

# Small Bandwidth Asymptotics for Density-Weighted Average Derivatives\*

MATIAS D. CATTANEO

DEPARTMENT OF ECONOMICS, UC BERKELEY

RICHARD K. CRUMP

DEPARTMENT OF ECONOMICS, UC BERKELEY

MICHAEL JANSSON

DEPARTMENT OF ECONOMICS, UC BERKELEY AND *CREATES*

May 8, 2008

**ABSTRACT.** This paper proposes (apparently) novel standard error formulas for the density-weighted average derivative estimator of Powell, Stock, and Stoker (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 microeconomic 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 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 requires that the

---

\*The authors thank Bryan Graham, Jim Powell, Tom Rothenberg, Paul Ruud, and seminar participants at Cornell, Harvard, and Penn State for comments. We thank Jasjeet Sekhon and Rocio Titiunik for providing access to the Calgrid cluster. The third author gratefully acknowledges the research support of *CREATES* (funded by the Danish National Research Foundation).

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).<sup>1</sup>

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,<sup>2</sup> suggesting that in samples of moderate size the inference procedures exhibit a certain “non-robustness” 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 which implies asymptotic linearity is violated, then PSS’s standard errors exhibit an upward bias that renders the associated inference procedures conservative.<sup>3</sup> In contrast, we show that valid (non-conservative) inference can be based on PSS’s estimator provided it is combined with a “robust” standard error which accommodates (but does not require) 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

---

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

<sup>2</sup>An extreme case is the one where the dimension of the explanatory variable exceeds unity and a non-negative kernel is employed. In that case, the lower and upper bounds on the bandwidth are mutually incompatible.

<sup>3</sup>As briefly discussed below, violation of the assumption on the bandwidth which implies asymptotic linearity also has implications for the efficiency properties of PSS’s estimator.

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 which 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 two 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, which would appear to be novel, 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.

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 Robinson (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.

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  $z_i = (y_i, x_i)'$  ( $i = 1, \dots, n$ ) are *i.i.d.* 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  $f(\cdot)$ . 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<sup>4</sup>

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

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

**Assumption 1.** (a)  $\mathbb{E}(y^4) < \infty$ .

(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

$$\sup_{x \in \mathbb{R}^d} [v(x) f(x) + v(x) \|\partial f(x)/\partial x\| + \|\partial v(x)/\partial x\|] < \infty,$$

where  $\|\cdot\|$  is the Euclidean norm and  $v(x) = \mathbb{E}(y^2|x)$ .

(f)  $\lim_{\|x\| \rightarrow \infty} [f(x) + |e(x)|] = 0$ .

---

<sup>4</sup>The parameter  $\theta$  is of interest partly because it is proportional to the vector of coefficients in an index model; that is,  $\theta$  is proportional to  $\beta$  if  $g(x) = G(x'\beta)$  for some function  $G(\cdot)$  and some parameter  $\beta$  (e.g., Stoker (1986) and PSS).

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  $K : \mathbb{R}^d \rightarrow \mathbb{R}$  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  $P \geq 2$ ,

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

and

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

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  $\theta$  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  $nh_n^{2\min(P,Q)} \rightarrow 0$  and  $nh_n^{d+2} \rightarrow \infty$ , then

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

where

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

(Here, and elsewhere in the paper, limits are taken as  $n \rightarrow \infty$  unless otherwise noted.) In the statement of this result, the conditions  $nh_n^{2P} \rightarrow 0$  and  $nh_n^{d+2} \rightarrow \infty$  are minimal in the sense that (2) can fail if one (or both) of the assumptions is (are) relaxed.<sup>5</sup> 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  $h_n \rightarrow 0$  and  $nh_n^{d+2} \rightarrow \infty$  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  $nh_n^{2\min(P,Q)} \rightarrow 0$  and  $nh_n^{d+2} \rightarrow \infty$ , then

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

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  $\min(nh_n^{d+2}, 1) nh_n^{2\min(P,Q)} \rightarrow 0$  and  $n^2h_n^d \rightarrow \infty$ , then (3) is true.*

The conditions of this theorem weaken those of PSS in two respects. First, the condition  $n^2h_n^d \rightarrow \infty$  is considerably weaker than the condition  $nh_n^{d+2} \rightarrow \infty$ . 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

---

<sup>5</sup>On the other hand, the assumption  $nh_n^{2Q} \rightarrow 0$  is not minimal: If  $K$  is a twicing kernel, then  $nh_n^8 \rightarrow 0$  and  $nh_n^{d+2} \rightarrow \infty$  can suffice even if  $Q = 2$  (e.g., Newey, Hsieh, and Robins (2004)).

equivalent to its Hájek projection. Second, due to the presence of the additional term  $\min(nh_n^{d+2}, 1)$  our “bias” condition is weaker than the condition  $nh_n^{2\min(P,Q)} \rightarrow 0$  of PSS. As usual, we need the bias of the estimator to be of smaller order of magnitude than the standard deviation. The term  $\min(nh_n^{d+2}, 1)$  in the “bias” condition reflects the fact (further discussed below) that the rate of convergence of the estimator is slower than  $\sqrt{n}$  when  $nh_n^{d+2} \rightarrow 0$ .

Partly due to the presence of  $\min(nh_n^{d+2}, 1)$  in the “bias” condition, Theorem 1 accommodates smaller values of  $P$  than do the results of PSS.<sup>6</sup> Indeed, for any value of  $d$  there exists a bandwidth sequence  $h_n$  compatible with the assumptions of Theorem 1 even if  $P = 2$ .<sup>7</sup> 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  $\mathbb{V}(\hat{\theta}_n)$  is replaced by a suitable estimator in (3)).

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

$$\hat{\theta}_n = \binom{n}{2}^{-1} \sum_{i=1}^{n-1} \sum_{j=i+1}^n U(z_i, z_j; h_n), \quad U(z_i, z_j; h) = -h^{-(d+1)} \dot{K}\left(\frac{x_i - x_j}{h}\right) (y_i - y_j).$$

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

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

where

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

with

$$\theta(h) = \mathbb{E}[U(z_i, z_j; h)], \quad L(z_i; h) = 2[\mathbb{E}(U(z_i, z_j; h) | z_i) - \theta(h)],$$

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

<sup>6</sup>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.

