# Generalized Jackknife Estimators of Weighted Average Derivatives*

Matias D. Cattaneo

Department of Economics, University of Michigan

Richard K. Crump

Federal Reserve Bank of New York

Michael Jansson

Department of Economics, UC Berkeley and *CREATES*

November 29, 2011

Abstract. With the aim of improving the quality of asymptotic distributional approximations for nonlinear functionals of nonparametric estimators, this paper revisits the large-sample properties of an important member of that class, namely a kernel-based weighted average derivative estimator. Asymptotic linearity of the estimator is established under weak conditions. Indeed, we show that the bandwidth conditions employed are necessary in some cases. A bias-corrected version of the estimator is proposed and shown to be asymptotically linear under yet weaker bandwidth conditions. Consistency of an analog estimator of the asymptotic variance is also established. To establish the results, a novel result on uniform convergence rates for kernel estimators is obtained.

*Keywords:* Semiparametric estimation, bias correction, uniform consistency.

*For comments and suggestions, we thank Enno Mammen, Whitney Newey, Jim Powell, Rocio Titiunik, and seminar participants at Brown, CEMFI/Universidad Carlos III de Madrid, Harvard/MIT, Mannheim, New York University, Toulouse School of Economics, and the 2010 World Congress of the Econometric Society. The first author gratefully acknowledges financial support from the National Science Foundation (SES 0921505 and SES 1122994). The third author gratefully acknowledges financial support from the National Science Foundation (SES 0920953 and SES 1124174) and the research support of *CREATES* (funded by the Danish National Research Foundation).

## 1. Introduction

Two-step semiparametric $m$-estimators are an important and versatile class of estimators whose conventional large-sample properties are by now well understood. These procedures are constructed by first choosing a preliminary nonparametric estimator, which is then "plugged in" in a second step to form the semiparametric estimator of the finite-dimensional parameter of interest. Although the precise nature of the high-level assumptions used in conventional approximations varies slightly, it is possible to formulate sufficient conditions so that the semiparametric estimator is $\sqrt{n}$-consistent (where $n$ denotes the sample size) and asymptotically linear (i.e., asymptotically equivalent to a sample average based on the influence function). These results lead to a Gaussian distributional approximation for the semiparametric estimator that, together with valid standard-error estimators, theoretically justify classical inference procedures, at least in large samples. Newey and McFadden (1994, Section 8), Ichimura and Todd (2007, Section 7) and Chen (2007, Section 4), among others, give detailed surveys on semiparametric inference in econometric theory, and further references in statistics and econometrics.

A widespread concern with these conventional asymptotic results is that the (finite sample) distributional properties of semiparametric estimators are widely believed to be much more sensitive to the implementational details of its nonparametric ingredient (e.g., bandwidth choice when the nonparametric estimator is kernel-based) than predicted by conventional asymptotic theory, according to which semiparametric estimators are asymptotically linear with influence functions that are invariant with respect to the choice of nonparametric estimator (e.g., Newey (1994a, Proposition 1)). Conventional approximations rely on sufficient conditions carefully tailored to achieve asymptotic linearity, thereby assuming away additional approximation errors that may be important in samples of moderate size. In particular, whenever the preliminary nonparametric estimator enters nonlinearly in the construction of the semiparametric procedure, a common approach is to linearly approximate the underlying estimating equation in order to characterize the contribution of the nonparametric ingredient to the distributional approximation. This approach leads to the familiar sufficient condition that requires the nonparametric ingredient to converge at a rate faster than $n^{1/4}$, effectively allowing one to proceed "as if" the semiparametric estimator depends linearly on its nonparametric ingredient, which in turn guarantees an asymptotic linear representation of the semiparametric estimator under appropriate sufficient conditions.

In this paper we study the large-sample properties of a kernel-based estimator of weighted average derivatives (Stoker (1986), Newey and Stoker (1993)), and propose a new first-order asymptotic approximation for the semiparametric estimator based on a quadratic expansion of the underlying estimating equation. The key idea is to relax the requirement that the convergence rate of the nonparametric estimator be faster than $n^{1/4}$, and to rely instead on a quadratic expansion to tease out further information about the dependence of the semiparametric estimator on its nonparametric ingredient, thereby improving upon the conventional (first-order) distributional approximation available in the literature. Although our idea leads to an improved understanding of the differences between linear and nonlinear functionals of nonparametric estimators in some generality, we focus attention on weighted average derivatives to keep the results as interpretable as possible, and because this estimand is popular in theoretical and empirical work. (We discuss below its importance and role in the literature.) Indeed, it should be conceptually straightforward to apply the methodology employed herein to other kernel-based semiparametric $m$-estimators at the expense of considerable more notation and technicalities.

We obtain four types of results for the kernel-based weighted average derivatives estimator. First, under standard kernel and bandwidth conditions we establish asymptotic linearity of the estimator and consistency of its associated "plug-in" variance estimator under a weaker-than-usual moment condition on the dependent variable. Indeed, the moment condition imposed would appear to be (close to) minimal, suggesting that these results may be of independent theoretical interest in the specific context of weighted average derivatives. More broadly, the results (and their derivation) may be of interest as they are achieved by judicial choice of estimator, and by employing a new uniform law of large numbers specifically designed with consistency proofs in mind.

Second, we also establish asymptotic linearity of the weighted average derivative estimator under weaker-than-usual bandwidth conditions. This relaxation of bandwidth conditions is of practical usefulness because it permits the employment of kernels of lower-than-usual order (and, relatedly, enables us to accommodate unknown functions of lower-than-usual degree of smoothness). More generally, the derivation of these results may be of interest because of its "generic" nature and because of its ability to deliver an improved understanding of the distributional properties of other semiparametric estimators that depend nonlinearly on a nonparametric component.

These results are based on a stochastic expansion retaining a "quadratic" term that is treated as a "remainder" term in conventional derivations. Retaining this term not only permits the relaxation of sufficient (bandwidth) conditions for asymptotic linearity, but also enables us to establish necessity of these sufficient conditions in some cases and, most importantly, characterize the consequences of further relaxing the bandwidth conditions. Indeed, the third (and possibly most important) type of result we obtain shows that in general the nonlinear dependence on a nonparametric estimator gives rise to a nontrivial "bias" term in the stochastic expansion of the semiparametric estimator. Being a manifestation of the well known curse of dimensionality of nonparametric estimators, this "nonlinearity bias" is a generic feature of nonlinear functionals of nonparametric estimators whose presence can have an important impact on distributional properties of such functionals.

Because the "nonlinearity bias" is due to the (large) variance of nonparametric estimators, attempting to remove it by means of conventional bias reduction methods aimed at reducing "smoothing" bias, such as increasing the order of the kernel, does not work. Nevertheless, it turns out that this "nonlinearity bias" admits a polynomial expansion (in the bandwidth), suggesting that it should be amenable to elimination by means of the method of generalized jackknifing (Schucany and Sommers (1977)). Making this intuition precise is the purpose of the final type of result presented herein. Although some details of this result are specific to our weighted average derivative estimator, the main message is of much more general validity. Indeed, an inspection of the derivation of the result suggests that the fact that removal of "nonlinearity bias" can be accomplished by means of generalized jackknifing is a property shared by most (if not all) kernel-based semiparametric two-step estimators.

Our results are closely related and contribute to the important literature on semiparametric averaged derivatives (Stoker (1986); see also, e.g., Härdle and Stoker (1989), Härdle, Hart, Marron, and Tsybakov (1992) and Horowitz and Härdle (1996)), in particular shedding new light on the problem of semiparametric weighted average derivative estimation (Newey and Stoker (1993)). This problem has wide applicability in statistics and econometrics, as we further discuss in the following section. This problem is conceptually and analytically different from the problem of semiparametric density-weighted average derivatives because a kernel-based density-weighted average derivative estimator depends on the nonparametric ingredient in a linear way (Powell, Stock, and Stoker (1989)), while the kernel-based weighted average derivative estimator has a nonlinear dependence on a nonparametric estimator. As a consequence, the alternative first-order distributional approximation obtained in Cattaneo, Crump and Jansson (2010, 2011) for a kernel-based density-weighted average derivatives estimator is not applicable to the estimator studied herein and our main findings are qualitatively different from those obtained in our earlier work. Indeed, a crucial finding in this paper is that considering "small bandwidth asymptotics" for the kernel-based weighted average derivative estimator leads to a first-order bias contribution to the distributional approximation (rather than a first-order variance contribution, as in the case of the kernel-based density-weighted average derivative estimator), which in turn requires bias-correction of the estimator (rather than adjustment of the standard-error estimates, as in the case of the kernel-based density-weighted average derivative estimator).

From a more general perspective, our findings are also connected to other results in the semiparametric literature. Mammen (1989) studies the large sample properties of a nonlinear least-squares estimator when the (effective) dimension of the parameter space is allowed to increase rapidly, and finds a first-order bias effect qualitatively similar to the one characterized herein. The "nonlinearity bias" we encounter is also analogous in source to the so-called "degrees of freedom bias" discussed by Ichimura and Linton (2005) for the case of a univariate semiparametric estimation problem, but due to the different nature of our asymptotic experiment its presence has first-order consequences herein. Non-negligible biases in models with covariates of large dimension (i.e., "curse of dimensionality" effects of first order) were also found by Abadie and Imbens (2006), but in the case of their matching estimator the bias in question does not seem to be attributable to nonlinearities. Finally, the recent work of Robins, Li, Tchetgen, and van der Vaart (2008) on higher-order influence functions is also related to our results insofar as it relaxes the underlying convergence rate requirement for the nonparametric estimator. Whereas Robins, Li, Tchetgen, and van der Vaart (2008) are motivated by a concern about the plausibility of the smoothness conditions needed to guarantee existence of $n^{1/4}$-consistent nonparametric estimators in models with large-dimensional covariates, our work seeks to relax this underlying convergence rate requirement for the nonparametric estimator in order to improve the accuracy of the distributional approximation even in cases where lots of smoothness is assumed. Indeed, our results highlight the presence of a leading, first-order bias term that is unrelated to the amount of smoothness assumed (but clearly related to the dimensionality of the covariates).

The paper proceeds as follows. Section 2 introduces the model and estimator(s) under study. Our main theoretical results are presented in Section 3, while some Monte Carlo results are given in Section 4. Section 5 offers concluding remarks. Appendix A contains proofs of the theoretical results, while Appendix B contains some auxiliary results (of possibly independent interest) about uniform convergence of kernel estimators.

## 2. Preliminaries

**2.1. Model and Estimand.** We assume that $z_i = (y_i, x_i')'$, $i = 1, \ldots, n$, are i.i.d. observed copies of a vector $z = (y, x')'$, where $y \in \mathbb{R}$ is a dependent variable and $x \in \mathbb{R}^d$ is a continuous explanatory variable with density $f(\cdot)$. A weighted average derivative of the regression function $g(x) = \mathbb{E}[y|x]$ is defined as

$$\theta = \mathbb{E}\left[w(x)\frac{\partial}{\partial x}g(x)\right], \qquad (1)$$

where $w(x)$ is a known scalar weight function. (Further restrictions on $w(\cdot)$ will be imposed below.) This is an important estimand which has been widely considered in both theoretical and empirical work, as we discuss in the following well-known examples.

**Example 1: Semi-linear Single-Index Models.** Let $x = (x_1', x_2')'$ and $g(x) = G(x_1'\beta, x_2)$ with $G(\cdot)$ unknown and partition $\theta$ conformably with $x$ as $\theta = (\theta_1', \theta_2')'$. Under appropriate assumptions, the parameter of interest $\beta$ is proportional to $\theta_1$ because

$$\theta_1 = \mathbb{E}\left[w(x)\dot{G}_1(x_1'\beta, x_2)\right]\beta, \qquad \dot{G}_1(u, x_2) = \frac{\partial}{\partial u}G(u, x_2).$$

This setup covers several problems of interest. For example, single-index limited dependent variable models (e.g., discrete choice, censored and truncated models) are included with $G(x_1'\beta, x_2) = \phi(x_1'\beta)$, possibly $x_1 = x$, and $\phi(\cdot)$ the so-called link function. Another class of problems fitting in this example are partially linear models where $G(x_1'\beta, x_2) = \phi_1(x_1'\beta + \phi_2(x_2))$ with $\phi_1(\cdot)$ a link function and $\phi_2(\cdot)$ another unknown function. For further discussion on these and related examples see, e.g., Stoker (1986), Härdle and Stoker (1989), Newey and Stoker (1993) and Powell (1994). □

**Example 2: Non-Separable Models.** Let $x = (x_1', x_2')'$ and $y = m(x_1, \varepsilon)$ with $m(\cdot)$ unknown and $\varepsilon$ an unobserved random variable. Under appropriate assumptions, including $x_1 \perp\!\!\!\perp \varepsilon \mid x_2$, a population parameter of interest is given by

$$\theta_1 = \mathbb{E}\left[w(x)\frac{\partial}{\partial x_1}m(x_1, \varepsilon)\right] = \mathbb{E}\left[w(x)\frac{\partial}{\partial x_1}g(x_1, x_2)\right],$$

which captures the (averaged) marginal effect of $x_1$ on $m(\cdot)$ over the population $(x_1', \varepsilon)'$. As in the previous example, $\theta_1$ is the first component of the weighted average derivative $\theta$ partitioned conformably with $x$. The parameter $\theta_1$ is of interest in policy analysis and treatment effect models. A canonical example is given by the linear random coefficients model $y = \beta_0(\varepsilon) + x_1'\beta_1(\varepsilon)$, where the parameter of interest reduces to $\theta_1 = \mathbb{E}\left[w(x)\beta_1(\varepsilon)\right]$ under appropriate assumptions. (When $w(x) = 1$, $\theta_1$ is the so-called average partial effect.) For further discussion on averaged derivatives in non-separable models see, e.g., Matzkin (2007) and Imbens and Newey (2009). $\square$

**Example 3: Applications in Economics.** In addition to the examples discussed above, which are statistical in nature, weighted average derivatives have also been employed in several specific economic applications that do not necessarily fit the previous setups. Some examples are: $(i)$ Stoker (1989) proposed several tests statistics based on averaged derivatives obtained from economic-theory restrictions such as homogeneity or symmetry of cost functions, $(ii)$ Härdle, Hildenbrand, and Jerison (1991) developed a test for the law of demand using weighted average derivatives, $(iii)$ Deaton and Ng (1998) employed averaged derivatives to estimate the effect of a tax and subsidy policy change on individuals' behavior, and $(iv)$ Coppejans and Sieg (2005) developed a test for non-linear pricing in labor markets based on averaged derivatives obtained from utility maximization. $\square$

The previous examples highlight the applicability of weighted average derivatives in statistics and econometrics. The next section introduces the kernel-based estimator studied in this paper, and reviews some known results in the literature.

**2.2. Estimator and Known Results.** Newey and Stoker (1993) studied estimands of the form (1) and gave conditions under which the semiparametric variance bound for $\theta$ is

$$\Sigma = \mathbb{E}[\psi(z)\psi(z)'], \tag{2}$$

where $\psi(\cdot)$, the pathwise derivative of $\theta$, is given by

$$\psi(z) = w(x)\frac{\partial}{\partial x}g(x) - \theta + [y - g(x)]\, s(x),$$

$$s(x) = -\frac{\partial}{\partial x}w(x) + w(x)\ell(x), \qquad \ell(x) = -\frac{\partial f(x)/\partial x}{f(x)}.$$

The following assumption, which we make throughout the paper, guarantees existence of the parameter $\theta$ and semiparametrically efficient estimators thereof.

**Assumption 1.** (a) For some $S \geq 2$, $\mathbb{E}[|y|^S] < \infty$ and $\mathbb{E}[|y|^S|x]f(x)$ is bounded.

(b) $\mathbb{E}[\psi(z)\psi(z)']$ is positive definite.

