On the Effect of Bias Estimation on Coverage Accuracy in Nonparametric Inference*

Sebastian Calonico† Matias D. Cattaneo‡ Max H. Farrell§

November 24, 2014

PRELIMINARY AND INCOMPLETE
COMMENTS WELCOME

Abstract

Nonparametric methods play a central role in modern empirical work in economics. While they provide inference procedures that are more robust to parametric misspecification biases, they may be quite sensitive to the specific choices of tuning parameters. In this paper, we examine this problem for kernel density and local polynomial regression estimation. We study the effects of bias correction on confidence interval coverage and show formally that bias correction may be preferred to undersmoothing when the goal is to minimize the coverage error of the confidence interval. This result is established using a novel, yet simple, studentization approach for inference, which leads to a new way of constructing kernel-based nonparametric confidence intervals. Bandwidth selection is discussed, and shown to be very simple to implement. Indeed, we show that MSE-optimal bandwidths deliver the fastest coverage error rates when second-order kernels are employed. Our results have important implications for empirical work as they indicate that bias correction, coupled with appropriate standard errors and easy bandwidth choices, delivers confidence intervals with theoretical properties that can be substantially better than those constructed using ad-hoc undersmoothing. These results are established using Edgeworth expansions and illustrated with simulated data. We also discuss their connection to several important findings in the statistics and econometrics literatures.

*We thank Bruce Hansen, Michael Jansson, Francesca Molinari, and Ulrich Müller for thoughtful comments and suggestions, as well as participants at the 2014 Cowles Conference on Econometrics. Financial support from the National Science Foundation (SES 1357561) is gratefully acknowledged.

†Department of Economics, University of Miami.

‡Department of Economics, University of Michigan.

§Corresponding Author. Booth School of Business, University of Chicago.

1 Introduction

Nonparametric methods are nowadays widely employed in applied work in economics, as they provide point estimators and inference procedures that are more robust to parametric misspecification bias. Kernel-based methods are commonly used to estimate densities, conditional expectations, and related functions nonparametrically in a wide variety of empirical problems. These methods require specifying a bandwidth and, like all nonparametric estimators, their performance in applications crucially relies on how this tuning parameter is chosen. In particular, valid inference requires the delicate balancing act of selecting a bandwidth small enough to remove smoothing bias, yet large enough to ensure adequate precision. Tipping the balance in one direction or another can greatly skew results. We investigate this problem for kernel density and local polynomial regression estimation and demonstrate that by coupling explicit bias reduction with a novel, yet simple, studentization, inference can be made substantially more robust to bandwidth choice, greatly easing implementability.

Perhaps the most common bandwidth selection approach in practice is to minimize the asymptotic mean-square error (MSE) of the point estimator, and then use this bandwidth choice even when the ultimate goal is inference. So difficult is bandwidth selection perceived to be, that despite the fact that the MSE-optimal bandwidth leads to *invalid* confidence intervals, even asymptotically, this method is still advocated, and is the default in most popular software. Indeed, Hall and Kang (2001, p. 1446) write: "there is a growing belief that the most appropriate approach to constructing confidence regions is to estimate [the density] in a way that is optimal for pointwise accuracy.... [I]t has been argued that such an approach has advantages of clarity, simplicity and easy interpretation."

The underlying issue, as formalized below, is that bias must be removed for valid inference, and (in particular) the MSE-optimal bandwidth is "too large", leaving a bias that is still first order. Two main methods have been proposed to address this, undersmoothing and explicit bias correction. We seek to compare these two, and offer concrete ways to better implement the latter. Undersmoothing amounts to choosing a bandwidth smaller than would be optimal for point estimation then arguing that the bias is smaller than the variability of the estimator asymptotically, leading to valid confidence intervals. In practice this method often involves simply shrinking the MSE-optimal bandwidth by an ad-hoc amount. The second approach is to bias-correct the estimator with the explicit goal of removing the first-order bias that caused the invalidity of the inference procedure in the first place.

It has been believed for some time that undersmoothing is preferable for two reasons. First, prior theoretical studies showed inferior asymptotic coverage properties of bias-corrected confidence intervals. Second, implementation of bias correction is more complex as a second bandwidth is required (for the bias correction itself), deterring practitioners. See Hall (1992b), Neumann (1997), Horowitz (2001), and Hall and Horowitz (2013). However, we show theoretically that bias correction can be as good, or better in some cases, than undersmoothing, if the new standard error formula that we derive is used. Further, our main findings have important implications for empirical work

because the resulting confidence intervals are more robust to bandwidth choice, including to the secondary bandwidth for bias estimation. Indeed, we recommend setting the two bandwidths equal; a simple, automatic choice that performs very well and is even optimal in certain senses. Our results justify using the popular MSE-optimal bandwidth choice (as do Hall and Horowitz (2013) with an entirely different approach), and furthermore we show that when second-order kernels are used (by far the most common choice), the coverage error vanishes at the "best" possible rate.[1] When higher-order kernels are used, we show that the MSE-optimal bandwidth leads to confidence intervals with quite suboptimal coverage error rates, though coverage is always asymptotically correct, and thus propose a simple adjustment to the MSE-optimal bandwidth to improve the coverage error rates. (See Section 3.4 for details.) In addition, we study the important related issue of asymptotic length of the new confidence intervals.

Our formal comparisons of the two methods are based on Edgeworth expansions for the conventional undersmoothed estimator and the bias-corrected estimator with and without the novel studentization formula, for both density estimation and local polynomial regression, both at interior and boundary points. We prove that explicit bias correction, coupled with our proposed standard errors, yields coverage that is as accurate, or better, than the best possible undersmoothing approach. Loosely speaking, this improvement is possible because explicit bias correction can remove more bias than undersmoothing, while our proposed standard errors capture not only the variability of the original estimator but also the additional variability introduced by the bias correction. Our findings contrast with well established recommendations: Hall (1992b) used Edgeworth expansions to show that undersmoothing produces more accurate intervals than explicit bias correction in the density case and Neumann (1997) repeated this finding for nonparametric regression. Their expansions, however, crucially relied on the assumption that the bias correction was asymptotically first-order negligible. (See Remark 6 below for details.) In contrast, we allow the bias estimator to potentially have a first-order impact on the distributional approximation, an alternative asymptotic experiment designed to more closely mimic the finite-sample behavior of bias correction.

Our standard error formulas are based on fixed-n calculations, as opposed to asymptotic ones, which also turns out to be important. We show that using asymptotic variance formulas can introduce further errors in coverage probability, with particularly negative consequences at boundary points. This turns out to be at the heart of the "quite unexpected" conclusion found by Chen and Qin (2002, Abstract) that local polynomial based confidence intervals are not boundary-adaptive in coverage error: we prove that this is not the case with proper studentization. Thus, as a by-product of our main theoretical work, we also establish higher-order boundary carpentry of local-polynomial-based confidence intervals at a boundary point whenever an appropriate fixed-n standard error formula is employed, a result that is of independent (but related) interest.

Our paper is connected with the well established statistical literature on nonparametric smoothing; see, among many others, Wand and Jones (1995), Fan and Gijbels (1996), and Ruppert, Wand,

[1] In this paper, we take the "best" possible coverage error rate to be the fastest achievable rate of contraction of coverage errors as obtain from the Edgeworgth expansions developed in Section 3.3 (Corollary 1). An alternative notion of "best" is briefly discussed in Section 3.6.

and Carroll (2009) for reviews. In econometrics, nonparametric smoothing also plays a crucial role; see, for example, the recent reviews of Ichimura and Todd (2007), Li and Racine (2007), and Horowitz (2009). Even more recently, there appears to be a renewed interest on (possibly non-standard) distributional approximations and inference procedures in nonparametrics econometrics. Also studying kernel estimation, Calonico, Cattaneo, and Titiunik (2014) propose an alternative first-order asymptotic approach to account for the effect of bias-correction in inference at a boundary point. In concurrent work, Hansen (2014) develops first-order robustness approaches to account for the effect of smoothing bias in series-based inference, while Armstrong (2014) and Armstrong and Kolesár (2014) discuss smoothness adaptive inference. The main findings reported in this paper are also in qualitative agreement with those in Jansson (2004) and Sun, Phillips, and Jin (2008), who studied the effects on coverage error of using the fixed-b asymptotic approximations of Kiefer and Vogelsang (2005) to conduct heteroskedasticity autocorrelation robust inference.

