Semiparametric Estimation of Weighted Average Derivatives

James L. Powell
James H. Stock
Thomas M. Stoker

WP# 1793-86 June 1986

* University of Wisconsin at Madison, Kennedy School of Government, Harvard University and Sloan School of Management, Massachusetts Institute of Technology respectively. The authors gratefully acknowledge funding for this research from National Science Foundation grants. The authors wish to thank L. Hansen, J. Hausman, J. Heckman, C. Manski, J. Marron, D. McFadden and C. Simms for helpful comments.

ABSTRACT

This paper studies the estimation of the density-weighted average derivative of a general regression function. Let y denote a dependent variable, x a vector of explanatory variables, $g(x)=E(y|x)$ the true regression function and $h(x)$ the marginal density of x. The density-weighted average derivative of the regression function is $\delta=E[h(x)(\partial g/\partial x)]$. We present a computationally simple, $\sqrt{N}$ consistent, asymptotically normal estimator $\hat{\delta}_N$ of δ, that relies on no functional form assumptions on $g(x)$, $h(x)$ or the joint distribution of (y,x). An estimator of the asymptotic variance of $\hat{\delta}_N$ is proposed. $\hat{\delta}_N$ provides a practical solution to the problem of estimating β up to scale when $g(x)$ can be written as $g(x)=F(x'\beta)$, as with many models of limited dependent variables. $\hat{\delta}_N$ also provides a solution to the problem of nonparametrically testing linear first derivative constraints on $g(x)$.

The estimator $\hat{\delta}_N$ is a sample analogue of the product-moment representation of the average derivative, that employs kernel estimates of the density of x. Extensions of classical U-statistic theorems are used to establish asymptotic normality of $\hat{\delta}_N$, and a jackknifing procedure is used to remove asymptotic bias. The correctly-scaled weighted average $E[h(x)\partial g/\partial x]/E[h(x)]$ is shown to be estimated by certain linear instrumental variable coefficients of y regressed on x. Issues in the estimation of more general weighted average derivatives are addressed. The relationship of the results to classical central limit theorems, as well as results on slow convergence rates of pointwise nonparametric estimators is discussed.

SEMIPARAMETRIC ESTIMATION OF WEIGHTED AVERAGE DERIVATIVES

by J. L. Powell, J. H. Stock and T. M. Stoker

1. Introduction

In this paper we consider the estimation of the density-weighted average derivative of a general regression function. Let y denote a dependent variable and x a vector of independent variables, where x is distributed with density h(x) and the true regression function is $E(y|x)=g(x)$. Our interest is in estimation of the weighted average derivative vector $\delta=E[h(x)\partial g/\partial x]$. The approach we take is semiparametric: we propose an estimator of the finite parameter vector δ whose properties are valid without restricting the joint distribution of (y,x). In particular, no functional form assumptions are applied to h(x) or g(x).

The primary interest in weighted average derivatives arises from their role in several semiparametric estimation problems in econometrics. One important example is the "scaled coefficient" problem of Ruud(1986) and Stoker(1986), where the conditional expectation is restricted to the single index form $g(x)=F(x'\beta)$, for some function F. For example, this form arises in many models with limited dependent variables. Here the weighted average derivative δ is proportional to the coefficients β, so an estimator of δ will measure β up to scale. Average derivatives are also useful in testing linear constraints on the first derivatives of g(x); Stoker(1985) discusses these sorts of applications.[1]

A second issue of interest which our approach addresses is the practical question of how to summarize empirically the "typical effects" of changes in x on y. In practice, standard OLS coefficients of y regressed on x carry this interpretation, in accordance with the widespread use of the linear model in empirical work. However, it is well known that OLS coefficients are inconsistent for average derivatives when the true model between y and x is nonlinear and/or the specification of x omits important behavioral variables. A general estimator of δ can be interpreted as a measure of typical effects regardless of whether the true model g(x) is linear or not, and thus is robust to one of the two sources of bias of OLS coefficients. In this context δ can be scaled to a true weighted average $\delta_{IV}=\delta/E[h(x)]=E[h(x)\partial g/\partial x]/E[h(x)]$, and we point out how δ_{IV} is estimated by certain instrumental variables coefficients of y regressed on x.

Formally, in this paper we propose an estimator $\hat{\delta}_N$ of the weighted average derivative $\delta=E[h(x)(\partial g/\partial x)]$, where h(x) is a continuous density function that vanishes on the boundary of the values of x but is otherwise unrestricted, and where N is the sample size. We show that $\hat{\delta}_N$ is a $\sqrt{N}$ consistent, asymptotically normal estimator of δ. We give an estimator of the asymptotic variance-covariance matrix of $\hat{\delta}_N$.

The estimator $\hat{\delta}_N$ is a sample analogue of a product-moment representation of density-weighted average derivatives, that is corrected for asymptotic bias. The representation involves derivatives of the density of x, which are nonparametrically estimated using the kernel density estimation technique of Parzen(1962) and others.[2] The estimator $\hat{\delta}_N$ is based on the appropriate average of the (pointwise) nonparametric estimators and is computed directly from the observed data, requiring no computational techniques for maximization or other types of equation solving.

The verification of the properties of $\hat{\delta}_N$ combines two classical tools of statistical theory. The asymptotic normality of averaged kernel density derivatives is based on an application of Hoeffding's(1948) projection method for U-statistics. The correction for asymptotic bias extends Bierens'(1985) proposal by applying the conceptual logic of Quenouille's(1949) jackknife to several averaged kernel estimators computed using differing bandwidths. While our primary interest is in the properties of $\hat{\delta}_N$, the methods are generally applicable to procedures based on averaged pointwise kernel estimators.

In Section 2, we present our notation and briefly review some properties of kernel estimators. Section 3 proposes our estimator for density-weighted average derivatives, and establishes $\sqrt{N}$ consistency and asymptotic normality. Section 4 addresses two topics: the correctly-scaled instrumental variables estimator, and the special role that the density h(x) plays as a weighting function in our estimation procedure. Section 5 concludes by discussing the role of our results in statistical theory, and topics for future research.

2. Notation, Assumptions and Technical Background

2.1 The Basic Framework and Examples

We consider an empirical problem where y denotes a dependent variable and x a k-vector of independent variables. The data consists of N observations (y_i, x_i'), i=1,...,N, which is assumed to be an i.i.d. random sample from a distribution that is absolutely continuous with respect to a σ-finite measure ν, with (Radon-Nikodym) density H(y,x). The marginal density of x is denoted as h(x), and the regression function of y given x is denoted as $g(x) \equiv E(y|x)$. Our interest is in the estimation of the density-weighted average derivative vector

$$(2.1) \qquad \delta \equiv E\left[h(x)\, \frac{\partial g}{\partial x}\right]$$

We consider alternative weighting functions (other than h(x)) in Section 4.2.

The weighted average derivative δ arises in several semiparametric estimation problems. Examples 1 and 2 indicate two problem areas where estimation of δ is valuable.

Example 1: Estimation of Scaled Coefficients - Suppose that $g(x)$ can be written in "single index" form as $g(x)=F(x'\beta)$, for some unknown function F and coefficients β.[3] This structure implies that the pointwise derivatives of $g(x)$ are proportional to β, as $\partial g/\partial x=[dF/d(x'\beta)]\beta$. δ is then proportional to the coefficients β, since $\delta=E[h(x)\partial g/\partial x]=E[h(x)dF/d(x'\beta)]\beta=\gamma\beta$. An estimate of δ is therefore an estimate of β up to scale, regardless of the form of the unknown function F. The correctly scaled weighted average $\delta_{IV}=\delta/E[h(x)]$ is likewise proportional to the coefficients β.[4] See Ruud(1986) and Stoker(1986) for specific examples of this structure, including many standard limited dependent variables models.

Example 2: Tests of Linear Derivative Constraints - It is often of interest to perform tests of hypotheses of the form $c'(\partial g/\partial x)-c_0=0$, where c is a k-vector of known constants and c_0 a known scalar. Letting $\Delta(x)=c'(\partial g/\partial x)-c_0$ denote the departure, the constraint can be summarized as $\Delta(x)=0$. The density-weighted average departure can be written as $E[h(x)\Delta(x)]=c'E[h(x)\partial g/\partial x]-E[h(x)]c_0=$ $c'\delta-E[h(x)]c_0$, which clearly must vanish if the constraint is valid. An estimate of δ can therefore be used to construct an estimate of the mean departure $E[h(x)\Delta(x)]$, and the asymptotic distribution of such an estimate can be used to construct a test of $E[h(x)\Delta(x)]=0$. The $E[h(x)]$ term of the average departure can be absorbed by considering the correctly scaled weighted departure $E[h(x)\Delta(x)]/E[h(x)]=c'\delta_{IV}-c_0$, so that a test of zero average departure can be based directly on an estimate of δ_{IV}. See Stoker(1985) for a discussion of tests of this type, as well as many statistical and economic examples of derivative constraints.

We now introduce the required assumptions on the the marginal density $h(x)$ and the regression function $g(x)$.

Assumption 1: The support Ω of h(x) is a convex (possibly unbounded) subset of R^k with nonempty interior Ω^o. The underlying measure ν can be written in product form as $\nu=\nu_y\times\nu_x$. where ν_x is Lebesgue measure on R^k. h(x) is continuously differentiable in the components of x for all $x\in\Omega^o$.

Assumption 2: h(x)=0 for all $x\in d\Omega$. where $d\Omega$ denotes the boundary of Ω.

Assumption 3: g(x) is continuously differentiable in the components of x for all $x\in\bar{\Omega}$. where $\bar{\Omega}$ differs from Ω by a set of measure 0. $E(y^2)$. $E(xx')$. $E[y(\partial h/\partial x)]$ and $E[h(x)\partial g/\partial x]$ exist.

Assumption 1 restricts x to be a continuously distributed random variable.[5] where no component of x is functionally determined by other components of x.[6] Assumption 2 is a boundary condition. that allows for unbounded x's where $\Omega=R^k$ and $d\Omega=\emptyset$. Assumption 3 states that the true regression function is a.e. differentiable. and that the weighted average derivative exists. Further regularity conditions are given in Assumption A1 of Appendix 1.

2.2 The Product-Moment Representation of δ

Our approach to the estimation of δ is based on the product-moment representation of the density-weighted average derivative. This representation is based on the following multivariate application of integration by parts:

$$(2.2)\qquad E\left[h(x)\frac{\partial g}{\partial x}\right] = \int \frac{\partial g}{\partial x}h(x)^2\,dx = -2\int g(x)\frac{\partial h}{\partial x}h(x)dx = -2\,E\left[y\frac{\partial h}{\partial x}\right]$$

where the boundary terms in the integration by parts formula vanish by Assumption 2. We formalize the result as Theorem 1:

Theorem 2.1: Given Assumptions 1-3,

$$(2.3) \qquad \delta = E\left[h(x)\ \frac{\partial g}{\partial x}\right] = -2\ E\left[y\ \frac{\partial h}{\partial x}\right] = -2\ \mathrm{Cov}\left[y,\ \frac{\partial h}{\partial x}\right] \quad .$$

We give a formal proof of Theorem 2.1 in Appendix 1.

