Econometric Theory, **30**, 2014, 176–200.
doi:10.1017/S0266466613000169

SMALL BANDWIDTH ASYMPTOTICS FOR DENSITY-WEIGHTED AVERAGE DERIVATIVES

MATIAS D. CATTANEO
University of Michigan

RICHARD K. CRUMP
Federal Reserve Bank of New York

MICHAEL JANSSON
UC Berkeley and CREATES

This paper proposes (apparently) novel standard error formulas for the density-weighted average derivative estimator of Powell, Stock, and Stoker (*Econometrica* 57, 1989). Asymptotic validity of the standard errors developed in this paper does not require the use of higher-order kernels, and the standard errors are "robust" in the sense that they accommodate (but do not require) bandwidths that are smaller than those for which conventional standard errors are valid. Moreover, the results of a Monte Carlo experiment suggest that the finite sample coverage rates of confidence intervals constructed using the standard errors developed in this paper coincide (approximately) with the nominal coverage rates across a nontrivial range of bandwidths.

1. INTRODUCTION

Semiparametric estimators employing nonparametric kernel estimators of unknown nuisance functions have been proposed for a variety of microeconometric estimands. Under suitable application-specific regularity conditions, many such estimators enjoy the properties of $\sqrt{n}$-consistency (where n is the sample size) and asymptotic normality, the variance of the limiting distribution being

The authors thank Oliver Linton, two referees, Bryan Graham, Jim Powell, Tom Rothenberg, Paul Ruud, and seminar participants at Arizona, Berkeley, Boston University, Cornell, Harvard, Michigan, Michigan State, Ohio State, Penn State, Western Ontario, the 2008 Latin American Meeting of the Econometric Society, the 2009 North American Summer Meeting of the Econometric Society, the 2009 All UC Econometrics Conference, and the 2010 International Symposium on Econometric Theory and Applications for comments. We thank Jasjeet Sekhon and Rocio Titiunik for providing access to the Calgrid cluster. Cattaneo gratefully acknowledges financial support from the National Science Foundation (SES 0921505). Jansson gratefully acknowledges financial support from the National Science Foundation (SES 0920953) and the research support of *CREATES* (funded by the Danish National Research Foundation). Address correspondence to Michael Jansson, University of California, Berkeley, Department of Economics, 530 Evans Hall #3880, Berkeley, CA 94720-3880; e-mail: mjansson@econ.berkeley.edu.

consistently estimable and invariant with respect to the kernel and bandwidth of the nonparametric estimator.

Achieving these properties often requires a delicate choice of the kernel and bandwidth of the nonparametric estimator. A prime example, and the one we focus on in this paper, is provided by the density-weighted average derivative estimator of Powell, Stock, and Stoker (1989, henceforth PSS). The validity of inference procedures based on this estimator and the standard errors proposed by PSS require that the bandwidth and the order of the kernel be chosen in a way that meets two distinct requirements. On the one hand, the bias of the estimator must be negligible relative to its standard deviation, a requirement that can be met by making the bandwidth "small enough" and the order of the kernel "large enough." At the same time, the bandwidth needs to be large enough to ensure that the estimator is asymptotically linear (i.e., asymptotically equivalent to a sample average).[1]

The range of bandwidths that are simultaneously small enough to meet the bias requirement and large enough to meet the asymptotic linearity requirement is often quite narrow,[2] suggesting that in samples of moderate size the inference procedures exhibit a certain "nonrobustness" with respect to the bandwidth. Although the tension between the lower and upper bounds on the bandwidth imposed by the bias and asymptotic linearity requirements can be eased by increasing the order of the kernel, estimators employing higher-order kernels are commonly believed to have poor small-sample properties (e.g., Robinson, 1988, p. 938; Hristache, Juditsky, and Spokoiny, 2001, p. 597). It would therefore appear to be of interest to explore alternative ways of achieving "robustness" with respect to the bandwidth.

In an attempt to achieve such robustness, this paper explores the consequences of employing bandwidth sequences that are not large enough for asymptotic linearity to hold (on the part of PSS's estimator). It turns out that if the assumption on the bandwidth that implies asymptotic linearity is violated, then PSS's standard errors exhibit an upward bias that renders the associated inference procedures conservative.[3] In contrast, we show that valid (nonconservative) inference can be based on PSS's estimator provided it is combined with a robust standard error that accommodates (but does not require) the failure of asymptotic linearity. Specifically, this paper proposes an apparently novel standard error (matrix) formula for PSS's estimator, and we give conditions under which asymptotic standard normality holds for PSS's estimator when centered at the truth and standardized by the robust standard error matrix proposed in this paper.

As do existing procedures, the procedure developed in this paper requires that the bandwidth be large enough for certain quantities to be asymptotically negligible, but the lower bound in this paper is considerably weaker than the bounds that have appeared elsewhere in the literature. In addition to (possibly) increasing our confidence in the standard normal approximation upon which inference procedures are based, the weakening of the lower bound on the bandwidth also has potentially interesting implications for our ability to control the bias of the estimator. Indeed, our results involve a weakening of the lower bound on the order of the kernel that enables us to provide a formal justification for the use of procedures that avoid the use of higher-order kernels altogether.

To achieve our goals, we first characterize the asymptotic distribution of PSS's estimator under conditions on the kernel and the bandwidth that are weaker than those entertained in the existing literature. Specifically, we show that PSS's estimator is asymptotically normal (with correct centering) across a wide range of bandwidths, with the rate of convergence and the variance of the limiting distribution depending on the bandwidth (and, in case of the variance, also the kernel) in those cases where the bandwidth violates the conditions imposed by PSS. Although a range of possibilities (indexed by the limiting behavior of the bandwidth) arise on the part of the asymptotic distribution of the estimator, a natural unification of the results is available: The estimation error premultiplied by the inverse of a square root of its variance matrix is asymptotically standard normal in all of the cases considered.

In addition to having the intuitively appealing feature that it captures (at least partially) the dependence of the distribution theory on some specifics of the kernel and the bandwidth, the unification is constructive insofar as it suggests how valid standard errors can be obtained, and we use it to obtain valid standard errors in three distinct ways. The first construction is conceptually straightforward and proceeds by replacing the unknown parameters in an asymptotic expansion of the variance by consistent analog estimators. A potential disadvantage of this approach is that a separate bandwidth parameter is needed to ensure consistency of the analog estimators employed. Our second construction circumvents this potential problem and exploits the intriguing fact that although PSS's variance estimator is inconsistent in general, a simple downward adjustment of this estimator produces standard errors that are valid in all of the cases considered. Finally, our third construction achieves validity by implementing PSS's variance estimator with a bandwidth given by a known, constant multiple of the bandwidth used when constructing the estimator of the parameter of interest.

In an obvious way, our work can be viewed as a continuation of the seminal work by PSS. As suggested by the title, our main contribution is to accommodate small values of the bandwidth parameter. Other work closely related to the present work is Robinson (1995) and Nishiyama and Robinson (2000, 2001, 2005). Our first-order asymptotic analysis is conceptually distinct from (and valid under weaker assumptions on the bandwidth and the kernel order than) the higher-order asymptotic theory developed in those papers, but our motivation is similar, and our proofs are facilitated by the fact that we are able to make heavy use of some of the technical results obtained in Robinson (1995) and Nishiyama and Robinsons (2000). Furthermore, and not unexpectedly in view of the fact that our analysis is based on a characterization of the joint limiting distribution of the terms in a stochastic expansion of PSS's estimator, it turns out that the results we obtain are in qualitative agreement with some of the findings of Nishiyama and Robinson (2000, 2001, 2005). Finally, the approach taken in this paper is similar in spirit to that of Kiefer, Vogelsang, and Bunzel (2000) and Kiefer and Vogelsang (2002a, 2002b, 2005), a common feature being that the effect of a nonparametric ingredient is accounted for by considering sequences of tuning parameters corresponding to undersmoothing that is sufficiently severe to affect the first-order asymptotic properties of the statistic of interest.[4]

The next section lists assumptions and presents our theoretical results. Section 3 reports Monte Carlo evidence, while Section 4 offers concluding remarks. Proofs of the theoretical results are collected in an Appendix.

2. ASSUMPTIONS AND RESULTS

2.1. Assumptions

Suppose $z_i = \left(y_i, x_i'\right)'$ $(i = 1, \ldots, n)$ are independent and identically distributed (i.i.d.) copies of a vector $z = \left(y, x'\right)'$, where $y \in \mathbb{R}$ is a dependent variable and $x \in \mathbb{R}^d$ is a continuous explanatory variable with density $f(\cdot)$. As pointed out by PSS, an interesting functional of the regression function $g(x) = \mathbb{E}(y|x)$ is its density-weighted average derivative vector, which is defined as[5]

$$\theta = \mathbb{E}\left[f(x)\frac{\partial}{\partial x}g(x)\right]. \quad \textbf{(1)}$$

The following assumption, adapted from Nishiyama and Robinsons (2000), ensures that θ is well defined and imposes additional regularity conditions that will facilitate the subsequent development of theoretical results.

Assumption 1.

(a) $\mathbb{E}\left(y^4\right) < \infty$.

(b) $\mathbb{E}[\mathbb{V}(y|x) f(x)] > 0$ and $\mathbb{V}[\partial e(x)/\partial x - y\partial f(x)/\partial x]$ is positive definite, where $e(x) = f(x)g(x)$.

(c) f is $(Q+1)$ times differentiable, and f and its first $(Q+1)$ derivatives are bounded, for some $Q \geq 2$.

(d) g is twice differentiable, and e and its first two derivatives are bounded.

(e) v is differentiable, and vf and its first derivative are bounded, where $v(x) = \mathbb{E}\left(y^2|x\right)$.