The rest of the paper proceeds as follows. The next two sections treat density estimation in detail, because the main ideas and results, and their implications, can be made clear with relative ease: Section 2 formalizes the basic ideas and main questions, while Section 3, the bulk of the paper, formally states the main results on error in coverage probability and its relationship to bias reduction and discusses first-order asymptotic properties, bandwidth choice, and interval length. Local polynomial estimation is taken up in Section 4, and there we spend additional time on standard errors and inference at the boundary due to its importance in empirical work (e.g., regression discontinuity designs). Numerical evidence is offered in Section 5, while Section 6 concludes. Technical material is collected in the supplemental appendix.

2 Setup and Basic Ideas

In this section and the next, we elaborate the main ideas and conclusions of the paper focused on the simple case of inference on the density at an interior point, which requires relatively little notation. Suppose we have a random sample $\{X_i : 1 \leq i \leq n\}$ from a large population that is continuously distributed with Lebesgue density f. (In the main paper we treat the univariate covariate case to minimize notation; the supplement summarizes how our results extend naturally to derivative estimation and $\dim(X_i) > 1$.) The classical kernel-based estimator of $f(x)$ is

$$\hat{f}(x) = \frac{1}{nh} \sum_{i=1}^{n} K\left(\frac{x - X_i}{h}\right)$$

for a kernel function K that integrates to 1 on the appropriate support and bandwidth $h \to 0$ as $n \to \infty$. For a comprehensive review, see Wand and Jones (1995). Under appropriate smoothness assumptions, it is straightforward to find that the bias is given by

$$\mathbb{E}[\hat{f}(x)] - f(x) = h^r f^{(r)}(x) \mu_{K,r} + o(h^r) \tag{1}$$

if a kernel of order r is used, where $f^{(r)}(x) := \partial^r f(x)/\partial x^r$. The leading bias has three factors: (i) a rate of convergence, depending on the bandwidth, (ii) a derivative of the unknown function, and (iii) a known quantity. The local polynomial bias is conceptually identical, see Eqn. (6). The fundamental question we seek to answer is this: if this is the bias, is one better off estimating the leading bias (explicit bias correction) or choosing $h \to 0$ fast enough to render the bias negligible (undersmoothing) when forming nonparametric confidence intervals?

Traditionally, Studentized statistics based on undersmoothing and explicit bias correction are, respectively,

$$T_{\mathtt{us}}(x) := \frac{\hat{f}(x) - f(x)}{\hat{\mathbb{V}}[\hat{f}(x)]^{1/2}} \qquad \text{and} \qquad T_{\mathtt{bc}}(x) := \frac{\hat{f}(x) - h^r \hat{f}^{(r)}(x)\mu_{K,r} - f(x)}{\hat{\mathbb{V}}[\hat{f}(x)]^{1/2}},$$

for a suitable estimator $\hat{\mathbb{V}}[\hat{f}(x)]$ and a (kernel-based) estimator $\hat{f}^{(r)}(x)$. These are the two statistics compared in the influential paper of Hall (1992b).

From the form of these statistics, two points are already clear. First, the numerator of $T_{\mathtt{us}}$ relies on choosing h vanishing fast enough so that the leading bias is asymptotically negligible after scaling, whereas explicit bias correction requires less to be removed by choice of bandwidth due to the manual estimation of the leading bias. Second, $T_{\mathtt{bc}}$ requires that the variance of $h^r \hat{f}^{(r)}(x)\mu_{K,r}$ must be first-order asymptotically negligible: $\hat{\mathbb{V}}[\hat{f}(x)]$ in the denominator only accounts for the variance of the main estimate, but $\hat{f}^{(r)}(x)$, being a kernel-based estimator, naturally has a variance controlled by its bandwidth. Thus, even though $\hat{\mathbb{V}}[\hat{f}(x)]$ is based on a fixed-n calculation, the variance of the numerator of $T_{\mathtt{bc}}$ only coincides with the denominator in the limit. Under this regime, Hall (1992b) employed Edgeworth expansions to show that the reduction in bias is too expensive in terms noise, and therefore argued that undersmoothing dominates explicit bias correction.

On the other hand, in this paper we argue that there need not be such a "mismatch" between the numerator of the bias-corrected statistic and the Studentization. We thus consider a third option corresponding to the idea of capturing the finite sample variability of $\hat{f}^{(r)}(x)$ directly:

$$T_{\mathtt{rbc}} := \frac{\hat{f}(x) - h^r \hat{f}^{(r)}(x)\mu_{K,r} - f(x)}{\hat{\mathbb{V}}[\hat{f}(x) - h^r \hat{f}^{(r)}(x)\mu_{K,r}]^{1/2}}.$$

That is, our proposed standard error estimate is based on a fixed-n calculation that captures the variability of both terms $\hat{f}(x)$ and $\hat{f}^{(r)}(x)$, as well as their covariance. This alternative approach to construct a test statistic, as we argue below, can be justified by using an alternative asymptotic experiment that allows (but does not require) the bias correction to be of first-order importance, after rescaling, and we indeed show that doing so yields more accurate confidence intervals (i.e., higher-order corrections).

3 Density Estimation

The present section formalizes the main conclusions as described above, continuing with the density case. The plan of presentation is as follows. We first make precise the leading bias to be removed and the approaches to valid inference we compare. We then give a terse treatment of first-order distributional properties, before turning to a lengthier discussion of coverage error and its implications for bandwidth choice and interval length. Particular attention will be paid to the smoothness assumptions placed on f. We will omit the dependence on the point x when there is no confusion.

The following two conditions, respectively governing the data generating process and the properties of the kernel functions K and L, are standard in nonparametrics, and are sufficient for characterizing the biases and first-order convergence. For any kernel K and integer k, define

$$\mu_{K,k} = \frac{(-1)^k}{k!} \int u^k K(u) du \qquad \text{and} \qquad \vartheta_{K,k} = \int K(u)^k du.$$

Assumption 3.1 (Data-generating process)**.** *$\{X_1, \ldots, X_n\}$ is a random sample with an absolutely continuous distribution with Lebesgue density f. In a neighborhood of x, $f > 0$, f is S-times continuously differentiable with bounded derivatives $f^{(k)}$, $k = 1, 2, \cdots, S$, and $f^{(S)}$ is Hölder continuous with exponent ς.*

Assumption 3.2 (Kernels)**.**

(a) The kernels K and L are bounded, even functions with compact support $[-1, 1]$, and are of order $r \geq 2$ and $s \geq 2$, respectively, where r and s are even. That is, $\mu_{K,0} = 1$, $\mu_{K,q} = 0$ for $1 \leq q < r$, and $\mu_{K,r} \neq 0$ and bounded, and similarly for $\mu_{L,q}$ with s in place of r. Further, L is r-times continuously differentiable.

(b) For all integers k and l such that $k + l = r - 1$, $f^{(k)}(x_0) L^{(l)}((x_0 - x)/b) = 0$ for x_0 in the boundary of the support.