We propose to estimate δ by the sample analogue of (2.3), where $\partial h/\partial x$ is replaced by a consistent nonparametric estimate. Specifically, let $\hat{h}(x)$ be an estimator of $h(x)$, and let $\partial\hat{h}(x)/\partial x$ denote the associated estimator of its derivative. Then an estimator of δ can be formed as the sample product-moment of (2.2), namely $(-2/N)\ \Sigma\ [y_i \partial\hat{h}(x_i)/\partial x]$.[7] Our specific estimator $\hat{\delta}_N$ of δ uses a kernel estimator of the marginal density $h(x)$. We now review kernel density estimators and some of their properties.

2.3 Kernel Estimators: Notation and Pointwise Convergence Properties

There are a number of methods for nonparametrically estimating an unknown function, as surveyed by Prakasa-Rao(1983). Kernel estimators arise from a particular method of local averaging.[8] A kernel estimator of the density $h(x)$ can be written in the form:

$$(2.4) \qquad \hat{h}(x) = \frac{1}{N} \sum_{i=1}^{N} \left[\frac{1}{\gamma_N}\right]^k W\left[\frac{x - x_i}{\gamma_N}\right]$$

where the "kernel" $W(.)$ is a density function and the "band (or window) width" γ_N is a smoothing parameter that depends on the sample size N. The contribution to $\hat{h}(x)$ of data points that are close to x is determined by $W(.)$, where "closeness" is determined by the bandwidth γ_N. The asymptotic properties of $\hat{h}(x)$ refer to the limiting properties obtained as the sample size N increases and the bandwidth γ_N declines. We make the following assumption on the kernel density $W(.)$:

Assumption 4: The support Ω_W of $W(u)$ is a convex (possibly unbounded) subset of R^k with nonempty interior, with the origin as an interior point. $W(u)$ is a differentiable function such that $W(u) \geq 0$ for all $u \in \Omega_W$, $\int W(u)du=1$ and $\int uW(u)du=0$. $W(u)=0$ for all $u \in d\Omega_W$, where $d\Omega_W$ denotes the boundary of Ω_W. $W(u)$ is a symmetric function; $W(u)=W(-u)$ for all $u \in \Omega_W$.

We denote the derivative of W as $V(u) \equiv \partial W/\partial u$. The symmetry of W implies that V is anti-symmetric: $V(-u)=-V(u)$ for all $u \in \Omega_W$.

Detailed studies of the pointwise properties of kernel density estimators can be found in the statistical literature.[9] For our purposes, some of the known pointwise properties on the convergence of $\hat{h}(x)$ to $h(x)$ are of interest for interpreting our results. First, $\hat{h}(x)$ is asymptotically unbiased for $h(x)$ as $N \to \infty$ and $\gamma_N \to 0$. Second, the variance of $\hat{h}(x)$ is $O(1/N\gamma_N^k)$, and therefore converges to 0 if $N\gamma_N^k \to \infty$. Third, if $N\gamma_N^k \to \infty$ and $N\gamma_N^{k+2} \to 0$, the mean square error of $\hat{h}(x)$ is $\hat{O}(1/N\gamma_N^k)$. While proofs of these properties can be found in the literature, for completeness we summarize their derivations in Appendix 2. These properties imply that the maximal rate of convergence of $\hat{h}(x)$ to $h(x)$ is $\sqrt{N\gamma_N^k}$, which is strictly slower than $\sqrt{N}$ since $\gamma_N \to 0$.

The same slow pointwise convergence is also displayed by kernel density derivative and kernel regression function estimators. The density derivative estimator associated with $\hat{h}(x)$ from (2.4) is

$$(2.5) \qquad \frac{\partial \hat{h}(x)}{\partial x} = \frac{1}{N} \sum_{i=1}^{N} \left[\frac{1}{\gamma_N}\right]^{k+1} V\left[\frac{x - x_i}{\gamma_N}\right] \quad .$$

By analogous arguments to those in Appendix 2, the density derivative estimator $\partial\hat{h}/\partial x$ obeys $E(\partial\hat{h}/\partial x) \to \partial h/\partial x$ and $MSE(\partial\hat{h}/\partial x)=O(1/N\gamma_N^{k+1})$ as $\gamma_N \to 0$, $N\gamma_N^{k+1} \to \infty$ and $N\gamma_N^{k+3} \to 0$. Similarly, the kernel regression function estimator

$$(2.6) \qquad \hat{g}(x) = \frac{1}{N}\,\hat{h}(x)^{-1} \sum_{i=1}^{N} \left[\frac{1}{\gamma_N}\right]^k W\left[\frac{x - x_i}{\gamma_N}\right] y_i$$

obeys $E(\hat{g}) \to g(x)$ and $MSE(\hat{g}(x)) = O(1/N\gamma_N^k)$ as $\gamma_N \to 0$, $N\gamma_N^k \to \infty$ and $N\gamma_N^{k+2} \to 0$.

Such slow rates of convergence imply that precise pointwise nonparametric characterizations of density functions, regression functions and derivatives of such functions will be feasible only for extremely large data sets. These problems are particularly severe for higher dimensional applications (larger k), reflecting a particular embodiment of the "curse of dimensionality" cited by Huber(1985) and others.

We have raised these issues to place our results in a particular context. In the next section, we produce a $\sqrt{N}$ consistent and asymptotically normal estimator of the weighted average derivative δ, that is based on averages of kernel density derivative estimators. Consequently, our results give an example of how the slow convergence rates of pointwise estimators can be speeded up when they are averaged to estimate a finite parameter vector, thereby avoiding the "curse of dimensionality."

3. The Weighted Average Derivative Estimator

3.1 The Average Kernel Estimator and Its Interpretation

A natural estimator of the weighted average derivative $\delta = E[h(x)\partial g/\partial x]$ is the sample analogue of the product-moment representation (2.3) defined as

$$(3.1) \qquad \tilde{\delta}_N = \frac{-2}{N} \sum_{i=1}^{N} \left[\frac{\partial \hat{h}_i(x_i)}{\partial x}\right] y_i \qquad .$$

where $\hat{h}_i(x)$ is the kernel density estimator

$$(3.2) \qquad \hat{h}_i(x) = \frac{1}{N-1} \sum_{\substack{j=1 \\ j \neq i}}^{N} \left[\frac{i}{\gamma_N}\right]^k W\left[\frac{x - x_j}{\gamma_N}\right] \quad .$$

(Note that $\hat{h}_i(x)$ differs from $\hat{h}(x)$ of (2.4) by omitting x_i in the estimation of the density $h(x)$, which we require for technical reasons.) Thus to compute the i^{th} summand of $\tilde{\delta}_N$, the density is estimated as $\hat{h}_i(x)$, the derivative is estimated as $\partial\hat{h}_i(x)/\partial x$, and the summand formed as $(\partial\hat{h}_i(x_i)/\partial x)y_i$.

Provided that $\partial\hat{h}_i(x)/\partial x$ is consistent for $\partial h(x)/\partial x$ (in some sense), $\tilde{\delta}_N$ will be a consistent estimator of δ by the law of large numbers. In Section 3.2 we show that $\sqrt{N}[\tilde{\delta}_N - E(\tilde{\delta}_N)]$ has a limiting normal distribution. In Section 3.3, we introduce the estimator $\hat{\delta}_N$, which is just $\tilde{\delta}_N$ corrected for asymptotic bias, and show that $\sqrt{N}(\hat{\delta}_N - \delta)$ has a limiting normal distribution.

In order to obtain a further interpretation of $\tilde{\delta}_N$, first insert (3.2) in (3.1) to obtain the explicit representation

$$(3.3) \qquad \tilde{\delta}_N = \frac{-2}{N(N-1)} \sum_{i=1}^{N} \sum_{\substack{j=1 \\ j \neq i}}^{N} \left[\frac{1}{\gamma_N}\right]^{k+1} V\left[\frac{x_i - x_j}{\gamma_N}\right] y_i \quad .$$

The estimator $\tilde{\delta}_N$ is easily rewritten in the standard "U-statistic" form as

$$(3.4) \qquad \tilde{\delta}_N = - \binom{N}{2}^{-1} \sum_{i=1}^{N-1} \sum_{j=i+1}^{N} \left[\frac{1}{\gamma_N}\right]^{k+1} V\left[\frac{x_i - x_j}{\gamma_N}\right] (y_i - y_j) \quad ,$$

using the fact that $V(u) = -V(-u)$.[10]

The average kernel estimator $\tilde{\delta}_N$ has a natural "slope" interpretation. Let the subscript ℓ denote a particular component of a k-vector, as in $x_i = (x_{1i}, \ldots, x_{\ell i}, \ldots, x_{ki})'$. The ℓ^{th} component of $\tilde{\delta}_N$ can be written as

$$(3.5) \qquad \tilde{\delta}_{\ell N} = - \begin{bmatrix} N \\ 2 \end{bmatrix}^{-1} \sum_{i=1}^{N-1} \sum_{j=i+1}^{N} \left[\frac{1}{\gamma_N}\right]^{k+1} w_\ell\left[\frac{x_i - x_j}{\gamma_N}\right] \left[\frac{y_i - y_j}{x_{i\ell} - x_{j\ell}}\right]$$

where $w_\ell(u) = -u_\ell \partial W/\partial u_\ell$. $w_\ell(u)$ is a weighting function, where an application of integration by parts gives $\int w_\ell(u)du=\int W(u)du=1$. Equation (3.5) shows that $\tilde{\delta}_{\ell N}$ is a weighted average of the slopes $(y_i-y_j)/(x_{i\ell}-x_{j\ell})$, $i,j=1,\dots,N$, with low weight given to observations with $\|x_i-x_j\|$ large.[11] Consequently, while our approach to estimation uses the product-moment representation instead of kernel estimates of the derivatives of the regression function $g(x)$, the estimator $\tilde{\delta}_N$ embodies the intuitive feature of combining the slope (derivative) estimates $(y_i-y_j)/(x_{i\ell}-x_{j\ell})$ for all i,j,ℓ.

