working paper
department
of economics

KERNEL ESTIMATION OF PARTIAL MEANS
AND A GENERAL VARIANCE ESTIMATOR

Whitney K. Newey

No. 93-3 Dec. 1992

massachusetts
institute of
technology

50 memorial drive
cambridge, mass. 02139

KERNEL ESTIMATION OF PARTIAL MEANS AND A GENERAL VARIANCE ESTIMATOR

Whitney K. Newey

No. 93-3 Dec. 1992

MIT Economics Working Paper 93-3

KERNEL ESTIMATION OF PARTIAL MEANS AND A GENERAL VARIANCE ESTIMATOR*

Whitney K. Newey

MIT Department of Economics

December, 1991

Revised: December, 1992

*Financial support was provided by the NSF. P. Robinson and T. Stoker provided useful comments.

Abstract

Econometric applications of kernel estimators are proliferating, suggesting the need for convenient variance estimates and conditions for asymptotic normality. This paper develops a general "delta method" variance estimator for functionals of kernel estimators. Also, regularity conditions for asymptotic normality are given, along with a guide to verifying them for particular estimators. The general results are applied to partial means, which are averages of kernel estimators over some of their arguments with other arguments held fixed. Partial means have econometric applications, such as consumer surplus estimation, and are useful for estimation of additive nonparametric models.

Keywords: Kernel estimation, partial means, standard errors, delta method, functional estimation.

1. Introduction

There are a growing number of applications where estimators use the kernel method in their construction, i.e. where functionals of kernel estimators are involved. Examples include average derivative estimation (Hardle and Stoker, 1989, and Powell, Stock, and Stoker, 1989), nonparametric policy analysis (Stock, 1989), consumer surplus estimation (Hausman and Newey, 1992), and others that are the topic of current research. An important example in this paper is a *partial mean*, which is an average of a kernel regression estimator over some components, holding others fixed. The growth of kernel applications suggests the need for a general variance estimator, that applies to many cases, including partial means. This paper presents one such estimator. Also, the paper gives general results on asymptotic normality of functionals of kernel estimators.

Partial means control for covariates by averaging over them. They are related to additive nonparametric models and have important uses in economics, as further discussed below. It is shown here that their convergence rate is determined by the number of components that are averaged out, being faster the more components that are averaged over.

The variance estimator is based on differentiating the functional with respect to the contribution of each observation to the kernel. A more common method is to calculate the asymptotic variance formula and then "plug-in" consistent estimators. This method can be quite difficult when the asymptotic formula is complicated, as often seems to be the case. In contrast, the approach described here only requires knowing the form of the functional and kernel. Also, it gives consistent standard errors even for fixed bandwidths (when the estimator is centered at its limit), unlike the more common approach. In this way it is like the Huber (1967) asymptotic variance for m-estimators. Also, it is a generalization of the "delta method" for functions of sample means.

An alternative approach to variance estimation, or confidence intervals, is the bootstrap. The bootstrap may give consistent confidence intervals (e.g. by the percentile method) for the same types of functionals considered here, although this does not appear to be known. In any case, variance estimates are useful for bootstrap improvements to the asymptotic distribution, as considered in Hall (1992).

The variance formula given here has antecedents in the literature. For a kernel density at a point it is equal to the sample variance of the kernel observations, as recently considered by Hall (1992). For a kernel regression at a point, a related estimator was proposed by Bierens (1987). Also, the standard errors for average derivatives in Hardle and Stoker (1989) and Powell, Stock, and Stoker (1989) are equal to this estimator when the kernel is symmetric. New cases included here are partial means and estimators that depend (possibly) nonlinearly on all of the density or regression function, and not just on its value at sample points.

Section 2 sets up m-estimators that depend on kernel densities or regressions, and gives examples. Section 3 gives the standard errors, i.e. the asymptotic variance estimator. Section 4 describes partial means and their estimators, and associated asymptotic theory. Section 5 gives some general lemmas that are useful for the asymptotic theory of partial means, and more generally for other nonlinear functionals of kernel estimators. The proofs are collected in Appendix A, and Appendix B contains some technical lemmas.

2. The Estimators

The estimators considered in this paper are two step estimators where the first step is a vector of kernel estimators. To describe the first step, let y be a $r \times 1$ vector of variables, x a $k \times 1$ vector of continuously distributed variables, and denote the product of the density $f_0(x)$ of x with $E[y|x]$ as

(2.1) $\quad h_0(x) = E[y|x]f_0(x).$

Let $\mathcal{K}(u)$ denote a kernel function satisfying $\int\mathcal{K}(u)du = 1$ and other conditions given in Section 4, where u is $k \times 1$. Let z_i, $(i = 1, \ldots, n)$, denote data observations, that include observations y_i and x_i on y and x. Then for a bandwidth $\sigma > 0$ and $K_\sigma(u) = \sigma^{-k}\mathcal{K}(u/\sigma)$, a kernel estimator of h_0 is

(2.2) $\quad \hat{h}(x) = n^{-1}\sum_{j=1}^{n} y_j K_\sigma(x-x_j).$

This estimator is the first step considered here.

A second step allowed for in this paper is an m-estimator that depends on the estimated function $\hat{h}$. To describe such an estimator, let β denote a vector of parameters, with true value β_0, and $m(z,\beta,h)$ a vector of functions that depend on the observation, parameter, and the function h. Here $m(z,\beta,h)$ is allowed to depend on the entire function h, and not just its value at observed points; see below for examples. Suppose that $E[m(z,\beta_0,h_0)] = 0$. A second step estimator $\hat{\beta}$ that solves a corresponding sample equation is

(2.3) $\quad n^{-1}\sum_{i=1}^{n} m(z_i,\beta,\hat{h}) = 0.$

This is a two-step m-estimator where the first step is the kernel estimator described above.

This estimator includes as special cases functions of kernel estimators evaluated at points, e.g. a kernel density estimator at a point. Some other interesting examples are as follows:

Partial Means: An example that is (apparently) new is an average of a nonparametric regression over some variables holding others fixed. Let q denote a random variable and $g_0(x) = E[q|x]$. Partition $x = (x_1,x_2)$ and let $\tilde{x}_2$ be a variable that is included in z and has the same dimension as x_2, and $\bar{x}_1$ be some fixed value for x_1. Let $\tau(x_2)$ be some weight function, possibly associated with fixed "trimming" that keeps a denominator bounded away from zero. A partial mean is

$$\beta_0 = E[\tau(\tilde{x}_2)g_0(\bar{x}_1,\tilde{x}_2)]. \tag{2.3a}$$

This object is an average over some conditioning variables holding others fixed. It can be estimated by substituting a kernel estimator for g_0 and a sample average for the expectation. Let $y = (1,q)$, $\hat{g}(x) = \hat{h}_2(x)/\hat{h}_1(x) = \hat{h}_2(x)/\hat{f}(x)$, for the kernel density estimator $\hat{f}(x) = \hat{h}_1(x)$, and $\bar{x}_i = (\bar{x}_1,\tilde{x}_{2i})$. Then the estimator is

$$\hat{\beta} = n^{-1}\sum_{i=1}^{n}\tau(\tilde{x}_{2i})\hat{g}(\bar{x}_i). \tag{2.4}$$

This estimator is a special case of equation (2.3) with $m(z,\beta,h) = \tau(\tilde{x}_2)h_2(\bar{x}_1,\tilde{x}_2)/h_1(\bar{x}_1,\tilde{x}_2) - \beta$. It shows how explicit estimators can be included as special cases of equation (2.3). Further discussion is given in Section 4.

Differential Equation Solution: An estimator with economic applications is one that solves a differential equation depending on a nonparametric regression. To describe this estimator, let $y = (1,q)$ and suppose x is two-dimensional (i.e. $k = 2$), with $x = (x_1,x_2)'$. Let $\bar{x}_1$ be some fixed value for x_1 and consider two possible values for x_2, denoted by p^0 and p^1, with $p^0 < p^1$. The estimator is given by

$$\hat{\beta} = S(p^0), \quad dS(p)/dp = -\hat{g}(\bar{x}_1-S(p),p), \quad S(p^1) = 0, \tag{2.5}$$

for $\hat{g}(x) = \hat{h}_2(x)/\hat{f}(x)$. It is a special case where the $m(z,\beta,h)$ of equation (2.3) is the solution of the differential equation minus β. This example shows one way that $m(z,\beta,h)$ can be allowed to depend on the entire function h. The economic interpretation of $\hat{\beta}$ is a nonparametric estimate of the cost of a change of price p, of a commodity q, from p^0 to p^1, for an individual with income $\bar{x}_1$ and demand function $g_0(x) = E[q|x]$. This example is analyzed in Hausman and Newey (1992), using results developed here.

Inverse Density Weighted Least Squares: An estimator that is useful for estimating the semiparametric generalized regression model $E[y|x] = t(x'\delta)$, where $t(\cdot)$ is an unknown transformation, is a weighted least squares estimator, that can be described as follows. Let $\tau(x)$ be a density of an elliptically symmetric distribution (i.e. $\tau(x)$ is a density that depends only on $(x-\mu)'\Sigma(x-\mu)$ for some μ and Σ), that has bounded support. The estimator solves

$$(2.6) \qquad \hat{\beta} \text{ minimizes } \sum_{i=1}^{n}\hat{f}(x_i)^{-1}\tau(x_i)[q_i-x_i'\beta]^2.$$

This estimator has the form given in equation (2.3), with $m(z,\beta,h) = h(x)^{-1}\tau(x)x[q-x'\beta]$. The weighting by the inverse density leads to $\hat{\beta}$ converging to the least squares projection of $E[q|x]$ on x under the density $\tau(x)$, which is consistent for scaled coefficients of a generalized regression model, as discussed in Ruud (1986) and Li and Duan (1991). This estimator is analyzed in Newey and Ruud (1991), using results developed here.

The results of this paper apply to each of these examples, as discussed below. They will also apply to other estimators, including those that minimize a quadratic form in a sample average depending on β and $\hat{h}$, or minimize a sample average, such as quasi-maximum likelihood estimators that depend on kernel estimators.