These two assumptions are essentially identical to those imposed by Hall (1991, 1992b). The precision of the Hölder condition is necessary to show how kernel order and bias correction interact with smoothness limits, and to quantify the fastest possible rates of decay in coverage error (given the underlying smoothness and the kernel order). The restriction that the orders of the kernels be even is not crucial for our results conceptually, but is required in order to characterize the leading bias terms. The boundary conditions of Assumption 3.2 are needed for the derivative estimation inherent in bias correction, and is satisfied for instance if the support of f is the whole real line.

Under these conditions we can make precise the bias of our estimators. It is important to be precise at this point, as the interaction between the order of the kernels K and L and the smoothness of the density will determine the rates of decay of coverage error and the feasibility of bandwidth selection. In particular, we must take care when the smoothness is fully utilized by choosing higher-order kernels, as this will lead to the best-possible rates, at the expense of feasibility. For the initial

density estimator, we have:

$$\mathbb{E}[\hat{f}] - f = \begin{cases} h^r f^{(r)} \mu_{K,r} + h^{r+2} f^{(r+2)} \mu_{K,r+2} + o(h^{r+2}) & \text{if } r \leq S-2 \\ h^r f^{(r)} \mu_{K,r} + O(h^{S+\varsigma}) & \text{if } r \in \{S-1, S\} \\ 0 + O(h^{S+\varsigma}) & \text{if } r > S. \end{cases} \tag{2}$$

In the first case the leading bias and the next term can be fully characterized, which will allow for bias correction with feasible bandwidth selection. In the second case, the leading bias is still characterizable, but the remainder is not ($r = S$ and $r = S-1$ yield the same remainder as r is even). In the third term the bias is only known up to order. Let B_f denote the leading bias, with the convention, emphasized in the third case above, that $B_f = 0$ for $r > S$. In any case, we can form the estimate

$$\hat{B}_f = h^r \hat{f}^{(r)} \mu_{K,r}, \qquad \text{where} \qquad \hat{f}^{(r)}(x) = \frac{1}{nb^{1+r}} \sum_{i=1}^{n} L^{(r)}\left(\frac{x - X_i}{b}\right),$$

for another kernel $L(\cdot)$ and bandwidth $b \to 0$ as $n \to \infty$. Note that $\hat{B}_f$ can, and will, take this form for any value of r, and in particular for $r > S$. The bias of $\hat{f} - \hat{B}_f$ has two pieces, the bias from approximating $\mathbb{E}[\hat{f}] - f$ by B_f and the bias of $\hat{B}_f$ as an estimator of B_f, as follows:

$$\mathbb{E}[\hat{f} - \hat{B}_f] - f = \begin{cases} h^{r+2} f^{(r+2)} \mu_{K,r+2} + h^r b^s f^{(r+s)} \mu_{K,r} \mu_{L,s} + o(h^{r+2} + h^r b^s) & \text{if } r + s \leq S \\ h^{r+2} f^{(r+2)} \mu_{K,r+2} + O(h^r b^{S-r+\varsigma}) + o(h^{r+2}) & \text{if } 2 \leq S - r < s \\ O(h^{S+\varsigma}) + O(h^r b^{S-r+\varsigma}) & \text{if } r \in \{S-1, S\} \\ O(h^{S+\varsigma}) + O(h^r b^{S-r}) & \text{if } r > S. \end{cases} \tag{3}$$

A key quantity is the ratio of the two bandwidths h and b, given by $\rho := h/b$. If this sequence vanishes asymptotically, then the bias correction is first-order negligible. On the other hand, if ρ converges to a positive, finite limit then the bias correction will be first order important. Our results allow for both cases, unlike prior work. (Remark 3 discusses $\rho \to \infty$.) This asymptotic experiment is designed to more accurately capture the fact that in finite samples the effect of bias correction is certainly not negligible, as make explicit in the sequel.

To complete the definitions of the three Studentized statistics $T_{\mathtt{us}}$, $T_{\mathtt{bc}}$, and $T_{\mathtt{rbc}}$, we formalize the variance estimators $\hat{\mathbb{V}}[\hat{f}]$ and $\hat{\mathbb{V}}[\hat{f} - \hat{B}_f]$ mentioned above. Straightforward calculations give

$$\sigma^2_{\mathtt{us}} := nh\mathbb{V}[\hat{f}] = \frac{1}{h}\left\{\mathbb{E}\left[K\left(\frac{x - X_i}{h}\right)^2\right] - \mathbb{E}\left[K\left(\frac{x - X_i}{h}\right)\right]^2\right\},$$

which is nonasymptotic: n and h are fixed. We use the natural estimator

$$\hat{\sigma}^2_{\mathtt{us}} = \frac{1}{h}\left\{\frac{1}{n}\sum_{i=1}^{n}\left[K\left(\frac{x - X_i}{h}\right)^2\right] - \frac{1}{n}\sum_{i=1}^{n}\left[K\left(\frac{x - X_i}{h}\right)\right]^2\right\},$$

which coincides with the estimator used in Hall (1992b). The collective variance of the density estimate and bias correction, $\mathbb{V}[\hat{f} - \hat{B}_f]$, will fit the same pattern, utilizing the ratio ρ. First, note that we may write

$$\hat{f} - h^r \hat{f}^{(r)} \mu_{K,r} = \frac{1}{nh} \sum_{i=1}^{n} M\left(\frac{x - X_i}{h}\right), \qquad M(u) := K(u) - \rho^{1+r} L^{(r)}(\rho u) \mu_{K,r}. \tag{4}$$

Written thusly, the only difference between $\hat{f}$ and $\hat{f} - \hat{B}_f$ is the change in "kernel" from K (a fixed function) to M (an n-varying, higher-order kernel), a link that will be useful at times to explain the intuition behind our results and simplifies some proofs.[2] With this notation, we define the variance

$$\sigma^2_{\mathtt{rbc}} := nh\mathbb{V}[\hat{f} - \hat{B}_f] = \frac{1}{h}\left\{\mathbb{E}\left[M\left(\frac{x - X_i}{h}\right)^2\right] - \mathbb{E}\left[M\left(\frac{x - X_i}{h}\right)\right]^2\right\},$$

and its estimator

$$\hat{\sigma}^2_{\mathtt{rbc}} = \frac{1}{h}\left\{\frac{1}{n}\sum_{i=1}^{n}\left[M\left(\frac{x - X_i}{h}\right)^2\right] - \frac{1}{n}\sum_{i=1}^{n}\left[M\left(\frac{x - X_i}{h}\right)\right]^2\right\}.$$

From Eqn. (4), it is clear that if $\rho \to 0$, the second term of M is dominated by the first, i.e. the bias correction is first-order negligible. However, in finite samples (and asymptotically if ρ is nonvanishing), the shape of the kernel M depends on ρ (a fact explored in Section 3.5 below) and $\sigma^2_{\mathtt{rbc}}$ captures this dependence explicitly. Notice that if $\rho \to 0$, making the bias correction higher-order, then $\sigma^2_{\mathtt{us}}$ and $\sigma^2_{\mathtt{rbc}}$ (and their estimators) will be first-order equivalent, but not higher-order equivalent. This is exactly the sense in which traditional bias correction relies on an asymptotic variance, instead of a fixed-n one, and pays the price in coverage error.

Formally then, the three statistics of interest are

$$T_{\mathtt{us}} = \frac{\sqrt{nh}(\hat{f} - f)}{\hat{\sigma}_{\mathtt{us}}}, \qquad T_{\mathtt{bc}} = \frac{\sqrt{nh}(\hat{f} - \hat{B}_f - f)}{\hat{\sigma}_{\mathtt{us}}}, \qquad T_{\mathtt{rbc}} = \frac{\sqrt{nh}(\hat{f} - \hat{B}_f - f)}{\hat{\sigma}_{\mathtt{rbc}}}.$$

This notation makes abundantly clear the "mismatch" in the traditional bias correction approach, $T_{\mathtt{bc}}$. We restrict attention to bounded ρ (see Remark 3), and hence the statistics share the common scaling of $\sqrt{nh}$ that is made explicit here, in contrast to the heuristic Introduction, because explicit scaling is more natural in the Edgeworth expansions.