3.2 Asymptotic Normality of the Averaged Kernel Estimator

In this section we establish that $\sqrt{N}[\tilde{\delta}_N - E(\tilde{\delta}_N)]$ has a limiting normal distribution with mean 0 and variance-covariance matrix Σ_δ, and derive an explicit formula for Σ_δ. The results follow from general theorems on the asymptotic behavior of U-statistics, that are extensions of the classical theorems of Hoeffding(1948) (see Serfling(1980) for a recent reference). We first prove the general results as Lemma 3.1 and Corollary 3.1 below, and then apply them to the average kernel estimator $\tilde{\delta}_N$.

Begin by considering a general U-statistic of the form

$$(3.6) \qquad U_N \equiv \begin{bmatrix} N \\ 2 \end{bmatrix}^{-1} \sum_{i=1}^{N-1} \sum_{j=i+1}^{N} p_N(z_i,z_j)$$

where $\{z_i,\ i=1,\dots,N\}$ is an i.i.d. random sample and p_N is a k-dimensional symmetric kernel; i.e. $p_N(z_i,z_j)=p_N(z_j,z_i)$. Also define

(3.7) $$r_N(z_i) = E[p_N(z_i,z_j)|z_i]$$

(3.8) $$\theta_N = E[r_N(z_i)] = E[p_N(z_i,z_j)]$$

(3.9) $$\hat{U}_N = \theta_N + \frac{2}{N} \sum_{i=1}^{N} [r_N(z_i) - \theta_N]$$

where we assume that θ_N exists. $\hat{U}_N$ is called the "projection" of the statistic U_N (Hoeffding(1948)).

Our first result is Lemma 3.1, which establishes the asymptotic equivalence of U_N and $\hat{U}_N$:[12]

Lemma 3.1: If $E[\|p_N(z_i,z_j)\|^2] = o(N)$, then $\sqrt{N}(U_N - \hat{U}_N) = o_p(1)$.

Proof: We prove the equivalence by showing that $NE[\|U_N-\hat{U}_N\|^2]=o(1)$, where $\|U_N-\hat{U}_N\|^2=(U_N-\hat{U}_N)'(U_N-\hat{U}_N)$. Define $q_N(z_i,z_j)=[p_N(z_i,z_j)-r_N(z_i)-r_N(z_j)+\theta_N]$, so that

(3.10) $$U_N - \hat{U}_N = \begin{bmatrix} N \\ 2 \end{bmatrix}^{-1} \sum_{i=1}^{N-1} \sum_{j=i+1}^{N} q_N(z_i,z_j)$$

The expectation of the squared length of the vector $U_N-\hat{U}_N$ is

(3.11) $$E[\|U_N - \hat{U}_N\|^2] = \begin{bmatrix} N \\ 2 \end{bmatrix}^{-2} \sum_{i=1}^{N-1} \sum_{j=i+1}^{N} \sum_{\ell=1}^{N-1} \sum_{m=\ell+1}^{N} E[q_N(z_i,z_j)'q_N(z_\ell,z_m)]$$

Because z_i, i=1,...,N are independent vectors, all terms with $(i,j)\neq(\ell,m)$ have zero expectations (if $i\neq\ell$ and $j\neq m$ this is obvious, and writing out the expectation for $i=\ell$, $j\neq m$ gives a quick verification). Therefore, (3.12) $$E[\|U_N - \hat{U}_N\|^2] = \left[\begin{matrix} N \\ 2 \end{matrix}\right]^{-2} \sum_{i=1}^{N-1} \sum_{j=i+1}^{N} E[\|q_N(z_i,z_j)\|^2]$$

The number of terms in this double summand is $O(N^2)$. The nonzero expectations are each $O(E\|q_N(z_i,z_j)\|^2)=O(E\|p_N(z_i,z_j)\|^2)=o(N)$, the latter equality by assumption. Consequently,

(3.13) $$NE[\|U_N - \hat{U}_N\|^2] = N \left[\begin{matrix} N \\ 2 \end{matrix}\right]^{-2} O(N^2)\ o(N)$$

$$= o(1)$$

as required. QED

Because the projection $\hat{U}_N$ is an average of independent random variables, we can immediately establish the asymptotic normality of U_N:

Corollary 3.1: If $E[\|p_N(z_i,z_j)\|^2] = o(N)$, $E[\|r_N(z_i)\|^2] < \bar{r} < \infty$ for some $\bar{r}>0$, and if $R_N \equiv 4E\{[r_N(z_i)-\theta_N][r_N(z_i)-\theta_N]'\}$ is nonsingular, then the limiting distribution of $R_N^{-1/2}\sqrt{N}(U_N - \theta_N)$ is multivariate normal with mean 0 and variance-covariance matrix I, where $R_N^{-1/2}$ is any square root of the inverse of R_N.

Proof: The follows directly from Lemma 3.1 and the Central Limit Theorem for triangular arrays (c.f. Chung(1974) or Serfling(1980, section 1.9.3), among others). QED

Lemma 3.1 and Corollary 3.1 provide sufficient technical machinery to establish the asymptotic normality of the average kernel estimator $\tilde{\delta}_N$ centered about $E(\tilde{\delta}_N)$. Let $z_i \equiv (y_i,x_i')'$, and rewrite (3.4) as

$$(3.14) \qquad \tilde{\delta}_N = \begin{bmatrix} N \\ 2 \end{bmatrix}^{-1} \sum_{i=1}^{N-1} \sum_{j=i+1}^{N} p_N(z_i, z_j)$$

with

$$(3.15) \qquad p_N(z_i, z_j) \equiv - \left[\frac{1}{\gamma_N}\right]^{k+1} V\left[\frac{x_i - x_j}{\gamma_N}\right] (y_i - y_j)$$

Also define

$$(3.16) \qquad \begin{aligned} \sigma^2(x_i) &\equiv E[(y_i - g(x_i))^2 | x_i] = \mathrm{Var}(y_i | x_i) \\ m(x_i) &\equiv E(y_i^2 | x_i) = \sigma^2(x_i) + [g(x_i)]^2 \end{aligned}$$

To apply Lemma 3.1, we require conditions on the bandwidth γ_N such that $E[\|p_N(z_i, z_j\|^2] = o(N)$. We have

$$\begin{aligned} (3.17) \quad E[\|p_N(z_i, z_j)\|^2] &= \int \left[\frac{1}{\gamma_N}\right]^{2k+2} \left\| V\left[\frac{x_i - x_j}{\gamma_N}\right] \right\|^2 [m(x_i) + m(x_j)] h(x_i) h(x_j) dx_i dx_j \\ &= \left[\frac{1}{\gamma_N}\right]^{k+2} \int \|V(u)\|^2 [m(x_i) + m(x_i + \gamma_N u)] h(x_i) h(x_i + \gamma_N u) dx_i du \\ &= O(\gamma_N^{-(k+2)}) = O[N(N\gamma_N^{k+2})^{-1}] \end{aligned}$$

where the second equality uses the change-of-variables from (x_i, x_j) to $(x_i, u = (x_j - x_i)/\gamma_N)$, with Jacobian γ_N^{-k}. Consequently, we have $E[\|p_N(z_i, z_j)\|^2] = o(N)$ if and only if $N\gamma_N^{k+2} \to \infty$ as $\gamma_N \to 0$, the required bandwidth condition for asymptotic normality of $\tilde{\delta}_N$.

To characterize the bias and asymptotic variance of $\tilde{\delta}_N$, note that for $p_N(z_i, z_j)$ of (3.15) we have

$$(3.18)\qquad r_N(z_i) = E[p_N(z_i,z_j)|z_i]$$

$$= -\int \left[\frac{1}{\gamma_N}\right]^{k+1} V\left[\frac{x_i - x}{\gamma_N}\right] (y_i - g(x))\, h(x)dx$$

$$= \int \left[\frac{1}{\gamma_N}\right] V(u)\ (y_i - g(x_i+\gamma_N u))\ h(x_i+\gamma_N u)du$$

$$= \int \frac{\partial g(x_i+\gamma_N u)}{\partial x}\ h(x_i+\gamma_N u)W(u)du$$

$$- \int [y_i - g(x_i+\gamma_N u)]\ \frac{\partial h(x_i+\gamma_N u)}{\partial x}\ W(u)du$$

where the third equality uses the change-of-variable $u=(x-x_i)/\gamma_N$ and the fourth equality involves integration by parts, where the boundary terms vanish by Assumptions 2 and 4.[13] Consequently, if we define

$$(3.19)\qquad r(z_i) = \frac{\partial g(x_i)}{\partial x}\ h(x_i) - [y_i - g(x_i)]\ \frac{\partial h(x_i)}{\partial x}$$

then $E(\|r_N(z_i)-r(z_i)\|)=O(\gamma_N)$, so that $r_N(z_i)-r(z_i)=O_p(\gamma_N)$.[14] Moreover, $E[r(z_i)]=E[h(x)\partial g/\partial x]=\delta$.

The bias in the average kernel estimator $\tilde{\delta}_N$ is therefore characterized as

$$(3.20)\qquad E(\tilde{\delta}_N) - \delta = E[r_N(z_i)] - \delta$$

$$= E[r(z_i)] - \delta + O(\gamma_N)$$

$$= O(\gamma_N)$$

For $\tilde{\delta}_N$, the variance-covariance matrix R_N of Corollary 3.1 is

$$(3.21)\qquad R_N = 4\ E\{[r_N(z_i)-E(r_N(z_i)][r_N(z_i)-E(r_N(z_i)]'\}$$

$$= 4\ E[r(z_i)r(z_i)'] - 4\delta\delta' + O(\gamma_N)$$

$$= \Sigma_\delta + O(\gamma_N)$$

where Σ_δ is the variance-covariance matrix of $2r(z_i)$. Explicitly, we have

$$(3.22) \qquad \Sigma_\delta = \Sigma_0 + \Sigma_1$$

where

$$(3.23) \qquad \Sigma_0 = 4\ E\left[[y_i - g(x_i)]^2 \frac{\partial h(x_i)}{\partial x} \frac{\partial h(x_i)}{\partial x'}\right]$$

$$(3.24) \qquad \Sigma_1 = 4\ E\left[[h(x_i)]^2 \frac{\partial g(x_i)}{\partial x} \frac{\partial g(x_i)}{\partial x'}\right] - 4\ \delta\delta'$$

We summarize this finding as Theorem 3.1:

Theorem 3.1: Given Assumptions 1-4 and A1, if $\gamma_N \to 0$ and $N\gamma_N^{k+2} \to \infty$, then the average kernel estimator $\tilde{\delta}_N$ of (3.1) is such that $\sqrt{N}[\tilde{\delta}_N - E(\tilde{\delta}_N)]$ has a limiting multivariate normal distribution with mean 0 and variance-covariance matrix Σ_δ of (3.22).

We now turn to a discussion of asymptotic bias. An estimator of the variance-covariance matrix is given in Section 3.4.

