KERNEL ESTIMATION
OF PARTIAL MEANS AND A
GENERAL VARIANCE ESTIMATOR

WHITNEY K. NEWEY
Massachusetts Institute of Technology

Econometric applications of kernel estimators are proliferating, suggesting
the need for convenient variance estimates and conditions for asymptotic nor-
mality. 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 verify them for particular estima-
tors. 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 esti-
mation, and are useful for estimation of additive nonparametric models.

1. INTRODUCTION

There are a growing number of applications where estimators use the ker-
nel method in their construction, that is, where functionals of kernel estima-
tors are involved. Examples include average derivative estimation [4,11],
nonparametric policy analysis [14], consumer surplus estimation [5], and oth-
ers that are the topic of current research. An important example in this pa-
paper is the partial mean, which is an average of a kernel regression estimator
over some components holding others fixed. The growth of kernel applica-
tions 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 ker-
nel estimators.

Partial means control for covariates by averaging over them. They are re-
lated to additive nonparametric models and have important uses in econom-
ics, as 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 re-
spect to the contribution of each observation to the kernel. A more common
method is to calculate the asymptotic variance formula and then "plug-in"

Financial support was provided by the NSF. D.W.K. Andrews, P. Robinson, T. Stoker, and the referees pro-
vided useful comments.
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 [6] 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 that are based on \( t \)-ratios. An asymptotic variance estimator is an integral part of the \( t \)-ratio, and hence of these improvements. See Hall [3] for a description and motivation of these bootstrap methods.

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 [3]. For a kernel regression at a point, a related estimator was proposed by Bierens [1]. Also, the standard errors for average derivatives in Härdle and Stoker [4], and Powell, Stock, and Stoker [11] 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, that is, the asymptotic variance estimator. Section 4 describes partial means and their 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

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

Let \( \mathcal{K}(u) \) denote a kernel function that satisfies \( \int \mathcal{K}(u) du = 1 \) 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 $y_t$ and $x_t$ on $y$ and $x$. Then for a bandwidth $\sigma > 0$ and $K_\sigma(u) = \sigma^{-d} K(u/\sigma)$, a kernel estimator of $h_0$ is

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

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 $\beta$ denote a vector of parameters, with true value $\beta_0$, 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

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

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

An important practical problem is the choice of bandwidth $\sigma$. The regularity conditions given below for $\sqrt{n}$-consistency will require that $\sigma$ be chosen to be smaller than the value that would minimize the asymptotic mean square error of $\hat{h}$, a feature that is often referred to as "undersmoothing." Thus, choosing the bandwidth from cross-validation, or any other method that minimizes the asymptotic mean square error, is not appropriate. It is beyond the scope of this paper to say much more about the theory of how to choose $\sigma$, but a practical method might be to start at a value obtained by cross-validation and decrease $\sigma$ until $\hat{\beta}$ does not change much relative to its estimated standard error.

The estimator $\hat{\beta}$ includes as special cases functions of kernel estimators evaluated at points, for example, 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 $\bar{x}_2$ be a variable that is included in $z$ and has the same dimension as $x_2$, and $x_1$ 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

$$\beta_0 = E[\tau(\bar{x}_2)g_0(\bar{x}_1, \bar{x}_2)]. \tag{2}$$

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 \( y = (1, q) \), so that \( h(x) = (h_1(x), h_2(x))' \) where \( h_{10}(x) \) is the density of \( x \) and \( h_{20}(x) = h_{10}(x)E[q|x] \). Also, let \( \hat{g}(x) = \hat{h}_2(x)/\hat{h}_1(x) = \hat{h}_2(x)/f(x) \), for the kernel density estimator \( f(x) = \hat{h}_1(x) \), and \( x_i = (x_{i1}, x_{i2}) \). Then the estimator is

\[
\hat{\beta} = n^{-1} \sum_{i=1}^n \tau(\hat{x}_{i2}) \hat{g}(\hat{x}_i)
\]  

(3)

This estimator is a special case of equation (1) with \( m(z,\beta,h) = \tau(\hat{x}_2) \times h_2(\hat{x}_1, \hat{x}_2)/h_1(\hat{x}_1, \hat{x}_2) - \beta \). It shows how explicit estimators can be included as special cases of equation (1). Also, as further discussed in Section 4, this estimator can be applied to estimation of approximate consumer surplus, an important problem in economics.

\textbf{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 \( \tilde{x}_1 \) 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

\[
\hat{\beta} = S(p^0), \quad dS(p)/dp = -\hat{g}(\tilde{x}_1 - S(p), p), \quad S(p^1) = 0,
\]  

(4)

for \( \hat{g}(x) = \hat{h}_2(x)/\hat{f}(x) \). It is a special case where the \( m(z,\beta,h) \) of equation (1) is the solution of the differential equation minus \( \beta \). 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 \( \tilde{x}_1 \) and demand function \( g_0(x) = E[q|x] \). This example is analyzed in Hausman and Newey [5] by using results developed here.

\textbf{Inverse density weighted least squares.} An estimator that is useful for estimating the semiparametric generalized regression model \( E[q|x] = t(x'\hat{\beta}) \), where \( t(\cdot) \) is an unknown transformation, is a weighted least-squares estimator, which can be described as follows. Let \( \tau(x) \) be a user-specified density of an elliptically symmetric distribution for \( x \) (i.e., \( \tau(x) \) is a density that depends only on \( (x - \mu)'\Sigma(x - \mu) \) for some \( \mu \) and \( \Sigma \) that has bounded support. The estimator solves

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

(5)

This estimator has the form given in equation (1), 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 by Ruud [13] and Li and Duan [8]. This estimator is analyzed in Newey and Ruud [10], 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 \( \beta \) 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 (1) around \( \beta_0 \) and solving for \( \hat{\beta} - \beta_0 \) gives

\[
\hat{\beta} - \beta_0 = -\left[ n^{-1} \sum_{i=1}^{n} \frac{\partial m(z_i, \hat{\beta}, \hat{h})}{\partial \beta} \right]^{-1} \hat{m}_n(\beta_0),
\]

where \( \hat{\beta} \) is a mean value. By the uniform law of large numbers discussed in Section 5, \( n^{-1} \sum_{i=1}^{n} \frac{\partial m(z_i, \hat{\beta}, \hat{h})}{\partial \beta} \) will converge in probability to

\[
M = \mathbb{E}[\frac{\partial m(z, \beta_0, h_0)}{\partial \beta}],
\]

so that the asymptotic distribution of \( \hat{\beta} \) will be determined by \( \hat{m}_n(\beta_0) \). In Section 5 conditions will be given for existence of \( \alpha \geq 0 \) such that

\[
\sqrt{n} \alpha \hat{m}_n(\beta_0) \xrightarrow{d} \mathcal{N}(0, V).
\]

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

\[
\sqrt{n} \alpha (\hat{\beta} - \beta_0) \xrightarrow{d} \mathcal{N}(0, M^{-1} V M^{-1}).
\]

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

\[
\hat{M} = n^{-1} \sum_{i=1}^{n} \frac{\partial m(z_i, \hat{\beta}, \hat{h})}{\partial \beta}.
\]

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 estima-
tor by substituting estimates for unknown functions, such as sample averages for expectations. This approach can be difficult when the asymptotic variance is very complicated.