<sup>7</sup>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.

The projection theorem for variable  $U$ -statistics (e.g., Lemma 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  $\sqrt{n}\bar{W}_n \rightarrow_p 0$ . To handle cases where this projection theorem provides insufficient technical machinery to establish asymptotic normality of  $\hat{\theta}_n$  (because  $\sqrt{n}\bar{W}_n \not\rightarrow_p 0$ ), 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  $h_n \rightarrow 0$  and  $n^2 h_n^d \rightarrow \infty$ , then

$$\left( \begin{array}{c} \sqrt{n}\bar{L}_n \\ \sqrt{\binom{n}{2} h_n^{d+2} \bar{W}_n} \end{array} \right) \rightarrow_d \mathcal{N} \left[ \begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \Sigma & 0 \\ 0 & \Delta \end{pmatrix} \right] \quad (4)$$

where

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

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  $n^2 h_n^d \rightarrow \infty$  on the bandwidth sequence and utilize some technical lemmas due to Robinson (1995) and Nishiyama and Robinson (2000). Because the condition  $n^2 h_n^d \rightarrow \infty$  is considerably weaker than the condition  $n h_n^{d+2} \rightarrow \infty$  needed for the result  $\sqrt{n}\bar{W}_n \rightarrow_p 0$ , we can accommodate a significantly wider range of bandwidths by basing the distribution theory on (4) rather than a result which requires  $\sqrt{n}\bar{W}_n \rightarrow_p 0$ .

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). Whereas higher-order asymptotic results for  $U$ -statistics whose kernels do not vary with the sample size suggest that no asymptotic refinements are achieved by standardizing by the actual variance (e.g., Jing and Wang (2003)), Theorem 1 demonstrates by example that the situation can be very different for a  $U$ -statistic whose kernel does vary with the sample size.<sup>8</sup>

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)

---

<sup>8</sup>An analogous result was obtained by Jammalamadaka and Janson (1986, Theorem 2.1) under a boundedness condition that is violated here.

$nh_n^{d+2} \rightarrow \infty$ , 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 Robinson (2000) suggest that the formulation (3) can be attractive for certain (small) values of the bandwidth. Specifically, if the assumptions of Nishiyama and Robinson (2000, Theorem 1) hold and if  $n^3 h_n^{2(d+2+\min(P,Q))} \rightarrow 0$  and  $nh_n^{2(d+2)} \rightarrow 0$ , then for any nonzero  $\lambda \in \mathbb{R}^d$  the leading term in the Edgeworth expansion of the distribution of  $\lambda'(\hat{\theta}_n - \theta) / \sqrt{n^{-1}\lambda'\Sigma\lambda}$  is a “variance” term that accounts for the variability of  $\bar{W}_n$ .<sup>9</sup> 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  $\hat{\theta}_n$ .<sup>10</sup> It is shown in the proof of Theorem 1 that

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

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

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

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

Because  $\Delta$  and  $\kappa$  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 (1994a, Proposition 1), as  $\hat{\theta}_n$  ceases to be asymptotically linear when the condition  $nh_n^{d+2} \rightarrow \infty$  is dropped.<sup>11</sup>

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

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

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

<sup>10</sup>In other words, the “variance” term does not appear in the Edgeworth expansion of the distribution of  $\lambda'(\hat{\theta}_n - \theta) / \sqrt{\lambda'(n^{-1}\Sigma + 2n^{-2}h_n^{-(d+2)}\Delta)\lambda}$ .

<sup>11</sup>Being a necessary condition for asymptotic efficiency, asymptotic linearity is an important condition for the results of Newey (1994a) to hold.

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

**Remarks.** (i) The asymptotic efficiency of  $\hat{\theta}_n$  is maximized by employing a bandwidth sequence satisfying  $n h_n^{d+2} \rightarrow \infty$ . Indeed, although  $\theta$  is not covered by the results of Newey and Stoker (1993), by proceeding as in the proof of Newey and Stoker (1993, Theorem 3.1) it can be shown that if certain regularity conditions hold, then  $L(\cdot)$  is the pathwise derivative of  $\theta$ . As a result,  $\hat{\theta}_n$  enjoys semiparametric efficiency properties if (and only if)  $n h_n^{d+2} \rightarrow \infty$ .

(ii) If  $n h_n^{d+2}$  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, Section 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 the statement (3) is that it involves the matrix  $\mathbb{V}(\hat{\theta}_n)$ , which is unknown. Replacing  $\mathbb{V}(\hat{\theta}_n)$  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

$$\hat{V}_n^{-1/2} (\hat{\theta}_n - \theta) \rightarrow_d \mathcal{N}(0, I_d). \quad (6)$$

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

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

then (6) holds.

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

$$\Sigma = \lim_{h \rightarrow 0} \mathbb{E} [L(z_i; h) L(z_i; h)'] \quad (8)$$

and

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

Analog estimators of  $\Sigma$  and  $\Delta$  suggested by these characterizations are given by

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

where  $H_n$  is a bandwidth sequence and

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

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

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. Indeed, the conditions on  $H_n$  required for the following result are violated by  $H_n = h_n$  in many of the cases covered by Theorem 1.

**Theorem 2.** *If the assumptions of Theorem 1 hold and if  $H_n \rightarrow 0$  and  $nH_n^{d+2} \rightarrow \infty$ , then (6) holds for*

$$\hat{V}_n = n^{-1} \hat{\Sigma}_n + \binom{n}{2}^{-1} h_n^{-(d+2)} \hat{\Delta}_n. \quad (10)$$

The theorem demonstrates in particular that valid inference on  $\theta$  can be based on  $\hat{\theta}_n$  under mild conditions on the kernel  $K$  and the bandwidth  $h_n$  provided the estimator of dispersion utilizes  $\{U(z_i, z_j; H_n) : 1 \leq i < j \leq n\}$  constructed with a (possibly) different bandwidth  $H_n$ .