3.3 Jackknifing the Asymptotic Bias

The limiting normal distribution of $\sqrt{N}[\tilde{\delta}_N - E(\tilde{\delta}_N)]$ arises from the variation of the weighted average of kernel density derivatives. The bias $E(\tilde{\delta}_N)-\delta$ is due the local averaging inherent to kernel estimators, and vanishes as $\gamma_N \to 0$. However, the asymptotic bias $\sqrt{N}[E(\tilde{\delta}_N)-\delta] = O(\sqrt{N}\gamma_N)$ does not vanish, since asymptotic normality requires that $N\gamma_N^{k+2} \to \infty$. In this section we indicate how to correct $\tilde{\delta}_N$ so that the asymptotic bias vanishes, without affecting the variance of its limiting normal distribution.

The bias of a pointwise kernel estimator can be studied by expanding it in a Taylor series in the bandwidth γ_N.[15] Bierens(1985) proposed eliminating the bias for one-dimensional pointwise kernel estimators by differencing two kernel estimators with appropriately chosen bandwidths, to subtract off the leading term of the Taylor series expansion. In higher dimensional settings, one must eliminate bias terms of order higher than one. Our bias elimination extends this approach to the higher order case, which utilizes differencing over multiple kernel estimators.

We begin by introducing conditions under which the bias can be expanded as a Taylor series in the bandwidth γ_N as

$$(3.25) \qquad E(\tilde{\delta}_N) - \delta = b_1\gamma_N + b_2\gamma_N^2 + \ldots + b_{P-1}\gamma_N^{P-1} + b_P\gamma_N^P + O(\gamma_N^{P+1})$$

where P=(k+4)/2 if k is even and P=(k+3)/2 if k is odd. By Young's version of Taylor's Theorem (c.f. Serfling(1980), among others), this representation is possible if the first P+1 derivatives of $E(\tilde{\delta}_N)$ with respect to γ_N exist at $\gamma_N=0$. Write $E(\tilde{\delta}_N)$ as

$$(3.26) \qquad E(\tilde{\delta}_N) = -2\int \left[\frac{1}{\gamma_N}\right]^{k+1} V\left[\frac{x_1 - x_2}{\gamma_N}\right] y_1\ h(x_1)h(x_2)dx_1dx_2$$

$$= 2\left[\frac{1}{\gamma_N}\right]\int V(u)\ y\ h(x)\ h(x + \gamma_N u)\ dx\ du$$

Expansion of the integral in (3.26) gives the representation (3.25), with[16]

$$(3.27) \qquad b_\rho = \sum_{\ell_1,\ldots,\ell_{\rho+1}=1}^{k} \int u_{\ell_1}\cdots u_{\ell_{\rho+1}}\ V(u)\ \ y\ \frac{\partial^{\rho+1}h(x)}{\partial x_{\ell_1}\ldots\partial x_{\ell_{\rho+1}}}\ h(x)\ dx\ du$$

Thus a sufficient condition for the existence of the expansion (3.25) is

Assumption 5: Let P=(k+4)/2 if k is even and P=(k+3)/2 if k is odd. All partial derivatives of h(x) of order P+2 exist. The expectation $E[y(\partial^\rho h(x)/\partial x_{\ell_1}\ldots\partial x_{\ell_\rho})]$ exists for all $\rho \leq$ P+2. All moments of W(u) of order P+1 exist.

To understand the sufficiency of the moment condition on W(u), consider the following application of integration by parts:

$$(3.28) \qquad \int u_{\ell_1}^{t_1} \cdots u_{\ell_j}^{t_j} \frac{\partial W}{\partial u_{\ell_j}}\, du = - t_j \int u_{\ell_1}^{t_1} \cdots u_{\ell_j}^{t_j-1} \, W(u)\, du$$

When Ω is unbounded, (3.28) can be inserted directly into (3.27). If Ω is bounded, the region of integration for u in (3.27) may be a subset of Ω_W,[17] which would introduce boundary terms into (3.28). Nevertheless, it is clear that the existence of the moments of W(u) guarantees the existence of the integrals involving u and V(u) in (3.27).

The asymptotic bias of $\tilde{\delta}_N$ arises from the terms up to order P-1 in (3.25). Multiply the expansion by $\sqrt{N}$ to express the asymptotic bias as

$$(3.29) \; \sqrt{N}[E(\tilde{\delta}_N) - \delta] = b_1\sqrt{N}\gamma_N + b_2\sqrt{N}\gamma_N^2 + \ldots + b_{P-1}\sqrt{N}\gamma_N^{P-1} + b_P\sqrt{N}\gamma_N^P + O(\sqrt{N}\gamma_N^{P+1})$$

For Theorem 2, we require $N\gamma^{k+2} \rightarrow \infty$, so that the $\sqrt{N}\gamma_N$ through $\sqrt{N}\gamma_N^{P-1}$ terms explode, and we can choose γ_N such that $\sqrt{N}\gamma_N^P \rightarrow 0$. Therefore, asymptotic bias correction can be performed if the leading P-1 terms can be removed, while retaining the order of the remainder.

The key insight for our approach is noting that the coefficients b_ρ, $\rho=1,\ldots,P$ do not depend on the bandwidth γ_N. In particular, we remove the bias by subtracting from $\tilde{\delta}_N$ a weighted sum of P kernel estimators with differing bandwidths: Let $\tilde{\delta}_{\rho N}$ denote the estimator

$$(3.30) \qquad \tilde{\delta}_{\rho N} = \frac{-2}{N} \sum_{i=1}^{N} \left[\frac{\partial \hat{h}_{\rho i}(x_i)}{\partial x}\right] y_i$$

where $\hat{h}_{\rho i}(x)$ is the kernel density estimator with bandwidth $\alpha_{\rho N}\gamma_N$:

$$\hat{h}_{\rho i}(x) = \frac{1}{N-1} \sum_{\substack{j=1 \\ j\neq i}}^{N} \left[\frac{1}{\alpha_{\rho N}\gamma_N}\right]^k W\left[\frac{x - x_j}{\alpha_{\rho N}\gamma_N}\right] \tag{3.31}$$

with $\alpha_{\rho N}=d_\rho \gamma_N^{-\eta}$, $\rho=1,\ldots,P$, $0 < \eta < 1/P$, where d_ρ, $\rho=1,\ldots,P$ are distinct positive constants. Let $c_N=(c_{1N},\ldots,c_{PN})'$ be defined as

$$c_N = D^{-1}\begin{bmatrix}\gamma_N^{\eta} \\ \vdots \\ \gamma_N^{P\eta}\end{bmatrix} \quad ; \qquad D = \begin{bmatrix} d_1 & \cdots & d_P \\ \vdots & & \vdots \\ d_1^P & \cdots & d_P^P\end{bmatrix} \tag{3.32}$$

Finally define the "jackknifed" estimator $\hat{\delta}_N$ of δ as

$$\hat{\delta}_N = \frac{\tilde{\delta}_N - \sum_\rho c_{\rho N}\tilde{\delta}_{\rho N}}{1 - \sum_\rho c_{\rho N}} \tag{3.33}$$

$\hat{\delta}_N$ is the desired estimator without asymptotic bias. We can now establish Theorem 3.2, the main result of this section:

<u>Theorem 3.2</u>: Given Assumptions 1-5 and A1, if γ_N obeys $N\gamma_N^{k+2}\to\infty$ and $N\gamma_N^{2P}\to 0$, then the "jackknifed" estimator $\hat{\delta}_N$ defined in (3.33) is such that $\sqrt{N}(\hat{\delta}_N - \delta)$ has a limiting multivariate normal distribution with mean 0 and variance covariance matrix Σ_δ of (3.22).

<u>Proof</u>: $N\gamma_N^{k+2}\to\infty$ implies that $N(\alpha_{\rho N}\gamma_N)^{k+2}=Nd_\rho^{k+2}\gamma_N^{(1-\eta)(k+2)}\to\infty$, so Theorem 2 implies that $\sqrt{N}[\tilde{\delta}_{\rho N} - E(\tilde{\delta}_{\rho N})]$ has a limiting normal distribution for each $\rho=1,\ldots,P$. If $\bar{\delta}_N = (\tilde{\delta}_N', \tilde{\delta}_{1N}', \ldots, \tilde{\delta}_{PN}')$, the Cramer-Wold device implies that $\sqrt{N}[\bar{\delta}_N - E(\bar{\delta}_N)]$ has a limiting normal distribution with mean 0 and variance-covariance matrix $\Sigma_{\bar{\delta}}$. Consequently, since $0 < \eta < 1/P$, we have $c_N\to 0$ as $\gamma_N\to 0$, and $\sqrt{N}[\hat{\delta}_N - E(\hat{\delta}_N)]$ has a limiting normal distribution with mean 0 and variance covariance matrix $\Sigma_\delta=\lim_{N\to\infty}[1-\Sigma c_{\rho N}]^{-2}[(1,c_N')'\otimes I_k]'\Sigma_{\bar{\delta}}[(1,c_N')'\otimes I_k]$, where I_k is the k×k identity matrix.

The result follows from $\lim \sqrt{N}[E(\hat{\delta}_N) - \delta] = 0$. This is verified directly: First evaluate (3.25) for each $\tilde{\delta}_{\rho N}$ as

$$(3.34) \qquad \begin{bmatrix} E(\tilde{\delta}_{1N}) - \delta \\ \vdots \\ E(\tilde{\delta}_{PN}) - \delta \end{bmatrix} = D' \begin{bmatrix} b_1 \gamma_N^{1-\eta} \\ \vdots \\ b_P \gamma_N^{P-P\eta} \end{bmatrix} + O(\gamma_N^{(P+1)(1-\eta)})$$

so that

$$(3.35) \qquad c_N' \begin{bmatrix} E(\tilde{\delta}_{1N}) - \delta \\ \vdots \\ E(\tilde{\delta}_{PN}) - \delta \end{bmatrix} = (\gamma_N^{\eta}, \ldots, \gamma_N^{P\eta})\ (D^{-1})'D' \begin{bmatrix} b_1 \gamma_N^{1-\eta} \\ \vdots \\ b_P \gamma_N^{P-P\eta} \end{bmatrix} + O(\gamma_N^{P+1-\eta P})$$

$$= b_1\gamma_N + b_2\gamma_N^2 + \ldots + b_P\gamma_N^P + O(\gamma_N^{P+1-\eta P})$$

where D^{-1} exists because the constants d_ρ, $\rho=1,\ldots,P$, are distinct. Consequently, we have that $\sqrt{N}[E(\hat{\delta}_N)-\delta]=\sqrt{N}(\{[E(\tilde{\delta}_N)-\Sigma c_{\rho N}E(\tilde{\delta}_{\rho N})]/(1-\Sigma c_{\rho N})\}-\delta) = O(\sqrt{N}\gamma_N^{P+1-\eta P}) = o(\sqrt{N}\gamma_N^{P})$. Since γ_N is such that $N\gamma_N^{2P}\to 0$, $\lim \sqrt{N}[E(\hat{\delta}_N) - \delta] = 0$. QED

Thus $\hat{\delta}_N$ is a $\sqrt{N}$ consistent, asymptotically normal estimator of the weighted average derivative δ, that depends on no specific functional form assumptions on the true regression function $g(x)$ or the density $h(x)$. We refer to $\hat{\delta}_N$ as a "jackknifed" estimator because the logic of bias removal is in line with Quenouille's(1949) jackknifing procedure for estimating bias (see Efron(1982) for a recent exposition). In particular, Quenouille's jackknife is based on the fact that the bias in many estimators depends on sample size, and that the bias can be estimated by taking differences in estimators computed from samples of varying sizes. Our estimator $\hat{\delta}_N$ is constructed by using the bandwidth γ_N in the same role as the sample size in the classical jackknife procedure.