Remark 1 (Asymptotic variances)**.** $T_{\mathtt{bc}}$ relies on an asymptotic variance, but one derived nonetheless from a fixed-n calculations. For both $\sigma^2_{\mathtt{us}}$ and $\sigma^2_{\mathtt{rbc}}$, there are also asymptotic forms that are first-order valid, but inadvisable to use. In particular, as $h \to 0$, $\sigma^2_{\mathtt{us}} \to f\vartheta_{K,2}$, and hence for first-order purposes we can replace $\hat{\sigma}^2_{\mathtt{us}}$ with $\hat{f}\vartheta_{K,2}$ ($\sigma^2_{\mathtt{rbc}}$ behaves similarly). However, doing so

[2] This observation is not new in the nonparametric density estimation literature; see, Fan and Hu (1992), Jones and Foster (1993), and Jones (1995), among others, and Jones and Signorini (1997) for simulation evidence comparing various "higher-order" methods.

will have consequences in finite samples that manifest as additional error terms in the Edgeworth expansion (see Hall, 1992a, p. 209, for discussion). In some cases, these additional terms can be the dominant error terms. In particular, for local polynomials, using an asymptotic instead of a fixed-n variance will lead to unnnecessary coverage error at boundary points. ■

Remark 2 (Confidence bands)**.** Our results concern pointwise intervals, but it may be of interest to extended them to simultaneous confidence bands, as discussed by, e.g., Hall (1993), Xia (1998), Gine and Nickl (2010) and Chernozhukov, Chetverikov, and Kato (2013). Indeed, employing the techniques of the last paper, we conjecture that our results could be demonstrated for bands. ■

3.1 First-Order Properties

Before presenting the higher-order expansions of coverage probability, it is worthwhile to discuss the first-order properties of the statistics $T_{\mathtt{us}}$, $T_{\mathtt{bc}}$, and $T_{\mathtt{rbc}}$, as formalized in the following result. Define the scaled biases $\eta_{\mathtt{us}} = \sqrt{nh}(\mathbb{E}[\hat{f}] - f)$ and $\eta_{\mathtt{bc}} = \sqrt{nh}(\mathbb{E}[\hat{f} - \hat{B}_f] - f)$.

Theorem 1 (First-order properties)**.** *Let Assumptions 3.1 and 3.2 hold, and* $nh \to \infty$.

(a) If $\eta_{\mathtt{us}} \to 0$*, then* $T_{\mathtt{us}} \to_d \mathcal{N}(0,1)$.

(b) If $\eta_{\mathtt{bc}} \to 0$ *and* $\rho \to 0$*, then* $T_{\mathtt{bc}} \to_d \mathcal{N}(0,1)$.

(c) If $\eta_{\mathtt{bc}} \to 0$ *and* $\rho \to \bar{\rho} < \infty$*, then* $T_{\mathtt{rbc}} \to_d \mathcal{N}(0,1)$.

Parts (a) and (b) are standard in the nonparametrics literature corresponding, respectively, to the undersmoothing and traditional explicit bias correction cases. Part (c) is nonstandard, and was put forth for local polynomial regression by Calonico, Cattaneo, and Titiunik (2014). The first-order convergence of $T_{\mathtt{rbc}}$ requires strictly weaker bandwidth conditions than the other results for a given r, which in turn suggests a potentially more robust distributional approximation. (We restrict attention to convergent ρ sequences for clarity of discussion, as there is little gained from considering non-convergent sequences $\rho = h/b$.) The bandwidth conditions on h and b are behind the generic assumption of each part of Theorem 1 that the scaled bias vanishes and can be read off of Equations (2) and (3). To see this, consider two leading cases.

First, suppose we take $r = S$. Then $\mathbb{E}[\hat{f}] - f = O(h^S)$ and part (a) requires $\sqrt{nh}h^S \to 0$. However, from (3), $\mathbb{E}[\hat{f} - \hat{B}_f] - f = O(h^S(h^s + b^s))$, and thus $\eta_{\mathtt{bc}} = O(\sqrt{nh}h^S(h^s + b^s))$. Second, suppose $r + s \leq S - 2$, in which case part (a) requires $\sqrt{nh}h^r \to 0$ whereas parts (b) and (c) require only $\sqrt{nh}h^r(h^2 \vee b^s) \to 0$. In both cases, $T_{\mathtt{bc}}$ and $T_{\mathtt{rbc}}$ allow for $\sqrt{nh}h^r \not\to 0$ or $b \not\to 0$, but not both, and thus have weaker bias requirements than $T_{\mathtt{us}}$. The coverage error in these two cases will be explored in detail below, but this formalizes the intuition that bias correction can remove more bias than undersmoothing.

However, bias correction requires a choice of $\rho = h/b$, related to variance. One easily finds that $\mathbb{V}[\hat{f}^{(r)}] = O(n^{-1}b^{-1-2r})$ and hence $\mathbb{V}[\sqrt{nh}\hat{B}_f] = O(\rho^{1+2r})$, so only if $\rho \to 0$ is the variance of $\hat{B}_f$ higher order, allowing for weak convergence of $T_{\mathtt{bc}}$. But $T_{\mathtt{rbc}}$ does not suffer from this requirement

because of the proposed, alternative Studentization. From a first-order point of view, traditional bias correction allows for a larger class of sequences h, but requires a delicate choice of ρ (or b), and Hall (1992b) shows that this constraint prevents $T_{\mathtt{bc}}$ from delivering improved inference. From this point of view, our novel standard errors effectively remove these constraints, allowing for improvements in bias to carry over to improvements in inference. These gains are quantified below.

Remark 3 ($\rho \to \infty$)**.** The case $\bar{\rho} = \infty$ can also covered by Theorem 1(c) with an even weaker bias rate restriction: $\eta_{\mathtt{bc}} = o(\rho^{1/2+r})$. In this case $\hat{B}_f$ dominates the first-order approximation and the rate of convergence is no longer $\sqrt{nh}$ (hence $nb \to \infty$ is needed now), but nonetheless $T_{\mathtt{rbc}}$ stills converges to standard Normal by virtue of the choice of $\sigma^2_{\mathtt{rbc}}$. From a coverage point of view, however, there seems to be no advantage: the bias rate can not be improved due to the second bias term ($\mathbb{E}[\hat{f}] - f - B_f$), and the variance can only be inflated. Thus, we restrict to bounded $\bar{\rho}$. ■

3.2 Generic Higher Order Expansions of Coverage Error

We now turn to Edgeworth expansions to formalize the improvements in inference, by examining the coverage accuracy of confidence intervals based on the above Normal approximation. To be concrete, we seek to compare the error in coverage probability of the following Gaussian-based $(1-\alpha)\%$ symmetric confidence intervals:

$$I_{\mathtt{us}} = \left[\hat{f} - z_{1-\frac{\alpha}{2}}\frac{\hat{\sigma}_{\mathtt{us}}}{\sqrt{nh}},\quad \hat{f} - z_{\frac{\alpha}{2}}\frac{\hat{\sigma}_{\mathtt{us}}}{\sqrt{nh}}\right], \qquad I_{\mathtt{bc}} = \left[\hat{f} - \hat{B}_f - z_{1-\frac{\alpha}{2}}\frac{\hat{\sigma}_{\mathtt{us}}}{\sqrt{nh}},\quad \hat{f} - \hat{B}_f - z_{\frac{\alpha}{2}}\frac{\hat{\sigma}_{\mathtt{us}}}{\sqrt{nh}}\right],$$

and

$$I_{\mathtt{rbc}} = \left[\hat{f} - \hat{B}_f - z_{1-\frac{\alpha}{2}}\frac{\hat{\sigma}_{\mathtt{rbc}}}{\sqrt{nh}},\quad \hat{f} - \hat{B}_f - z_{\frac{\alpha}{2}}\frac{\hat{\sigma}_{\mathtt{rbc}}}{\sqrt{nh}}\right].$$

For higher order expansions, the conditions of Theorem 1 must be augmented with the n-varying analogue of Cramér's condition.