3. The Asymptotic Variance Estimator

To form approximate confidence intervals and test statistics it is important to have consistent standard errors. To motivate the form of the asymptotic variance estimator, it is helpful to briefly sketch the asymptotic distribution theory. Expanding the left-hand side of equation (2.3) around β_0 and solving for $\hat{\beta} - \beta_0$ gives

$$(3.1) \qquad \hat{\beta} - \beta_0 = -[n^{-1}\sum_{i=1}^{n}\partial m(z_i,\bar{\beta},\hat{h})/\partial\beta]^{-1}\hat{m}_n(\beta_0), \quad \hat{m}_n(\beta) = \sum_{i=1}^{n}m(z_i,\beta,\hat{h})/n,$$

where $\bar{\beta}$ is a mean value. By the uniform law of large numbers discussed in Section 4, $n^{-1}\sum_{i=1}^{n}\partial m(z_i,\bar{\beta},\hat{h})/\partial\beta$ will converge in probability to

$$M = E[\partial m(z,\beta_0,h_0)/\partial\beta].$$

so that the asymptotic distribution of $\hat{\beta}$ will be determined by $\hat{m}_n(\beta_0)$. In Section 4 conditions will be given for existence of $\alpha \geq 0$ such that

$$(3.1a) \qquad \sqrt{n}\sigma^{\alpha}\hat{m}_n(\beta_0) \xrightarrow{d} N(0,V).$$

The magnitude of α will be determined by the form of $E[m(z,\beta_0,h)]$ as a function of h, with α being smaller the more dimensions being integrated over in $E[m(z,\beta_0,h)]$. By the Slutzky theorem the asymptotic distribution for $\hat{\beta}$ will be

$$(3.1b) \qquad \sqrt{n}\sigma^{\alpha}(\hat{\beta} - \beta_0) \xrightarrow{d} N(0,M^{-1}VM^{-1\prime}).$$

A consistent asymptotic variance estimator can be constructed by substituting estimates for true values in the formula $M^{-1}VM^{-1\prime}$. It is easy to construct an estimator of M, as

$$(3.1c) \qquad \hat{M} = n^{-1}\sum_{i=1}^{n}\partial m(z_i,\hat{\beta},\hat{h})/\partial\beta.$$

Finding a consistent estimator of V is more difficult, because of the need to account for the presence of $\hat{h}$ in $\hat{m}_n(\beta_0)$. One common approach to this problem is to calculate the asymptotic variance, and then form an estimator by substituting estimates for unknown functions, such as sample averages for expectations. This approach can be difficult when the asymptotic variance is very complicated. Also, it may be sensitive to the bandwidth parameter, because the variance formula is only valid in the limit as $\sigma \longrightarrow 0$.

The asymptotic variance estimator here is constructed by estimating the influence of each observation in $\hat{h}$ on $\sum_{i=1}^{n} m(z_i,\hat{\beta},\hat{h})/n$. Let ζ denote a scalar number and let

$$(3.2) \qquad \hat{\delta}_i = \partial[n^{-1}\sum_{j=1}^{n} m(z_j,\hat{\beta},\hat{h} + \zeta y_i K_\sigma(\cdot - x_i))]/\partial\zeta|_{\zeta=0}.$$

The interpretation of $\hat{\delta}_i$ is that it estimates the first-order effect of the i^{th} observation in $\hat{h}$ on $\sum_{j=1}^{n} m(z_j,\beta_0,\hat{h})/n$. In this sense it is an "influence function" estimator. The variance can be estimated by including this term with $m(z_i,\hat{\beta},\hat{h})$ in a sample variance, as in

$$(3.3) \qquad \hat{V} = \sum_{i=1}^{n}\hat{\psi}_i\hat{\psi}_i'/n, \quad \hat{\psi}_i = m(z_i,\hat{\beta},\hat{h}) + \hat{\delta}_i - \sum_{j=1}^{n}\hat{\delta}_j/n.$$

An asymptotic variance estimator for $\hat{\beta}$ can then be constructed by combining $\hat{V}$ with a Jacobian estimator in the usual way, as in

$$(3.4) \qquad \hat{Var}(\hat{\beta}) = \hat{M}^{-1}\hat{V}\hat{M}^{-1\prime}, \quad \hat{M} = n^{-1}\sum_{i=1}^{n}\partial m(z_i,\hat{\beta},\hat{h})/\partial\beta.$$

In Section 5 conditions will be given that are sufficient for $\sigma^{2\alpha}\hat{Var}(\hat{\beta}) \xrightarrow{p} M^{-1}VM^{-1\prime}$. Consequently, inference procedures based on $\hat{\beta} - \beta_0$ being normally distributed with mean 0 and variance $\hat{Var}(\hat{\beta})/n$ will be asymptotically valid. For example, $\hat{\beta}_j \pm 1.96[\hat{Var}(\hat{\beta})_{jj}/n]^{1/2}$ will be an asymptotic 95 percent confidence interval. It is interesting to note that the form of $\hat{Var}(\hat{\beta})$ does not depend on the convergence rate for $\hat{\beta}$ (i.e. on α), but that its large sample behavior will.

This asymptotic variance estimator accounts for the presence of $\hat{h}$ by including the terms $\hat{\delta}_i$ in $\hat{\psi}_i$. These terms are straightforward to compute, requiring only knowledge of the form of $m(z,\beta,h)$ and the kernel. In particular, $\hat{\delta}_i$ can be calculated by analytic differentiation with respect to the scalar ζ. Alternatively, if the analytic formula is very hard to construct, $\hat{\delta}_i$ can be calculated as the numerical derivative of

$\sum_{j=1}^{n} m(z_j,\hat\beta,\hat h + \zeta y_i K_\sigma(\cdot - x_i))/n$ with respect to ζ.

Here $\hat V = \sum_{i=1}^{n}\hat\psi_i\hat\psi_i'/n$ is a "delta-method" variance for kernel estimators. It is exactly analogous to delta method variances for parametric estimators. For example, if $\hat h$ was a sample mean rather than a kernel estimator, say $\hat h = \sum_{i=1}^{n} y_i/n$, then the analog of $\hat\delta_i$ is $\partial\{\sum_{j=1}^{n} m(z_j,\hat\beta,\hat h + \zeta y_i)/n\}/\partial\zeta = \overline{m}_h y_i$, where $\overline{m}_h = n^{-1}\sum_{j=1}^{n}\partial a(z_j,\hat\beta,\hat h)/\partial h$. Thus, the analogous influence function estimator would be $\hat\psi_i = m(z_i,\hat\beta,\hat h) + \overline{m}_h y_i - (\sum_{j=1}^{n}\overline{m}_h y_j/n) = a(z_i,\hat\beta,\hat h) + \overline{m}_h(y_i - \overline{y})$, the usual delta-method formula for the presence of a sample average in an m-estimator. Another feature of $\hat{\mathrm{Var}}(\hat\beta)$ is that it does not rely on the bandwidth shrinking to zero for its validity. If the bandwidth were held fixed it would be a consistent estimator of the asymptotic variance of $\sqrt{n}(\hat\beta - \beta_\sigma)$, where β_σ is the limit of $\hat\beta$ when the bandwidth is held fixed at σ.

The terms $m(z_i,\hat\beta,\hat h)$ and $\sum_{j=1}^{n}\hat\delta_j/n$ are asymptotically negligible in $\hat\psi_i$ when the convergence rate of $\hat\beta$ is slower than $1/\sqrt{n}$. They are retained because they are easy to compute and could conceivably improve the asymptotic approximation. Also, for analogous reasons the formula for $\hat\delta_i$ does not distinguish between elements of $\hat h$ that affect the asymptotic distribution and those that do not (e.g. between pointwise density levels and derivatives, where the slower convergence rate of the derivative will dominate).

Some examples may serve to illustrate the form of this estimator. The simplest example is a density estimator $\hat\beta = \hat f(\overline{x})$ at some $\overline{x}$, where the asymptotic variance estimator is $\hat{\mathrm{Var}}(\hat\beta) = \sum_{i=1}^{n} K_\sigma(\overline{x}-x_i)^2/n - [\sum_{j=1}^{n} K_\sigma(\overline{x}-x_j)/n]^2$, the sample variance of $K_\sigma(\overline{x}-x_i)$. This estimator was recently considered by Hall (1992). Other examples are:

Partial Means: Here $\hat\delta_i$ can be obtained by explicit differentiation of $n^{-1}\sum_{j=1}^{n}\tau(\tilde x_{2j})[\hat h_2(\overline{x}_j) + \zeta q_i K_\sigma(\overline{x}_j - x_i)]/[\hat f(\overline{x}_j) + \zeta K_\sigma(\overline{x}_j - x_i)]$, as

$$\hat\delta_i = n^{-1}\sum_{j=1}^{n}\tau(\tilde x_{2j})\hat f(\overline{x}_j)^{-1}[q_i - \hat f(\overline{x}_j)^{-1}\hat h_2(\overline{x}_j)]K_\sigma(\overline{x}_j - x_i).$$

The asymptotic variance estimator can then be formed as

(3.5) $\hat{Var}(\hat{\beta}) = \sum_{i=1}^{n}\hat{\psi}_i^2/n, \quad \hat{\psi}_i = \tau(\tilde{x}_{2i})\hat{h}_2(\bar{x}_i)/\hat{f}(\bar{x}_i) - \hat{\beta} + \hat{\delta}_i - \sum_{j=1}^{n}\hat{\delta}_j/n.$

Differential Equation Solution: It is possible to derive an analytical expression for $\hat{\delta}_i$, but this expression is quite complicated and difficult to evaluate. An alternative approach that is used by Hausman and Newey (1992), is to numerically differentiate the numerical solution to $dS(p)/dp = -[\hat{h}_2(p,\bar{x}_1-S(p)) + \zeta q_i K_\sigma((p,\bar{x}_1-S(p))-x_i)]/[\hat{f}(p,\bar{x}_1-S(p)) + \zeta K_\sigma((p,\bar{x}_1-S(p))-x_i)]$ with respect to ζ to form $\hat{\delta}_i$. This approach is quite feasible using existing fast and accurate numerical algorithms for ordinary differential equations.

Inverse Density Weighted Least Squares: Let $\hat{u}_i = y_i - x_i'\hat{\beta}$. The variance estimator is

(3.5a) $\hat{Var}(\hat{\beta}) = \hat{M}^{-1}(n^{-1}\sum_{i=1}^{n}\hat{\psi}_i\hat{\psi}_i')\hat{M}^{-1}, \quad \hat{M} = n^{-1}\sum_{i=1}^{n}\hat{f}(x_i)^{-1}\tau(x_i)x_ix_i'$

$$\hat{\psi}_i = \tau(x_i)x_i\hat{u}_i + \hat{\delta}_i - \sum_{j=1}^{n}\hat{\delta}_j, \quad \hat{\delta}_i = -n^{-1}\sum_{j=1}^{n}\tau(x_j)\hat{f}(x_j)^{-2}x_j\hat{u}_jK_\sigma(x_j-x_i).$$

For partial means, asymptotic normality and consistency of the variance estimator are shown in Section 4, using the Lemmas of Section 5. Corresponding theoretical results for the other examples are given elsewhere.

4. Partial Means

Partial means have a number of applications in economics. For example, they can be used to approximate the solution to the differential equation described in Section 2. Dropping the $S(p)$ term from inside $g_0(\bar{x}_1-S(p),p)$ leads to an approximation as $\beta_0 = \int_{p^0}^{p^1} g_0(\bar{x}_1,p)dp = E[(p^1-p^0)g_0(\bar{x}_1,\tilde{x}_2)]$, where $\tilde{x}_2$ is distributed uniformly on $[p^0,p^1]$. It is known that this approximation is quite good in many economic examples, where $S(p)$ is a small proportion of $\bar{x}_1$ (see Willig, 1978). This is a partial mean as described in Section 2. It can be estimated by

(4.1) $\hat{\beta} = (p^1-p^0)\sum_{i=1}^n \hat{g}(\bar{x}_1,p_i)/n,$

where p_i is drawn from a uniform distribution on $[p^0,p^1]$. This is a simulation estimator similar in spirit to that of Lerman and Manski (1981).

Partial means are also of interest from a purely statistical point of view, as dimension attenuation devices. Like $E[q|x_1]$, the partial mean is a function of a smaller dimensional argument. Consequently, partial mean estimators will converge faster than estimators of $g_0(x)$. However, unlike $E[q|x_1]$, a partial mean controls for the covariates x_2, in an average way.

