j-x_j)]$; the canonical reparameterization is in fact $F[F^{-1}(\alpha_1)+\sum_{j=1}^d\frac{\alpha_{j+1}}{F'\{F^{-1}(\alpha_1)\}}(z_j-x_j)]$.

(3) *The functions* $m_\alpha m_\alpha^T(X,\alpha)\{Y - m(X,\alpha)\}$, $m_{\alpha\alpha}(X,\alpha)\{Y - m(X,\alpha)\}$ *and* $m_{\alpha xx}(X,\alpha)$ *are element-wise dominated for all* $\alpha \in \mathcal{A}_x$ *by* $s_{1x}(Y,X)$, $s_{2x}(Y,X)$ *and* $s_{3x}(X)$ *respectively, where* $E\{s_{jx}\} \leq S_{jx} < \infty$ *for* $j = 1,2,3$.

(4) *Let* $\nu_j(k) = \int t^j k^2(t)dt$ *and* $\mu_j(k) = \int t^j k(t)dt$ *for any integer* j. *The kernel* k *is symmetric about zero and* $0 < \nu_0(k), \nu_2(k), \mu_2(k) < \infty$.

We require only two moments for u [see A0], which is less than the $2+\delta$ commonly used − Bierens (1987) and Härdle (1990) − see Lemma CLT in Appendix A. The other conditions are strengthenings of the differentiability and boundedness conditions of Assumption A and are fairly standard to the nonparametric literature. The requirement that $\alpha^0(x)$ be an interior point of $\mathcal{A}_x$ in Assumption B1 is trivially satisfied: given the boundedness of g and its first order derivatives, we can always find a compact $\mathcal{A}_x$ with $\alpha^0(x)$ in the interior. Our theorem gives the properties of both a regression function estimator and a derivative estimator. Usually, these results are stated separately because (a) if one is only interested in the function itself one can get away with weaker smoothness conditions (twice not thrice continuous differentiability) and (b) the optimal bandwidth for these two problems is of different magnitude. The latter reason means that there are really two separate procedures: one optimized for the regression function and one for its derivatives.

THEOREM 2. *Let A0-A1, with* $\mathcal{X}_0 = \{x\}$, *A3 and B1-B4 hold. Then,*

(a) *If also* $\lim_{n\to\infty} h^2(nh^d)^{1/2} = c$, *with* $0 \leq c < \infty$, *and B2 holds with* $r = 2$, *the regression estimate* $\widehat{\alpha}_1(x)$ *satisfies*

$$(nh^d)^{1/2}\left\{\widehat{\alpha}_1(x) - \alpha_1^0(x)\right\} \Rightarrow N\left\{cb_1(x), \nu_0(k)\frac{\sigma^2(x)}{f_X(x)}\right\},$$

where $b_1(x) = \frac{1}{2}\mu_2(k)\mathrm{tr}(\Delta_{xx})$.

(b) *If also* $\lim_{n\to\infty} h^2(nh^{d+2})^{1/2} = c$, *with* $0 \leq c < \infty$, *and B2 holds with* $r = 3$, *the partial derivative estimates* $\widehat{\alpha}_j(x)$ *satisfy*

$$(nh^{d+2})^{1/2}\left\{\widehat{\alpha}_j(x) - \alpha_j^0(x)\right\} \Rightarrow N\left\{cb_j(x), \frac{\nu_2(k)\sigma^2(x)}{\mu_2^2(k)f_X(x)}\right\}, \quad j = 2,\ldots,p,$$

where $b_j(x)$ *is defined in* (31) *in Appendix A.*

(c) *Let* $V(x)$ *denote the asymptotic variance matrix of* $\{(nh^d)^{1/2}H\widehat{\alpha}(x)\}$, *with* $H = \mathrm{diag}(1,h,\ldots,h)$ *a* $p \times p$ *diagonal matrix. That is,* $v_{11}(x)$ *is the asymptotic variance of* $\{(nh^d)^{1/2}\widehat{\alpha}_{11}(x)\}$,

$v_{j1}(x) = v_{1j}(x)$ is the asymptotic covariance of $\{(nh^d)^{1/2}\widehat{\alpha}_1(x),\ (nh^{d+2})^{1/2}\widehat{\alpha}_j(x)\}$, $j = 2, \ldots, p,$, and $v_{ij}(x)$ is the asymptotic covariance of $\{(nh^{d+2})^{1/2}\widehat{\alpha}_i(x), (nh^{d+2})^{1/2}\widehat{\alpha}_j(x)\}$, $i, j = 2, \ldots, p$. Then, for h satisfying the conditions of either part (a) *or* (b)*, $v_{ij}(x) = 0, i \neq j, i, j = 1, \ldots, p$.*

(d) *Let $\widehat{V} = \widehat{A}^{-1}\widehat{B}\widehat{A}^{-1}$, where $\widehat{A} = H^{-1}n^{-1}\sum_{i=1}^n m_\alpha m_\alpha^T(X_i, \widehat{\alpha}(x))K_h(X_i - x)H^{-1}$ and $\widehat{B} = h^d H^{-1}n^{-1}\sum_{i=1}^n \widehat{u}_i^2 m_\alpha m_\alpha^T(X_i, \widehat{\alpha}(x))K_h^2(X_i - x)H^{-1}$, with $\widehat{u}_i = Y_i - m(X_i, \widehat{\alpha}(x))$. Then, for h satisfying the conditions of part* (a)*, $v_{11}(x)$ is consistently estimated by the $(1,1)$ element of $\widehat{V}$. For h satisfying the conditions of part* (b)*, $v_{jj}(x)$ is consistently estimated by the (j,j) element of $\widehat{V}$.*

Remarks

1. The asymptotic variance of $\widehat{g}(x)$ $[= \widehat{\alpha}_1(x)]$ is independent of the parametric model $m(x, \alpha)$ used to generate the estimate.[8] Furthermore, the bias of $\widehat{g}(x)$ does not depend on $f_X(x)$ to first order, i.e., the procedure is design adaptive in the sense of Fan (1992). Our work demonstrates that this is not due to the specific functional form chosen by Fan but solely to the *number* of parameters $[p = d + 1]$ chosen in the approximating parametric model.

2. The extension to the case $p = p_s$ for any integer $s = 0, 1, 2, \ldots$ is relatively straightforward and parallels closely the theory for local polynomial estimators described in Masry (1996). Namely, under corresponding smoothness conditions and bandwidth rates, the asymptotic variance of the reparameterized $\widehat{\alpha}_1(x)$ $[= \widehat{g}(x)]$ is the same as that for the corresponding local polynomial [i.e., the local polynomial of order s] estimator of $g(x)$, while the bias is the same as the corresponding local polynomial estimator except that derivatives of $g(x) - m(x, \alpha^0(x))$ replace derivatives of $g(x)$; the results for general local polynomials are stated in Theorem 5 of Masry (1996).

3. The extension to the unsaturated case with $p_{s-1} < p < p_s$ can also be briefly described. Suppose that we reparameterize the first p_{s-1} parameters to correspond to the first $s - 1$ partial derivatives, and then reparameterize the remaining parameters to match up with an arbitrary list of order s partial derivatives of length $p - p_{s-1}$. Then, the asymptotics parallel that for local polynomial estimators in which only some of the partial derivative parameters are included in the polynomial. In particular, the asymptotic variance of the reparameterized $\widehat{\alpha}_1(x)$ $[= \widehat{g}(x)]$ is the same as in remark 2 except for the kernel constant, which is now the first element of the matrix $M_p^{-1}\Gamma_p M_p^{-1}$, where M_p and Γ_p are the $p \times p$ matrices of kernel moments [each are submatrices of the kernel moment matrices appearing in Theorem 5 of Masry (1996)]. Finally, the bias contains only the derivatives

[8] It is already known that the variances of the Nadaraya-Watson and the local linear estimators are the same; our general class of models m includes the local linear (with $p = d + 1$) and the Nadaraya-Watson (with $p = 1$).

of $g(x) - m(x, \alpha^0(x))$ corresponding to the unmatched derivatives and has kernel constants similarly modified as in the variance.

A major advantage of our method arises when the parametric model is true or approximately true. If for some fixed α^0, $g(x) = m(x, \alpha^0)$ for all x, then $\widehat{g}(x)$ is essentially unbiased as the derivatives of all orders of $g(x) - m(x, \alpha^0)$ with respect to x equal zero for all x. In this case, there is no upper bound on feasible bandwidth and one could widen h to achieve faster convergence. In fact, parametric asymptotic theory applies to $\widehat{\alpha}(x)$ on substituting nh^d for n, since Q_n is then just a subsample weighted least squares criterion. More generally, the performance of our procedure is directly related to the value of the information contained in the parametric model m; specifically, the bias of $\widehat{g}(x)$ is proportional to the distance of m from g as measured by Δ_{xx}. If $g(x)$ is close to $m(x, \alpha^0)$ in the sense that

$$|\mathrm{tr}\,(\Delta_{xx})| \leq |\mathrm{tr}\,(g_{xx})|,$$

then the bias of $\widehat{g}(x)$ will be smaller than that of the comparable local linear estimators. For the local linear estimator, $m_{xx} \equiv 0$, and so this procedure has reduced bias only when $|\mathrm{tr}\,(g_{xx})|$ is small. If the regression function were closer to the Cobb-Douglas functional form than to a linear function, then the procedure with Cobb-Douglas m would have better performance as measured by smaller bias. The performance gain is directly proportional to the value of the information used in the construction of $\widehat{g}(x)$, unlike for conventional bias reduction techniques.[9]

The choice of parametric model is clearly important here. In some situations, there are many alternative plausible models such as logit or probit for binary data. We examine whether the sensible, but misspecified, choice of a logit regression function works better than plain local linear fitting, when the true regression is probit. The comparison is made with respect to the theoretical bias of the local linear and local logit procedures for the probit regression function $\Phi(1 + 0.5x)$, where $\Phi(\cdot)$ is the standard normal c.d.f., while x is a standard normal covariate. These parameter values are taken

[9]Fan (1993) establishes the minimax superiority of the local linear estimator over the Nadaraya-Watson estimator in the one-dimensional case; this result was established over essentially the following class of joint distributions D,

$$\mathcal{C} = \big\{D(\cdot,\cdot)\colon\ |g_{xx}(x)| \leq C, \quad f_X(x) \geq a, \quad \sigma^2(x) \leq b, \quad f \ \text{ is Lipschitz}\big\}$$

for positive finite constants C, a, and b. In fact, the local linear procedure is minimax superior to any other given parametric pilot procedure over this class. However, consider the modified class $\mathcal{C}'$ that replaces the bound on g_{xx} by $\big|g_{xx}(x) - m_{xx}(x, \alpha^0(x))\big| \leq C$. Then, one has the symmetric conclusion that the pilot $m(x, \alpha)$ generates a regression estimator with better minimax performance over $\mathcal{C}'$ than the local linear estimator.

(approximately) from a real dataset, see section 4 below. Figure 1 shows $|g_{xx}|$ and $|g_{xx} - m_{xx}|$ as a function of x. Local logit would appear to be considerably better.