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 \( \xi \) denote a scalar number and let

\[
\hat{\delta}_i = \frac{\partial}{\partial \xi} \left[ n^{-1} \sum_{j=1}^{n} m(z_j, \hat{\beta}, \hat{h} + \xi y_j K_n(\cdot - x_i)) \right]_{\xi=0}.
\]

The interpretation of \( \hat{\delta}_i \) is that it estimates the first-order effect of the \( i \)th observation in \( \hat{h} \) on \( \sum_{j=i}^{n} m(z_j, \hat{\beta}_0, \hat{h})/n \). 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

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

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

\[
\text{var}(\hat{\beta}) = \hat{M}^{-1} \hat{V} \hat{M}^{-1}, \quad \hat{M} = n^{-1} \sum_{i=1}^{n} \frac{\partial m(z_i, \hat{\beta}, \hat{h})}{\partial \beta}.
\]

In Section 5, conditions will be given that are sufficient for \( \sigma^2 \text{var}(\hat{\beta}) \overset{p}{\rightarrow} M^{-1} VM^{-1} \). Consequently, inference procedures based on \( \hat{\beta} - \beta_0 \) being normally distributed with mean 0 and variance \( \text{var}(\hat{\beta})/n \) will be asymptotically valid. For example, \( \hat{\beta} \pm 1.96(\text{var}(\hat{\beta})_n)^{1/2} \) will be an asymptotic 95% confidence interval. It is interesting to note that the form of \( \text{var}(\hat{\beta}) \) does not depend on the convergence rate for \( \hat{h} \) (i.e., on \( \alpha \)), 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 \( \xi \). Alternatively, if the analytic formula is very hard to construct, \( \hat{\delta}_i \) can be calculated as the numerical derivative of \( \sum_{j=1}^{n} m(z_j, \hat{\beta}, \hat{h} + \xi y_j K_n(\cdot - x_i))/\xi \) with respect to \( \xi \).

Here, \( V = \sum_{i=1}^{n} \hat{\psi}_i \hat{\psi}_i/n \) is a "delta-method" variance for kernel estimators. It is exactly analogous to delta-method variances for parametric estimators. For example, if \( \hat{h} \) were a sample mean rather than a kernel estimator, say \( \hat{h} = \sum_{i=1}^{n} y_i/n \), then the analog of \( \hat{\delta}_i \) would be \( \frac{\partial}{\partial \xi} \left[ \sum_{j=1}^{n} m(z_j, \hat{\beta}, \hat{h} + \xi y_j)/n \right] = m_n y_i \), where \( m_n = n^{-1} \sum_{j=1}^{n} \partial m(z_j, \hat{\beta}, \hat{h})/\partial \beta \). Thus, the analogous influence function estimator would be \( \psi_i = m(z_i, \hat{\beta}, \hat{h}) + m_n y_i - (\sum_{j=1}^{n} m_n y_j/n) = m(z_i, \hat{\beta}, \hat{h}) + m_n (y_i - \bar{y}) \), the usual delta-method formula for the presence of a sample average in an \( m \)-estimator. Another feature of \( \text{var}(\hat{\beta}) \) is that it
KERNELS, PARTIAL MEANS, AND VARIANCE ESTIMATION 239

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 \( \sqrt{n}(\hat{\beta} - \beta_0) \), where \( \beta_0 \) is the limit of \( \hat{\beta} \) when the bandwidth is held fixed at \( \sigma \).

The terms \( m(z_i, \hat{\beta}, h) \) 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 \( \hat{\delta}_j \) does not distinguish between elements of \( h \) 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 \( \hat{\beta} = f(x) \) at some \( x \), where the asymptotic variance estimator is \( \text{var}(\hat{\beta}) = \frac{1}{n} \sum_{j=1}^{n} K_\sigma(x_j - x)^2/n - \frac{1}{n} \left[ \sum_{j=1}^{n} K_\sigma(x_j - x_i)/n \right]^2 \), the sample variance of \( K_\sigma(x_j - x_i) \). This estimator was recently considered by Hall [3]. Other examples are as follows.

**Partial means.** Here \( \hat{\delta}_j \) can be obtained by explicit differentiation of \( n^{-1} \sum_{i=1}^{n} \tau(x_j) \frac{\hat{\beta}(x_j) + \sum_{i=1}^{n} K_\sigma(x_j - x_i)}{\hat{f}(x_j) + \sum_{i=1}^{n} K_\sigma(x_j - x_i)} \), as

\[
\hat{\delta}_j = n^{-1} \sum_{i=1}^{n} \tau(x_j) \frac{\hat{\beta}(x_j) - \hat{f}(x_j) - \hat{\delta}_j}{\hat{f}(x_j) + \sum_{i=1}^{n} K_\sigma(x_j - x_i)}. 
\]

The asymptotic variance estimator can then be formed as

\[
\text{var}(\hat{\beta}) = \frac{1}{n} \sum_{i=1}^{n} \hat{\psi}_i^2/n, \quad \hat{\psi}_i = \tau(x_j) \frac{\hat{\beta}(x_j) - \hat{f}(x_j) - \hat{\delta}_j}{\hat{f}(x_j) + \sum_{i=1}^{n} K_\sigma(x_j - x_i)}. 
\]

**Differential equation solution.** It is possible to derive an analytical expression for \( \hat{\delta}_j \), but this expression is quite complicated and difficult to evaluate. An alternative approach that is used by Hausman and Newey [5] is to numerically differentiate the numerical solution to \( dS(p)/dp = -\left[ \frac{\hat{h}_2(p, x_j - S(p))}{\hat{f}(p, x_j - S(p)) + \sum_{i=1}^{n} K_\sigma((p, x_j - S(p)) - x_i)} \right] \) with respect to \( \tau \) to form \( \hat{\delta}_j \). This approach is quite feasible by using existing fast and accurate numerical algorithms for ordinary differential equations.

**Inverse density weighted least squares.** Let \( \hat{u}_i = x_j - x_i \hat{\beta} \). The variance estimator is

\[
\text{var}(\hat{\beta}) = \hat{M}^{-1} \left( n^{-1} \sum_{i=1}^{n} \hat{\psi}_i \hat{\psi}_i' \right) \hat{M}^{-1}, \quad \hat{M} = n^{-1} \sum_{i=1}^{n} \hat{f}(x_j)^{-1} \tau(x_j) x_i x_i', \\
\hat{\psi}_i = \tau(x_j) x_i \hat{u}_i + \hat{\delta}_j - \sum_{j=1}^{n} \hat{\delta}_j n, \quad \hat{\delta}_j = -n^{-1} \sum_{j=1}^{n} \tau(x_j) \hat{f}(x_j)^{-2} x_j \hat{u}_j K_\sigma(x_j - x_i). 
\]
For partial means, asymptotic normality and consistency of the variance estimator are shown in Section 4 by using the lemmas of Section 5. Corresponding theoretical results for the differential equation and weighted least-squares examples are given in Hausman and Newey [5] and Newey and Ruud [10], respectively.