The way in which partial means control for covariates is illustrated by their relationship to additive nonparametric models. Suppose that the conditional expectation takes an additive form, $E[q|x] = g_{10}(x_1) + g_{20}(x_2)$, and that $E[\tau(\tilde{x}_2)] = 1$. Then

(4.2) $E[\tau(\tilde{x}_{2i})g_0(\bar{x}_1,\tilde{x}_{2i})] = g_{10}(\bar{x}_1) + E[\tau(\tilde{x}_{2i})g_{20}(\tilde{x}_{2i})].$

Thus, as a function of $\bar{x}_1$, the partial mean estimates the corresponding component of an additive model, up to a constant.

In comparison with other additive model estimators, partial means are easier to compute but may be less asymptotically efficient. Unlike alternating conditional expectation estimator for additive models (ACE, Breiman and Friedman, 1985), the partial mean is an explicit functional, so the kernel estimator will not require iteration. However, because the partial mean does not impose additivity it may be a less efficient estimator. Also, the partial mean depends on the full conditional expectation, so the curse of dimensionality may result in slower convergence to the limiting distribution.

The partial mean and ACE are different statistical objects when no restrictions are placed on $E[q|x]$. The partial mean is given in equation (2.3a). The ACE object is the mean-square projection of $E[q|x]$ on the set of functions of the form $g_1(x_1) + g_2(x_2)$. These estimators summarize different features of $E[q|x]$. If one is interested in of $\hat{\delta}_i$ is $\partial\{\sum_{j=1}^n m(z_j,\hat{\beta},\hat{h} + \zeta y_i)/n\}/\partial\zeta = \overline{m}_h y_i$, where $\overline{m}_h = n^{-1}\sum_{j=1}^n \partial a(z_j,\hat{\beta},\hat{h})/\partial h$. Thus, the analogous influence function estimator would be $\hat{\psi}_i = m(z_i,\hat{\beta},\hat{h}) + \overline{m}_h y_i - (\sum_{j=1}^n \overline{m}_h y_j/n) = a(z_i,\hat{\beta},\hat{h}) + \overline{m}_h(y_i-\overline{y})$, the usual delta-method formula for the presence of a sample average in an m-estimator. Another feature of $\hat{\text{Var}}(\hat{\beta})$ is that it does not rely on the bandwidth shrinking to zero for its validity. If the bandwidth were held fixed it would be a consistent estimator of the asymptotic variance of $\sqrt{n}(\hat{\beta} - \beta_\sigma)$, where β_σ is the limit of $\hat{\beta}$ when the bandwidth is held fixed at σ.

The terms $m(z_i,\hat{\beta},\hat{h})$ and $\sum_{j=1}^n \hat{\delta}_j/n$ are asymptotically negligible in $\hat{\psi}_i$ when the convergence rate of $\hat{\beta}$ is slower than $1/\sqrt{n}$. They are retained because they are easy to compute and could conceivably improve the asymptotic approximation. Also, for analogous reasons the formula for $\hat{\delta}_i$ does not distinguish between elements of $\hat{h}$ that affect the asymptotic distribution and those that do not (e.g. between pointwise density levels and derivatives, where the slower convergence rate of the derivative will dominate).

Some examples may serve to illustrate the form of this estimator. The simplest example is a density estimator $\hat{\beta} = \hat{f}(\overline{x})$ at some $\overline{x}$, where the asymptotic variance estimator is $\hat{\text{Var}}(\hat{\beta}) = \sum_{i=1}^n K_\sigma(\overline{x}-x_i)^2/n - [\sum_{j=1}^n K_\sigma(\overline{x}-x_j)/n]^2$, the sample variance of $K_\sigma(\overline{x}-x_i)$. This estimator was recently considered by Hall [3]. Other examples are:

Partial Means: Here $\hat{\delta}_i$ can be obtained by explicit differentiation of $n^{-1}\sum_{j=1}^n \tau(\tilde{x}_{2j})[\hat{h}_2(\overline{x}_j) + \zeta q_i K_\sigma(\overline{x}_j-x_i)]/[\hat{f}(\overline{x}_j) + \zeta K_\sigma(\overline{x}_j-x_i)]$, as

$$\hat{\delta}_i = n^{-1}\sum_{j=1}^n \tau(\tilde{x}_{2j})\hat{f}(\overline{x}_j)^{-1}[q_i - \hat{f}(\overline{x}_j)^{-1}\hat{h}_2(\overline{x}_j)]K_\sigma(\overline{x}_j-x_i).$$

The asymptotic variance estimator can then be formed as

$$\hat{\text{Var}}(\hat{\beta}) = \sum_{i=1}^n \hat{\psi}_i^2/n, \quad \hat{\psi}_i = \tau(\tilde{x}_{2i})\hat{h}_2(\overline{x}_i)/\hat{f}(\overline{x}_i) - \hat{\beta} + \hat{\delta}_i - \sum_{j=1}^n \hat{\delta}_j/n. \qquad (10)$$

Differential Equation Solution: It is possible to derive an analytical expression for $\hat{\delta}_i$, but this expression is quite complicated and difficult to evaluate. An alternative normal. Let the u argument of $\mathcal{K}(u)$ be partitioned conformably with x and $\tilde{f}_0(\tilde{x}_2)$ denote the true density of $\tilde{x}_2$. The asymptotic variance of the partial mean estimator will be

$$(4.3) \qquad V = [\int\{\int\mathcal{K}(u_1,u_2)du_2\}^2du_1]\cdot\int f_0(\bar{x}_1,t)^{-1}\tau(t)^2\tilde{f}_0(t)^2\mathrm{Var}(q|x=(\bar{x}_1,t))dt.$$

Theorem 4.1: Suppose that i) $E[|q|^4] < \infty$, $E[|q|^4|x]f_0(x)$ *and* $f_0(x)$ *are bounded; ii) Assumptions H and K are satisfied for* $d \geq s$*; iii)* $\tau(\tilde{x}_2)$ *is bounded and zero except on a compact set where* $f_0(\bar{x}_1,\tilde{x}_2)$ *is bounded away from zero; iv)* $\tau(\tilde{x}_2)$ *and* $\tilde{f}_0(\tilde{x}_2)$ *are continuous a.e.,* $\tilde{f}_0(\tilde{x}_2)$ *is bounded,* $E[q|x]$ *and* $E[q^2|x]$ *are continuous, and for some* $\varepsilon > 0$, $\int\sup_{\|\eta\|\leq\varepsilon}\{1+E[q^4|x=(\bar{x}_1+\eta,x_2)]\}f(\bar{x}_1+\eta,x_2)dx_2 < \infty$*; v)* $n\sigma^{2k-k_1}/\ln(n)^2 \longrightarrow \infty$ *and* $n\sigma^{k_1+2s} \longrightarrow 0$*. Then for* $\hat{\beta}$ *in equation (2,4),* $\sqrt{n}\sigma^{k_1/2}(\hat{\beta} - \beta_0) \xrightarrow{d} N(0,V)$*. If, in addition,* $n\sigma^{3k-k_1} \longrightarrow \infty$*, then* $\sigma^{k_1}\hat{V} \xrightarrow{p} V$.

The conditions here embody "undersmoothing," meaning that the bias goes to zero faster than the variance. Undersmoothing is reflected in the conclusion, where the limiting distribution is centered at zero, rather than at a bias term.

An improved convergence rate for partial means over pointwise estimators is embodied in the normalizing factor $\sqrt{n}\sigma^{k_1/2}$ for the asymptotic distribution. The rate implied by the asymptotic distribution result is $(n\sigma^{k_1})^{-1/2}$ while the corresponding rate from the usual asymptotic normality result for pointwise estimators is $(n\sigma^{k})^{-1/2}$, which converges to zero slower by σ going to zero. Furthermore, the rate for partial means is exactly the nonparametric pointwise rate when the dimension is k_1. Thus, the more components are averaged out, the smaller will be k_1, and hence the faster will be the convergence rate.

One important feature of this result is hypothesis iii), that amounts to a "fixed trimming" condition, where the density of x is bounded away from zero where τ is nonzero. This condition is theoretically convenient because it avoids the "denominator problem." It is used here because it is not restrictive in many cases (e.g. pointwise estimation) and because the resulting theory roughly corresponds to trimming based on a large auxiliary sample, which is often available. It might be possible to modify the results to allow trimming to depend on the sample size, e.g. as in Robinson (1988), but this modification would be very complicated.

Estimators will be $\sqrt{n}$-consistent when they are full means, i.e. are averages over all components. There are many interesting examples of such estimators, such as the policy analysis estimator of Stock (1989). The general conditions given here are slightly different for that case, so it is helpful to describe the estimator and result in a slightly different way. Suppose $\beta_0 = E[a_0(z)g_0(\tilde{x})]$, where $g_0(x) = E[q|x]$, $a_0(z)$ is some function of the data, and $\tilde{x}$ is a continuously distributed variable that may be different than x. A kernel estimator of β_0, along with the associated asymptotic variance estimator from equation (2.6), for $\hat{g}(x) = \hat{h}_2(x)/\hat{h}_1(x)$ as above, is

$$(4.4) \qquad \hat{\beta} = \textstyle\sum_{i=1}^{n} a_0(z_i)\hat{g}(\tilde{x}_i)/n, \quad \hat{\mathrm{Var}}(\hat{\beta}) = \sum_{i=1}^{n}\hat{\psi}_i^2/n, \quad \hat{\psi}_i = a_0(z_i)\hat{g}(\tilde{x}_i) - \hat{\beta} + \hat{\delta}_i - \sum_{j=1}^{n}\hat{\delta}_j/n,$$

$$\hat{\delta}_i = \textstyle\sum_{j=1}^{n} a_0(z_j)\hat{f}(\tilde{x}_j)^{-1}[q_i - \hat{g}(\tilde{x}_j)]K_\sigma(\tilde{x}_j - x_i)/n.$$

The asymptotic variance of this estimator will be

$$(4.5) \qquad V = E[\psi_i^2], \quad \psi_i = a_0(z_i)g_0(\tilde{x}_i) - \beta_0 + E[a_0(z)|\tilde{x}]|_{\tilde{x}=x_i} f_0(x_i)^{-1}\tilde{f}_0(x_i)[q_i - g_0(x_i)].$$

Theorem 4.2: Suppose that i) $E[|q|^4] < \infty$, $E[|q|^4|x]f_0(x)$ and $f_0(x)$ are bounded, and $E[|m_0(z)|^2] < \infty$; ii) Assumptions K and H are satisfied for $d \geq s$; iii) $a_0(z)$ is zero if $\tilde{x}$ is not in a compact set, and $f_0(\tilde{x})$ is bounded away from zero on that compact set; iv) $E[a_0(z)|\tilde{x}]$, and $\tilde{f}_0(\tilde{x})$ are continuous a.e. and bounded for $\tilde{x}$ inside the compact set of iii); v) $n\sigma^{2k}/\ln(n)^2 \longrightarrow \infty$ and $n\sigma^{2s} \longrightarrow 0$. Then $\sqrt{n}(\hat{\beta}-\beta_0)-\sum_{i=1}^{n}\psi_i/\sqrt{n} \xrightarrow{p} 0$ and $\sqrt{n}(\hat{\beta} - \beta_0) \xrightarrow{d} N(0, V)$. If, in addition, $n\sigma^{3k} \longrightarrow \infty$, then $\sum_{i=1}^{n}\|\hat{\psi}_i-\psi_i\|^2/n \xrightarrow{p} 0$ and $\hat{V} \xrightarrow{p} V$.