*** FIGURE 1 HERE ***

Often there is interest in the original parameters γ. We can derive the asymptotic distribution of $\widehat{\gamma}(x)$ by an application of the delta method. Recalling the relationship $\gamma = \gamma(\alpha, x)$, (so that $\widehat{\gamma}_j(x) = \gamma_j(\widehat{\alpha}(x), x)$, and $\gamma_j^0(x) = \gamma_j(\alpha^0(x), x)$, $j = 1, \ldots, p,$) we obtain

$$\widehat{\gamma}_j(x) - \gamma_j^0(x) = \Gamma_j^T(x)\{\widehat{\alpha}(x) - \alpha^0(x)\} + o_p(\{nh^{d+2}\}^{-1}),$$

where $\Gamma_j(x)$ denotes the $p \times 1$ vector $\partial\gamma_j(\alpha, x)/\partial\alpha$ evaluated at $\alpha^0(x)$. An important consequence of this is that $\widehat{\gamma}_{nj}(x)$ inherits the slow convergence rates of the derivatives estimates $\widehat{\alpha}_2(x), \ldots, \widehat{\alpha}_p(x)$, unless all elements of $\Gamma_j(x)$ except the first one equal zero. The following corollary summarizes the results.

COROLLARY. *Suppose that the conditions of Theorem 2 hold and that $\gamma(\alpha, x)$ is continuously differentiable in α. Then,*

(a) *If* $\Gamma_j(x) = (\Gamma_{j1}(x), 0, \ldots, 0)^T$, $\Gamma_{j1}(x) \neq 0$, *then*

$$(nh^d)^{1/2}\left\{\widehat{\gamma}_j(x) - \gamma_j^0(x)\right\} \Rightarrow N\{c\Gamma_{j1}(x)b_1(x), \Gamma_{j1}^2(x)v_{11}(x)\}, \quad j = 1, \ldots, p,$$

where h satisfies the conditions of Theorem 2(a)*, and $b_1(x)$ and $v_{11}(x)$ are as defined in Theorem* 2.

(b) *If* $\Gamma_j(x) = (\Gamma_{j1}(x), \ldots, \Gamma_{jp}(x))^T$, *with* $\Gamma_{j\ell}(x) \neq 0$, *for some* $\ell = 2, \ldots, p$, *then*

$$(nh^{d+2})^{1/2}\left\{\widehat{\gamma}_j(x) - \gamma_j^0(x)\right\} \Rightarrow N\left\{c\sum_{\ell=2}^{p}\Gamma_{j\ell}(x)b_\ell(x), \sum_{\ell=2}^{p}\Gamma_{j\ell}^2(x)v_{\ell\ell}(x)\right\}, \quad j = 1, \ldots, p,$$

where h satisfies the conditions of Theorem 2(b)*, and $b_\ell(x)$ and $v_{\ell\ell}(x)$ are as defined in Theorem* 2.

(c1) *If* $\Gamma_i(x) = (\Gamma_{i1}(x), 0, \ldots, 0)^T$ *and* $\Gamma_j(x) = (\Gamma_{j1}(x), 0, \ldots, 0)^T$ *with* $\Gamma_{i1}(x) \neq 0$ *and* $\Gamma_{j1}(x) \neq 0$, *then the asymptotic covariance of* $\{(nh^d)^{1/2}\widehat{\gamma}_i(x),\ (nh^d)^{1/2}\widehat{\gamma}_j(x)\}$ *is* $\Gamma_{i1}(x)\Gamma_{j1}(x)v_{11}(x)$, $i, j = 1, \ldots, p$, *with h satisfying the conditions of Theorem* 2(a).

(c2) *If* $\Gamma_i(x) = (\Gamma_{i1}(x), \ldots, \Gamma_{ip}(x))^T$ *and* $\Gamma_j(x) = (\Gamma_{j1}(x), \ldots, \Gamma_{jp}(x))^T$, *with* $\Gamma_{i\ell}(x) \neq 0$, *for some* $\ell = 2, \ldots, p$, *and* $\Gamma_{j\ell}(x) \neq 0$, *for some* $\ell = 2, \ldots, p$, *then, the asymptotic covariance of* $\{(nh^{d+2})^{1/2}\widehat{\gamma}_i(x),\ (nh^{d+2})^{1/2}\widehat{\gamma}_j(x)\}$ *is* $\sum_{\ell=2}^{p} \Gamma_{i\ell}(x)\Gamma_{j\ell}(x)v_{\ell\ell}(x), \quad i, j = 1, \ldots, p.$

(c3) *If* $\Gamma_i(x) = (\Gamma_{i1}(x), 0, \ldots, 0)^T$, $\Gamma_{i1}(x) \neq 0$, *and* $\Gamma_j(x) = (\Gamma_{j1}(x), \ldots, \Gamma_{jp}(x))^T$ *with* $\Gamma_{j\ell}(x) \neq 0$, *for some* $\ell = 2, \ldots, p$, *then, the asymptotic covariance of* $\{(nh^d)^{1/2}\widehat{\gamma}_i(x),\ (nh^{d+2})^{1/2}\widehat{\gamma}_j(x)\}$, $i, j = 1, \ldots, p$, *is zero, where* h *satisfies the conditions of Theorem* 2(a) *for* $\widehat{\gamma}_i(x)$ *and Theorem* 2(b) *for* $\widehat{\gamma}_j(x)$.

(d1) *Let* $\widehat{V}(\widehat{\gamma})$ *be defined in identical way to* $\widehat{V}$ *after replacing* $\widehat{\alpha}(x)$ *by* $\widehat{\gamma}(x)$. *Then, for* $\Gamma_i(x)$, $\Gamma_j(x)$ and h *satisfying the conditions of part* (c1), *the* (i, j) *element of* $\widehat{V}(\widehat{\gamma})$ *is a consistent estimator of the asymptotic variance–covariance* $\Gamma_{i1}(x)\Gamma_{j1}(x)v_{11}(x), \quad i, j = 1, \ldots, p,$ *of part* (c1).

(d2) *For* $\Gamma_i(x)$, $\Gamma_j(x)$ and h *satisfying the conditions of part* (c2), *the* (i, j) *element of* $\widehat{V}(\widehat{\gamma})$ *is a consistent estimator of the asymptotic variance–covariance* $\sum_{\ell=2}^{p} \Gamma_{i\ell}(x)\Gamma_{j\ell}(x)v_{\ell\ell}(x)$, $i, j = 1, \ldots, p,$ *of part* (c2).

4 Empirical Illustration

We implemented our procedures on the transport choice dataset described in Horowitz (1993). There are 842 observations on transport mode choice for travel to work sampled randomly from the Washington D.C. area transportation study. The purpose of our exercise (and Horowitz's) is to model individuals choice of transportation method (DEP, which is 0 for transit and 1 for automobile) as it relates to the covariates: number of cars owned by the traveler's household (AUTOS), transit out-of-vehicle travel time minus automobile out-of-vehicle travel time in minutes (DOVTT), transit in-vehicle travel time minus automobile in-vehicle travel time in minutes (DIVTT), and transit fare minus automobile travel cost in 1968 cents (DCOST).

Horowitz compared several parametric and semiparametric procedures suggested by the following models:

H1: $\Pr(Y = 1|X = x) = \Phi(\beta^T x)$

H2: $\Pr(Y = 1|X = x) = F(\beta^T x)$

H3: $\Pr(Y=1|X=x)=\Phi\left(\frac{\beta^T x}{V(x)^{1/2}}\right)$

H4: $Y=1(\beta^T X+u>0)$,

where in the single index model (H2) the scalar function $F(\cdot)$ is of unknown form, while in the random coefficient probit model (H3) $V(x)=x'\Sigma x$ in which Σ is the covariance matrix of the random coefficients, and in (H4) the distribution $F_{u|X}(\cdot)$ of u given X is of unknown form with median$(u|X)=0$.[10] To estimate H2 he used the Klein and Spady (1993) procedure, while to estimate H4 he used both the Manski (1975) maximum score and the Horowitz (1992) smoothed maximum score methods. We fit a local probit to the dataset; thus our local model can be written as

$$E(Y|X=x)=\Phi(\gamma(x)^T x),$$

where $\gamma(\cdot)$ is of arbitrary unknown form. This can be interpreted as a random coefficient probit, except that the variation in parameters is driven by the conditioning information rather than by some arbitrary distribution unrelated to the covariates. Note that if H1 were true, then we should find that our estimated $\gamma's$ are constant with respect to x, while if either H2 or H3 were true, we should find that ratios of slope parameters are constant, i.e., $\gamma_j(x)/\gamma_\ell(x)$, $j,\ell=2,\ldots,p$, does not depend on x.[11]

For each point x, we found $\widehat{\gamma}(x)$, and hence $\widehat{g}(x)=\Phi(\widehat{\gamma}^T x)$, by minimizing

$$\sum_{i=1}^{n}\left\{Y_i-\Phi(\gamma^T X_i)\right\}^2 K_H(X_i-x)$$

with respect to γ. The parametric probit values provided starting points and a BHHH algorithm was used to locate the maximum. Convergence was typically rapid. The kernel was a product of

[10]For notational consistency with the rest of the paper we let the first element of x be a constant and call its parameter the intercept and the remaining parameters slopes.

[11]Note that in our model, $\gamma_j(x)/\gamma_\ell(x)=\alpha_j(x)/\alpha_\ell(x)$, for $j,\ell=2,\ldots,p$.

The comparison between H4 and our nonparametric regression is less clear as H4 is based on conditional median restrictions rather than conditional mean restrictions. If the data were generated by H4, then is hard to see what the regression function might be since the distribution of $(u|X)$ is arbitrary. What can be said here is that the H4 model provides interpretable parameters for this particular latent model, while the nonparametric regression formulation is better geared to evaluating effects in the observed data.

univariate Gaussian densities and the bandwidth was of the form $H = \widehat{h}\widehat{\Sigma}^{1/2}$, where $\widehat{\Sigma}$ was the sample covariance matrix of the regressors. The constant $\widehat{h}$ was chosen to minimize

$$CV(h) = \sum_{j=1}^{n} \{Y_j - \widehat{g}_{-j}(X_j)\}^2 \pi(X_j),$$

where $\pi(\cdot)$ is a trimming function and $\widehat{g}_{-j}$ is the leave-one-out estimate of g, see Härdle and Linton (1994). Figure 2 shows the cross-validation curve plotted against the logarithm of bandwidth for three different trimming rates. In each case, approximately the same bandwidth, $\widehat{h} = 0.917$, was chosen.[12] In Figure 3 we plot the fitted function against the fitted parametric regression $\Phi(\widetilde{\gamma}^T X_i)$. We present different curves for households with zero autos, one auto and two or more autos. The differences between these subsamples are quite pronounced. Figures 4 show the estimated local parameters (together with the corresponding parametric $\widetilde{\gamma}_j$ shown as the horizontal line) versus their own regressors along with a smooth of these points and 99% symmetric pointwise confidence intervals.

*** Figures 2-4 here ***

That these parameters are widely dispersed is consistent with Horowitz's findings against the fixed coefficient probit. There appears in some cases, notably in-vehicle time, to be a pronounced trend in the parameters.