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

Under Assumption 1, it follows from integration by parts that the density-weighted average derivative vector in (1) admits the representation

$$\theta = -2\mathbb{E}\left[y\frac{\partial}{\partial x}f(x)\right],$$

PSS's analog estimator of which is given by

$$\hat{\theta}_n = -2n^{-1}\sum_{i=1}^{n} y_i \frac{\partial}{\partial x}\hat{f}_{n,i}(x_i),$$

where $\hat{f}_{n,i}(\cdot)$ is a "leave one out" kernel density estimator defined as

$$\hat{f}_{n,i}(x) = (n-1)^{-1} \sum_{\substack{j=1 \\ j \neq i}}^{n} h_n^{-d} K\left(\frac{x - x_j}{h_n}\right)$$

for some kernel $K : \mathbb{R}^d \to \mathbb{R}$ and some positive (bandwidth) sequence h_n.

On the part of the kernel, we make the following assumption.

Assumption 2.

(a) K is even.

(b) K is differentiable, and K and its first derivative are bounded.

(c) $\int_{\mathbb{R}^d} \dot{K}(u) \dot{K}(u)' du$ is positive definite, where $\dot{K}(u) = \partial K(u) / \partial u$.

(d) For some $P \geq 2$,

$$\int_{\mathbb{R}^d} |K(u)| \left(1 + \|u\|^P\right) du + \int_{\mathbb{R}^d} \left\|\dot{K}(u)\right\| \left(1 + \|u\|^2\right) du < \infty$$

and

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

When $P > 2$, Assumption 2 implies that K is a higher-order kernel. The use of such kernels is standard in the existing literature on density-weighted average derivatives (e.g., PSS, Powell and Stoker (1996), Robinson (1995), Nishiyama and Robinson (2000, 2001, 2005), and Newey, Hsieh, and Robins (2004)). Among other things, this paper addresses the question of whether valid inference on θ can be based on $\hat{\theta}_n$ even if $P = 2$ (e.g., if a Gaussian kernel is employed).

2.2. Distribution Theory

To motivate the question of whether the use of a higher-order kernel can be avoided, recall (e.g., from Theorem 3.3 of PSS) that if Assumptions 1 and 2 hold and if $nh_n^{2\min(P,Q)} \to 0$ and $nh_n^{d+2} \to \infty$, then

$$\sqrt{n}\left(\hat{\theta}_n - \theta\right) \to_d \mathcal{N}(0, \Sigma), \tag{2}$$

where

$$\Sigma = \mathbb{E}\left[L(z) L(z)'\right], \qquad L(z) = 2\left[\frac{\partial}{\partial x} e(x) - y \frac{\partial}{\partial x} f(x) - \theta\right].$$

(Here and elsewhere in the paper, limits are taken as $n \to \infty$ unless otherwise noted.) In the statement of this result, the conditions $nh_n^{2P} \to 0$ and $nh_n^{d+2} \to \infty$ are minimal in the sense that (2) can fail if one (or both) of the assumptions is (are) relaxed.[6] Because a necessary condition for existence of a bandwidth sequence h_n compatible with both assumptions is that $P > (d+2)/2$, it may appear that the use of a higher-order kernel is unavoidable unless $d = 1$.

Under Assumptions 1 and 2, the assumptions $h_n \to 0$ and $nh_n^{d+2} \to \infty$ imply that

$$n\mathbb{V}\left(\hat{\theta}_n\right) = \Sigma + o(1).$$

Therefore, an alternative statement of PSS's Theorem 3.3 is the following: If Assumptions 1 and 2 hold and if $nh_n^{2\min(P,Q)} \to 0$ and $nh_n^{d+2} \to \infty$, then

$$\mathbb{V}\left(\hat{\theta}_n\right)^{-1/2}\left(\hat{\theta}_n - \theta\right) \to_d \mathcal{N}(0, I_d). \tag{3}$$

As it turns out, the conditions on h_n can be weakened considerably without invalidating this convergence result.

THEOREM 1. *If Assumptions 1 and 2 hold and if* $\min(nh_n^{d+2}, 1)nh_n^{2\min(P,Q)} \to 0$ *and* $n^2h_n^d \to \infty$, *then (3) is true.*

The conditions of this theorem weaken those of PSS in two respects. First, the condition $n^2h_n^d \to \infty$ is considerably weaker than the condition $nh_n^{d+2} \to \infty$. As further explained below, this relaxation of the lower bound on the bandwidth is possible because our method of proof accommodates cases where $\hat{\theta}_n$ is not asymptotically equivalent to its Hájek projection. Second, due to the presence of the additional term $\min\left(nh_n^{d+2}, 1\right)$, our "bias" condition is weaker than the condition $nh_n^{2\min(P,Q)} \to 0$ of PSS. As usual, we need the bias of the estimator to be of smaller order of magnitude than the standard deviation. The term $\min\left(nh_n^{d+2}, 1\right)$ in the bias condition reflects the fact (further discussed below) that the rate of convergence of the estimator is slower than $\sqrt{n}$ when $nh_n^{d+2} \to 0$.

Partly due to the presence of $\min\left(nh_n^{d+2}, 1\right)$ in the bias condition, Theorem 1 accommodates smaller values of P than do the results of PSS.[7] Indeed, for any value of d there exists a bandwidth sequence h_n compatible with the assumptions of Theorem 1 even if $P = 2$.[8] In other words, Theorem 1 suggests that the use of higher-order kernels can be avoided irrespective of the value of d. As will be shown below, this positive message remains true also when studentized statistics are considered (i.e., when $\mathbb{V}\left(\hat{\theta}_n\right)$ is replaced by a suitable estimator in (3)).

As in PSS, the starting point for our analysis is the variable U-statistic (i.e., U-statistic with an n-dependent kernel) representation of $\hat{\theta}_n$,

$$\hat{\theta}_n = \binom{n}{2}^{-1}\sum_{i=1}^{n-1}\sum_{j=i+1}^{n} U\left(z_i, z_j; h_n\right),$$

$$U\left(z_i, z_j; h\right) = -h^{-(d+1)}\dot{K}\left(\frac{x_i - x_j}{h}\right)\left(y_i - y_j\right).$$

The Hoeffding decomposition of $\hat{\theta}_n$ is

$$\hat{\theta}_n = \theta_n + \bar{L}_n + \bar{W}_n,$$

where

$$\theta_n = \theta(h_n), \qquad \bar{L}_n = n^{-1}\sum_{i=1}^{n} L(z_i; h_n),$$
$$\bar{W}_n = \binom{n}{2}^{-1}\sum_{i=1}^{n-1}\sum_{j=i+1}^{n} W\left(z_i, z_j; h_n\right),$$

with

$$\theta(h) = \mathbb{E}\left[U\left(z_i, z_j; h\right)\right], \qquad L(z_i; h) = 2\left[\mathbb{E}\left(U\left(z_i, z_j; h\right)|z_i\right) - \theta(h)\right],$$
$$W\left(z_i, z_j; h\right) = U\left(z_i, z_j; h\right) - \frac{1}{2}\left[L(z_i; h) + L\left(z_j; h\right)\right] - \theta(h).$$

The projection theorem for variable U-statistics (e.g., Lem. 3.1 of PSS) gives sufficient conditions for $\bar{W}_n$, the difference between $\hat{\theta}_n$ and its Hájek projection, to be asymptotically negligible in the sense that $\sqrt{n}\bar{W}_n \to_p 0$. To handle cases where this projection theorem provides insufficient technical machinery to establish asymptotic normality of $\hat{\theta}_n$ (because $\sqrt{n}\bar{W}_n \nrightarrow_p 0$), the proof of Theorem 1 obtains a characterization of the joint limiting distribution of $\bar{L}_n$ and $\bar{W}_n$. Specifically, it is shown in the Appendix that if Assumptions 1 and 2 hold and if $h_n \to 0$ and $n^2 h_n^d \to \infty$, then

$$\begin{pmatrix} \sqrt{n}\bar{L}_n \\ \sqrt{\binom{n}{2} h_n^{d+2}}\bar{W}_n \end{pmatrix} \to_d \mathcal{N}\left[\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \Sigma & 0 \\ 0 & \Delta \end{pmatrix}\right] \tag{4}$$

where

$$\Delta = 2\mathbb{E}\left[\mathbb{V}(y|x)f(x)\right]\int_{\mathbb{R}^d}\dot{K}(u)\dot{K}(u)'du.$$

The proof of (4) employs a central limit theorem for sample averages and degenerate U-statistics due to Eubank and Wang (1999). To verify the conditions of this central limit theorem, we impose the lower bound $n^2 h_n^d \to \infty$ on the bandwidth sequence and utilize some technical lemmas due to Robinson (1995) and Nishiyama and Robinsons (2000). Because the condition $n^2 h_n^d \to \infty$ is considerably weaker than the condition $nh_n^{d+2} \to \infty$ needed for the result $\sqrt{n}\bar{W}_n \to_p 0$, we can accommodate a significantly wider range of bandwidths by basing the distribution theory on (4) rather than a result that requires $\sqrt{n}\bar{W}_n \to_p 0$. Indeed, because $n^2 h_n^{d+2}$ can fail to diverge (and may even vanish) without violating the condition $n^2 h_n^d \to \infty$, distribution theory based on (4) even covers certain cases where $\bar{W}_n \to_p 0$ fails to hold.