(c) $w$ is continuously differentiable, and $w$ and its first derivative are bounded.

(d) $\inf_{x\in\mathcal{W}} f(x) > 0$, where $\mathcal{W} = \left\{x \in \mathbb{R}^d : w(x) > 0\right\}$.

(e) For some $P_f \geq 2$, $f$ is $(P_f + 1)$ times differentiable, and $f$ and its first $(P_f + 1)$ derivatives are bounded.

(f) $g$ is continuously differentiable, and $e$ and its first derivative are bounded, where $e(x) = f(x)g(x)$.

(g) $\lim_{\|x\|\to\infty} [f(x) + |e(x)|] = 0$, where $\|\cdot\|$ is the Euclidean norm.

The restrictions imposed by Assumption 1 are fairly standard and relatively mild, with the possible exception of the "fixed trimming" condition in part (d). This condition simplifies the exposition in our paper, allowing us to avoid tedious technical arguments. It may be relaxed to allow for non-random asymptotic trimming, but we decided not to pursue this extension to avoid cumbersome notation and other associated technical distractions.

Under Assumption 1 it follows from integration by parts that $\theta = \mathbb{E}\left[ys\left(x\right)\right]$. A kernel-based analog estimator of $\theta$ is therefore given by

$$\hat{\theta}_n(h_n) = \frac{1}{n}\sum_{i=1}^{n} y_i \hat{s}_n(x_i; h_n), \qquad \hat{s}_n(x_i; h_n) = -\frac{\partial}{\partial x}w(x) - w(x)\frac{\partial \hat{f}_n(x; h_n)/\partial x}{\hat{f}_n(x; h_n)},$$

where

$$\hat{f}_n(x; h_n) = \frac{1}{nh_n^d}\sum_{j=1}^{n} K\left(\frac{x - x_j}{h_n}\right)$$

for some kernel $K : \mathbb{R}^d \to \mathbb{R}$ and some positive (bandwidth) sequence $h_n$. As defined, $\hat{\theta}_n$ depends on the user-chosen objects $K$ and $h_n$, but because our main interest is in the sensitivity of the properties of $\hat{\theta}_n$ with respect to the bandwidth $h_n$, we suppress the dependence of $\hat{\theta}_n$ on $K$ in the notation (and make the dependence on $h_n$ explicit). The following assumption about the kernel $K$ will be assumed to hold.

**Assumption 2.** (a) $K$ is even.

(b) $K$ is twice differentiable, and $K$ and its first two derivatives are bounded.

(c) $\int_{\mathbb{R}^d} \|\dot{K}\left(u\right)\|(1 + \|u\|^2)du < \infty$, where $\dot{K}(u) = \partial K(u)/\partial u$.

(d) For some $P_K \geq 2$, $\int_{\mathbb{R}^d} |K\left(u\right)|(1 + \|u\|^{P_K+1})du < \infty$ and

$$\int_{\mathbb{R}^d} u_1^{l_1} \cdots u_d^{l_d} K(u)du = \begin{cases} 1, & \text{if } l_1 = \cdots = l_d = 0, \\ 0, & \text{if } (l_1, \ldots, l_d)' \in \cup_{k=1}^{P_K-1}\mathbb{Z}_+^d(k), \end{cases}$$

where $\mathbb{Z}_+^d(k) = \left\{(l_1, \ldots, l_d)' \in \mathbb{Z}_+^d : l_1 + \ldots + l_d = k\right\}$.

(e) $\int_{\mathbb{R}^d} \bar{K}(u)du < \infty$, where $\bar{K}(u) = \sup_{\|r\|\geq u} \|\partial(K(r), \dot{K}(r)')/\partial r\|$.

With the possible exception of Assumption 2 (e), the restrictions imposed on the kernel are fairly standard. Assumption 2 (e) is inspired by Hansen (2008) and holds if $K$ has bounded support or if $K$ is a normal density-based higher-order kernel obtained as in, e.g., Robinson (1988).

If Assumptions 1 and 2 hold (with $P_f$ and $P_K$ large enough) it is easy to give conditions on the bandwidth $h_n$ under which $\hat{\theta}_n$ is asymptotically linear with influence function $\psi(\cdot)$. For instance, proceeding as in Newey (1994a, 1994b) it can be shown that if Assumptions 1 and 2 hold and if

$$nh_n^{2P} \to 0, \qquad P = \min(P_f, P_K) \tag{3}$$

and

$$\frac{nh_n^{2d+4}}{(\log n)^2} \to \infty, \tag{4}$$

then

$$\hat{\theta}_n(h_n) = \theta + n^{-1}\sum_{i=1}^{n} \psi(z_i) + o_p\left(n^{-1/2}\right). \tag{5}$$

Moreover, under the same conditions the variance $\Sigma$ in (2) is consistently estimable, as we discussed in more detail in Section 3.3. The lower bound on $h_n$ implied by the condition (4) helps ensure that the estimation error of the nonparametric estimator $\hat{f}_n$ is $o_p\left(n^{-1/4}\right)$ in an appropriate (Sobolev) norm, which in turn is a high-level assumption featuring prominently in Newey's (1994a) work on asymptotic normality of semiparametric $m$-estimators (and in more recent refinements thereof, such as Chen, Linton, and van Keilegom (2003)).

This paper explores the consequences of employing bandwidths that are "small" in the sense that (4) is violated. Four types of results will be derived. The first result, given in Theorem 1 below, gives sufficient conditions for (5) that involve a weaker lower bound on $h_n$ than (4). For $d \geq 3$, the weaker lower bound takes the form $nh_n^{2d} \to \infty$. The second result, given in Theorem 2 below, shows that $nh_n^{2d} \to \infty$ is also necessary for (5) to hold (if $d \geq 3$). More specifically, Theorem 2 finds that if $d \geq 3$, then $\hat{\theta}_n$ has a non-negligible bias when $nh_n^{2d} \nrightarrow \infty$. The third result, given in Theorem 3 below, shows that while $nh_n^{2d} \to \infty$ is necessary for asymptotic linearity of $\hat{\theta}_n$ (when $d \geq 3$), a bias-corrected version of $\hat{\theta}_n$ enjoys the property of asymptotic linearity under the weaker condition

$$\frac{nh_n^{\frac{3}{2}d+1}}{(\log n)^{3/2}} \to \infty. \tag{6}$$

Finally, Theorem 4 shows that a modest strengthening of Assumption 1 (a) is sufficient to obtain consistency of the conventional plug-in standard-error estimator even when the lower bound on the bandwidth is given by (6).

**Remark.** Newey and McFadden (1994, pp. 2212-2214) establish asymptotic linearity of the alternative kernel-based estimator

$$\breve{\theta}_n(h_n) = \frac{1}{n}\sum_{i=1}^{n} w(x_i)\frac{\partial}{\partial x}\hat{g}_n(x_i; h_n), \qquad \hat{g}_n(x; h_n) = \frac{1}{nh_n^d}\sum_{j=1}^{n} y_j K\left(\frac{x - x_j}{h_n}\right)/\hat{f}_n(x; h_n),$$

under (3)−(4) and assumptions similar to Assumptions 1 and 2. Their analysis requires $S \geq 4$ in order to handle the presence of $\hat{g}_n$. The fact that $\hat{\theta}_n$ does not involve $\hat{g}_n$ enables us to develop distribution theory for it under the seemingly minimal condition $S = 2$.

## 3. Results

Validity of the stochastic expansion (5) can be established by exhibiting an approximation $\hat{\theta}_n^A$ (say) to $\hat{\theta}_n$ satisfying the following trio of conditions:

$$\hat{\theta}_n(h_n) - \hat{\theta}_n^A = o_p\left(n^{-1/2}\right), \tag{7}$$

$$\hat{\theta}_n^A - \mathbb{E}[\hat{\theta}_n^A] = n^{-1}\sum_{i=1}^{n}\psi(z_i) + o_p\left(n^{-1/2}\right), \tag{8}$$

$$\mathbb{E}[\hat{\theta}_n^A] - \theta = o\left(n^{-1/2}\right). \tag{9}$$

Variations of this approach have been used in numerous papers, the typical choice being to obtain $\hat{\theta}_n^A$ by "linearizing" $\hat{\theta}_n$ with respect to the nonparametric estimator $\hat{f}_n$ and then establishing (7) by showing in particular that the estimation error of $\hat{f}_n$ is $o_p\left(n^{-1/4}\right)$ in a suitable norm. This general approach is now well-established in semiparametrics; see, e.g., Newey and McFadden (1994, Section 8), Ichimura and Todd (2007, Section 7), Chen (2007, Section 4), and references therein.

**3.1. Asymptotic Linearity: Linear vs. Quadratic Approximations.** In the context of averaged derivatives, conventional "linearization" amounts to setting $\hat{\theta}_n^A$ equal to

$$\hat{\theta}_n^*(h_n) = n^{-1}\sum_{i=1}^{n} y_i \hat{s}_n^*(x_i; h_n),$$

where

$$\hat{s}_n^*(x_i; h_n) = s(x) - \frac{w(x)}{f(x)}\left[\frac{\partial}{\partial x}\hat{f}_n(x; h_n) + \ell(x)\hat{f}_n(x; h_n)\right]$$

is obtained by linearizing $\hat{s}_n$ with respect to $\hat{f}_n$. With this choice of $\hat{\theta}_n^A$, conditions (7) – (9) will hold if Assumptions 1 and 2 are satisfied and if (3) – (4) hold. In particular, (4) serves as part of what would appear to be the best known sufficient condition for the estimation error of $\hat{f}_n$ (and its derivative) to be $o_p\left(n^{-1/4}\right)$, a property which in turn is used to establish (7) when $\hat{\theta}_n^A = \hat{\theta}_n^*(h_n)$.

In an attempt to establish (7) under a bandwidth condition weaker than (4), we set $\hat{\theta}_n^A$ equal to a "quadratic" approximation to $\hat{\theta}_n(h_n)$ given by

$$\hat{\theta}_n^{**}(h_n) = n^{-1}\sum_{i=1}^{n} y_i \hat{s}_n^{**}(x_i; h_n),$$

where

$$\hat{s}_n^{**}(x_i;h_n) = \hat{s}_n^{*}(x_i;h_n) + \frac{w(x)}{f(x)^2}\left[\hat{f}_n(x;h_n) - f(x)\right]\left[\frac{\partial}{\partial x}\hat{f}_n(x;h_n) + \ell(x)\hat{f}_n(x;h_n)\right].$$

The use of a quadratic approximation to $\hat{\theta}_n$ gives rise to a "cubic" remainder in (7), suggesting that it suffices to require that the estimation error of $\hat{f}_n$ (and its derivative) be $o_p\left(n^{-1/6}\right)$. In fact, the proof of the following result shows that the somewhat special structure of the estimator (i.e., the fact that $\hat{s}_n$ is linear in the derivative of $\hat{f}_n$) can be exploited to establish sufficiency of a slightly weaker condition.

**Theorem 1.** *Suppose Assumptions 1 and 2 are satisfied and suppose* (3) *holds. Then* (5) *is true if either (i)* $d = 1$ *and* $nh_n^3 \to \infty$*, (ii)* $d = 2$ *and* $nh_n^4/\left(\log n\right)^{3/2} \to \infty$*, or (iii)* $d \geq 3$ *and* $nh_n^{2d} \to \infty$*.*

The proof of Theorem 1 verifies $(7)-(9)$ for $\hat{\theta}_n^A = \hat{\theta}_n^{**}(h_n)$. Because the lower bounds on $h_n$ imposed in cases (i) through (iii) are weaker than (4) in all cases, working with $\hat{\theta}_n^{**}$ when analyzing $\hat{\theta}_n$ has the advantage that it enables us to weaken the sufficient conditions for asymptotic linearity to hold on the part of $\hat{\theta}_n$. Notably, existence of a bandwidth sequence satisfying the assumptions of Theorem 1 holds whenever $P > d$, a weaker requirement than the restriction $P > d + 2$ implied by the conventional conditions $(3)-(4)$. In other words, Theorem 1 justifies the use of kernels of lower order, and thus requires less smoothness on the part of the density $f$, than do analogous results obtained using $\hat{\theta}_n^A = \hat{\theta}_n^{*}(h_n)$. Moreover, working with $\hat{\theta}_n^{**}$ enables us to derive necessary conditions for (5) in some cases.

**Theorem 2.** *Suppose Assumptions 1 and 2 are satisfied and suppose* (3) *and* (6) *hold. Then*

$$\mathbb{E}[\hat{\theta}_n^{**}(h_n)] - \theta = n^{-1}h_n^{-d}\mathcal{B}_0 + o\left(n^{-1/2} + n^{-1}h_n^{-d}\right), \tag{10}$$

*where*

$$\mathcal{B}_0 = \left(-K(0)I_d + \int_{\mathbb{R}^d}\left[K(u)^2 I_d + K(u)\dot{K}(u)u'\right]du\right)\int_{\mathbb{R}^d} g(r)w(r)\ell(r)dr.$$

*Moreover,*

$$\hat{\theta}_n(h_n) - \mathbb{E}[\hat{\theta}_n^{**}(h_n)] = n^{-1}\sum_{i=1}^{n}\psi(z_i) + o_p\left(n^{-1/2}\right)$$

*if either (i)* $d = 1$ *and* $nh_n^3 \to \infty$ *or (ii)* $d \geq 2$.

The first part of Theorem 2 is based on an asymptotic expansion of the approximate bias $\mathbb{E}[\hat{\theta}_n^{**}(h_n)] - \theta$ and shows that, in general, the condition $nh_n^{2d} \to \infty$ is necessary for (9) to hold when $\hat{\theta}_n^A = \hat{\theta}_n^{**}(h_n)$. (We know of no "popular" kernels and/or "plausible" examples of $g(\cdot)$, $w(\cdot)$, and $\ell(\cdot)$ for which $\mathcal{B}_0 = 0$.) The second part of Theorem 2 verifies (7) – (8) for $\hat{\theta}_n^A = \hat{\theta}_n^{**}(h_n)$ and can be combined with the first part to yield the result that the sufficient condition $nh_n^{2d} \to \infty$ obtained in Theorem 1 (iii) is also necessary (in general) when $d \geq 3$.

To interpret the matrix $\mathcal{B}_0$ in the (approximate) bias expression (10), it is instructive to decompose it as $\mathcal{B}_0 = \mathcal{B}_0^* + \mathcal{B}_0^{**}$, where

$$\mathcal{B}_0^* = -K(0)\int_{\mathbb{R}^d} g(r)w(r)\ell(r)dr$$

and

$$\mathcal{B}_0^{**} = \left(\int_{\mathbb{R}^d}\left[K(u)^2 I_d + K(u)\dot{K}(u)u'\right]du\right)\int_{\mathbb{R}^d} g(r)w(r)\ell(r)dr.$$

The term $\mathcal{B}_0^*$ is a "leave in" bias term arising because each $\hat{s}_n(x_i; h_n)$ employs a nonparametric estimator $\hat{s}_n$ which uses the own observation $x_i$. The other bias term, $\mathcal{B}_0^{**}$, is a "nonlinearity" bias term reflecting the fact that $\hat{s}_n^{**}$ involves a nonlinear function of $\hat{f}_n$. The magnitude of this nonlinearity bias is $n^{-1}h_n^{-d}$. This magnitude is exactly the magnitude of the pointwise variance of $\hat{f}_n$, which is no coincidence because $\hat{s}_n^{**}$ involves a term which is "quadratic" in

$\hat{f}_n$. (The approximation $\hat{s}_n^{**}$ also involves a cross-product term in $\hat{f}_n$ and its derivative which, as shown in the proof of Lemma A-3, gives rise to a bias term of magnitude $n^{-1}h_n^{-d}$ when $K$ is even.)