Finally, we investigated the ratios $\widehat{\gamma}_j(x)/\widehat{\gamma}_\ell(x)$ and found many to exhibit non-constancy. Four of these plots are shown in Figures 5 where the dependence of the local slope parameter ratios on the regressors is clear.[13] This dependence suggests the presence of interactions (lack of separability) among the regressors in the utility function of these individuals, a feature not captured by the other models.

*** Figure 5 Here ***

5 Concluding Remarks

Our procedure provides a link between parametric and nonparametric methods. It allows one to shrink towards a favorite nonlinear shape, rather than towards only constants or polynomials as was previously the only available options. This is particularly relevant for binary data, where polynomials

[12] We chose the bandwidth from the least trimmed dataset because it is most representative of the data itself. In any event, the results are not greatly affected.

[13] As in Figures 4, the corresponding parametric probit ratio is shown as the horizontal line for comparison.

violate data restrictions, and in nonlinear time series estimation and prediction problems where parametric information is useful.

As with any smoothing procedure, an important practical issue that needs further attention is the bandwidth choice. In section 4 we used cross-validation. An alternative is to use the expression for the optimal bandwidth implied by the bias and variance expressions of Theorem 2 and its corollary to construct plug-in bandwidth selection methods. To implement these procedures, one must obtain estimates of g_{xx} using in place of m a model that includes m as a special case. Proper study of this issue goes beyond the scope of this paper. Interested readers can consult Fan and Gijbels (1992) for further insights into this problem.

A Appendix

PROOF OF THEOREM 1. We present the main argument for the case that $s = 0$. The proof follows the main steps of similar results in the literature (see for example Pötscher and Prucha (1991, Lemmas 3.1 and 4.2) or Andrews (1993, Lemma A1)). It relies on the use of the identification assumption A2 and a uniform strong law of large numbers (Lemma USLLN) which is proved in Appendix B. Lemma USLLN in particular guarantees that

$$\sup_{x\in\mathcal{X}_0,\alpha\in\mathcal{A}} |Q_n(x,\alpha) - \overline{Q}_n(x,\alpha)| \longrightarrow 0 \ a.s., \tag{14}$$

where

$$\begin{aligned}\overline{Q}_n(x,\alpha) &= E\left[\{Y - m(X,\alpha)\}^2 K_h(X-x)\right] \\ &= E\left[\{g(X) - m(X,\alpha)\}^2 K_h(X-x)\right] + E\left\{\sigma^2(X)K_h(X-x)\right\}.\end{aligned}$$

Now rewrite

$$\overline{Q}_n(x,\alpha) = \int \{g(x-vh) - m(x-vh,\alpha)\}^2 f_X(x-vh)K(v)dv + \int \sigma^2(x-vh)f_X(x-vh)K(v)dv$$

by a change of variables. Then, by dominated convergence and continuity (assumption A1),

$$\sup_{x\in\mathcal{X}_0,\alpha\in\mathcal{A}} |\overline{Q}_n(x,\alpha) - Q(x,\alpha)| \longrightarrow 0, \tag{15}$$

where convergence is uniform in α by virtue of the uniform continuity of m over $\mathcal{A}$.

Assumption A2 now guarantees that: for any δ-neighborhood $\Phi_0^\delta(x)$ of $\Phi_0(x)$, $\delta > 0$, there is an $\varepsilon > 0$, such that for any $A(x) \subseteq \mathcal{A}_x \setminus \Phi_0^\delta(x)$,

$$\inf_{x \in \mathcal{X}_0} \inf_{\alpha \in A(x)} Q(x, \alpha) - Q\{x, \alpha^0(x)\} \geq \varepsilon. \tag{16}$$

But this implies that there exist an η, $0 < \eta \leq \varepsilon$, such that

$$\Pr\left(\rho_H[\mathcal{M}_n(x), \Phi_0(x)] \geq \delta, \text{ for some } x \in \mathcal{X}_0\right) \leq$$

$$\Pr\left(Q(x, \mathcal{M}_n(x)) - Q(x, \Phi_0(x)) \geq \eta, \text{ for some } x \in \mathcal{X}_0\right) \longrightarrow 0 \ a.s., \tag{17}$$

where $Q(x, \mathcal{M}_n(x))$ denotes $Q(x, \widehat{\alpha}(x))$ for any $\widehat{\alpha}(x) \in \mathcal{M}_n(x)$ (similarly for $Q(x, \Phi_0(x))$), and "$\longrightarrow$ 0 $a.s.$" holds provided $\sup_{x \in \mathcal{X}_0} |Q(x, \mathcal{M}_n(x)) - Q(x, \Phi_0(x))| \longrightarrow 0$ $a.s.$ Using (14), (15) and (16), the latter follows from

$$\begin{aligned}
0 \ &\leq \ \inf_{x \in \mathcal{X}_0} [Q(x, \mathcal{M}_n(x)) - Q(x, \Phi_0(x))] \leq \sup_{x \in \mathcal{X}_0} [Q(x, \mathcal{M}_n(x)) - Q(x, \Phi_0(x))] \\
&\leq \ \sup_{x \in \mathcal{X}_0} [Q(x, \mathcal{M}_n(x)) - Q_n(x, \mathcal{M}_n(x))] + \sup_{x \in \mathcal{X}_0} [Q_n(x, \mathcal{M}_n(x)) - Q(x, \Phi_0(x))] \\
&\leq \ \sup_{x \in \mathcal{X}_0} [Q(x, \mathcal{M}_n(x)) - Q_n(x, \mathcal{M}_n(x))] + \sup_{x \in \mathcal{X}_0} [Q_n(x, \Phi_0(x)) - Q(x, \Phi_0(x))] \\
&\leq \ 2 \sup_{x \in \mathcal{X}_0, \alpha \in \mathcal{A}} |Q_n(x, \alpha) - \ Q(x, \alpha)| \longrightarrow 0 \ a.s.
\end{aligned}$$

Therefore, (17) follows. This, together with the uniform continuity of $m(\cdot, \alpha)$ and $Q_n(\cdot, \alpha)$ in α imply that for any $\widehat{\alpha}(x) \in \mathcal{M}_n$, $m\{x, \widehat{\alpha}(x)\} \longrightarrow g(x)$ $a.s.$, and $Q_n\{x, \widehat{\alpha}(x)\} \longrightarrow \sigma^2(x) f_X(x)$ $a.s.$

Assuming now a general $s \geq 0$, we modify the above argument as follows. First apply an s^{th} order Taylor theorem to $g(x - vh) - m(x - vh, \alpha)$ and write $\overline{Q}_n(x, \alpha)$ as

$$\begin{aligned}
\overline{Q}_n(x, \alpha) \ &= \ \int \left[\sum_{j=0}^{s} h^j \sum_{\{a : |a| = j\}} \frac{v^a}{a!} \{D^a g(x) - D_x^a m(x, \alpha)\} + h^s R_n(x, \alpha) \right]^2 f_X(x - vh) K(v) dv \\
&\quad + \int \sigma^2(x - vh) f_X(x - vh) K(v) dv,
\end{aligned}$$

where

$$R_n(x,\alpha) = \sum_{\{a:|a|=s\}} \frac{v^a}{a!}\left[\{D^a g(x^*(v,h)) - D_x^a m(x^*(v,h),\alpha)\} - \{D^a g(x) - D_x^a m(x,\alpha)\}\right],$$

where $x^*(v,h)$ are intermediate between x and $x - vh$. Note that $R_n(x,\alpha) = o(1)$ uniformly in x and α. Let $\overline{Q}_n^s(x,\alpha)$ be the same as $\overline{Q}_n(x,\alpha)$ but without $R_n(x,\alpha)$. Then,

$$\sup_{x\in\mathcal{X}_0,\alpha\in\mathcal{A}} \left|\overline{Q}_n(x,\alpha) - \overline{Q}_n^s(x,\alpha)\right| = o(h^s)$$

by our continuity and compactness assumptions. Furthermore, by our uniform convergence result, the rate in (14) is $\left(\log n/nh^d\right)^{1/2}$, which is smaller order than h^s under the stated bandwidth conditions. To minimize $\overline{Q}_n^s(x,\alpha)$ one proceeds recursively, first minimizing $C_0(x,\alpha)$ with respect to α, then $C_1(x,\alpha)$ with respect to $\alpha \in \Phi_0(x)$ etc. This is because the cross-product terms are zero and for each j, $f_X(x - vh)$ only enters proportionately and is independent of v to first order. Under our conditions in part (iv), there is a unique minimum $\alpha^0(x)$ to this higher order. In any case, condition A2 ensures that the set $\Phi_s(x)$ is identified in this generalized sense. ■

PROOF OF THEOREM 2. We deal with bias and variance terms separately. Let $H = \text{diag}(1,h,\ldots,h)$ be a $p\times p$, $p = d+1$, diagonal scaling matrix, and let e_1 and E_2 denote the following $p\times 1$ and $p\times d$ arrays $e_1 = (1,0_d^T)^T$, $E_2 = (0_d, I_d)^T$, where 0_d is a $d\times 1$ vector of zeros and I_d is the $d\times d$ identity matrix. Our assumptions guarantee the following

Given Assumptions A0-A1, A3, B1-B2, and our bandwidth conditions, $\widehat{\alpha}(x) \to \alpha^0(x)$ a.s., $\alpha^0(x)$ a unique interior point of $\mathcal{A}_x$, x interior to $\mathcal{X}$. Therefore, there exists a sequence $\{\overline{\alpha}_n(x)\}$ such that $\overline{\alpha}_n(x) = \widehat{\alpha}(x)$ a.s. for n sufficiently large, and each $\overline{\alpha}_n(x)$ takes its values in a convex compact neighborhood of $\alpha^0(x)$ interior to $\mathcal{A}_x$. In what follows we eliminate the dependence on x of the coefficients α to simplify notation.

Given Assumption B2, an element-by-element mean value expansion of $\partial Q_n(x,\overline{\alpha}_n)/\partial\alpha$ about α^0 gives (Jennrich (1969, Lemma 3)):

$$\frac{\partial Q_n}{\partial\alpha}(x,\overline{\alpha}_n) = \frac{\partial Q_n}{\partial\alpha}(x,\alpha^0) + \frac{\partial^2 Q_n}{\partial\alpha\partial\alpha^T}(x,\alpha_n^*)(\overline{\alpha}_n - \alpha^0),$$

where α_n^* is a random variable such that α_{nj}^* lies on the line segment joining $\overline{\alpha}_{nj}$ and α_j^0, $j = 1,\ldots,p$, and hence, $\alpha_n^* \to \alpha^0$, a.s. Since $\overline{\alpha}_n = \widehat{\alpha}$ a.s. for n sufficiently large, and since $\partial Q_n(x,\widehat{\alpha})/\partial x = 0$ whenever $\widehat{\alpha}$ is interior to $\mathcal{A}_x$, it follows that the left-hand side of the expansion vanishes a.s. for n sufficiently large. Premultiplying by $(nh^d)^{1/2}H^{-1}$, inserting the identity matrix $H^{-1}H$ and rearranging gives

$$
\begin{aligned}
(nh^d)^{1/2}H(\overline{\alpha}_n - \alpha^0) &= -\{H^{-1}\tfrac{\partial^2 Q_n}{\partial\alpha\partial\alpha^T}(x,\alpha_n^*)H^{-1}\}^{-1}(nh^d)^{1/2}H^{-1}\tfrac{\partial Q_n}{\partial\alpha}(x,\alpha^0) \\
&= A_n(x,\alpha_n^*)^{-1}\ (nh^d)^{1/2}S_n(x,\alpha^0),
\end{aligned} \tag{18}
$$

where $A_n(x,\alpha_n^*) = H^{-1}\partial^2 Q_n(x,\alpha_n^*)/\partial\alpha\partial\alpha^T H^{-1}$, and $S_n(x,\alpha^0) = -H^{-1}\partial Q_n(x,\alpha^0)/\partial\alpha$.

