

Robust Data-Driven Inference for Density-Weighted Average Derivatives*

MATIAS D. CATTANEO

DEPARTMENT OF ECONOMICS, UNIVERSITY OF MICHIGAN

RICHARD K. CRUMP

FEDERAL RESERVE BANK OF NEW YORK

MICHAEL JANSSON

DEPARTMENT OF ECONOMICS, UC BERKELEY AND *CREATES*

February 10, 2010

ABSTRACT. This paper presents a novel data-driven bandwidth selector compatible with the small bandwidth asymptotics developed in Cattaneo, Crump, and Jansson (2009) for density-weighted average derivatives. The new bandwidth selector is of the plug-in variety, and is obtained based on a mean squared error expansion of the estimator of interest. An extensive Monte Carlo experiment shows a remarkable improvement in performance when the bandwidth-dependent robust inference procedures proposed by Cattaneo, Crump, and Jansson (2009) are coupled with this new data-driven bandwidth selector. The resulting robust data-driven confidence intervals compare favorably to the alternative procedures available in the literature. The online supplemental material to this paper contains further results from the simulation study.

Keywords: Averaged derivatives, bandwidth selection, robust inference, small bandwidth asymptotics.

*The authors thank Sebastian Calonico, Lutz Kilian, seminar participants at Georgetown, Michigan, Penn State and Wisconsin, and conference participants at the 2009 Latin American Meeting of the Econometric Society and 2010 North American Winter Meeting of the Econometric Society for comments. We also thank the editor, associate editor and a referee for comments and suggestions that improved this paper. The first author gratefully acknowledges financial support from the National Science Foundation (SES 0921505). The third author gratefully acknowledges financial support from the National Science Foundation (SES 0920953) and the research support of CREATES (funded by the Danish National Research Foundation).

1. INTRODUCTION

Semiparametric models, which include both a finite dimensional parameter of interest and an infinite dimensional nuisance parameter, play a central role in modern statistical and econometric theory, and are potentially of great interest in empirical work. However, the applicability of semiparametric estimators is seriously hampered by the sensitivity of their performance to seemingly ad hoc choices of “smoothing” and “tuning” parameters involved in the estimation procedure. Although classical large sample theory for semiparametric estimators is now well developed, these theoretical results are typically invariant to the particular choice of parameters associated with the nonparametric estimator employed, and usually require strong untestable assumptions (e.g., smoothness of the infinite dimensional nuisance parameter). As a consequence, inference procedures based on these estimators are in general not robust to changes in the choice of tuning and smoothing parameters underlying the nonparametric estimator, and to departures from key unobservable model assumptions. These facts suggest that classical asymptotic results for semiparametric estimators may not always accurately capture their behavior in finite samples, posing considerable restrictions on the overall applicability they may have for empirical work.

This paper proposes two robust data-driven inference procedures for the semiparametric density-weighted average derivatives estimator of Powell, Stock, and Stoker (1989). The averaged derivatives is a simple yet important semiparametric estimand of interest, which naturally arises in many statistical and econometric models such as (nonadditive) single-index models (see, e.g., Powell (1994) and Matzkin (2007) for review). Moreover, this estimand has been considered in a variety of empirical problems, including nonparametric demand estimation (Härdle, Hildenbrand, and Jerison (1991)), policy analysis of tax and subsidy reform (Deaton and Ng (1998)) and nonlinear pricing in labor markets (Coppejans and Sieg (2005)). This paper focuses on the density-weighted average derivatives estimator not only because of its own importance, but also because it admits a particular U -statistic

representation. As discussed in detail below, this representation is heavily exploited in the theoretical developments presented here, which implies that the results in this paper may be extended to cover other estimators having a similar representation.

The main idea is to develop a novel data-driven bandwidth selector compatible with the small bandwidth asymptotic theory presented in Cattaneo, Crump, and Jansson (2009). This alternative (first-order) large sample theory encompasses the classical large sample theory available in the literature, and also enjoys several robustness properties. In particular, (i) it provides valid inference procedures for (small) bandwidth sequences that would render the classical results invalid, (ii) it permits the use of a second-order kernel regardless of the dimension of the regressors and therefore removes strong smoothness assumptions, and (iii) it provides a limiting distribution that is in general not invariant to the particular choices of smoothing and tuning parameters, without necessarily forcing a slower than root- n rate of convergence (where n is the sample size). The key theoretical insight behind these results is to accommodate bandwidth sequences that break down the asymptotic linearity of the estimator of interest, leading to a more general first-order asymptotic theory that is no longer invariant to the particular choices of parameters underlying the preliminary nonparametric estimator. Consequently, it is expected that an inference procedure based on this alternative asymptotic theory would (at least partially) “adapt” to the particular choices of these parameters.

The preliminary simulation results in Cattaneo, Crump, and Jansson (2009) show that this alternative asymptotic theory opens the possibility for the construction of a robust inference procedure, providing a range of (small) bandwidths for which the appropriate test statistic enjoys approximately correct size. However, the bandwidth selectors available in the literature turn out to be incompatible with these new results in the sense that they would not deliver a bandwidth choice within the robust range. The new data-driven bandwidth selector presented here achieves this goal, thereby providing a robust automatic (i.e., fully

data-driven) inference procedure for the estimand of interest. These results are corroborated by an extensive Monte Carlo experiment, which shows that the asymptotic theory developed in Cattaneo, Crump, and Jansson (2009) coupled with the data-driven bandwidth selector proposed here leads to remarkable improvements in inference when compared to the alternative procedures available in the literature. In particular, the resulting confidence intervals exhibit close-to-correct empirical coverage across all designs considered. Among other advantages, these data-driven statistical procedures allow for the use of a second-order kernel, which is believed to deliver more stable results in applications (see, e.g., Horowitz and Härdle (1996)), and appear to be considerably more robust to the additional variability introduced by the estimation of the bandwidth selectors. Furthermore, these results are important because the standard nonparametric bootstrap is not a valid alternative in general to the large sample theory employed in this paper (Cattaneo, Crump, and Jansson (2010)).

Another interesting feature of the analysis presented here is related to the well known trade-off between efficiency and robustness in statistical inference. In particular, the novel procedures presented here are considerably more robust while in general (semiparametric) inefficient. This feature is captured by the behavior of the new robust confidence intervals in the simulation study, where they are seen to have correct size and less bias but larger length on average. For example, when the classical procedure is valid (i.e., when using a higher-order kernel), the efficiency loss is found to be around 10% on average, while the bias of the estimator is reduced by about 60% on average.

This paper contributes to the important literature of semiparametric inference for weighted average derivatives. This population parameter was originally introduced by Stoker (1986), and has been intensely studied since then. Härdle and Stoker (1989) and Härdle, Hart, Marron, and Tsybakov (1992) study general weighted average derivatives estimators, although their results are considerably complicated by the fact that their representation requires handling stochastic denominators and appears to be very sensitive to the choice of trimming

parameters. The density-weighted average derivatives estimator circumvents this problem, while retaining the desirable properties of the general weighted average derivative, and leads to a simple and useful semiparametric estimator. Powell, Stock, and Stoker (1989) study the first-order large sample properties of this estimator and provide sufficient (but not necessary) conditions for root- n consistency and asymptotic normality. Under appropriate restrictions, Newey and Stoker (1993) discuss semiparametric efficiency of weighted average derivatives. Nishiyama and Robinson (2000, 2005) study the second-order large sample properties of density-weighted average derivatives by deriving valid Edgeworth expansions for the estimator considered in this paper (see also Robinson (1995)), while Härdle and Tsybakov (1993) and Powell and Stoker (1996) provide second-order mean squared error expansions for this estimator (see also Newey, Hsieh, and Robins (2004)). Both types of higher-order expansions provide simple plug-in bandwidth selectors targeting different properties of this estimator, and are compatible with the classical large sample theory available in the literature. Ichimura and Todd (2007) provide a recent survey with particular emphasis on implementation.

The rest of the paper is organized as follows. Section 2 describes the model and reviews the main results available in the literature regarding first-order large sample inference for density-weighted average derivatives. Section 3 presents the higher-order mean squared error expansion and develops the new (infeasible) theoretical bandwidth selector, while Section 4 describes how to construct a feasible (i.e., data-driven) bandwidth selector and establishes its consistency. Section 5 summarizes the results of an extensive Monte Carlo experiment. Section 6 discusses how the results may be generalized and concludes.

2. MODEL AND PREVIOUS RESULTS

Let $z_i = (y_i, x_i)'$, $i = 1, \dots, n$, be a random sample from a vector $z = (y, x)'$, where $y \in \mathbb{R}$ is a dependent variable and $x = (x_1, \dots, x_d)' \in \mathbb{R}^d$ is a continuous explanatory variable with a density $f(\cdot)$. The population parameter of interest is the density-weighted average derivative

given by

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

where $g(x) = \mathbb{E}[y|x]$ denotes the population regression function. For example, this estimand is a popular choice for the estimation of the coefficients (up to scale) in a single-index model with unknown link function. To see this, note that $\theta \propto \beta$ when $g(x) = \tau(x'\beta)$ for an unknown (link) function $\tau(\cdot)$, a semiparametric problem that arises in a variety of contexts, including discrete choice and censored models.

The following assumption collects typical regularity conditions imposed on this model.

- Assumption 1.** (a) $\mathbb{E}[y^4] < \infty$, $\mathbb{E}[\sigma^2(x) f(x)] > 0$ and $\mathbb{V}[\partial e(x)/\partial x - y \partial f(x)/\partial x]$ is positive definite, where $\sigma^2(x) = \mathbb{V}[y|x]$ and $e(x) = f(x)g(x)$.
- (b) f is $(Q + 1)$ times differentiable, and f and its first $(Q + 1)$ derivatives are bounded, for some $Q \geq 2$.
- (c) g is twice differentiable, and e and its first two derivatives are bounded.
- (d) 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 $v(x) = \mathbb{E}[y^2|x]$.
- (e) $\lim_{\|x\| \rightarrow \infty} [f(x) + |e(x)|] = 0$.