The second part of Theorem 2 suggests that if $d \geq 3$, then a bias-corrected version of $\hat{\theta}_n$ might be asymptotically linear even if the condition $nh_n^{2d} \to \infty$ is violated.

**Remarks.** (i) The leave-in-bias can be avoided simply by employing a "leave-one-out" estimator of $f$ when forming $\hat{s}_n$. ($\mathcal{B}_0^* = 0$ when $K(0) = 0$.)

(ii) Merely removing leave-in-bias does not automatically render $\hat{\theta}_n$ asymptotically linear unless $nh_n^{2d} \to \infty$, however, as the nonlinearity bias of the leave-one-out version of $\hat{\theta}_n$ is identical to that of $\hat{\theta}_n$ itself. ($\mathcal{B}_0 = \mathcal{B}_0^{**} \neq 0$ when $K(0) = 0$.)

(iii) Manipulating the order of the kernel ($P_K$) does not eliminate the nonlinearity bias, as its magnitude is invariant with respect to the order of the kernel. ($\mathcal{B}_0^{**} \neq 0$ for all $P_K \geq 2$.)

**3.2. Asymptotic Linearity under Non-standard Conditions.** The method of generalized jackknifing can be used to arrive at an estimator $\tilde{\theta}_n$ (say) whose (approximate) bias is sufficiently small also when $nh_n^{2d} \nrightarrow \infty$. It can be shown that if the assumptions of Theorem 2 hold, then the (approximate) bias $\mathbb{E}[\hat{\theta}_n^{**}(h_n)] - \theta$ admits a polynomial (in $h_n$) expansion of the form

$$\mathbb{E}[\hat{\theta}_n^{**}(h_n)] - \theta = n^{-1}h_n^{-d}\mathcal{B}_0 + \sum_{j=1}^{\lfloor (P-1)/2 \rfloor} n^{-1}h_n^{2j-d}\mathcal{B}_j^{**} + o\left(n^{-1/2}\right), \tag{11}$$

where $\left\{\mathcal{B}_j^{**} : 1 \leq j \leq \lfloor (P-1)/2 \rfloor\right\}$ are constants capturing (higher order) nonlinearity bias. Accordingly, let $J$ be a positive integer with $J < 1 + d/2$, let $c = (c_0, \ldots, c_J)'$ be a vector of distinct constants with $c_0 = 1$, and define

$$\begin{pmatrix} \lambda_0(c) \\ \lambda_1(c) \\ \vdots \\ \lambda_J(c) \end{pmatrix} = \begin{pmatrix} 1 & 1 & \cdots & 1 \\ 1 & c_1^{-d} & \cdots & c_J^{-d} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & c_1^{2(J-1)-d} & \cdots & c_J^{2(J-1)-d} \end{pmatrix}^{-1} \begin{pmatrix} 1 \\ 0 \\ \vdots \\ 0 \end{pmatrix}.$$

It follows from (11) that if the assumptions of Theorem 2 hold and if $J \geq (d-2)/8$, then

$$\sum_{j=0}^{J} \lambda_j(c)\, \mathbb{E}[\hat{\theta}_n^{**}(c_j h_n)] - \theta = o\left(n^{-1/2}\right).$$

As a consequence, we have the following result about the (generalized jackknife) estimator

$$\tilde{\theta}_n(h_n, c) = \sum_{j=0}^{J} \lambda_j(c)\hat{\theta}_n(c_j h_n).$$

**Theorem 3.** *Suppose Assumptions 1 and 2 are satisfied and suppose* (3) *and* (6) *hold. If* $(d-2)/8 \leq J < 1 + d/2$, *then*

$$\tilde{\theta}_n(h_n, c) = \theta + n^{-1} \sum_{i=1}^{n} \psi(z_i) + o_p\left(n^{-1/2}\right)$$

*if either (i)* $d = 1$ *and* $nh_n^3 \to \infty$ *or (ii)* $d \geq 2$.

Theorem 3 gives a simple recipe for constructing an estimator of $\theta$ which is semiparametrically efficient under relatively mild restrictions on the rate at which the bandwidth $h_n$ vanishes.

**Remarks.** (i) An alternative, and perhaps more conventional, method of bias correction would employ (nonparametric) estimators of $\mathcal{B}_0$ and $\left\{\mathcal{B}_j^{**}\right\}$ and subtract an estimator of $\mathbb{E}[\hat{\theta}_n^{**}(h_n)] - \theta$ from $\hat{\theta}_n(h_n)$. In our view, generalized jackknifing is attractive from a practical point of view precisely because there is no need to explicitly (characterize and) estimate complicated functionals such as $\mathcal{B}_0$ and $\{\mathcal{B}_j^{**}\}$.

(ii) Our results demonstrate by example that a more nuanced understanding of the bias properties of $\hat{\theta}_n$ can be achieved by working with a "quadratic" (as opposed to "linear") approximation to it. It is conceptually straightforward to go further and work with a "cubic" approximation (say) to $\hat{\theta}_n$. Doing so would enable a further relaxation of the bandwidth condition at the expense of a more complicated "bias" expression, but would not alter the fact that generalized jackknifing could be used to eliminate also the bias terms that become non-negligible under the relaxed bandwidth conditions. The small simulation evidence presented in Section 4 suggests that eliminating the biases characterized in (11) suffices for the purposes of rendering the bias of the estimator negligible relative to its standard deviation in many cases, so for brevity we omit results based on a "cubic" approximation to $\hat{\theta}_n$.

**3.3. Standard Errors.** The results presented above describe a novel approach to obtain a first-order asymptotic linear approximation for $\hat{\theta}_n(h_n)$ even when the classical conditions imposed in the literature are not satisfied. For inference purposes it is important to also have a consistent standard-error estimator. If Assumptions 1 and 2 hold, and under the conventional bandwidth restrictions (3) – (4), it is not difficult to show that the asymptotic variance $\Sigma$ in (2) is consistently estimable. Specifically, it follows from Theorem 4 below that

$$\hat{\Sigma}_n = \frac{1}{n}\sum_{i=1}^{n} \hat{\psi}_n(z_i)\hat{\psi}_n(z_i)' \to_p \Sigma, \tag{12}$$

where

$$\hat{\psi}_n(z) = \hat{\psi}_n(z; h_n) = w(x)\frac{\partial}{\partial x}\hat{g}_n(x; h_n) - \hat{\theta}_n(h_n) + [y - \hat{g}_n(x; h_n)]\,\hat{s}_n(x; h_n),$$

$$\hat{g}_n(x;h_n) = \frac{1}{nh_n^d} \sum_{j=1}^{n} y_j K\left(\frac{x-x_j}{h_n}\right) / \hat{f}_n(x;h_n).$$

Importantly, parts (ii) and (iii) of the following result establishes consistency of the variance estimator $\hat{\Sigma}_n$ under the same weaker conditions on the bandwidth entertained in the previous section.

**Theorem 4.** *Suppose Assumptions 1 and 2 are satisfied and suppose* (3) *and* (6) *hold. Then* (12) *is true if either (i)* $S = 2$ *and* $nh_n^{2d+2}/(\log n)^2 \to \infty$, *(ii)* $d = 1$, $nh_n^3 \to \infty$, *and* $S > 3$, *or (iii)* $S \geq 3 + 2/d$.

Part (i) of the theorem gives a condition (on $h_n$) for consistency of $\hat{\Sigma}_n$ under the (seemingly) minimal moment requirement that $S = 2$, while parts (ii) and (iii) gives conditions (on $S$) for consistency of $\hat{\Sigma}_n$ to hold under the assumptions of Theorem 3. The proof of Theorem 4 utilizes a (seemingly) novel uniform consistency result kernel estimators (and their derivatives), given in Appendix B. It does not seem possible to establish part (i) using existing uniform consistency results for kernel estimators, as we are unaware of any such results (for objects like $\hat{g}_n$) that require only $S = 2$. For instance, a proof of (12) based on Newey (1994b, Lemma B.1) requires $S > 4 - 4/(d+2)$ when the lower bound on the bandwidth is of the form $nh_n^{2d+2}/(\log n)^2 \to \infty$. (When the lower bound on the bandwidth is of the form (6), Newey (1994b, Lemma B.1) can be applied if $d \geq 2$ and $S > 6 - 8/(d+2)$.)

## 4. Small Simulation Study

We conducted a small Monte Carlo experiment to investigate the finite-sample properties of our procedure for weighted average derivatives. We report results for both the conventional estimator $\hat{\theta}_n(h_n)$ and the generalized jackknife estimator $\tilde{\theta}_n(h_n, c)$.

The data generating process is a Tobit model $y_i = \tilde{y}_i \mathbf{1}\{\tilde{y}_i \geq 0\}$ with $\tilde{y}_i = x_i'\beta + \varepsilon_i$, so that $\theta = \beta\, \mathbb{E}[w(x)\Phi(x'\beta)]$, where $\Phi(\cdot)$ is the standard normal cdf. We assume that $\varepsilon_i \sim$ i.i.d. $\mathcal{N}(0,1)$ and are independent of the covariates. The dimension of the covariates, $d$, is set equal to three and all three components of $\beta$ are set to unity. The vector of covariates is generated as $x_i \sim$ i.i.d. $\mathcal{N}(0, I_3)$. For simplicity, only results for the first component of $\theta = (\theta_1, \theta_2, \theta_3)'$ are reported. As for the choice of weight function, we use

$$w(x;\gamma,\kappa) = \prod_{j=1}^{d} \exp\left[-\frac{x_j^{2\kappa}}{\tau(\gamma)^{2\kappa}\left(\tau(\gamma)^{2\kappa} - x_j^{2\kappa}\right)}\right] \mathbf{1}\{|x_j| < \tau(\gamma)\}.$$

The parameter $\kappa$ governs the degree of approximation between $w(\cdot)$ and the rectangular function, the approximation becoming more precise as $\kappa$ grows. (Being discontinuous, $w(\cdot)$ violates Assumption 1(c), so strictly speaking our theory does not cover the chosen weight function.) For specificity, we set $\kappa = 2$. Keeping in mind that the covariates are jointly standard normal, the trimming parameter $\tau(\gamma)$ is given by $\tau(\gamma) = \Phi^{-1}\left(1 - (1 - \sqrt[d]{1-\gamma})/2\right)$, where $\gamma$ is the (symmetric) nominal amount of trimming (i.e., $\gamma = 0.15$ implies a nominal trimming of 15% of the observations).

The number of simulations is set to $S = 1,000$, and we consider samples of size $n = 200$. We report results implemented by Gaussian density-based multiplicative kernels with $P = 4$. (Note that since $d = 3$, choice of $P = 4$ not would not be available under the conventional conditions (3) – (4).) In these simulations we choose a value of $\gamma$ equal to $\gamma = 0.15$. Finally, when implementing the generalized jackknife estimator we consider pairs of constants of the form, $(c_1, c_2) = (\exp(-\delta), \exp(\delta))$ where $\delta \in \{0.05, 0.10\}$; however, it should be noted that the qualitative conclusions are little changed for other choices of jackknife constants.

Figure 1 presents graphs of the standardized bias of each estimator, $\hat{\theta}_n(h_n)$ and $\tilde{\theta}_n(h_n, c)$, for a grid of bandwidth choices $h_n$. The standardized bias is defined as the bias divided by the standard deviation of the estimator across all $S$ simulations, where the purpose of the rescaling is to improve the interpretability of the bias results. Specifically, this rescaling ensures that the severity (or otherwise) of bias problems can be gauged simply by looking at the graph and utilizing well known facts about the standard normal distribution used for approximation purposes when constructing the confidence intervals. Consistent with our theory, the conventional estimator is severely biased whereas there is a region of small bandwidths for which the generalized jackknife estimator has negligible (normalized) bias. These results highlight the potential sensitivity of the conventional estimator to perturbations of the bandwidth choice, which in the case of the weighted average derivatives leads to a non-trivial bias for "small" bandwidths, and therefore a clear need for bias correction.

Figure 2 illustrates the quality of the normal approximation to the distribution of the $t$-statistic. Here we estimate a smoothed density of the $t$-statistic which has been normalized by its (simulation) standard deviation so that the variance is one. In each figure, we consider a choice of bandwidth that leads to the best possible empirical coverage rate for the corresponding confidence interval. For example, for a sample size of $n = 200$ the $t$-statistic density is estimated using the a choice of bandwidth of $h_n = 0.275$ and $h_n = 0.85$ for the generalized jackknife estimator ($\delta = 0.05$) and conventional estimator, respectively. For simplicity, in this simulations we did not explore the performance of "optimal" bandwidth selectors, but rather decided to focus on the "best case scenario" for this Monte Carlo experiment. Both figures suggest that the densities are well-approximated by the normal distribution. Moreover, and consistent with the evidence presented in Figure 1, the estimated density for the normalized $t$-statistic based on $\tilde{\theta}_n\left(h_n, c\right)$ is approximately centered correctly while this is not the case for the conventional estimator $\hat{\theta}_n\left(h_n\right)$.

Finally, we also explored the empirical coverage rates of the conventional and bias-corrected $t$-statistics. We found that neither the conventional nor the jackknife estimator succeeded in achieving empirical coverage rates near the nominal rate. This finding, together with the results reported above, suggests that the lack of good empirical coverage of the associated confidence intervals for the generalized jackknife procedure is due to the poor performance of the classical variance estimator commonly employed in the literature. Indeed, in the case of the conventional procedure, we found that both the bias properties and

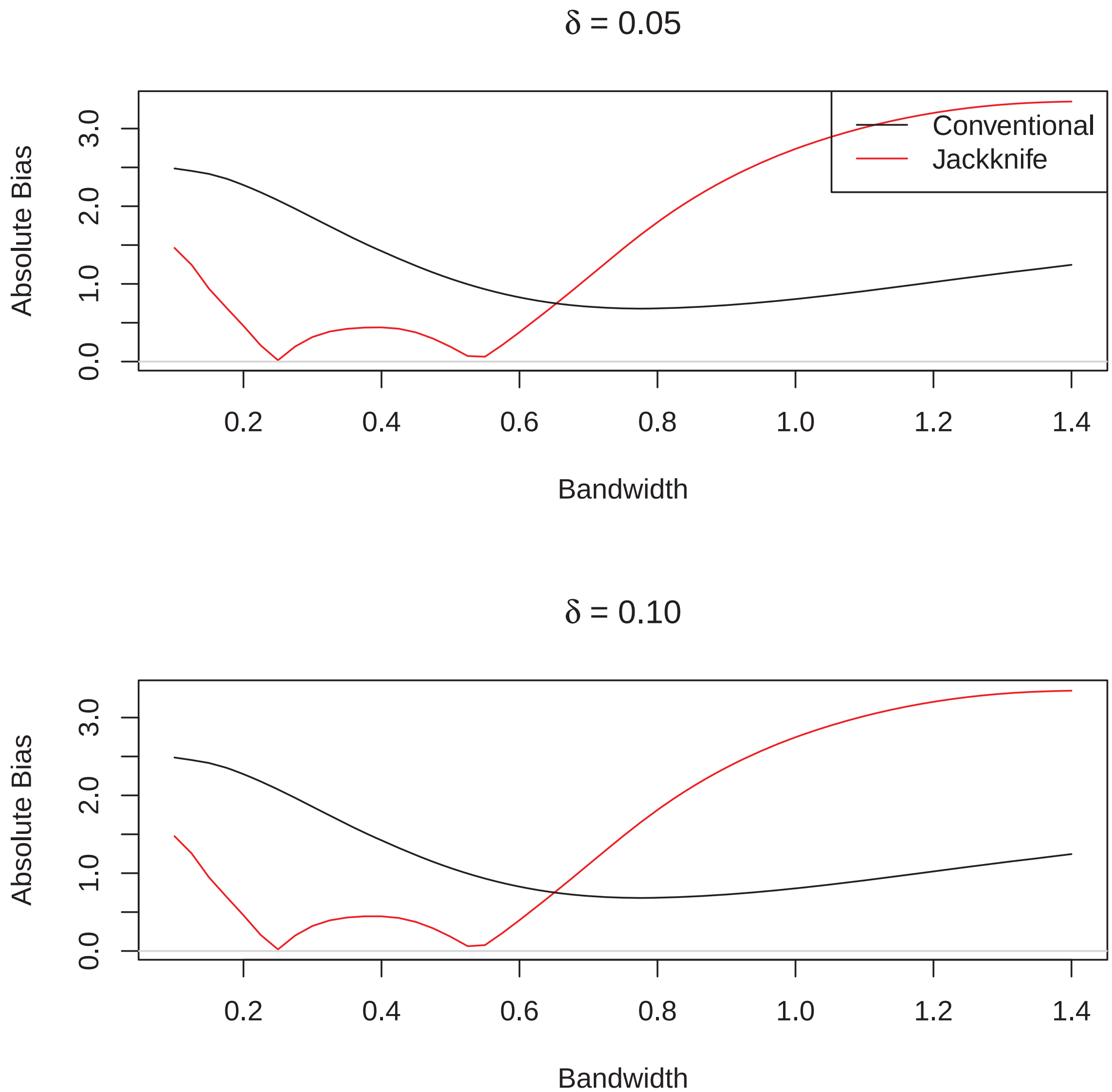


Figure 1: Standardized Bias

the performance of this variance estimator seem to be at fault for the disappointing empirical coverage rates found in the simulations. Further investigation into alternative variance estimation procedures, although beyond the scope of this paper, is underway.