The two main steps of the proof consist in showing that **[1]** $A_n(x,\alpha^*)$ converges to a positive definite limit matrix, and **[2]** $(nh^d)^{1/2}S_n(x,\alpha^0)$ satisfies a multivariate central limit theorem.

PROOF OF **[1]** Write $A_n(x,\alpha_n^*)$ as

$$
A_n(x,\alpha_n^*) = H^{-1}R_{n1}(x,\alpha^0)H^{-1} + \{R_{n1}(x,\alpha_n^*) - R_{n1}(x,\alpha^0)\}H^{-1} + H^{-1}R_{n2}(x,\alpha_n^*)H^{-1},
$$

where $R_{n1}(x,\alpha) = 2n^{-1}\sum_{i=1}^n m_\alpha(X_i,\alpha)m_\alpha^T(X_i,\alpha)K_h(X_i - x)$, and $R_{n2}(x,\alpha) = 2n^{-1}\sum_{i=1}^n\{Y_i - m(X_i,\alpha)\}m_{\alpha\alpha}(X_i,\alpha)K_h(X_i - x)$.

We will show below that

$$
||H^{-1}R_{n1}(x,\alpha^0)H^{-1} - A_n(x,\alpha^0)|| = o_p(h), \tag{19}
$$

$$
||H^{-1}\{R_{n1}(x,\alpha_n^*) - R_{n1}(x,\alpha^0)\}H^{-1}|| = o_p(1), \text{ and} \tag{20}
$$

$$
||H^{-1}R_{n2}(x,\alpha_n^*)H^{-1}|| = o_p(1), \tag{21}
$$

where

$$
A_n(x,\alpha^0) = 2\begin{bmatrix} a_{11}(x,\alpha^0) & h\mathbf{a}_{12}(x,\alpha^0) \\ h\mathbf{a}_{21}(x,\alpha^0) & A_{22}(x,\alpha^0) \end{bmatrix}, \tag{22}
$$

with $a_{11}(x,\alpha^0) = f_X(x)$, $A_{22}(x,\alpha^0) = f_X(x)\mu_2(k)I_d$, where I_d is the $d\times d$ identity matrix, and $\mathbf{a}_{12}(x,\alpha^0) = \mathbf{a}_{21}^T(x,\alpha^0) = \mu_2(k)\{D^T f_X(x) + \frac{1}{2}f_X(x)[\mathrm{tr}\{D_2 m_{xx}(x,\alpha^0)\},\ldots,\mathrm{tr}\{D_p m_{xx}(x,\alpha^0)\}]\}$, where $Df_X(x) = \partial f_X(x)/\partial x$. Here, $D_\ell m_{xx}(x,\alpha)$, $\ell = 1,\ldots,p$, denotes a $d\times d$ matrix with (i,j) element $\partial^3 m(x,\alpha)/\partial\alpha_\ell\partial x_i\partial x_j$. Note that $A_n(x,\alpha^0)$ is a positive definite matrix as $n\to\infty$ given Assumptions A1 and B4.

PROOF OF **[2]** Write $S_n(x,\alpha^0)$ as

$$
S_n(x,\alpha^0) = S_{n1}(x,\alpha^0) + S_{n2}(x,\alpha^0) + S_{n3}(x,\alpha^0),
$$

where

$$S_{n1}(x,\alpha^0) = 2H^{-1}n^{-1}\sum_{i=1}^{n} u_i\{e_1 + E_2(X_i - x)\}K_h(X_i - x), \tag{23}$$

$$S_{n2}(x,\alpha^0) = 2H^{-1}n^{-1}\sum_{i=1}^{n} u_i\{m_\alpha(X_i,\alpha^0) - e_1 - E_2(X_i - x)\}K_h(X_i - x), \tag{24}$$

$$S_{n3}(x,\alpha^0) = 2H^{-1}n^{-1}\sum_{i=1}^{n}[g(X_i) - m(X_i,\alpha^0)]m_\alpha(X_i,\alpha^0)K_h(X_i - x). \tag{25}$$

We will show below that

$$(nh^d)^{1/2}S_{n1}(x,\alpha^0) \Rightarrow N(0,B), \tag{26}$$

where $B = 4\sigma^2(x)f_X(x)\mathrm{diag}(\nu_0(k),\nu_2(k),\ldots,\nu_2(k))$. Furthermore,

$$||(nh^d)^{1/2}S_{n2}(x,\alpha^0)|| = O_p(h), \text{ and} \tag{27}$$

$$||S_{n3}(x,\alpha^0) - h^2Hc(x,\alpha^0)|| = O_p(h^4), \tag{28}$$

where

$$c(x,\alpha^0) = (c_1(x,\alpha^0),\ldots,c_p(x,\alpha^0))^T, \tag{29}$$

with

$$c_1(x,\alpha^0) = \mu_2(k)f_X(x)\mathrm{tr}(\Delta_{xx}),$$

$$\begin{aligned} c_j(x,\alpha^0) = {} & \mu_4(k)\{D_jf_X(x)\Delta_{jj} + f_X(x)[\frac{1}{2}\sum_{r=1}^{d} m_{jrr}\Delta_{rr} + \frac{1}{3}\sum_{r=1}^{d}\Delta_{jjr}]\} \\ & + \mu_2^2(k)\{D_jf_X(x)\sum_{r\neq j}\Delta_{rr} + \sum_{s\neq j} D_sf_X(x)[\Delta_{js} + \Delta_{sj}] + f_X(x)[\frac{1}{2}(\sum\sum_{r\neq s} m_{jrr}\Delta_{ss} \\ & + \sum\sum_{r\neq s} m_{jrs}\Delta_{rs} + \sum\sum_{r\neq s} m_{jrs}\Delta_{sr}) + \frac{1}{3}\sum_{r\neq j}\sum_{s=1}^{d}\Delta_{jrs}]\}, \end{aligned}$$

for $j = 2,\ldots,p$, where $D_jf_X(x) = \partial f_X(x)/\partial x_j$, $\Delta_{ij} = \partial^2[g(x) - m(x,\alpha^0)]/\partial x_i\partial x_j$ is the (i,j) element of Δ_{xx}, $\sum_{r\neq j}$ denotes $\sum_{r=1,r\neq j}^{d}$, $\sum\sum_{r\neq s}$ denotes $\sum_{s=1}^{d}\sum_{r=1,r\neq s}^{d}$, $m_{ijr} = \partial^3 m(x,\alpha^0)/\partial x_i\partial x_j\partial x_r$, and $\Delta_{ijr} = \partial^3[g(x) - m(x,\alpha^0)]/\partial x_i\partial x_j\partial x_r$.

Using these results, we can now calculate the bias and variance of the regression and first derivative estimates :

[**BIAS**]: From (18)-(29), and the equivalence between $\widehat{\alpha}$ and $\overline{\alpha}_n$ for n sufficiently large, it follows that the (asymptotic)bias is

$$\begin{aligned} E(\widehat{\alpha}-\alpha^0) &= H^{-1}A_n(x,\alpha^0)^{-1}h^2Hc(x,\alpha^0)\} + O(h^4), \\ &= h^2\frac{1}{2}\begin{bmatrix} f_X^{-1}(x) & -h\mu_2^{-1}(k)f_X^{-2}(x)\mathbf{a}_{12} \\ -\mu_2^{-1}(k)f_X^{-2}(x)\mathbf{a}_{12}^T & h^{-1}\mu_2^{-1}(k)f_X^{-1}(x)I_d \end{bmatrix}\begin{bmatrix} c_1(x,\alpha^0) \\ hc_2(x,\alpha^0) \\ hc_p(x,\alpha^0) \end{bmatrix} + O(h^4) \end{aligned}$$

where the last equality follows from the inversion of the block matrix $A_n(x,\alpha^0)$. Therefore, the bias of $\widehat{\alpha}$ is given by $h^2b(x)+O(h^4) = h^2(b_1(x),\ldots,b_p(x))^T + O(h^4)$, where

$$b_1(x) = \frac{1}{2}\mu_2(k)\mathrm{tr}(\Delta_{xx}) \tag{30}$$

is the constant of the regression estimate bias at x, while the constant of the bias $E(\widehat{\alpha}_j-\alpha_j^0)$ at x of the first derivative estimate with respect to $x_{j-1},\ j=2,\ldots,p$, is given by

$$\begin{aligned} b_j(x) &= \frac{1}{2}\left\{\frac{c_j(x,\alpha^0)}{\mu_2(k)f_X(x)} - \frac{\mu_2(k)\mathrm{tr}(\Delta_{xx})}{f_X(x)}[D_jf_X(x)+\frac{1}{2}f_X(x)\mathrm{tr}\{D_jm_{xx}(x,\alpha^0)\}]\right\} \\ &= \frac{1}{2}\left\{\left(\frac{\mu_4(k)-\mu_2^2(k)}{\mu_2(k)}\right)\left\{\frac{D_jf_X(x)}{f_X(x)}\Delta_{jj}+\frac{1}{2}\sum_{r=1}^{d}m_{jrr}\Delta_{rr}\right\}+\frac{1}{3}\frac{\mu_4(k)}{\mu_2(k)}\sum_{r=1}^{d}\Delta_{jjr}\right. \\ &\quad +\mu_2(k)[\sum_{s\neq j}\frac{D_sf_X(x)}{f_X(x)}(\Delta_{js}+\Delta_{sj})+\frac{1}{2}(\sum\sum_{r\neq s}m_{jrs}\Delta_{rs}+\sum\sum_{r\neq s}m_{jrs}\Delta_{sr}) \\ &\quad \left.+\frac{1}{3}\sum_{r\neq j}\sum_{s=1}^{d}\Delta_{jrs}]\right\}. \end{aligned} \tag{31}$$

[**VARIANCE**]: From (18)-(29), we have (asymptotically)

$$\mathrm{Var}\left\{(nh^d)^{1/2}H(\overline{\alpha}_n-\alpha^0)\right\} = A_n(x,\alpha^0)^{-1}BA_n(x,\alpha^0)^{-1} = V(x),$$

where

$$V(x) = [\sigma^2(x)/f_X(x)]\mathrm{diag}(\nu_0(k),\nu_2(k)/\mu_2^2(k),\ldots,\nu_2(k)/\mu_2^2(k)) + O(h). \tag{32}$$

It now follows easily from the above results, and the equivalence between $\widehat{\alpha}$ and $\overline{\alpha}_n$ for n sufficiently large, that

$$(nh^d)^{1/2}H(\widehat{\alpha}-\alpha^0-h^2b(x)) \Rightarrow N\left\{0,V(x)\right\}.$$

The scaling $(nh^d)^{1/2}H$ shows the different rate of convergence for the regression estimator, $(nh^d)^{1/2}$, and for the partial derivatives, $(nh^{d+2})^{1/2}$. This implies we must choose a different rate for h, see

parts (a) and (b) of the theorem, depending on whether we are interested in estimating the regression function or the partial derivatives. Note that the fastest rate of convergence in distribution in parts (a) and (b) of the theorem is given by $0 < c < \infty$ (i.e., no undersmoothing.)