Although there are many cases covered by Theorem 1 in which the lower bound on  $H_n$  implied by the condition  $nH_n^{d+2} \rightarrow \infty$  cannot be satisfied by setting  $H_n = h_n$ , it is not inconceivable that (7) can be satisfied by estimators  $\hat{V}_n$  based solely on  $\{U(z_i, z_j; h_n) : 1 \leq i < j \leq n\}$  even if  $nh_n^{d+2} \rightarrow \infty$ . It turns out that if Assumptions 1 and 2 hold and if  $H_n \rightarrow 0$  and  $n^2 H_n^d \rightarrow \infty$ , then

$$n^{-1}\hat{\Sigma}_n = n^{-1}\Sigma + 2 \binom{n}{2}^{-1} H_n^{-(d+2)} \Delta + o_p(n^{-1} + n^{-2}H_n^{-(d+2)}). \quad (11)$$

In addition to implying that  $\hat{\Sigma}_n$  is an inconsistent estimator of  $\Sigma$  when  $nH_n^{d+2} \not\rightarrow \infty$ , the stochastic expansion (11) shows that even if  $H_n = h_n$ , the requirement (7) can be met by an estimator that combines  $\hat{\Sigma}_n$  with a consistent estimator of  $\Delta$ . Indeed, the following result is an immediate consequence of (11) and the fact that if Assumptions 1 and 2 hold and if  $H_n \rightarrow 0$  and  $n^2H_n^d \rightarrow \infty$ , then

$$\hat{\Delta}_n \rightarrow_p \Delta. \quad (12)$$

**Theorem 3.** *If the assumptions of Theorem 1 hold and if  $H_n = h_n$ , then (6) holds for*

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

In view of this theorem, the construction (13) gives a simple recipe for achieving (6) under very mild conditions on (the kernel  $K$  and) the bandwidth  $h_n$ . Moreover, although the requirement (7) reduces to  $n\hat{V}_n = \Sigma + o_p(1)$  (and is met by  $\hat{V}_n = n^{-1}\hat{\Sigma}_n$ ) when  $nh_n^{d+2} \rightarrow \infty$ , the results of Nishiyama and Robinson (2000, 2001) suggest that the construction (13) may enjoy higher-order advantages over the standard construction  $\hat{V}_n = n^{-1}\hat{\Sigma}_n$ . Specifically, if the assumptions of Nishiyama and Robinson (2000, Theorem 3) hold and if  $n^3h_n^{2(d+2+\min(P,Q))} \rightarrow 0$  and  $nh_n^{2(d+2)} \rightarrow 0$ , then (for any nonzero  $\lambda \in \mathbb{R}^d$ ) the leading term in the Edgeworth expansion of the distribution of  $\lambda'(\hat{\theta}_n - \theta) / \sqrt{n^{-1}\lambda'\hat{\Sigma}_n\lambda}$  is a ‘‘variance’’ term that accounts for the fact that  $n^{-1}\hat{\Sigma}_n$  overestimates the variance of  $\hat{\theta}_n$ . This term can be removed by subtracting the term  $2n^{-2}h_n^{-(d+2)}\Delta$  from  $n^{-1}\hat{\Sigma}_n$ , which (up to estimation error introduced by replacing  $\Delta$  with a consistent estimator) is exactly what the construction (13) does.<sup>12</sup> As a result, the construction (13) seems attractive even in those cases where  $nh_n^{d+2} \rightarrow \infty$ .

**Remarks.** (i) When  $H_n = h_n$ ,  $\hat{\Sigma}_n$  is PSS’s estimator of  $\Sigma$ . 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

<sup>12</sup>This qualitative finding also holds (albeit with slightly different rate restrictions on  $h_n$ ) when  $\left| \lambda'(\hat{\theta}_n - \theta) \right| / \sqrt{n^{-1}\lambda'\hat{\Sigma}_n\lambda}$  is considered, as would be appropriate in the context of two-sided hypothesis testing. Specifically, if the assumptions of Nishiyama and Robinson (2005, Theorem 5) hold and if  $n^2h_n^{2\min(P,Q)+d+2} \rightarrow 0$  and  $nh_n^{d+2+\min(P,Q)} \rightarrow 0$ , then the leading term in the Edgeworth expansion can be removed by replacing  $n^{-1}\hat{\Sigma}_n$  with  $n^{-1}\hat{\Sigma}_n - 2n^{-2}h_n^{-(d+2)}\Delta$ .

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

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  $nh_n^{d+2} \not\rightarrow \infty$ .

(ii) The proof of (12) implicitly establishes consistency of two additional estimators of  $\Delta$ , namely

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

and

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

These are also analog estimators because

$$\Delta = \lim_{h \rightarrow 0} h^{d+2} \mathbb{V} [U(z_i, z_j; h)] = \lim_{h \rightarrow 0} h^{d+2} \mathbb{E} [U(z_i, z_j; h) U(z_i, z_j; h)'] \quad (i < j),$$

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

### 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 Robinson (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

$$y_i = \tau(y_i^*), \quad y_i^* = x_i' \beta + \varepsilon_i,$$

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  $x_{2i} \sim \varkappa$ , where  $\varkappa$  is a normalized chi-square random variable with 4 degrees of freedom (i.e.,  $\varkappa = (\chi_4^2 - 4) / \sqrt{8}$ ).<sup>13</sup> 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 I: 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: $\theta_1 = 0.03906$

Table I summarizes the Monte Carlo models, reports the value of the population parameter of interest, and provides the corresponding label of each model considered. In Models 4 and 6 a tidy closed-form expression is unavailable for  $\theta_1$  and we therefore report a numerical approximation instead. Models 1 through 4 were studied by PSS in their simulation study,<sup>14</sup> while Model 5 corresponds to the one employed in the simulation study of Nishiyama and Robinson (2000).