Assumption 3.3 (Cramér's Condition)**.** *For each $\delta > 0$ and all sufficiently small h*

$$\sup_{t\in\mathbb{R}^2, t_1^2+t_2^2>\delta}\left|\int \exp\{i(t_1 M(u) + t_2 M(u)^2)\} f(x-uh)du\right| \leq 1 - C(x,\delta)h,$$

where $C(x,\delta) > 0$ is a fixed constant and $i = \sqrt{-1}$.

This technical assumption is essentially Lemma 4.1 in Hall (1991), which he establishes by way of restricting the class of kernels, essentially requiring piecewise monotonicity of $K^{(1)}$ and $L^{(r+1)}$.[3]

The following result gives generic formulas for the coverage error of the three confidence intervals introduced above, which follow from the valid Edgeworth expansions for the distribution functions

[3] A similar condition is imposed in virtually all papers employing Edgeworth expansions techniques; see, e.g., Hall and Horowitz (1996), Andrews (2002), Nishiyama and Robinson (2005), Kline and Santos (2012), among others.

of the three statistics established in the supplemental appendix. For any kernel K and quantile z, define

$$q_1(z;K) = \vartheta_{K,2}^{-2}\vartheta_{K,4}(z^3-3z)/6 - \vartheta_{K,2}^{-3}\vartheta_{K,3}^2[2z^3/3 + (z^5-10z^3+15z)/9],$$
$$q_2(z;K) = -\vartheta_{K,2}^{-1}(z), \qquad \text{and} \qquad q_3(z;K) = \vartheta_{K,2}^{-2}\vartheta_{K,3}(2z^3/3).$$

All that is conceptually important is that these functions are known, odd polynomials in z with coefficients that depend only on integrals of the kernel, but not on the sample size. Let $\phi(z)$ be the standard Normal density.

Theorem 2 (Coverage error)**.** *Let Assumptions 3.1, 3.2, and 3.3 hold, and* $nh/\log(n) \to \infty$.

(a) If $\eta_{\mathtt{us}} \to 0$, *then*

$$\mathbb{P}[f \in I_{\mathtt{us}}] = 1-\alpha + \left\{ \frac{1}{nh} q_1(z_{\frac{\alpha}{2}};K) + \eta_{\mathtt{us}}^2 q_2(z_{\frac{\alpha}{2}};K) + \frac{\eta_{\mathtt{us}}}{\sqrt{nh}} q_3(z_{\frac{\alpha}{2}};K) \right\} \frac{\phi(z_{\frac{\alpha}{2}})}{f} + o\left((nh)^{-1} + \eta_{\mathtt{us}}^2 + \eta_{\mathtt{us}}(nh)^{-1/2}\right).$$

(b) If $\eta_{\mathtt{bc}} \to 0$ *and* $\rho \to 0$, *then*

$$\mathbb{P}[f \in I_{\mathtt{bc}}] = 1-\alpha + \left\{ \frac{1}{nh} q_1(z_{\frac{\alpha}{2}};K) + \eta_{\mathtt{bc}}^2 q_2(z_{\frac{\alpha}{2}};K) + \frac{\eta_{\mathtt{bc}}}{\sqrt{nh}} q_3(z_{\frac{\alpha}{2}};K) \right\} \frac{\phi(z_{\frac{\alpha}{2}})}{f} + \rho^{1+r}(\Sigma_1 + \rho^r \Sigma_2)\phi(z_{\frac{\alpha}{2}}) z_{\frac{\alpha}{2}} + o\left((nh)^{-1} + \eta_{\mathtt{bc}}^2 + \eta_{\mathtt{bc}}(nh)^{-1/2} + \rho^{1+2r}\right).$$

(c) If $\eta_{\mathtt{bc}} \to 0$ *and* $\rho \to \bar{\rho} < \infty$, *then*

$$\mathbb{P}[f \in I_{\mathtt{rbc}}] = 1-\alpha + \left\{ \frac{1}{nh} q_1(z_{\frac{\alpha}{2}};M) + \eta_{\mathtt{bc}}^2 q_2(z_{\frac{\alpha}{2}};M) + \frac{\eta_{\mathtt{bc}}}{\sqrt{nh}} q_3(z_{\frac{\alpha}{2}};M) \right\} \frac{\phi(z_{\frac{\alpha}{2}})}{f} + o\left((nh)^{-1} + \eta_{\mathtt{bc}}^2 + \eta_{\mathtt{bc}}(nh)^{-1/2}\right).$$

Some implications of this result are obscured by its genericness, in particular regarding the scaled biases $\eta_{\mathtt{us}}$ and $\eta_{\mathtt{bc}}$. However, two features are immediately clear. First, the first three error terms in each case all take the same form.[4] Hence, putting aside part (b) for the moment, comparing (in terms of rates) undersmoothing to robust bias correction amounts to comparing, for a given n and h, the biases. It is clear from Equations (2) and (3) that in many situations $\eta_{\mathtt{bc}} = O(\eta_{\mathtt{us}})$ and $\eta_{\mathtt{bc}} = o(\eta_{\mathtt{us}})$, and thus Theorem 2 immediately proves that robust bias correction can equal, or outperform, undersmoothing in terms of coverage error. This is one of the main insights of our

[4]The kernel K appears in the polynomials $q_j(z_{\frac{\alpha}{2}};\cdot)$ in part (b) because $\rho \to 0$ is assumed, which implies $q_j(z_{\frac{\alpha}{2}};M) = q_j(z_{\frac{\alpha}{2}};K) + o(1)$, not because of the "mismatch" in studentization of $T_{\mathtt{bc}}$. For part (c), the dependence on (the limit of) ρ is retained.

paper. The following subsections explore this point further, first assuming S is large enough to allow all leading constants to be characterized and then later in the regime where smoothness is known and exploited fully.

The second point is the notable presence of $\rho^{1+r}(\Sigma_1 + \rho^r\Sigma_2)$ in part (b) of the theorem. This is the limit of $\sigma^2_{\texttt{rbc}}/\sigma^2_{\texttt{us}} - 1$, and hence consists of the covariance of $\hat{f}$ and $\hat{B}_f$ (denoted by Σ_1) and the variance of $\hat{B}_f$ (Σ_2). Crucially, terms only appear for $T_{\texttt{bc}}$, and not for robust bias correction because they are entirely due to the "mismatch" in $T_{\texttt{bc}}$. That is, although $\sigma_{\texttt{us}}$ is a valid first-order standardization for traditional bias correction (as $\rho \to 0$), it fails to account for any variability in $\hat{B}_f$ that would naturally be present in any finite sample, and may in fact be first order if $\rho \to 0$ is not a good approximation; of course, $\rho = h/b > 0$ for each $n \geq 0$.

Hall (1992b) showed how these terms prevent bias correction from performing as well as undersmoothing in terms of coverage. In essence, the potential for improved bias properties do not translate into improved inference properties because the variance is not well-controlled beyond first order. We emphasize that the new Studentization does not simply remove the leading ρ terms; the entire sequence is absent. Hence, $T_{\texttt{rbc}}$ capitalizes fully on the improvements from bias correction without any higher-order variance penalty. The remainder of our discussion will compare undersmoothing to robust bias correction, and we will not further repeat Hall (1992b)'s convincing argument against traditional bias correction.