Assumption 1 and integration by parts lead to $\theta = -2\mathbb{E}[y \partial f(x)/\partial x]$, which in turn motivates the analogue estimator of Powell, Stock, and Stoker (1989) given by

$$\hat{\theta}_n = -2 \frac{1}{n} \sum_{i=1}^n y_i \frac{\partial}{\partial x} \hat{f}_{n,i}(x_i), \quad \hat{f}_{n,i}(x) = \frac{1}{n-1} \sum_{j=1, j \neq i}^n \frac{1}{h_n^d} K\left(\frac{x_j - x}{h_n}\right),$$

where $\hat{f}_{n,i}(\cdot)$ is a “leave-one-out” kernel density estimator for some kernel function $K : \mathbb{R}^d \rightarrow \mathbb{R}$ and some positive (bandwidth) sequence h_n . Typical regularity conditions imposed on the kernel-based nonparametric estimator are given in the following assumption.

Assumption 2. (a) K is even and differentiable, and K and its first derivative are bounded.

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

(c) For some $P \geq 2$, $\int_{\mathbb{R}^d} [|K(u)| (1 + \|u\|^P) + \|\dot{K}(u)\| (1 + \|u\|^2)] du < \infty$, and

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

Powell, Stock, and Stoker (1989) showed that, under appropriate restrictions on the bandwidth sequence and kernel function, the estimator $\hat{\theta}_n$ is asymptotically linear with influence function given by $L(z) = 2[\partial e(x) / \partial x - y \partial f(x) / \partial x - \theta]$. Thus, the asymptotic variance of this estimator is given by $\Sigma = \mathbb{E}[L(z)L(z)']$. Moreover, although not covered by the results in Newey and Stoker (1993), it is possible to show that $L(z)$ is the efficient influence function for θ , and hence Σ is the semiparametric efficiency bound for this estimand. The following result describes the exact conditions and summarizes the main conclusion. (Limits are taken as $n \rightarrow \infty$ unless otherwise noted.)

Result 1. (Powell, Stock, and Stoker (1989)) *If Assumptions 1 and 2 hold, and if $nh_n^{2\min(P,Q)} \rightarrow 0$ and $nh_n^{d+2} \rightarrow \infty$, then*

$$\sqrt{n}(\hat{\theta}_n - \theta) = \frac{1}{\sqrt{n}} \sum_{i=1}^n L(z_i) + o_p(1) \rightarrow_d \mathcal{N}(0, \Sigma).$$

Result 1 follows from noting that the estimator $\hat{\theta}_n$ admits a n -varying U -statistic representation given by

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

which leads to the Hoeffding decomposition $\hat{\theta}_n = \theta_n + \bar{L}_n + \bar{W}_n$, where

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

with $L(z_i; h) = 2[\mathbb{E}[U(z_i, z_j; h)|z_i] - \mathbb{E}[U(z_i, z_j; h)]]$ and $W(z_i, z_j; h) = U(z_i, z_j; h) - (L(z_i; h) + L(z_j; h))/2 - \mathbb{E}[U(z_i, z_j; h)]$. This decomposition shows that the estimator admits a bilinear form representation in general, which clearly justifies the conditions imposed on the bandwidth sequence and the kernel function: (i) condition $nh_n^{2\min(P,Q)} \rightarrow 0$ ensures that the bias of the estimator is asymptotically negligible because $\theta_n - \theta = O(h_n^{\min(P,Q)})$, and (ii) condition $nh_n^{d+2} \rightarrow \infty$ ensures that the ‘‘quadratic term’’ of the Hoeffding decomposition is also asymptotically negligible because $\bar{W}_n = O_p(n^{-1}h_n^{-(d+2)/2})$. Under the same conditions, Powell, Stock, and Stoker (1989) also develop a simple consistent estimator for Σ , which is given by the analogue estimator

$$\hat{\Sigma}_n = \frac{1}{n} \sum_{i=1}^n \hat{L}_{n,i} \hat{L}'_{n,i}, \quad \hat{L}_{n,i} = 2 \left[\frac{1}{n-1} \sum_{j=1, j \neq i}^n U(z_i, z_j; h_n) - \hat{\theta}_n \right].$$

Consequently, under the conditions imposed in Result 1, it is straightforward to form a studentized version of $\hat{\theta}_n$, leading to an asymptotically pivotal test statistic given by $\sqrt{n} \hat{\Sigma}_n^{-1/2} (\hat{\theta}_n - \theta) \rightarrow_d \mathcal{N}(0, I_d)$, with $\hat{\Sigma}_n \rightarrow_p \Sigma$. This test statistic may be used in the usual way to construct a confidence interval for θ (or, equivalently, to carry out the corresponding dual hypothesis test).

As discussed in Newey (1994), asymptotic linearity of a semiparametric estimator has several distinct features that may be considered attractive from a theoretical point of view. In particular, asymptotic linearity is a necessary condition for semiparametric efficiency and leads to a limiting distribution of the statistic of interest that is invariant to the choice of the nonparametric estimator used in the construction of the semiparametric procedure. In

other words, regardless of the particular choice of preliminary nonparametric estimator, the limiting distribution will not depend on the nonparametric estimator whenever the semiparametric estimator admits an asymptotic linear representation.

However, achieving an asymptotic linear representation of a semiparametric estimator imposes several strong model assumptions and leads to a large sample theory that may not accurately represent the finite sample behavior of the estimator. In the case of $\hat{\theta}_n$, asymptotic linearity would require $P > 2$ unless $d = 1$, which in turn requires strong smoothness conditions ($Q \geq P$). Consequently, classical asymptotic theory will require the use of a higher-order kernel whenever more than one covariate is included. In addition, classical asymptotic theory (whenever valid) leads to a limiting experiment which is invariant to the particular choices of smoothing (K) and tuning (h_n) parameters involved in the construction of the estimator, and therefore it is unlikely to be able to “adapt” to changes in these parameters. In other words, inference based on classical asymptotic theory is silent with respect to the impact that these parameters may have on the finite sample behavior of $\hat{\theta}_n$.

In an attempt to better characterize the finite sample behavior of $\hat{\theta}_n$, Cattaneo, Crump, and Jansson (2009) show that it is possible to increase the robustness of this estimator by considering a different asymptotic experiment. In particular, instead of forcing asymptotic linearity of the estimator, the authors develop an alternative first-order asymptotic theory that accommodates weaker assumptions than those imposed in the classical first-order asymptotic theory discussed above. Intuitively, the idea is to characterize the (joint) asymptotic behavior of both the linear (\bar{L}_n) and quadratic (\bar{W}_n) terms. The following result collects the main findings.

Result 2. (Cattaneo, Crump, and Jansson (2009)) *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*

$$(\mathbb{V}[\hat{\theta}_n])^{-1/2}(\hat{\theta}_n - \theta) \rightarrow_d \mathcal{N}(0, I_d),$$

where

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

with $\Delta = 2\mathbb{E}[\sigma^2(x) f(x)] \int_{\mathbb{R}^d} \dot{K}(u) \dot{K}(u)' du$. In addition,

$$\frac{1}{n} \hat{\Sigma}_n = \frac{1}{n} [\Sigma + o_p(1)] + 2 \binom{n}{2}^{-1} h_n^{-(d+2)} [\Delta + o_p(1)].$$

Result 2 shows that the conditions on the bandwidth sequence may be considerably weakened without invalidating the limiting Gaussian distribution. In particular, whenever h_n is chosen so that nh_n^{d+2} is bounded, the limiting distribution will cease to be invariant with respect to the underlying preliminary nonparametric estimator because $\hat{\theta}_n$ is no longer asymptotically linear. (In particular, note that $nh_n^{d+2} \rightarrow \kappa > 0$ retains the root- n consistency of $\hat{\theta}_n$.) In addition, because h_n is allowed to be “smaller” than usual, the bias of the estimator is controlled in a different way, removing the need for higher-order kernels. In particular, Result 2 remains valid even in cases when the estimator is not consistent. Finally, this result also highlights the well known trade-off between robustness and efficiency in the context of semiparametric estimation. In particular, the estimator $\hat{\theta}_n$ is semiparametric efficient if and only if $nh_n^{d+2} \rightarrow \infty$, while it is possible to construct more robust inference procedures under considerably weaker conditions.

It follows from Result 2 that the feasible classical testing procedure based on $\sqrt{n} \hat{\Sigma}_n^{-1/2} (\hat{\theta}_n - \theta)$ will be invalid unless $nh_n^{d+2} \rightarrow \infty$, which corresponds to the classical large sample theory case (Result 1). To solve this problem, Cattaneo, Crump, and Jansson (2009) propose two alternative corrections to the standard error matrix $\hat{\Sigma}_n$, leading to two options for “robust” standard errors. To construct the first “robust” standard error formula, the authors introduce a simple consistent estimator for Δ , under the same conditions of Result 2, which is

given by the analogue estimator

$$\hat{\Delta}_n = h_n^{d+2} \binom{n}{2}^{-1} \sum_{i=1}^{n-1} \sum_{j=i+1}^n \hat{W}_{n,ij} \hat{W}'_{n,ij}, \quad \hat{W}_{n,ij} = U(z_i, z_j; h_n) - \frac{1}{2} (\hat{L}_{n,i} + \hat{L}_{n,j}) - \hat{\theta}_n.$$

Thus, using this estimator,

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

yields a consistent standard error estimate under small bandwidth asymptotics (i.e., under the weaker conditions imposed in Result 2, which include in particular those imposed in Result 1). To describe the second “robust” standard error formula, let $\hat{\Sigma}_n(H_n)$ be the estimator $\hat{\Sigma}_n$ constructed using a bandwidth sequence H_n (e.g., $\hat{\Sigma}_n = \hat{\Sigma}_n(h_n)$ by definition). Then, under the same conditions of Result 2,

$$\hat{V}_{2,n} = \frac{1}{n} \hat{\Sigma}_n(2^{1/(d+2)} h_n)$$

also yields a consistent standard error estimate under small bandwidth asymptotics.

Consequently, under the conditions imposed in Result 2, it is straightforward to form a studentized version of $\hat{\theta}_n$, leading to two simple, robust and pivotal test statistics of the form $\hat{V}_{k,n}^{-1/2}(\hat{\theta}_n - \theta) \rightarrow_d \mathcal{N}(0, I_d)$, with $\hat{V}_{k,n}^{-1} \mathbb{V}[\hat{\theta}_n] \rightarrow_p I_d$, $k = 1, 2$. These test statistics may also be used to construct (asymptotically equivalent) confidence intervals for θ under the (weaker) conditions imposed in Result 2, and constitute alternative procedures to the classical confidence interval introduced above.