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 $g_0(\tilde{x}_1 - S(p), p)$ leads to an approximation as $\hat{\beta}_0 = \int_{p^0}^{p^1} g_0(\tilde{x}_1, p) \, dp = E[(p^1 - p^0) g_0(\tilde{x}_1, \tilde{x}_2)]$, 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 $x_t$ (see [16]). This is a partial mean as described in Section 2. It can be estimated by

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

(12)

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 [7].

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 $E[\tau(x_2)] = 1$. Then

$$E[\tau(\tilde{x}_2) g_0(\tilde{x}_1, \tilde{x}_2)] = E[\tau(\tilde{x}_2) g_{10}(\tilde{x}_1)] + E[\tau(\tilde{x}_2) g_{20}(\tilde{x}_2)].$$

(13)

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

In comparison with other estimators for additive models, partial means are easier to compute but may be less asymptotically efficient. Unlike the alternating conditional expectation estimator for additive models (ACE, see [2]), 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 estimates a different statistical object than ACE when $E[q|x]$ is not additive. The partial mean estimates the object in equation (2).
The object estimated by ACE 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 describing dependence of $E[q | x]$ on $x_1$, while controlling for the other variables in an average way, then the partial mean would be of interest.

An important property of partial mean estimators is that their convergence rates will depend on the number of components that are being averaged out. The more components that are included in the average, the higher will be the convergence rate, with $\sqrt{n}$-consistency resulting when all components have been averaged out. Also, for making an inference about a partial mean, it is useful for $\beta$ to be asymptotically normal and $\var(\beta)$ to be a consistent asymptotic variance estimator. These results can be shown under fairly standard regularity conditions.

The first condition imposes restrictions on the kernel.

Assumption K. There are positive integers $A$ and $s$ such that $\mathcal{K}(u)$ is differentiable of order $A$, the derivatives of order $A$ are Lipschitz, $\mathcal{K}(u)$ is zero outside a bounded set, $\int \mathcal{K}(u) \, du = 1$, and for all $j < s$, $\int \mathcal{K}(u) \left[ \otimes_{t=1}^j u \right] \, du = 0$.

The last condition requires that the kernel be a higher order (bias reducing) kernel of order $s$. It will be used here to guarantee that the limiting distribution of the estimators are centered at the true value.

The next condition imposes smoothness on the functions $h(x)$.

Assumption H. There is a non-negative integer $d$ and an extension of $h_0(x)$ to all of $\mathbb{R}^d$ that is continuously differentiable to order $d$ on $\mathbb{R}^d$.

This condition is used in conjunction with Assumption K to make sure the bias of the estimator is small. It rules out cases where the density of $x$ and its derivatives are nonzero on the boundary of the support by requiring smoothness everywhere.

Under these conditions and certain others, partial means will be asymptotically normal. Let the $u$ argument of $\mathcal{K}(u)$ be partitioned conformably with $x$ and $f_0(\tilde{x}_2)$ denote the true density of $\tilde{x}_2$. The asymptotic variance of the partial mean estimator will be

$$V = \left[ \int \left( \int \mathcal{K}(u_1, u_2) \, du_2 \right)^2 \, du_1 \right] \times \int f_0(\tilde{x}_1, t)^{-1} \tau(t)^2 f_0(t)^2 \, \text{var}(q | x = (\tilde{x}_1, t)) \, dt. \quad (14)$$

**THEOREM 4.1.** Suppose that (i) $E[|q|^4] < \infty$, $E[|q|^4 | x] f_0(\tilde{x})$, 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 $f_0(\tilde{x}_1, \tilde{x}_2)$ is bounded away from zero; (iv) $\tau(\tilde{x}_2)$ and $f_0(\tilde{x}_2)$ are continuous a.e., $f_0(\tilde{x}_2)$ is bounded, $E[q | x]$ and $E[q^2 | x]$ are continuous, and for some $c > 0$,
The fourth moment condition is a dominance condition for the density and the product of the density and the fourth conditional moment, which treats \( x_1 \) as a fixed parameter and integrates over \( x_2 \). The dominance over the small variation in \( x_1 \) is useful because of the convolution form of the kernel estimator. The bandwidth 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 \( \sqrt{n\sigma_k/2} \) for the asymptotic distribution. The normalizing factor for a kernel estimator at a point is \( \sqrt{n\sigma_k/2} \), so that integrating over \( x_2 \) has increased the convergence rate to the pointwise rate for \( x_1 \).

One important feature of this result is hypothesis (iii), which amounts to a “fixed-trimming” condition, where the density of \( x \) is bounded away from zero where \( r \) 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, for example, as in Robinson [12], but this modification would be very complicated.

Estimators will be \( \sqrt{n} \)-consistent when they are full means, that is, are averages over all components. There are many interesting examples of such estimators, such as the policy analysis estimator of Stock [14]. 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 \( \beta_0 = E[a_0(z)g_0(x)] \), where \( g_0(x) = E[q|x] \), \( a_0(z) \) is some function of the data, and \( x \) is a continuously distributed variable that may be different than \( x \). A kernel estimator of \( \beta_0 \), along with the associated asymptotic variance estimator from equations (7) and (8), for \( \hat{g}(x) = \hat{h}_2(x)/\hat{h}_1(x) \) as above, is

\[
\hat{\beta} = \frac{1}{n} \sum_{i=1}^{n} a_0(z_i) \hat{g}(\hat{x}_i), \quad \text{vár}(\hat{\beta}) = \frac{1}{n} \sum_{i=1}^{n} \frac{\hat{\beta}_i^2}{n},
\]

\[
\hat{\psi}_i = a_0(z_i) \hat{g}(\hat{x}_i) - \hat{\beta} + \hat{\delta}_i - \frac{1}{n} \sum_{j=1}^{n} \frac{\hat{\delta}_j}{n},
\]

\[
\hat{\delta}_i = \frac{1}{n} \sum_{j=1}^{n} a_0(z_j) \hat{f}(\hat{x}_j)^{-1} [q_i - \hat{g}(\hat{x}_i)] K_s(\hat{x}_j - x_i) \frac{q_i - \hat{g}(\hat{x}_i)}{n}. \quad (15)
\]
The asymptotic variance of this estimator will be

\[ V = E[\psi^2], \]

\[ \psi_i = a_0(z_i)g_0(\tilde{x}_i) - \beta_0 + E[a_0(z)|x_i]f_0(x_i)^{-1}f_0(x_i)[q_i - g_0(x_i)]. \]  

(16)

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

This result gives asymptotic normality for a trimmed version of Stock's [14] 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 \( \hat{\beta} \)-estimator described in Section 2. Uniform convergence results are useful for showing consistency of \( \hat{\beta} \) and of the Jacobian term in the expansion of equation (6). Asymptotic normality of \( \hat{\beta} \) will follow from \( \sqrt{n}\alpha^m m(z, \beta_0, h_0) \overset{d}{\to} N(0, V). \) Also, \( \sigma^2W \overset{p}{\to} V. \) 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 \( \psi_i \) such that \( \sqrt{n}\alpha m(z, \beta_0, h_0) = \sum_{i=1}^n \psi_i/\sqrt{n} + o_p(1) \) and \( \sum_{i=1}^n ||\psi_i - \psi||^2/n \overset{p}{\to} 0. \) 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 [5], [9], and [10].