It remains to show (19)-(21) and (26)-(28):

PROOF OF [**(19)**]: An element-by-element second order mean value expansion of $m_\alpha(X_i, \alpha^0)$ about x gives:

$$\begin{aligned} m_\alpha(X_i, \alpha^0) &= e_1 + E_2(X_i - x) + \tfrac{1}{2}(I_p \otimes (X_i - x)^T) m_{\alpha xx}(X_i^*, \alpha^0)(X_i - x) \\ &= \begin{bmatrix} 1 \\ (X_i - x) \end{bmatrix} + \tfrac{1}{2}(I_p \otimes (X_i - x)^T) m_{\alpha xx}(X_i^*, \alpha^0)(X_i - x) \\ &= a_i + b_i, \end{aligned} \tag{33}$$

where the first equality follows from the reparameterization chosen, and $m_{\alpha xx}(x, \alpha)$ is a $pd \times d$ matrix consisting of p $d \times d$ blocks $D_\ell m_{xx}(x, \alpha)$, $\ell = 1, \ldots, p$, with $D_\ell m_{xx}(x, \alpha)$ as defined after equation (22). Thus,

$$\begin{aligned} R_{n1}(x, \alpha^0) &= 2n^{-1} \textstyle\sum_{i=1}^n (a_i + b_i)(a_i + b_i)^T K_h(X_i - x) \\ &= 2n^{-1} \textstyle\sum_{i=1}^n [a_i a_i^T + a_i b_i^T + b_i a_i^T + b_i b_i^T] K_h(X_i - x). \end{aligned} \tag{34}$$

Now,

$$\begin{aligned} & n^{-1} \sum_{i=1}^n a_i a_i^T K_h(X_i - x) \\ &= n^{-1} \sum_{i=1}^n \begin{bmatrix} 1 & (X_i - x)^T \\ (X_i - x) & (X_i - x)(X_i - x)^T \end{bmatrix} K_h(X_i - x) \\ &= \begin{bmatrix} f_X(x) & h^2 \mu_2(k) D^T f_X(x) \\ h^2 \mu_2(k) D f_X(x) & h^2 \mu_2(k) f_X(x) I_d \end{bmatrix} + \begin{bmatrix} O_p(h^2) & o_p(h^2) \\ o_p(h^2) & o_p(h^2) \end{bmatrix}, \end{aligned} \tag{35}$$

by standard results from kernel density estimation, see Wand and Jones (1995, Chapter 4).

Similar calculations with the other terms in (34) yield

$$n^{-1} \sum_{i=1}^n a_i b_i^T K_h(X_i - x)$$

$$
\begin{aligned}
&= n^{-1}\sum_{i=1}^{n} b_i a_i^T K_h(X_i - x) \\
&= \begin{bmatrix} \frac{1}{2}h^2\mu_2(k) f_X(x)\mathrm{tr}\{D_1 m_{xx}(x,\alpha^0)\} & \ldots & \frac{1}{2}h^2\mu_2(k) f_X(x)\mathrm{tr}\{D_p m_{xx}(x,\alpha^0)\} \\ O_p(h^4) & & O_p(h^4) \end{bmatrix}
\end{aligned} \tag{36}
$$

and

$$
n^{-1}\sum_{i=1}^{n} b_i b_i^T K_h(X_i - x) = O_p(h^4). \tag{37}
$$

Equation (19) now follows easily from (34)–(37).

PROOF OF [**(20)**]: Equation (20) follows directly by dominated convergence given $\alpha_n^* \to \alpha^0$ a.s. and the boundedness of $m_\alpha(x,\alpha)$ uniformly in α implied by Assumption A1.

PROOF OF [**(21)**]: Write

$$
\begin{aligned}
R_{n2}(x,\alpha) &= 2n^{-1}\sum_{i=1}^{n} u_i m_{\alpha\alpha}(X_i,\alpha) K_h(X_i - x) \\
&\quad + 2n^{-1}\sum_{i=1}^{n} [g(X_i) - m(X_i,\alpha)] m_{\alpha\alpha}(X_i,\alpha) K_h(X_i - x) \\
&= T_{n1}(x,\alpha) + T_{n2}(x,\alpha).
\end{aligned}
$$

$$
E[T_{n1}(x,\alpha)] = 2\int E(u|X) m_{\alpha\alpha}(X,\alpha) K_h(X - x) f_X(X) dX = 0,
$$

by (1), while

$$
\begin{aligned}
E[T_{n2}(x,\alpha)] &= 2\textstyle\int [g(X) - m(X,\alpha)] m_{\alpha\alpha}(X,\alpha) K_h(X - x) f_X(X) dX \\
&\to 2[g(x) - m(x,\alpha)] m_{\alpha\alpha}(X,\alpha) f_X(x)
\end{aligned}
$$

uniformly in α by dominated convergence and continuity (Assumptions A1 and B2-B3). In particular, $||ET_{n2}(x,\alpha_n)|| \to 0$ for any $\alpha_n \to \alpha^0$ since $g(x) - m(x,\alpha^0) = 0$. Then, Assumptions A0-A1, A3, B1-B4, and our bandwidth conditions provide enough regularity conditions to apply Lemma USLLN (with $q_n(Z,\theta)$ replaced by $u m_{\alpha\alpha}(X,\alpha) K\{h^{-1}(X - x)\}$ and $[g(X) - m(X,\alpha)] m_{\alpha\alpha}(X,\alpha) K\{h^{-1}(X - x)\}$, respectively, with $\mathcal{X}_0 = \{x\}$) to show that

$$
\sup_{\rho\{\alpha,\Phi_0(x)\}<\delta} ||T_{n1}(x,\alpha)|| \to 0 \text{ a.s., and}
$$

$$
\sup_{\rho\{\alpha,\Phi_0(x)\}<\delta} ||T_{n2}(x,\alpha) - E[T_{n2}(x,\alpha)]|| \to 0 \text{ a.s.}
$$

This implies (21).

PROOF OF [**(26)**]: It is clear from (23) and (1) that $S_{n1}(x,\alpha^0)$ is a sum of independent random variables with mean zero and

$$\begin{aligned}\text{Var}\left\{(nh^d)^{1/2}S_{n1}\right\} &= 4h^d\int\sigma^2(X)\begin{bmatrix}1\\(\frac{X-x}{h})\end{bmatrix}[1\ \ (\frac{X-x}{h})^T]K_h^2(X-x)f_X(X)dX\\ &= 4\int\sigma^2(x+hv)f_X(x+hv)\begin{bmatrix}1 & v^T\\ v & vv^T\end{bmatrix}K^2(v)dv\\ &= B+O(h).\end{aligned}$$

Finally, S_{n1} obeys the Lindeberg-Feller central limit theorem, see Lemma CLT below.

PROOF OF [**(27)**]: From (24) and (33)

$$S_{n2}(x,\alpha^0)=2H^{-1}n^{-1}\sum_{i=1}^n u_i\{\frac{1}{2}(I_p\otimes(X_i-x)^T)m_{\alpha xx}(X_i^*,\alpha^0)(X_i-x)\}K_h(X_i-x).$$

Thus, $E\{S_{n2}(x,\alpha^0)\}=0$ by (1), while

$$\begin{aligned}&\text{Var}\left\{(nh^d)^{1/2}S_{n2}(x,\alpha^0)\right\}\\ &= h^4H^{-1}\{\int\sigma^2(x+hv)f_X(x+hv)\\ &\qquad\times[(I_p\otimes v^T)m_{xxx}(x+hv^*)vv^Tm_{xxx}^T(x+hv^*)(I_p\otimes v)]K(v)dv\}H^{-1}\\ &= O(h^2).\end{aligned}$$

PROOF OF [**(28)**]: Doing a third order Taylor expansion of $g(X_i)-m(X_i,\alpha^0)$ about x, and recalling that $g(x)=m(x,\alpha^0)$ and $g_x(x)=m_x(x,\alpha^0)$ thanks to the reparameterization, we get

$$\begin{aligned}g(X_i)-m(X_i,\alpha^0) &= \tfrac{1}{2}(X_i-x)^T\Delta_{xx}(X_i-x)\\ &\quad+\tfrac{1}{6}((X_i-x)^T\otimes(X_i-x)^T)\Delta_{xxx}(X_i-x)+r(X_i,\alpha^0)\\ &= d_{2i}+d_{3i}+r(X_i,\alpha^0),\end{aligned}$$

where $r(X_i,\alpha^0)$ denotes the remainder term, and Δ_{xxx} is the $d^2\times d$ matrix of third order partial derivatives with respect to x. Using this expansion and (25) and (33) we can write

$$S_{n3}(x,\alpha^0)=2H^{-1}n^{-1}\sum_{i=1}^n(d_{2i}+d_{3i})(a_i+b_i)K_h(X_i-x). \tag{38}$$

We will show that $E\{S_{n3}(x,\alpha^0)\} = h^2 Hc(x,\alpha^0) + O(h^4)$, while $\mathrm{Var}\{S_{n3}(x,\alpha^0)\} = O(h^8)$.

An analysis of (38) similar to that carried out in (34)-(37) yields

$$\begin{aligned}
&2H^{-1}\sum_{i=1}^{n} d_{2i}a_i K_h(X_i - x) \\
&= H^{-1}n^{-1}\sum_{i=1}^{n}(X_i - x)^T \Delta_{xx}(X_i - x)\begin{bmatrix} 1 \\ (X_i - x) \end{bmatrix} K_h(X_i - x) \\
&= H^{-1}\begin{bmatrix} h^2 \int v^T \Delta_{xx} v K(v) f_X(x + hv)dv + o_p(h^2) \\ h^4 \int v^T \Delta_{xx} vvv^T D f_X(x + hv)K(v)dv + o_p(h^4) \end{bmatrix} \\
&= h^2 H\mathbf{w} + o_p(h^2),
\end{aligned}$$

say, where $\mathbf{w} = (w_1, \ldots, w_p)^T$ is a $p\times 1$ vector with $w_1 = \mu_2(k) f_X(x)\mathrm{tr}(\Delta_{xx})$, and $w_j = \mu_4(k) D_j f_X(x)\Delta_{jj} + \mu_2^2(k)\{D_j f_X(x)\sum_{r\neq j}\Delta_{rr} + \sum_{s\neq j} D_s f_X(x)[\Delta_{js} + \Delta_{sj}]\}$, $j = 2, \ldots, p$. Similarly,

$$\begin{aligned}
&2H^{-1}n^{-1}\sum_{i=1}^{n} d_{2i}b_i K_h(X_i - x) \\
&= \frac{1}{2}H^{-1}n^{-1}\sum_{i=1}^{n}(X_i - x)^T \Delta_{xx}(X_i - x)(I_p \otimes (X_i - x)^T) m_{\alpha xx}(X_i^*, \alpha^0)(X_i - x)K_h(X_i - x) \\
&= \frac{1}{2}H^{-1}\begin{bmatrix} h^4 \int v^T \Delta_{xx} v(I_p \otimes v^T) & D_1 m_{xx}(x + hv^*, \alpha^0) & vK(v)f(x + hv)dv \\ & \vdots & \\ h^4 \int v^T \Delta_{xx} v(I_p \otimes v^T) & D_p m_{xx}(x + hv^*, \alpha^0) & vK(v)f(x + hv)dv \end{bmatrix} \\
&= h^2 \mathrm{diag}(h^2, h, \ldots, h)\mathbf{z} + o_p(h^3),
\end{aligned}$$

say, where $\mathbf{z}$ is a $p \times 1$ vector with

$$\begin{aligned}
z_j &= \frac{1}{2}\sum_{r=1}^{d}\sum_{t=1}^{d}\sum_{\ell=1}^{d}\sum_{s=1}^{d} f_X(x) m_{js\ell}\Delta_{rt} v_r v_t v_\ell v_s K(v)dv \\
&= \frac{1}{2}f_X(x)\{\mu_4(k)\sum_{r=1}^{d} m_{jrr}\Delta_{rr} + \mu_2^2(k)[\sum\sum_{r\neq s} m_{jrr}\Delta_{ss} + \sum\sum_{r\neq s} m_{jrs}\Delta_{rs} + \sum\sum_{r\neq s} m_{jrs}\Delta_{sr}]\}.
\end{aligned}$$