These results, however, have the obvious drawback of being dependent on the choice of h_n , which is unrestricted beyond the rate restrictions imposed in Result 2. A preliminary Monte Carlo experiment reported in Cattaneo, Crump, and Jansson (2009) shows that the new, robust standard error formulas have the potential to deliver good finite sample behavior if

the initial bandwidth is chosen to be small enough. Unfortunately, the plug-in rules available in the literature for h_n fail to deliver a choice of bandwidth that would enjoy the robustness property introduced by the new asymptotic theory described in Result 2. This is not too surprising, since these bandwidth selectors are typically constructed to balance (higher-order) bias and variance in a way that is “appropriate” for the classical large sample theory.

3. MSE EXPANSION AND “OPTIMAL” BANDWIDTH SELECTORS

This paper considers the mean squared error expansion of $\hat{\theta}_n$ as the starting point for the construction of the plug-in “optimal” bandwidth selector. To derive this expansion it is necessary to strengthen the assumptions concerning the data generating process. The following assumption describes these additional mild sufficient conditions.

Assumption 3. (a) $\mathbb{E}[\|\partial g(x)/\partial x\|^2 f(x)] < \infty$.

(b) g is $(Q + 1)$ times differentiable, and e and its first $(Q + 1)$ derivatives are bounded.

(c) v is three times differentiable, and vf and its first three derivatives are bounded.

(d) $\lim_{\|x\| \rightarrow \infty} [\sigma(x) f(x) + \|\partial \sigma(x)/\partial x\| f(x)] = 0$.

Assumption 3(a) is used to ensure that the higher-order mean squared expansion is valid up to the order needed in this paper. Assumptions 3(b) and 3(c) are in agreement with those imposed in Powell and Stoker (1996) and Nishiyama and Robinson (2000, 2005), while Assumption 3(d) is slightly stronger than the analogue restriction imposed in those papers.

Theorem 1. *If Assumptions 1, 2 and 3 hold, then for $s = \min(P, Q)$ and $\dot{f}(x) = \partial f(x)/\partial x$,*

$$\begin{aligned} \mathbb{E} \left[(\hat{\theta}_n - \theta)(\hat{\theta}_n - \theta)' \right] &= \frac{1}{n} \Sigma + \binom{n}{2}^{-1} h_n^{-(d+2)} \Delta + \binom{n}{2}^{-1} h_n^{-d} \mathcal{V} + h_n^{2s} \mathcal{B}\mathcal{B}' \\ &\quad + O(n^{-1} h_n^s) + o(n^{-2} h^{-d} + h_n^{2s}), \end{aligned}$$

where

$$\mathcal{B} = -\frac{2(-1)^s}{s!} \sum_{\substack{0 \leq l_1, \dots, l_d \leq s \\ l_1 + \dots + l_d = s}} \left[\int_{\mathbb{R}^d} u_1^{l_1} \cdots u_d^{l_d} K(u) \, du \right] \mathbb{E} \left[\left(\frac{\partial^{(l_1 + \dots + l_d)}}{\partial x_1^{l_1} \cdots \partial x_d^{l_d}} \dot{f}(x) \right) g(x) \right]$$

and

$$\mathcal{V} = \int_{\mathbb{R}^d} \dot{K}(u) \dot{K}(u)' \left(u' \mathbb{E} \left[\sigma^2(x) \frac{\partial^2}{\partial x \partial x'} f(x) + \left(\frac{\partial}{\partial x} g(x) \right) \left(\frac{\partial}{\partial x} g(x) \right)' f(x) \right] u \right) \, du.$$

The result in Theorem 1 is similar to the one obtained by Härdle and Tsybakov (1993) and Powell and Stoker (1996), the key difference being that the additional term of order $O(n^{-2}h_n^{-d})$ is explicitly retained here. (Recall that Result 2 requires $n^2h_n^d \rightarrow \infty$.)

To motivate the new “optimal” bandwidth selector, recall that the “robust” variance matrix in Result 2 is given by the first two terms of the mean squared error expansion presented in Theorem 1, which suggests considering the next two terms of the expansion to construct an “optimal” bandwidth selector. (Note that, as it is common in the literature, this approach implicitly assumes that both \mathcal{B} and \mathcal{V} are non-zero.) Intuitively, balancing these terms corresponds to the case of $nh_n^{d+2} \rightarrow \kappa < \infty$, and therefore pushes the selected bandwidth to the “small bandwidth region”. This approach may be considered “optimal” in a mean square error sense because it makes the leading terms ignored in the general large sample approximation presented in Result 2 as small as possible.

To describe the new bandwidth selector, let $\lambda \in \mathbb{R}^d$ and consider (for simplicity) a bandwidth that minimizes the next two terms of $\mathbb{E}[(\lambda'(\hat{\theta}_n - \theta))^2]$. This “optimal” bandwidth selector is given by

$$h_{CCJ}^* = \begin{cases} \left(\frac{d(\lambda' \mathcal{V} \lambda)}{s(\lambda' \mathcal{B})^2 n^2} \right)^{\frac{1}{2s+d}} & \text{if } \lambda' \mathcal{V} \lambda > 0 \\ \left(\frac{2|\lambda' \mathcal{V} \lambda|}{(\lambda' \mathcal{B})^2 n^2} \right)^{\frac{1}{2s+d}} & \text{if } \lambda' \mathcal{V} \lambda < 0 \end{cases}.$$

This new theoretical bandwidth selector is consistent with the small bandwidth asymptotics described in Result 2 because $n^2 (h_{CCJ}^*)^d \rightarrow \infty$. In addition, observe that $n^{-1}h_n^s = o(n^{-2}h_n^{-d})$ whenever $nh_n^{s+d} \rightarrow 0$, which is satisfied when $h_n = h_{CCJ}^*$.

This new bandwidth selector may be compared to the two competing plug-in bandwidth selectors available in the literature, proposed by Powell and Stoker (1996) and Nishiyama and Robinson (2005), and given by

$$h_{PS}^* = \left(\frac{(d+2)(\lambda' \Delta \lambda)}{s(\lambda' \mathcal{B})^2 n^2} \right)^{\frac{1}{2s+d+2}} \quad \text{and} \quad h_{NR}^* = \left(\frac{2(\lambda' \Delta \lambda)}{(\lambda' \mathcal{B})^2 n^2} \right)^{\frac{1}{2s+d+2}},$$

respectively. Inspection of these bandwidth selectors shows that $h_{CCJ}^* \prec h_{PS}^* \asymp h_{NR}^*$, leading to a bandwidth selection of smaller order.¹

4. DATA-DRIVEN BANDWIDTH SELECTORS

The previous section described a new (infeasible) plug-in bandwidth selector that is compatible with the small bandwidth asymptotic theory introduced in Result 2. In order to implement this selector in practice, as well as its competitors h_{PS}^* and h_{NR}^* , it is necessary to construct consistent estimates for each of the leading constants. These estimates would lead to a data-driven (i.e., automatic) bandwidth selector, denoted \hat{h}_{CCJ} . This section introduces easy to implement, consistent nonparametric estimators for \mathcal{B} , Δ and \mathcal{V} .²

To describe the data-driven plug-in bandwidth selectors, let b_n be a preliminary positive bandwidth sequence, which may be different for each estimator. A simple analogue estimator of Δ was introduced in Section 2. In particular, let $\hat{\Delta}_n(b_n)$ be the estimator $\hat{\Delta}_n$ constructed

¹Nishiyama and Robinson (2000) derive a third alternative bandwidth selector which is not explicitly discussed here because this procedure is targeted to one-sided hypothesis testing. Nonetheless, inspection of this alternative bandwidth selection procedure, denoted h_{NR00}^* , shows that $h_{CCJ}^* \prec h_{NR00}^*$ whenever $d+8 > 2s$. Therefore, h_{CCJ}^* is of smaller order unless strong smoothness assumptions are imposed in the model and a corresponding higher-order kernel is employed.

²Alternatively, a straightforward bandwidth selector may be constructed using a “rule-of-thumb” estimator based on some ad-hoc distributional assumptions.

using a bandwidth sequence b_n (e.g., $\hat{\Delta}_n = \hat{\Delta}_n(h_n)$ by definition). Note that this estimator is a n -varying U -statistic as well. Theorem 1 and the calculations provided in Cattaneo, Crump, and Jansson (2009) show that, if Assumptions 1, 2 and 3 hold, then

$$\hat{\Delta}_n(b_n) = \Delta + b_n^2 \mathcal{V} + O_p(b_n^3 + n^{-1/2} + n^{-1}b_n^{-d/2}),$$

which gives the consistency of this estimator if $b_n \rightarrow 0$ and $n^2 b_n^d \rightarrow \infty$.

Next, consider the construction of consistent estimators of \mathcal{B} and \mathcal{V} , the two parameters entering the new bandwidth selector h_{CCJ}^* . To this end, let k be a kernel function, which may be different for each estimator, and may be different from K . The following assumption collects a set of sufficient conditions to establish consistency of the plug-in estimators proposed in this paper for \mathcal{B} and \mathcal{V} .

Assumption 4. (a) f , v and e are $(s + 1 + S)$ times differentiable, and f , vf , e and their first $(s + 1 + S)$ derivatives are bounded, for some $S \geq 1$.

(b) k is even and M times differentiable, and k and its first M derivatives are bounded, for some $M \geq 0$.

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

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

For the bias \mathcal{B} , a plug-in estimator is given by

$$\hat{\mathcal{B}}_n = -\frac{2(-1)^s}{s!} \sum_{\substack{0 \leq l_1, \dots, l_d \leq s \\ l_1 + \dots + l_d = s}} \left[\int_{\mathbb{R}^d} u_1^{l_1} \cdots u_d^{l_d} K(u) du \right] \hat{\vartheta}_{l_1, \dots, l_d, n},$$

where

$$\hat{\vartheta}_{l_1, \dots, l_d, n} = \frac{1}{n(n-1)} \sum_{i=1}^n \sum_{j=1, j \neq i}^n b_n^{-(d+1)} \left(\frac{\partial^{(l_1 + \dots + l_d)}}{\partial x_1^{l_1} \dots \partial x_d^{l_d}} \dot{k} \left(\frac{x_i - x_j}{b_n} \right) \right) y_i.$$

The estimator $\hat{\vartheta}_{l_1, \dots, l_d, n}$ is the sample analogue estimator of $\mathbb{E}[(\partial^{(l_1 + \dots + l_d)} \dot{f}(x) / \partial x_1^{l_1} \dots \partial x_d^{l_d}) y]$, and is also a n -varying U -statistic estimator employing a leave-one-out kernel-based density estimator.