The formulation (3) is by no means without antecedents. Indeed, in his seminal paper on U-statistics Hoeffding (1948, p. 307) argues that in many applications it is desirable to standardize a U-statistic by its actual variance (rather than its asymptotic variance, namely the variance of its Hájek projection). Jing and Wang (2003, Thm. 1.1) can be used to show that for U-statistics whose kernels do not vary with the sample size, no asymptotic refinements are achieved by standardizing by the actual variance. In other words, the one-term Edgeworth expansions for a U-statistic standardized by its actual variance and asymptotic variance, respectively, coincide. Theorem 1 demonstrates by example that the situation can be very different for a U-statistic whose kernel does vary with the sample size.[9]

In view of (4), the situations covered by Theorem 1 can be classified according to the rate of decay of the bandwidth in the following way. First, if (and only if) $nh_n^{d+2} \to \infty$, then the first-order asymptotic behavior of $\hat{\theta}_n$ is dominated by $\bar{L}_n$ and the conventional result (2) holds. Even in this case, the results of Nishiyama and Robinsons (2000) suggest that formulation (3) can be attractive for certain (small) values of the bandwidth. Specifically, if the assumptions of Nishiyama and Robinsons (2000, Thm. 1) hold, and if $n^3 h_n^{2(d+2+\min(P,Q))} \to 0$ and $nh_n^{2(d+2)} \to 0$, then for any nonzero $\lambda \in \mathbb{R}^d$, the leading term in the Edgeworth expansion of the distribution of $\lambda' \left(\hat{\theta}_n - \theta\right) / \sqrt{n^{-1} \lambda' \Sigma \lambda}$ is a variance term that accounts for the variability of $\bar{W}_n$.[10] In other words, the leading term accounts for the fact that $n^{-1}\Sigma$ underestimates the variance of $\hat{\theta}_n$. This term can be removed by incorporating the term $2n^{-2} h_n^{-(d+2)} \Delta$ into the (approximate) variance of $\hat{\theta}_n$.[11] It is shown in the proof of Theorem 1 that

$$\mathbb{V}\left(\hat{\theta}_n\right) = n^{-1}[\Sigma + o(1)] + \binom{n}{2}^{-1} h_n^{-(d+2)} [\Delta + o(1)], \tag{5}$$

so it seems plausible that there are conditions under which an Edgeworth correction is achieved by the standardization used in (3).

Next, if $nh_n^{d+2} \to \kappa \in (0, \infty)$, then neither $\bar{L}_n$ nor $\bar{W}_n$ dominates the asymptotic behavior of $\hat{\theta}_n$ and the result becomes

$$\sqrt{n}\left(\hat{\theta}_n - \theta\right) \to_d \mathcal{N}\left(0, \Sigma + \frac{2}{\kappa}\Delta\right).$$

Because Δ and κ depend on the kernel and the bandwidth sequence, respectively, this result demonstrates by example that semiparametric estimators can be $\sqrt{n}$-consistent and asymptotically normally distributed without the limiting distribution being invariant with respect to the nonparametric estimator. This finding does not contradict Newey (1994, Prop. 1), as $\hat{\theta}_n$ ceases to be asymptotically linear when the condition $nh_n^{d+2} \to \infty$ is dropped.[12]

Finally, if $nh_n^{d+2} \to 0$, then $\bar{L}_n$ is asymptotically negligible, and we have

$$\sqrt{\binom{n}{2} h_n^{d+2}} \left(\hat{\theta}_n - \theta\right) \to_d \mathcal{N}(0, \Delta).$$

Even in this case, $\hat{\theta}_n$ is asymptotically normally distributed, but the rate of convergence is slower than $\sqrt{n}$. Indeed, if $n^2 h_n^{d+2} \nrightarrow \infty$, then $\hat{\theta}_n$ is not even consistent.

Remark 1. The asymptotic efficiency of $\hat{\theta}_n$ is maximized by employing a bandwidth sequence satisfying $nh_n^{d+2} \to \infty$. Indeed, although θ is not covered by the results of Newey and Stoker (1993) (because the weight function $f(\cdot)$ in (1) is unknown), by proceeding as in the proof of Newey and Stoker (Thm. 3.1) it can be shown that if certain regularity conditions hold, then $L(\cdot)$ is the pathwise derivative of θ. (See also Severini and Tripathi, 2001.) As a result, $\hat{\theta}_n$ enjoys semiparametric efficiency properties if (and only if) $nh_n^{d+2} \to \infty$.

Remark 2. If nh_n^{d+2} converges (in $\mathbb{R}$), then the asymptotic efficiency of $\hat{\theta}_n$ depends on the kernel through the functional $\int_{\mathbb{R}^d} \dot{K}(u)\dot{K}(u)' du$. The scalar counterpart of this functional arises in the context of estimation of the mode of a probability density (e.g., Parzen, 1962) and the results of Eddy (1980, Sec. 3) can be used to construct kernels minimizing $\int_{\mathbb{R}^d} \dot{K}(u)\dot{K}(u)' du$ (subject to certain conditions).

2.3. Variance Estimation

From a practical point of view, a shortcoming of statement (3) is that it involves the matrix $\mathbb{V}\left(\hat{\theta}_n\right)$, which is unknown. Replacing $\mathbb{V}\left(\hat{\theta}_n\right)$ by an estimator $\hat{V}_n$ (say), we obtain a studentized version of $\hat{\theta}_n$, and it is of interest to characterize conditions under which

$$\hat{V}_n^{-1/2}\left(\hat{\theta}_n - \theta\right) \to_d \mathcal{N}(0, I_d). \tag{6}$$

If (2) holds, then so does (6) provided $n\hat{V}_n$ is a consistent estimator of Σ, a requirement that is easily met (e.g., see Thm. 3.4 of PSS). More generally, it follows from (5) that if the assumptions of Theorem 1 hold and if $\hat{V}_n$ satisfies

$$\hat{V}_n = n^{-1}\Sigma + \binom{n}{2}^{-1} h_n^{-(d+2)}\Delta + o_p\left(n^{-1} + n^{-2}h_n^{-(d+2)}\right), \tag{7}$$

then (6) holds.

The requirement (7) can be met in various ways. Perhaps the most natural construction proceeds by first obtaining consistent estimators of Σ and Δ and then combining these in the manner suggested by (7). To that end, the following characterizations of Σ and Δ are useful:

$$\Sigma = \lim_{h \to 0} \mathbb{E}\left[L(z_i; h) L(z_i; h)'\right] \tag{8}$$

and

$$\lim_{h\to 0} h^{d+2}\mathbb{E}\left[W\left(z_i, z_j; h\right) W\left(z_i, z_j; h\right)'\right] = \Delta \qquad (i < j). \tag{9}$$

Analog estimators of Σ and Δ suggested by these characterizations are given by

$$\hat{\Sigma}_n = n^{-1}\sum_{i=1}^{n}\hat{L}_{n,i}\hat{L}_{n,i}', \qquad \hat{\Delta}_n = H_n^{d+2}\binom{n}{2}^{-1}\sum_{i=1}^{n-1}\sum_{j=i+1}^{n}\hat{W}_{n,ij}\hat{W}_{n,ij}',$$

where H_n is a bandwidth sequence and

$$\tilde{\theta}_n = \binom{n}{2}^{-1}\sum_{i=1}^{n-1}\sum_{j=i+1}^{n} U\left(z_i, z_j; H_n\right),$$

$$\hat{L}_{n,i} = 2\left[(n-1)^{-1}\sum_{\substack{j=1\\ j\neq i}}^{n} U\left(z_i, z_j; H_n\right) - \tilde{\theta}_n\right],$$

$$\hat{W}_{n,ij} = U\left(z_i, z_j; H_n\right) - \frac{1}{2}\left(\hat{L}_{n,i} + \hat{L}_{n,j}\right) - \tilde{\theta}_n.$$

The preceding definitions involve a bandwidth H_n that may differ from h_n. This generality is not merely spurious, as there are cases where it seems desirable to let the bandwidths H_n and h_n vanish at different rates. For instance, it turns out that if Assumptions 1 and 2 hold and if $H_n \to 0$ and $n^2 H_n^d \to \infty$, then

$$\hat{\Delta}_n \to_p \Delta \tag{10}$$

and

$$n^{-1}\hat{\Sigma}_n = n^{-1}\Sigma + 2\binom{n}{2}^{-1} H_n^{-(d+2)}\Delta + o_p\left(n^{-1} + n^{-2}H_n^{-(d+2)}\right), \tag{11}$$

so $\hat{\Sigma}_n$ is a consistent estimator of Σ only if $nH_n^{d+2} \to \infty$, a condition that is violated by $H_n = h_n$ in many of the cases covered by Theorem 1. The following result is an immediate consequence of (10) and (11).

THEOREM 2. *Suppose the assumptions of Theorem 1 hold.*

(a) If $H_n \to 0$ and $nH_n^{d+2} \to \infty$, then (6) holds for

$$\hat{V}_n = n^{-1}\hat{\Sigma}_n + \binom{n}{2}^{-1} h_n^{-(d+2)}\hat{\Delta}_n. \tag{12}$$

(b) If $H_n = h_n$, then (6) holds for

$$\hat{V}_n = n^{-1}\hat{\Sigma}_n - \binom{n}{2}^{-1} h_n^{-(d+2)}\hat{\Delta}_n. \tag{13}$$

(c) If $H_n = 2^{1/(d+2)} h_n$, then (6) holds for

$$\hat{V}_n = n^{-1} \hat{\Sigma}_n. \tag{14}$$

The theorem gives three distinct recipes for achieving (6). Of these, the most attractive would appear to be the constructions in parts (b) and (c), as these do not require the selection of a separate bandwidth. A possible disadvantage of the construction in part (b) is that it may fail to deliver a positive (semi-)definite variance estimator. In the Monte Carlo experiments reported below, this phenomenon occurred only very rarely, however, so this feature is probably not a cause for major concern.

Remark 3. When $H_n = h_n$, $\hat{\Sigma}_n$ is PSS's estimator of Σ. It follows from (11) and Theorem 1 that although (this estimator is inconsistent and) $\hat{V}_n = n^{-1}\hat{\Sigma}_n$ does not satisfy (7), it does enjoy the property that if the assumptions of Theorem 1 hold and if nh_n^{d+2} converges (in $\mathbb{R}$), then

$$\hat{V}_n^{-1/2} \left(\hat{\theta}_n - \theta\right) \to_d \mathcal{N}(0, I_d - J),$$

where J is some positive definite matrix (the value of which depends on the limiting value of nh_n^{d+2}). As a consequence, inference based on PSS's standard error matrix is asymptotically conservative when $nh_n^{d+2} \nrightarrow \infty$.