Remark 4 (Bootstrap)**.** It is possible to use the bootstrap to obtain quantiles for $T_{\texttt{us}}$, $T_{\texttt{bc}}$, and $T_{\texttt{rbc}}$, rather than the Normal approximation as we do here. It is well known that the bootstrap does not estimate the smoothing bias in nonparametric problems, and thus employing it for $T_{\texttt{us}}$ does not materially affect the main conclusions of this paper. Indeed, Hall (1992b) employs bootstrap quantiles for $T_{\texttt{us}}$ and then studies the higher order distributional properties of $T_{\texttt{us}}$ and $T_{\texttt{bc}}$. This affects the constants, but not the rates, of the coverage error. ■

3.3 Undersmoothing vs. Bias-Correction with Nonbinding Smoothness

In this section and the following, we assume that for a given choice of kernel order, all constants may be characterized explicitly. This amounts to the underlying smoothness S being large enough to be of no direct consequence. This setting seeks to mimic empirical practice, where smoothness is unknown but taken to be large, and the researcher chooses first the order of the kernel and then conducts inference based on that choice. As we discuss in Section 3.6, this approach is suboptimal if the theoretical goal is to conduct inference that exhausts the unknown smoothness.

The question we seek to answer is: if the bias is given as in Eqn. (1) (i.e. the top two cases of (2)), is one better off estimating the leading term (bias correction) or choosing $h \to 0$ fast enough to render the bias negligible (undersmoothing)? Theorem 2 provides an immediate, unambiguous answer to this: robust bias correction is superior if $b \to 0$ because $\eta_{\texttt{bc}} = o(\eta_{\texttt{us}})$, while the leading variance is order $(nh)^{-1}$ in both cases. Allowing large S as a theoretical device we can capture the improvement and offer concrete recommendations for implementation. We will use our expansions to derive optimal bandwidth choices, and we will find that the optimal choice will involve setting ρ

to a positive, finite constant (with $\rho = 1$ being an automatic choice leading to some demonstrable optimality properties for the associated confidence intervals). Intuitively, this is because letting $b \to 0$ faster (larger ρ) removes more bias, and as long as ρ is bounded, the variance is not inflated.

For a basis of comparison, let us first state the coverage error of $I_{\mathtt{us}}$ in this context. This result is most directly comparable to Hall (1992b, §3.4).

Corollary 1 (Undersmoothing)**.** *Let the conditions of Theorem 2(a) hold and fix $r \leq S$. Then*

$$\mathbb{P}[f \in I_{\mathtt{us}}] = 1 - \alpha + \left\{ \frac{1}{nh} q_1(z_{\frac{\alpha}{2}}; K) + nh^{1+2r} (f^{(r)})^2 \mu_{K,r}^2 q_2(z_{\frac{\alpha}{2}}; K) + h^r f^{(r)} \mu_{K,r} q_3(z_{\frac{\alpha}{2}}; K) \right\} \frac{\phi(z_{\frac{\alpha}{2}})}{f} + o\left((nh)^{-1} + nh^{1+2r} + h^r\right).$$

In particular, if $h^*_{\mathtt{us}} = H^*_{\mathtt{us}} n^{-1/(1+r)}$, then $\mathbb{P}[f \in I_{\mathtt{us}}] = 1 - \alpha + O(n^{-r/(1+r)})$, where

$$H^*_{\mathtt{us}} = \arg\min_H \left\{ H^{-1} q_1(z_{\frac{\alpha}{2}}; K) + H^{1+2r} (f^{(r)})^2 \mu_{K,r}^2 q_2(z_{\frac{\alpha}{2}}) + H^r f^{(r)} \mu_{K,r} q_3(z_{\frac{\alpha}{2}}; K) \right\}.$$

This result establishes a benchmark for comparison of coverage errors and confirms that $T_{\mathtt{us}}$ must be undersmoothed. Indeed, here the optimal h balances variance against bias (in rates), rather than squared bias as in mean square error. The MSE-optimal bandwidth is $h_{\mathtt{mse}} \propto n^{-1/(1+2r)}$, which is not allowed for in this expansion (or Theorem 1(a)) because it is too large, i.e. $h^*_{\mathtt{us}} = o(h_{\mathtt{mse}})$.

Turning to robust bias correction, let us simplify the discussion by taking $s = 2$, reflecting the widespread use of symmetric kernels.[5] With this choice, Eqn. (4) yields the tidy expression

$$\eta_{\mathtt{bc}} = \sqrt{nh}(\mathbb{E}[\hat{\theta}_2] - f) = \sqrt{nh} h^{r+2} f^{(r+2)} \left(\mu_{K,r+2} - \rho^{-2} \mu_{K,r} \mu_{L,2}\right). \tag{5}$$

First, we argue that $\bar{\rho} = 0$ is suboptimal because this does not exploit the full power of the variance correction. Intuitively, the standard errors $\hat{\sigma}^2_{\mathtt{rbc}}$ control variance up to order $(nh)^{-1}$, while letting $b \to 0$ faster removes more bias. If b vanishes too fast, the variance is no longer controlled. Setting $\bar{\rho} \in (0, \infty)$ balances these two. The following result makes this intuition precise.

Corollary 2 (Robust bias correction: $\rho \to 0$)**.** *Let the conditions of Theorem 2(c) hold, with $\bar{\rho} = 0$, and fix $s = 2$ and $r \leq S - 2$. Then*

$$\mathbb{P}[f \in I_{\mathtt{rbc}}] = 1 - \alpha + \left\{ \frac{1}{nh} q_1(z_{\frac{\alpha}{2}}; K) + nh^{1+2(r+2)} (f^{(r+2)})^2 \left(\mu_{K,r+2}^2 + \rho^{-4} \mu_{K,r}^2 \mu_{L,2}^2\right) q_2(z_{\frac{\alpha}{2}}; K) + h^{r+2} f^{(r+2)} \left(\mu_{K,r+2} + \rho^{-2} \mu_{K,r} \mu_{L,2}\right) q_3(z_{\frac{\alpha}{2}}; K) \right\} \frac{\phi(z_{\frac{\alpha}{2}})}{f} + o\left((nh)^{-1} + nh^{1+2(r+2)}(1 + \rho^{-4}) + h^{r+2}(1 + \rho^{-2})\right).$$

[5] This does not affect the conclusions in any conceptual way, but considerably simplifies the notation. The coverage error decay rates and optimal bandwidth order of Corollaries 2 and 3 will of course change slightly, and constants such as $\mu_{L,2}$ would have to replaced by their generic counterparts. When S is assumed known, it would be standard to take $s = 2$ and $r = S - 2$.

By virtue of our new studentization, the leading variance remains order $(nh)^{-1}$ and the problematic correlation terms are absent, however by forcing $\rho \to 0$, the ρ^{-2} terms of $\eta_{\mathtt{bc}}$ are dominant (the bias of $\hat{B}_f$), and in light of our results, unnecessarily inflated. Hall (1992b, p. 682) remarked that if $\mathbb{E}[\hat{f}] - f - B_f$ is (part of) the leading bias term, then "the explicit bias correction method is even less attractive relative to undersmoothing due to the appearance of additional, dominant error terms." Our analysis shows that, on the contrary, when using our proposed Studentization, we would like to take $\bar{\rho} \in (0, \infty)$ so that both biases are the same order.[6]

Thus we give the following result, which quantifies the rate improvement due to robust bias correction when smoothness is nonbinding. Being bounded and positive, ρ does not affect the rates of convergence, only the shape of the kernel M, and we make this explicit by writing $M = M_\rho$.