It is also possible to form an obvious plug-in estimator for the new higher-order term \mathcal{V} . However, this estimator would have the unappealing property of requiring the estimation of several nonparametric objects ($\sigma^2(x)$, $\partial^2 f(x) / \partial x \partial x'$, $\partial g(x) / \partial x$, $f(x)$). Moreover, this direct plug-in approach is likely to be less stable when implemented because it would require handling stochastic denominators. Fortunately, it is possible to construct an alternative, indirect estimator much easier to implement in practice. This estimator is intuitively justified as follows: the results presented above show that, under appropriate regularity conditions, $b_n^{-2}(\hat{\Delta}_n(b_n) - \Delta) = \mathcal{V} + O_p(b_n + n^{-1/2}b_n^{-2} + n^{-1}b_n^{-d/2-2})$, and therefore an estimator satisfying $\tilde{\Delta}_n = \Delta + o_p(b_n^2)$ would lead to

$$\hat{\mathcal{V}}_n = b_n^{-2}(\hat{\Delta}_n(b_n) - \tilde{\Delta}_n) = \mathcal{V} + o_p(1),$$

if $b_n \rightarrow 0$, $nb_n^4 \rightarrow 0$ and $n^2b_n^{d+4} \rightarrow 0$. Under appropriate conditions, an estimator having these properties is given by

$$\tilde{\Delta}_n = \hat{\delta}_n \int_{\mathbb{R}^d} \dot{K}(u) \dot{K}(u)' du, \quad \hat{\delta}_n = \binom{n}{2}^{-1} \sum_{i=1}^{n-1} \sum_{j=i+1}^n b_n^{-d} k \left(\frac{x_j - x_i}{b_n} \right) (y_i - y_j)^2.$$

In this case, $\hat{\delta}_n$ is a sample analogue estimator of $2\mathbb{E}[\sigma^2(x) f(x)]$, which is also a n -varying U -statistic estimator employing a leave-one-out kernel-based density estimator.

Theorem 2. *If Assumptions 1, 3 and 4 hold, then:*

(i) For $M \geq s + 1$,

$$\hat{\vartheta}_{l_1, \dots, l_d, n} = \mathbb{E} \left[\left(\frac{\partial^{(l_1 + \dots + l_d)}}{\partial x_1^{l_1} \dots \partial x_d^{l_d}} \dot{f}(x) \right) y \right] + O_p \left(b_n^{\min(R, S)} + n^{-1/2} + n^{-1} b_n^{-(d+2+2s)/2} \right).$$

(ii) For $R \geq 3$,

$$\hat{\delta}_n = 2\mathbb{E} [\sigma^2(x) f(x)] + O_p \left(b_n^{\min(R, s+1+S)} + n^{-1/2} + n^{-1} b_n^{-d/2} \right).$$

This theorem gives simple sufficient conditions to construct a robust data-driven bandwidth selector consistent with the small bandwidth asymptotics derived in Cattaneo, Crump, and Jansson (2009). In particular, define

$$\hat{h}_{CCJ} = \begin{cases} \left(\frac{d(\lambda' \hat{\mathcal{V}}_n \lambda)}{s(\lambda' \hat{\mathcal{B}}_n)^2 n^2} \right)^{\frac{1}{2s+d}} & \text{if } \lambda' \hat{\mathcal{V}}_n \lambda > 0 \\ \left(\frac{2|\lambda' \hat{\mathcal{V}}_n \lambda|}{(\lambda' \hat{\mathcal{B}}_n)^2 n^2} \right)^{\frac{1}{2s+d}} & \text{if } \lambda' \hat{\mathcal{V}}_n \lambda < 0 \end{cases}.$$

The following corollary establishes the consistency of the new bandwidth selector \hat{h}_{CCJ} .

Corollary 1. *If Assumptions 1, 2, 3 and 4 hold with $M \geq s + 1$ and $R \geq 3$, and if $b_n \rightarrow 0$ and $n^2 b_n^{\max(8, d+2+2s)} \rightarrow \infty$, then for $\lambda \in \mathbb{R}^d$ such that $\lambda' \mathcal{B} \neq 0$ and $\lambda' \mathcal{V} \lambda \neq 0$,*

$$\frac{\hat{h}_{CCJ}}{h_{CCJ}^*} \rightarrow_p 1.$$

(The analogous result also holds for \hat{h}_{PS} and \hat{h}_{NR} .)

The results presented so far are silent about the selection of the initial bandwidth choice b_n in applications, beyond the rate restrictions imposed by Corollary 1. A simple choice for the preliminary bandwidth b_n may be based on some data-driven bandwidth selector

developed for a nonparametric object present in the corresponding target estimands \mathcal{B} , Δ and \mathcal{V} . Typical examples of such procedures include simple rule-of-thumbs, plug-in bandwidth selectors and (smoothed) cross-validation.

As shown in the simulations presented in the next section, it appears that a simple data-driven bandwidth selector from the literature of nonparametric estimation works well for the choice of b_n . Nonetheless, it may be desirable to improve upon this preliminary bandwidth selector in order to obtain better finite sample behavior. Although beyond the scope of this paper, a conceptually feasible (but computationally demanding) idea would be to compute second-order mean squared error expansions for $\hat{\vartheta}_{l_1, \dots, l_d, n}$, $\hat{\Delta}_n$ and $\hat{\delta}_n$. Since these three estimators are n -varying U -statistics, the results from Powell and Stoker (1996) may be applied to obtain a corresponding set of “optimal” bandwidth choices. These procedures will, in turn, also depend on a preliminary bandwidth when implemented empirically, which again would need to be chosen in some way. This idea mimics, in the context of semiparametric estimation, the well-known second-generation direct plug-in bandwidth selector (of level 2) from the literature of nonparametric density estimation. (See, e.g., Wand and Jones (1995) for a detailed discussion.) Although the validity of such bandwidth selectors would require stronger assumptions, by analogy from the nonparametric density estimation literature, they would be expected to improve the finite sample properties of the bandwidth selector for h_n and, in turn, the performance of the semiparametric inference procedure.

5. MONTE CARLO EXPERIMENT

This section summarizes the main findings from an extensive Monte Carlo experiment conducted to analyze the finite sample properties of the new robust data-driven procedures and their relative merits when compared to the other procedures available. The online supplemental material includes a larger set of results from this simulation study, which shows that the findings reported here are consistent across all designs considered.

Following the results reported in Cattaneo, Crump, and Jansson (2009), the Monte Carlo experiment considers six different models of the “single index” form $y_i = \tau(y_i^*)$, where $y_i^* = x_i' \beta + \varepsilon_i$, $\tau(\cdot)$ is a nondecreasing (link) function and $\varepsilon_i \sim \mathcal{N}(0, 1)$ is independent of the vector of regressors $x_i \in \mathbb{R}^d$. Three different link functions are considered: $\tau(y^*) = y^*$, $\tau(y^*) = \mathbf{1}(y^* > 0)$ and $\tau(y^*) = y^* \mathbf{1}(y^* > 0)$, which correspond to a linear regression, probit, and Tobit model, respectively. ($\mathbf{1}(\cdot)$ represents the indicator function.) The vector of regressors is generated using independent random variables and standardized to have $\mathbb{E}[x_i] = 0$ and $\mathbb{E}[x_i x_i'] = I_d$, with the first component x_{1i} having either a Gaussian distribution or a chi-squared distribution with 4 degrees of freedom (denoted χ_4), while the remaining components have a Gaussian distribution throughout the experiment. All the components of β are set equal to unity, and for simplicity only results for the first component θ_1 are considered.

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: $\theta_1 = \frac{1}{8\pi}$
$x_{1i} \sim \frac{\chi_4 - 4}{\sqrt{8}}$	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. (Whenever unavailable in closed form, the population parameters are computed by a numerical approximation.) The simulation study considers three sample sizes ($n = 100, n = 400, n = 700$), two dimensions of the regressors vector ($d = 2, d = 4$), and two kernel orders ($P = 2, P = 4$). The kernel function $K(\cdot)$ is chosen to be a Gaussian product kernel, and the preliminary kernel function $k(\cdot)$ is chosen to be a fourth-order Gaussian product kernel as required by Corollary 1. For each combination of parameters 10,000 replications are carried out. To conserve space this section only includes the results for $d = 2$ and $n = 400$.

The simulation experiment considers the three (infeasible) population bandwidth choices derived in Section 3 (h_{PS}^* , h_{NR}^* , h_{CCJ}^*), and their corresponding data-driven estimates (\hat{h}_{PS} , \hat{h}_{NR} , \hat{h}_{CCJ}). The three estimated bandwidths are obtained using the results described in Section 4 with a common initial bandwidth plug-in estimate used to construct $\hat{\mathcal{B}}_n$, $\hat{\Delta}_n$ and $\hat{\mathcal{V}}_n$. To provide a parsimonious data-driven procedure, an estimate of the initial bandwidth b_n is constructed as a sample average of a second-generation direct plug-in level-two estimate for the (marginal) density of each dimension of the regressors vector (see, e.g., Wand and Jones (1995)). Confidence intervals for θ_1 are constructed using the classical test statistic $\sqrt{n}\hat{\Sigma}_n^{-1/2}(\hat{\theta}_n - \theta)$, denoted PSS, and the two alternative robust test statistics $\hat{V}_{k,n}^{-1/2}(\hat{\theta}_n - \theta)$, $k = 1, 2$, denoted by CCJ1 and CCJ2, respectively. The classical inference procedure PSS is only theoretically valid when $P = 4$, while the robust procedures CCJ1 and CCJ2 are always valid across all simulation designs.