Remark 4. The proof of (10) implicitly establishes the consistency of two additional estimators of Δ, namely

$$\hat{\Delta}_{2,n} = H_n^{d+2}\left[\binom{n}{2}^{-1} \sum_{i=1}^{n-1} \sum_{j=i+1}^{n} U\left(z_i, z_j; H_n\right) U\left(z_i, z_j; H_n\right)' - \tilde{\theta}_n \tilde{\theta}_n'\right],$$

and

$$\hat{\Delta}_{3,n} = H_n^{d+2} \binom{n}{2}^{-1} \sum_{i=1}^{n-1} \sum_{j=i+1}^{n} U\left(z_i, z_j; H_n\right) U\left(z_i, z_j; H_n\right)'.$$

These are also analog estimators because

$$\begin{aligned}\Delta &= \lim_{h\to 0} h^{d+2} \mathbb{V}\left[U\left(z_i, z_j; h\right)\right] \\ &= \lim_{h\to 0} h^{d+2} \mathbb{E}\left[U\left(z_i, z_j; h\right) U\left(z_i, z_j; h\right)'\right] \qquad (i < j),\end{aligned}$$

where the first equality follows from (8) and (9), while the second equality uses the fact that $\lim_{h\to 0} \theta(h) = \theta$.

Remark 5. Being a variable U-statistic, PSS's estimator can be represented as a minimizer of a variable U-process,

$$\hat{\theta}_n = \arg\min_t \binom{n}{2}^{-1} \sum_{i=1}^{n-1} \sum_{j=i+1}^{n} Q\left(z_i, z_j, t; h_n\right),$$

$$Q\left(z_i, z_j, t; h\right) = \left\| U\left(z_i, z_j; h\right) - t \right\|^2.$$

It seems plausible that results analogous to those derived in this paper can be obtained for other estimators that can be represented as minimizers of variable U-processes, such as the pairwise difference estimators discussed in Aradillas-Lopez, Honore, and Powell (2007, pp. 1120–1122). On the other hand, for semiparametric two-step estimators of the form studied by Newey and McFadden (1994, Sec. 8.3), the results of Ichimura and Linton (2005) suggest that linearity with respect to the first-step kernel estimator is crucial for our results. Research currently under way has confirmed this suspicion in the case of the kernel-based weighted average derivative estimators studied by Newey and Stoker (1993).

Remark 6. It would be of interest to develop a higher-order approximation to the distribution of $\hat{V}_n^{-1/2}\left(\hat{\theta}_n - \theta\right)$, especially in the case where $nh_n^{d+2} \to 0$. In this case $\hat{\theta}_n$ is asymptotically equivalent to the degenerate U-statistic $\bar{W}_n$, so it seems plausible that the approach of Fan and Linton (2003) can be used.

3. MONTE CARLO EVIDENCE

We conducted a Monte Carlo experiment to investigate the finite-sample properties of our procedure and the procedures of PSS and Nishiyama and Robinsons (2000). Specifically, to assess whether the robustness property of our procedure holds in small samples, we provide results on the coverage rate of 95% confidence intervals constructed using a variety of bandwidths.

3.1. Setup

We consider six different models. The models are all of the (single index) form

$$y_i = \tau\left(y_i^*\right), \qquad y_i^* = x_i'\beta + \varepsilon_i,$$

where $\tau(\cdot)$ is a nondecreasing (link) function and $\varepsilon_i \sim \mathcal{N}(0,1)$ is independent of the bivariate regressor $x_i = (x_{1i}, x_{2i})'$. Three different link functions are considered, namely $\tau(y^*) = y^*$, $\tau(y^*) = \mathbf{1}\{y^* > 0\}$, and $\tau(y^*) = y^*\mathbf{1}\{y^* > 0\}$, where $\mathbf{1}(\cdot)$ is the indicator function. (These specifications correspond to a linear regression, probit, and Tobit model, respectively.). Two specifications of the regressors are considered. In both cases, the regressors have mean zero, unit variance, and are independent. Specifically, $x_{2i} \sim \mathcal{N}(0,1)$ throughout, while two distinct distributions are considered for x_{1i}, namely $x_{1i} \sim \mathcal{N}(0,1)$ and $x_{1i} \sim \varkappa$, where $\varkappa$ is a normalized chi-square random variable with 4 degrees of freedom (i.e., $\varkappa = \left(\chi_4^2 - 4\right)/\sqrt{8}$).[13] The latter choice of distribution was included to ensure that our results were not unduly influenced by the joint normality of the regressors. Throughout the experiment we set $\beta = (1,1)'$ and concentrate on the first component of $\theta = (\theta_1, \theta_2)'$, since the results for the second component were very similar.

Table 1 summarizes the Monte Carlo models, reports the value of the population parameter of interest, and provides the corresponding label of each model

TABLE 1. Monte carlo models

	$y_i = y_i^*$	$y_i = \mathbf{1}\{y_i^* > 0\}$	$y_i = y_i^* \mathbf{1}\{y_i^* > 0\}$
$x_{1i} \sim \mathcal{N}(0,1)$	Model 1: $\theta_1 = \frac{1}{4\pi}$	Model 3: $\theta_1 = \frac{1}{8\pi^{3/2}}$	Model 5: $\theta_1 = \frac{1}{8\pi}$
$x_{1i} \sim \varkappa$	Model 2: $\theta_1 = \frac{1}{4\sqrt{2\pi}}$	Model 4: $\theta_1 = 0.02795$	Model 6: $\theta_1 = 0.03906$

considered. In Models 4 and 6 a tidy closed-form expression is unavailable for θ_1, and we therefore report a numerical approximation instead. Models 1 through 4 were studied by PSS in their simulation study,[14] while Model 5 corresponds to the one employed in the simulation study of Nishiyama and Robinsons (2000).

We consider two sample sizes, $n = 100$ and $n = 400$, and for each case we carry out $S = 10,000$ simulations. We report results utilizing a second-order kernel ($P = 2$) implemented by a standard Gaussian product kernel, and a higher-order kernel ($P = 4$) constructed using a Gaussian density-based multiplicative kernel as discussed in Nishiyama and Robinsons (2000, pp. 943–944). We also explored other choices of kernel functions, such as a twiced Gaussian kernel (e.g., Newey et al., 2004), but the results were qualitatively similar and therefore we omit them to conserve space.

We consider four competing procedures for inference. First, using the results of PSS we constructed confidence intervals employing $\hat{V}_n = n^{-1}\hat{\Sigma}_n$ (see the remark at the end of Section 2). Second, following Nishiyama and Robinsons (2000, p. 958) we computed higher-order corrected (asymmetric) confidence bounds, which required the estimation of additional correction terms. We estimated these additional quantities using sample analogues and choices of tuning parameters as discussed in Nishiyama and Robinsons (2000). Finally, the third and fourth inference procedures are the ones described in parts (b) and (c) of Theorem 2. We investigate the relative virtues of each procedure by implementing them for an array of bandwidths ranging from 0.01 to 1.

3.2. Results

In Figures 1 to 4 we plot the empirical coverage for the competing 95% confidence intervals as a function of the choice of bandwidth for each of the six models. As discussed previously, we report four inference procedures: PSS's procedure, Nishiyama and Robinsons (2000) higher-order corrected procedure, abbreviated "NR" for simplicity, and the procedures introduced in parts (b) and (c) of Theorem 2. The latter are denoted by "CCJ1" and "CCJ2", respectively. To facilitate comparison, we plot the results only for a restricted range of bandwidths and include two additional horizontal lines at 0.90 and at the nominal coverage rate 0.95 for reference.

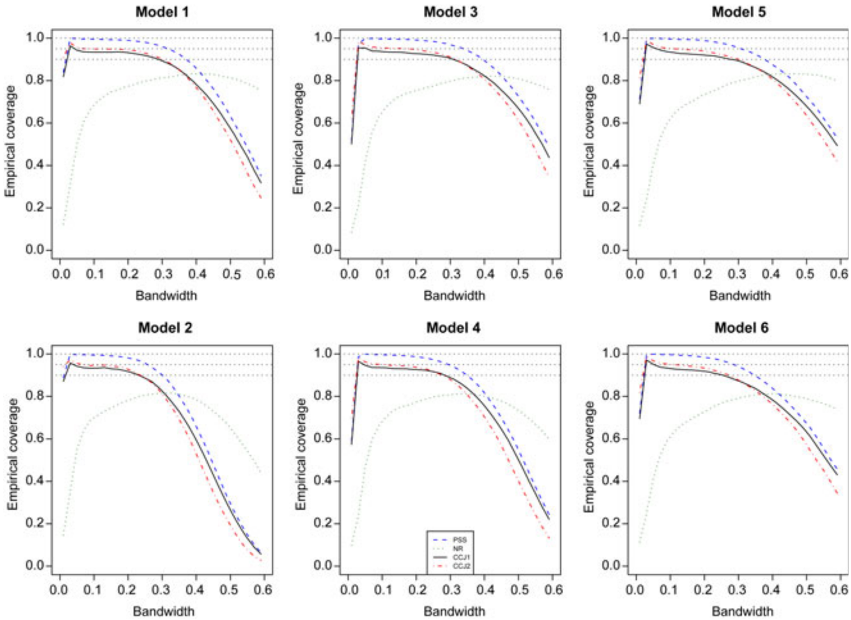

FIGURE 1. Coverage rates for 95% confidence intervals; $P = 2$ and $n = 100$

Figure 1 reports the simulation results when using a second-order Gaussian product kernel ($P = 2$) and $n = 100$. With this choice of kernel the assumptions underlying the results of PSS and Nishiyama and Robinsons (2000) are violated, but we include them in Figure 1 (and in Figure 2 below) to show the effect of the (nonvanishing) bias on empirical coverage in small samples for both procedures under a (too-low) kernel order $P = 2$. Figure 1 shows that for a range of (small) bandwidths and in all models, CCJ1 and CCJ2 exhibit approximately correct empirical coverage, although CCJ1 tends to deliver a slightly liberal inference procedure for this particular sample size and choice of kernel. Nonetheless, the results are encouraging in the sense that the coverage rates of our confidence intervals are close to the nominal coverage rate for a range of (small) bandwidths in a case where technically there are no alternative procedures to be used. Moreover, even though CCJ1 has the potential drawback of failing to deliver a positive-definite matrix $\hat{V}_n$, we note that in this case at most 39 replications (out of 10,000) for each bandwidth had this problem.