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

For the asymptotic theory, it is 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 deriv-
atives, 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, which makes it an easy norm to work with.

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 non-negative integer $j$, let

$$\|B\|_j = \max_{\mathcal{X}} \sup_{x \in \mathcal{X}} \left| \frac{\partial^j B(x)}{\partial x^j} \right|,$$

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 $\beta \in \Theta$ with probability 1, where $\Theta$ is compact, and $E[\sup_{\beta \in \Theta} |m(z,\beta, h_0)|] < \infty$; (ii) Assumptions K, H, and Y are satisfied with $d \geq \Delta + 1$, $\ln(n)/(n \sigma^{k+2\Delta}) \to 0$ and $\sigma \to 0$, and there is $b(z)$ and $\varepsilon > 0$ such that $E[b(z)] < \infty$, and for all $\beta \in \Theta$ and $\|h - h_0\|_\Delta < \varepsilon$, $|m(z,\beta, h) - m(z,\beta, h_0)| \leq b(z)(\|h - h_0\|_\Delta)^\varepsilon$. Then $E[m(z,\beta, h_0)]$ is continuous on $\Theta$ and

$$\sup_{\beta \in \Theta} \left| \frac{1}{n} \sum_{i=1}^{n} m(z_i, \beta, \hat{h}) - E[m(z,\beta, h_0)] \right| \to 0. \quad (17)$$

The uniform convergence conclusion of equation (17) is a well-known condition for consistency of the solution to equation (1). Also, equation (17) 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 the derivative.

Asymptotic normality of $\sqrt{n} \sigma \hat{m}_n(\beta_0)$ 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 $\sqrt{n} \sigma \hat{m}_n(\beta_0)$ 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
KERNELS, PARTIAL MEANS, AND VARIANCE ESTIMATION 245

magnitude of $\alpha$) 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 mo-
ment, 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
$E[\sup_{|x| \leq t} |\nu(x + u)|^2 E[\|y\|^2 | x]| < \infty$, and $\sqrt{n} \eta \rightarrow 0$, then for $\delta_i = \nu(x_i) y_i$,
$\sqrt{n} (m(h) - m(h_0)) = \sum_{i=1}^n [\delta_i - E[\delta_i]]/\sqrt{n} + o_p(1)$.

Cases where convergence is slower than $1/\sqrt{n}$ are somewhat more com-
plicated. The following assumption is useful for these cases. For the mo-
ment, let $\ell$ be a non-negative integer and let $\partial^\ell h(x)/\partial x^\ell$ be ordered so that
$\partial^\ell [y, K_x(x - x_i)]/\partial x^\ell = y_i \otimes [\partial^\ell K_x(x - x_i)]/\partial x^\ell$.

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

The key condition here is the integral representation of $m(h)$. The dimen-
sion of the argument being integrated and the order of the derivative lead to
the convergence rate for $m(h)$, that is $\sqrt{n} \eta^{k_1/2 + \ell}$. Thus, every additional
dimension of integration increases the convergence rate by $\sqrt{\varepsilon}$; whereas, every
additional derivative decreases the rate by a factor of $1/\varepsilon$. This hypothesis
also leads to a specific form for the asymptotic variance of $m(h)$, which for
$\bar{R}(u_1, t) = \int \omega(t) \left[ \int \bar{R}(u_1, t) \bar{R}(u_1, t') \, du_1 \right] \omega(t)f_0(x(t)) \, dt$.

LEMMA 5.3. If Assumptions K, H, and Y are satisfied, $\mathcal{C}$ is
compact, there is a vector of functionals $D(z, h)$, and non-negative constants
$\alpha, \Delta \leq \Delta, (\ell = 1, 2), \varepsilon > 0$ such that $d \geq \max\{\Delta + 1, \Delta + s, \Delta_2 + s\}$ and
(i) $D(z, h)$ is linear in $h$ on $h: \|h\|_\Delta < \infty$; (ii) for all $h$ with $\|h - h_0\|_\Delta < \varepsilon$,
\[ \| m(z,h) - m(z,h_0) - D(z,h - h_0) \| \leq b(z) \| h - h_0 \|_{\Delta_1} ; \]
\[ (iii) \| D(z,h) \| \leq \tilde{b}(z) \| h \|_{\Delta_1} \text{ and } E[\tilde{b}(z)^4] < \infty ; \]
\[ (iv) \text{ for } \eta_n = [\ln(n)/ (n\sigma^{k+2})]^{1/2} + \sigma^2, \eta_n^2 \to 0, \sqrt{n} \sigma^2 E[\tilde{b}(z)] \eta_n^{2+2} \to 0, \text{ and } \sqrt{n} \sigma^{k+\Delta_1} \to \infty. \]

Then, for \( m(h) = \int D(z,h) dF(z) \),
\[ \sqrt{n} \sigma^2 \sum_{i=1}^{n} [m(z_i,h) - m(z_i,h_0)]/n = \sqrt{n} \sigma^2 [m(h) - m(h_0)] + o_p(1). \]

The conditions of this result imply Frechet differentiability at \( h_0 \) of \( m(z,h) \)
as a function of \( h \), in the Sobolev norm \( \| A \|_{\Delta_1} \). 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 \( \sigma^2 \sum_{i=1}^{n} m(z_i,h)/\sqrt{n} \) 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 \( \sum_{i=1}^{n} \{ m(z_i,h_0) + \delta_i - E[\delta_i] \}/\sqrt{n} + o_p(1) \), so that asymptotic normality, with asymptotic variance \( \text{var}(m(z_i,h_0) + \delta_i) \), 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 \( \sigma^2 \sum_{i=1}^{n} m(z_i,h_0)/\sqrt{n} \overset{D}{\to} 0 \), so that \( \sigma^2 \sum_{i=1}^{n} m(z_i,h)/\sqrt{n} \to N(0,V) \).