Furthermore,

$$\begin{aligned}
&2H^{-1}n^{-1}\sum_{i=1}^{n} d_{3i}a_i K_h(X_i - x) \\
&= \frac{1}{3}H^{-1}n^{-1}\sum_{i=1}^{n}((X_i - x)^T \otimes (X_i - x)^T \Delta_{xxx}(X_i - x)\begin{bmatrix} 1 \\ (X_i - x) \end{bmatrix} K_h(X_i - x)
\end{aligned}$$

$$
\begin{aligned}
&= \frac{1}{3}H^{-1}\begin{bmatrix} h^4\int(v^T\otimes v^T)\Delta_{xxx}vv^TDf(x+hv)K(v)dv \\ h^4\int(v^T\otimes v^T)\Delta_{xxx}vvf_X(x+hv)K(v)dv \end{bmatrix} \\
&= \frac{1}{3}h^2\mathrm{diag}(h^2,h,\ldots,h)\mathbf{q}+o_p(h^3),
\end{aligned}
$$

say, where $\mathbf{q}$ is a $p\times 1$ vector with $q_j = \frac{1}{3}f_X(x)\{\mu_4(k)\sum_{r=1}^d \Delta_{jjr}+\mu_2^2(k)\sum_{r\neq j}\sum_{s=1}^d \Delta_{jrs}\}$, $j=2,\ldots,p$ (the first element q_1 is not given in detail for being of smaller order of magnitude).

Similar analysis shows that $n^{-1}\sum_{i=1}^n d_{3i}b_iK_h(X_i-x) = O_p(h^6)$, and hence can be ignored. Thus, we have

$$
E\{S_{n3}(x,\alpha^0)\} = h^2H\mathbf{w}+h^2\mathrm{diag}(h^2,h,\ldots,h)\{\mathbf{z}+\mathbf{q}\}+O(h^4) = h^2Hc(x,\alpha^0)+O(h^4).
$$

Furthermore, $S_{n3}(x,\alpha^0)-E\{S_{n3}(x,\alpha^0)\} = O_p(h^4)$, hence

$$
\mathrm{Var}\{S_{n3}(x,\alpha^0)\} = O(h^8),
$$

which implies (28). ■

LEMMA CLT. *The standardized sum S_{n1} obeys the Lindeberg-Feller central limit theorem, specifically*

$$
(nh^d)^{1/2}S_{n1} \Rightarrow N(0,B).
$$

PROOF. For any $p\times 1$ vector c, let $c^TS_{n1} = \sum_{i=1}^n t_{ni}$, where $t_{ni} = 2n^{-1}c^TH^{-1}\{e_1+E_2(X_i-x)\}K_h(X_i-x)u_i$. Note that for each n, t_{ni} are i.i.d for $i=1,\ldots,n$, with common variance

$$
s_n^2 = 4n^{-1}h^{-d}c^T\left\{\int(e_1+E_2v)(e_1+E_2v)^TK^2(v)\sigma^2(x-vh)f_X(x-vh)dv\right\}c.
$$

Furthermore, $|t_{n1}|\leq 2n^{-1}h^{-d}|\tau|$, where $\tau = c^T\{e_1\sup K(v)+\sup\left(|v_1|,\ldots,|v_d|\right)^TK(v)\}u$ is a random variable not depending on n. Therefore, for any $\varepsilon>0$,

$$
s_n^{-2}\sum_{i=1}^n E[t_{ni}^2\mathbb{I}_{\{|t_{ni}|>\varepsilon s_n\}}] \leq s_0^{-2}E[h^{-d}\Theta^T\Theta u^2\ \mathbb{I}_{\left\{|\tau|>\varepsilon' n^{1/2}h^{d/2}\right\}}],
$$

for some $\varepsilon'>0$, where $\Theta^T = c^TH^{-1}\{e_1+E_2(X-x)\}K\{h^{-1}(X-x)\}$ and $s_0^2 = nh^ds_n^2/4 = O(1)$. Here, $\mathbb{I}_{\{A\}}$ denotes the indicator function of the event A.

Since

$$
E[\Theta^T\Theta u^2h^{-d}\mathbb{I}_{\left\{|\tau|>\varepsilon' n^{1/2}h^{d/2}\right\}}] \leq E[\Theta^T\Theta u^2h^{-d}] < \infty,
$$

we can apply Fubini's theorem to show that

$$E[\Theta^T\Theta u^2 h^{-d}\mathbb{I}_{\left\{|\tau|>\varepsilon' n^{1/2}h^{d/2}\right\}}] = \int \Theta^T\Theta h^{-d} J_n(X) dP_X,$$

where for some ε'',

$$J_n(X) = E\left[u^2\mathbb{I}_{\left\{|u|>\varepsilon'' n^{1/2}h^{d/2}\right\}}|X\right] \to 0 \quad a.s. \quad X$$

by dominated convergence. Therefore, the Lindeberg condition is satisfied. Since c was arbitrary, the asymptotic normality of S_{n1} follows by the Cramér-Wold device. ■

B Appendix

Here, we use linear functional notation and write $P\zeta = \int \zeta dP$ for any probability measure P and random variable $\zeta(Z)$. We use P_n to denote the empirical probability measure of the observations $\{Z_1, \ldots, Z_n\}$ sampled randomly from the distribution P, in which case $P_n\zeta = n^{-1}\sum_{i=1}^n \zeta(Z_i)$. Similarly, let P_X be the marginal distribution of X under P, and let P_{Xn} be the corresponding empirical measure. For a class of functions $\mathcal{F}$, the *envelope* F is defined as $F = \sup_{f\in\mathcal{F}}|f|$. For $1 \leq s < \infty$, and G some probability measure on $\mathbb{R}^q$, we denote by $L^s(G)$ the space of measurable real functions on $\mathbb{R}^q$ with $(\int |f|^s dG)^{1/s} < \infty$. In what follows, G will usually be the population measure P or the empirical measure P_n. Moreover, for $\mathcal{F} \subset L^s(G)$, we define the *covering number* $N_s(\epsilon, G, \mathcal{F})$ as the smallest value of N for which there exist functions $g_1, \ldots, g_N$ (not necessarily in $\mathcal{F}$) such that $\min_{j\leq N}(G(f-g_j)^s)^{1/s} \leq \epsilon$ for each f in $\mathcal{F}$. Furthermore, the ϵ-*entropy* of $\mathcal{F}$ with respect to the $L^s(G)$ metric is defined as $\log N_s(\epsilon, G, \mathcal{F})$. Finally, the notation $x_n \ll y_n$ means $x_n/y_n \to 0$ as $n \to \infty$. By virtue of our assumptions we have the following facts: $\sup_{\alpha\in\mathcal{A}}\sup_{z\in\mathcal{X}_0^\delta} C_j(z,\alpha) < \infty$ for all relevant j, $\sup_{z\in\mathcal{X}_0^\delta}\sigma^2(z) \leq \overline{\sigma}^2 < \infty$, and $\sup_{z\in\mathcal{X}_0^\delta} f_X(z) \leq \overline{f} < \infty$, for some $\delta > 0$.

Jennrich (1969) proves consistency of the parametric nonlinear least squares estimator for the parametric regression functions $m(\cdot,\theta)$: $\theta \in \Theta$, under the assumptions that Θ is compact, and that the envelope condition $\int \sup_{\theta\in\Theta}|m(x,\theta)|^2 dP_X(x) < \infty$ holds. For the uniform consistency of our estimator we require a similar envelope condition, and we need to show that the covering numbers do not grow exponentially fast, namely that $n^{-1}\log N_1(\epsilon, G, \mathcal{F}) \xrightarrow{P} 0$, where $\xrightarrow{P}$ denotes

convergence in outer probability.[14] This entropy condition is satisfied by many classes of functions. For example, if the functions in $\mathcal{F}$ form a finite-dimensional vector space, then $\mathcal{F}$ satisfies the entropy condition (see Pollard (1984, Lemmas II.28 and II.25). In the proof of our next lemma we will use the entropy lemma of Pakes and Pollard (1989, Lemma 2.13) which establishes the equivalence between the entropy condition $N_1(\epsilon GF, G, \mathcal{F}) \leq A\epsilon^{-W}$ of a class of functions $\mathcal{F}$ indexed by a parameter satisfying a Lipschitz condition on that parameter (Assumption A1), and the compactness of the parameter space.

LEMMA (ENTROPY). *Let G_X denote an arbitrary probability measure on $\mathcal{X}$, and $\mathcal{F} = \{f(\cdot, \theta) : \theta \in \Theta\}$ be a class of real-valued functions on $\mathcal{X}$ indexed by a bounded subset Θ of $\mathbb{R}^p$. Suppose that $f(\cdot, \theta)$ is Lipschitz in θ, that is, there exists a $\lambda > 0$ and non-negative bounded function $\phi(\cdot)$, such that*

$$|f(x, \theta) - f(x, \theta^*)| \leq \phi(x) \|\theta - \theta^*\|^\lambda \text{ for all } x \in \mathcal{X} \text{ and } \theta, \theta^* \in \Theta.$$

Then, for the envelope $F(\cdot) = |f(\cdot, \theta_0)| + M\phi(\cdot)$, where $M = (2\sqrt{p} \sup_{\theta \in \Theta} \|\theta - \theta_0\|)^\lambda$ with θ_0 an arbitrary point of Θ, and for any $0 < \epsilon \leq 1$, we have

$$N_1(\epsilon GF, G, \mathcal{F}) \leq A\epsilon^{-W} \ ,$$

where A and W are positive constants not depending on n.

PROOF. Pakes and Pollard (1989, Lemma 2.13).