Figures 1 and 2 plot the empirical coverage for the three competing 95% confidence intervals as a function of the choice of bandwidth for each of the six models. To facilitate the analysis two additional horizontal lines at 0.90 and at the nominal coverage rate 0.95 are included for reference, and the three population bandwidth selectors (h_{PS}^* , h_{NR}^* , h_{CCJ}^*) are plotted as vertical lines. (Note that $h_{PS}^* = h_{NR}^*$ for the case $d = 2$ and $P = 2$.) These figures highlight the potential robustness properties that the test statistics CCJ1 and CCJ2 may have when using the new data-driven plug-in bandwidth selector. In particular, the theoretical bandwidth selector h_{CCJ}^* lays within the robust region for which both CCJ1 and CCJ2 have correct empirical coverage for a range of bandwidths. For example, this suggests that (at least) some of the variability introduced by the estimation of this bandwidth selector will not affect the performance of the robust test statistics CCJ1 and CCJ2, a property unlikely to hold for the classical procedure PSS. Table 1 reports the empirical coverage of each possible confidence intervals (PSS, CCJ1, CCJ2) when using each possible population bandwidth selector (h_{PS}^* , h_{NR}^* , h_{CCJ}^*).

Figures 3 and 4 plot corresponding kernel density estimates for the test statistic PSS coupled with either h_{PS}^* and h_{NR}^* , and for the test statistics CCJ1 and CCJ2 coupled with h_{CCJ}^* . To facilitate the comparison the density of the standard normal is also depicted. These figures show that the Gaussian approximation of the robust test statistics using the new bandwidth selector is considerably better than the corresponding approximation for PSS when constructed using either of the classical bandwidth selectors. In particular, the empirical distribution of the classical procedure appears to be more biased and more concentrated than the empirical distributions of either CCJ1 or CCJ2. These findings highlight the well known trade-off between efficiency and robustness previously discussed. These results are verified in Table 2, where the average empirical bias and average empirical interval length are reported for each competing confidence interval when coupled with each possible population bandwidth selector.

To analyze the performance of the new data-driven bandwidth selector, and the resulting robust data-driven confidence intervals, Table 3 presents the empirical coverage of each possible confidence interval (PSS, CCJ1, CCJ2) when using each possible estimated bandwidth selector (\hat{h}_{PS} , \hat{h}_{NR} , \hat{h}_{CCJ}). These tables provide concrete evidence of the superior performance (in terms of achieving correct coverage) of the robust test statistics when coupled with the new estimated bandwidth. Both robust confidence intervals (CCJ1, CCJ2) using \hat{h}_{CCJ} provide close-to-correct empirical coverage across all designs, a property not enjoyed by the classical confidence interval (PSS) using either \hat{h}_{PS} or \hat{h}_{NR} .

The good performance of CCJ1 and CCJ2 is maintained not only when using a second-order kernel ($P = 2$), but also when the dimension of x is larger ($d = 4$), which provides simulation evidence of the relatively low sensitivity of the new robust data-driven procedures to the so-called “curse of dimensionality.” This finding may be (heuristically) justified by the fact that under the small bandwidth asymptotics, the limiting distribution is not invariant to the “parameter” d , which in turn may lead to the additional robustness properties found.

In addition, as suggested by the superior distributional approximation reported in Figures 3 and 4, the main findings continue to hold if other nominal confidence levels are considered.

6. EXTENSIONS AND FINAL REMARKS

This paper introduced a novel data-driven plug-in bandwidth selector compatible with the small bandwidth asymptotics developed in Cattaneo, Crump, and Jansson (2009) for density-weighted average derivatives. This new bandwidth selector is of the plug-in variety, and is obtained based on a mean squared error expansion of the estimator of interest. An extensive Monte Carlo experiment showed a remarkable improvement in performance of the resulting new robust data-driven inference procedure. In particular, the new confidence intervals provide approximately correct coverage in cases where there is no valid alternative inference procedures (i.e., using a second-order kernel with at least two regressors), and also compares favorably to the alternative, classical confidence intervals when they are theoretically justified.

Since these results are derived by exploiting the n -varying U -statistic representation of $\hat{\theta}_n$, it is plausible that similar results could be obtained for other estimators having an analogous representation. For example, the class of estimands considered in Newey, Hsieh, and Robins (2004, Section 2) have this representation, and therefore it seems possible that the results presented here could be generalized to cover that class. More generally, as suggested in Cattaneo, Crump, and Jansson (2009), an n -varying U -statistic may be represented as a minimizer of the U -process:

$$\hat{\theta}_n = \arg \min_{\theta} \binom{n}{2}^{-1} \sum_{i=1}^{n-1} \sum_{j=i+1}^n Q(z_i, z_j; \theta, h_n), \quad Q(z_i, z_j; \theta, h) = \|U(z_i, z_j, h) - \theta\|^2,$$

which also suggests that the results presented here may be extended to cover this class of estimators (see, e.g., Aradillas-Lopez, Honore, and Powell (2007, pp. 1120–1122)).

7. APPENDIX

Proof of Theorem 1. To save notation, for any function $a : \mathbb{R}^d \rightarrow \mathbb{R}$ let $\dot{a}(x) = \partial a(x) / \partial x$ and $\ddot{a}(x) = \partial^2 a(x) / \partial x \partial x'$. A Hoeffding decomposition of $\hat{\theta}_n$ gives

$$\begin{aligned} \mathbb{E} [(\hat{\theta}_n - \theta)(\hat{\theta}_n - \theta)'] &= \mathbb{V}[\hat{\theta}_n] + \left(\mathbb{E}[\hat{\theta}_n] - \theta \right) \left(\mathbb{E}[\hat{\theta}_n] - \theta \right)' \\ &= \mathbb{V}[\bar{L}_n] + \mathbb{V}[\bar{W}_n] + h_n^{2s} \mathcal{B}\mathcal{B}' + o(h_n^{2s}), \end{aligned}$$

where the bias expansion follows immediately by a Taylor series expansion.

For $\mathbb{V}[\bar{L}_n]$, using integration by parts,

$$\mathbb{E}[U_n(z_i, z_j) | z_i] = \int_{\mathbb{R}^d} \dot{e}(x_i + uh_n) K(u) du - y_i \int_{\mathbb{R}^d} \dot{f}(x_i + uh_n) K(u) du,$$

and therefore $\mathbb{V}[\bar{L}_n] = 4n^{-1} \mathbb{V}[\mathbb{E}[U_n(z_i, z_j) | z_i] - \theta_n] = n^{-1} \Sigma + O(n^{-1} h_n^s)$.

For $\mathbb{V}[\bar{W}_n]$, by standard calculations,

$$\begin{aligned} \mathbb{V}[\bar{W}_n] &= \binom{n}{2}^{-1} \mathbb{E}[U_n(z_i, z_j) U_n(z_i, z_j)'] + O(n^{-2}) \\ &= \binom{n}{2}^{-1} h_n^{-(d+2)} \int_{\mathbb{R}^d} \dot{K}(u) \dot{K}(u)' T(x, uh_n) dx du + O(n^{-2}), \end{aligned}$$

with $T(x, u) = (v(x) + v(x+u) - 2g(x)g(x+u))f(x)f(x+u)$. Then, using a Taylor series expansion, $T(x, uh_n) = T_1(x) + T_2(x)'uh_n + u'T_3(x)uh_n^2 + o(h_n^2)$, where $T_1(x) = 2\sigma^2(x)f(x)^2$, $T_2(x) = 2\sigma^2(x)f(x)\dot{f}(x) + f(x)^2\dot{\sigma}^2(x)$, and $T_3(x) = \sigma^2(x)f(x)\ddot{f}(x) + f(x)\dot{\sigma}^2(x)\dot{f}(x) + (\ddot{v}(x)/2 - g(x)\ddot{g}(x))f(x)^2$.

Note that $\int_{\mathbb{R}^d} \int_{\mathbb{R}^d} \dot{K}(u) \dot{K}(u)' T_1(x) dx du = \int_{\mathbb{R}^d} \int_{\mathbb{R}^d} \dot{K}(u) \dot{K}(u)' 2\sigma^2(x) f(x)^2 dx du = \Delta$ and, using integration by parts,

$$\begin{aligned} &h_n \int_{\mathbb{R}^d} \int_{\mathbb{R}^d} \dot{K}(u) \dot{K}(u)' (T_2(x)' u) dx du \\ &= h_n \int_{\mathbb{R}^d} \dot{K}(u) \dot{K}(u)' \left[\left(\int_{\mathbb{R}^d} \left[\sigma^2(x) 2f(x)\dot{f}(x) + f(x)^2 \dot{\sigma}^2(x) \right] dx \right)' u \right] du = 0. \end{aligned}$$

Finally, using integration by parts and the fact that $\ddot{\sigma}^2(x) = \ddot{v}(x) - 2\dot{g}(x)\dot{g}(x)' - 2g(x)\ddot{g}(x)$,

$$\begin{aligned} & h_n^2 \int_{\mathbb{R}^d} \dot{K}(u) \dot{K}(u)' (u'T_3(x)u) dx du \\ &= h_n^2 \int_{\mathbb{R}^d} \dot{K}(u) \dot{K}(u)' \left[u' \left(\int_{\mathbb{R}^d} \sigma^2(x) \ddot{f}(x) f(x) dx + \int_{\mathbb{R}^d} \dot{g}(x) \dot{g}(x)' f(x)^2 dx \right) u \right] du. \end{aligned}$$

Therefore, $\mathbb{V}[\bar{W}_n] = \binom{n}{2}^{-1} h_n^{-(d+2)} \Delta + \binom{n}{2}^{-1} h_n^{-d} \mathcal{V} + o(n^{-2} h_n^{-d})$. \blacksquare

Proof of Theorem 2. For part (i), note that $\hat{\vartheta}_{l_1, \dots, l_d, n}$ may be written as a n -varying U -statistic (assuming without loss of generality that s is even), given by

$$\hat{\vartheta}_{l_1, \dots, l_d, n} = \binom{n}{2}^{-1} \sum_{i=1}^{n-1} \sum_{j=i+1}^n u_1(z_i, z_j; b_n),$$

with (recall that $s = l_1 + \dots + l_d$)

$$u_1(z_i, z_j; b) = b^{-(d+1+s)} \left(\frac{\partial^s}{\partial x_1^{l_1} \dots \partial x_d^{l_d}} \dot{k}(x) \Big|_{x=(x_i-x_j)/b} \right) (y_i - y_j).$$