This result gives asymptotic normality for a trimmed version of Stock's (1989) estimator, as well as being a general result on the asymptotic normality of sample moments that are random linear functions of kernel regressions.

5. Useful Lemmas

Several intermediate results of a familiar type are useful in developing asymptotic theory for the m-estimator $\hat{\beta}$ described in Section 2. Uniform convergence results are useful for showing consistency of $\hat{\beta}$ and of the Jacobian term in the expansion of (3.1). Asymptotic normality of $\hat{\beta}$ will follow from $\sqrt{n}\sigma^{\alpha}\hat{m}_n(\beta_0) \xrightarrow{d} N(0,V)$. Also, $\sigma^{2\alpha}\hat{V} \xrightarrow{p} V$ is very important for consistent estimation of the asymptotic variance. In addition, when $\alpha = 0$, corresponding to $\sqrt{n}$-consistency of $\hat{\beta}$, it can be shown that there is ψ_i such that $\sqrt{n}\hat{m}_n(\beta_0) = \sum_{i=1}^n \psi_i/\sqrt{n} + o_p(1)$ and $\sum_{i=1}^n \|\hat{\psi}_i - \psi_i\|^2/n \xrightarrow{p} 0$. Primitive conditions for each of these results are given in this Section. Examples of how these results can be used to derive results for particular functionals are given in the proofs of Theorems 4.1 and 4.2, and in the proofs of results in Hausman and Newey (1992), Matzkin and Newey (1992), and Newey and Ruud (1991).

A number of additional regularity conditions are used in the analysis to follow. The first regularity conditions imposes some moment assumptions. For a matrix B let $\|B\| = [tr(B'B)]^{1/2}$, where $tr(\cdot)$ denotes the trace of a square matrix.

Assumption Y: For $p \geq 4$, $E[\|y\|^p] < \infty$, $E[\|y\|^p|x]f_0(x)$ is bounded, $E[\|m(z,\beta_0,h_0)\|^2] < \infty$.

This condition, like Assumptions K and H, is a standard type of condition. The fourth moment condition for y is useful for obtaining optimal convergence rates for $\hat{h}$.

For the asymptotic theory, it useful to impose smoothness conditions on $m(z,\beta,h)$ as a function of h, in terms of a metric on the set of possible functions. Here, the metric is the supremum norm on the function and its derivatives, a Sobolev norm. The supremum norm is quite strong, but uniform convergence rates for a kernel estimator and its derivatives are either well known or straightforward to derive (see Appendix B), and are not very much slower than L_p convergence rates (there is only an additional "log term" in the uniform rates). Consequently, conditions for remainder terms to go to zero fast enough to achieve asymptotic normality will not be much stronger with the supremum norm than they will be with L_p norms. Furthermore, it is quite easy to show smoothness in supremum norm for many functionals.

To define the norm, for a matrix of functions $B(x)$ let $\partial^j B(x)/\partial x^j$ denote any vector consisting of all distinct j^{th} order partial derivatives of all elements of $B(x)$. Also, let $\mathcal{X}$ denote a set that is contained in the support of x, and for any nonnegative integer j let

$$\|B\|_j \equiv \max_{\ell \le j} \sup_{x \in \mathcal{X}} \|\partial^\ell B(x)/\partial x^\ell\|,$$

where $\|B\|_j$ is taken equal to infinity if the derivatives do not exist for some $x \in \mathcal{X}$. This is a Sobolev supremum norm of order j.

One useful type of result is uniform convergence in probability, as in the conclusion of the following result. Let $m_0(\beta) = E[m(z,\beta,h_0)]$.

Lemma 5.1: Suppose that i) $m(z,\beta,h_0)$ is continuous at each $\beta \in \mathcal{B}$ with probability one, where $\mathcal{B}$ is compact, and $E[\sup_{\beta\in\mathcal{B}}\|m(z,\beta,h_0)\|] < \infty$; ii) Assumptions K, H, and Y are satisfied with $d \ge \Delta+1$, $\ln(n)/(n\sigma^{k+2\Delta}) \longrightarrow 0$ and $\sigma \longrightarrow 0$, and there is $b(z)$ and $\varepsilon > 0$ such that $E[b(z)] < \infty$, and for all $\beta \in \mathcal{B}$ and $\|h-h_0\|_\Delta < \varepsilon$, $\|m(z,\beta,h)-m(z,\beta,h_0)\| \le b(z)(\|h-h_0\|_\Delta)^\varepsilon$. Then $E[m(z,\beta,h_0)]$ is continuous on $\mathcal{B}$ and

$$(5.1) \qquad \sup_{\beta\in\mathcal{B}}\|n^{-1}\textstyle\sum_{i=1}^n m(z,\beta,\hat{h}) - E[m(z,\beta,h_0)]\| \xrightarrow{p} 0.$$

The uniform convergence conclusion of equation (5.1) is a well known condition for consistency of the solution to equation (2.3). Also, equation (5.1) is useful in showing consistency of an estimator that maximizes an objective function $n^{-1}\sum_{i=1}^n m(z_i,\beta,\hat{h})$, where m is a scalar, and is useful for showing consistency of the Jacobian term $n^{-1}\sum_{i=1}^{n}\partial m(z_i,\beta,\hat{h})/\partial\beta$, by letting the m in the statement of the Lemma be each column of of the derivative.

Asymptotic normality of $\sqrt{n}\sigma^{\alpha}\hat{m}_n(\beta_0)$ is essential for asymptotic normality of $\hat{\beta}$. This result has two components, which are a linearization around the true h_0 and asymptotic normality of the linearization. It is useful to state these two components separately.

Asymptotic normality of the linearization will follow from asymptotic normality of $\sqrt{n}\sigma^{\alpha}\hat{m}_n(\beta_0)$ when $m(z,\beta_0,h)$ is a linear functional that does not depend on z, say $m(h) = m(z,\beta_0,h)$. The rate of convergence (i.e. the magnitude of α) will depend on the nature of $m(h)$. Here the results are grouped into two main ones, the first involving $\sqrt{n}$-consistency. For the moment, assume that $m(h)$ is a scalar

Lemma 5.2: If $m(h) = \int \nu(x)'h(x)dx$ *where* $\nu(x)$ *is zero outside a compact set, continuous almost everywhere, there is* $\varepsilon > 0$ *such that* $E[\sup_{\|u\|\le\varepsilon}\|\nu(x+u)\|^2 E[\|y\|^2|x]] < \infty$, *and* $\sqrt{n}\sigma^{s} \longrightarrow 0$, *then for* $\delta_i = \nu(x_i)'y_i$, $\sqrt{n}[m(\hat{h}) - m(h_0)] = \sum_{i=1}^{n}\{\delta_i - E[\delta_i]\}/\sqrt{n} + o_p(1)$.

Cases where convergence is slower than $1/\sqrt{n}$ are somewhat more complicated. The following assumption is useful for these cases. For the moment let ℓ be a nonnegative integer and let $\partial^j h(x)/\partial x^j$ be ordered so that $\partial^{\ell}[y_i K_{\sigma}(x-x_i)]/\partial x^{\ell} = y\otimes[\partial^{\ell}K_{\sigma}(x-x_i)/\partial x^{\ell}]$.

Assumption 5.1: Suppose that $k = k_1 + k_2$, there is a matrix of functions $\omega(t)$ with domain $\mathbb{R}^{k_2}$, $0 \le k_2 < k$, a vector of functions $x_1(t)$ in $\mathbb{R}^{k_1}$, such that i) $m(h) = \int \omega(t)[\partial^{\ell} h(x(t))/\partial x^{\ell}]dt$ for $x(t) = (x_1(t)', t')'$; ii) $\omega(t)$ is bounded and continuous almost everywhere and zero outside a compact set $\mathcal{T}$, and $x_1(t)$ is continuously differentiable with bounded partial derivatives on a convex, compact set $\tilde{\mathcal{T}}$ containing $\mathcal{T}$ in its interior; iii) $\Sigma(x) = E[yy'|x]$ is continuous a.e., and for $\varepsilon > 0$ and $v(x) = E[\|y\|^4|x]$, $\int_{\tilde{\mathcal{T}}} \sup_{\|\eta\| \le \varepsilon} [\{1 + v(x_1(t)+\eta, t)\} f_0(x_1(t)+\eta, t)]dt < \infty$.

The key condition here is the integral representation of $m(h)$. The dimension of the argument being integrated and the order of the derivative lead to the convergence rate for $m(\hat{h})$, that is $\sqrt{n}\sigma^{k_1/2 + \ell}$. Thus, every additional dimension of integration increases the convergence rate by $\sqrt{\sigma}$ while every additional derivative decreases the rate by a factor of $1/\sigma$. This hypothesis also leads to a specific form for the asymptotic variance of $m(\hat{h})$, which for $\tilde{\mathcal{K}}(u_1,t) = \int \partial^{\ell}\mathcal{K}(u+[\partial x_1(t)/\partial t]v, v)/\partial u^{\ell} dv$ is

$$(5.2) \qquad V = \int \omega(t)[\Sigma(x(t)) \otimes \{\int \tilde{\mathcal{K}}(u_1,t)\tilde{\mathcal{K}}(u_1,t)' du_1\} \omega(t)' f_0(x(t))dt.$$

Lemma 5.3: If Assumptions K, H, Y, and 5.1 is satisfied with $d \ge \ell + s$ and for $\alpha = k_1/2 + \ell$, $\sqrt{n}\sigma^{k_1/2} \longrightarrow \infty$, and $\sqrt{n}\sigma^{\alpha+s} \longrightarrow 0$ then $\sqrt{n}\sigma^{\alpha}[m(\hat{h}) - m(h_0)] \xrightarrow{d} N(0,V)$.

Asymptotic normality in the more general case where $m(z, \beta_0, h)$ depends on z and is nonlinear in h can be reduced to the previous cases by a linearization. The following assumption is useful for the linearization. Let $m(z,h) = m(z,\beta_0,h)$, and again assume that this is a scalar.

Lemma 5.4: Suppose that Assumptions K, H, and Y are satisfied, $\mathcal{X}$ *is compact, there is a vector of functionals* $D(z,h)$, *and nonnegative constants* α, $\Delta_\ell \le \Delta$, $(\ell = 1, 2)$, $\varepsilon > 0$ such *that* $d \ge \max\{\Delta+1, \Delta_1+s, \Delta_2+s\}$ *and i)* $D(z,h)$ *is linear in* h *on* $\{h : \|h\|_\Delta < \infty\}$; ii) *for all* h *with* $\|h-h_0\|_\Delta < \varepsilon$, $\|m(z,h)-m(z,h_0)-D(z,h-h_0)\| \le b(z)\|h-h_0\|_{\Delta_1}\|h-h_0\|_{\Delta_2}$; iii) $\|D(z,h)\| \le \tilde{b}(z)\|h\|_{\Delta_1}$ *and* $E[\tilde{b}(z)^4] < \infty$; *iv) for* $\eta_n^j = [\ln(n)/(n\sigma^{k+2j})]^{1/2} + \sigma^s$, $\eta_n^\Delta \longrightarrow 0$, $\sqrt{n}\sigma^\alpha E[b(z)]\eta_n^{\Delta_1}\cdot\eta_n^{\Delta_2} \longrightarrow 0$ *and* $\sqrt{n}\sigma^{k+\Delta_1-\alpha} \longrightarrow \infty$. *Then for* $m(h) = \int D(z,h)dF(z)$,

$$\sqrt{n}\sigma^\alpha\textstyle\sum_{i=1}^n[m(z_i,\hat{h})-m(z_i,h_0)]/n = \sqrt{n}\sigma^\sigma[m(\hat{h})-m(h_0)] + o_p(1).$$

The conditions of this result imply Frechet differentiability at h_0 of $m(z,h)$ as a function of h, in the Sobolev norm $\|h\|_{\max\{\Delta_1,\Delta_2\}}$. The remainder bounds are formulated with different norms, rather than $\Delta = \Delta_1 = \Delta_2$, to allow weaker conditions for asymptotic normality in some cases.