The class of functions $\mathcal{F}$ satisfying Assumption A1 is what Andrews (1994) calls a *type II class*. Classes of functions satisfying $N_1(\epsilon GF, G, \mathcal{F}) \leq A\epsilon^{-W}$ are said to be *Euclidean classes* (c.f. Nolan and Pollard (1987, p.789)).

We will also use Pollard (1984, Theorem II.37) (with his δ_n^2 replaced with h^d) in order to derive the rate of the uniform convergence of our estimator.

[14] Given an underlying probability space $(\Omega, \mathcal{G}, \mathcal{P})$, the outer probability for $A \subset \Omega$ is defined as $\mathcal{P}^*(A) = \inf\{\mathcal{P}(B) : A \subset B, B \in \mathcal{G}\}$. The reason for needing outer probability is due to the fact that the random covering numbers need not be measurable with respect to $\mathcal{P}$, even if the class of functions is permissible in the sense of Pollard (1984).

LEMMA (POLLARD). *For each n, let $\mathcal{F}_n$ be a permissible class of functions whose covering numbers satisfy $N_1(\epsilon, G, \mathcal{F}) \leq A\epsilon^{-W}$ for $0 < \epsilon \leq 1$, where G is an arbitrary probability measure, and A and W are positive constants not depending on n. If $nh^d\alpha_n^2 \gg \log n$, $|f_n| \leq 1$, and $(Pf_n^2)^{1/2} \leq h^{d/2}$ for each f_n in $\mathcal{F}_n$, then*

$$\sup_{f_n \in \mathcal{F}_n} |P_n f_n - P f_n| \ll h^d \alpha_n \quad a.s.$$

PROOF. Pollard (1984, Theorem II.37) with his δ_n^2 replaced with h^d. ■

The following lemma provides a uniform strong law of large numbers for our criterion function which is needed in the consistency proof of the estimator.

LEMMA (USLLN). *Let $\theta = (x, \alpha)$ be an element of $\Theta = \mathcal{X}_0 \times \mathcal{A}$, where $\mathcal{X}_0$ is a bounded subset of $\mathbb{R}^d$. Let $\mathcal{Q}_n = \{q_n(\cdot, \theta) : \theta \in \Theta\}$ be a class of functions with $q_n(Z, \theta) = \{Y - m(X, \alpha)\}^2 K\{h^{-1}(X - x)\}$. Moreover, let $Q_n(\theta) = h^{-d} P_n q_n(Z, \theta)$ and $\overline{Q}_n(\theta) = h^{-d} P q_n(Z, \theta)$. Under assumptions A0-A4,*

$$\sup_{\theta \in \Theta} |Q_n(\theta) - \overline{Q}_n(\theta)| = O_p\left(\left\{\frac{\log n}{nh^d}\right\}^{1/2}\right) \quad a.s. \tag{39}$$

PROOF. Without loss of generality we replace $q_n(\cdot, \theta)$ by

$$q_n(Z, \theta) = \left[\{g(X) - m(X, \alpha)\}^2 + u^2\right] K\{h^{-1}(X - x)\},$$

where equation (1) was used to eliminate the cross product term. Observe that the functions $q_n(\cdot, \theta)$ depend on n through h , and that they are not necessarily uniformly bounded.

In order to facilitate comparison with similar arguments in the literature, we break the proof into several steps.

PERMISSIBILITY. The boundedness of the index set $\Theta = \mathcal{X}_0 \times \mathcal{A}$ and the measurability of the relevant expressions are sufficient conditions to guarantee that the class of functions $\mathcal{Q}_n = \{q_n(\cdot, \theta) : \theta \in \Theta\}$ is *permissible* in the sense of Pollard (1984, Appendix C) for each n, which suffices to ensure that $\mathcal{Q}_n$ is permissible. Permissibility imposes enough regularity to ensure measurability of

the supremum (and other functions needed in the uniform consistency proof) when taken over an uncountable class of measurable functions.

ENVELOPE INTEGRABILITY. Let $\overline{q}_n = \{\sup_{\alpha\in\mathcal{A}} |g(X) - m(X,\alpha)|^2 + |u|^2\}\, |K\{h^{-1}(X-x)\}|$ be the envelope of the class of functions $\mathcal{Q}_n$. Assumptions A1 and A3 are sufficient to guarantee that

$$P\overline{q}_n < \infty. \tag{40}$$

To see this note that

$$\begin{aligned}
P\overline{q}_n &< P_X R(X)|K\{h^{-1}(X-x)\}| + Pu^2|K\{h^{-1}(X-x)\}| \\
&= h^d \int R(x+hv) f_X(x+hv)|K(v)|dv + h^d \int \sigma^2(x-vh) f_X(x+hv)|K(v)|dv \\
&= h^d R(x) f_X(x) \int |K(v)|dv + h^d\{\sigma^2(x) f_X(x) \int |K(v)|dv + o(1)\} \\
&\leq h^d \hat{K}^d \overline{f}(R+\overline{\sigma}^2)\{1+o(1)\} = O(h^d),
\end{aligned}$$

where the first line follows by assumption A1, and the second line follows by a change of variables and noting that $Pu^2|K\{h^{-1}(X-x)\}| = P_X\sigma^2(X)|K\{h^{-1}(X-x)\}|$ by iterated expectations. The third line follows by dominated convergence and A4 ($h \to 0$), and the last line by assumptions A1 and A3. This establishes (40).

Given the permissibility and envelope integrability of $\mathcal{Q}_n$, a.s. convergence to zero of the random variable $\sup_{\mathcal{Q}_n} |Q_n(\theta) - \overline{Q}_n(\theta)|$ is equivalent to their convergence in probability to zero. See Giné and Zinn (1984, Remark 8.2(1)) or Pollard (1984, Proof of Theorem II.24) and references therein.

TRUNCATION. The envelope integrability (40) allows us to truncate the functions to a finite range. Let β_n be a sequence of constants such that $\beta_n \geq 1$, $\beta_n \to \infty$. Note that

$$\begin{aligned}
\sup_{\mathcal{Q}_n} |P_n q_n - P q_n| \leq \sup_{\mathcal{Q}_n} |P_n q_n \, \mathbb{I}_{\{\overline{q}_n \leq \beta_n\}} - P q_n \mathbb{I}_{\{\overline{q}_n \leq \beta_n\}}| \\
+ \sup_{\mathcal{Q}_n} P_n |q_n| \mathbb{I}_{\{\overline{q}_n > \beta_n\}} + \sup_{\mathcal{Q}_n} P|q| \mathbb{I}_{\{\overline{q}_n > \beta_n\}}.
\end{aligned}$$

Since $|q_n| \leq \overline{q}_n$ for all $q_n \in \mathcal{Q}_n$, the last two terms sum to less than $P_n\overline{q}_n \mathbb{I}_{\{\overline{q}_n > \beta_n\}} + P\overline{q}_n \mathbb{I}_{\{\overline{q}_n > \beta_n\}}$. This converges almost surely to $2P\overline{q}_n \mathbb{I}_{\{\overline{q}_n > \beta_n\}}$. Since $\overline{q}_n \mathbb{I}_{\{\overline{q}_n > \beta_n\}}$ are dominated by $\overline{q}_n$ and $\overline{q}_n \mathbb{I}_{\{\overline{q}_n > \beta_n\}} \to 0$

as $\beta_n \to \infty$, it follows that $P\overline{q}_n \mathbb{I}_{\{\overline{q}_n > \beta_n\}} = o(1)$ by dominated convergence.[15] Thus we can concentrate on the truncated class

$$\mathcal{Q}_{\beta(n)} \equiv \left\{ q_{\beta(n)} = q_n \mathbb{I}_{\{\overline{q}_n \leq \beta_n\}} \ : \ q_n \in \mathcal{Q}_n \right\}.$$

SCALING. Given that the class $\mathcal{Q}_{\beta_n}$ is uniformly bounded, that is $|q_n \mathbb{I}_{\{\overline{q}_n \leq \beta_n\}}| \leq \beta_n$ for all functions in $\mathcal{Q}_{\beta_n}$, we can consider, without loss of generality, the scaled class $\mathcal{Q}^*_{\beta(n)} = \left\{ q^*_{\beta(n)} = q_{\beta(n)}/\beta_n \ : \ q_{\beta(n)} \in \mathcal{Q}_{\beta(n)} \right\}$, where $|q^*_{\beta(n)}| < 1$ for all $q^*_{\beta(n)} \in \mathcal{Q}^*_{\beta(n)}$. Note that $P\overline{q}_n \mathbb{I}_{\{\overline{q}_n > \beta_n\}}/\beta_n = o(1/\beta_n)$, so that the truncation step is not affected by the scaling. It then follows from Pollard (1984, Theorem II.37) applied to the class $\mathcal{Q}^*_{\beta(n)}$ and A4 that

$$\sup_{\mathcal{Q}^*_{\beta(n)}} |P_n q^*_{\beta(n)} - P q^*_{\beta(n)}| \ll h^d \alpha_n \text{ a.s.}, \tag{41}$$

provided

$$\sup_{\mathcal{Q}^*_{\beta(n)}} \left\{ P q^{*2}_{\beta(n)} \right\}^{1/2} < h^{d/2} \tag{42}$$

and

$$\sup N_1(\varepsilon, H, \mathcal{Q}^*_{\beta(n)}) \leq A \varepsilon^{-W} \text{ for } 0 < \varepsilon \leq 1, \tag{43}$$

where the supremum is taken over all probability measures H, and A and W are constants independent of n.

Consider condition (42). Since $|q^*_{\beta(n)}| \leq 1$ uniformly over $\mathcal{Q}^*_{\beta(n)}$, it follows that

$$\begin{aligned}
P q^{*2}_{\beta(n)} &\leq P|q^*_{\beta(n)}| \\
&= \int \beta_n^{-1} \left| \left[\{g(X) - m(X,\alpha)\}^2 + u^2 \right] K\{h^{-1}(X - x)\} \right| \mathbb{I}_{\{\overline{q}_n \leq \beta_n\}} dP \\
&= \int \beta_n^{-1} \left| \left[\{g(t) - m(t,\alpha)\}^2 + \sigma^2(t) \right] K(h^{-1}(t - x)) \right| f_X(t) \mathbb{I}_{\{\overline{q}_n \leq \beta_n\}} dt \\
&= h^d \int \beta_n^{-1} \left| \left[\{g(x + hv) - m(x + hv, \alpha)\}^2 + \sigma^2(x + hv) \right] K(v) \right| f_X(x + hv) \mathbb{I}_{\{\overline{q}_n \leq \beta_n\}} dt \\
&\leq h^d,
\end{aligned}$$

[15] Note that $PQ_n = O(h^d)$ implies that any β_n would suffice for large enough n.

since the integral is an average of functions uniformly bounded by 1.