One natural explanation for the observed difference between nominal and empirical coverage is that the sample size is too small for our asymptotic results to provide a good approximation. Thus, in Figure 2 we report simulation results when using the same second-order Gaussian product kernel but with a sample size of $n = 400$. These coverage rates improved considerably when compared with those in Figure 1. In particular, now we obtain close-to-correct empirical

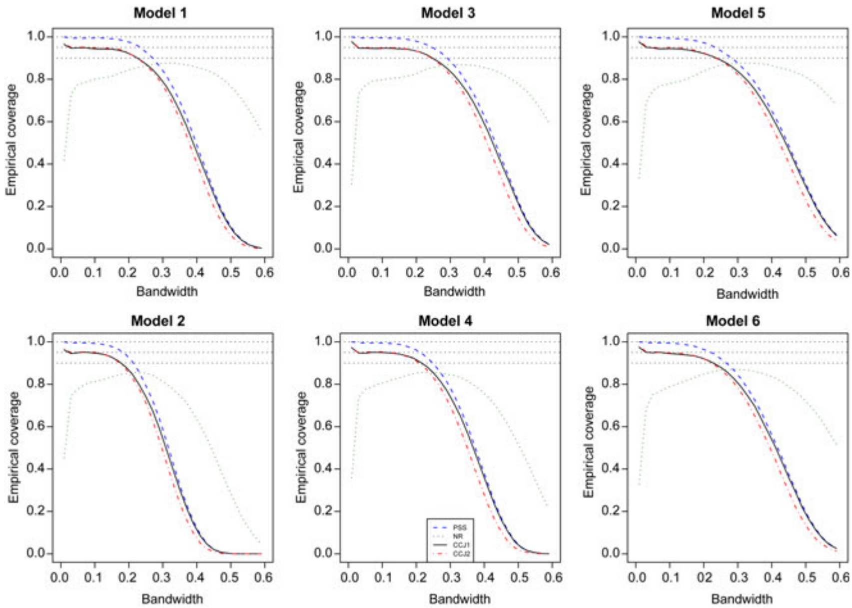

FIGURE 2. Coverage rates for 95% confidence intervals; $P = 2$ and $n = 400$

coverage for a range of (small) bandwidths as the theory predicts. The range of bandwidths for which our procedures work best varies with each model, although in general we see that the lower bound is nearly always the same (i.e., $h > 0.01$). It is interesting to note that while in Figure 1 the smallest bandwidth considered was in fact "too small" (in the sense that the procedure broke down), in this case even this very small bandwidth generally exhibits reasonable properties in terms of empirical coverage. Furthermore, in this case we obtained a positive-definite matrix $\hat{V}_n$ in all replications. These results are very encouraging and suggest that our procedures work well even for a modest sample size of $n = 400$. The last result, coupled with our choice of a commonly used (Gaussian) kernel, suggests that approximately correct, robust confidence bounds may be constructed using our procedure in a relatively straightforward way.

Next, we turn to a (technically) valid comparison between our procedures and those suggested by PSS and NR. Figure 3 reports the simulation results when using a fourth-order kernel ($P = 4$) and a sample size of $n = 100$. When compared to Figure 1, these procedures appear to work better (note the difference in the range of bandwidths plotted in Figures 1 and 2 relative to Figures 3 and 4). The range of bandwidths for which our procedures deliver approximately correct empirical coverage has been extended. This suggests that the use of higher-order kernels provides more robust results. It is interesting to note that PSS appears to have only one bandwidth choice that would provide correct coverage, while

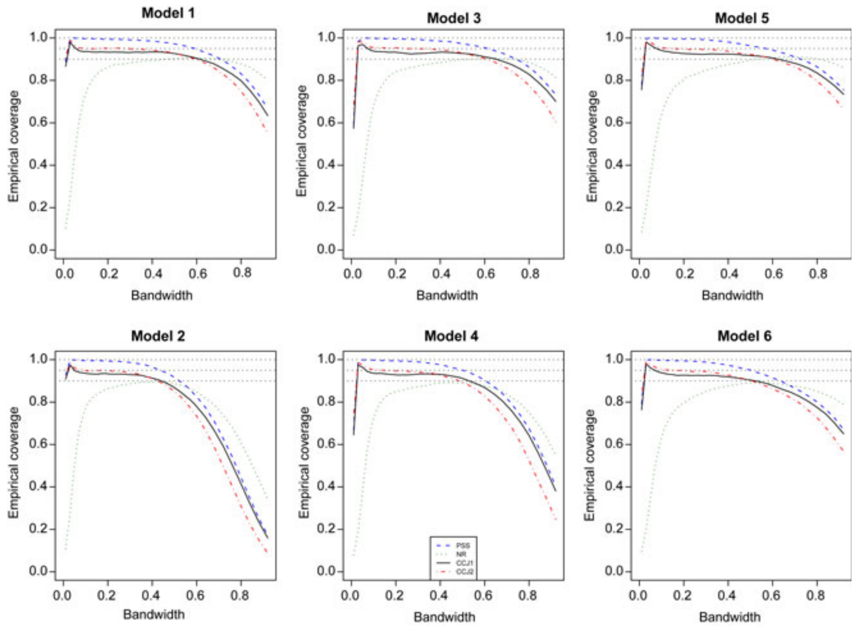

FIGURE 3. Coverage rates for 95% confidence intervals; $P = 4$ and $n = 100$

NR is considerably liberal for all bandwidth choices and models considered. Our confidence intervals are still slightly liberal in this case, although less so than when using a second-order kernel.

Finally, Figure 4 reports the simulation results for the same choice of (higher-order) kernel as in Figure 3 but with a sample size of $n = 400$. As in the case of Figure 2, this sample size appears to be sufficient to deliver close-to-correct coverage over a range of bandwidths for our procedure. In this case as well, the range of bandwidth choices for which CCJ1 and CCJ2 work well has been extended. PSS exhibits very similar behavior to that in Figure 3, while the results for NR suggest that this sample size and kernel choice is insufficient to achieve correct coverage.

The Monte Carlo evidence presented in Figures 1 to 4 suggests that our procedures may be preferred to both PSS and NR, since they justify the use of a second-order kernel while providing approximately valid inference for an array of (sufficiently small) bandwidth choices.[15] Bandwidth selection methods have been developed for density-weighted averaged derivatives by Powell and Stoker (1996) and Nishiyama and Robinsons (2000, 2005).[16] Unfortunately, we found that the population analogue of these three alternative methods did not provide bandwidth choices compatible with the range of bandwidths that were appropriate for our procedure.[17] For example, in the case of Model 5, with $P = 4$ (higher-order kernel) and a sample size of $n = 100$, the population bandwidth

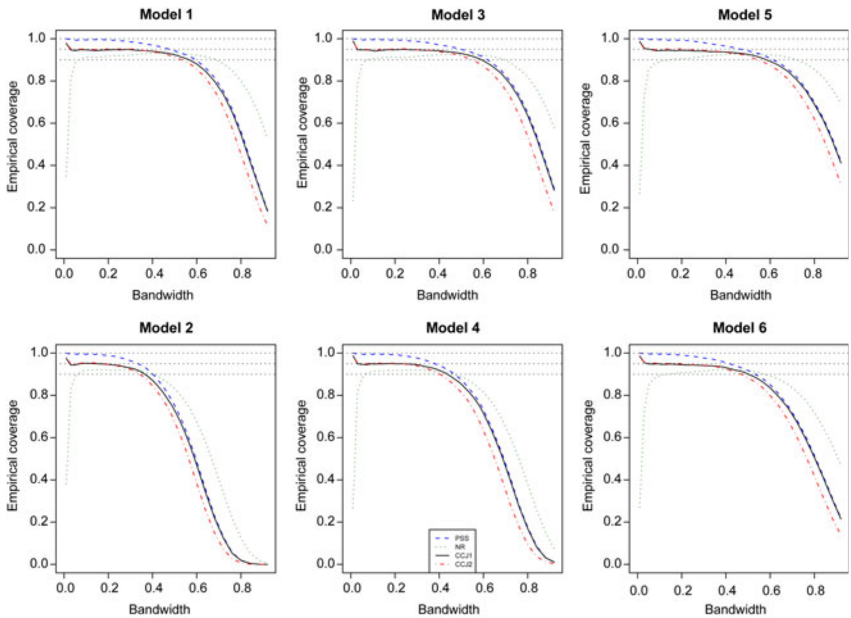

FIGURE 4. Coverage rates for 95% confidence intervals; $P = 4$ and $n = 400$

values are 0.61, 0.51, and 0.65, for the rule-of-thumb formulas in Powell and Stoker (1996) and Nishiyama and Robinsons (2000, 2005), respectively. (For $n = 400$, the corresponding population bandwidth values are 0.49, 0.40, and 0.52.) In all cases, these choices of bandwidths appear to be too high for us to recommend them to be used with our procedures. On the other hand, the bandwidth selection procedures developed in Cattaneo, Crump, and Jansson (2010) were found in that paper to perform very well and we would recommend those to be used.

4. CONCLUSION

This paper has proposed (apparently) novel standard error formulas for the density-weighted average derivative estimator of PSS. Asymptotic validity of the standard errors developed in this paper does not require the use of higher-order kernels, and the standard errors are robust in the sense that they accommodate (but do not require) bandwidths that are smaller than those for which conventional standard errors are valid. Moreover, the results of a Monte Carlo experiment suggest that the finite sample coverage rates of confidence intervals constructed using the standard errors developed in this paper coincide (approximately) with the nominal coverage rates across a nontrivial range of bandwidths, a property not enjoyed by existing procedures.