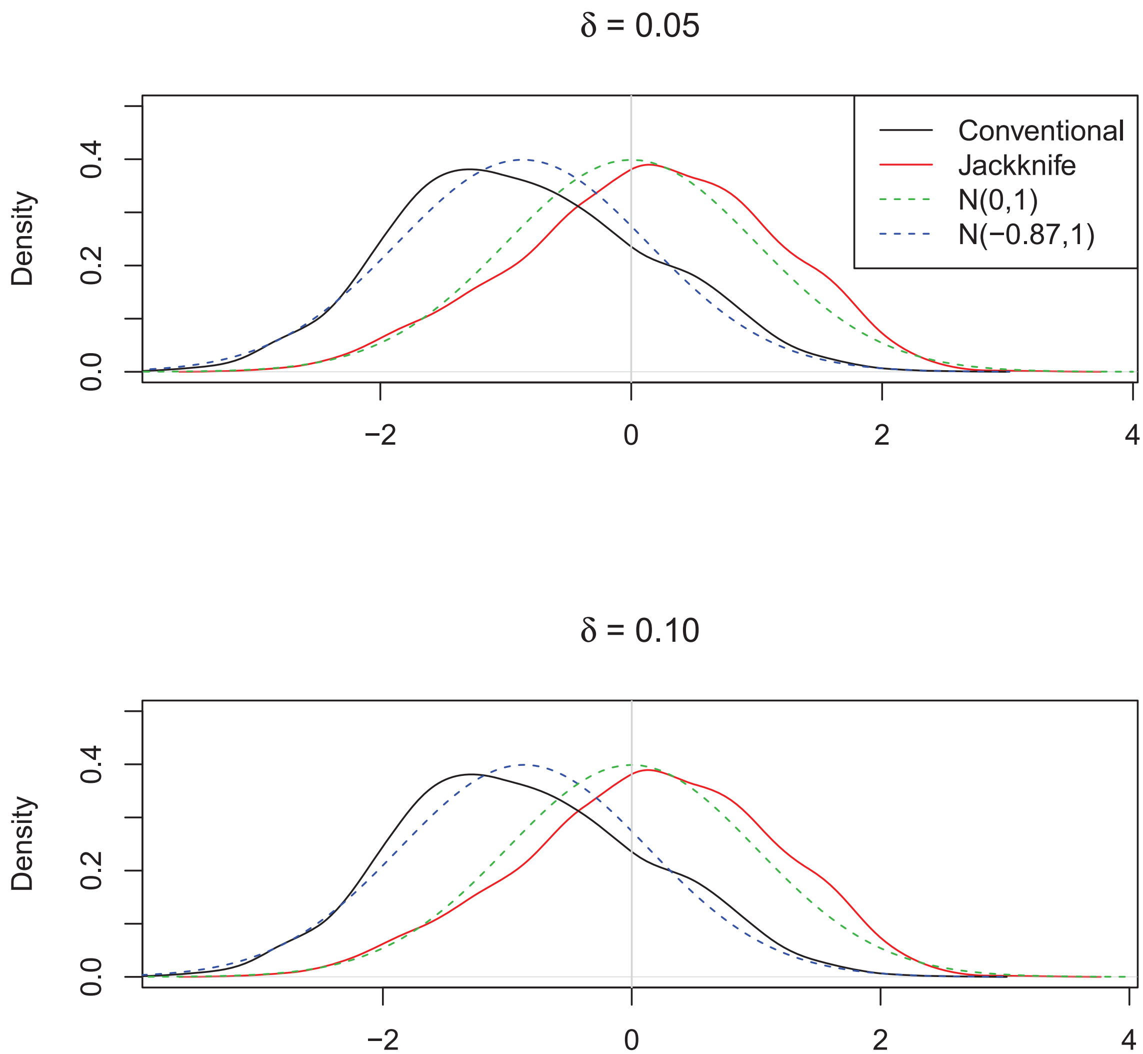


Figure 2: Normal Approximation

## 5. Conclusion

This paper has revisited the large-sample properties of a kernel-based weighted average derivative estimator. In important respects this estimator can be viewed as a representative member of the much larger class of (kernel-based) semiparametric $m$-estimators. In particular, the "nonlinearity bias" highlighted by our development of asymptotics with smaller-than-usual bandwidths (i.e., larger-than-usual undersmoothing) is a generic feature of nonlinear functionals of nonparametric estimators and is likely to be quantitatively important in samples of moderate size also for estimators other than the one studied in this paper.

To remove this "nonlinearity bias", we have employed the method of generalized jackknifing. Being "semi-automatic" in the sense that it requires knowledge only of the magnitudes of the terms in an asymptotic expansion of the "nonlinearity bias", that same method should be easily applicable whenever the nonparametric ingredient is a kernel estimator, as the variance properties of kernel estimators are very well understood. Partly because certain popular nonparametric estimators (notably series estimators) have variance properties that seem harder to analyze than those of kernel estimators, it would be useful to know if the validity of certain "fully automatic" bias correction methods and/or distributional approximations can be established under assumptions similar to those entertained in this paper.

## 6. Appendix A: Proofs

This appendix gives the proofs of Theorems 1-3. We first state four lemmas, the proofs of which are available in the supplemental appendix. We then employ these lemmas, together with the results for kernel-based estimators outlined in Appendix B, to prove the main theorems.

**6.1. Useful lemmas.** The first lemma gives sufficient conditions for (7) in terms of the magnitudes of

$$\Delta_{0,n}(h_n) = \sup_{x \in \mathcal{W}} \left| \hat{f}_n(x; h_n) - f(x) \right|$$

and

$$\Delta_{1,n}(h_n) = \max\left\{ \Delta_{0,n}(h_n),\ \sup_{x \in \mathcal{W}} \left\| \frac{\partial}{\partial x} \hat{f}_n(x; h_n) - \frac{\partial}{\partial x} f(x) \right\| \right\}.$$

**Lemma A-1.** *Suppose Assumption 1 is satisfied and suppose $\Delta_{0,n}(h_n) = o_p(1)$. Then (7) is true if either (i) $\hat{\theta}_n^A = \hat{\theta}_n^{**}(h_n)$ and $\Delta_{0,n}(h_n)^2 \Delta_{1,n}(h_n) = o_p\left(n^{-1/2}\right)$ or (ii) $\hat{\theta}_n^A = \hat{\theta}_n^{*}(h_n)$ and $\Delta_{0,n}(h_n)\Delta_{1,n}(h_n) = o_p\left(n^{-1/2}\right)$.*

The next result gives sufficient conditions for (8).

**Lemma A-2.** *Suppose Assumptions 1 and 2 are satisfied and suppose $h_n \to 0$ and $nh_n^{d+2} \to \infty$. Then (8) is true for $\hat{\theta}_n^A = \hat{\theta}_n^{**}(h_n)$ and $\hat{\theta}_n^A = \hat{\theta}_n^{*}(h_n)$.*

The following result can be used to evaluate $\mathbb{E}[\hat{\theta}_n^A(h_n)] - \theta$.

**Lemma A-3.** *Suppose Assumptions 1 and 2 are satisfied and suppose $h_n \to 0$. Then*

$$\mathbb{E}\left[\hat{\theta}_n^{*}(h_n)\right] - \theta = n^{-1} h^{-d} \mathcal{B}_0^{*} + O\left(h_n^P\right),$$

*and*

$$\mathbb{E}\left[\hat{\theta}_n^{**}(h_n) - \hat{\theta}_n^{*}(h_n)\right] = \sum_{j=0}^{\lfloor (P-1)/2 \rfloor} n^{-1} h_n^{2j-d} \mathcal{B}_j^{**} + O\left(n^{-1} h_n^{P-d} + n^{-2} h_n^{-2d} + h_n^{2P}\right),$$

*where, for $j \geq 1$,*

$$\mathcal{B}_j^{**} = \frac{1}{(2j)!} \sum_{l \in \mathbb{Z}_+^d(2j)} B_K(l)\, B_z(l) + \frac{1}{(2j+1)!} \sum_{l \in \mathbb{Z}_+^d(2j+1)} \dot{B}_K(l)\, \dot{B}_z(l),$$

$$B_K(l) = \int_{\mathbb{R}^d} u_1^{l_1} \cdots u_d^{l_d} K(u)^2 \, du, \qquad B_z(l) = \int_{\mathbb{R}^d} g(r) \frac{w(r)}{f(r)} \ell(r) \frac{\partial^j}{\partial r_1^{l_1} \cdots \partial r_d^{l_d}} f(r) \, dr,$$

$$\dot{B}_K(l) = \int_{\mathbb{R}^d} u_1^{l_1} \cdots u_d^{l_d} K(u) \dot{K}(u) \, du, \qquad \dot{B}_z(l) = -\int_{\mathbb{R}^d} g(r) \frac{w(r)}{f(r)} \frac{\partial^j}{\partial r_1^{l_1} \cdots \partial r_d^{l_d}} f(r) \, dr.$$

The last lemma collects basic results about kernels-based integrals. Let $\mathcal{K}(x; h) = h^{-d} K(x/h)$ and $\dot{\mathcal{K}}(x; h) = \partial \mathcal{K}(x; h) / \partial x$.

**Lemma A-4.** *Suppose Assumptions 1 and 2 are satisfied and suppose $h_n \to 0$. Then*
*(a) Uniformly in $x \in \mathcal{W}$,*

$$b(x; h_n) = \int_{\mathbb{R}^d} \mathcal{K}(x - r; h_n) f(r) \, dr - f(x) = O\left(h_n^P\right),$$

$$\dot{b}(x; h_n) = \int_{\mathbb{R}^d} \dot{\mathcal{K}}(x - r; h_n) f(r) \, dr - \partial f(x) / \partial x = O\left(h_n^P\right).$$

*(b) For any function $F$ with $\mathbb{E}[F(z)^2] < \infty$,*

$$\mathbb{E}\left[F(z_1)^2 \mathcal{K}(x_1 - x_2; h_n)^2\right] = O\left(h_n^{-d}\right),$$

$$\mathbb{E}\left[F(z_1)^2 \left\| \dot{\mathcal{K}}(x_1 - x_2; h_n) \right\|^2\right] = O\left(h_n^{-(d+2)}\right).$$

*(c) For any function $F$ with $\mathbb{E}[F(z)^2] < \infty$,*

$$\mathbb{E}\left[F(z_1)^2 \mathcal{K}(x_1 - x_2; h_n)^2 \mathcal{K}(x_1 - x_3; h_n)^2\right] = O\left(h_n^{-2d}\right),$$

$$\mathbb{E}\left[F(z_1)^2 \mathcal{K}(x_1 - x_2; h_n)^2 \left\| \dot{\mathcal{K}}(x_1 - x_3; h_n) \right\|^2\right] = O\left(h_n^{-2(d+1)}\right).$$

**6.2. Proof of Theorems 1-3.** Under the assumptions of the theorems, (7) – (8) hold for $\hat{\theta}_n^A = \hat{\theta}_n^{**}(h_n)$. Validity of (8) follows from Lemma A-2, while (7) follows from Lemma

A-1 because it can be shown that

$$\sup_{x\in\mathcal{W}}\left|\hat{f}_n(x;h_n)-f(x)\right|=O_p\left(h_n^P+\sqrt{\frac{\log n}{nh_n^d}}\right) \tag{A-1}$$

and

$$\sup_{x\in\mathcal{W}}\left\|\frac{\partial}{\partial x}\hat{f}_n(x;h_n)-\frac{\partial}{\partial x}f(x)\right\|=O_p\left(h_n^P+\sqrt{\frac{\log n}{nh_n^{d+2}}}\right). \tag{A-2}$$

Specifically, (A-1) holds because $\sup_{x\in\mathcal{W}}\left|\mathbb{E}[\hat{f}_n(x;h_n)]-f(x)\right|=O\left(h_n^P\right)$ by Lemma A-4 (a) and because

$$\sup_{x\in\mathcal{W}}\left|\hat{f}_n(x;h_n)-\mathbb{E}[\hat{f}_n(x;h_n)]\right|=O_p\left(\sqrt{\frac{\log n}{nh_n^d}}\right)$$

by Lemma B-1 with $(Y,X)=(1,x)$, $\kappa=K$, and $\mathcal{X}_n=\mathcal{W}$. Similarly, (A-2) can be shown by applying Lemma A-4 (a) and Lemma B-1 (with $\kappa(u)=h_n\partial K(u)/\partial u_l$ for $l=1,\ldots,d$).

Theorem 1 is a special case of Theorem 2. To complete the proof of Theorem 2, use Lemma A-3 to verify (9). Similarly, the proof of Theorem 3 can be completed by using Lemma A-3 to verify (11). ■

**6.3. Proof of Theorem 4.** It suffices to show that

$$\frac{1}{n}\sum_{i=1}^{n}\left\|\hat{\psi}_n(z_i)-\psi(z_i)\right\|^2=o_p(1).$$

To do so, it suffices to show that

$$\hat{\theta}_n(h_n)-\theta=o_p(1), \tag{A-3}$$

$$\sup_{x\in\mathcal{W}}\|\hat{s}_n(x;h_n)-s(x)\|=o_p(1), \tag{A-4}$$

$$\sup_{x\in\mathcal{W}}\|\hat{g}_n(x;h_n)-g(x)\|=o_p(1), \tag{A-5}$$

$$\sup_{x\in\mathcal{W}}\left\|\frac{\partial}{\partial x}\hat{g}_n\left(x;h_n\right)-\frac{\partial}{\partial x}g\left(x\right)\right\|=o_p\left(1\right), \tag{A-6}$$

It follows from Theorem 2 and its proof that (A-3) – (A-4) hold. Also, Lemma B-1 (with $(Y,X)=(y,x)$, $s=S$, $\kappa=K$, and $\mathcal{X}_n=\mathcal{W}$) and routine arguments can be used to show that if Assumptions 1 and 2 are satisfied and if (3) and (6) hold, then (A-5) will be implied by $n^{1-1/S}h_n^d/\log n\to\infty$. Similarly, (A-6) can be established under the condition $n^{1-1/S}h_n^{d+1}/\log n\to\infty$. The latter holds if condition (i), (ii), or (iii) in the statement of the theorem is satisfied. ■

## 7. Appendix B: Uniform Convergence Rates for Kernel Estimators

This Appendix derives uniform convergence rates for kernel estimators. Lemma B-1 is used in the proofs of the main results of this paper. Because this result may be of independent interest, it is stated at a (slightly) greater level of generality than needed in the proofs of the other results in this paper.