Assumption 5.2. (i) \[ \| m(z,\beta,\tilde{h}) - m(z,\beta_0,\tilde{h}_0) \| \leq b(\beta) \| \beta - \beta_0 \|_{\Delta_1} \] and \( E[b(\beta)^2] < \infty \) (ii) For \( \epsilon > 0 \) and \( \| \beta - \beta_0 \|_{\Delta_1} < \epsilon \) and \( \| \tilde{h} - \tilde{h}_0 \|_{\Delta_1} < \epsilon \), there is \( D(z,h;\beta,\tilde{h}) \) that is linear on \( h \) satisfying \( \| D(z,\beta,\tilde{h}) - D(z,\beta,\tilde{h}) \|_{\Delta_1} \leq b(\beta) \| \tilde{h} - \tilde{h}_0 \|_{\Delta_1} \) and \( D(z,h;\beta_0,\tilde{h}_0) \leq b(\beta) \| \tilde{h} - \tilde{h}_0 \|_{\Delta_1} \) and \( E[\tilde{b}(z)^4] < \infty \). (iv) \( \tilde{b} = \beta_0 + O_p(\delta_{\beta n}), \delta_{\beta n} \to 0, \sigma^{k+\Delta_1} \delta_{\beta n} \to 0, \alpha + s > k + \Delta_1, n\sigma^{k+\Delta_1+2\Delta_2-2n}/(n\sigma^{k+\Delta_1}) \to \infty. \)

**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 \( \delta_i = \nu(x_i) \beta^i, \sum_{i=1}^{n} [\delta_i - \tilde{\delta}_i]/\sqrt{n} \overset{D}{\to} 0 \) and \( V \overset{P}{=} V = \text{var}(\delta_i) \). If \( m(h) = \int D(z,h;\beta_0,h_0) dF(z) \) satisfies the conditions of Lemma 5.4, \( \sigma^2 V \overset{D}{=} V \), for \( V \) in equation (18).

**References**

1. Bierens, H.J. Kernel estimators of regression functions. In T.F. Bewley (ed.), *Advances in Econometrics: Fifth World Congress*. Cambridge: Cambridge University Press, 1987.
2. Breiman, L. & J.H. Friedman. Estimating optimal transformations for multiple regression and correlation. *Journal of the American Statistical Association* 80 (1985): 580–598.
3. Hall, P. *The Bootstrap and Edgeworth Approximations*. New York: Springer-Verlag, 1992.
4. Härdle, W. & T. Stoker. Investigation of smooth multiple regression by the method of average derivatives. *Journal of the American Statistical Association* 84 (1989): 986–995.
5. Hausman, J.A. & W.K. Newey. Nonparametric estimation of exact consumer surplus and deadweight loss. Working paper, MIT Department of Economics, 1992.
APPENDIX A: PROOFS OF THEOREMS

Throughout the Appendix, \( C \) will denote a generic constant that may be different in different uses and \( \sum_{i=1}^{n} = \sum_{i=1}^{n} x_i \). 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 \( x = (x_1, x_2, ..., x_n) \), \( X \) be the compact set of hypothesis (iii), and \( |A| = \sup_{x \in X} |A(x)| \). Let \( m(x, h) = m(x) = T(x) f_0(x) \), \( D(z, h; \bar{h}) = r(x) h_1(x) - h_2(x) h_1(x) h_2(x) \), and \( D(z, h) = D(z, h; \bar{h}) \). Choose \( \varepsilon > 0 \) small enough that \( h_1(x) \) is bounded below by \( \varepsilon \) for all \( x \in X \). Then for \( |h - h_1| < \varepsilon \), \( |m(z, h_1 - h_2) - m(z, h_1) - D(z, h - h_2)| = \left| [h_1(x) - h_2(x)] h_1(x) h_2(x) - h_1(x) h_1(x) h_2(x) \right| \leq C \|h_1(x) - h_2(x)\| \leq C \|h - h_1\| \) and \( |D(z, h; h)| \leq C \|h\| \leq C \|h_1\| \). Let \( \alpha = k/2 \). Then for \( n \alpha = [\ln(n)/\alpha \phi(k)]^{1/2} + s \), \( \sqrt{n} \alpha n^2 = \ln(n) \alpha \phi(k) / \sqrt{n} + 2 \ln(n) \) and \( \sqrt{n} \alpha \phi(k) / \sqrt{n} \to 0 \) by \( \ln(n) \phi(k) \to 0 \), \( \sqrt{n} \phi(k) \to 0 \), implying \( \alpha \to 0 \) faster than some power of \( n \), and by \( \alpha + s > k/2 \). Also, \( \ln(n) \alpha \phi(k) / \sqrt{n} \to 0 \) implies that \( \sqrt{n} \alpha \phi(k) / \sqrt{n} \to 0 \), so that the rate hypotheses of Lemma 5.4 are satisfied. Thus, the conclusion of Lemma 5.4 holds, with \( m(h) = \int D(z, h; h_2) dF(z) = \int r(x(t)) f_0(x(t))^{-1} [h_2(x(t)) - g_0(x(t))] h_1(x(t)) f_0(x(t)) \right| \right| \). Let \( \omega(t) = r(x(t)) f_0(x(t))^{-1} f_0(t) [g_0(x(t)) - g_0(x(t))] \). This function is bounded and continuous a.e. and zero outside a compact set by continuity of \( f_0 \), \( f_0 \), and \( g_0 \), and by the assumption about \( r \). The other conditions of Assumption 5.1 are also satisfied by hypothesis. Furthermore, \( \sqrt{n} \alpha \phi(k) / \sqrt{n} \to 0 \) and
\[ \sqrt[\nu_0]{a^{1/2}} = \sqrt[\nu]{a} \to \infty \text{ by hypothesis and } \alpha \leq k - \alpha = k_2 + k_1/2. \] Thus, the conclusion of Lemma 5.3 holds, for \( V \) in (14). Then, by the triangle inequality, and \( \beta_0 = E[m(z, h_0)] \),

\[ \sqrt{\nu}\sum_i |m(z_i, h) - E[m(z, h_0)]|/n \]

\[ = \sqrt{\nu}\sum_i |m(z_i, h_0) - E[m(z, h_0)]|/n \]

\[ + \sqrt{\nu}\sum_i |m(h) - m(h_0)| + o_p(1) \xrightarrow{a} N(0, V). \]

because \( \sqrt{\nu}\sum_i |m(z, h) - E[m(z, h_0)]|/n \leq 0 \text{ as } a \to 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 \( D(z, h; \beta, h) = D(z, h; \beta) \), 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 \( \delta_{00} = 0, n^{a^{1/2-\alpha}/\ln(n)} \to \infty \), and the fact that this last condition implies \( s > 3k/2 - 2\alpha 3k/2 + k_1/2 > k_2 + k_1/2 = k - \alpha. \) The second conclusion then follows by the conclusion of Lemma 5.5. \( \blacksquare \)

**Proof of Theorem 4.2.** Let \( m(z, h) = a_0(z)h_1(x)^{-1}h_2(x) \) and \( D(z, h; \beta, h) = a_0(z)h_1(x)^{-1}h_2(x) - [h_1(x)/h_1(x)]h_1(x) \). 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 \( b_0(z) \). Also, here \( m(h) = E[f(z, h, h_0)] df(z) = E[a_0(z)f_0(x)|h_2(x) - g_0(h_1(x))] = E[E[a_0(z)|h_2(x)] f_0(x)|h_2(x) - g_0(h_1(x))] \).