First, change of variables and integration by parts give

$$\mathbb{E}[u_1(z_i, z_j; b_n) | z_i] = \int_{\mathbb{R}^d} k(u) \left(\frac{\partial^s}{\partial x_1^{l_1} \dots \partial x_d^{l_d}} \dot{f}(x) \Big|_{x=x_i-ub_n} \quad y_i - \frac{\partial^s}{\partial x_1^{l_1} \dots \partial x_d^{l_d}} \dot{e}(x) \Big|_{x=x_i-ub_n} \right) du.$$

Second, a Taylor series expansion gives $\mathbb{E}[u_1(z_i, z_j; b_n)] = \vartheta_{l_1, \dots, l_d} + O(b_n^{\min(R, S)})$. Next, letting $\hat{\vartheta}_n = \hat{\vartheta}_{l_1, \dots, l_d, n}$ to save notation, a Hoeffding decomposition gives $\mathbb{V}[\hat{\vartheta}_n] = \mathbb{V}[\hat{\vartheta}_{1, n}] + \mathbb{V}[\hat{\vartheta}_{2, n}]$, where

$$\hat{\vartheta}_{1, n} = \frac{1}{n} \sum_{i=1}^n 2 [\mathbb{E}[u_1(z_i, z_j; b_n) | z_i] - \mathbb{E}[u_1(z_i, z_j; b_n)]],$$

and

$$\hat{\vartheta}_{2, n} = \binom{n}{2}^{-1} \sum_{i=1}^{n-1} \sum_{j=i+1}^n [u_1(z_i, z_j; b_n) - \mathbb{E}[u_1(z_i, z_j; b_n) | z_i] - \mathbb{E}[u_1(z_i, z_j; b_n) | z_j] + \mathbb{E}[u_1(z_i, z_j; b_n)]].$$

Finally, using standard calculations, $\mathbb{V}[\hat{\vartheta}_{1,n}] = O(n^{-1})$ and $\mathbb{V}[\hat{\vartheta}_{2,n}] = O(n^{-2}b_n^{-(d+2+2s)})$, and the conclusion follows by Markov's Inequality.

For part (ii), note that $\hat{\delta}_n$ is also a n -varying U -statistic, given by

$$\hat{\delta}_n = \binom{n}{2}^{-1} \sum_{i=1}^{n-1} \sum_{j=i+1}^n u_2(z_i, z_j; b_n), \quad u_2(z_i, z_j; b) = b^{-d} k\left(\frac{x_j - x_i}{b}\right) (y_i - y_j)^2.$$

First, change of variables gives

$$\mathbb{E}[u_2(z_i, z_j; b_n) | z_i] = \int_{\mathbb{R}^d} k(u) (y_i^2 f(x_i - ub_n) + v(x_i - ub_n) f(x_i - ub_n) - 2y_i e(x_i - ub_n)) du.$$

Second, a Taylor's expansion gives $\mathbb{E}[\hat{\delta}_n] = 2\mathbb{E}[\sigma^2(x) f(x)] + O(b_n^{\min(R, s+1+S)})$. Next, a Hoeffding decomposition gives $\mathbb{V}[\hat{\delta}_n] = \mathbb{V}[\hat{\delta}_{1,n}] + \mathbb{V}[\hat{\delta}_{2,n}]$, where

$$\hat{\delta}_{1,n} = \frac{1}{n} \sum_{i=1}^n 2 [\mathbb{E}[u_2(z_i, z_j; b_n) | z_i] - \mathbb{E}[u_2(z_i, z_j; b_n)]],$$

and

$$\hat{\delta}_{2,n} = \binom{n}{2}^{-1} \sum_{i=1}^{n-1} \sum_{j=i+1}^n [u_2(z_i, z_j; b_n) - \mathbb{E}[u_2(z_i, z_j; b_n) | z_i] - \mathbb{E}[u_2(z_i, z_j; b_n) | z_j] + \mathbb{E}[u_2(z_i, z_j; b_n)]].$$

Finally, using standard calculations, $\mathbb{V}[\hat{\delta}_{1,n}] = O(n^{-1})$ and $\mathbb{V}[\hat{\delta}_{2,n}] = O(n^{-2}b_n^d)$, and the conclusion follows by Markov's Inequality. \blacksquare

8. SUPPLEMENTAL MATERIAL

Further Simulation Results This document contains a comprehensive set of results from the Monte Carlo experiment summarized in Section 5. These results include all combinations of sample sizes ($n = 100, n = 400, n = 700$), dimension of regressors vector ($d = 2, d = 4$), and kernel orders ($P = 2, P = 4$).

REFERENCES

- ARADILLAS-LOPEZ, A., B. E. HONORE, AND J. L. POWELL (2007): “Pairwise Difference Estimation with Nonparametric Control Variables,” *International Economic Review*, 48, 1119–1158.
- CATTANEO, M. D., R. K. CRUMP, AND M. JANSSON (2009): “Small Bandwidth Asymptotics for Density-Weighted Average Derivatives,” working paper.
- (2010): “On the Validity of the Bootstrap for Density-Weighted Average Derivatives,” working paper.
- COPPEJANS, M., AND H. SIEG (2005): “Kernel Estimation of Average Derivatives and Differences,” *Journal of Business and Economic Statistics*, 23, 211–225.
- DEATON, A., AND S. NG (1998): “Parametric and Nonparametric Approaches to Price and Tax Reform,” *Journal of the American Statistical Association*, 93, 900–909.
- HÄRDLE, W., J. HART, J. MARRON, AND A. TSYBAKOV (1992): “Bandwidth Choice for Average Derivative Estimation,” *Journal of the American Statistical Association*, 87, 218–226.
- HÄRDLE, W., W. HILDENBRAND, AND M. JERISON (1991): “Empirical Evidence on the Law of Demand,” *Econometrica*, 59, 1525–1549.
- HÄRDLE, W., AND T. STOKER (1989): “Investigating Smooth Multiple Regression by the Method of Average Derivatives,” *Journal of the American Statistical Association*, 84, 986–995.
- HÄRDLE, W., AND A. TSYBAKOV (1993): “How Sensitive are Average Derivatives?,” *Journal of Econometrics*, 58, 31–48.
- HOROWITZ, J., AND W. HÄRDLE (1996): “Direct Semiparametric Estimation of Single-Index Models with Discrete Covariates,” *Journal of the American Statistical Association*, 91, 1632–1640.
- ICHIMURA, H., AND P. E. TODD (2007): “Implementing Nonparametric and Semiparametric Estimators,” in *Handbook of Econometrics, Volume VIB*, ed. by J. Heckman, and E. Leamer, pp. 5370–5468. Elsevier Science B.V.
- MATZKIN, R. L. (2007): “Nonparametric Identification,” in *Handbook of Econometrics, Volume VIB*, ed. by J. Heckman, and E. Leamer, pp. 5307–5368. Elsevier Science B.V.
- NEWHEY, W. K. (1994): “The Asymptotic Variance of Semiparametric Estimators,” *Econometrica*, 62, 1349–1382.

- NEWWEY, 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.
- (2005): “The Bootstrap and the Edgeworth Correction for Semiparametric Averaged Derivatives,” *Econometrica*, 73, 197–240.
- POWELL, J. L. (1994): “Estimation of Semiparametric Models,” in *Handbook of Econometrics, Volume IV*, ed. by R. Engle, and D. McFadden, pp. 2443–2521. Elsevier Science B.V.
- POWELL, J. L., J. H. STOCK, AND T. M. STOKER (1989): “Semiparametric Estimation of Index Coefficients,” *Econometrica*, 57, 1403–1430.
- POWELL, J. L., AND T. M. STOKER (1996): “Optimal Bandwidth Choice for Density-Weighted Averages,” *Journal of Econometrics*, 75, 291–316.
- ROBINSON, P. M. (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.
- WAND, M., AND M. JONES (1995): *Kernel Smoothing*. Chapman & Hall/CRC, Florida.

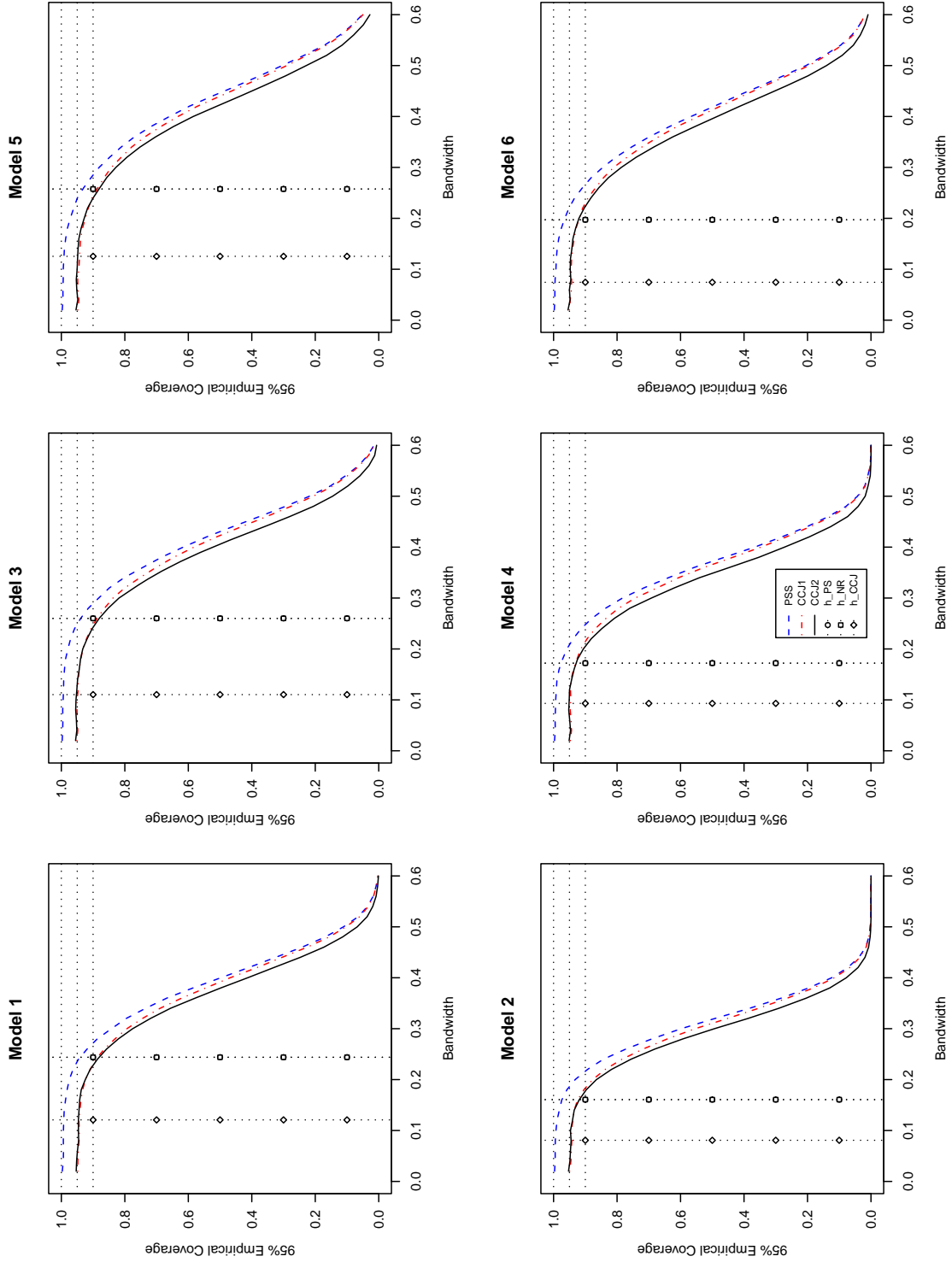


Figure 1: Empirical Coverage Rates for 95% Confidence Intervals: $d = 2$, $P = 2$, $n = 400$