Corollary 3 (Robust bias correction: bounded, positive ρ)**.** *Let the conditions of Theorem 2(c) hold, with $\bar{\rho} \in (0, \infty)$ and fix $s = 2$ and $r \leq S - 2$. Then*

$$\begin{aligned}\mathbb{P}[f \in I_{\mathtt{rbc}}] = 1 - \alpha + \Bigg\{ &\frac{1}{nh} q_1(z_{\frac{\alpha}{2}}; M_{\bar{\rho}}) + nh^{1+2(r+2)} (f^{(r+2)})^2 \left(\mu_{K,r+2} + \bar{\rho}^{-2} \mu_{K,r} \mu_{L,2}\right)^2 q_2(z_{\frac{\alpha}{2}}; M_{\bar{\rho}}) \\ &+ h^{r+2} f^{(r+2)} \left(\mu_{K,r+2} + \bar{\rho}^{-2} \mu_{K,r} \mu_{L,2}\right) q_3(z_{\frac{\alpha}{2}}; M_{\bar{\rho}}) \Bigg\} \frac{\phi(z_{\frac{\alpha}{2}})}{f} \\ &+ o\left((nh)^{-1} + nh^{1+2(r+2)} + h^{r+2}\right).\end{aligned}$$

In particular, if $h^*_{\mathtt{rbc}} = H^*_{\mathtt{rbc}}(\rho) n^{-1/(1+(r+2))}$, then $\mathbb{P}[f \in I_{\mathtt{rbc}}] = 1 - \alpha + O(n^{-(r+2)/(1+(r+2))})$, where

$$\begin{aligned}H^*_{\mathtt{rbc}}(\bar{\rho}) = \arg\min_H \Big\{ &H^{-1} q_1(z_{\frac{\alpha}{2}}; M_{\bar{\rho}}) + H^{1+2(r+2)} (f^{(r+2)})^2 \left(\mu_{K,r+2} + \bar{\rho}^{-2} \mu_{K,r} \mu_{L,2}\right)^2 q_2(z_{\frac{\alpha}{2}}; M_{\bar{\rho}}) \\ &+ H^{r+2} f^{(r+2)} \left(\mu_{K,r+2} + \bar{\rho}^{-2} \mu_{K,r} \mu_{L,2}\right) q_3(z_{\frac{\alpha}{2}}; M_{\bar{\rho}}) \Big\}.\end{aligned}$$

The most notable feature of this result, beyond the formalization of the coverage improvement, is that the coverage error terms share the same structure as those of Corollary 1, with r replaced by $r + 2$, and represent the same conceptual objects. This is due at heart to the fact that M can be viewed as a higher order kernel, as discussed in Remark 6 further below.

3.4 Bandwidth Choices and Implications for Empirical Practice

We now use Corollaries 1 and 3 to give concrete methodological recommendations for empirical work, both in terms of coverage and interval length. In practice, and because the smoothness of f is unknown, employing robust bias correction to construct confidence intervals for f seems a desirable approach because these intervals will exhibit coverage error decay rates that are never slower than those offered by undersmoothing. To make this point precise, let $I_{\mathtt{us}}(h)$ and $I_{\mathtt{rbc}}(h)$ denote the intervals $I_{\mathtt{us}}$ and $I_{\mathtt{rbc}}$, respectively, when constructed using the bandwidth $h \to 0$. To quantify precisely the differences between undersmoothing and bias-correction on inference, we

[6] This reasoning is not an artifact of choosing r even and $s = 2$, but in other cases $\rho \to 0$ can be optimal if the convergence is sufficiently slow to equalize the two bias terms.

first need to choose a bandwidth sequence. As discussed below, the simple choice of $\rho = 1$ ($b = h$) performs very well, and is optimal in certain senses. Thus, we focus on the consequences of choice of h, for which we have three sensible options. Throughout, r is fixed. We focus only on rates for now, suppressing the constants.

1. (MSE Optimal.) The MSE-optimal bandwidth is $h^*_{\texttt{mse}} \propto n^{-1/(1+2r)}$. This choice of bandwidth is simple and popular, but leads to first-order bias in $T_{\texttt{us}}$, as shown in Theorem 1(a), rendering $I_{\texttt{us}}$ invalid. However, $I_{\texttt{rbc}}$ is still valid, and we can quantify the rate of coverage error decay:

$$\mathbb{P}[f \in I_{\texttt{us}}(h^*_{\texttt{mse}})] - (1-\alpha) \asymp 1 \qquad \text{vs.} \qquad \mathbb{P}[f \in I_{\texttt{rbc}}(h^*_{\texttt{mse}})] - (1-\alpha) \asymp n^{-\min\{4,r+2\}/(1+2r)}.$$

2. (Coverage Optimal for Undersmoothing.) While ad-hoc undersmoothing of $h^*_{\texttt{mse}}$ is a popular method for fixing the above first-order distortion, a more theoretically founded choice is $h^*_{\texttt{us}} \propto n^{-1/(1+r)}$, which is also a valid choice for $I_{\texttt{rbc}}$, and in fact yields the same rates:

$$\mathbb{P}[f \in I_{\texttt{us}}(h^*_{\texttt{us}})] - (1-\alpha) \asymp n^{-r/(1+r)} \qquad \text{vs.} \qquad \mathbb{P}[f \in I_{\texttt{rbc}}(h^*_{\texttt{us}})] - (1-\alpha) \asymp n^{-r/(1+r)}.$$

3. (Coverage Optimal for Robust Bias Correction.) Using $h^*_{\texttt{rbc}} \propto n^{-1/(1+(r+2))}$ again leads to a first-order coverage distortion of $I_{\texttt{us}}$, but $I_{\texttt{rbc}}$ shows improvements in coverage:

$$\mathbb{P}[f \in I_{\texttt{us}}(h^*_{\texttt{rbc}})] - (1-\alpha) \asymp 1 \qquad \text{vs.} \qquad \mathbb{P}[f \in I_{\texttt{rbc}}(h^*_{\texttt{rbc}})] - (1-\alpha) \asymp n^{-(r+2)/(1+(r+2))}.$$

The first point formalizes that an MSE-optimal bandwidth is always a valid choice for robust bias correction, however, the coverage error rates depend on the kernel order. In particular, the robust bias-corrected interval $I_{\texttt{rbc}}(h^*_{\texttt{mse}})$ will achieve the fastest decay in coverage error only when second-order kernels are employed. If higher-order kernels are employed, then the rate is suboptimal, and interestingly, the rate slows as r increases. This finding is important for empirical work, as it implies that either second order kernels should be used, or if higher-order kernels are required, either a coverage-optimal bandwidth should be used or, at the very least, the MSE-optimal bandwidth should be modified to

$$\tilde{h}_{\texttt{mse}} = h^*_{\texttt{mse}} n^{-(r-2)/((1+2r)(1+(r+2)))} \propto n^{-1/(1+(r+2))}.$$

Note that this is not an artifact of taking $s = 2$, because, as can be seen from Eqn. (3), coverage will be optimal when the leading bias is order h^{r+2}, for any s and $\bar{\rho} \in [0, \infty)$.

After considering coverage error in detail, it is natural to examine is interval length. An obvious concern is that the improvements in coverage offered by robust bias correction may come at the expense of a wider interval. By its symmetry, the length of the intervals $I_{\texttt{us}}$ and $I_{\texttt{rbc}}$ take the same form:

$$|I_{\texttt{us}}(h)| = 2z_{\frac{\alpha}{2}} \frac{\hat{\sigma}_{\texttt{us}}}{\sqrt{nh}} \qquad \text{and} \qquad |I_{\texttt{rbc}}(h)| = 2z_{\frac{\alpha}{2}} \frac{\hat{\sigma}_{\texttt{rbc}}}{\sqrt{nh}}.$$

Thus, comparing length amounts to first examining the rate of contraction, which do presently, and then the variance constants, which depend on $\bar{\rho} \in (0, \infty)$ and are addressed in the next section. Broadly, the answer for rates follows immediately from the above discussion: Robust bias correction can accommodate (and will optimally employ) a larger bandwidth (i.e. $h \to 0$ more slowly), and hence $I_{\texttt{rbc}}$ will contract more quickly (i.e. $nh \to \infty$ faster than with undersmoothing). To be precise, we have $|I_{\texttt{rbc}}(h^*_{\texttt{rbc}})|^2 \asymp n^{-(r+2)/(1+(r+2))}$ compared to $|I_{\texttt{us}}(h^*_{\texttt{us}})|^2 \asymp n^{-r/(1+r)}$. It is also instructive to note that $|I_{\texttt{rbc}}(h^*_{\texttt{mse}})|^2 \asymp n^{-2r/(1+r)}$ and $|I_{\texttt{rbc}}(h^*_{\texttt{us}})|^2 \asymp n^{-r/(1+r)}$, which agrees with the above discussion regarding use of $h^*_{\texttt{mse}}$ and $h^*_{\texttt{us}}$. The intervals $I_{\texttt{us}}(h^*_{\texttt{mse}})$ and $I_{\texttt{us}}(h^*_{\texttt{rbc}})$ do not have correct asymptotic coverage.