While the justification for our asymptotic bias elimination is involved, the actual correction is computationally straightforward and applicable to general empirical situations. All that is required is the existence of the expansion (3.25) - knowledge of the values of the expansion coefficients b_ρ, $\rho=1,\ldots,P$ is not required. For a k-dimensional problem, P additional estimators are subtracted for asymptotic bias correction, where P is the smallest integer greater than $(k+3)/2$.[18]

In basic terms, the asymptotic bias arises for the estimator $\tilde{\delta}_N$ because the kernel density estimator $\hat{h}_i(x)$ is a local average estimator of the nonlinear function $h(x)$. Further insight into the structure of the corrected estimator $\hat{\delta}_N$ is available by noting how the local averaging is altered in its definition. Now $\hat{\delta}_N$ can be written in "pseudo average kernel" form as:

$$(3.36) \qquad \hat{\delta}_N = \frac{-2}{N} \sum_{i=1}^{N} \left[\frac{\partial \hat{h}_{iN}(x_i)}{\partial x}\right] y_i$$

with the density estimator $\hat{h}_{iN}(x)$ defined as

$$(3.37) \qquad \hat{h}_{iN}(x) = \frac{\hat{h}_i(x) - \sum_\rho c_{\rho N} \hat{h}_{\rho i}(x)}{1 - \sum_\rho c_{\rho N}} = \frac{1}{N-1} \sum_{\substack{j=1 \\ j\neq i}}^{N} \left[\frac{1}{\gamma_N}\right]^k W_N\left[\frac{x - x_j}{\gamma_N}\right]$$

using the "pseudo kernel" $W_N(u)=[W(u)-\Sigma_\rho c_{\rho N}\alpha_{\rho N}^{-k}W(u/\alpha_{\rho N})]/[1-\Sigma_\rho c_{\rho N}]$.[19] $W_N(u)$ is the weighted difference of similar density functions with different spreads, and will involve positive weights for some points nearby x and negative weights for others. Thus the bias inherent to estimating nonlinear functions with local averaging is removed by averaging with both positive and negative weights, and varying the local weight pattern with sample size N.

3.4 Estimation of the Asymptotic Variance-Covariance Matrix

While one can propose several methods for estimating the asymptotic variance-covariance matrix Σ_δ of $\hat{\delta}_N$, we consider the direct sample analogue estimator that employs kernel representations of the density h(x) and the regression function g(x). In particular, recall that $\Sigma_\delta = 4E[r(z_i)r(z_i)'] - 4\delta\delta'$, where $r(z_i)$ is defined in (3.19). We consider the estimator

$$(3.38) \qquad \hat{\Sigma}_\delta = 4\,\frac{\Sigma_i\,\hat{r}(z_i)\hat{r}(z_i)'}{N} - 4\,\hat{\delta}_N\hat{\delta}_N'$$

where $\hat{r}(z_i)$ is the sample analogue of $r(z_i)$ defined as

$$(3.39) \qquad \hat{r}(z_i) = \frac{\partial\hat{g}(x_i)}{\partial x}\,\hat{h}(x_i) - [y_i - \hat{g}(x_i)]\,\frac{\partial\hat{h}(x_i)}{\partial x}$$

where $\hat{h}(x)$ and $\hat{g}(x)$ are the kernel estimators defined in (2.4) and (2.6) respectively. Clearly $\hat{r}(z_i)$ is a pointwise consistent estimator of $r(z_i)$.

We conjecture that $\hat{\Sigma}_\delta$ is a consistent estimator of Σ_δ, although we do not establish it here. Consider the "estimator" $\tilde{\tilde{\Sigma}}_\delta = 4\Sigma r(z_i)r(z_i)'/N - 4\delta\delta'$. By the weak law of large numbers we have plim $(\tilde{\tilde{\Sigma}}_\delta - \Sigma_\delta) = 0$. Consistency of $\hat{\Sigma}_\delta$ would follow from plim $(\hat{\Sigma}_\delta - \tilde{\tilde{\Sigma}}_\delta) = 0$. We have not established the minimal regularity conditions required for this condition. From the triangle inequality, it is easy to see that sufficient conditions are that $\hat{r}(z_i)$ is a uniformly weakly consistent estimator of $r(z_i)$ and that $r(z_i)$ is bounded in probability on Ω.[20]

Given this consistency, hypothesis tests on the value of some or all of the components of δ can be performed with standard Wald statistics using $\hat{\delta}_N$ and $\hat{\Sigma}_\delta$.[21] In particular, if $R\delta=\delta^0$ is a coefficient restriction of interest, where R is a $k_1 \times k$ matrix of full rank $k_1 \leq k$, then the limiting distribution of $N(R\hat{\delta}_N-\delta^0)'\hat{\Sigma}_\delta^{-1}(R\hat{\delta}_N-\delta^0)$ is χ^2 with k_1 degrees of freedom.

In general situations, the asymptotic distribution of $\hat{\delta}_N$ will have larger variance than that of a parametric estimator of δ, reflecting the efficiency cost of using a nonparametric density characterization. We close this section by noting that there are situations where $\hat{\delta}_N$ may have smaller asymptotic variance than a parametric estimator.

To see this point, consider the product moment estimator when the density h(x) is known exactly. In this case δ will be estimated up to sampling error by:

$$(3.40) \qquad \delta_N^* = \frac{-2}{N} \sum_{i=1}^{N} \left[\frac{\partial h(x_i)}{\partial x}\right] y_i$$

Clearly δ_N^* is a strongly consistent estimator of δ by the strong law of large numbers, and by the central limit theorem, $\sqrt{N}(\delta_N^* - \delta)$ has a limiting normal distribution with mean 0 and variance-covariance matrix

$$(3.41) \qquad \Sigma_\delta^* = \Sigma_0 + \Sigma_2$$

where Σ_0 is defined as in (3.23), and

$$(3.42) \qquad \Sigma_2 = 4\ E\left[[g(x_i)]^2 \frac{\partial h(x_i)}{\partial x} \frac{\partial h(x_i)}{\partial x'}\right] - 4\ \delta\delta'$$

For comparing $\hat{\delta}_N$ and δ_N^*, recall that the asymptotic-covariance of $\hat{\delta}_N$ is $\Sigma_\delta = \Sigma_0 + \Sigma_1$, where Σ_1 is defined in (3.24) as

$$(3.43) \qquad \Sigma_1 = 4\ E\left[[h(x_i)]^2 \frac{\partial g(x_i)}{\partial x} \frac{\partial g(x_i)}{\partial x'}\right] - 4\ \delta\delta'$$

The difference $\Sigma_1 - \Sigma_2$ is in general nondefinite, so that there exists situations where $\hat{\delta}_N$ is more efficient than δ_N^*. From the forms of Σ_1 and Σ_2 above, one circumstance is where g(x) is nearly a constant function, but where h(x) is variable. In this case $\Sigma_1 - \Sigma_2$ will have negative diagonal elements, so that $\hat{\delta}_N$ will give a more precise estimate of δ than δ_N^*.

This counterintuitive feature arises because in a large sample the components of $\hat{\delta}_N$ differ from those of δ_N^* by $2[r(z_i)-(-y_i\partial h(x_i)/\partial x)] = 2[h(x_i)\partial g(x_i)/\partial x + g(x_i)\partial h(x_i)/\partial x]$, a term attributable to the kernel estimation of h(x). Because of the generality of the our framework, these components may be sufficiently negatively correlated with the leading components $(-y_i\partial h(x_i)\partial x)$ to produce an efficiency gain from estimating the density h(x).

4. Related Discussion

4.1 The Instrumental Variables Estimator

In this section we consider the estimation of the correctly scaled average derivative $\delta_{IV}=E[h(x)\partial g/\partial x]/E[h(x)]$. δ_{IV} is arguably a better measure than δ of the "typical effects" of changes in x on y, since it is a true weighted average of the "effects" $\partial g/\partial x$.

It is straightforward to show that δ_{IV} is consistently estimated by $\hat{\delta}_N/\bar{h}_N$, where $\bar{h}_N$ is any consistent estimator of $E[h(x)]$. Likewise, $\sqrt{N}[\hat{\delta}_N/\bar{h}_N - \delta_{IV}]$ will have a limiting normal distribution[22] with mean 0 and variance-covariance matrix $\{E[h(x)]\}^{-2}\Sigma_\delta$. Consequently, the use of any consistent estimator $\bar{h}_N$ of $E[h(x)]$ will permit consistent, asymptotically normal estimation of δ_{IV}.