Asymptotic normality of $\sigma^\alpha\sum_{i=1}^n m(z_i,\hat{h})/\sqrt{n}$ can be shown by combining Lemma 5.4 with either Lemma 5.2 or 5.3. In the $\sqrt{n}$-consistent case of Lemma 5.2, it will follow from Lemmas 5.2 and 5.3 that $\sum_{i=1}^n m(z_i,\hat{h})/\sqrt{n} = \sum_{i=1}^n\{m(z_i,h_0) + \delta_i - E[\delta_i]\}/\sqrt{n} + o_p(1)$, so that asymptotic normality, with asymptotic variance $\mathrm{Var}(m(z_i,h_0)+\delta_i)$ follows by the central limit theorem. In the slower than $\sqrt{n}$-consistent case, where $m(h) = \int D(z,h)dF(z)$ satisfies the conditions of Lemma 5.3 and $\alpha > 0$, it will be the case that $\sigma^\alpha\sum_{i=1}^n m(z_i,h_0)/\sqrt{n} \xrightarrow{p} 0$, so that $\sigma^\alpha\sum_{i=1}^n m(z_i,\hat{h})/\sqrt{n} \longrightarrow N(0,V)$.

Assumption 5.2: i) $\|m(z,\beta,\tilde{h})-m(z,\beta_0,h_0)\| \le b(z)[\|\beta-\beta_0\|^{\varepsilon} + (\|\tilde{h}-h_0\|_{\Delta})^{\varepsilon}]$ and $E[b(z)^2] < \infty$; ii) For $\varepsilon > 0$ and $\|\beta-\beta_0\| < \varepsilon$ and $\|\tilde{h}-h_0\|_{\Delta} < \varepsilon$, there is $D(z,h;\beta,\tilde{h})$ that is linear on $\|h\|_{\Delta} < \infty$ satisfying $|m(z,\beta,h)-m(z,\beta,\tilde{h})-D(z,h-\tilde{h};\beta,\tilde{h})| = o(\|h-\tilde{h}\|_{\Delta})$ as $\|h-\tilde{h}\|_{\Delta} \longrightarrow 0$ for fixed β and $\tilde{h}$; iii) $\|D(z,h;\beta,\tilde{h})-D(z,h;\beta_0,h_0)\| \le b(z)\|h\|_{\Delta_1}(\|\beta-\beta_0\| + \|\tilde{h}-h_0\|_{\Delta_2})$ and $\|D(z,h;\beta_0,h_0)\| \le \tilde{b}(z)\|h\|_{\Delta_3}$ and $E[\tilde{b}(z)^4] < \infty$; iv) $\hat{\beta} = \beta_0 + O_p(\delta_{\beta n})$, $\delta_{\beta n} \longrightarrow 0$, $\sigma^{\alpha-k-\Delta_1}\cdot\delta_{\beta n} \longrightarrow 0$, $\alpha+s > k+\Delta_1$, $n\sigma^{3k+2\Delta_1+2\Delta_2-2\alpha}/\ln(n) \longrightarrow \infty$, $n\sigma^{2k+2\Delta_3-2\alpha} \longrightarrow \infty$.

Lemma 5.5: Suppose that Assumption 5.2 is satisfied. If $m(h) = \int D(z,h;\beta_0,h_0)dF(z)$ *satisfies the conditions of Lemma 5.2 then, for* $\delta_i = v(x_i)y_i$, $\sum_{i=1}^{n}\|\hat{\delta}_i-\delta_i\|^2/n \xrightarrow{p} 0$ *and* $\hat{V} \xrightarrow{p} V = Var(\delta_i)$. *If* $m(h) = \int D(z,h;\beta_0,h_0)dF(z)$ *satisfies the conditions of Lemma 5.4,* $\sigma^{2\alpha}\hat{V} \xrightarrow{p} V$, *for* V *in equation (5.2).*

Throughout the appendix C will denote a generic constant that may be different in different uses and $\sum_i = \sum_{i=1}^n$. Also, CS, M, and T will refer to the Cauchy-Schwartz, Markov, and triangle inequalities, respectively, and DCT to the dominated convergence theorem. Before proving the results in the body of the paper it is useful to state and prove some intermediate results.

Proof of Theorem 4.1: The proof proceeds by checking the conditions of Lemmas 5.3 - 5.5. Let $\bar{x} = (\bar{x}_1,\tilde{x}_2)$, $\tau(\bar{x}) = \tau(\tilde{x}_2)$, $\mathcal{X}$ be the compact set of hypothesis iii), and $\|h\| = \|h\|_0 = \sup_{x\in\mathcal{X}}\|h(x)\|$. Let $m(z,h) = \tau(\bar{x})h_2(\bar{x})/h_1(\bar{x})$, $D(z,h;\tilde{h}) = \tau(\bar{x})\tilde{h}_1(\bar{x})^{-1}[h_2(\bar{x}) - \{\tilde{h}_2(\bar{x})/\tilde{h}_1(\bar{x})\}h_1(\bar{x})]$, and $D(z,h) = D(z,h;h_0)$. Choose ε small enough that $h_{10}(\bar{x})$ is bounded below by ε for all $\bar{x} \in \mathcal{X}$. Then for $\|\tilde{h}_1-h_{10}\| < \varepsilon$, $|m(z,\beta,h) - m(z,\beta,\tilde{h}) - D(z,h-\tilde{h};\tilde{h})| = |[h_1(\bar{x})^{-1}\tilde{h}_1(\bar{x}) - 1]D(z,h-\tilde{h};\tilde{h})| \le C\|h-\tilde{h}\|^2$ and $|D(z,h;\tilde{h})| \le C\|h\|$. Let $\alpha = k_1/2$. Then for $\eta_n = [\ln(n)/(n\sigma^k)]^{1/2} + \sigma^s$, $\sqrt{n}\sigma^\alpha\eta_n^2 = \ln(n)\sigma^{\alpha-k}/\sqrt{n} + 2[\ln(n)]^{1/2}\sigma^{\alpha+s-k/2} + \sqrt{n}\sigma^{2s+\alpha} \longrightarrow 0$ by $\ln(n)\sigma^{\alpha-k}/\sqrt{n} \longrightarrow 0$, $\sqrt{n}\sigma^{2s+\alpha} \longrightarrow 0$, implying σ goes to zero faster than some power of n, and by $\alpha+s > k/2$. Also, $\ln(n)\sigma^{\alpha-k}/\sqrt{n} \longrightarrow 0$ implies that $\sqrt{n}\sigma^{k-\alpha} \longrightarrow \infty$, so that the rate hypotheses of Lemma 5.4 are satisfied. Thus, the conclusion of Lemma 5.4 holds, with $m(h) = \int D(z,h;h_0)dF(z) = \int\tau(x(t))f_0(x(t))^{-1}[h_2(x(t)) - g_0(x(t))h_1(x(t))]\tilde{f}_0(t)dt$, for $t = \tilde{x}_2$ and $x(t) = (\bar{x}_1,t)$. Let $\omega(t) = \tau(x(t))f_0(x(t))^{-1}\tilde{f}_0(t)[-g_0(x(t)),1]$. This function is bounded and continuous a.e. and zero outside a compact set by continuity of f_0, $\tilde{f}_0$, and g_0, and by the assumption about τ. The other conditions of Assumption 5.1 are also satisfied by hypothesis. Furthermore, $\sqrt{n}\sigma^{\alpha+s} \longrightarrow 0$ and $\sqrt{n}\sigma^{k_1/2} = \sqrt{n}\sigma^\alpha \longrightarrow \infty$ by hypothesis and $\alpha \le k-\alpha = k_2 + k_1/2$. Thus, the conclusion of Lemma 5.3 holds, for V in equation (4.3). Then by the triangle inequality, and $\beta_0 = E[m(z,h_0)]$

$$\sqrt{n}\sigma^\alpha(\hat{\beta}-\beta_0) = \sqrt{n}\sigma^\alpha\sum_i\{m(z_i,\hat{h})-E[m(z,h_0)]\}/n$$

$= \sqrt{n}\sigma^{\alpha}\sum_i\{m(z_i,h_0)-E[m(z,h_0)]\}/n + \sqrt{n}\sigma^{\sigma}[m(\hat{h})-m(h_0)] + o_p(1) \xrightarrow{d} N(0,V),$

because $\sqrt{n}\sigma^{\alpha}\sum_i\{m(z_i,h_0)-E[m(z,h_0)]\}/n \xrightarrow{p} 0$ by $\sigma \longrightarrow 0$.

To finish the proof, note that it follows from the above arguments and by hypothesis that for $m(z,h,\beta) = m(z,h)$ and $D(z,h;\beta,\tilde{h}) = D(z,h;\tilde{h})$, as specified above, conditions i) - iii) of Assumption 5.2 are satisfied, with $\Delta = \Delta_1 = \Delta_2 = \Delta_3 = 0$. Furthermore, condition iv) is satisfied by $\delta_{\beta n} = 0$, $n\sigma^{3k-2\alpha}/\ln(n) \longrightarrow \infty$, and the fact that this last condition implies $s > 3k/2 - 2\alpha = 3k_2/2 + k_1/2 > k_2 + k_1/2 = k-\alpha$. The second conclusion then follows by the conclusion of Lemma 5.5. QED.

Proof of Theorem 4.2: Let $m(z,h) = a_0(z)h_1(\tilde{x})^{-1}h_2(\tilde{x})$ and $D(z,h;\tilde{h}) = a_0(z)\tilde{h}_1(\tilde{x})^{-1}[h_2(\tilde{x}) - \{\tilde{h}_2(\tilde{x})/\tilde{h}_1(\tilde{x})\}h_1(\tilde{x})]$. The proof that the conditions of Lemma 5.4 are satisfied proceeds exactly as in the proof of Theorem 4.1, except that $\alpha = 0$ and the function $b(z)$ of Lemma 5.4 is taken to be $\|a_0(z)\|$. Also, here $m(h) = \int D(z,h;h_0)dF(z) = E[a_0(z)f_0(\tilde{x})^{-1}\{h_2(\tilde{x}) - g_0(\tilde{x})h_1(\tilde{x})\}] = E[E[a_0(z)|\tilde{x}]f_0(\tilde{x})^{-1}\{h_2(\tilde{x}) - g_0(\tilde{x})h_1(\tilde{x})\}] = \int\nu(x)h(x)dx$ for $\nu(x) = E[a_0(z)|\tilde{x}]|_{\tilde{x}=x}f_0(x)^{-1}\tilde{f}_0(x)(-g_0(x),1)$. By hypothesis, the conditions of Lemma 5.3 are satisfied for this $\nu(x)$, so that by the conclusion of Lemma 5.3, for $\delta_i = \nu(x_i)y_i = \psi_i - a_0(z_i)g_0(\tilde{x}_i)+\beta_0$, one obtains $\sqrt{n}[m(\hat{h})-m(h_0)] = \sum_i\{\delta_i-E[\delta_i]\}/\sqrt{n} + o_p(1)$. The first conclusion then follows. Also, the second conclusion follows from Lemma 5.5 similarly to the proof of Theorem 4.1. QED.