Suppose $(Y_i,X_i')'$, $i=1,\ldots,n$, are i.i.d. copies of $(Y,X')'$, where $X\in\mathbb{R}^d$ is continuous with density $f_X(\cdot)$. Consider the nonparametric estimator

$$\hat{\Psi}_n\left(x\right)=n^{-1}h_n^{-d}\sum_{j=1}^{n}Y_j\kappa\left(\frac{x-X_j}{h_n}\right),$$

where $h_n$ is a bandwidth sequence and $\kappa:\mathbb{R}^d\to\mathbb{R}$ is a kernel-like function. To obtain uniform convergence rates for $\hat{\Psi}_n$, we make the following assumptions.

**Assumption B1.** For some $s\geq 2$, $\mathbb{E}[|Y|^s]+\sup_{x\in\mathbb{R}^d}\mathbb{E}[|Y|^s|X=x]f_X\left(x\right)<\infty$.

**Assumption B2.** (a) $\sup_{u\in\mathbb{R}^d}|\kappa(u)|+\int_{\mathbb{R}^d}|\kappa(u)|\,du<\infty$.

(b) $\kappa$ admits a $\delta_\kappa>0$ and a function $\kappa^*:\mathbb{R}^d\to\mathbb{R}_+$ with $\sup_{u\in\mathbb{R}^d}\kappa^*(u)+\int_{\mathbb{R}^d}\kappa^*(u)du<\infty$, such that $|\kappa(u)-\kappa(u^*)|\leq\|u-u^*\|\kappa^*(u^*)$ whenever $\|u-u^*\|\leq\delta_\kappa$.

**Remark.** Assumption B2(b) is adapted from Hansen (2008). It holds if $\kappa$ is differentiable with $\bar{\kappa}(0) + \int_{\mathbb{R}^d} \bar{\kappa}(u)du < \infty$, where $\bar{\kappa}(u) = \sup_{\|r\| \geq u} \|\partial \kappa(r)/\partial r\|$.

The first result gives an upper bound on the convergence rate of $\hat{\Psi}_n$ on (possibly) expanding sets of the form $\mathcal{X}_n = \left\{x \in \mathbb{R}^d : \|x\| \leq C_{X,n}\right\}$, where $C_{X,n}$ is a positive sequence satisfying

$$\overline{\lim}_{n\to\infty} \frac{\log(C_{X,n})}{\log n} < \infty. \tag{B-1}$$

**Lemma B-1.** *Suppose Assumptions B1 and B2 are satisfied and suppose (B-1) holds. If $h_n \to 0$ and $n^{1-1/s}h_n^d/\log n \to \infty$, then*

$$\sup_{x\in\mathcal{X}_n} \left|\hat{\Psi}_n(x) - \Psi_n(x)\right| = O_p(\rho_n), \qquad \rho_n = \sqrt{\frac{\log n}{nh_n^d}} \max\left(1, \sqrt{\frac{\log n}{n^{1-2/s}h_n^d}}\right),$$

*where* $\Psi_n(x) = \mathbb{E}\left[\hat{\Psi}_n(x)\right]$.

**Remark.** The natural "$s = \infty$" analog of Lemma B-1 holds if $Y$ is bounded (e.g., if $Y \equiv 1$, as in the case of density estimation). In other words, the lower bound $nh_n^d/\log n \to \infty$ suffices and $\rho_n$ can be set equal to $\sqrt{\log n/(nh_n^d)}$ when $Y$ is bounded.

Lemma B-1 generalizes Newey (1994b, Lemma B.1) in a couple of respects. First, by borrowing ideas from Hansen (2008) we are able to accommodate kernels with unbounded support and to establish uniform convergence over certain types of expanding sets. More importantly (for our purposes at least), Lemma B-1 relaxes the condition $n^{1-2/s}h_n^d/\log n \to \infty$ imposed by Newey (1994b, Lemma B.1). In typical applications of Newey (1994b, Lemma B.1), a condition like $s \geq 4$ is imposed in order to ensure that $n^{1-2/s}h_n^d/\log n \to \infty$ is implied by "natural" conditions on $h_n$, such as $nh_n^{2d}/(\log n)^2 \to \infty$ (e.g., Newey (1994b, Theorem 4.2), Newey and McFadden (1994, Theorem 8.11)). In contrast, only $s \geq 2$ is required for the condition imposed in Lemma B-1 to be implied by $nh_n^{2d}/(\log n)^2 \to \infty$.

If $n^{1-2/s}h_n^d/\log n \to 0$, then the uniform rate obtained in Lemma B-1 falls short of the "usual" rate $\sqrt{nh_n^d/\log n}$. This is potentially problematic if Lemma B-1 is used to establish uniform convergence with a certain rate (e.g., $n^{1/4}$ or $n^{1/6}$, as in proofs of results such as (7)). On the other hand, the slower rate of convergence is of no concern when any rate of convergence will do (as in proofs of consistency results such as (12)).

Because of their ability to control bias in some cases, leave one out estimators of the form

$$\hat{\Psi}_{n,i}(x) = \frac{1}{(n-1)\,h_n^d} \sum_{j=1, j\neq i}^{n} Y_j \kappa\left(\frac{x - X_j}{h_n}\right)$$

are sometimes of interest. The next result extends Lemma B-1 to such estimators.

**Lemma B-2.** *Suppose Assumptions B1 and B2 are satisfied and suppose (B-1) holds. If $h_n \to 0$ and $n^{1-1/s}h_n^d/\log n \to \infty$, then*

$$\max_{1\leq i\leq n} \sup_{x\in\mathcal{X}_n} \left|\hat{\Psi}_{n,i}(x) - \Psi_{n,i}(x)\right| = O_p(\rho_n), \qquad \Psi_{n,i}(x) = \mathbb{E}[\hat{\Psi}_{n,i}(x)].$$

Another corollary of Lemma B-1 is the following result, which can be useful when uniform convergence on the support of the empirical distribution of $X$ suffices.

**Lemma B-3.** *Suppose $\mathbb{E}[\|X\|^{s_X}] < \infty$ for some $s_X > 0$ and suppose Assumptions B1 and B2 are satisfied. If $h_n \to 0$ and $n^{1-1/s}h_n^d/\log n \to \infty$, then*

$$\max_{1\leq i\leq n} \left|\hat{\Psi}_n(X_i) - \Psi_n(X_i)\right| = O_p(\rho_n)$$

*and*

$$\max_{1\leq i\leq n} \left|\hat{\Psi}_{n,i}(X_i) - \Psi_{n,i}(X_i)\right| = O_p(\rho_n).$$

**Remark.** Lemmas B-2 and B-3 are not used elsewhere in the paper. We have included them because they may be of independent interest.

## References

Abadie, A., and G. W. Imbens (2006): "Large Sample Properties of Matching Estimators for Average Treatment Effects," *Econometrica*, 74(1), 235–267.

Cattaneo, M. D., R. K. Crump, and M. Jansson (2010): "Robust Data-Driven Inference for Density-Weighted Average Derivatives," *Journal of the American Statistical Association*, 105(491), 1070–1083.

——— (2011): "Small Bandwidth Asymptotics for Density-Weighted Average Derivatives," forthcoming *Econometric Theory*.

Chen, X. (2007): "Large Sample Sieve Estimation of Semi-Nonparametric Models," in *Handbook of Econometrics, Volume VI*, ed. by J. J. Heckman, and E. Leamer, pp. 5549–5632. Elsevier Science B.V., New York.

Chen, X., O. Linton, and van Keilegom (2003): "Estimation of Semiparametric Models when The Criterion Function Is Not Smooth," *Econometrica*, 71(5), 1591–1608.

Coppejans, M., and H. Sieg (2005): "Kernel Estimation of Average Derivatives and Differences," *Journal of Business and Economic Statistics*, 23, 211–225.

Deaton, A., and S. Ng (1998): "Parametric and Nonparametric Approaches to Price and Tax Reform," *Journal of the American Statistical Association*, 93, 900–909.

Hansen, B. E. (2008): "Uniform Convergence Rates for Kernel Estimation with Dependent Data," *Econometric Theory*, 24, 726–748.

Härdle, W., J. Hart, J. Marron, and A. Tsybakov (1992): "Bandwidth Choice for Average Derivative Estimation," *Journal of the American Statistical Association*, 87, 218–226.

Härdle, W., W. Hildenbrand, and M. Jerison (1991): "Empirical Evidence on the Law of Demand," *Econometrica*, 59, 1525–1549.

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

Horowitz, J., and W. Härdle (1996): "Direct Semiparametric Estimation of Single-Index Models with Discrete Covariates," *Journal of the American Statistical Association*, 91, 1632–1640.

Ichimura, H., and O. Linton (2005): "Asymptotic Expansions for some Semiparametric Program Evaluation Estimators," in *Identification and Inference in Econometric Models: Essays in Honor of Thomas J. Rothenberg*, ed. by D. W. K. Andrews, and J. H. Stock, pp. 149–170. Cambridge University Press, New York.

Ichimura, H., and P. E. Todd (2007): "Implementing Nonparametric and Semiparametric Estimators," in *Handbook of Econometrics, Volume VIB*, ed. by J. Heckman, and E. Leamer, pp. 5370–5468. Elsevier Science B.V.

Imbens, G. W., and W. K. Newey (2009): "Identification and estimation of triangular simultaneous equations models without additivity," *Econometrica*, 77(5), 1481–1512.

Mammen, E. (1989): "Asymptotics with Increasing Dimension for Robust Regression with Applications to the Bootstrap," *Annals of Statistics*, 17, 382–400.

Matzkin, R. L. (2007): "Nonparametric Identification," in *Handbook of Econometrics, Volume VIB*, ed. by J. Heckman, and E. Leamer, pp. 5307–5368. Elsevier Science B.V.

Newey, W. K. (1994a): "The Asymptotic Variance of Semiparametric Estimators," *Econometrica*, 62, 1349–1382.

——— (1994b): "Kernel Estimation of Partial Means and a General Variance Estimator," *Econometric Theory*, 10, 233–253.

NEWEY, W. K., AND D. L. MCFADDEN (1994): "Large Sample Estimation and Hypothesis Testing," in *Handbook of Econometrics, Volume IV*, ed. by R. F. Engle, and D. L. McFadden, pp. 2111–2245. Elsevier Science B.V., New York.

NEWEY, W. K., AND T. M. STOKER (1993): "Efficiency of Weighted Average Derivative Estimators and Index Models," *Econometrica*, 61, 1199–1223.

POWELL, J. L. (1994): "Estimation of Semiparametric Models," in *Handbook of Econometrics, Volume IV*, ed. by R. Engle, and D. McFadden, pp. 2443–2521. Elsevier Science B.V.

POWELL, J. L., J. H. STOCK, AND T. M. STOKER (1989): "Semiparametric Estimation of Index Coefficients," *Econometrica*, 57, 1403–1430.

ROBINS, J., L. LI, E. TCHETGEN, AND A. VAN DER VAART (2008): "Higher Order Influence Functions and Minimax Estimation of Nonlinear Functionals," in *Probability and Statistics: Essays in Honor of David A. Freedman*, ed. by D. Nolan, and T. Speed, pp. 335–421. Institute of Mathematical Statistics, Beachwood, OH.

ROBINSON, P. M. (1988): "Root-$N$-Consistent Semiparametric Regression," *Econometrica*, 56, 931–954.

SCHUCANY, W. R., AND J. P. SOMMERS (1977): "Improvement of Kernel Type Density Estimators," *Journal of the American Statistical Association*, 72, 420–423.

STOKER, T. M. (1986): "Consistent Estimation of Scaled Coefficients," *Econometrica*, 54, 1461–1481.

——— (1989): "Tests of Additive Derivative Constraints," *Review of Economic Studies*, 56(4), 535–552.

# Supplemental Appendix to "Generalized Jackknife Estimators of Weighted Average Derivatives"

(Intended for web-publication.)

This supplement provides brief proofs for the Lemmas stated in the main text. Further details on these proofs, and the proofs of Theorems 1 – 3, are available upon request from the authors.

## 1. Appendix A: Proofs

**1.1. Proof of Lemma A-1.** Expanding $\hat{s}_n(x;h_n)$ around $s(x)$, we have

$$\hat{s}_n(x;h_n) = \hat{s}^{**}(x;h_n) - \frac{w(x)}{f(x)^2 \hat{f}_n(x;h_n)} \delta_n(x;h_n)^2 \left[\dot{\delta}_n(x;h_n) + \ell(x)\,\delta_n(x;h_n)\right],$$

where

$$\delta_n(x;h_n) = \hat{f}_n(x;h_n) - f(x), \qquad \dot{\delta}_n(x;h_n) = \frac{\partial}{\partial x}\hat{f}_n(x;h_n) - \frac{\partial}{\partial x} f(x).$$

Because $\Delta_{0,n}(h_n) = o_p(1)$ it follows from a simple bounding argument that for any $\varepsilon > 0$ there exists a constant $C_\varepsilon$ such that, for $n$ sufficiently large,

$$\sup_{x\in\mathcal{W}} \|\hat{s}_n(x;h_n) - \hat{s}^{**}(x;h_n)\| \leq C_\varepsilon \Delta_{0,n}(h_n)^2 \Delta_{1,n}(h_n) \tag{1}$$

with probability no less than $1-\varepsilon$. If (1) holds and $\Delta_{0,n}(h_n)^2 \Delta_{1,n}(h_n) = o_p\left(n^{-1/2}\right)$, then

$$\left\|\hat{\theta}_n(h_n) - \hat{\theta}_n^{**}(h_n)\right\| \leq C_\varepsilon \left(n^{-1}\sum_{i=1}^{n}|y_i|\right) \Delta_{0,n}(h_n)^2 \Delta_{1,n}(h_n) = o_p\left(n^{-1/2}\right),$$

where the equality uses $\mathbb{E}\left(|y|\right)<\infty$. This establishes (7) in case (i).