One such estimator arises naturally as the estimated slope coefficients of a linear regression of y_i on x_i. Begin by applying Theorem 2.1 to $I_k E[h(x)]$, where I_k is the k×k identity matrix, as

$$(4.1) \qquad I_k\ E[h(x)] = -2\ E\left[\frac{\partial h}{\partial x}\ x'\right]$$

Therefore, $I_k E[h(x)]$ is consistently estimated by the density-weighted average derivative estimator replacing y_i by x_i as

$$(4.2) \qquad \hat{\delta}_{xN} = \frac{-2}{N} \sum_{i=1}^{N} \left[\frac{\partial \hat{h}_{iN}(x_i)}{\partial x}\right] x_i'$$

where the "pseudo kernel" density $\hat{h}_{iN}(x)$ is defined in (3.37). Note also that δ_{IV} can be expressed as

$$(4.3) \qquad \delta_{IV} = \left[E\left[\frac{\partial h}{\partial x} x'\right]\right]^{-1} E\left[\frac{\partial h}{\partial x} y\right]$$

Equation (4.3) motivates the following estimator of δ_{IV}. Consider the slope coefficients of the linear equation:

$$(4.4) \qquad y_i = x_i'\hat{d}_N + \hat{u}_i \qquad\qquad i=1,\ldots,N$$

estimated using the "pseudo kernel" density derivatives $\partial\hat{h}_{iN}(x_i)/\partial x$ as instrumental variables, or:

$$(4.5) \qquad \hat{d}_N = \hat{\delta}_{xN}^{-1}\,\hat{\delta}_N = \left[\sum_{i=1}^{N} \left[\frac{\partial \hat{h}_{iN}(x_i)}{\partial x}\right] x_i'\right]^{-1} \left[\sum_{i=1}^{N} \left[\frac{\partial \hat{h}_{iN}(x_i)}{\partial x}\right] y_i\right]$$

The above remarks have established that $\hat{d}_N$ is a $\sqrt{N}$ consistent, asymptotically normal estimator of δ_{IV}, summarized as:[23]

Corollary 4.1: Under the conditions of Theorem 3.2, $\hat{d}_N$ is a consistent estimator of δ_{IV}, and $\sqrt{N}(\hat{d}_N - \delta_{IV})$ has a limiting normal distribution with mean 0 and variance-covariance matrix $\{E[h(x)]\}^{-2}\Sigma_\delta$.

If $\hat{\Sigma}_\delta$ is a consistent estimator of Σ_δ, then $(\hat{\delta}_{xN}')^{-1}\hat{\Sigma}_\delta(\hat{\delta}_{xN})^{-1}$ is a consistent estimator of $\{E[h(x)]\}^{-2}\Sigma_\delta$.

The omission of a constant term from the linear equation (4.4) facilitates the instrumental variables formula (4.5), but is otherwise inconsequential. The covariance representation of (2.3) of Theorem 2.1 is also valid for (4.1), so that the slope coefficient estimator from the linear equation with y_i regressed on x_i and a constant will also be a consistent and asymptotically normal estimator of δ_{IV}, where $(1, \partial\hat{h}_{iN}(x_i)/\partial x')'$ is used as the instrumental variable.

4.2 Statistical Issues in Kernel Estimation of General Weighted Average Derivatives

In this paper we have proposed an estimator of the density-weighted average derivative $\delta=E[h(x)\partial g/\partial x]$, without addressing the question of why the density h(x) is a natural weighting function. A useful extension of the results would be to the estimation of the general weighted average derivative $\delta_\omega=E[\omega(x)\partial g/\partial x]$, where $\omega(x)$ is a known weighting function. Specific choices of $\omega(x)$ could be used to estimate average derivatives over specific subsets of the data. Moreover, for $\omega(x)=1$, $\delta_\omega=E(\partial g/\partial x)$ is the unweighted average derivative studied by Stoker(1986). In this section we give an overview of the issues inherent to estimating δ_ω using kernel techniques, deferring formal analysis to later research. This overview illustrates how our framework can be applied, as well as some problems inherent to the case where $\omega(x)=1$.

We consider the case where $\omega(x)$ is a bounded differentiable function with support Ω_ω (with nonempty interior), restricted so that $\omega(x)h(x)$ vanishes on the boundary of $\Omega_\omega \cap \Omega$.[24] By applying integration by parts, δ_ω can be written in a product-moment representation as

$$(4.6) \qquad \delta_\omega = E\left[\omega(x)\,\frac{\partial g}{\partial x}\right] = -\,E\left[\frac{\omega(x)}{h(x)}\,\frac{\partial h}{\partial x}\,y\right] - E\left[\frac{\partial \omega}{\partial x}\,y\right]$$

A substantive difference between δ_ω and δ of (2.3) is the appearance of the density h(x) in the denominator of the first term above.

A consistent, asymptotically normal estimator of the second term of (4.6) is given by the sample average of the components $\{-[\partial\omega(x_i)/\partial x]y_i\}$, so we concentrate on the estimation of the first term of (4.6). Define $\tilde{\delta}_{\omega N}$ as the sample analogue estimator, using kernel estimates of the density h(x) as

$$(4.7) \qquad \tilde{\delta}_{\omega N} = -\frac{1}{N} \sum_{i=1}^{N} \frac{\omega(x_i)}{\hat{h}_i(x_i)} \frac{\partial \hat{h}_i(x_i)}{\partial x} y_i$$

where $\hat{h}_i(x)$ is the kernel density estimator defined in (3.2). Write $\tilde{\delta}_{\omega N}$ out as

$$(4.8) \qquad \tilde{\delta}_{\omega N} = Q_1 + Q_2 + Q_3$$

where

$$(4.9) \qquad Q_1 = -\frac{1}{N} \sum_{i=1}^{N} \frac{\omega(x_i)}{h_i(x_i)} \frac{\partial \hat{h}_i(x_i)}{\partial x} y_i$$

$$(4.10) \qquad Q_2 = -\frac{1}{N} \sum_{i=1}^{N} \left[\frac{\omega(x_i)}{\hat{h}_i(x_i)} - \frac{\omega(x_i)}{h(x_i)}\right] \frac{\partial h(x_i)}{\partial x} y_i$$

$$(4.11) \qquad Q_3 = -\frac{1}{N} \sum_{i=1}^{N} \left[\frac{\omega(x_i)}{\hat{h}_i(x_i)} - \frac{\omega(x_i)}{h(x_i)}\right] \left[\frac{\partial \hat{h}_i(x_i)}{\partial x} - \frac{\partial h(x_i)}{\partial x}\right] y_i$$

Focus first on the case where Ω_ω is contained in the interior of Ω, such that $h(x) > \varepsilon > 0$ for all $x \in \Omega_\omega$. Q_1 is immediately amenable to analysis using our U-statistic results. In particular, write Q_1 in U-statistic form as

$$(4.12) \qquad Q_1 = \left[\begin{matrix} N \\ 2 \end{matrix}\right]^{-1} \sum_{i=1}^{N-1} \sum_{j=i+1}^{N} \tilde{p}_N(z_i, z_j)$$

with

$$(4.13) \qquad \tilde{p}_N(z_i, z_j) \equiv -\frac{1}{2} \left[\frac{1}{\gamma_N}\right]^{k+1} V\left[\frac{x_i - x_j}{\gamma_N}\right] \left[\frac{\omega(x_i)y_i}{h(x_i)} - \frac{\omega(x_j)y_j}{h(x_j)}\right]$$

Because h(x) is bounded away from 0 on Ω_ω, $\omega(x)/h(x)$ is bounded, so that under regularity conditions we will have $E[\|\tilde{p}_N(z_i, z_j)\|^2] = o(N)$ if $N\gamma_N^{k+2} \to \infty$ as $\gamma_N \to 0$.

Thus Lemma 3.1 and Corollary 3.1 can be applied. For Q_2, consider the Taylor series expansion:

$$(4.14)\qquad \left[\frac{\omega(x)}{\hat{h}_i(x)} - \frac{\omega(x)}{h(x)}\right] = -\frac{\omega(x)}{h(x)^2}\,[\hat{h}_i(x) - h(x)] + \frac{2\omega(x)}{h(x)^3}\,[\hat{h}_i(x) - h(x)]^2 + \ldots$$

Since h(x) is bounded away from 0 on Ω_ω, the leading term of this expansion will uniformly dominate (or that the remaining terms are $O([\hat{h}_i(x)-h(x)]^2)$). (4.14) can then be inserted into (4.10), with Q_2 asymptotically equivalent to a weighted average of $\hat{h}_i(x_i)$, i=1,...,N. In this case our U-statistic theorems could be applied to show that Q_2 is asymptotically normal. Moreover, under regularity conditions one could show that $Q_3=o_p(1/\sqrt{N})$, so that Q_3 does not impact on the asymptotic distribution of $\tilde{\delta}_{\omega N}$. This outlines how the U-statistic results could be used to show asymptotic normality of $\tilde{\delta}_{\omega N}$, when the support of ω(x) lies strictly inside the support of h(x). The asymptotic bias correction can applied directly as outlined in Section 3.3.

Notice however, for the case of the unweighted average derivative with ω(x)=1, the conditions for our results are not met. For ω(x)=1, we will in general have $E[\|\tilde{p}_N(z_i,z_j)\|^2]=\infty$, so that Lemma 3.1 cannot be applied to Q_1. Moreover, when ω(x)=1, the higher order coefficients of (4.14) will in general explode as x approaches the boundary of Ω, eliminating this approach. Consequently, our results are not applicable for this case.

The main point of these remarks is that our focus on the density-weighted average derivative $\delta=E[h(x)\partial g/\partial x]$ avoids estimating expectations of functions which are ratios with the density h(x) in the denominator. Without careful consideration, division by h(x) can induce fundamental violations of the regularity conditions required for our approach. As outlined above, there is ample reason to believe that our approach can be applied to estimation of δ_ω where the support of ω(x) is strictly inside the support of h(x). Our results are not directly applicable to the case where $\omega(x)=1$, the estimation of the unweighted average derivative vector $E(\partial g/\partial x)$.

5. Conclusion

In this paper we have proposed an estimator $\hat{\delta}_N$ of the density-weighted average derivative $\delta=E[h(x)\partial g/\partial x]$. This estimator is based on averaging of nonparametric kernel estimates of the density h(x), and can be computed directly from the data, requiring no computational techniques for maximization or other types of equation solving. We have shown that $\hat{\delta}_N$ is $\sqrt{N}$ consistent and asymptotically normal, and proposed an estimator of its asymptotic variance-covariance matrix. We have also proposed a general estimator $\hat{d}_N$ of the correctly-scaled weighted average $\delta_{IV}=E[h(x)\partial g/\partial x]/E[h(x)]$, as the estimated slope coefficients of the linear regression of y regressed on x, using estimated density derivatives as instrumental variables. $\hat{\delta}_N$ and $\hat{d}_N$ provide fully implimentable solutions to several semiparametric estimation problems, such as the "scaled coefficient" problem of estimating β up to scale in any single index model with $g(x)=F(x'\beta)$.

The characterization of the asymptotic distribution of $\hat{\delta}_N$ uses an extension of the classical U-statistic theory of Hoeffding(1948) as well as a bandwidth jackknifing procedure along the lines proposed by Bierens(1985). These two steps may provide a valuable approach in analyzing general estimators based on nonparametric (kernel) characterizations of unknown density and regression functions.