Proof of Lemma 5.1: It follows by standard results (e.g. Tauchen, 1985) that $\sup_{\beta\in\mathcal{B}}\|n^{-1}\sum_i m(z_i,\beta,h_0) - E[m(z,\beta,h_0)]\| \xrightarrow{p} 0$ and $E[m(z,\beta,h_0)]$ is continuous in β. Also, by Theorem B.2, $\|\hat{h}-h_0\|_\Delta = O_p(\ln(n)^{1/2}(n\sigma^{k+2\Delta})^{-1/2} + \sigma) = o_p(1)$. Therefore, $\sup_{\beta\in\mathcal{B}}\|n^{-1}\sum_i[m(z_i,\beta,\hat{h}) - m(z_i,\beta,h_0)]\| \le n^{-1}\sum_i b(z_i)(\|\hat{h}-h_0\|_\Delta)^{\varepsilon} \xrightarrow{p} 0$ so the conclusion follows by T. QED.

Proof of Lemma 5.2: By the Fubini theorem, $E[m(\hat{h})] = m(E[\hat{h}])$. Also, by standard results, $\sup_{x\in\mathcal{C}}\|E[\hat{h}](x) - h_0(x)\| = O(\sigma^s)$ for any compact set $\mathcal{C}$. Then by $\nu(x)$ zero outside a compact set $\mathcal{C}$, $\sqrt{n}[E[m(\hat{h})] - m(h_0)] \le \sqrt{n}C\sup_{x\in\mathcal{C}}\|E[\hat{h}](x) - h_0(x)\| = O(\sqrt{n}\sigma^s) \longrightarrow 0$. Let $\delta_i^\sigma = [\int\nu(x)K_\sigma(x-x_i)dx]y_i = [\int\nu(x_i+\sigma u)\mathcal{K}(u)du]y_i$, where the last equality follows by a change of variables $u = (x-x_i)/\sigma$, so that $m(\hat{h}) = \sum_{i=1}^n\delta_i^\sigma/n$. By $\mathcal{K}(u)$ having bounded support, $\nu(x_i+\sigma u)\mathcal{K}(u) \le b(x_i)|\mathcal{K}(u)|$ for all small enough σ and $\int b(x_i)|\mathcal{K}(u)|du < \infty$. Then by DCT, $\delta_i^\sigma \longrightarrow \delta_i$ with probability one as $\sigma \longrightarrow 0$. Also, $|\delta_i^\sigma| \le Cb(x)\|y\|$, so by DCT $E[|\varepsilon_i^\sigma|^2] \longrightarrow 0$ for $\varepsilon_i^\sigma = \delta_i^\sigma-\delta_i$. Then by M, $\sum_i\{\varepsilon_i^\sigma-E[\varepsilon_i^\sigma]\}/\sqrt{n} \xrightarrow{p} 0$, so $\sqrt{n}[m(\hat{h})-m(h_0)] = \sqrt{n}\{m(\hat{h})-E[m(\hat{h})]\} + o(1) = \sum_i\{\delta_i-E[\delta_i]\}/\sqrt{n} + \sum_i\{\varepsilon_i^\sigma-E[\varepsilon_i^\sigma]\}/\sqrt{n} + o(1) = \sum_i\{\delta_i-E[\delta_i]\}/\sqrt{n} + o_p(1)$. QED.

Proof of Lemma 5.3: Note that $E[m(\hat{h})] = m(E[\hat{h}])$, so by $\omega(t)$ bounded and zero outside $\mathcal{T}$, and by $x(t)$ bounded on $\mathcal{T}$, it follows that $\sqrt{n}\sigma^\alpha\{E[m(\hat{h})] - m(h_0)\} \le \sqrt{n}\sigma^\alpha\|E[\hat{h}]-h_0\|_\ell = O(\sqrt{n}\sigma^{\alpha+s}) \longrightarrow 0$. Therefore, it suffices to show that $\sqrt{n}\sigma^\alpha\{m(\hat{h})-E[m(\hat{h})]\} \xrightarrow{d} N(0,V)$. Let $\mathcal{K}^\ell(u)$ denote $\partial^\ell\mathcal{K}(u)/\partial u^\ell$ and $\rho_\sigma(x) = \sigma^{-k-\ell}\int\omega(t)[I\otimes\mathcal{K}^\ell((x(t)-x)/\sigma)]dt = \sigma^{-k_1-\ell}\int\omega(x_2+\sigma v)[I\otimes\mathcal{K}^\ell((x_1(x_2+\sigma v)-x_1)/\sigma,v)]dv$, where I is an identity matrix with the same dimension as y and the last equality follows by a the change of variables $v = (t-x_2)/\sigma$. Then $m(\hat{h}) = \sum_i\rho_\sigma(x_i)y_i/n$. Thus, to show $\sqrt{n}\sigma^\alpha\{m(\hat{h})-E[m(\hat{h})]\} \xrightarrow{d} N(0,V)$ it suffices, by the Liapunov central limit theorem, to show that $\sigma^{2\alpha}\mathrm{Var}(\rho_\sigma(x_i)y_i) \longrightarrow V$ and $\sigma^{4\alpha}E[\|\rho_\sigma(x_i)y_i\|^4]/n \longrightarrow 0$. By i.i.d. data and $\sqrt{n} \longrightarrow \infty$, $\sigma^\alpha\|E[\rho_\sigma(x_i)y_i]-m(h_0)\| = \sigma^\alpha\|E[m(\hat{h})] - m(h_0)\| \longrightarrow 0$, and hence $\sigma^\alpha\|E[\rho_\sigma(x_i)y_i]\| \longrightarrow 0$. Therefore, to show $\sigma^{2\alpha}\mathrm{Var}(\rho_\sigma(x_i)y_i) \longrightarrow V$ it suffices to show that $\sigma^{2\alpha}E[\rho_\sigma(x_i)y_iy_i'\rho_\sigma(x_i)'] \longrightarrow V$. By $\mathcal{K}(u)$ having bounded support, $\mathcal{K}(u_1,v)$ is zero for all v outside a bounded set $\mathcal{V}$. Let $\bar{\mathcal{T}}$ be a compact, convex set containing $\mathcal{T}$ in its interior. Then, for small enough σ, if $x_2 \notin \bar{\mathcal{T}}$ then $x_2+\sigma v \notin \mathcal{T}$ for all $v \in \mathcal{V}$, so $\rho_\sigma(x)$ is zero for $x_2 \notin \bar{\mathcal{T}}$. For $x_2 \in \bar{\mathcal{T}}$ and $x_2+\sigma v \in \bar{\mathcal{T}}$, continuous differentiability of $x_1(t)$ and a mean value expansion give $[x_1(x_2+\sigma v)-x_1(x_2)]/\sigma = [\partial x_1(x_2+\bar{\sigma}v)/\partial t]v$, which is bounded over for $v \in \mathcal{V}$ and converges to $J(x_2)v$ as $\sigma \longrightarrow 0$. Therefore, $\sigma^{k_1+\ell}\rho_\sigma(x_1(x_2)-\sigma u,x_2) = \int\omega(x_2+\sigma v)[I\otimes\mathcal{K}^\ell([x_1(x_2+\sigma v)-x_1(x_2)]/\sigma + u,v)]dv$ is zero for all u outside a compact set, is bounded, and converges to $\int\omega(x_2)[I\otimes\mathcal{K}^\ell(J(x_2)v + u,v)]dv$ by the dominated convergence theorem, for $J(t) = \partial x_1(t)/\partial t$. Then by the change of variables $u = [x_1(x_2)-x_1]/\sigma$ and $t = x_2$,

(A.6) $$\sigma^{k_1+2\ell}E[\rho_\sigma(x_i)y_iy_i'\rho_\sigma(x_i)']$$

$$= \sigma^{2k_1+2\ell}\int\rho_\sigma(x_1(t)-\sigma u,t)\Sigma(x_1(t)-\sigma u,t)\rho_\sigma(x_1(t)-\sigma u,t)'f_0(x_1(t)-\sigma u,t)dtdu \longrightarrow V.$$

Also, $\sigma^{4\alpha}E[\|\rho_\sigma(x_i)y_i\|^4]/n \le \sigma^{4\alpha}E[\|\rho_\sigma(x_i)\|^4\|y_i\|^4]/n \le \sigma^{4\alpha}E[\|\rho_\sigma(x_i)\|^4\{1+\nu(x_i)\}] =$ $C\sigma^{3k_1+4\ell}\int_{\tilde{\mathcal{J}}}\|\rho_\sigma(x_1(t)-\sigma u,t)\|^4\{1+\nu(x_1(t)-\sigma u,t)\}f_0(x_1(t)-\sigma u,t)dtdu/n \le C/(n\sigma^{k_1}) \longrightarrow 0$. The conclusion then follows by the Liapunov central limit theorem. QED.

Proof of Lemma 5.4: By $d \ge \Delta + 1$ and Lemma B.3, $\|\hat{h}-h_0\|_\Delta \xrightarrow{p} 0$. Then by hypothesis iv) $\|\sigma^\alpha n^{-1/2}\sum_i\{m(z_i,\hat{h})-m(z_i,h_0)-D(z_i,\hat{h}-h_0)\}\| \le \sigma^\alpha\sqrt{n}[\sum_i b(z_i)/n]\|\hat{h}-h_0\|_{\Delta_1}\|\hat{h}-h_0\|_{\Delta_2} =$ $O_p(\sqrt{n}\sigma^\alpha E[b(z)]\eta_n^{\Delta_1}\cdot\eta_n^{\Delta_2}) = o_p(1)$. By linearity of $D(z,h)$, $\sum_i D(z_i,\hat{h})/n = \sum_{ij}D_{ij}/n^2$ for $D_{ij} = D(z_i,y_jK_\sigma(\cdot-x_j))$ and $\sum_{ij} = \sum_{i=1}^n\sum_{j=1}^n$. Let $D_{\cdot i} = E[D_{ji}|z_i]$ and $D_{i\cdot} = E[D_{ij}|z_i]$ for $j \ne i$. Note that $E[D_{ii}^2] \le E[b(z)^2\|y\|^2]\sigma^{-2k-2\Delta_1}$ and $E[D_{ij}^2] \le$ $E[b(z)^2]E[\|y\|^2]\sigma^{-2k-2\Delta_1}$. Then by a V-statistic projection on the basic observations (e.g. Serfling, Lemma 5.2.2b), $\sqrt{n}\sigma^\alpha|n^{-2}\sum_{ij}\{D_{ij}-D_{i\cdot}-D_{\cdot i}+E[D_{i\cdot}]\}| =$ $O_p(\sqrt{n}\sigma^\alpha\{(E[D_{ii}^2])^{1/2}+(E[D_{ij}^2])^{1/2}\}/n) = O_p(\sigma^{\alpha-k-\Delta_1}/\sqrt{n}) \xrightarrow{p} 0$. By Lemma B.4, $D_{i\cdot} =$ $D(z_i,\bar{h})$. By $\sigma \longrightarrow 0$, $\|\bar{h}-h_0\|_{\Delta_1} \longrightarrow 0$. Then $E[D(z_i,\bar{h}-h_0)^2] \le E[\tilde{b}(z)^2](\|\bar{h}-h_0\|_{\Delta_1})^2 \longrightarrow 0$. Thus, by Chebyshev's inequality, $\sigma^\alpha\sum_i\{D_{i\cdot}-E[D_{i\cdot}]-D(z_i,h_0)+E[D(z_i,h_0)]\}/\sqrt{n} \xrightarrow{p} 0$. Then the conclusion follows by T. QED.