Next, suppose $\Delta_{0,n}\left(h_n\right)\Delta_{1,n}\left(h_n\right)=o_p\left(n^{-1/2}\right)$. Then, by the triangle inequality and the result for case (i),

$$\begin{aligned}\left\|\hat{\theta}_n\left(h_n\right)-\hat{\theta}_n^*\left(h_n\right)\right\| &\leq \left\|\hat{\theta}_n\left(h_n\right)-\hat{\theta}_n^{**}\left(h_n\right)\right\|+\left\|\hat{\theta}_n^{**}\left(h_n\right)-\hat{\theta}_n^*\left(h_n\right)\right\| \\ &= \left\|\hat{\theta}_n^{**}\left(h_n\right)-\hat{\theta}_n^*\left(h_n\right)\right\|+o_p\left(n^{-1/2}\right),\end{aligned}$$

so validity of (7) in case (ii) follows from the fact that

$$\left\|\hat{\theta}_n^{**}\left(h_n\right)-\hat{\theta}_n^*\left(h_n\right)\right\|\leq C\left(n^{-1}\sum_{i=1}^n|y_i|\right)\Delta_{0,n}\left(h_n\right)\Delta_{1,n}\left(h_n\right)=o_p\left(n^{-1/2}\right),$$

where the inequality uses the elementary bound

$$\sup_{x\in\mathcal{W}}\left\|\hat{s}_n^{**}\left(x;h_n\right)-\hat{s}^*\left(x;h_n\right)\right\|\leq C\Delta_{0,n}\left(h_n\right)\Delta_{1,n}\left(h_n\right),$$

in which

$$C=\sup_{x\in\mathcal{W}}\left[\frac{|w\left(x\right)|}{f\left(x\right)^2}\left(1+|\ell\left(x\right)|\right)\right]<\infty. \qquad \blacksquare$$

**1.2. Proof of Lemma A-4.** Part (a) is a standard result on the bias of kernel estimators (e.g., Newey (1994b, Lemma B.2)), while parts (b) and (c) follow from change of variables and simple bounding arguments. For instance,

$$\begin{aligned}&\mathbb{E}\left[F\left(z_1\right)^2\mathcal{K}\left(x_1-x_2;h_n\right)^2\left\|\dot{\mathcal{K}}\left(x_1-x_3;h_n\right)\right\|^2\right] \\ &=\mathbb{E}\left[\int_{\mathbb{R}^d}\int_{\mathbb{R}^d}F\left(z_1\right)^2\mathcal{K}\left(x_1-s;h_n\right)^2\left\|\dot{\mathcal{K}}\left(x_1-t;h_n\right)\right\|^2f\left(s\right)f\left(t\right)dtds\right] \\ &=h_n^{-2(d+1)}\mathbb{E}\left[\int_{\mathbb{R}^d}\int_{\mathbb{R}^d}F\left(z_1\right)^2K\left(u\right)^2\left\|\dot{K}\left(v\right)\right\|^2f\left(x_1-uh_n\right)f\left(x_1-vh_n\right)dvdu\right] \\ &\leq h_n^{-2(d+1)}C_f^2\mathbb{E}\left[F\left(z\right)^2\right]\int_{\mathbb{R}^d}K\left(u\right)^2du\int_{\mathbb{R}^d}\left\|\dot{K}\left(v\right)\right\|^2dv=O\left(h_n^{-2(d+1)}\right),\end{aligned}$$

where $C_f = \sup_{x \in \mathbb{R}^d} f(x)$. ■

**1.3. Proof of Lemma A-2.** Defining

$$V_i^{\mu} = V_i - \mathbb{E}(V_i) = y_i s(x_i) - \theta, \qquad V_i = y_i s(x_i),$$

$$V_{ij}^{\mu}(h) = V_{ij}(h) - \mathbb{E}[V_{ij}(h)], \qquad V_{ij}(h) = -y_i \frac{w(x_i)}{f(x_i)} \left[\dot{\mathcal{K}}(x_i - x_j; h) + \ell(x_i) \mathcal{K}(x_i - x_j; h)\right],$$

we have the decomposition

$$\begin{aligned} \hat{\theta}_n^*(h) &= n^{-1} \sum_{i=1}^n V_i + n^{-2} \sum_{i=1}^n \sum_{j=1}^n V_{ij}(h) \\ &= \mathbb{E}\left[\hat{\theta}_n^*(h)\right] + n^{-1} \sum_{i=1}^n V_i^{\mu} + n^{-2} \sum_{i=1}^{n-1} \sum_{j=i+1}^n \left[V_{ij}^{\mu}(h) + V_{ji}^{\mu}(h)\right] + n^{-2} \sum_{i=1}^n V_{ii}^{\mu}(h), \end{aligned}$$

where $n^{-2} \sum_{i=1}^n V_{ii}^{\mu}(h_n) = o_p\left(n^{-1/2}\right)$ because

$$\mathbb{V}\left[n^{-2} \sum_{i=1}^n V_{ii}^{\mu}(h_n)\right] = n^{-3} \mathbb{V}[V_{11}(h_n)] = n^{-1}\left(n h_n^d\right)^{-2} K(0)^2 \, \mathbb{V}\left[y \frac{w(x)}{f(x)} \ell(x)\right] = o\left(n^{-1}\right).$$

The proof for $\hat{\theta}_n^A = \hat{\theta}_n^*(h_n)$ will be completed by showing that

$$n^{-1} \sum_{i=1}^{n-1} \sum_{j=i+1}^n \left[V_{ij}^{\mu}(h_n) + V_{ji}^{\mu}(h_n)\right] = n^{-1} \sum_{i=1}^n \varphi(z_i) + o_p\left(n^{-1/2}\right),$$

where

$$\varphi(z) = \psi(z) - [y s(x) - \theta] = \frac{\partial}{\partial x}[w(x) g(x)] - w(x) g(x) \ell(x).$$

To do so, let $\mathbb{E}_i$ denote conditional expectation given $z_i$ and for any positive sequence $\{r_n\}$, let $X_n = O_2(r_n)$ and $X_n = o_2(r_n)$ be shorthand for $\overline{\lim}_{n\to\infty} \mathbb{E}\left(X_n^2\right)/r_n^2 < \infty$ and $\lim_{n\to\infty} \mathbb{E}\left(X_n^2\right)/r_n^2 = 0$, respectively.

Because $h_n \to 0$ and $nh_n^{d+2} \to \infty$,

$$V_{ij}(h_n) = -y_i \frac{w(x_i)}{f(x_i)} \left[\dot{\mathcal{K}}(x_i - x_j; h_n) + \ell(x_i)\mathcal{K}(x_i - x_j; h_n)\right] = O_2\left(h_n^{-(d+2)/2}\right) = o_2\left(\sqrt{n}\right),$$

where the second equality uses Lemma A-4 (b). Therefore, by the projection theorem for variable $U$-statistics (e.g., Powell, Stock, and Stoker (1989, Lemma 3.1)),

$$n^{-2}\sum_{i=1}^{n-1}\sum_{j=i+1}^{n}\left[V_{ij}^{\mu}(h_n) + V_{ji}^{\mu}(h_n)\right] = n^{-1}\sum_{i=1}^{n}\mathbb{E}_i\left[V_{ij}^{\mu}(h_n) + V_{ji}^{\mu}(h_n)\right] + o_p\left(n^{-1/2}\right),$$

where, by Lemma A-4 (a),

$$\mathbb{E}_i V_{ij}(h_n) = -y_i \frac{w(x_i)}{f(x_i)}\left[\dot{b}(x_i; h_n) + \ell(x_i) b(x_i; h_n)\right] = O_2\left(h_n^P\right) = o_2(1)$$

and, using integration by parts and change of variables,

$$\begin{aligned}
\mathbb{E}_i V_{ji}(h_n) &= -\int_{\mathbb{R}^d} g(r) w(r)\left[\dot{\mathcal{K}}(r - x_i; h_n) + \ell(r)\mathcal{K}(r - x_i; h_n)\right] dr \\
&= \int_{\mathbb{R}^d}\left(\frac{\partial}{\partial r}\left[g(r) w(r)\right]\right)\mathcal{K}(r - x_i; h_n)\, dr - \int_{\mathbb{R}^d} g(r) w(r) \ell(r)\mathcal{K}(r - x_i; h_n)\, dr \\
&= \int_{\mathbb{R}^d}\frac{\partial}{\partial x}\left[g(x_i + th_n) w(x_i + th_n)\right] K(t)\, dt \\
&\qquad - \int_{\mathbb{R}^d} g(x_i + th_n) w(x_i + th_n) \ell(x_i + th_n) K(t)\, dt \\
&= \varphi(z_i) + o_2(1).
\end{aligned}$$

Using these results and the fact that $\mathbb{E}[\varphi(z)] = 0$ it is easy to show that

$$n^{-1}\sum_{i=1}^{n}\mathbb{E}_i\left[V_{ij}^{\mu}(h_n) + V_{ji}^{\mu}(h_n)\right] = n^{-1}\sum_{i=1}^{n}\varphi(z_i) + o_p\left(n^{-1/2}\right),$$

completing the proof for $\hat{\theta}_n^A = \hat{\theta}_n^*(h_n)$.

Finally, having established the result for $\hat{\theta}_n^A = \hat{\theta}_n^* (h_n)$ the result for $\hat{\theta}_n^A = \hat{\theta}_n^{**} (h_n)$ will follow if it can be shown that $\mathbb{V}\left[\hat{\theta}_n^{**} (h_n) - \hat{\theta}_n^* (h_n)\right] = o(n^{-1})$. To do so, we employ the decomposition

$$\begin{aligned}
\hat{\theta}_n^{**} (h) - \hat{\theta}_n^* (h) &= n^{-3} \sum_{i=1}^{n} \sum_{j_1=1}^{n} \sum_{j_2=1}^{n} V_{ij_1j_2} (h) \\
&= \mathbb{E}\left[\hat{\theta}_n^{**} (h) - \hat{\theta}_n^* (h)\right] + n^{-3} \sum_{i=1}^{n} \sum_{j_1=1}^{n} \sum_{j_2=1}^{n} V_{ij_1j_2}^{\mu} (h),
\end{aligned}$$

where $V_{ij_1j_2}^{\mu} (h) = V_{ij_1j_2} (h) - \mathbb{E}[V_{ij_1j_2} (h)]$ and

$$V_{ij_1j_2} (h) = y_i \frac{w(x_i)}{f(x_i)^2} [\mathcal{K}(x_i - x_{j_1}; h) - f(x_i)] \left[\dot{\mathcal{K}}(x_i - x_{j_2}; h) + \ell(x_i) \mathcal{K}(x_i - x_{j_2}; h)\right].$$

The Hoeffding decomposition yields

$$\mathbb{V}\left[\sum_{i=1}^{n} \sum_{j_1=1}^{n} \sum_{j_2=1}^{n} V_{ij_1j_2}^{\mu} (h)\right] = \sum_{p=1}^{3} \binom{n}{p} \mathbb{V}\left[\sum_{i=1}^{n} \sum_{j_1=1}^{n} \sum_{j_2=1}^{n} H_{ij_1j_2} (p; h)\right],$$

where

$$H_{ij_1j_2} (1; h) = \mathbb{E}_1 [V_{ij_1j_2} (h)] - \mathbb{E}[V_{ij_1j_2} (h)],$$

$$H_{ij_1j_2} (2; h) = \mathbb{E}_{1,2} [V_{ij_1j_2} (h)] - \mathbb{E}_1 [V_{ij_1j_2} (h)] - \mathbb{E}_2 [V_{ij_1j_2} (h)] + \mathbb{E}[V_{ij_1j_2} (h)],$$

$$\begin{aligned}
H_{ij_1j_2} (3; h) &= \mathbb{E}_{1,2,3} [V_{ij_1j_2} (h)] - \mathbb{E}_{1,2} [V_{ij_1j_2} (h)] - \mathbb{E}_{1,3} [V_{ij_1j_2} (h)] - \mathbb{E}_{2,3} [V_{ij_1j_2} (h)] \\
&\quad + \mathbb{E}_1 [V_{ij_1j_2} (h)] + \mathbb{E}_2 [V_{ij_1j_2} (h)] + \mathbb{E}_3 [V_{ij_1j_2} (h)] - \mathbb{E}[V_{ij_1j_2} (h)],
\end{aligned}$$

with $\mathbb{E}_{1,2,3} [V_{ij_1j_2} (h)] = \mathbb{E}[V_{ij_1j_2} (h) | z_1, z_2, z_3]$, $\mathbb{E}_{2,3} [V_{ij_1j_2} (h)] = \mathbb{E}[V_{ij_1j_2} (h) | z_2, z_3]$, and so on. It therefore suffices to show that

$$\mathbb{V}\left[\sum_{i=1}^{n}\sum_{j_1=1}^{n}\sum_{j_2=1}^{n} H_{ij_1j_2}\left(p;h_n\right)\right] = o\left(n^{5-p}\right), \qquad p \in \{1,2,3\}. \tag{2}$$

The proof of (2) for $p = 1$ will be based on the relation

$$\mathbb{V}\left[\sum_{i=1}^{n}\sum_{j_1=1}^{n}\sum_{j_2=1}^{n} H_{ij_1j_2}\left(1;h\right)\right] = \mathbb{V}\left[\mathcal{H}_n\left(1;h\right)\right],$$

where

$$\begin{aligned}
\mathcal{H}_n\left(1;h\right) &= H_{111}\left(1;h\right) + \left(n-1\right)\left[H_{112}\left(1;h\right) + H_{121}\left(1;h\right) + H_{211}\left(1;h\right)\right] \\
&\quad + \left(n-1\right)\left[H_{122}\left(1;h\right) + H_{212}\left(1;h\right) + H_{221}\left(1;h\right)\right] \\
&\quad + \left(n-1\right)\left(n-2\right)\left[H_{123}\left(1;h\right) + H_{213}\left(1;h\right) + H_{231}\left(1;h\right)\right].
\end{aligned}$$

Because $\mathbb{V}\left[H_{ijk}\left(1;h\right)\right] \le \mathbb{V}\left(\mathbb{E}_1\left[V_{ijk}\left(h\right)\right]\right)$ for each $(i,j,k)$, the result $\mathbb{V}\left[\mathcal{H}_n\left(1;h_n\right)\right] = o\left(n^4\right)$ can be established by means of polynomial (in $n$) bound on the second moment of each $\mathbb{E}_1\left[V_{ijk}\left(h_n\right)\right]$.

First,

$$\begin{aligned}
\mathbb{E}_1\left[V_{111}\left(h_n\right)\right] &= y_1\frac{w\left(x_1\right)}{f\left(x_1\right)^2}\left[\mathcal{K}\left(0;h_n\right) - f\left(x_1\right)\right]\ell\left(x_1\right)\mathcal{K}\left(0;h_n\right) \\
&= h_n^{-2d}K\left(0\right)^2 y_1\frac{w\left(x_1\right)}{f\left(x_1\right)^2}\ell\left(x_1\right) - h_n^{-d}K\left(0\right)y_1\frac{w\left(x_1\right)}{f\left(x_1\right)^2}f\left(x_1\right)\ell\left(x_1\right) \\
&= O_2\left(h_n^{-2d}\right) = o_2\left(n^4\right).
\end{aligned}$$

Next, using Lemma A-4 (a), change of variables, and simple bounding arguments,

$$\begin{aligned}
\mathbb{E}_1\left[V_{112}\left(h_n\right)\right] &= y_1\frac{w\left(x_1\right)}{f\left(x_1\right)^2}\mathcal{K}\left(0;h_n\right)\int_{\mathbb{R}^d}\left[\dot{\mathcal{K}}\left(x_1-s;h_n\right)+\ell\left(x_1\right)\mathcal{K}\left(x_1-s;h_n\right)\right]f\left(s\right)ds \\
&\quad -y_1\frac{w\left(x_1\right)}{f\left(x_1\right)^2}f\left(x_1\right)\int_{\mathbb{R}^d}\left[\dot{\mathcal{K}}\left(x_1-s;h_n\right)+\ell\left(x_1\right)\mathcal{K}\left(x_1-s;h_n\right)\right]f\left(s\right)ds \\
&= y_1\frac{w\left(x_1\right)}{f\left(x_1\right)^2}\left[h_n^{-d}K\left(0\right)-f\left(x_1\right)\right]\left[\dot{b}\left(x_1;h_n\right)+\ell\left(x_1\right)b\left(x_1;h_n\right)\right] \\
&= O_2\left(h_n^{P-d}\right)=o_2\left(n^2\right).
\end{aligned}$$

Similarly, it can be shown that

$$\mathbb{E}_1\left[V_{121}\left(h_n\right)\right]=O_2\left(h_n^{P-d}\right)=o_2\left(n^2\right),\qquad \mathbb{E}_1\left[V_{211}\left(h_n\right)\right]=O_2\left(h_n^{-(d+1)}\right)=o_2\left(n^2\right),$$

$$\mathbb{E}_1\left[V_{122}\left(h_n\right)\right]=O_2\left(h_n^{-(d+1)}\right)=o_2\left(n^2\right),\qquad \mathbb{E}_1\left[V_{212}\left(h_n\right)\right]=O_2\left(h_n^{-d}\right)=o_2\left(n^2\right),$$

$$\mathbb{E}_1\left[V_{221}\left(h_n\right)\right]=O_2\left(h_n^{-(d+1)}\right)=o_2\left(n^2\right),\qquad \mathbb{E}_1\left[V_{123}\left(h_n\right)\right]=O_2\left(h_n^{2P}\right)=o_2\left(1\right),$$

$$\mathbb{E}_1\left[V_{213}\left(h_n\right)\right]=O_2\left(h_n^{P}\right)=o_2\left(1\right),\qquad \mathbb{E}_1\left[V_{231}\left(h_n\right)\right]=O_2\left(h_n^{P-1}\right)=o_2\left(1\right),$$

from which (2) follows for $p=1$.