We consider two sample sizes,  $n = 100$  and  $n = 400$ , and for each case we carry out  $S = 1,000$  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 Robinson (2000, pp. 943-944). We also explored other choices of kernel functions, such as a twiced Gaussian kernel (e.g., Newey, Hsieh, and Robins (2004)), but the results were qualitatively similar and therefore we omit them to conserve space.

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

<sup>14</sup>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).

We consider three 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 Robinson (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 Robinson (2000). Finally, the third inference procedure is the one developed in Theorem 3. 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 three inference procedures: PSS’s procedure, Nishiyama and Robinson’s (2000) higher-order corrected procedure, abbreviated “NR” for simplicity, and our procedure introduced in Theorem 3, which is denoted by “CCJ”. 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.

**FIGURES 1-2 ABOUT HERE**

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 Robinson (2000) are violated, but we include them in Figure 1 (and in Figure 2 below) to show the effect of the (non-vanishing) 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, CCJ exhibits approximately correct empirical coverage. Our correction, however, 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 our procedure has the potential drawback of failing to deliver a positive-definite matrix  $\hat{V}_n$ , we note that in this case at most 2 replications (out of 1,000) 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 coverage for a range of (small)

bandwidths as the theory predicts. The range of bandwidths for which our procedure works 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 procedure works 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.

**FIGURES 3-4 ABOUT HERE**

Next, we turn to a (technically) valid comparison between our procedure 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 procedure delivers 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 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 CCJ works well has been extended. PSS exhibits very similar behavior as 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 procedure may be preferred to both PSS and NR, since it justifies the use of a second order kernel while providing approximately valid inference for an array of (sufficiently small) bandwidth choices.<sup>15</sup> However, it is still unclear how to choose a bandwidth

---

<sup>15</sup>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

within that range in applications. One possibility would be to use the available rule-of-thumb bandwidth choices developed for density-weighted averaged derivatives by Powell and Stoker (1996), Nishiyama and Robinson (2000), and Nishiyama and Robinson (2005). 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.<sup>16</sup> For example, in the case of Model 5, with  $P = 4$  (higher-order kernel) and a sample size of  $n = 100$  the population bandwidth values are 0.58, 0.51 and 0.65, for the rule-of-thumb formulas in Powell and Stoker (1996), Nishiyama and Robinson (2000), and Nishiyama and Robinson (2005), respectively. (For  $n = 400$  the corresponding population bandwidth values are 0.46, 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 procedure. Cross-validation may provide a feasible alternative for choosing an appropriate (small) bandwidth in applications. Fully data-driven procedures for selecting bandwidths compatible with our procedure is a topic of future research and beyond the scope of this paper.

#### 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. The latter property is not enjoyed by existing procedures, so it would be very useful to develop “automatic” bandwidth selection methods to accompany the new standard errors.

The theoretical analysis conducted in this paper has been greatly facilitated by the algebraic simplicity of PSS’s estimator and the availability of sophisticated technical results for it. Whereas the higher order asymptotic theory for density-weighted average derivative estimators developed by Nishiyama and Robinson (2000, 2001, 2005) is conjectured in those papers to be difficult to extend to semiparametric estimators that do not enjoy the algebraic simplicity of PSS’s estimator, it seems quite plausible that results analogous to those derived in this paper can be obtained (with a not unreasonable amount of technical effort) also for certain semiparametric estimators that depend on a kernel estimator in a “nonlinear” way (e.g., those considered in Newey

---

this conjecture.

<sup>16</sup>In particular, using  $10^6$  replications, we numerically approximated the population higher-order bias and variance terms needed to compute the rule-of-thumbs considered.

(1994b) or some interesting subset thereof). In future work, we intend to explore the extent to which this conjecture is correct.

## 5. APPENDIX: PROOFS

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

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

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

Because

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

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

$$\lim_{h \rightarrow 0} \mathbb{E}\left(\|L(z_i; h) - L(z_i)\|^2\right) = 0. \quad (14)$$

Now, (14) and (9) are variants of Nishiyama and Robinson (2000, Lemma 3) and Nishiyama and Robinson (2000, Lemma 12), respectively, and can be shown in exactly the same way.

A further implication of (14) 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

$$\sqrt{n}\bar{l}_n + \sqrt{\binom{n}{2} h_n^{d+2} \bar{w}_n} \rightarrow_d \mathcal{N}(0, \sigma^2 + \delta^2) \quad (15)$$

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

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

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

Assuming without loss of generality that  $\lambda_L$  and  $\lambda_W$  are both non-zero, we establish (15) by employing the theorem of Eubank and Wang (1999). In our notation, conditions (1.3) – (1.6) of Eubank and Wang (1999) are

$$h_n^{d+2} \binom{n}{2}^{-1} \max_{1 \leq j \leq n} \sum_{i=1}^n \mathbb{E} [w_n(z_i, z_j)^2] \rightarrow 0, \quad (16)$$

$$\left[ \binom{n}{2} h_n^{d+2} \right]^2 \mathbb{E} [\bar{w}_n^4] \rightarrow 3\delta^4, \quad (17)$$

$$n^{-2} \sum_{i=1}^n \mathbb{E} [l(z_i)^4] \rightarrow 0, \quad (18)$$

$$\binom{n}{2}^{-1} n^{-1} h_n^{d+2} \mathbb{E} \left[ \left( \sum_{j=2}^n \sum_{i=1}^{j-1} \mathbb{E} [w_n(z_i, z_j) l(z_j) | z_1, \dots, z_{j-1}] \right)^2 \right] \rightarrow 0. \quad (19)$$

Because  $z_i \sim i.i.d.$ , (16) is equivalent to

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

which is satisfied because (9) holds.