NOTES

1. A lucid discussion, with precise statements of the conditions on the kernel and the bandwidth, can be found in Section 3 of PSS.

2. An extreme case is the one where the dimension of the explanatory variable exceeds unity and a nonnegative kernel is employed. In that case, the lower and upper bounds on the bandwidth are mutually incompatible.

3. As briefly discussed below, violation of the assumption on the bandwidth that implies asymptotic linearity also has implications for the efficiency properties of PSS's estimator.

4. In turn, the approach of Kiefer, Vogelsang, and Bunzel (2000) and Kiefer and Vogelsang (2002a, 2002b, 2005) can be traced back to Neave (1970).

5. The parameter θ is of interest partly because it is proportional to the vector of coefficients in an index model; that is, θ is proportional to β if $g(x) = G(x'\beta)$ for some function $G(\cdot)$ and some parameter β (e.g., Stoker, 1986; PSS).

6. On the other hand, the assumption $nh_n^{2Q} \to 0$ is not minimal: If K is a twicing kernel, then $nh_n^8 \to 0$ and $nh_n^{d+2} \to \infty$ can suffice even if $Q = 2$ (e.g., Newey et al., 2004).

7. Similarly, the amount of smoothness (indexed by Q) on the part of the density f of the covariates that is required by Theorem 1 is relatively mild.

8. If $h_n \sim n^{-\alpha}$ for some $\alpha \in (\min[2/(d+6), 1/4], 2/d)$, then the assumptions of Theorem 1 hold.

9. An analogous result was obtained by Jammalamadaka and Janson (1986, Thm. 2.1) under a boundedness condition that is violated here.

10. The assumptions of Nishiyama and Robinsons (2000, Thm. 1) include a Cramér condition on $L(z_i)$ and the condition $nh_n^{d+2}/(\log n)^9 \to \infty$, but are otherwise very similar to the assumptions entertained here.

11. In other words, the "variance" term (involving κ_2) does not appear in the Edgeworth expansion of the distribution of $\lambda'\left(\hat{\theta}_n - \theta\right)/\sqrt{\lambda'\left(n^{-1}\Sigma + 2n^{-2}h_n^{-(d+2)}\Delta\right)\lambda}$, as can be seen by inspecting the proof of Nishiyama and Robinsons (2000, Thm. 1), noting that their κ_2 is $\lambda'\Delta\lambda/\lambda'\Sigma\lambda$ in our notation.

12. Being a necessary condition for asymptotic efficiency, asymptotic linearity is an important condition for the results of Newey to hold.

13. We also explored other distributional assumptions for x_{1i}, and in all cases the qualitative results were the same as those reported here.

14. Note that PSS actually used a normalized chi-square random variable with 3 degrees of freedom rather than 4. We changed the distributional assumption to avoid violating Assumption 1(c).

15. Being proportional to $h_n^{-(d+2)}\hat{\Delta}_n$, the correction term in $\hat{V}_n$ depends explicitly on both the bandwidth h_n and the dimension d of the regressor. As a result, it is conceivable that our procedure enjoys the additional robustness property of suffering less from the "curse of dimensionality" than does the procedure of PSS. Preliminary Monte Carlo results (not reported here) are consistent with this conjecture.

16. Because we consider bandwidth sequences corresponding to larger–than-usual undersmoothing, the use of bandwidth selection methods that do not deliver undersmoothing (e.g., cross-validation) cannot be justified using our theory.

17. In particular, using 10^6 replications, we numerically approximated the population higher-order bias and variance terms needed to compute the rules of thumb considered.

REFERENCES

Aradillas-Lopez, A., B.E. Honore, & J.L. Powell (2007) Pairwise difference estimation with nonparametric control variables. *International Economic Review* 48, 1119–1158.

Cattaneo, M.D., R.K. Crump, & M. Jansson (2010) Robust data-driven inference for density-weighted average derivatives. *Journal of the American Statistical Association* 105, 1070–1083.

de Jong, P. (1987) A central limit theorem for generalized quadratic forms. *Probability Theory and Related Fields* 75, 261–277.

Eddy, W.F. (1980) Optimum kernel estimators of the mode. *Annals of Statistics* 8, 870–882.

Eubank, R.L. & S. Wang (1999) A central limit theorem for the sum of generalized linear and quadratic forms. *Statistics* 33, 85–91.

Fan, Y. & O. Linton (2003) Some higher-order theory for a consistent non-parametric model specification test. *Journal of Statistical Planning and Inference* 109, 125–154.

Hoeffding, W. (1948) A class of statistics with asymptotically normal distribution. *Annals of Mathematical Statistics* 19, 293–325.

Hristache, M., A. Juditsky, & V. Spokoiny (2001) Direct estimation of the index coefficient in a single-index model. *Annals of Statistics* 29, 595–623.

Ichimura, H. & O. Linton (2005) Asymptotic expansions for some semi-parametric program evaluation estimators. In D.W.K. Andrews & J.H. Stock (eds.), *Identification and Inference in Econometric Models: Essays in Honor of Thomas J. Rothenberg*, pp. 149–170. Cambridge University Press.

Jammalamadaka, S.R. & S. Janson (1986) Limit theorems for a triangular scheme of U-statistics with applications to inter-point distances. *Annals of Probability* 14, 1347–1358.

Jing, B.-Y. & Q. Wang (2003) Edgeworth expansion for U-statistics under miminal conditions. *Annals of Statistics* 31, 1376–1391.

Kiefer, N.M. & T.J. Vogelsang (2002a) Heteroskedasticity-autocorrelation robust standard errors using the Bartlett kernel without truncation. *Econometrica* 70, 2093–2095.

Kiefer, N.M. & T.J. Vogelsang (2002b) Heteroskedasticity-Autocorrelation robust testing using bandwidth equal to sample size. *Econometric Theory* 18, 1350–1366.

Kiefer, N.M. & T.J. Vogelsang (2005) A new asymptotic theory for heteroskedasticity-autocorrelation robust tests. *Econometric Theory* 21, 1130–1164.

Kiefer, N.M., T.J. Vogelsang, & H. Bunzel (2000) Simple robust testing of regression hypotheses. *Econometrica* 68, 695–714.

Neave, H.R. (1970) An improved formula for the asymptotic variance of spectrum estimates. *Annals of Mathematical Statistics* 41, 70–77.

Newey, W.K. (1994) The asymptotic variance of semiparametric estimators. *Econometrica* 62, 1349–1382.

Newey, W.K., F. Hsieh, & J.M. Robins (2004) Twicing kernels and a small bias property of semiparametric estimators. *Econometrica* 72, 947–962.

Newey, W.K. & D. McFadden (1994) Large sample estimation and hypothesis testing. In R.F. Engle, & D.L. McFadden (eds.), *Handbook of Econometrics*, vol. 4, pp. 2111–2245. North Holland.

Newey, W.K. & T.M. Stoker (1993) Efficiency of weighted average derivative estimators and index models. *Econometrica* 61, 1199–1223.

Nishiyama, Y. & P.M. Robinson (2000) Edgeworth expansions for semiparametric averaged derivatives. *Econometrica* 68, 931–979.

Nishiyama, Y. & P.M. Robinson (2001) Studentization in edgeworth expansions for estimates of semiparametric index models. In C. Hsiao, K. Morimune, & J.L. Powell (eds.), *Nonlinear Statistical Modeling: Essays in Honor of Takeshi Amemiya*, pp. 197–240. Cambridge University Press.

Nishiyama, Y. & P.M. Robinson (2005) The bootstrap and the Edgeworth correction for semiparametric averaged derivatives. *Econometrica* 73, 903–948.

Parzen, E. (1962) On estimation of a probability density function and mode. *Annals of Mathematical Statistics* 33, 1065–1076.

Powell, J.L., J.H. Stock, & T.M. Stoker (1989) Semiparametric estimation of index coefficients. *Econometrica* 57, 1403–1430.

Powell, J.L. & T.M. Stoker (1996) Optimal bandwidth choice for density-weighted averages. *Journal of Econometrics* 75, 291–316.

Robinson, P.M. (1988) Root-N-Consistent semiparametric regression. *Econometrica* 56, 931–954.

Robinson, P.M. (1995) The normal approximation for semiparametric averaged derivatives. *Econometrica* 63, 667–680.

Severini, T.A. & G. Tripathi (2001) A simplified approach to computing efficiency bounds in semiparametric models. *Journal of Econometrics* 102, 23–66.
Stoker, T.M. (1986) Consistent estimation of scaled coefficients. *Econometrica* 54, 1461–1481.

APPENDIX: Proofs

Proof of Theorem 1. Suppose Assumptions 1 and 2 hold. If $h_n \to 0$, then it follows from (Robinson, 1995, Lem. 1) that $\theta_n = \theta + O\left(h_n^{\min(P,Q)}\right)$. As a consequence,

$$\mathbb{V}\left(\hat{\theta}_n\right)^{-1/2}(\theta_n - \theta) \to 0$$

if $\min\left(nh_n^{d+2}, 1\right) nh_n^{2\min(P,Q)} \to 0$ and if (5) holds. To complete the proof, it therefore suffices to show that if $h_n \to 0$ and $n^2 h_n^d \to \infty$, then (4) and (5) hold.

Because

$$\mathbb{V}\left(\hat{\theta}_n\right) = n^{-1}\mathbb{V}\left[L(z_i;h_n)\right] + \binom{n}{2}^{-1}\mathbb{V}\left[W\left(z_i,z_j;h_n\right)\right] \qquad (i<j),$$

the validity of (5) follows from (8) and (9). In turn, (8) holds provided

$$\lim_{h\to 0}\mathbb{E}\left(\|L(z_i;h) - L(z_i)\|^2\right) = 0. \tag{A.1}$$

Now, (A.1) and (9) are variants of Nishiyama and Robinsons (2000, Lem. 3) and Nishiyama and Robinsons (2000, Lem. 12), respectively, and can be shown in exactly the same way.