Proof of Lemma 5.5: It follows by a standard argument, similar to the proof of Lemma 5.1, that $\sum_{i=1}^n\|m(z_i,\hat{\beta},\hat{h})-m(z_i,\beta_0,h_0)\|^2/n \xrightarrow{p} 0$. Let $\hat{D}_{ij} = D(z_i,y_jK_\sigma(\cdot-x_j);\hat{\beta},\hat{h})$ and $D_{ij} = D(z_i,y_jK_\sigma(\cdot-x_j);\beta_0,h_0)$, $\tilde{\delta}_i = \sum_{j=1}^n D_{ji}/n$, and $\bar{\delta}_i = E[D_{ji}|z_i] =$ $\int D(z,,y_jK_\sigma(\cdot-x_j);\beta_0,h_0)dF(z)$. By Lemma B.5, $\hat{\delta}_i = \sum_{j=1}^n\hat{D}_{ji}/n$. Also, by Assumption 5.2, $\|\hat{D}_{ij}-D_{ij}\| \le b(z_i)\|y_iK_\sigma(\cdot-x_i)\|_{\Delta_1}(\|\hat{\beta}-\beta_0\| + \|\hat{h}-h_0\|_{\Delta_2}) \le b(z_i)\|y_j\|\sigma^{-k-\Delta_1}O_p(\delta_{\beta n}+\eta_n^{\Delta_2})$. Then by CS,

$$\sigma^{2\alpha}\sum_{i=1}^{n}\|\hat{\delta}_i-\tilde{\delta}_i\|^2/n \le C\sigma^{2\alpha}\sum_{ij}\|\hat{D}_{ij}-D_{ij}\|^2/n^2 \le C\sigma^{2\alpha}[\sum_i b(z_i)^2/n](\sum_i\|y_i\|^2/n) \times$$

$$\sigma^{-2k-2\Delta_1}(\|\hat{\beta}-\beta_0\| + \|\hat{h}-h_0\|_{\Delta_2})^2 = O_p([\sigma^{\alpha-k-\Delta_1}(\delta_{\beta n}+\eta_n^{\Delta_2})]^2) = o_p(1).$$

By the data i.i.d.,

$$E[\sigma^{2\alpha}\sum_i\|\tilde{\delta}_i-\bar{\delta}_i\|^2/n] \le \sigma^{2\alpha}E[\|\tilde{\delta}_1-\delta_1\|^2] \le C\sigma^{2\alpha}(E[\|n^{-1}\sum_{i\ne 1}(D_{i1}-\delta_1)\|^2] + n^{-1}E[\|D_{11}\|^2]$$

$$+ n^{-1}E[\|\delta_1\|^2] \le C\sigma^{2\alpha}n^{-1}(E[\|D_{21}\|^2] + E[\|D_{11}\|^2])$$

$$\le C\sigma^{2\alpha}n^{-1}\sigma^{-2k-2\Delta_3} \longrightarrow 0,$$

so by M, $\sigma^{2\alpha}\sum_i\|\tilde{\delta}_i-\bar{\delta}_i\|^2/n \xrightarrow{p} 0$. Under the conditions of lemma 5.2, it was shown in the proof of Lemma 5.2 that $E[\|\bar{\delta}_i-\delta_i\|^2] \longrightarrow 0$, so that $\sum_i\|\bar{\delta}_i-\delta_i\|^2/n \xrightarrow{p} 0$ follows by T. Then $\hat{V} \xrightarrow{p} V$ follows by T and the law of large numbers. Under the conditions of Lemma 5.3 note that $\bar{\delta}_i = \rho_\sigma(x_i)y_i$ for $\rho_\sigma(x)$ defined in the proof of Lemma 5.3. As shown in that proof, $\sigma^{\alpha}E[\bar{\delta}_i] \longrightarrow 0$, $\sigma^{2\alpha}E[\bar{\delta}_i\bar{\delta}_i'] \longrightarrow V$, and $n^{-1}\sigma^{4\alpha}E[\|\bar{\delta}_i\|^4] \longrightarrow 0$. Therefore, by M, $\sigma^{2\alpha}n^{-1}\sum_i(\bar{\delta}_i-\sum_j\bar{\delta}_j/n)(\bar{\delta}_i-\sum_j\bar{\delta}_j/n)' \xrightarrow{p} V$, so the conclusion follows by T. QED.

This appendix derives rates for uniform convergence in probability in Sobolev norms for derivatives of kernel estimators. Recall from the text that for a closed set $\mathcal{X}$, $\|h\|_j = \sup_{\ell\leq j}\sup_{x\in\mathcal{X}}\|\partial^\ell h(x)/\partial x^\ell\|$.

Lemma B.1: Suppose that $E[\|y\|^p] < \infty$ *for* $p > 2$, $E[\|y\|^p|x]f_0(x)$ *is bounded,* $\mathcal{X}$ *is compact, Assumption K is satisfied for* $\Delta \geq j$, *and* $\sigma = \sigma(n)$ *such that* $\sigma(n)$ *is bounded and* $n^{1-(2/p)}\sigma(n)^k/ln(n) \longrightarrow \infty$. *Then*

$$\text{(B.1)} \qquad \|\hat{h}-E[\hat{h}]\|_j = O_p(ln(n)^{1/2}(n\sigma^{k+2j})^{-1/2}).$$

Proof: It suffices to prove the result for y a scalar. For each $\ell \leq j$, by $\mathcal{K}(u)$ having bounded support the order of differentiation and integration can be interchanged to obtain $E[\partial^\ell\hat{h}(x)/\partial x^\ell] = \partial^\ell E[\hat{h}](x)/\partial x^\ell$. Next, let $\hat{H}(x)$ denote and ℓ^{th} order partial derivative of $\hat{h}(x)$, and $k(x)$ the corresponding derivative of $\mathcal{K}(x)$, so that $\hat{H}(x) = n^{-1}\sigma^{-k-\ell}\sum_{i=1}^n y_i k((x-x_i)/\sigma)$, and $\partial^\ell E[\hat{h}](x)/\partial x^\ell = E[\hat{H}(x)]$, where the n argument of $\sigma(n)$ is suppressed for notational convenience. Also, for a constant P, let $y_{in} = y_i$, $|y_i| \leq Pn^{1/p}$; $y_{in} = Pn^{1/p}$, $y_i > Pn^{1/p}$; $y_{in} = -Pn^{1/p}$, $y_i < Pn^{1/p}$. Let $\tilde{H}(x) = n^{-1}\sigma^{-k-\ell}\sum_{i=1}^n y_{in}k((x-x_i)/\sigma)$. Note that by Bonferonni's inequality,

$$\text{(B.2)} \qquad \text{Prob}(\tilde{H}(x) \neq \hat{H}(x) \text{ for some } x) \leq \text{Prob}(y_{in} \neq y_i \text{ for some } i \leq n)$$

$$\leq n\text{Prob}(y_{in} \neq y_i) \leq n\text{Prob}(|y_i| > Pn^{1/p}) \leq E[|y_i|^p]/P^{1/p}.$$

Let $\delta = [\ln(n)/(n\sigma^{k+2\ell})]^{1/2}$. For $c(x) = E[|y_i|^p|x_i=x]$ and P fixed, by $c(x)f_0(x)$ bounded and $p > 2$, (B.3) $$\delta^{-1}|E[\tilde{H}(x)] - E[\hat{H}(x)]| \le \delta^{-1}\sigma^{-k-\ell}E[1(|y_i|>Pn^{1/p})|y_i||\mathcal{k}((x-x_i)/\sigma)|]$$

$$\le C\delta^{-1}\sigma^{-k-\ell}n^{(1/p)-1}E[|y_i|^p|x_i]|\mathcal{k}((x-x_i)/\sigma)|]$$

$$= C\delta^{-1}\sigma^{-\ell}n^{(1/p)-1}\int|\mathcal{k}(v)|c(x-\sigma v)f_0(x-\sigma v)dv = O(\sigma^{k/2}n^{(1/p)-1/2}/\ln(n)^{1/2}) = o(1).$$

Next, by $\mathcal{k}(x)$ Lipschitz, $\sup_{\|x-\tilde{x}\|<1/n^3}|\tilde{H}(x)-\tilde{H}(\tilde{x})| \le Cn^{(1/p)-3}\sigma^{-k-\ell-1}$. Also, by $\mathcal{X}$ compact, it can be covered by less than Cn^{3k} open balls of radius n^{-3}. Let $x_{j\varepsilon}$ denote the centers of these open balls, $(j = 1, \ldots, J(\varepsilon))$, $J(\varepsilon) \le Cn^{-3k}$. Then for $x_{j\varepsilon}(x)$ equal to the center of an open ball containing x, by $|E[\tilde{H}(x)]-E[\tilde{H}(\tilde{x})]| \le E[|\tilde{H}(x)-\tilde{H}(\tilde{x})|]$ it follows that

(B.4) $$\sup_{\mathcal{X}}|\hat{H}(x)-E[\hat{H}(x)]| \le \sup_{\mathcal{X}}|\hat{H}(x)-E[\hat{H}(x)]-\{\hat{H}(x_{j\varepsilon}(x))-E[\hat{H}(x_{j\varepsilon}(x))]\}|$$

$$+ \sup_{\mathcal{X}}|\{\hat{H}(x_{j\varepsilon}(x))-E[\hat{H}(x_{j\varepsilon}(x))]\}| \le Cn^{(1/p)-3}\sigma^{-k-\ell-1} + \sup_j|\hat{H}(x_{j\varepsilon})-E[\hat{H}(x_{j\varepsilon})]|.$$

Note that $\ln(n)^2 < Cn$ and, by $\sigma(n)$ bounded, $\sigma^{3k} < C\sigma^{k+2}$. Then for the constant C in eq. (B.4), it follows by $p \ge 2$ and $n\sigma^k \longrightarrow \infty$ that for all M, n big enough, $M\delta - Cn^{(1/p)-3}\sigma^{-k-\ell-1} = M\delta(1 - C/[M^2\ln(n)n^{5-(2/p)}\sigma^{k+2}]^{1/2}) > M\delta(1 - C/[M^2\ln(n)n\sigma^k]^{3/2}) > M\delta/2$. Also, note that $n^{1/p}\sigma^{\ell}\delta = [n^{1-(2/p)}\sigma^k/\ln(n)]^{-1/2} \longrightarrow 0$. As usual for kernel estimators, $\sigma^{-2k-2\ell}E[y_{in}^2\mathcal{k}((x-x_i)/\sigma)^2] \le \sigma^{-2k-2\ell}E[y_i^2\mathcal{k}((x-x_i)/\sigma)^2] \le \sigma^{-k-2\ell}\int\mathcal{k}(u)^2E[y_i^2|x_i=x-\sigma u]f_0(x-\sigma u)du \le C\sigma^{-k-2\ell}$ by $E[y^2|x]f_0(x)$ bounded. Then by eq. (B.4), $y_{in}\sigma^{-k-\ell}\mathcal{k}((x-x_i)/\sigma)$ bounded by $Cn^{1/p}\sigma^{-k-\ell}$, and Bernstein's inequality, for M and n large enough,

(B.5) $$\text{Prob}(\sup_{\mathcal{X}}|\tilde{H}(x)-E[\tilde{H}(x)]| > M\delta) \le \text{Prob}(\sup_j|\hat{H}(x_{j\varepsilon})-E[\hat{H}(x_{j\varepsilon})]| > M\delta/2)$$

$$\le \textstyle\sum_{j=1}^{J(\varepsilon)}\text{Prob}(|\hat{H}(x_{j\varepsilon})-E[\hat{H}(x_{j\varepsilon})]| > M\delta/2)\ ]$$

$$\le 2\textstyle\sum_{j=1}^{J(\varepsilon)}\exp(-n^2\delta^2/[2n\text{Var}(\sigma^{-k-\ell}y_{in}\mathcal{k}((x_{j\varepsilon}-x_i)/\sigma) + Cn^{1+(1/p)}\sigma^{-k-\ell}\delta])$$

$$\le Cn^{3k}\exp(-n\delta^2/C[\sigma^{-k-2\ell} + n^{1/p}\sigma^{-k-\ell}\delta]) \le Cn^k\exp(-Cn\sigma^{k+2\ell}\delta^2/(1 + n^{1/p}\sigma^{\ell}\delta))$$

$\leq Cn^{3k}\exp(-CM^2\cdot\ln(n)) \leq C\exp(-[CM^2-3k]\ln(n))$.