The proofs of (2) are very similar for $p=2$ and $p=3$, so we give only the proof for $p=3$, which is based on the relation

$$\mathbb{V}\left[\sum_{i=1}^{n}\sum_{j_1=1}^{n}\sum_{j_2=1}^{n}H_{ij_1j_2}\left(3;h\right)\right]=\mathbb{V}\left[\mathcal{H}\left(3;h\right)\right],$$

where

$$\mathcal{H}\left(3;h\right)=H_{123}\left(3;h\right)+H_{132}\left(3;h\right)+H_{213}\left(3;h\right)+H_{231}\left(3;h\right)+H_{312}\left(3;h\right)+H_{321}\left(3;h\right)$$

and $\mathbb{V}\left[H_{ijk}\left(3;h\right)\right]\leq\mathbb{V}\left(\mathbb{E}_{1,2,3}\left[V_{ijk}\left(h\right)\right]\right)$ for each $(i,j,k)$.

Using Lemma A-4 (c),

$$\begin{aligned}\mathbb{E}_{1,2,3}\left[V_{123}\left(h_n\right)\right] &= V_{123}\left(h_n\right) \\ &= y_1\frac{w\left(x_1\right)}{f\left(x_1\right)^2}\left[\mathcal{K}\left(x_1-x_2;h_n\right)-f\left(x_1\right)\right]\left[\dot{\mathcal{K}}\left(x_1-x_3;h_n\right)+\ell\left(x_1\right)\mathcal{K}\left(x_1-x_3;h_n\right)\right] \\ &= O_2\left(h_n^{-(d+1)}\right)=o\left(n^2\right).\end{aligned}$$

The result $\mathbb{V}\left[\mathcal{H}_n\left(3;h_n\right)\right]=o\left(n^2\right)$ follows from this and the fact that $V_{123}\left(3;h\right)$, $V_{132}\left(3;h\right)$, $V_{213}\left(3;h\right)$, $V_{231}\left(3;h\right)$, $H_{312}\left(3;h\right)$, and $V_{321}\left(3;h\right)$ are identically distributed. ■

**1.4. Proof of Lemma A-3.** Using the same notation as in the proof of Lemma A-2, we have

$$\begin{aligned}\mathbb{E}\left[\hat{\theta}_n^*\left(h\right)\right] &= n^{-1}\sum_{i=1}^{n}\mathbb{E}\left(V_i\right)+n^{-2}\sum_{i=1}^{n}\sum_{j=1}^{n}\mathbb{E}\left[V_{ij}\left(h\right)\right] \\ &= \mathbb{E}\left(V_1\right)+n^{-1}\mathbb{E}\left[V_{11}\left(h\right)\right]+\left(1-n^{-1}\right)\mathbb{E}\left[V_{12}\left(h\right)\right],\end{aligned}$$

where $\mathbb{E}\left(V_1\right)=\theta$, $\mathbb{E}\left[V_{11}\left(h\right)\right]=h^{-d}\mathcal{B}_0^*$, and, using Lemma A-4 (a),

$$\mathbb{E}\left[V_{12}\left(h_n\right)\right]=-\int_{\mathbb{R}^d}g\left(r\right)w\left(r\right)\left[\dot{b}\left(r;h_n\right)+\ell\left(r\right)b\left(r;h_n\right)\right]dr=O\left(h_n^P\right).$$

Next,

$$\begin{aligned}\mathbb{E}\left[\hat{\theta}_n^{**}\left(h_n\right)-\hat{\theta}_n^*\left(h_n\right)\right] &= n^{-3}\sum_{i=1}^{n}\sum_{j_1=1}^{n}\sum_{j_2=1}^{n}\mathbb{E}\left[V_{ij_1j_2}\left(h_n\right)\right] \\ &= n^{-2}\mathbb{E}\left[V_{111}\left(h_n\right)\right]+n^{-1}\left(1-n^{-1}\right)\left(\mathbb{E}\left[V_{112}\left(h_n\right)\right]+\mathbb{E}\left[V_{121}\left(h_n\right)\right]\right) \\ &\qquad +n^{-1}\left(1-n^{-1}\right)\mathbb{E}\left[V_{122}\left(h_n\right)\right]+\left(1-n^{-1}\right)\left(1-2n^{-1}\right)\mathbb{E}\left[V_{123}\left(h_n\right)\right] \\ &= n^{-1}\left(1-n^{-1}\right)\mathbb{E}\left[V_{122}\left(h_n\right)\right]+O\left(n^{-1}h_n^{P-d}+n^{-2}h_n^{-2d}+h_n^{2P}\right)\end{aligned}$$

because it follows from Lemma A-4 (a) and simple bounding arguments that

$$\mathbb{E}\left[V_{111}\left(h_n\right)\right]=O\left(h_n^{-2d}\right), \qquad \mathbb{E}\left[V_{112}\left(h_n\right)\right]=O\left(h_n^{P-d}\right),$$

and

$$\mathbb{E}\left[V_{121}\left(h_n\right)\right]=O\left(h_n^{P-d}\right), \qquad \mathbb{E}\left[V_{123}\left(h_n\right)\right]=O\left(h_n^{2P}\right).$$

Moreover,

$$\begin{aligned}
\mathbb{E}\left[V_{122}\left(h_n\right)\right] &= \int_{\mathbb{R}^d}\int_{\mathbb{R}^d} g\left(r\right)\frac{w\left(r\right)}{f\left(r\right)^2}\mathcal{K}\left(r-s;h_n\right)\dot{\mathcal{K}}\left(r-s;h_n\right)f\left(r\right)f\left(s\right)dsdr \\
&\quad +\int_{\mathbb{R}^d}\int_{\mathbb{R}^d} g\left(r\right)\frac{w\left(r\right)}{f\left(r\right)^2}\ell\left(r\right)\mathcal{K}\left(r-s;h_n\right)^2 f\left(r\right)f\left(s\right)dsdr \\
&\quad -\int_{\mathbb{R}^d}\int_{\mathbb{R}^d} g\left(r\right)\frac{w\left(r\right)}{f\left(r\right)}\left[\dot{\mathcal{K}}\left(r-s;h_n\right)+\ell\left(r\right)\mathcal{K}\left(r-s;h_n\right)\right]f\left(r\right)f\left(s\right)dsdr \\
&= h_n^{-(d+1)}\int_{\mathbb{R}^d} g\left(r\right)\frac{w\left(r\right)}{f\left(r\right)}\left[\int_{\mathbb{R}^d} K\left(t\right)\dot{K}\left(t\right)f\left(r-th_n\right)dt\right]dr \\
&\quad +h_n^{-d}\int_{\mathbb{R}^d} g\left(r\right)\frac{w\left(r\right)}{f\left(r\right)}\ell\left(r\right)\left[\int_{\mathbb{R}^d} K\left(t\right)^2 f\left(r-th_n\right)dt\right]dr+O\left(h_n^{P}\right),
\end{aligned}$$

where Taylor's theorem can be used to show that

$$\int_{\mathbb{R}^d} g\left(r\right)\frac{w\left(r\right)}{f\left(r\right)}\left[\int_{\mathbb{R}^d} K\left(t\right)\dot{K}\left(t\right)f\left(r-th_n\right)dt\right]dr=\sum_{j=0}^{P}\dot{B}_j h_n^j+O\left(h_n^{P+1}\right),$$

$$\int_{\mathbb{R}^d} g\left(r\right)\frac{w\left(r\right)}{f\left(r\right)}\ell\left(r\right)\left[\int_{\mathbb{R}^d} K\left(t\right)^2 f\left(r-th_n\right)dt\right]dr=\sum_{j=0}^{P}B_j h_n^j+O\left(h_n^{P+1}\right),$$

$$\dot{B}_j=\frac{(-1)^{j+1}}{j!}\sum_{l\in\mathbb{Z}_+^d(j)}\dot{B}_K\left(l\right)\dot{B}_z\left(l\right), \qquad B_j=\frac{(-1)^{j}}{j!}\sum_{l\in\mathbb{Z}_+^d(j)}B_K\left(l\right)B_z\left(l\right).$$

Because $K$ is even, $B_K\left(l\right)=0$ whenever $l\in\mathbb{Z}_+^d\left(j\right)$ for $j$ odd and $\dot{B}_K\left(l\right)=0$ whenever

$l \in \mathbb{Z}_{+}^{d}(j)$ for $j$ even. As a consequence,

$$\begin{aligned} \mathbb{E}\left[V_{122}\left(h_n\right)\right] &= h_n^{-(d+1)} \sum_{j=0}^{P} \dot{B}_j h_n^j + h_n^{-d} \sum_{j=0}^{P} B_j h_n^j + O\left(h_n^{P-d} + h_n^P\right) \\ &= \sum_{j=0}^{\lfloor (P-1)/2 \rfloor} h_n^{2j-d} \mathcal{B}_j^{**} + O\left(h_n^{P-d} + h_n^P\right), \end{aligned}$$

where $\mathcal{B}_j^{**} = B_{2j} + \dot{B}_{2j+1}$. ■

## 2. Appendix B: Uniform Convergence Rates for Kernel Estimators

**2.1. Proof of Lemma B-1.** Similarly to the proof of Newey (1994b, Lemma B.1), the proof consists of three steps, of which the first step is a truncation step, the second step is a discretization step, and the final step uses Bernstein's inequality to bound certain tail probabilities. To accommodate kernels with unbounded support, the second step borrows ideas from Hansen (2008). In the third step, we use Bernstein's inequality in two distinct ways (and employ a subsequence argument) in order to accommodate bandwidths that do not satisfy $n^{1-2/s} h_n^d / \log n \to \infty$.

Given a sequence $\tau_n$, let

$$\tilde{\Psi}_n(x) = \frac{1}{n h_n^d} \sum_{j=1}^{n} Y_{jn} \kappa\left(\frac{x - X_j}{h_n}\right), \qquad Y_{jn} = Y_j \mathbf{1}\left(|Y_j| \le \tau_n\right),$$

denote a version of $\hat{\Psi}_n$ obtained by replacing $Y_j$ with the truncated variable $Y_{jn}$. The processes $\hat{\Psi}_n(\cdot)$ and $\tilde{\Psi}_n(\cdot)$ coincide with a probability that can be made arbitrarily close to one (uniformly in $n$) by setting $\tau_n = C_\tau n^{1/s}$ for some large $C_\tau$ because

$$\begin{aligned} \Pr\left[\hat{\Psi}_n(\cdot) \neq \tilde{\Psi}_n(\cdot)\right] &\le \Pr\left[Y_j \neq Y_{jn} \text{ for some } j\right] = \Pr\left[|Y_j| > \tau_n \text{ for some } j\right] \\ &\le n \Pr\left[|Y| > \tau_n\right] \le n \tau_n^{-s} C_Y(s), \end{aligned}$$

where $C_Y(r) = \mathbb{E}(|Y|^r) + \sup_{x\in\mathbb{R}^d} \mathbb{E}(|Y|^r | X = x) f_X(x)$ and the last inequality uses Markov's inequality. Also,

$$\begin{aligned}
\left|\mathbb{E}\left[\hat{\Psi}_n(x) - \tilde{\Psi}_n(x)\right]\right| &= \left|\mathbb{E}\left[Y 1(|Y| > \tau_n) h_n^{-d} \kappa\left(\frac{x-X}{h_n}\right)\right]\right| \\
&= \left|\int_{\mathbb{R}^d} \mathbb{E}[Y 1(|Y| > \tau_n) | X = r] h_n^{-d} \kappa\left(\frac{x-r}{h_n}\right) f_X(r) dr\right| \\
&\leq \tau_n^{-(s-1)} \int_{\mathbb{R}^d} \mathbb{E}[|Y|^s \mathbf{1}(|Y| > \tau_n) | X = r] h_n^{-d} \left|\kappa\left(\frac{r-x}{h_n}\right)\right| f_X(r) dr \\
&\leq \tau_n^{-(s-1)} C_Y(s) C_\kappa, \qquad C_\kappa = \sup_{u\in\mathbb{R}^d} |\kappa(u)| + \int_{\mathbb{R}^d} |\kappa(u)| du,
\end{aligned}$$

so if $\tau_n = C_\tau n^{1/s}$, then

$$\sup_{x\in\mathbb{R}^d} \left|\mathbb{E}\left[\hat{\Psi}_n(x)\right] - \mathbb{E}\left[\tilde{\Psi}_n(x)\right]\right| = O\left(n^{1/s-1}\right) = o(\rho_n).$$

To complete the proof, it therefore suffices to show that

$$\sup_{x\in\mathcal{X}_n} \left|\tilde{\Psi}_n(x) - \mathbb{E}\left[\tilde{\Psi}_n(x)\right]\right| = O_p(\rho_n), \qquad \tau_n = C_\tau n^{1/s}.$$

**Remark.** Hansen (2008, p. 740) employs $\tau_n = \rho_n^{-1/(s-1)} = o\left(n^{1/s}\right)$ in his truncation argument and shows that with this choice of $\tau_n$

$$\left|\left(\tilde{\Psi}_n(x) - \mathbb{E}\left[\tilde{\Psi}_n(x)\right]\right) - \left(\hat{\Psi}_n(x) - \mathbb{E}\left[\hat{\Psi}_n(x)\right]\right)\right| = O_p(\rho_n)$$

for every $x$. It is unclear whether this pointwise rate of convergence holds uniformly in $x \in \mathcal{X}_n$, so we err on the side of caution and set $\tau_n = C_\tau n^{1/s}$.