A further implication of (A.1) is that

$$\sqrt{n}\bar{L}_n = \frac{1}{\sqrt{n}}\sum_{i=1}^{n} L(z_i) + o_p(1).$$

Therefore, (4) holds if it can be shown that

$$\sqrt{n}\bar{l}_n + \sqrt{\binom{n}{2} h_n^{d+2}}\,\bar{w}_n \to_d \mathcal{N}\left(0, \sigma^2 + \delta^2\right) \tag{A.2}$$

for any vectors $\lambda_L \in \mathbb{R}^d$ and $\lambda_W \in \mathbb{R}^d$, where

$$\bar{l}_n = \frac{1}{n}\sum_{i=1}^{n} l(z_i), \qquad l(z_i) = \lambda_L' L(z_i), \qquad \sigma^2 = \lambda_L' \Sigma \lambda_L,$$

$$\bar{w}_n = \binom{n}{2}^{-1}\sum_{i=1}^{n-1}\sum_{j=i+1}^{n} w_n\left(z_i,z_j\right), \qquad w_n\left(z_i,z_j\right) = \lambda_W' W\left(z_i,z_j;h_n\right),$$

$$\delta^2 = \lambda_W' \Delta \lambda_W.$$

Assuming without loss of generality that λ_L and λ_W are both nonzero, we establish (A.2) by employing the theorem of Eubank and Wang (1999). In our notation, conditions (1.3)–(1.6) of Eubank and Wang are

$$h_n^{d+2}\binom{n}{2}^{-1}\max_{1\le j\le n}\sum_{i=1}^{n}\mathbb{E}\left[w_n\left(z_i,z_j\right)^2\right]\to 0, \tag{A.3}$$

$$\left[\binom{n}{2}h_n^{d+2}\right]^2\mathbb{E}\left[\bar{w}_n^4\right]\to 3\delta^4, \tag{A.4}$$

$$n^{-2}\sum_{i=1}^{n}\mathbb{E}\left[l\left(z_i\right)^4\right]\to 0, \tag{A.5}$$

$$\binom{n}{2}^{-1}n^{-1}h_n^{d+2}\mathbb{E}\left[\left(\sum_{j=2}^{n}\sum_{i=1}^{j-1}\mathbb{E}\left[w_n\left(z_i,z_j\right)l\left(z_j\right)|z_1,\ldots z_{j-1}\right]\right)^2\right]\to 0. \tag{A.6}$$

Because $z_i\sim iid$, (A.3) is equivalent to

$$n^{-1}h_n^{d+2}\mathbb{E}\left[w_n\left(z_i,z_j\right)^2\right]\to 0 \qquad (i<j),$$

which is satisfied because (9) holds.

Similarly, (A.5) is equivalent to

$$n^{-1}\mathbb{E}\left[l\left(z_i\right)^4\right]\to 0,$$

which holds because $\mathbb{E}\left[l\left(z_i\right)^4\right]<\infty$ under Assumption 1.

By de Jong (1987, Prop. 3.1), condition (A.4) is satisfied if

$$n^{-2}h_n^{2d+4}\mathbb{E}\left[w_n\left(z_i,z_j\right)^4\right]\to 0 \qquad (i<j), \tag{A.7}$$

$$n^{-1}h_n^{2d+4}\mathbb{E}\left[w_n\left(z_i,z_j\right)^2w_n\left(z_i,z_k\right)^2\right]\to 0 \qquad (i<j<k), \tag{A.8}$$

$$h_n^{2d+4}\mathbb{E}\left[w_n\left(z_i,z_j\right)w_n\left(z_i,z_k\right)w_n\left(z_j,z_m\right)w_n\left(z_k,z_m\right)\right]\to 0 \qquad (i<j<k<m), \tag{A.9}$$

$$h_n^{d+2}\mathbb{E}\left[w_n\left(z_i,z_j\right)^2\right]\to\delta^2 \qquad (i<j). \tag{A.10}$$

Now, Robinson (1995, Lem. 4) implies that $\mathbb{E}\left[w_n\left(z_i,z_j\right)^4\right]=O\left(h_n^{-3d-4}\right)$, so (A.7) holds because $n^2h_n^d\to\infty$. Also, the fact that $z_i\sim iid$ implies that

$$\begin{aligned}\mathbb{E}\left[w_n\left(z_i,z_j\right)^2w_n\left(z_i,z_k\right)^2|z_i\right]&=\mathbb{E}\left[w_n\left(z_i,z_j\right)^2|z_i\right]\mathbb{E}\left[w_n\left(z_i,z_k\right)^2|z_i\right]\\&=\mathbb{E}\left[w_n\left(z_i,z_j\right)^2|z_i\right]^2 \qquad (i<j<k),\end{aligned}$$

so (A.8) holds because

$$\mathbb{E}\left[w_n\left(z_i,z_j\right)^2w_n\left(z_i,z_k\right)^2\right]=\mathbb{E}\left(\mathbb{E}\left[w_n\left(z_i,z_j\right)^2|z_i\right]^2\right)=O\left(h_n^{-2d-4}\right) \qquad (i<j<k),$$

where the first equality uses the law of iterated expectations and the last equality uses Robinson (1995, Lem. 5). Similarly,

$$\begin{aligned}
&\mathbb{E}\left[w_n\left(z_i,z_j\right)w_n\left(z_i,z_k\right)w_n\left(z_j,z_m\right)w_n\left(z_k,z_m\right)|z_j,z_k\right] \\
&\quad= \mathbb{E}\left[w_n\left(z_i,z_j\right)w_n\left(z_i,z_k\right)|z_j,z_k\right]\mathbb{E}\left[w_n\left(z_j,z_m\right)w_n\left(z_k,z_m\right)|z_j,z_k\right] \\
&\quad= \mathbb{E}\left[w_n\left(z_i,z_j\right)w_n\left(z_i,z_k\right)|z_j,z_k\right]^2 \qquad (i<j<k<m),
\end{aligned}$$

so (A.9) follows from the law of iterated expectations and the fact that

$$\mathbb{E}\left(\mathbb{E}\left[w_n\left(z_i,z_j\right)w_n\left(z_i,z_k\right)|z_j,z_k\right]^2\right) = O\left(h_n^{-d-4}\right) \qquad (i<j<k)$$

under our assumptions, the latter being a variant of Nishiyama and Robinsons (2000, Lem. 6). Finally, (A.10) is a consequence of (9).

Condition (A.6) is equivalent to

$$h_n^{d+2}\mathbb{V}\left(\mathbb{E}\left[w_n\left(z_i,z_j\right)l\left(z_j\right)|z_i\right]\right)\to 0 \qquad (i<j).$$

Using the relation

$$\mathbb{V}\left(\mathbb{E}\left[w_n\left(z_i,z_j\right)l\left(z_j\right)|z_i\right]\right) = \mathbb{V}\left(\mathbb{E}\left[\lambda_W' U\left(z_i,z_j;h_n\right)l\left(z_j\right)|z_i\right]\right) \qquad (i<j),$$

change of variables, integration by parts, and simple bounding arguments, it can be shown that if the assumptions of Theorem 1 hold, then

$$\mathbb{V}\left(\mathbb{E}\left[w_n\left(z_i,z_j\right)l\left(z_j\right)|z_i\right]\right) = O(1) \qquad (i<j),$$

implying in particular that (A.6) is satisfied. ■

Proof of Theorem 2. Suppose the assumptions of Theorem 1 hold and suppose $H_n \to 0$ and $n^2 H_n^d \to \infty$. It suffices to show that (10) and (11) hold.

To establish (10), it suffices to show that

$$\hat{\Delta}_n = \hat{\Delta}_{2,n} + o_p(1) = \hat{\Delta}_{3,n} + o_p(1) = \Delta + o_p(1). \tag{A.11}$$

The last equality in (A.11) holds because $\Delta = \lim_{h\to 0} h^{d+2}\mathbb{E}\left[U\left(z_i,z_j;h\right)U\left(z_i,z_j;h\right)'\right]$ $(i<j)$ and because it follows from straightforward moment calculations (utilizing Robinson (1995, App. B) and Nishiyama and Robinsons (2000, App. C)) that

$$\mathbb{E}\left\|\hat{\Delta}_{3,n} - H_n^{d+2}\mathbb{E}\left[U\left(z_i,z_j;H_n\right)U\left(z_i,z_j;H_n\right)'\right]\right\|^2 = O\left(n^{-1}+n^{-2}H_n^{-d}\right) \qquad (i<j).$$

(Specifically, letting $\lambda \in \mathbb{R}^d$ be arbitrary, defining $u\left(z_i,z_j;H_n\right) = \lambda' U\left(z_i,z_j;H_n\right)$, and using the Hoeffding decomposition, we have

$$\begin{aligned}
\mathbb{V}\left(\lambda'\hat{\Delta}_{3,n}\lambda\right) &= H_n^{2d+4}\mathbb{V}\left[\binom{n}{2}^{-1}\sum_{i=1}^{n-1}\sum_{j=i+1}^{n}u\left(z_i,z_j;H_n\right)^2\right] \\
&= H_n^{2d+4}O\left[n^{-1}\mathbb{V}\left(\mathbb{E}\left[u\left(z_i,z_j;H_n\right)^2|z_i\right]\right)+n^{-2}\mathbb{E}\left[u\left(z_i,z_j;H_n\right)^4\right]\right] \\
&= O\left(n^{-1}+n^{-2}H_n^{-d}\right),
\end{aligned}$$

where the last equality uses Nishiyama and Robinsons (2000, Lem. 4) and Robinson (1995, Lem. 4).