Consider now the covering number condition (43) that require $\mathcal{Q}^*_{\beta(n)}$ to be a Euclidean class. Note that the functions in $\mathcal{Q}^*_{\beta(n)}$, $q^*_{\beta(n)} = \beta_n^{-1}\left[\{m(X,\alpha^0) - m(X,\alpha)\}^2 + u^2\right] K\{h^{-1}(X-x)\}\mathbb{I}_{\{\overline{q}_n \leq \beta_n\}}$, are composed from the classes of functions $\mathcal{M} = \{\delta m(\cdot,\alpha) \ : \ \alpha \in \mathcal{A},\ \delta \in \mathbb{R}\}$, $\mathcal{U} = \{\gamma u^2 \ : \ \gamma \in \mathbb{R}\}$, $\mathcal{K} = \{K(x^T\rho + \eta) \ : \rho \in \mathbb{R}^d,\ \eta \in \mathbb{R}\}$ with K a measurable real valued function of bounded variation on $\mathbb{R}$, and the class of indicator functions of the envelope $\mathcal{I} = \{\mathbb{I}(\xi\overline{q}_n \leq 1)\colon\ \xi \in \ \mathbb{R}\}$. The Euclidean property of $\mathcal{Q}^*_{\beta(n)}$ will follow if each of the functions used to construct $q^*_{\beta(n)}$ themselves form Euclidean classes, as sums and products of Euclidean classes preserve the Euclidean property. See Nolan and Pollard (1987, Section 5) and Pakes and Pollard (1989, Lemmas 2.14 and 2.15) among others (see also the stability results of Pollard (1984) and Andrews (1994, Theorems 3 and 6)).

The Euclidean property of the class $\mathcal{M}$ follows directly from Pakes and Pollard (1989, Lemma 2.13) and A1. The class $\mathcal{U}$ forms a VC (Vapnik-Červonenkis)–graph class[16] by Pollard (1984, Lemma II.28), and it follows from Pollard's (1984, Lemma II.25) Approximation Lemma that $\mathcal{U}$ is a Euclidean class. The Euclidean property of $\mathcal{K}$ follows directly from Nolan and Pollard (1987, Lemma 22(ii)) (see also Pakes and Pollard (1989, Example 2.10)). Finally, the Euclidean property of $\mathcal{M}$, $\mathcal{U}$, and $\mathcal{K}$ imply that the half-spaces defined by the inequalities $\overline{q}_n \leq \beta_n$ form a VC class and therefore the class of indicator functions $\mathcal{I}$ with envelope 1 is a VC–graph class, so that another application of Pollard's (1984, Lemma II.25) ensures its Euclidean property.

In view of the definition of $Q_n(\theta) = h_d P_n q(\cdot\,,\cdot)$, and $\overline{Q}_n(\theta) = h_d P q(\cdot\,,\cdot)$, of (41), and of the truncation argument, it follows that

$$\sup_{\theta\in\Theta} |Q_n(\theta) - \overline{Q}_n(\theta)| = o_p\left(\alpha_n\right) \quad a.s. \tag{44}$$

Pollard'ss Lemma imposes the constraint $\alpha_n \gg \{\log n/(nh^d)\}^{1/2}$. Choosing $\alpha_n = \{\log n/(nh^d)\}^{1/2}$ gives the final result of the lemma

$$\sup_{\theta\in\Theta} |Q_n(\theta) - \overline{Q}_n(\theta)| = O_p\left(\alpha_n\right) = O_p\left(\left\{\frac{\log n}{nh^d}\right\}^{1/2}\right) \quad a.s. \tag{45}$$

■

Acknowledgements

We would like to thank Don Andrews, Nils Hjort, Chris Jones, and David Pollard for helpful discussions. The comments of a referee corrected several lacunae in our argument and greatly improved

[16] VC classes are also known as classes with polynomial discrimination.

this paper. We thank Joel Horowitz for providing the dataset used in section 4. Financial support from the National Science Foundation and the North Atlantic Treaty Organization is gratefully acknowledged.

References

[1] Anand, S., C.J. Harris, and O. Linton (1993): "On the concept of ultrapoverty," Harvard Center for Population Studies Working paper 93-02.

[2] Andrews, D.W.K. (1993): "Tests for parameter instability and structural change with unknown change point." *Econometrica.* **61**, 821-856.

[3] Andrews, D.W.K. (1994): "Empirical process methods in econometrics." *The Handbook of Econometrics,* Vol. IV. Eds. R.F. Engle III and R.F. McFadden. North Holland.

[4] Ansley, C.F., R. Kohn, and C. Wong (1993): "Nonparametric spline regression with prior information." *Biometrika.* **80**, 75-88.

[5] Bierens, H.J. (1987): "Kernel Estimators of Regression Functions." in *Advances in Econometrics: Fifth World Congress,* Vol 1. Ed. by T.F. Bewley. Cambridge University Press.

[6] Bierens, H.J., and H.A. Pott-Buter (1990): "Specification of Household Engel Curves by Nonparametric Regression." *Econometric Reviews.* **9**, 123-184.

[7] Copas, J.B. (1994): "Local likelihood based on kernel censoring." *Journal of the Royal Statistical Society, Series* B, **57**, 221-235.

[8] Cox, D.R., and N. Reid (1987): "Parameter orthogonality and Approximate conditional inference," (with discussion) *Journal of the Royal Statistical Society, Series* B, **49**, 1-39.

[9] Deaton, A. (1986): "Demand Analysis." in *The Handbook of Econometrics*, Vol. III. Eds Z. Griliches and M.D. Intriligator. North Holland.

[10] Fan, J. (1992): "Design-Adaptive Nonparametric Regression." *Journal of the American Statistical Association,* **87** , 998-1004.

[11] Fan, J. (1993): "Local Linear Regression Smoothers and their Minimax Efficiencies," *The Annals of Statistics,* **21**, 196-216.

[12] Fan, J. and I. Gijbels (1992): "Variable Bandwidth and Local Linear Regression Smoothers." *Annals of Statistics,* **20**, 2008-2036.

[13] Fan, J., N. Heckman and M. Wand (1995): "Local polynomial kernel regression for generalized linear models and quasi-likelihood functions." *Journal of the American Statistical Association,* **90**, 141-150.

[14] Fenton, V.M. and A.R. Gallant (1996): "Convergence rates for SNP density estimators." *Econometrica,* **64**, 719-727.

[15] Giné, E. and J. Zinn (1984): "Some limit theorems for empirical processes." *Annals of Probability,* **12**, 929-989.

[16] Gourieroux, C., Monfort, A., and A. Tenreiro (1994): "Kernel M-Estimators: Nonparametric diagnostics for structural models." Manuscript, CEPREMAP, Paris.

[17] Härdle, W., and O.B. Linton (1994): "Applied nonparametric methods." in *The Handbook of Econometrics*, Vol. IV, pp2295-23339, eds D.F. McFadden and R.F. Engle III. North Holland.

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

[19] Hastie, T., and R. Tibshirani (1993): "Varying-coefficient models," (with discussion) *Journal of the Royal Statistical Society, Series B.* **55**, 757-796.

[20] Hjort, N.L. (1993): "Dynamic likelihood hazard estimation." *Biometrika.* To appear.

[21] Hjort, N.L., and M.C. Jones (1994): "Local fitting of regression models by likelihood: what is important?" Institute of Mathematics, University of Oslo, Statistical Research Report no. 10.

[22] Hjort, N.L., and M.C. Jones (1996): "Locally parametric nonparametric density estimation." *Annals of Statistics* **24**, 1619-1647.

[23] Horowitz, J., (1992): "A smoothed maximum score estimator for the binary response model," *Econometrica* **60**, 505-531.

[24] Horowitz, J.L. (1993): "Semiparametric estimation of a work-trip mode choice model." *Journal of Econometrics.* **58**, 49-70.

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

[26] Kelley, J.L. (1955): *General Topology.* D.Van Nostrand: New York.

[27] Klein, R.W., and R.H. Spady (1993): "An efficient semiparametric estimator for discrete choice models," *Econometrica* **61**, 387-421.

[28] Loader, C.R. (1996): "Local likelihood density estimation." *Annals of Statistics* **24**, 1602-1618.

[29] McManus, D.A. (1994): "Making the Cobb-Douglas functional form an efficient nonparametric estimator through localization," Working Paper no. 94-31, *Finance and Economics Discussion Series*, Federal Reserve Board, Washington D.C.

[30] Manski, C.F. (1975): "The Maximum Score Estimation of the Stochastic Utility Model of Choice," *Journal of Econometrics* **3**, 205-228.

[31] Masry, E., (1996): "Multivariate regression estimation: Local polynomial fitting for time series," *Stochastic Processes and their Applications* **65**, 81-101.

[32] Müller, H.G. (1988): *Nonparametric regression analysis of longitudinal data.* Springer Verlag.

[33] Nolan, D. and D. Pollard (1987): "U-Processes: Rates of convergence." *Annals of Statistics,* **15**, 780-799.

[34] Pakes, A. and D. Pollard (1989): "Simulation and the asymptotics of optimization estimators." *Econometrica,* **57**, 1027-1057.

[35] Pollard, D. (1984): *Convergence of Stochastic Processes.* Springer Verlag.

[36] Pötscher, B.M. and I.R. Prucha (1991): "Basic structure of the asymptotic theory in dynamic nonlinear econometric models I. Consistency and Approximation concepts." *Econometric Reviews,* **10**, 125-216.

[37] Robinson, P.M. (1989): "Time varying nonlinear regression." in *Statistical Analysis and Forecasting of Economic Structural Change.* Eds. P. Hackl and A. Westland. Springer-Verlag.

[38] Ruppert, D. and M.P. Wand (1994): "Multivariate Weighted Least Squares Regression." *The Annals of Statistics,* **22**, 1346-1370.

[39] Schuster, E.F. and S. Yakowitz (1979): "Contributions to the theory of nonparametric regression with application to system identification." *Annals of Statistics,* **7**, 139-145.

[40] Silverman, B.W. (1978): "Weak and strong uniform consistency of the kernel estimate of a density and its derivatives." *Annals of Statistics,* **6**, 177-184.

[41] Staniswalis, J.G. (1989): "The Kernel Estimate of a Regression Function in Likelihood Based Models." *Journal of the American Statistical Association,* **84**, 276-283.

[42] Staniswalis, J.G. and T.A. Severini (1991): "Diagnostics for Assessing Regression Models." *Journal of the American Statistical Association,* **86**, 684-691.

[43] Stoker, T.M. (1986): "Consistent Estimation of Scaled Coefficients." *Econometrica,* **54**, 1461-1481.

[44] Stone, C.J. (1977): "Consistent Nonparametric Regression." *Annals of Statistics,* **5**, 595-620.

[45] Stone, C.J. (1982). "Optimal global rates of convergence for nonparametric regression." *Annals of Statistics,* **8**, 1040-1053.

[46] Tibshirani, R. (1984): "Local Likelihood estimation." PhD Thesis, Stanford University.

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

Figure 1. Comparison of local linear bias term $|g_{xx}(x)|$ and local logit bias term $|g_{xx}(x) - m_{xx}(x)|$ against covariate x, when the true regression is probit.

Figure 2. Crossvalidation curve against logarithm of bandwidth for four different trimming rates.

Figure 3. Local probit $\widehat{g}(x)$ against parametric probit fit $\Phi(\widetilde{\gamma}^T x)$, where

$x = \{1, AUTOS, DOVTT, DIVTT, DCOST\}$

for zero autos (0), one auto (1), and two auto (2) households.

Figure 4abcd. Local $\widehat{\gamma}_j(x)$ against x_j with local linear smooth of points and 99 confidence interval. Parametric $\widetilde{\gamma}_j$ (horizontal line) is shown for comparison.

Figure 5abcd.