Continuing with the proof of Lemma B-1, we discretize by employing a sequence $G_n$ (depending on $C_{X,n}$ and $h_n$) and associated points $\left\{x_{g,n}^* : j = 1, \ldots, G_n\right\}$ such that

$$\overline{\lim}_{n\to\infty} \log\left(G_n\right) / \log n < \infty \tag{3}$$

and

$$\mathcal{X}_n \subseteq \cup_{g=1}^{G_n} \mathcal{X}_{g,n}, \qquad \mathcal{X}_{g,n} = \left\{x : \left\|x - x_{g,n}^*\right\| \leq \min\left(1, \delta_\kappa\right) h_n\right\}. \tag{4}$$

It follows from (3) that $G_n = o\left(n^R\right)$ for some $R < \infty$, while (4) implies that, for any $M$,

$$\Pr\left[\sup_{x\in\mathcal{X}_n}\left|\tilde{\Psi}_n\left(x\right) - \mathbb{E}\tilde{\Psi}_n\left(x\right)\right| > M\rho_n\right] \leq G_n \max_{1\leq g\leq G_n} \Pr\left[\sup_{x\in\mathcal{X}_{g,n}}\left|\tilde{\Psi}_n\left(x\right) - \mathbb{E}\tilde{\Psi}_n\left(x\right)\right| > M\rho_n\right].$$

To complete the proof it therefore suffices to show that for any $R < \infty$, there is an $M$ such that

$$\max_{1\leq g\leq G_n} \Pr\left[\sup_{x\in\mathcal{X}_{g,n}}\left|\tilde{\Psi}_n\left(x\right) - \mathbb{E}\tilde{\Psi}_n\left(x\right)\right| > M\rho_n\right] = O\left(n^{-R}\right). \tag{5}$$

If $x \in \mathcal{X}_{g,n}$ and $\rho_n \leq \delta_\kappa$, then

$$\left|\kappa\left(\frac{x - X_j}{h_n}\right) - \kappa\left(\frac{x_{g,n}^* - X_j}{h_n}\right)\right| \leq \rho_n\kappa^*\left(\frac{x_{g,n}^* - X_j}{h_n}\right) \qquad \left(j = 1, \ldots, n\right),$$

so

$$\left|\tilde{\Psi}_n\left(x\right) - \tilde{\Psi}_n\left(x_{g,n}^*\right)\right| \leq \rho_n\tilde{\Psi}_n^*\left(x_{g,n}^*\right), \qquad \tilde{\Psi}_n^*\left(x\right) = \frac{1}{nh_n^d}\sum_{j=1}^n Y_{jn}\kappa^*\left(\frac{x - X_j}{h_n}\right).$$

Therefore, if $\rho_n \leq \delta_\kappa$, then

$$\begin{aligned}\sup_{x\in\mathcal{X}_{g,n}}\left|\tilde{\Psi}_n\left(x\right) - \mathbb{E}\left[\tilde{\Psi}_n\left(x\right)\right]\right| \leq\ & \left|\tilde{\Psi}_n\left(x_{g,n}^*\right) - \mathbb{E}\left[\tilde{\Psi}_n\left(x_{g,n}^*\right)\right]\right| \\ & +\rho_n\left|\tilde{\Psi}_n^*\left(x_{g,n}^*\right) - \mathbb{E}\left[\tilde{\Psi}_n^*\left(x_{g,n}^*\right)\right]\right| \\ & +2\rho_n\mathbb{E}\left(\left|\tilde{\Psi}_n^*\left(x_{g,n}^*\right)\right|\right),\end{aligned}$$

where

$$\begin{aligned}\mathbb{E}\left(\left|\tilde{\Psi}_n^*\left(x_{g,n}^*\right)\right|\right) &\leq \int_{\mathbb{R}^d} \mathbb{E}\left[|Y|\,|X=x\right] h_n^{-d} \kappa^*\left(\frac{x_{g,n}^* - x}{h_n}\right) f_X(x)\, dx \\ &\leq C_Y(1)\, C_{\kappa^*}, \qquad C_{\kappa^*} = \sup_{u\in\mathbb{R}^d} \kappa^*(u) + \int_{\mathbb{R}^d} \kappa^*(u)\, du.\end{aligned}$$

As a consequence, if $\rho_n \leq \min(1, \delta_\kappa)$ and $M \geq 4C_Y(1)\, C_{\kappa^*}$, then

$$\begin{aligned}\Pr\left[\sup_{x\in\mathcal{X}_{g,n}} \left|\tilde{\Psi}_n(x) - \mathbb{E}\left[\tilde{\Psi}_n(x)\right]\right| > M\rho_n\right] &\leq \Pr\left[\left|\tilde{\Psi}_n\left(x_{g,n}^*\right) - \mathbb{E}\left[\tilde{\Psi}_n\left(x_{g,n}^*\right)\right]\right| > M\rho_n/4\right] \\ &\quad + \Pr\left[\left|\tilde{\Psi}_n^*\left(x_{g,n}^*\right) - \mathbb{E}\left[\tilde{\Psi}_n^*\left(x_{g,n}^*\right)\right]\right| > M\rho_n/4\right].\end{aligned}$$

Because

$$\left|h_n^{-d} Y_{jn} \kappa\left(\frac{x - X_j}{h_n}\right) - \mathbb{E}\left[h_n^{-d} Y_{jn} \kappa\left(\frac{x - X_j}{h_n}\right)\right]\right| \leq 2\tau_n h_n^{-d} C_\kappa = 2C_\tau n^{1/s} h_n^{-d} C_\kappa,$$

and

$$\begin{aligned}\mathbb{V}\left[h_n^{-d} Y_{jn} \kappa\left(\frac{x - X_j}{h_n}\right)\right] &\leq h_n^{-d} \mathbb{E}\left[Y_{jn}^2 h_n^{-d} \kappa\left(\frac{x - X_j}{h_n}\right)^2\right] \\ &\leq h_n^{-d} \int_{\mathbb{R}^d} \mathbb{E}\left[|Y|^2\,|X=r\right] h_n^{-d} \kappa\left(\frac{x - r}{h_n}\right)^2 f_X(r)\, dr \\ &\leq h_n^{-d} C_Y(2) \int_{\mathbb{R}^d} \kappa(t)^2\, dt \leq h_n^{-d} C_Y(2)\, C_\kappa^2,\end{aligned}$$

it follows from Bernstein's inequality that

$$\Pr\left[\left|\tilde{\Psi}_n\left(x_{g,n}^*\right) - \mathbb{E}\tilde{\Psi}_n\left(x_{g,n}^*\right)\right| > M\rho_n/4\right] \leq 2\exp\left[-\frac{n h_n^d \rho_n^2 M^2/32}{C_Y(2)\, C_\kappa^2 + \frac{1}{6} M C_\tau C_\kappa \rho_n n^{1/s}}\right].$$

Similarly,

$$\Pr\left[\left|\tilde{\Psi}_n^*\left(x_{g,n}^*\right)-\mathbb{E}\tilde{\Psi}_n^*\left(x_{g,n}^*\right)\right|>M\rho_n/4\right]\leq 2\exp\left[-\frac{nh_n^d\rho_n^2M^2/32}{C_Y\left(2\right)C_{\kappa^*}^2+\frac{1}{6}MC_\tau C_{\kappa^*}\rho_n n^{1/s}}\right],$$

so if $\rho_n\leq\min\left(1,\delta_\kappa\right)$ and $M\geq 4C_Y\left(1\right)C_{\kappa^*}$, then

$$\begin{aligned}&\max_{1\leq g\leq G_n}\Pr\left[\sup_{x\in\mathcal{X}_{g,n}}\left|\tilde{\Psi}_n\left(x\right)-\mathbb{E}\tilde{\Psi}_n\left(x\right)\right|>M\rho_n\right]\\&\leq 4\exp\left[-\frac{nh_n^d\rho_n^2M^2/32}{C_Y\left(2\right)\max\left(C_\kappa,C_{\kappa^*}\right)^2+\frac{1}{6}MC_\tau\max\left(C_\kappa,C_{\kappa^*}\right)\rho_n n^{1/s}}\right].\end{aligned}$$

To complete the proof, we let $R<\infty$ be given and use the bound just obtained to exhibit an $M$ such that (5) holds.

First, suppose $\underline{\lim}_{n\to\infty}n^{1-2/s}h_n^d/\log n>0$, in which case there exists a $\underline{C}_h>0$ such that

$$\rho_n n^{1/s}=\sqrt{\frac{\log n}{n^{1-2s}h_n^d}}\max\left(1,\sqrt{\frac{\log n}{n^{1-2/s}h_n^d}}\right)\leq\frac{1}{\underline{C}_h}$$

for all $n$ large enough. For any such $n$,

$$\begin{aligned}&\frac{nh_n^d\rho_n^2M^2/32}{C_Y\left(2\right)\max\left(C_\kappa,C_{\kappa^*}\right)^2+\frac{1}{6}MC_\tau\max\left(C_\kappa,C_{\kappa^*}\right)\rho_n n^{1/s}}\\&\geq\frac{M^2/32}{C_Y\left(2\right)\max\left(C_\kappa,C_{\kappa^*}\right)^2+\frac{1}{6}MC_\tau\max\left(C_\kappa,C_{\kappa^*}\right)/\underline{C}_h}\log n,\end{aligned}$$

so if $n$ is large enough and if $M\geq 4C_Y\left(1\right)C_{\kappa^*}$, then

$$\begin{aligned}&\max_{1\leq g\leq G_n}\Pr\left[\sup_{x\in\mathcal{X}_{g,n}}\left|\tilde{\Psi}_n\left(x\right)-\mathbb{E}\tilde{\Psi}_n\left(x\right)\right|>M\rho_n\right]\\&\leq 4n^{-M^2/32\left[C_Y(2)\max(C_\kappa,C_{\kappa^*})^2+\frac{1}{6}MC_\tau\max(C_\kappa,C_{\kappa^*})/\underline{C}_h\right]},\end{aligned}$$

implying in particular that (5) holds if $M$ is large enough.

Next, suppose $\overline{\lim}_{n\to\infty} n^{1-2/s}h_n^d/\log n < \infty$, in which case there exists a $\overline{C}_h < \infty$ such that

$$\frac{n^{1-2/s}h_n^d}{\log n} \leq \overline{C}_h, \qquad \frac{n^{1-2/s}h_n^d}{\log n}\rho_n n^{1/s} = \max\left(1, \sqrt{\frac{n^{1-2/s}h_n^d}{\log n}}\right) \leq \overline{C}_h$$

for all $n$ large enough. For any such $n$,

$$\begin{aligned}
&\frac{nh_n^d\rho_n^2 M^2/32}{C_Y(2)\max(C_\kappa, C_{\kappa^*})^2 + \frac{1}{6}MC_\tau \max(C_\kappa, C_{\kappa^*})\rho_n n^{1/s}} \\
&\geq \frac{M^2/32}{C_Y(2)\max(C_\kappa, C_{\kappa^*})^2 \frac{n^{1-2/s}h_n^d}{\log n} + \frac{1}{6}MC_\tau \max(C_\kappa, C_{\kappa^*})\frac{n^{1-2/s}h_n^d}{\log n}\rho_n n^{1/s}}\log n \\
&\geq \frac{M^2/32}{C_Y(2)\max(C_\kappa, C_{\kappa^*})^2\overline{C}_h + \frac{1}{6}MC_\tau \max(C_\kappa, C_{\kappa^*})\overline{C}_h}\log n,
\end{aligned}$$

so if $n$ is large enough and if $M \geq 4C_Y(1)C_{\kappa^*}$, then

$$\begin{aligned}
&\max_{1\leq g\leq G_n} \Pr\left[\sup_{x\in\mathcal{X}_{g,n}}\left|\tilde{\Psi}_n(x) - \mathbb{E}\tilde{\Psi}_n(x)\right| > M\rho_n\right] \\
&\leq 4n^{-M^2/32\left[C_Y(2)\max(C_\kappa,C_{\kappa^*})^2\overline{C}_h + \frac{1}{6}MC_\tau\max(C_\kappa,C_{\kappa^*})\overline{C}_h\right]},
\end{aligned}$$

implying once again that (5) holds if $M$ is large enough.

Finally, suppose $\overline{\lim}_{n\to\infty} n^{1-2/s}h_n^d/\log n = \infty$ and $\underline{\lim}_{n\to\infty} n^{1-2/s}h_n^d/\log n = 0$. Suppose that for some $\varepsilon > 0$ and for every $M$, there exists a subsequence $n'$ with

$$\Pr\left[\sup_{x\in\mathcal{X}_{n'}}\left|\tilde{\Psi}_{n'}(x) - \mathbb{E}\left[\tilde{\Psi}_{n'}(x)\right]\right| > M\rho_{n'}\right] > \varepsilon$$

for every $n'$. Given $\varepsilon > 0$, pick an $M \geq 4C_Y(1)C_{\kappa^*}$ satisfying

$$\overline{\lim}_{n\to\infty} G_n n^{-M^2/32\left[C_Y(2)\max(C_\kappa,C_{\kappa^*})^2 + \frac{1}{6}MC_\tau\max(C_\kappa,C_{\kappa^*})\right]} < \varepsilon/4.$$

Any subsequence $n'$ contains a further subsubsequence $n''$ along which

$$\overline{\lim}_{n''\to\infty}\left(n''\right)^{1-2/s}h_{n''}^{d}/\log n''=\underline{\lim}_{n''\to\infty}\left(n''\right)^{1-2/s}h_{n''}^{d}/\log n''\in[0,\infty].$$

Along such subsubsequences the previous results can be used to show that

$$\overline{\lim}_{n''\to\infty}\Pr\left[\sup_{x\in\mathcal{X}_{n''}}\left|\tilde{\Psi}_{n''}(x)-\mathbb{E}\tilde{\Psi}_{n''}(x)\right|>M\rho_{n''}\right]<\varepsilon,$$

a contradiction. ■

**2.2. Proof of Lemma B-2.** Because $\Psi_{n,i}(x)=\Psi_n(x)$ and

$$\hat{\Psi}_{n,i}(x)=\frac{n}{n-1}\hat{\Psi}_n(x)-\frac{1}{(n-1)h_n^d}Y_i\kappa\left(\frac{x-X_i}{h_n}\right),$$

we have the elementary bound

$$\begin{aligned}\left|\hat{\Psi}_{n,i}(x)-\Psi_{n,i}(x)\right| \leq\ & (1-n^{-1})^{-1}\left|\hat{\Psi}_n(x)-\Psi_n(x)\right|+(n-1)^{-1}\mathbb{E}\left[\left|\hat{\Psi}_n(x)\right|\right]\\ &+(n-1)^{-1}h_n^{-d}\left|Y_{in}\kappa\left(\frac{x-X_i}{h_n}\right)\right|\\ &+(n-1)^{-1}h_n^{-d}\left|\left(Y_i-Y_{in}\right)\kappa\left(\frac{x-X_i}{h_n}\right)\right|,\end{aligned}$$

where $Y_{in}=Y_i\mathbf{1}(|Y_i|\leq\tau_n)$ with $\tau_n=O\left(n^{1/s}\right)$. The first term on the right is covered by Lemma B-1, the second term is $O\left(n^{-1}\right)$, and the third term satisfies

$$(1-n^{-1})^{-1}h_n^{-d}\left|Y_{in}\kappa\left(\frac{x-X_i}{h_n}\right)\right|\leq(n-1)^{-1}h_n^{-d}\tau_nC_\kappa=O\left(n^{1/s-1}h_n^{-d}\right),$$

where

$$n^{1/s-1}h_n^{-d}=\sqrt{\frac{1}{nh_n^d}}\sqrt{\frac{1}{n^{1-2/s}h_n^d}}=o\left(\rho_n\right).$$

Finally, the fourth term is negligible because

$$\Pr\left[\max_{1\le i\le n}(n-1)^{-1}h_n^{-d}\left|(Y_i-Y_{in})\,\kappa\left(\frac{x-X_i}{h_n}\right)\right|>0\right]=\Pr\left[Y_i\neq Y_{in}\text{ for some }i\right]$$

can be made arbitrarily close to zero. ■

**2.3. Proof of Lemma B-3.** By Markov's inequality,

$$\Pr\left[\max_{1\le i\le n}\|X_i\|>n^{2/s_X}\right]\le n\Pr\left[\|X\|^{s_X}>n^2\right]\le n^{-1}\mathbb{E}[\|x\|^{s_X}]=o\left(1\right).$$

Setting $C_{X,n}=n^{2/s_X}$, we therefore have

$$\max_{1\le i\le n}\left|\hat{\Psi}_n(X_i)-\Psi_n(X_i)\right|\le\sup_{x\in\mathcal{X}_n}\left|\hat{\Psi}_n(x)-\Psi_n(x)\right|$$

and

$$\max_{1\le i\le n}\left|\hat{\Psi}_{n,i}(X_i)-\Psi_{n,i}(X_i)\right|\le\max_{1\le i\le n}\sup_{x\in\mathcal{X}_n}\left|\hat{\Psi}_{n,i}(x)-\Psi_{n,i}(x)\right|$$

with probability approaching one. The result now follows from Lemmas B-1 and B-2. ■