By hypothesis, the conditions of Lemma 5.3 are satisfied for this \( f(x) \), so that by the conclusion of Lemma 5.3, for \( \delta = \delta_0(z, h_0) \), \( \delta_0(z, h_0) \), \( \delta_0(z, h_0) \), and \( \delta_0(z, h_0) \), one obtains \( \sqrt{\nu}\sum_i |m(h) - m(h_0)| = \sum_i |\delta_i - E[\delta_i]|/\sqrt{\nu} + o_p(1) \). The first conclusion then follows. Also, the second conclusion follows from Lemma 5.5 similarly to the proof of Theorem 4.1. \( \blacksquare \)

**Proof of Lemma 5.1.** It follows by standard results (e.g., [15]) that \( \sup_{\beta \in \mathcal{B}} |n^{-1/2} \sum_{i=1}^n m(z_i, \beta, h_0) - E[m(z, \beta, h_0)]| \xrightarrow{P} 0 \) and \( E[m(z, \beta, h_0)] \) is continuous in \( \beta \). Also, by Theorem 2.2, \( \sup_{\beta \in \mathcal{B}} |n^{-1/2} \sum_{i=1}^n m(z_i, \beta, h_0) - m(z_i, \beta, h_0)| \leq n^{-1} \sum_i |\delta_i|/\sqrt{n} \to 0 \) so the conclusion follows by \( T \).

**Proof of Lemma 5.2.** By the Frobenius theorem, \( E[m(h)] = m(E[h]) \). Also, by standard results, \( \sup_{\beta \in \mathcal{B}} E[m(h)](x) - m(h_0)] = O(\sigma) \) for any compact set \( \mathcal{B} \). Then by \( \sigma(x) \) zero outside a compact set \( \mathcal{E} \), \( \sum_i |m(z_i, \beta, h_0) - E[m(z, \beta, h_0)]| \leq \sqrt{\nu}\mathcal{C} \sup_{\beta \in \mathcal{B}} |E[m(h)](x) - m(h_0)]| = O(\sqrt{\nu}) \to 0 \). Let \( \delta_i' = [\int f(x)K(x-x)dx]y_i = [\int f(x)\kappa(x)dx]y_i \), where the last equality follows by a change of variables \( u = (x - x)/\alpha \), so that \( m(h) = \sum_{i=1}^n \delta_i'/\alpha \). By \( \kappa(u) \) having bounded support, \( \nu(x_0 + \alpha u)K(u) \leq b(x_0)K(u) \) for all small enough \( \sigma \) and \( b(x_0)K(u) \) \( du < \infty \). Then by DCT, \( \delta_i' \to \delta_i \) with probability 1 as \( \sigma \to 0 \). Also, \( \delta_i' \to \delta_i \) with probability 1 as \( \sigma \to 0 \). Then by M, \( \sum_i \delta_i' - E(\delta_i') \to 0 \), so \( \sqrt{n}[m(h) - m(h_0)] = \sqrt{n}[\delta_i - E(\delta_i')]/\sqrt{n} + o(1) = \sum_i (\delta_i - E(\delta_i))/\sqrt{n} + \sum_i (\delta_i - E(\delta_i'))/\sqrt{n} + o(1) = \sum_i (\delta_i - E(\delta_i))/\sqrt{n} + o_p(1). \)

**Proof of Lemma 5.3.** Note that \( E[m(h)] = m(E[h]) \), so by \( \omega(t) \) bounded and zero outside \( T \), and by \( x(t) \) bounded on \( T \), it follows that \( \sqrt{\nu}\sum_i |E[m(h)](x) - m(h_0)| \leq \sqrt{\nu}\sum_i |E[m(h)](x) - m(h)| \to 0 \). Therefore, it suffices to show that \( \sqrt{\nu} \times \)
Let $\mathcal{K}(u)$ denote $\delta'\mathcal{K}(u)/\partial u'$ and $\rho_s(x) = x^{a_0 - k-1} \times f_\omega(t) [I \otimes \mathcal{K}'((x(t) - x)/\alpha)] dt = \sigma^{a_0 - k-1} f_\omega(t) [I \otimes \mathcal{K}'((x(t) + \alpha v - x(t))/\alpha)] dv$, where $I$ is an identity matrix with the same dimension as $y$ and the last equality follows by the change of variables $v = (t - x(t))/\alpha$. Then $m(\hat{h}) = \sum_i \rho_s(x_i)y_i/n$. Thus, to show $\sqrt{n}\sigma^a [m(\hat{h}) - E[m(\hat{h})]] \stackrel{d}{\rightarrow} N(0, V)$, it suffices, by the Liapunov central limit theorem, to show that $\sigma^a \text{var}(\rho_s(x_i)y_i) \rightarrow V$ and $\sigma^a E[\|\rho_s(x_i)y_i\|^4]/n \rightarrow 0$. By i.i.d. data and $\sqrt{n} \rightarrow \infty$, $\sigma^a E[\rho_s(x_i)y_i - m(h_0)] = \sigma^a E[m(h)] - m(h_0) \rightarrow 0$, and hence $\sigma^a E[\rho_s(x_i)y_i]\| \rightarrow 0$. Therefore, to show $\sigma^a \text{var}(\rho_s(x_i)y_i) \rightarrow V$, it suffices to show that $\sigma^a E[\rho_s(x_i)y_i]\| \rightarrow V$. By $\mathcal{K}(u)$ having a bounded support, $\mathcal{K}(u_i,v)$ is zero for all $v$ outside a bounded set $\mathcal{V}$. Let $\mathcal{T}$ be a compact, convex set containing $\mathcal{V}$ in its interior. Then, for small enough $\sigma$, if $x_2 \notin \mathcal{T}$ then $x_2 + \sigma e \notin \mathcal{T}$ for all $v \in \mathcal{V}$, so $\rho_s(x)$ is zero for $x_2 \notin \mathcal{T}$. For $x_2 \in \mathcal{T}$ and $x_2 + \sigma e \notin \mathcal{T}$, continuous differentiability of $x_1(t)$ and a mean value expansion give $[x_1(x_2 + \sigma e) - x_1(x_2)]/\sigma = [\delta x_1(x_2 + \sigma e)/\sigma] v$, which is bounded over $v \in \mathcal{V}$ and converges to $J(x_1)v$ as $\sigma \rightarrow 0$. Therefore, $\sigma^{a_0 + 2k} \rho_s(x_1(t) - \alpha u_1,t) = \rho_s(x_2 + \alpha u_1,t) = \rho_s(x_1(t) - \alpha u_1,t) f_\omega(t) / f_\omega(t)$, is zero for all $u$ outside a compact set, is bounded, and converges to $\rho_s(x_1(t))/ J(x_1)v$ by the dominated convergence theorem, for $J(t) = \delta x_1(t)/\delta t$. Therefore, by the change of variables $u = [x_1(x_2) - x_1(t)]/\sigma$ and $t = x_2$,

$$\sigma^{a_0 + 2k} E[\rho_s(x_1(t) - \alpha u_1,t) \delta x_1(t)/\delta t] dt \rightarrow V.$$
\[ a^{2a} \sum_{i=1}^{n} (\hat{\delta}_i - \delta_i)^2/n \leq C a^{2a} \sum_{i} \| \hat{D}_i - D_i \|^2/n^2 \leq C a^{2a} \left[ \sum_{i} b_2(z_i)^2/n \right] \left( \sum_{i} \| y_i \|^2/n \right) \]
\[ \times \sigma^{-2a-2}\lambda_1(\| \hat{\beta} - \beta_0 \| + \| \hat{\theta} - \theta_0 \|_{\Delta_2})^2 \]
\[ = O_\rho(\sigma^{-a-2-\lambda_1}(\delta_{2a} + \eta_{2a}^2))^2 = o_\rho(1). \]