In broader statistical terms, our results have an interesting role in the general theory of estimation, as a bridge between known distributional properties of nonparametric estimators. On the one hand, nonparametric pointwise estimates of density, density derivatives or regression functions, such as kernel estimators, are consistent for the true values at rates that are necessarily slower than $\sqrt{N}$.[25]. On the other hand, the Central Limit Theorem states that sample average statistics, which are clearly nonparametric estimators, are $\sqrt{N}$ consistent for their expectations. Our results give a nontrivial situation where averaging of nonparametric pointwise estimates permits $\sqrt{N}$ consistency (and asymptotic normality) to be attained.[26]

This feature implies that semiparametric estimators can be placed on the same footing as parametric estimators, in the following sense. The main results of the classical Cramer-Rao theory state that maximum likelihood techniques produce a $\sqrt{N}$ consistent, asymptotically normal parameter estimator for any (sufficiently regular) finitely parameterized statistical model, and that the maximum-likelihood estimator is the best estimator in terms of asymptotic efficiency. Our results give a situation where nonparametric pointwise estimates can be used to estimate a finite parameter vector to yield the same rate of convergence. In other words, for estimating density-weighted average derivatives (the finite parameter vector δ), all specific model restrictions can be relaxed at an efficiency cost of a fixed percentage of the data, that does not increase as the sample size increases. Thus, the enormous data requirements for obtaining precise pointwise estimates are relaxed here, because of the (semiparametric) focus on estimation of a finite vector of parameters.

These remarks give our results as much of an "existence theorem" flavor as a set of practical estimation instructions. Indeed, our results pose a large number of practical future research questions as to the best way to impliment the average derivative estimator. While we have established the proper asymptotic behavior of the kernel bandwidth to establish attractive statistical properties for $\hat{\delta}_N$, future research is necessary to indicate the best way to set bandwidth size in applications, such as whether there exist desirable "cross-validation" techniques[27] for averaged kernel estimators.

Moreover, future research is necessary to indicate specific instructions for choosing the weights in the jackknifing procedure, to best control the bias in estimation.

Appendix 1: Regularity Conditions and Proof of Theorem 2.1

Further regularity conditions are summarized as:

Assumption A1:

i) For all y, x and γ, $W(u)[y - g(x + \gamma u)]h(x + \gamma u) \to 0$ as $\|u\| \to \infty$.

ii) h, $\partial h/\partial x$, g and $\partial g/\partial x$ are Lipschitz continuous in their arguments.

Proof of Theorem 2.1: Let x_1 denote the first component of x, and x_0 the other components, so that $x=(x_1,x_0')'$. For a given value of x_0, denote the range of x_1 as $\omega(x_0)=\{x_1|(x_1,x_0')'\in\Omega\}$. Now apply Fubini's Theorem (c.f. Billingsley(1979), among others) to write $E(h(x)\partial g/\partial X_1)$ as

$$\text{(A1.1)} \quad \int_{\Omega} \frac{\partial g(x)}{\partial x_1} (h(x))^2 dx = \int\left[\int_{\omega(x_0)} \frac{\partial g(x)}{\partial x_1} (h(x))^2 dx_1\right] dx_0)$$

The result follows from the validity of the following equation:

$$\text{(A1.2)} \quad \int_{\omega(x_0)} \frac{\partial g(x)}{\partial x_1} (h(x))^2 dx_1 = - 2 \int_{\omega(x_0)} g(x) \frac{\partial h(x)}{\partial x_1} h(x) dx_1$$

By inserting (A1.2) into (A1.1), $E(h(x)\partial g/\partial x_1)=-2E(g(x)\partial h/\partial x_1)$ is established, and by iterated expectation, $E(g(x)\partial h/\partial x_1)=E(y(\partial h/\partial x_1))$.

To establish (A1.2), note first that the convexity of Ω implies that $\omega(x_0)$ is either a finite interval $[a,b]$ (where a, b depend on x_0), or an infinite interval of the form $[a,\infty)$, $(-\infty,b]$ or $(-\infty,\infty)$. Supposing first that $\omega(x_0)=[a,b]$, integrate the LHS of (A1.2) by parts (c.f. Billingsley(1979)) as

$$\text{(A1.3)}\qquad \int_a^b \frac{\partial g(x)}{\partial x_1}\,(h(x))^2 dx_1 = -\,2\int_a^b g(x)\ \frac{\partial h(x)}{\partial x_1}\,h(x)dx_1$$
$$+\ g(b,x_0)(h(b,x_0))^2 - g(a,x_0)(h(a,x_0))^2$$

The latter two terms represent gh^2 evaluated at boundary points, which vanish by Assumption 2, so that (A1.2) is established for $\omega(x_0)=[a,b]$.

For the unbounded case $\omega(x_0)=[a,\infty)$, note first that the existence of $E(h(x)y)$, $E(h(x)\partial g/\partial x_1)$ and $E(g(x)\partial h/\partial x_1)$ respectively imply the existence of $E(h(x)g(x)|x_0)$, $E(h(x)\partial g/\partial x_1|x_0)$ and $E(g(x)\partial h/\partial x_1|x_0)$ (c.f. Kolmogorov(1950)). Now consider the limit of (A1.3) over intervals $[a,b]$, where $b\to\infty$, rewritten as

$$\text{(A1.4)}\qquad \lim_{b\to\infty} g(b,x_0)(h(b,x_0))^2 = g(a,x_0)(h(a,x_0))^2 + \lim_{b\to\infty}\int_a^b \frac{\partial g(x)}{\partial x_1}\,h(x)^2 dx_1$$
$$+\ 2\lim_{b\to\infty}\int_a^b g(x)\ \frac{\partial h(x)}{\partial x_1}\,h(x)\ dx_1$$

$$= g(a,x_0)(h(a,x_0))^2 + h_0(x_0)E\left[\frac{\partial g}{\partial x_1}\,h(x)\,\middle|\,x_0\right] + 2\ h_0(x_0)E\left[g(x)\frac{\partial h}{\partial x_1}\,\middle|\,x_0\right]$$

so that $C \equiv \lim g(b,x_0)h(b,x_0)^2$ exists, where $h_0(x_0)$ is the marginal density of x_0. Now suppose that $C>0$. Then there exists scalars ε and B such that $0<\varepsilon<C$ and for all $b \geq B$, $|g(b,x_0)h(b,x_0)^2-C|<\varepsilon$. Therefore $g(x_1,x_0)h(x_1,x_0)^2 > (C-\varepsilon)I_{[B,\infty)}$, where $I_{[B,\infty)}$ is the indicator function of $[B,\infty)$. But this implies that $h_0(x_0)E(g(x)h(x)|x_0) = \int g(x_1,x_0)h(x_1,x_0)^2dx_1 > (C-\varepsilon)\int I_{[B,\infty)}dx_1 = \infty$, which contradicts the existence of $E(g(X)h(x)|x_0)$. Consequently, $C>0$ is ruled out. $C<0$ similarly contradicts the existence of $E(g(x)h(x)|x_0)$.

Since $C \equiv \lim g(b,x_0)h(b,x_0)^2 = 0$, and $g(a,x_0)h(a,x_0)^2 = 0$ by Assumption 2, equation (A1.2) is valid for $\omega(x_0)=[a,\infty)$. Analogous arguments establish the validity of (A1.2) for $\omega(x_0)=(-\infty,a]$ and $\omega(x_0)=(-\infty,\infty)$.

The covariance representation $E[y(\partial h/\partial x)]=\mathrm{Cov}(y,\partial h/\partial x)$ is implied by $E(\partial h/\partial x)=0$, which follows by applying (A1.2) with $g(x)=1$. QED

Appendix 2: Pointwise Convergence Properties of Kernel Estimators

In this Appendix, we present brief derivations of the pointwise convergence properties of the kernel density estimator (2.4). First note that $\hat{h}(x)$ is asymptotically unbiased if $\gamma_N \to 0$, as

$$(\text{A2.1}) \qquad \begin{aligned} E(\hat{h}(x)) &= \frac{1}{N} \sum_{i=1}^{N} \int \left[\frac{1}{\gamma_N}\right]^k W\left[\frac{x - z}{\gamma_N}\right] h(z)dz \\ &= \int W(u)h(x - \gamma_N u)du \to h(x) \int W(u)du = h(x) \quad \text{as } \gamma_N \to 0 \end{aligned}$$

where the second equality employs the standard algebraic device for studying kernel estimators, namely a change of variables from z to $u=(x-z)/\gamma_N$, with Jacobian $J(dz/du)=\gamma_N^{\ k}$.

Our primary interest is in the rate of convergence of $\hat{h}(x)$ to $h(x)$. To obtain the maximal rate, we bound the mean square error of $\hat{h}(x)$ by first examining its bias and then its variance. The (absolute) bias can be written as

$$(\text{A2.2}) \qquad \begin{aligned} |E[\hat{h}(x)] - h(x)| &= \left| \int W(u) \; [h(x - \gamma_N u) - h(x) \; du \right| \\ &\leq \gamma_N \sup_x \left\| \frac{\partial h}{\partial x} \right\| \int \|u\| \; W(u) \; du \\ &= O(\gamma_N) \end{aligned}$$

provided that $h(x)$ obeys a Lipschitz condition. The variance of $\hat{h}(x)$ can be written as

$$\begin{aligned} &\text{Var}[\hat{h}(x)] = E(\hat{h}(x) - E[\hat{h}(x)])^2 \\ &= E \left[\frac{1}{N} \sum_{i=1}^{N} \left[\frac{1}{\gamma_N}\right]^k W\left[\frac{x-x_i}{\gamma_N}\right] - E[\hat{h}(x)]\right] \left[\frac{1}{N} \sum_{j=1}^{N} \left[\frac{1}{\gamma_N}\right]^k W\left[\frac{x-x_j}{\gamma_N}\right] - E[\hat{h}(x)]\right] \end{aligned}$$

$$= \frac{1}{N^2} \sum_{i=1}^{N} \sum_{j=1}^{N} E \left[\left[\left(\frac{1}{\gamma_N}\right)^k W\left[\frac{x-x_i}{\gamma_N}\right] - E[\hat{h}(x)]\right]\left[\left(\frac{1}{\gamma_N}\right)^k W\left[\frac{x-x_j}{\gamma_N}\right] - E[\hat{h}(x)]\right]\right]$$

$$= \frac{1}{N^2} \sum_{i=1}^{N} E \left(\left[\left(\frac{1}{\gamma_N}\right)^k W\left[\frac{x-x_i}{\gamma_N}\right] - E[\hat{h}(x)]\right]\right)^2$$

$$= \frac{1}{N^2} \sum_{i=1}^{N} E \left(\frac{1}{\gamma_N}\right)^{2k} W\left[\frac{x-x_i}{\gamma_N}\right]^2 - \frac{1}{N} (E[\hat{h}(x)])^2$$

$$= \frac{1}{N} \left[\frac{1}{\gamma_N}\right]^k \int W(u)^2 h(x-\gamma_N u)\, du - \frac{1}{N} (E[\hat{h}(x)])^2$$

$$= O(1/N\gamma_N^{\ k}) + O(1/N)$$

$$= O(1/N\gamma_N^{\ k})$$

since $\gamma_N \to 0$. Thus

$$\begin{aligned} MSE(\hat{h}(x)) &= E[\hat{h}(x) - h(x)]^2 \\ &= Var(\hat{h}(x)) + (E[\hat{h}(x)] - h(x))^2 \\ &= O(1/N\gamma_N^{\ k}) + O(\gamma_N^{\ 2}) \\ &= O(1/N\gamma_N^{\ k}) \end{aligned}$$

provided that $N\gamma_N^{\ k} \to \infty$ and $(N\gamma_N^{\ k})\gamma_N^{\ 2} = N\gamma_N^{\ k+2} \to 0$. Thus the maximum rate of convergence of $\hat{h}(x)$ to $h(x)$ is $\sqrt{N\gamma_N^{\ k}} < \sqrt{N}$, since $\gamma_N \to 0$ is required. The properties of kernel regression function estimators and density derivative estimators discussed in Section 2.3 are verified by analogous derivations.