Since these inequalities hold for any M, n large enough, it follows that $\sup_{\mathcal{X}}|\tilde{H}(x)-E[\tilde{H}(x)]| = O_p(\delta)$. Then by eq. (B.3) and the triangle inequality, $\sup_{\mathcal{X}}|\tilde{H}(x)-E[\hat{H}(x)]| = O_p(\delta)$. Consider any $\varepsilon > 0$. Choose P so that $E[|y_i|^P]/P^{1/P} < \varepsilon/2$, so that by eq. (B.2), $\mathrm{Prob}(\delta^{-1}\sup_{\mathcal{X}}|\tilde{H}(x)-\hat{H}(x)| > M/2) < \varepsilon/2$ for all n. For this fixed P, by $\sup_{\mathcal{X}}|\tilde{H}(x)-E[\hat{H}(x)]| = O_p(\delta)$ there exists M such that $\mathrm{Prob}(\delta^{-1}\sup_{\mathcal{X}}|\tilde{H}(x)-E[\hat{H}(x)]| > M/2) < \varepsilon/2$ for all n. Then by the triangle inequality, $\mathrm{Prob}(\delta^{-1}\sup_{\mathcal{X}}|\hat{H}(x)-E[\hat{H}(x)]| > M) \leq \mathrm{Prob}(\delta^{-1}\sup_{\mathcal{X}}|\tilde{H}(x)-\hat{H}(x)| > M/2) + \mathrm{Prob}(\delta^{-1}\sup_{\mathcal{X}}|\tilde{H}(x)-E[\hat{H}(x)]| > M/2) < \varepsilon$. Therefore, by eq. (B.2) and the triangle inequality, $\sup_{\mathcal{X}}|\hat{H}(x)-H(x)| = O_p(\delta)$. The conclusion then follows by applying this conclusion to each derivative of up to order j and by σ bounded. QED.

Lemma B.2: If Assumptions K, H, and Y are satisfied for $d \geq j+s$ then $\|E[\hat{h}]-h_0\|_j = O(\sigma^m)$.

Proof: Note that $E[\hat{h}](x) = E[y_iK_\sigma(x-x_i)] = \int h(t)[\mathcal{K}((x-t)/\sigma)/\sigma^k]dt = \int\mathcal{K}(u)h(x+u\sigma)du$, so that by $\mathcal{K}(u)$ having finite support, $\partial^jE[\hat{h}](x)/\partial x^j = \int\mathcal{K}(u)\partial^jh(x+u\sigma)/\partial x^jdu$. Also, by $\int\mathcal{K}(u)du = 1$ it follows that $h_0(x) = \int\mathcal{K}(u)h_0(x)du$. Then by a Taylor expansion in σ around $\sigma = 0$, for constant matrices C_ℓ, $(\ell = 1, \ldots, j)$,

$$\text{(B.6)}\qquad \|\partial^j\bar{h}/\partial x^j - \partial^jh_0(x)/\partial x^j\| = \|\textstyle\sum_{\ell=1}^{m-1}\sigma^\ell C_\ell\int\mathcal{K}(u)\{\otimes_{p=1}^{\ell}u\}\otimes\{\partial^{j+\ell}h(x)/\partial x^{j+\ell}\}du$$

$$+ C_m\sigma^m\int\mathcal{K}(u)\{\otimes_{r=1}^{m}u\}\otimes\{\partial^{j+m}h(x+\bar{\sigma}u)/\partial x^{j+m}\}du\|$$

$$\leq C\sigma^m[\int|\mathcal{K}(u)|\,\|u\|^mdu]\|\sup_x\|\partial^{j+m}h(x)/\partial x^{j+m}\| \leq C\sigma^m. \qquad \text{QED.}$$

Lemma B.3: If the hypotheses of Lemmas B.1 and B.2 are satisfied and Assumption H is satisfied with $d \geq j+s$ then

$$\text{(B.7)}\qquad \|\hat{h}-h_0\|_j = O_p(\ln(n)^{1/2}(n\sigma^{k+2j})^{-1/2} + \sigma^s).$$

Proof: Follows by Lemmas B.1 and B.2 and the triangle inequality. QED.

Lemma B.4: If Assumption K is satisfied, $m(h)$ *is linear,* $|m(h)| \le C\|h\|_\Delta$, *then* $E[m(\hat{h})] = m(E[\hat{h}])$.

Proof: By $\mathcal{K}(u)$ having finite support and $\mathcal{X}$ compact there is a compact set $\mathcal{C}$ such that $\|K_\sigma(\cdot-x)\|_\Delta = 0$, and hence $m(g(x)K_\sigma(\cdot-x)) = 0$, for all $x \notin \mathcal{C}$. Hence, by linearity of $m(h)$, $E[m(\hat{h})] = \int_{\mathcal{C}} m(g(x)K_\sigma(\cdot-x))f_0(x)dx$ and $m(E[\hat{h}]) = m(\int_{\mathcal{C}} g(x)K_\sigma(\cdot-x)f_0(x)dx)$. Let $F_J(x)$ be a sequence of measures with finite support, that converge in distribution to the distribution of x on $\mathcal{C}$ (e.g. the empirical measure from a sequence of i.i.d. draws) as $J \longrightarrow \infty$. Then, since $m(g(x)K_\sigma(\cdot-x))$ is continuous and bounded on $\mathcal{C}$, it follows that $\int_{\mathcal{C}} m(g(x)K_\sigma(\cdot-x))F_J(dx) \longrightarrow E[m(\hat{h})]$. Also, since each derivative of $g(x)K_\sigma(\tilde{x}-x)f_0(x)$ with respect to $\tilde{x}$ of up to order Δ is bounded and continuous on $\mathcal{C}$, it follows that $\|\int_{\mathcal{C}} g(x)K_\sigma(\cdot-x)F_J(dx)-E[\hat{h}]\|_\Delta \longrightarrow 0$, and hence $m(\int_{\mathcal{C}} g(x)K_\sigma(\cdot-x)F_J(dx)) \longrightarrow m(E[\hat{h}])$. Furthermore, by F_J having finite support, $m(\int_{\mathcal{C}} g(x)K_\sigma(\cdot-x)F_J(dx)) = \int m(g(x)K_\sigma(\cdot-x))F_J(dx)$. Then T gives the conclusion. QED.

Lemma B.5: If Assumption K is satisfied and for given $\tilde{h}$ *with* $\|\tilde{h}\|_\Delta < \infty$ *there is linear* $D(h)$ *with* $|m(h)-m(\tilde{h})-D(h-\tilde{h})| = o(\|h-\tilde{h}\|_\Delta)$ *as* $\|h-\tilde{h}\|_\Delta \longrightarrow 0$, *then* $dm(\tilde{h} + \zeta yK_\sigma(\cdot-x))/d\zeta|_{\zeta=0} = D(yK_\sigma(\cdot-x))$.

Proof: Let $h_\zeta = h + \zeta yK_\sigma(\cdot-x)$, so that $\|h_\zeta-\tilde{h}\|_\Delta \le \zeta\|yK_\sigma(\cdot-x)\| \le C\zeta$. Then $|m(h_\zeta)-m(\tilde{h})-\zeta D(yK_\sigma(\cdot-x))|/\zeta = |m(h_\zeta)-m(\tilde{h})-D(h_\zeta-\tilde{h})|/\zeta = o(\|h_\zeta-\tilde{h}\|_\Delta/|\zeta|) = o(1)$. QED.

References

Bierens, H.J. (1987): "Kernel Estimators of Regression Functions," in Bewley, T.F., ed., *Advances in Econometrics: Fifth World Congress*, Cambridge: Cambridge University Press.

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

Hall, P. (1992): *The Bootstrap and Edgeworth Approximation*, New York: Springer-Verlag.

Hardle, W. and T. Stoker (1989): "Investigation of Smooth Multiple Regression by the Method of Average Derivatives," *Journal of the American Statistical Association,* 84, 986-995.

Hausman, J.A. and W.K. Newey (1992): "Nonparametric Estimation of Exact Consumer Surplus and Deadweight Loss," working paper, MIT Department of Economics.

Huber, P. (1967): "The Behavior of Maximum Likelihood Estimates Under Nonstandard Conditions," *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability*, Berkeley: University of California Press.

Lerman, S.R. and C.F. Manski (1981): "On the Use of Simulated Frequencies to Approximate Choice Probabilities," in Manski, C.F. and D. McFadden, eds., *Structural Analysis of Discrete Data with Econometric Applications*, Cambridge: MIT Press.

Li, K.C. and Duan, N. "Regression Analysis Under Link Violation," *Annals of Statistics*, 17, 1009-1052.

Matzkin, R. and W.K. Newey (1992): "Nonparametric Estimation of Structural Discrete Choice Models," mimeo, Department of Economics, Yale University

Newey, W.K. and P.A. Ruud (1992): "Density Weighted Least Squares Estimation," working paper, MIT Department of Economics.

Powell, J.L., J.H. Stock, and T.M. Stoker (1989): "Semiparametric Estimation of Index Coefficients," *Econometrica,* 57, 1403-1430.

Robinson, P. (1988): "Root-N-Consistent Semiparametric Regression," *Econometrica,* 56, 931-954.

Ruud, P.A. (1986): "Consistent Estimation of Limited Dependent Variable Models Despite Misspecification of Distribution," *Journal of Econometrics*, 32, 157-187.

Stock, J.H. (1989): "Nonparametric Policy Analysis," *Journal of the American Statistical Association,* 84, 567-575.

Tauchen, G.E. (1985): "Diagnostic Testing and Evaluation of Maximum Likelihood Models," *Journal of Econometrics*, 30, 415-443.

Willig, R. (1976): "Consumer's Surplus without Apology," *American Economic Review,* 66, 589-597.