By the data i.i.d.,

\[ E\left[ a^{2a} \sum_{i=1}^{n} (\hat{\delta}_i - \delta_i)^2/n \right] \leq a^{2a} E[\| \hat{\delta}_i - \delta_i \|^2] \leq C a^{2a} \]
\[ \times \left( E\left[ \left( \sum_{i=1}^{n} (D_{i1} - \delta_1) \right)^2 \right] \right) \]
\[ + n^{-1} E[\| D_{i1} \|^2] + n^{-1} E[\| \delta_1 \|^2] \]
\[ \leq C a^{2a} n^{-1}(E[\| D_{i1} \|^2] + E[\| D_{i1} \|^2]) \]
\[ \leq C a^{2a} n^{-1} a^{-2k-2\lambda_1} \rightarrow 0, \]

so by M, \( a^{2a} \sum_{i=1}^{n} (\hat{\delta}_i - \delta_i)^2/n \rightarrow 0 \). Under the conditions of Lemma 5.2, it was shown in the proof of Lemma 5.2 that \( E[\| \hat{\delta}_i - \delta_i \|^2] \rightarrow 0 \), so that \( \sum_{i} (\hat{\delta}_i - \delta_i)^2/n \rightarrow 0 \) follows by T. Then \( V \rightarrow V \) follows by T and the law of large numbers. Under the conditions of Lemma 5.3, note that \( \hat{\delta}_i = \rho_\delta(x_i) y_i \) for \( \rho_\delta(x) \) defined in the proof of Lemma 5.3. As shown in that proof, \( \sigma^2 E[\| \hat{\delta}_i \|^2] \rightarrow 0 \), \( \sigma^2 E[\| \hat{\delta}_i \|]\rightarrow V \), and \( n^{-1} \sigma^2 E[\| \delta_i \|^4] \rightarrow 0 \). Therefore, by M, \( \sigma^2 n^{-1} \sum_{i} (\hat{\delta}_i - \delta_i)^2/n \rightarrow V \), so the conclusion follows by T.

\section*{APPENDIX B: TECHNICAL DETAILS}

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 \( \mathcal{X} \), \( h(x) = \sup_{x \in \mathcal{X}} |\partial^\ell h(x)/\partial x^\ell| \).

\textbf{LEMMA B.1.} Suppose that \( E[\| y \|^p] < \infty \) for \( p > 2 \), \( E[\| y \|^p | x \]|_0(x) \) is bounded, \( \mathcal{X} \) is compact, Assumption K is satisfied for \( \Delta \geq \delta \), and \( \sigma = \sigma(n) \) such that \( \sigma(n) \) is bounded and \( n^{1-(2/p)} \sigma(n)^2/\ln(n) \rightarrow \infty \). Then
\[ \| \hat{h} - E[\hat{h}] \|_p = O_\rho(\ln(n)^{1/2}(n^{k+2})^{-1/2}). \] \hfill (B.1)

\textbf{Proof.} It suffices to prove the result for \( y \) a scalar. For each \( \ell \leq j \), by \( K(u) \) 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 an \( \ell \)th order partial derivative of \( \hat{h}(x) \), and \( k(x) \) the corresponding derivative of \( K(x) \), 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 \( y_{i1} = y_{i1} |y_i| \leq P_{n^{1/p}} \), \( y_{i1} = P_{n^{1/p}} \), \( y_{i1} > P_{n^{1/p}} \), \( y_{i1} = -P_{n^{1/p}} \), \( y_{i1} < -P_{n^{1/p}} \). Let \( \hat{H}(x) = n^{-1} \sigma^{-k+\ell} \sum_{i=1}^{n} y_i k((x-x_i)/\sigma) \). Note that by Bonferroni's inequality,
\[ \text{prob}(\bar{H}(x) \neq \bar{H}(x)) \quad \text{for some } (x) \leq \text{prob}(y_n \neq y_i \text{ for some } i \leq n) \]
\[ \leq n \text{prob}(y_n \neq y_i) \leq n \text{prob}(|y_i| > Pn^{1/p}) \]
\[ \leq E(|y_i|^p)/P^{1/p}. \]  
(B.2)

Let \( \delta = \left[ \ln(n)/(na^{k+2}) \right]^{1/2} \). For \( c(x) = E(|y_i|^p|x_i = x|) \) and \( P \) fixed, by \( c(x)f_0(x) \)
bounded and \( p > 2, \)
\[ \delta^{-1}|E[\bar{H}(x)] - E[\bar{H}(x)]| \leq \delta^{-1}\sigma^{-k-l}E[|y_i|^p|x_i = x| |k((x - x_i)/\sigma)|] \]
\[ \leq C\delta^{-1}\sigma^{-k-l}n^{1/(p-1)}E[|y_i|^p|x_i]|k((x - x_i)/\sigma)|] \]
\[ = C\delta^{-1}\sigma^{-l}n^{1/(p-1)}\int |k(v)|c(x - \sigma v)f_0(x - \sigma v) dv \]
\[ = O(\sigma^{-k/2}n^{1/(p-1)/2}/\ln(n)^{1/2}) = o(1). \]  
(B.3)

Next, by \( \bar{k}(x) \) Lipschitz, \( \sup_{x\in E_{x(t)}}|\bar{H}(x) - \bar{H}(x) - \bar{H}(x)| \leq C\sigma(n)^{-1}e^{-k-l-1}. \) Also, by \( \Sigma \) compact, it can be covered by less than \( Cn^{2k} \) open balls of radius \( n^{-1}. \) Let \( x_{jk} \) denote the centers of these open balls, \( (j = 1, \ldots, J(x)) \). Then for \( x_{jk}(x) \) equal to the center of an open ball containing \( x, \) by \( E[\bar{H}(x)] = E[\bar{H}(x)] \)
\[ \sup_{x}[\bar{H}(x) - E[\bar{H}(x)]] \leq \sup_{x}[\bar{H}(x) - E[\bar{H}(x)]] - \sup_{x}[\bar{H}(x) - E[\bar{H}(x)]] \]
\[ = Cn^{1/(p-1)}\sigma^{-k-l-1} + \sup_{j}[\bar{H}(x_{jk}) - E[\bar{H}(x_{jk})]]. \]  
(B.4)