Next, the penultimate equality in (A.11) holds because

$$\hat{\Delta}_{2,n} - \hat{\Delta}_{3,n} = -H_n^{d+2}\tilde{\theta}_n\tilde{\theta}_n' = o_p(1),$$

where the last equality uses $\tilde{\theta}_n = O_p\left(1 + n^{-1/2} + n^{-1}H_n^{-(d+2)/2}\right)$. Finally,

$$\hat{\Delta}_{1,n} - \hat{\Delta}_{2,n} = \hat{\Delta}_{1,n}^{(2)} + \hat{\Delta}_{1,n}^{(3)},$$

where

$$\hat{\Delta}_{1,n}^{(2)} = \frac{1}{4}\binom{n}{2}^{-1} H_n^{d+2}\sum_{i=1}^{n-1}\sum_{j=i+1}^{n}\left[\hat{L}_{n,i} + \hat{L}_{n,j}\right]\left[\hat{L}_{n,i} + \hat{L}_{n,j}\right]',$$

$$\hat{\Delta}_{1,n}^{(3)} = \frac{1}{2}\binom{n}{2}^{-1} H_n^{d+2}\sum_{i=1}^{n-1}\sum_{j=i+1}^{n}\left[U\left(z_i, z_j; H_n\right) - \tilde{\theta}_n\right]\left[\hat{L}_{n,i} + \hat{L}_{n,j}\right]'$$
$$+\frac{1}{2}\binom{n}{2}^{-1} H_n^{d+2}\sum_{i=1}^{n-1}\sum_{j=i+1}^{n}\left[\hat{L}_{n,i} + \hat{L}_{n,j}\right]\left[U\left(z_i, z_j; H_n\right) - \tilde{\theta}_n\right]'.$$

Using the fact that

$$\hat{\Sigma}_n = n^{-1}\sum_{i=1}^{n}\hat{L}_{n,i}\hat{L}_{n,i}' = O_p\left(1 + n^{-1}H_n^{-(d+2)}\right),$$

it is easy to show that $\hat{\Delta}_{1,n}^{(2)} = O_p\left(H_n^{d+2} + n^{-1}\right) = o_p(1)$. Also, because $\hat{\Delta}_{2,n} = O_p(1)$ and $\hat{\Delta}_{1,n}^{(2)} = o_p(1)$, it follows from the Cauchy-Schwarz inequality that $\hat{\Delta}_{1,n}^{(3)} = o_p(1)$. Therefore, $\hat{\Delta}_{1,n} - \hat{\Delta}_{2,n} = o_p(1)$ and the validity of the first equality in (A.11) has been established.

Next, letting

$$\hat{\mu}_{n,i} = (n-1)^{-1}\sum_{\substack{j=1\\ j\neq i}}^{n} U\left(z_i, z_j; H_n\right), \qquad \mu\left(z_i; h\right) = \mathbb{E}\left(U\left(z_i, z_j; h\right)|z_i\right),$$

and expanding $\hat{L}_{n,i}$ as

$$\hat{L}_{n,i} = 2\left[\hat{\mu}_{n,i} - \mu\left(z_i; H_n\right) + \mu\left(z_i; H_n\right) - \theta\left(H_n\right) + \theta\left(H_n\right) - \tilde{\theta}_n\right],$$

we arrive at the expansion of $\hat{\Sigma}_n$,

$$\hat{\Sigma}_n = n^{-1}\sum_{i=1}^{n}\hat{L}_{n,i}\hat{L}_{n,i}' = \sum_{j=1}^{6}\hat{\Sigma}_n^{(j)},$$

where

$$\hat{\Sigma}_n^{(1)} = 4n^{-1} \sum_{i=1}^{n} \left[\hat{\mu}_{n,i} - \mu(z_i; H_n)\right] \left[\hat{\mu}_{n,i} - \mu(z_i; H_n)\right]',$$

$$\hat{\Sigma}_n^{(2)} = 4n^{-1} \sum_{i=1}^{n} \left[\mu(z_i; H_n) - \theta(H_n)\right] \left[\mu(z_i; H_n) - \theta(H_n)\right]',$$

$$\hat{\Sigma}_n^{(3)} = 4n^{-1} \sum_{i=1}^{n} \left[\theta(H_n) - \tilde{\theta}_n\right] \left[\theta(H_n) - \tilde{\theta}_n\right]',$$

$$\begin{aligned}\hat{\Sigma}_n^{(4)} &= 4n^{-1} \sum_{i=1}^{n} \left[\hat{\mu}_{n,i} - \mu(z_i; H_n)\right] \left[\mu(z_i; H_n) - \theta(H_n)\right]' \\ &\quad + 4n^{-1} \sum_{i=1}^{n} \left[\mu(z_i; H_n) - \theta(H_n)\right] \left[\hat{\mu}_{n,i} - \mu(z_i; H_n)\right]',\end{aligned}$$

$$\begin{aligned}\hat{\Sigma}_n^{(5)} &= 4n^{-1} \sum_{i=1}^{n} \left[\hat{\mu}_{n,i} - \mu(z_i; H_n)\right] \left[\theta(H_n) - \tilde{\theta}_n\right]' \\ &\quad + 4n^{-1} \sum_{i=1}^{n} \left[\theta(H_n) - \tilde{\theta}_n\right] \left[\hat{\mu}_{n,i} - \mu(z_i; H_n)\right]',\end{aligned}$$

$$\begin{aligned}\hat{\Sigma}_n^{(6)} &= 4n^{-1} \sum_{i=1}^{n} \left[\mu(z_i; H_n) - \theta(H_n)\right] \left[\theta(H_n) - \tilde{\theta}_n\right]' \\ &\quad + 4n^{-1} \sum_{i=1}^{n} \left[\theta(H_n) - \tilde{\theta}_n\right] \left[\mu(z_i; H_n) - \theta(H_n)\right]'.\end{aligned}$$

To establish (11), it suffices to show that

$$n^{-1} \hat{\Sigma}_n^{(1)} = 2 \binom{n}{2}^{-1} H_n^{-(d+2)} \Delta + o_p\left(n^{-2} H_n^{-(d+2)}\right), \quad \textbf{(A.12)}$$

$$\hat{\Sigma}_n^{(2)} = \Sigma + o_p(1), \quad \textbf{(A.13)}$$

$$\hat{\Sigma}_n^{(j)} = o_p\left(1 + n^{-1} H_n^{-(d+2)}\right) \qquad (j = 3, 4, 5, 6). \quad \textbf{(A.14)}$$

Using the relation

$$U\left(z_i, z_j; H_n\right) - \mu(z_i; H_n) = W\left(z_i, z_j; H_n\right) + \frac{1}{2} L\left(z_j; H_n\right)$$

and straightforward moment calculations (utilizing Robinson (1995, App. B) and Nishiyama and Robinsons (2000, App. C), as in the proof of (A.11)), it can be shown that

$$n^4 H_n^{2(d+2)} \mathbb{E} \left\| n^{-1} \hat{\Sigma}_n^{(1)} - 2 \binom{n}{2}^{-1} H_n^{-(d+2)} \tilde{\Delta}_n \right\|^2 = o(1)$$

and

$$\mathbb{E} \left\| \tilde{\Delta}_n - \mathbb{E}\left(\tilde{\Delta}_n\right) \right\|^2 = o(1),$$

where

$$\tilde{\Delta}_n = H_n^{d+2} \binom{n}{2}^{-1} \sum_{i=1}^{n-1} \sum_{j=i+1}^{n} W\left(z_i, z_j; H_n\right) W\left(z_i, z_j; H_n\right)'.$$

The result (A.12) follows from this and the fact that

$$\tilde{\Delta}_n \to_p \lim_{n\to\infty} \mathbb{E}\left(\tilde{\Delta}_n\right) = \Delta,$$

where the equality uses (9).

Next, (A.13) holds because

$$\hat{\Sigma}_n^{(2)} = n^{-1} \sum_{i=1}^{n} L(z_i; H_n) L(z_i; H_n)' = n^{-1} \sum_{i=1}^{n} L(z_i) L(z_i)' + o_p(1) = \Sigma + o_p(1),$$

where the second equality uses

$$\begin{aligned}\mathbb{E}\left(\left\|\hat{\Sigma}_n^{(2)} - n^{-1} \sum_{i=1}^{n} L(z_i) L(z_i)'\right\|^2\right) &\le \mathbb{E}\left(\left\|L(z_i; H_n) L(z_i; H_n)' - L(z_i) L(z_i)'\right\|^2\right) \\ &= o(1),\end{aligned}$$

the equality being a consequence of (A.1).

The condition (A.14) holds for

$$\hat{\Sigma}_n^{(3)} = 4\left[\theta(H_n) - \tilde{\theta}_n\right]\left[\theta(H_n) - \tilde{\theta}_n\right]'$$

because it follows from (A.2) that

$$\begin{aligned}\tilde{\theta}_n - \theta(H_n) &= n^{-1} \sum_{i=1}^{n} L(z_i; H_n) + \binom{n}{2}^{-1} \sum_{i=1}^{n-1} \sum_{j=i+1}^{n} W\left(z_i, z_j; H_n\right) \\ &= O_p\left(n^{-1/2} + n^{-1} H_n^{-(d+2)/2}\right).\end{aligned}$$

Furthermore, (A.14) holds for $\hat{\Sigma}_n^{(4)}$ because straightforward moment calculations (once again utilizing Robinson (1995, App. B) and Nishiyama and Robinsons (2000, App. C)) can be used to show that

$$\min\left(1, n^2 H_n^{2(d+2)}\right) \mathbb{E}\left(\left\|\hat{\Sigma}_n^{(4)}\right\|^2\right) = o(1).$$

Finally, because (A.12)–(A.13) hold and because (A.14) holds for $\hat{\Sigma}_n^{(3)}$, it follows from the Cauchy-Schwarz inequality that (A.14) holds for $\hat{\Sigma}_n^{(5)}$ and $\hat{\Sigma}_n^{(6)}$. ■