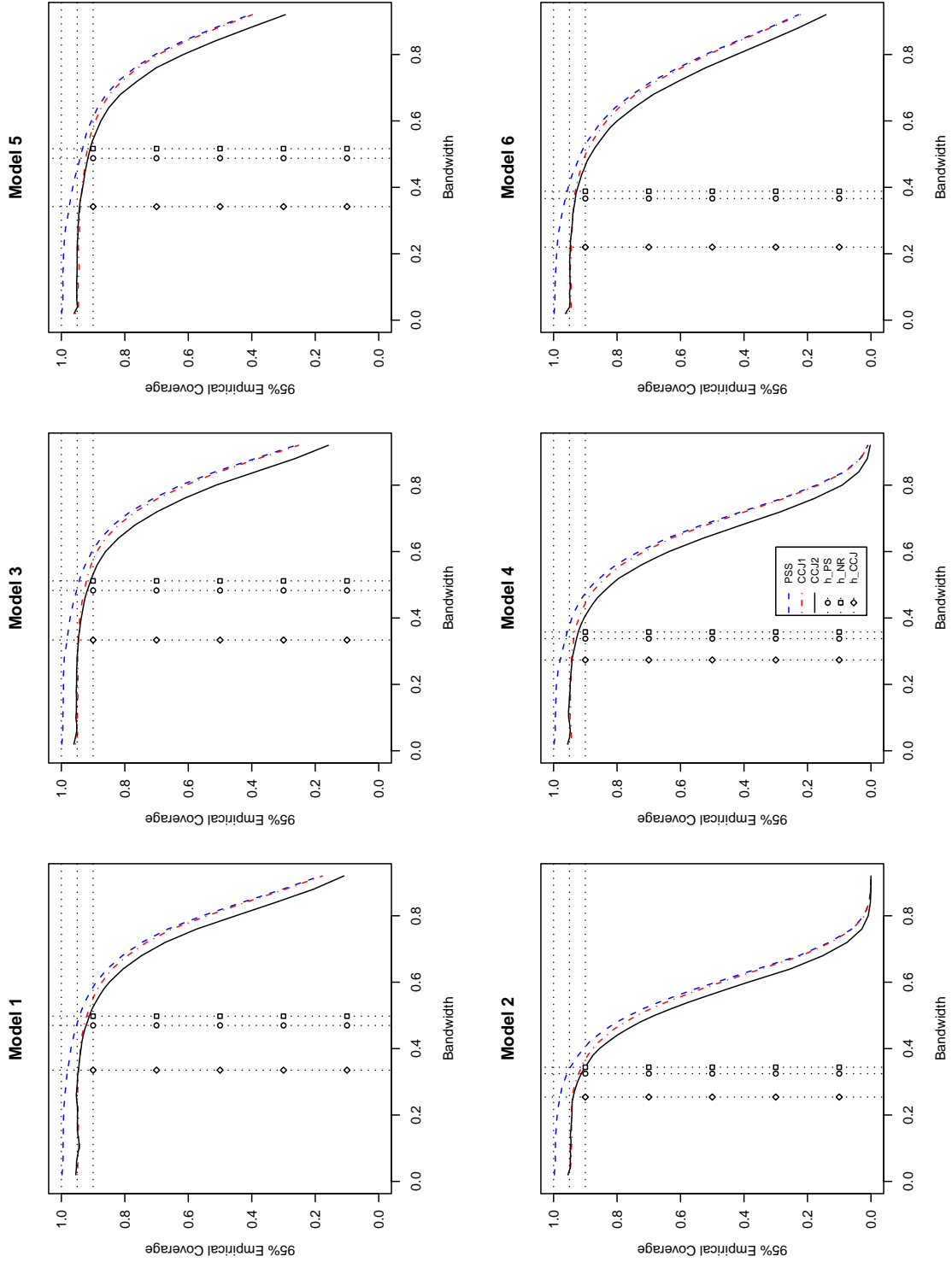


Figure 2: Empirical Coverage Rates for 95% Confidence Intervals: $d = 2$, $P = 4$, $n = 400$

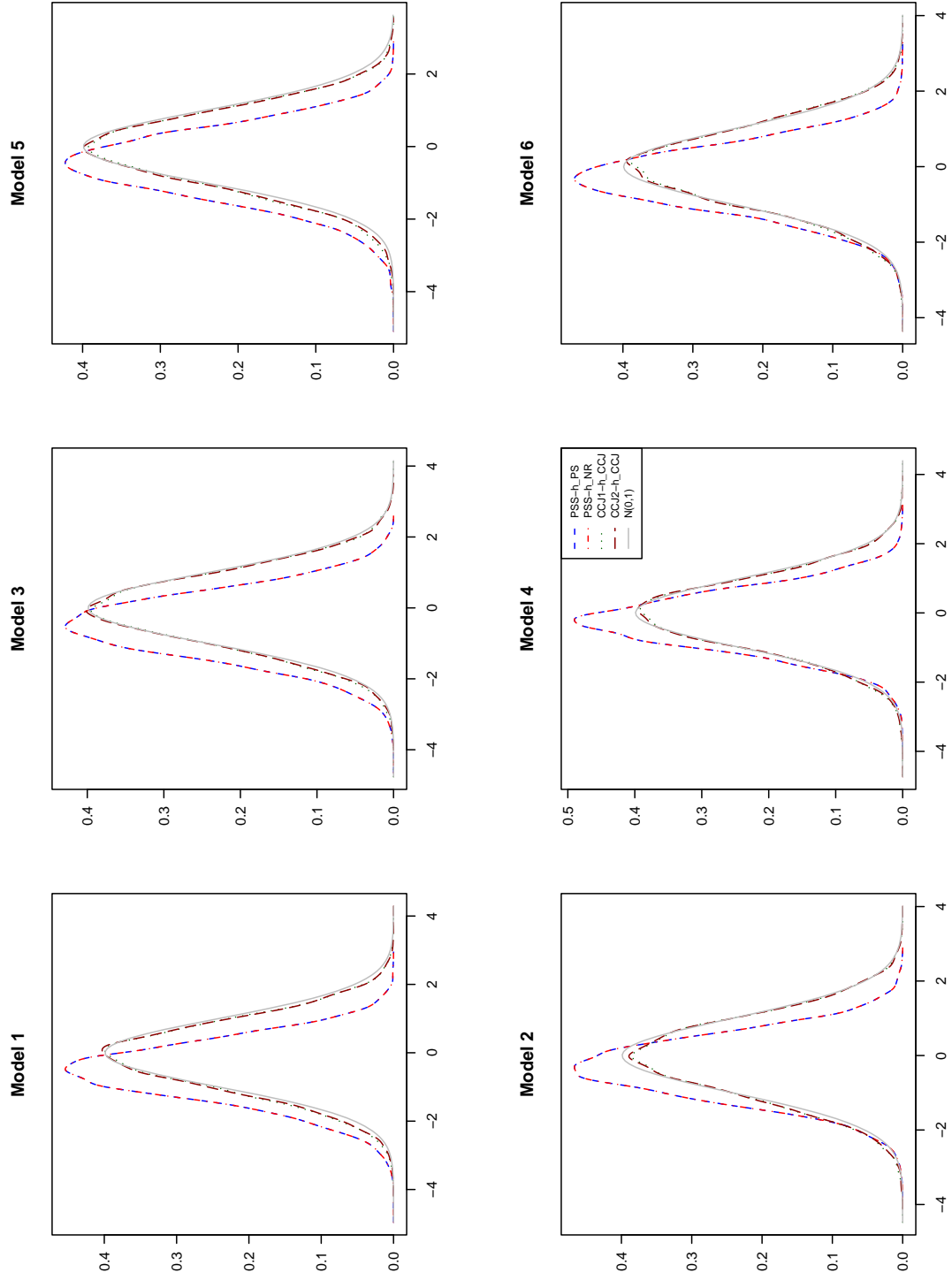


Figure 3: Empirical Gaussian Approximation with Population Bandwidth: $d = 2, P = 2, n = 400$