Similarly, (18) is equivalent to

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

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

By de Jong (1987, Proposition 3.1), condition (17) is satisfied if

$$n^{-2} h_n^{2d+4} \mathbb{E} [w_n(z_i, z_j)^4] \rightarrow 0 \quad (i < j), \quad (20)$$

$$n^{-1} h_n^{2d+4} \mathbb{E} [w_n(z_i, z_j)^2 w_n(z_i, z_k)^2] \rightarrow 0 \quad (i < j < k), \quad (21)$$

$$h_n^{2d+4} \mathbb{E} [w_n(z_i, z_j) w_n(z_i, z_k) w_n(z_j, z_m) w_n(z_k, z_m)] \rightarrow 0 \quad (i < j < k < m), \quad (22)$$

$$h_n^{d+2} \mathbb{E} [w_n(z_i, z_j)^2] \rightarrow \delta^2 \quad (i < j). \quad (23)$$

Now, Robinson (1995, Lemma 4) implies that  $\mathbb{E} [w_n(z_i, z_j)^4] = O(h_n^{-3d-4})$ , so (20) holds because  $n^2 h_n^d \rightarrow \infty$ . Also, the fact that  $z_i \sim i.i.d.$  implies that

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

so (21) holds because

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

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

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

so (22) follows from the law of iterated expectations and the fact that

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

under our assumptions, the latter being a variant of Nishiyama and Robinson (2000, Lemma 6). Finally, (23) is a consequence of (9).

Condition (19) is equivalent to

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

Using the relation

$$\mathbb{V}(\mathbb{E}[w_n(z_i, z_j) l(z_j) | z_i]) = \mathbb{V}(\mathbb{E}[\lambda'_W U(z_i, z_j; h_n) l(z_j) | z_i]) \quad (i < j),$$

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

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

implying in particular that (19) is satisfied.  $\blacksquare$

**Proof of Theorems 2 and 3.** Suppose the assumptions of Theorem 1 hold and suppose  $H_n \rightarrow 0$  and  $n^2 H_n^d \rightarrow \infty$ . It suffices to show that (11) and (12) hold. To do so, let

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

Expanding  $\hat{L}_{n,i}$  as

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

we arrive at the following 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

$$\begin{aligned} \hat{\Sigma}_n^{(1)} &= 4n^{-1} \sum_{i=1}^n [\hat{\mu}_{n,i} - \mu(z_i; H_n)] [\hat{\mu}_{n,i} - \mu(z_i; H_n)]', \\ \hat{\Sigma}_n^{(2)} &= 4n^{-1} \sum_{i=1}^n [\mu(z_i; H_n) - \theta(H_n)] [\mu(z_i; H_n) - \theta(H_n)]', \\ \hat{\Sigma}_n^{(3)} &= 4n^{-1} \sum_{i=1}^n [\theta(H_n) - \tilde{\theta}_n] [\theta(H_n) - \tilde{\theta}_n]', \end{aligned}$$

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

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

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

The establish (11), it suffices to show that

$$n^{-1} \hat{\Sigma}_n^{(1)} = 2 \binom{n}{2}^{-1} H_n^{-(d+2)} \Delta + o_p(n^{-2} H_n^{-(d+2)}), \quad (24)$$

$$\hat{\Sigma}_n^{(2)} = \Sigma + o_p(1), \quad (25)$$

$$\hat{\Sigma}_n^{(j)} = o_p(1 + n^{-1} H_n^{-(d+2)}) \quad (j = 3, 4, 5, 6). \quad (26)$$

Using the relation

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

and straightforward moment calculations (utilizing Nishiyama and Robinson (2000, Appendix C)), 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

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

The result (24) follows from this and the fact that

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

where the equality uses (9).

Next, (25) 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

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

the equality being a consequence of (14).

The condition (26) 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 (15) that

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

Furthermore, (26) holds for  $\hat{\Sigma}_n^{(4)}$  because straightforward moment calculations (utilizing Nishiyama and Robinson (2000, Appendix C)) can be used to show that

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

Finally, because (24) – (25) hold and because (26) holds for  $\hat{\Sigma}_n^{(3)}$ , it follows from the Cauchy-Schwarz inequality that (26) holds for  $\hat{\Sigma}_n^{(5)}$  and  $\hat{\Sigma}_n^{(6)}$ .

To establish (12), it suffices to show that

$$\hat{\Delta}_n = \hat{\Delta}_{2,n} + o_p(1) = \hat{\Delta}_{3,n} + o_p(1) = \Delta + o_p(1). \quad (27)$$

The last equality in (27) holds because it follows from straightforward moment calculations (utilizing Nishiyama and Robinson (2000, Appendix C)) that

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

and because  $\Delta = \lim_{h \rightarrow 0} h^{d+2} \mathbb{E} [U(z_i, z_j; h) U(z_i, z_j; h)']$  ( $i < j$ ). Next, the penultimate equality in (27) 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  $\tilde{\theta}_n = O_p(1 + n^{-1/2} + n^{-1} H_n^{-(d+2)/2})$ . Finally,

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

where

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

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

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  $\hat{\Delta}_{1,n}^{(2)} = O_p(H_n^{d+2} + n^{-1}) = o_p(1)$ . 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 (27) has been established. ■

## REFERENCES

- 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., AND S. WANG (1999): “A Central Limit Theorem for the Sum of Generalized Linear and Quadratic Forms,” *Statistics*, 33, 85–91.
- HOEFFDING, W. (1948): “A Class of Statistics with Asymptotically Normal Distribution,” *Annals of Mathematical Statistics*, 19, 293–325.
- HRISTACHE, M., A. JUDITSKY, AND V. SPOKOINY (2001): “Direct Estimation of the Index Coefficient in a Single-Index Model,” *Annals of Statistics*, 29, 595–623.
- JAMMALAMADAKA, S. R., AND 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., AND Q. WANG (2003): “Edgeworth Expansion for  $U$ -Statistics under Minimal Conditions,” *Annals of Statistics*, 31, 1376–1391.
- KIEFER, N. M., AND T. J. VOGELSANG (2002a): “Heteroskedasticity-Autocorrelation Robust Standard Errors Using the Bartlett Kernel Without Truncation,” *Econometrica*, 70, 2093–2095.
- (2002b): “Heteroskedasticity-Autocorrelation Robust Testing Using Bandwidth Equal to Sample Size,” *Econometric Theory*, 18, 1350–1366.
- (2005): “A New Asymptotic Theory for Heteroskedasticity-Autocorrelation Robust Tests,” *Econometric Theory*, 21, 1130–1164.
- KIEFER, N. M., T. J. VOGELSANG, AND H. BUNZEL (2000): “Simple Robust Testing of Regression Hypotheses,” *Econometrica*, 68, 695–714.
- NEWBY, 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.
- NEWBY, W. K., F. HSIEH, AND J. M. ROBINS (2004): “Twicing Kernels and a Small Bias Property of Semiparametric Estimators,” *Econometrica*, 72, 947–962.