Notes

1. These examples are formally reviewed in Section 2.

2. Prakasa-Rao(1983) provides a survey of these methods.

3. For binary response models, Manski(1986) refers to this assumption as "index sufficiency." Note that any intrinsic single index model with $E(y|X)=F_1[F_2(X)'\beta]$ for some vector X is included in this example by setting $x=F_2(X)$. For instance, the model $E(y|X)=X_1^{\beta_1}\exp(\beta_2X_2)$ is included by setting $x=(\ln X_1,X_2)'$, and by the following remarks, δ is proportional to $\beta=(\beta_1,\beta_2)'$.

4. This scaling retains the property that if $g(x)=\alpha+x'\beta$, then $\delta_{IV}=\beta$.

5. Continuity of x is essential in this context because of the generality of the dependent variables considered. For example, for the binary response model, Manski(1985) points out how continuity of the regressors is useful for identification with index restrictions.

We make use of many integrals over Ω and affine transformations of Ω. Since all of these integrals utilize Lebegue measure, for clarity we indicate the argument of integration instead of the appropriate measure. For instance, we will write $\int h(x)dx$ instead of $\int h(x)d\nu_x$.

6. This condition is used for the application of integration by parts, but is otherwise inconsequential. Because we are estimating δ nonparametrically, including functionally related variables is superfluous.

7. Other methods of semiparametric estimation of δ can be proposed. For instance, a more direct approach could begin by forming nonparametric estimators of the density h(x) and the regression function g(x), say $\hat{h}(x)$ and $\hat{g}(x)$. An estimate of δ could then be defined as the sample average of the derivatives of the estimated function, namely $(1/N)\ \Sigma\ \hat{h}(x_i)\partial\hat{g}(x_i)/\partial x$. This estimator cannot be written as a U-statistic, and so our approach is not directly applicable (see Section 3.2); the relationship of this approach to our approach raises interesting questions for future research.

8. Other methods include nearest neighbor estimation, as studied by Stone(1977) and others. For a review of nonparametric estimators in the context of econometric problems, see McFadden(1986).

9. See, for example, Spiegelman and Sacks(1980), Stone(1984) and Bierens(1983).

10. This represents the main use of the symmetry of W(.). We can dispense with the symmetry of W by using a "symmetrized" representation of the kernel,

as described in Serfling(1980, p.172). We assume the symmetry of W only to avoid nonessential complications in the notation and exposition.

11. Equations (3.3-5) point out that computing $\tilde{\delta}_N$ involves a calculation of order N^2 for sample size N. In general, the weight w_ℓ of (3.5) is nonmonotonic in the difference $\|x_i-x_j\|$, where $\|.\|$ denotes the standard Euclidean norm on R^k, namely $\|u\|=\sqrt{\Sigma u_\ell^2}$. The fact that w_ℓ integrates to 1 implies that slopes associated with $\|x_i-x_j\|$ suitably large are allocated small weight. For instance, if k=1 and $W(u)=\phi(u)$, the univariate normal density function, then $w(u)=u^2\phi(u)$, and $w[(x_i-x_j)/\gamma_N]$ is increasing in $|x_i-x_j|$ for $|x_i-x_j|<2\gamma_N$ and decreasing for $|x_i-x_j|>2\gamma_N$. A similar structure exists for W(u) equal to the k-variate normal density function.

12. Lemma 3.1 is a straightforward extension of the results given in Serfling(1980, p. 186-188), generalizing his results to the case in which $p_N(z_i,z_j)$ varies with N.

13. This also utilizes condition i) of Assumption A1.

14. This utilizes condition ii) of Assumption A1.

15. For a general approach to bias reduction by series expansion, see Cox and Hinkley(1974).

16. All boundary terms in the derivatives vanish by Assumptions 2 and 4.

17. This is because $u=(x_2-x_1)/\gamma_N$ by the original change-of-variables.

18. Often the choice of kernel W(u) will permit asymptotic bias correction with a smaller number of additional estimators. For instance, if Ω is unbounded and W(u) is the k-dimensional normal density (with covariance matrix I_k), then the odd moments of W(u) in (3.28) are zero, so that b_ρ of (3.27) is zero when ρ is even. If this case, correction will require substracting off one estimator for every b_ρ term with ρ odd, or roughly half the additional estimators required in Theorem 3.

19. $W_N(u)$ is not a true kernel density because it depends explicitly on sample size N.

20. Bierens(1983) establishes conditions under which $\hat{g}(x)$ is uniformly consistent for g(x). His results suggest that uniformity with require that the kernel estimators comprising $\hat{r}(z)$ be computed with bandwidth $\bar{\gamma}_N$, where $N\bar{\gamma}_N^{2k+2}\to\infty$ as $\bar{\gamma}_N\to 0$. Of course, under this condition for any given sample size one could take $\bar{\gamma}_N=\gamma_N$, however the asymptotic theory would require that $\hat{r}(z)$ be recomputed with more slowly shrinking bandwidths as the sample size increases.

21. For the single index framework where $g(x)=F(x'\beta)$, scale-free restrictions on the value of β correspond with restrictions on the components

of δ (c.f. Stoker(1986)).

22. It is easy to verify that asymptotic bias correction can be performed after scaling by $\bar{h}_N$: $\tilde{\delta}_N/\bar{h}_N$ has an asymptotic normal distribution centered at $E(\tilde{\delta}_N)/E[h(x)]$, and the asymptotic bias is corrected by subtracting off the c_N weighted average of $\tilde{\delta}_{\rho N}/\bar{h}_N$, $\rho=1,\ldots,P$.

23. If $\tilde{d}_N$ denotes the slope coefficient estimator using the true kernel density derivative estimators $\partial\hat{h}_i(x_i)/\partial x$ as instrumental variables, then $\tilde{d}_N$ is a consistent, asymptotically normal estimator of $E(\tilde{\delta}_N)/E[h(x)]$, which can be bias corrected to yield an estimator of δ_{IV}, as indicated in note 22 above.

24. This does not permit $\omega(x)$ to be an indicator function over a convex subset of Ω, however $\omega(x)$ could be a (say Gaussian) smoothed version of an indicator function.

25. See Stone(1977) and McFadden(1986) among others.

26. Stock(1985) proves asymptotic normality for a specific average kernel estimator (centered around its mean), and analyzes the asymptotic bias via simulation.

27. See, for example, Rice(1984) and Marron(1985).

References

Bierens, H.J.(1983), "Uniform Consistency of Kernel Estimators of a Regression Function Under Generalized Conditions," Journal of the American Statistical Association, 78, 699-707.

Bierens, H.J.(1985), "Kernel Estimators of Regression Functions," paper presented in the Invited Symposium on Nonparametric and Robust Estimation at the Fifth World Congress of the Econometric Society, Cambridge, Massachusetts.

Billingsley, P.(1979), Probability and Measure, Wiley, New York.

Chung, K.L.(1974), A Course in Probability Theory, 2nd ed., Academic Press, New York.

Cox, D.R. and D.V. Hinkley(1974), Theoretical Statistics, Chapman Hall Ltd., London.

Efron, B.(1982), The Jackknife, the Bootstrap and Other Resampling Plans, CBMS Regional Conference Series in Applied Mathematics, 38, Society for Industrial and Applied Mathematics, Philadelphia.

Hoeffding, W.(1948), "A Class of Statistics with Asymptotically Normal Distribution," Annals of Mathematical Statistics, 19, 293-325.

Huber, P.J.(1985), "Projection Pursuit," Annals of Statistics, 13.

Kolgomorov, A.N.(1950), Foundations of the Theory of Probability, (German edition 1933), Chelsea, New York.

Marron, J.J.(1985), "An Asymptotically Efficient Solution to the Bandwidth Problem of Kernel Density Estimation," Annals of Statistics, 13, 1011-1023.

Manski, C.F.(1986), "Identification of Binary Response Models," Social Science Research Institute Working Paper No. 8602, February.

McFadden, D.(1985), "Specification of Econometric Models," Presidential Address to the Fifth World Congress of the Econometric Society, Cambridge, Massachusetts.

Parzen, E.(1962), "On Estimation of a Probability Density Function and Mode," Annals of Mathematical Statistics, 33, 1065-1076.

Prakasa Rao, B.L.S.(1983), Nonparametric Functional Estimation, Academic Press, New York.

Quenouille, M.(1949), "Approximate Tests of Correlation in Time Series," Journal of the Royal Statistical Society, Series B, 11, 18-84.

Rice, J.(1984), "Bandwidth Choice for Nonparametric Regression," Annals of Statistics, 12, 1215-1230.

Ruud, P.A.(1986), "Consistent Estimation of Limited Dependent Variables Models Despite Misspecification of Distribution," Journal of Econometrics, 18.

Serfling, R.J.(1980), Approximation Theorems of Mathematical Statistics, John Wiley and Sons. New York.

Spiegelman, C. and J. Sacks(1980), "Consistent Window Estimation in Nonparametric Regression," Annals of Statistics, 8, 240-246.

Stock, J.H.(1985), "Nonparametric Policy Analysis," Kennedy School of Government Working Paper, revised November.

Stoker, T.M.(1985), "Tests of Derivative Constraints," MIT Sloan School of Management Working Paper No. 1649-85, April.

Stoker, T.M.(1986), "Consistent Estimation of Scaled Coefficients," forthcoming in Econometrica.

Stone, C.J.(1977), "Consistent Nonparametric Regression," Annals of Statistics, 5, 595-620.

Stone, C.J.(1984), "An Asymptotically Optimal Window Selection Rule for Kernel Density Estimates," Annals of Statistics, 12, 1285-1298.