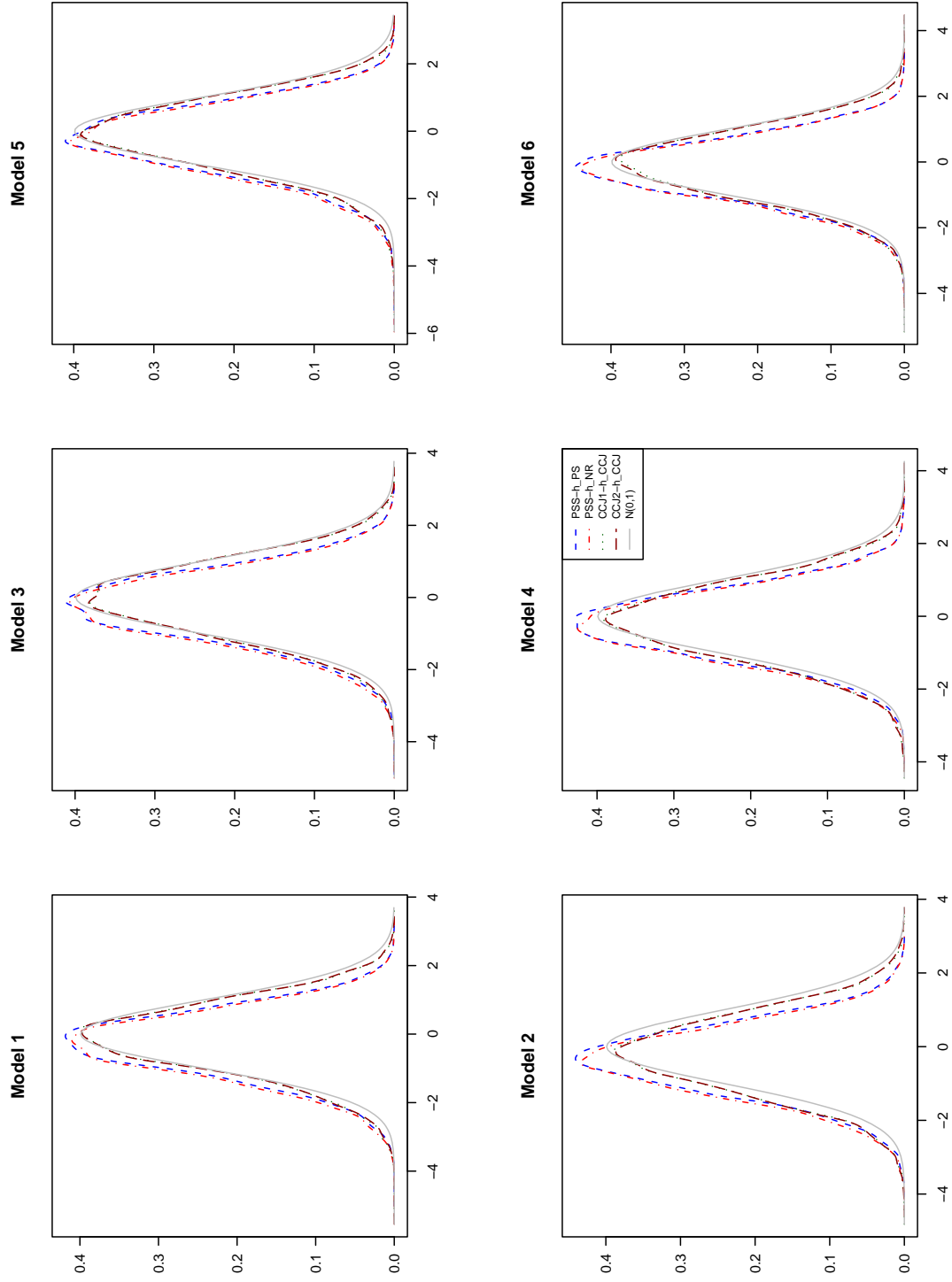


Figure 4: Empirical Gaussian Approximation with Population Bandwidth: $d = 2, P = 4, n = 400$

Table 1: Empirical Coverage Rates of 95% Confidence Intervals with Population Bandwidth: $d = 2, n = 400$.

		Model 1				Model 3				Model 5			
		BW	PSS	CCJ1	CCJ2	BW	PSS	CCJ1	CCJ2	BW	PSS	CCJ1	CCJ2
$P = 2$	h_{PS}^*	0.244	0.931	0.878	0.876	0.260	0.939	0.887	0.881	0.258	0.929	0.885	0.880
	h_{NR}^*	0.244	0.931	0.878	0.876	0.260	0.939	0.887	0.881	0.258	0.929	0.885	0.880
	h_{CCJ}^*	0.121	0.994	0.948	0.952	0.110	0.995	0.947	0.954	0.125	0.993	0.947	0.951
$P = 4$	h_{PS}^*	0.470	0.949	0.926	0.920	0.483	0.951	0.930	0.921	0.488	0.941	0.925	0.918
	h_{NR}^*	0.498	0.940	0.920	0.912	0.512	0.943	0.925	0.912	0.517	0.935	0.918	0.910
	h_{CCJ}^*	0.335	0.978	0.942	0.943	0.333	0.981	0.945	0.945	0.342	0.975	0.940	0.941

		Model 2				Model 4				Model 6			
		BW	PSS	CCJ1	CCJ2	BW	PSS	CCJ1	CCJ2	BW	PSS	CCJ1	CCJ2
$P = 2$	h_{PS}^*	0.161	0.970	0.921	0.916	0.172	0.978	0.935	0.931	0.197	0.968	0.920	0.919
	h_{NR}^*	0.161	0.970	0.921	0.916	0.172	0.978	0.935	0.931	0.197	0.968	0.920	0.919
	h_{CCJ}^*	0.081	0.994	0.944	0.946	0.093	0.993	0.947	0.949	0.074	0.995	0.946	0.950
$P = 4$	h_{PS}^*	0.325	0.951	0.917	0.907	0.338	0.964	0.938	0.927	0.366	0.962	0.936	0.931
	h_{NR}^*	0.344	0.940	0.909	0.897	0.358	0.958	0.933	0.922	0.388	0.956	0.931	0.926
	h_{CCJ}^*	0.254	0.977	0.940	0.939	0.273	0.982	0.945	0.943	0.220	0.990	0.945	0.949

Note: Column BW reports population bandwidths.

Table 2: Empirical Average Length of 95% Confidence Intervals with Population Bandwidth: $d = 2, n = 400$.

		Model 1				Model 3				Model 5			
		BIAS	PSS	CCJ1	CCJ2	BIAS	PSS	CCJ1	CCJ2	BIAS	PSS	CCJ1	CCJ2
$P = 2$	h_{PS}^*	0.005	0.036	0.031	0.030	0.002	0.013	0.011	0.011	0.003	0.022	0.019	0.019
	h_{NR}^*	0.005	0.036	0.031	0.030	0.002	0.013	0.011	0.011	0.003	0.022	0.019	0.019
	h_{CCJ}^*	0.002	0.110	0.080	0.080	0.000	0.053	0.038	0.038	0.001	0.064	0.047	0.047
$P = 4$	h_{PS}^*	0.183	3.096	2.842	2.755	0.050	1.184	1.091	1.043	0.090	1.981	1.849	1.782
	h_{NR}^*	0.221	2.971	2.762	2.657	0.065	1.133	1.057	1.001	0.112	1.909	1.802	1.723
	h_{CCJ}^*	0.070	4.302	3.566	3.556	0.002	1.720	1.417	1.409	0.021	2.696	2.279	2.268

		Model 2				Model 4				Model 6			
		BIAS	PSS	CCJ1	CCJ2	BIAS	PSS	CCJ1	CCJ2	BIAS	PSS	CCJ1	CCJ2
$P = 2$	h_{PS}^*	0.006	0.080	0.065	0.064	0.002	0.029	0.023	0.023	0.002	0.030	0.025	0.025
	h_{NR}^*	0.006	0.080	0.065	0.064	0.002	0.029	0.023	0.023	0.002	0.030	0.025	0.025
	h_{CCJ}^*	0.002	0.270	0.193	0.194	0.000	0.083	0.060	0.060	0.000	0.168	0.119	0.120
$P = 4$	h_{PS}^*	0.483	5.983	5.292	5.093	0.114	2.229	1.973	1.905	0.125	2.558	2.266	2.227
	h_{NR}^*	0.551	5.651	5.077	4.853	0.132	2.108	1.896	1.818	0.143	2.432	2.190	2.142
	h_{CCJ}^*	0.270	7.995	6.555	6.454	0.061	2.843	2.353	2.317	0.042	5.024	3.796	3.821

Note: Column BIAS reports absolute difference between average of $\hat{\theta}_n$ (across simulations) and θ_0 . All figures times 100.

Table 3: Empirical Coverage Rates of 95% Confidence Intervals with Estimated Bandwidth: $d = 2, n = 400$.

		Model 1				Model 3				Model 5			
		BW	PSS	CCJ1	CCJ2	BW	PSS	CCJ1	CCJ2	BW	PSS	CCJ1	CCJ2
$P = 2$	\hat{h}_{PS}	0.248	0.870	0.817	0.809	0.255	0.883	0.819	0.809	0.252	0.887	0.833	0.823
	\hat{h}_{NR}	0.248	0.870	0.817	0.809	0.255	0.883	0.819	0.809	0.252	0.887	0.833	0.823
	\hat{h}_{CCJ}	0.113	0.980	0.937	0.940	0.132	0.976	0.932	0.932	0.120	0.981	0.938	0.941
$P = 4$	\hat{h}_{PS}	0.290	0.978	0.921	0.924	0.290	0.980	0.922	0.923	0.290	0.979	0.923	0.926
	\hat{h}_{NR}	0.308	0.975	0.921	0.922	0.307	0.977	0.921	0.921	0.308	0.975	0.921	0.922
	\hat{h}_{CCJ}	0.187	0.993	0.949	0.953	0.198	0.994	0.948	0.954	0.192	0.995	0.949	0.954

		Model 2				Model 4				Model 6			
		BW	PSS	CCJ1	CCJ2	BW	PSS	CCJ1	CCJ2	BW	PSS	CCJ1	CCJ2
$P = 2$	\hat{h}_{PS}	0.201	0.858	0.796	0.780	0.208	0.903	0.851	0.838	0.212	0.920	0.860	0.854
	\hat{h}_{NR}	0.201	0.858	0.796	0.780	0.208	0.903	0.851	0.838	0.212	0.920	0.860	0.854
	\hat{h}_{CCJ}	0.104	0.972	0.916	0.919	0.119	0.973	0.929	0.930	0.105	0.986	0.943	0.946
$P = 4$	\hat{h}_{PS}	0.239	0.975	0.912	0.911	0.241	0.981	0.925	0.925	0.241	0.986	0.922	0.925
	\hat{h}_{NR}	0.254	0.967	0.908	0.906	0.255	0.976	0.925	0.921	0.256	0.981	0.919	0.921
	\hat{h}_{CCJ}	0.166	0.991	0.942	0.945	0.175	0.993	0.943	0.948	0.164	0.995	0.951	0.958

Note: Column BW reports sample mean of estimated bandwidths.