Note that \( \ln(n)^2 \leq Cn \) and, by \( \sigma(n) \) bounded, \( \sigma^{2k} \leq C\sigma^{k+2} \). Then for the constant \( C \) in equation (B.4), it follows by \( p \geq 2 \) and \( na^{k} \to \infty \) that for all \( M, n \) big enough, \( M\delta - Cn^{1/(p-1)}\sigma^{-k-l-1} = M\delta(1 - C/M^2\ln(n)\ln^{2-2/(p-1)}\sigma^{k+2}) > M\delta/2. \) Also, that \( n^{1/(p-1)}\delta = n^{1/(2(p-1))}\sigma^{k}/\ln(n)^{1/2} \to 0. \) As usual for kernel estimators, \( \sigma^{-k-l}E[y_{jk}^2k((x - x_i)/\sigma)] \leq \sigma^{-k-l}E[y_{jk}^2((x - x_i)/\sigma)] \]
\[ \leq \sigma^{-k-l}E[y_{jk}^2((x - x_i)/\sigma)] \leq \sigma^{-k-l}E[y_{jk}^2((x - x_i)/\sigma)] \leq \sigma^{-k-l}E[y_{jk}^2((x - x_i)/\sigma)] \]
\[ \leq \sigma^{-k-l}E[y_{jk}^2((x - x_i)/\sigma)] \leq \sigma^{-k-l}E[y_{jk}^2((x - x_i)/\sigma)] \]
\[ \leq Cn^{1/(p)}\sigma^{-k-l} \]
and Bernstein’s inequality, for \( M \) and \( n \) large enough, \[ \text{prob}(\sup_{x}[\bar{H}(x) - E[\bar{H}(x)]] > M\delta) \]
\[ \leq \text{prob}(\sup_{j}[\bar{H}(x_{jk}) - E[\bar{H}(x_{jk})]] > M\delta/2) \]
\[ \leq 2\sum_{j=1}^{J(x)}\text{prob}(|\bar{H}(x_{jk}) - E[\bar{H}(x_{jk})]] > M\delta/2)| \]
\[ \leq 2\sum_{j=1}^{J(x)}\exp\left(-n^2\delta^2/2n\var{y_{jk}k((x_{jk} - x_i)/\sigma)} + Cn^{1/(p)}\sigma^{-k-l}\delta\right) \]
\[ \leq Cn^{2k}\exp\left(-n\delta^2/C[\sigma^{-k-l} + n^{1/(p)}\sigma^{-k-l}\delta]\right) \]
\[ \leq Cn^{2k}\exp\left(-Cn^2\ln(n)/C\right) \leq C \exp\left(-[CM^2 - 3k] \ln(n)\right). \]  
(B.5)
Because these inequalities hold for any $M, n$ large enough, it follows that $\sup_x |H(x) - E[H(x)]| = O_p(\delta)$. Then by equation (B.3) and the triangle inequality, $\sup_x |H(x) - E[H(x)]| = O_p(\delta)$. Consider any $\epsilon > 0$. Choose $P$ so that $E[|y_i|^2] / P^{1/2} < \epsilon / 2$, so that by equation (B.2), $\prob(\delta^{-1} \sup_x |H(x) - E[H(x)]| > M/2) < \epsilon / 2$ for all $n$. For this fixed $P$, by $\sup_x |H(x) - E[H(x)]| = O_p(\delta)$, there exists $M$ such that $\prob(\delta^{-1} \sup_x |H(x) - E[H(x)]| > M/2 + \prob(\delta^{-1} \sup_x |H(x) - E[H(x)]| > M/2) < \epsilon$. Therefore, by equation (B.2) and the triangle inequality, $\sup_x |H(x) - E[H(x)]| = O_p(\delta)$. The conclusion then follows by applying this conclusion to each derivative of up to order $j$ and by $\sigma$ bounded.

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

**Proof.** Note that $E[H](x) = E[y_i, K_q(x - x_i)] = \int h(t) \{K((-t)/\sigma)/\sigma^k\} dt = \int K(u) h(x + u\sigma) du$ so that by $K(u)$ having finite support, $\partial^j E[H](x) / \partial x^j = \int K(u) \partial^j h(x + u\sigma) / \partial x^j du$. Also, by $\int K(u) du = 1$, it follows that $h_0(x) = \int K(u) x h_0(x) du$ then by a Taylor expansion in $\sigma$ around $\sigma = 0$, for constant matrices $C_t$, $(t = 1, \ldots, j)$,

\[
\|\partial^j h / \partial x^j - \partial^j h_0 / \partial x^j\| \leq \sum_{t=1}^{m-1} \sigma^t C_t \int K(u) \{ \otimes_{t=1}^j u \} \otimes \{ \partial^{j+t} h(x) / \partial x^{j+t} \} du + C_m \sigma^m \int K(u) \{ \otimes_{t=1}^m u \} \otimes \{ \partial^{j+m} h(x + u\sigma) / \partial x^{j+m} \} du \leq C\sigma^m \int \|K(u)\| \|u\|^m du \sup_x \|\partial^{j+m} h(x) / \partial x^{j+m}\| \leq C\sigma^m. \tag{B.6}
\]

**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 $\|h - h_0\|_j = O_p(\ln(n)^{1/2} (n\sigma^k)^{-1/2} + \sigma^s)$.

**Proof.** Follows by Lemmas B.1 and B.2 and the triangle inequality. ■

**Lemma B.4.** If Assumption $K$ is satisfied, $m(h)$ is linear, $|m(h)| \leq C_1 \|h\|_k$, then $E[m(h)] = m(E[h])$.

**Proof.** By $K(u)$ having finite support and $K$ compact, there is a compact set $\mathcal{C}$ such that $\|K_q(\cdot - x)\|_k = 0$, and hence, $m(g(x)K_q(\cdot - x)) = 0$, for all $x \notin \mathcal{C}$. Hence, by linearity of $m(h)$, $E[m(h)] = \int m(g(x)K_q(\cdot - x)f_0(x) dx$ and $E[H(\hat{h})] = \int \int g(x)K_q(\cdot - x)f_0(x) dx$. 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 $J \to \infty$. Then, since $m(g(x)K_q(\cdot - x))$ is continuous and bounded on $\mathcal{C}$, it follows that $\int m(g(x)K_q(\cdot - x))F_j(dx) \to E[m(h)]$. Also, since each derivative of $g(x)K_q(\cdot - x)$ with respect to $\tilde{x}$ of up to order $\Delta$ is bounded and continuous on $\mathcal{C}$, it follows that $\int m(g(x)K_q(\cdot - x) \times \cdots \times $
Further, by $F_j$ having finite support, $m(\int e g(x)K_a(\cdot - x)F_j(dx)) = \int m(g(x) \times K_a(\cdot - x))F_j(dx)$. Then $T$ gives the conclusion.

**Lemma B.5.** If Assumption K is satisfied and for given $\beta$ with $\| \beta \|_A < \infty$, there is linear $D(h)$ with $|m(h) - m(\beta) - D(h - \beta)| = o(\| h - \beta \|_A)$ as $\| h - \beta \|_A \to 0$.

**Proof.** Let $h_t = h + \tau y K_a(\cdot - x)$, so that $|h_t - h_t|_{\Lambda} \leq \tau \| y K_a(\cdot - x) \| \leq C\tau$. Then $|m(h_t) - m(h) - \tau D(y K_a(\cdot - x))| / \tau = |m(h_t) - m(h) - D(h_t - h)| / \tau = o(\| h_t - h_t \| / \tau) = o(1)$. 

$\square$