- NEWWEY, W. K., AND T. M. STOKER (1993): "Efficiency of Weighted Average Derivative Estimators and Index Models," *Econometrica*, 61, 1199–1223.
- NISHIYAMA, Y., AND P. M. ROBINSON (2000): "Edgeworth Expansions for Semiparametric Averaged Derivatives," *Econometrica*, 68, 931–979.
- (2001): "Studentization in Edgeworth Expansions for Estimates of Semiparametric Index Models," in *Nonlinear Statistical Modeling: Essays in Honor of Takeshi Amemiya*, ed. by C. Hsiao, K. Morimune, and J. L. Powell. New York: Cambridge University Press, 197–240.
- (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, AND T. M. STOKER (1989): "Semiparametric Estimation of Index Coefficients," *Econometrica*, 57, 1403–1430.
- POWELL, J. L., AND 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.
- (1995): "The Normal Approximation for Semiparametric Averaged Derivatives," *Econometrica*, 63, 667–680.
- STOKER, T. M. (1986): "Consistent Estimation of Scaled Coefficients," *Econometrica*, 54, 1461–1481.

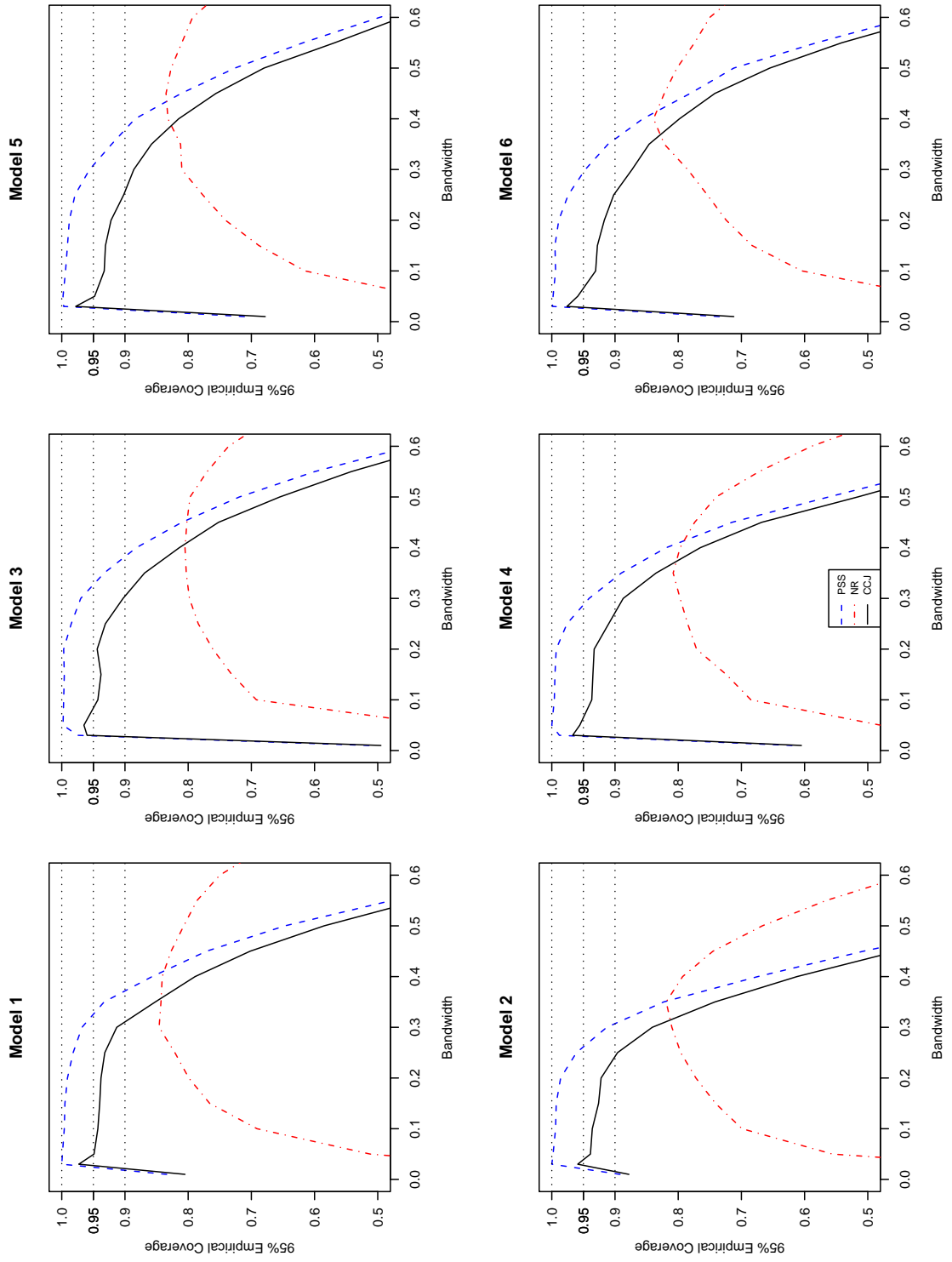


Figure 1: Coverage Rates for 95% Confidence Intervals;  $P = 2$  and  $n = 100$

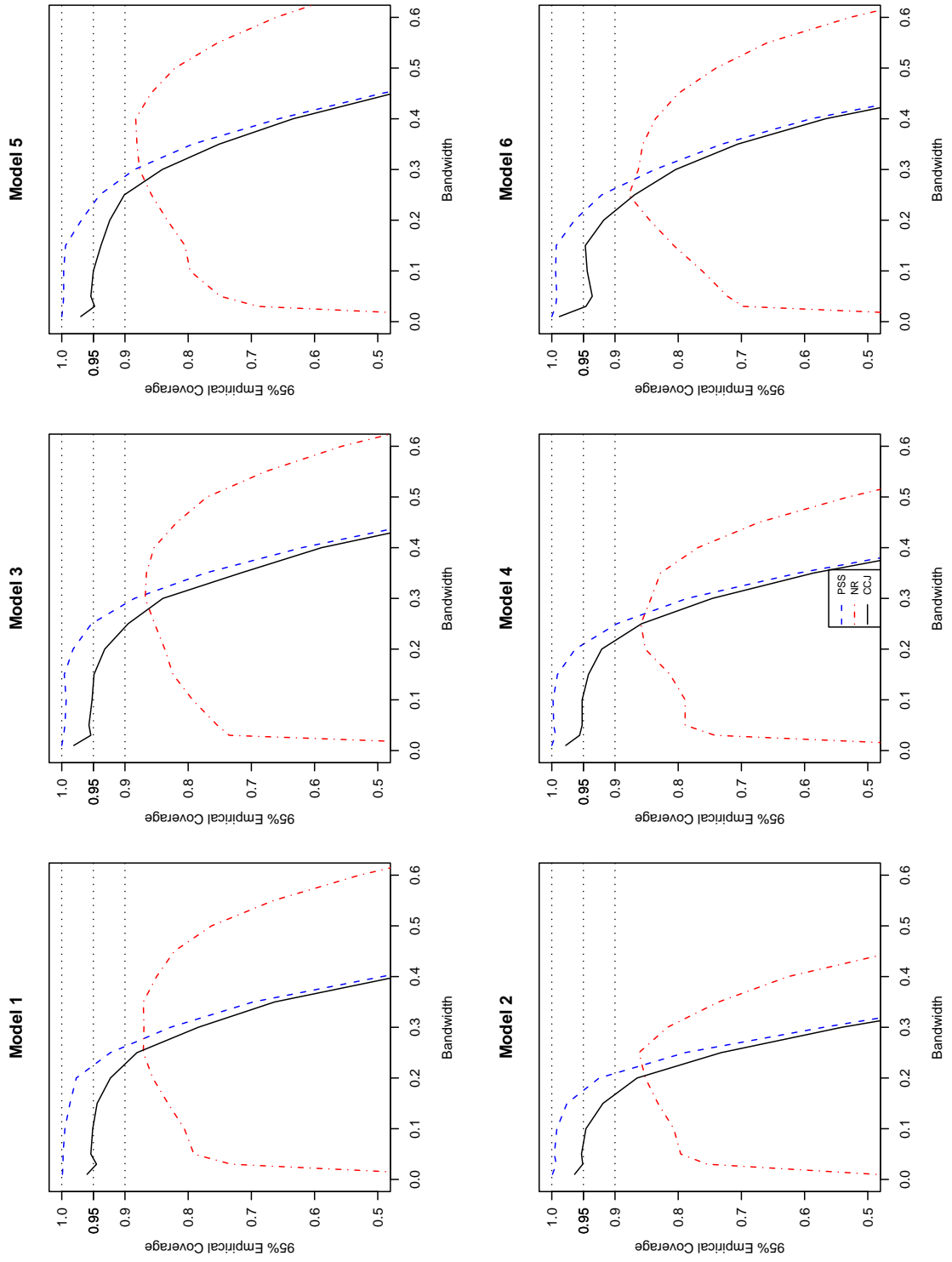


Figure 2: Coverage Rates for 95% Confidence Intervals;  $P = 2$  and  $n = 400$

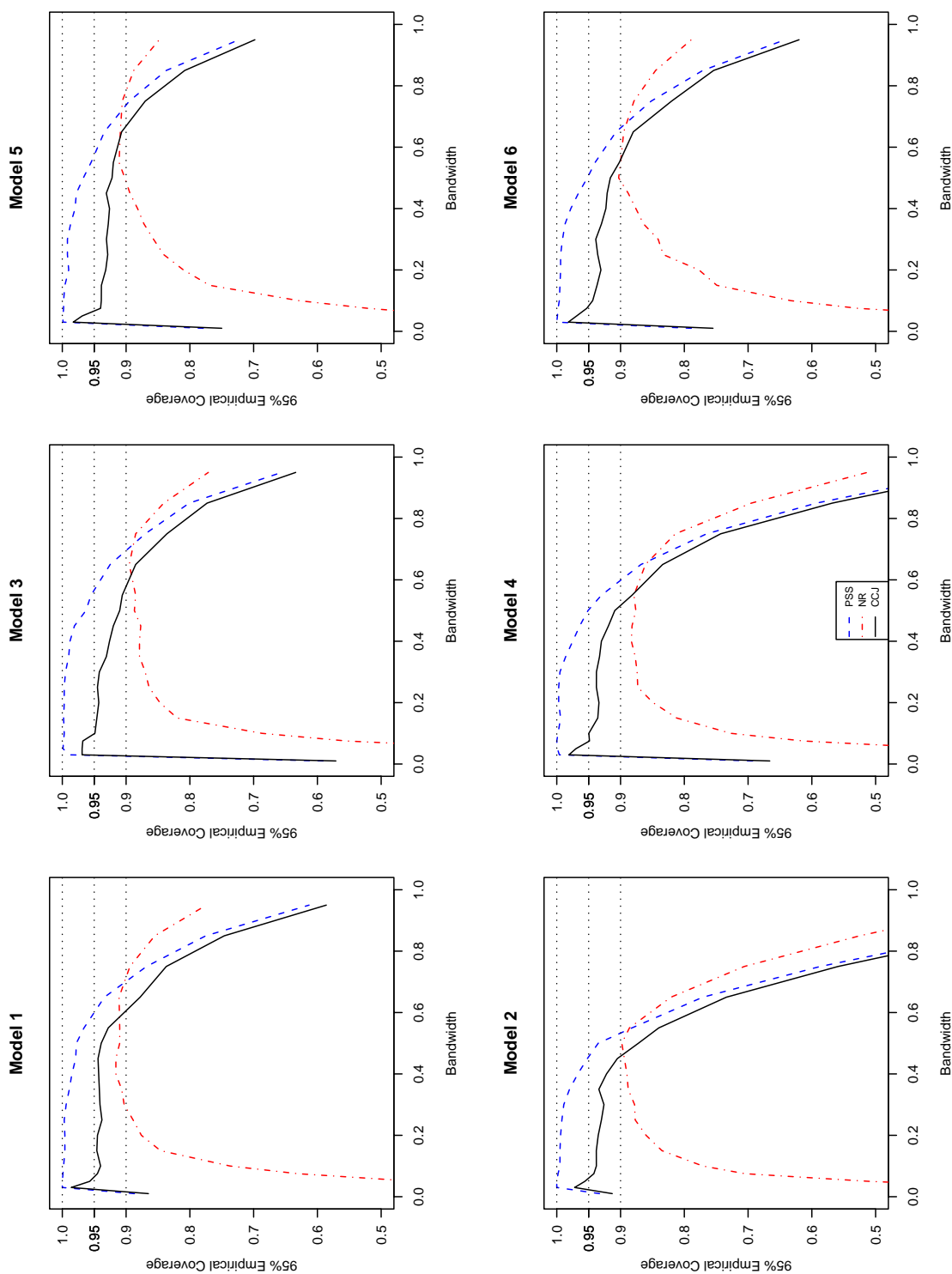


Figure 3: Coverage Rates for 95% Confidence Intervals;  $P = 4$  and  $n = 100$

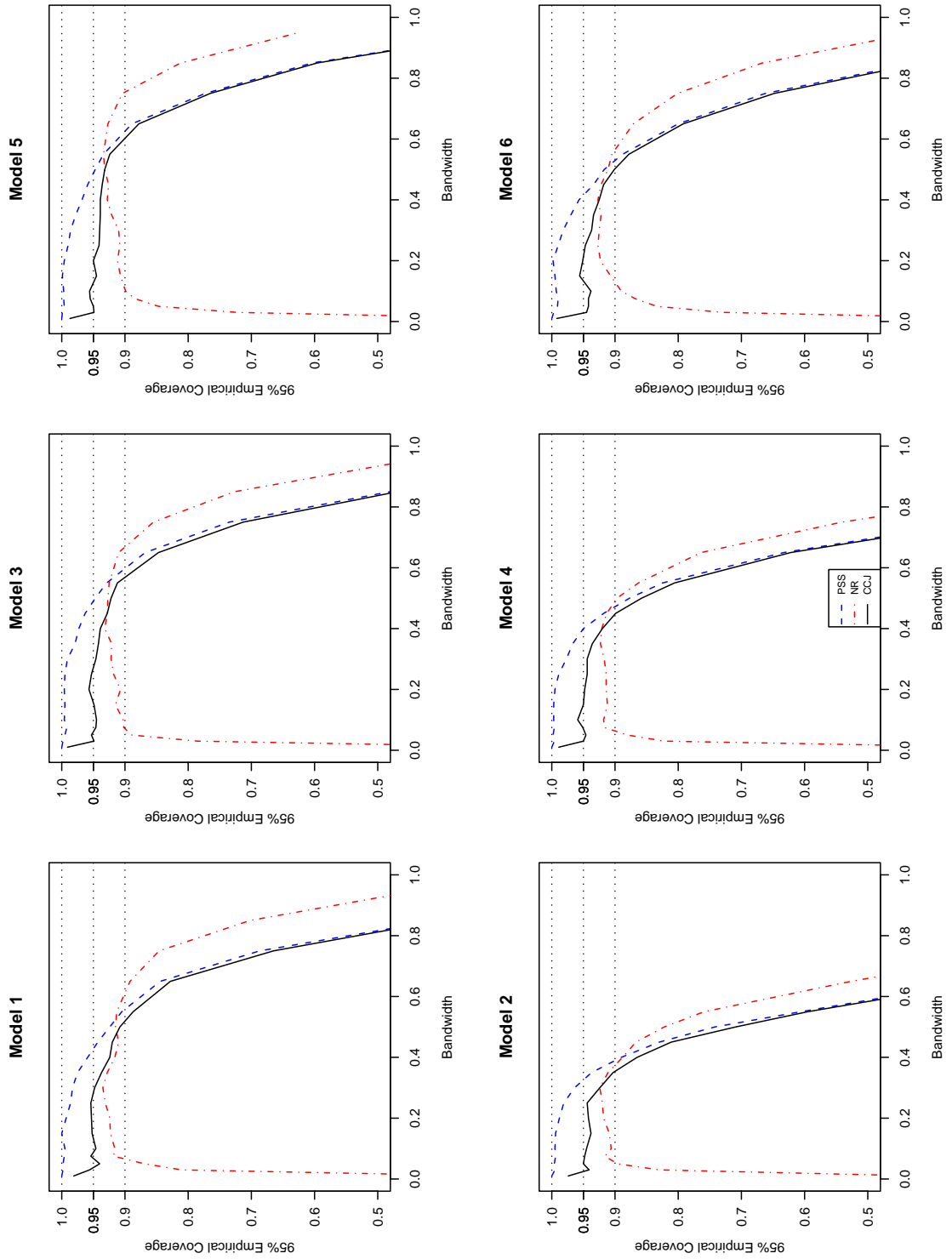


Figure 4: Coverage Rates for 95% Confidence Intervals;  $P = 4$  and  $n = 400$