3.5 Choice of ρ

We now turn attention to choosing ρ. Recall that bounded, positive ρ impacts the shape of the "kernel" $M_\rho(u) = K(u) - \rho^{1+r} L^{(r)}(\rho u)\mu_{K,r}$, and hence choice of ρ depends on what properties are desired for the kernel. We continue in the regime with large S and $s = 2$. There are several cogent possibilities here, and can depend on user preferences: we will show that for both minimizing interval length and MSE, $\rho = 1$ is the optimal choice, provided K and L are chosen appropriately. Thus, we recommend the simple choice of $\rho = 1$, which has good theoretical properties and performs very well numerically. As a result, from the practitioner's point of view, choice of ρ (or b) is completely automatic. The same choice is intuitive in local polynomials; see Section 4.3.

We explore two choices for ρ: first minimizing interval length and second MSE (in terms of constants, rates are compared above). Both of these depend on well-understood optimal kernel shapes. Given the rate results in the prior section, we employ a second order kernel for bias correction and compare against the benchmark of the optimal fourth-order kernel. We take the optimal shapes from Gasser, Muller, and Mammitzsch (1985).[7] In both cases, the simple choice of $\rho = 1$ is optimal.[8]

Following up on the prior section's discussion of the rate of interval length contraction, we can choose ρ to minimize the constant portion of (asymptotic) interval length. This follows because $\sigma^2_{\texttt{us}} \to f\vartheta_{K,2}$ and $\sigma^2_{\texttt{rbc}} \to f\vartheta_{M,2}$, and so choosing a kernel to minimize interval length, for given bandwidth h, is asymptotically equivalent to finding the minimum variance kernel of a given order. The fourth order minimum variance kernel is $K_{\texttt{mv}}(u) = (3/8)(-5u^2 + 3)$, and hence this is the benchmark. Perhaps surprisingly, by setting $\rho = 1$ and choosing K and $L^{(2)}$ to be the second-order minimum variance kernels for estimating f and $f^{(2)}$ respectively, the resulting $M_1(u)$ is exactly $K_{\texttt{mv}}(u)$. In this case, we choose K to be the uniform kernel and $L^{(2)} = (15/4)(3u^2 - 1)$.

For MSE we obtain a similar result. The benchmark fourth order kernel is $K_{\texttt{mse}}(u) = (15/32)(7u^4 - 10u^2 + 3)$ and this is exactly the kernel $M_1(u)$ we find by setting $\rho = 1$ and choosing K and $L^{(2)}$

[7] As discussed in more detail in the supplement, the optimal shapes for derivative estimation given in Gasser, Muller, and Mammitzsch (1985) belong to a slightly different class of kernels than those defined by Assumption 3.2(a), and differ chiefly in how they achieve limiting unbiasedness. Our results easily extend to this class, though we maintain 3.2(a) for simplicity and comparability to prior work.

[8] The optimality properties do not extend to higher order kernels.

Table 1: Results for other kernel shapes

Kernel K	Kernel $L^{(2)}$	$\tilde{\mu}_{M,4}$	$\vartheta_{M,2}$	MSE
Epanechnikov	$(105/16)(6u^2-5u^4-1)$	-0.0476	1.2500	0.6199
Uniform	$(105/16)(6u^2-5u^4-1)$	-0.0222	1.4722	0.6052
Biweight	$(105/16)(6u^2-5u^4-1)$	-0.0476	1.2500	0.6199
Triweight	$(105/16)(6u^2-5u^4-1)$	-0.0438	1.2774	0.6202
Tricube	$(105/16)(6u^2-5u^4-1)$	-0.0506	1.2332	0.6207
Cosine	$(105/16)(6u^2-5u^4-1)$	-0.0476	1.2503	0.6199
Epanechnikov	$(15/4)(3u^2-1)$	-0.0857	1.1250	0.6432
Uniform	$(15/4)(3u^2-1)$	-0.0857	1.1250	0.6432
Biweight	$(15/4)(3u^2-1)$	-0.0748	1.1352	0.6291
Triweight	$(15/4)(3u^2-1)$	-0.0649	1.1631	0.6229
Tricube	$(15/4)(3u^2-1)$	-0.0780	1.1319	0.6333
Cosine	$(15/4)(3u^2-1)$	-0.0836	1.1254	0.6399
Biweight	Biweight$^{(2)}$	-0.0748	1.1352	0.6291
Tricube	Tricube$^{(2)}$	-0.0790	1.1993	0.6687
Gaussian	Gaussian$^{(2)}$	-3.0000	0.4760	0.6599

to be the MSE-optimal kernels for their respective problems. Here $K(u) = (3/4)(1-u^2)$ and $L^{(2)}(u) = (105/16)(6u^2-5u^4-1)$. A practitioner may be interested in using MSE-optimal kernels (perhaps along with $h^*_{\texttt{mse}}$) to obtain the best possible point estimate. Our results then give a natural measure of uncertainty to accompany the point estimate, which has correct coverage and the attractive feature of using the same effective samples as well as the rate and constant optimality.

For other choices of K and L, Table 1 shows the resulting interval length (measured by $\vartheta_{M,2}$) and MSE (measured by $(\vartheta^8_{M,2}\tilde{\mu}^2_{M,4})^{1/9}$; see Wand and Jones (1995)) of the implied kernel $M_1(u)$. We also include a measure of bias given by $\tilde{\mu}_{M,4}$[9]. Of course, whenever the underlying kernels or orders are not the above, one could choose ρ numerically to minimize some notation of distance (e.g., L_2) between the resulting kernel M_ρ and the optimal kernel shape already available in the literature. This procedure is general, but perhaps more cumbersome, and consequently we recommend $\rho = 1$ as a simple rule-of-thumb for implementation. As shown in the Table, the relative performance does not suffer greatly from setting $\rho = 1$ for all kernel shapes.

Remark 5 (Coverage Error Optimal Kernels)**.** The above discussion focuses on two notions of optimality for kernel shapes, as these have been studied in the literature. But our results hint at a further possibility: finding the optimal kernel for coverage error. This kernel, for a fixed order r, would minimize the constants found in Corollary 1. In that result, h is chosen to optimize the rate and the constant $H^*_{\texttt{us}}$ gives the minimum for a fixed kernel K. A step further would be to view $H^*_{\texttt{us}}$ as a function of the kernel K, and optimizing. To our knowledge, such a derivation has not been done and it may be of interest. ■

[9]Notat that $\tilde{\mu}_{M,4} = (k!)(-1)^k\mu_{M,4}$ according to our notation.

3.6 Undersmoothing vs. Bias-Correction with Binding Smoothness

A theoretical problem that has attracted some recent attention is related to adaptibity to unknown smoothness, that is, constructing point estimators and inference procedures that utilized all the unknown smoothness optimally (or nearly so). See Tsybakov (2003) for a review, and Armstrong (2014) and Armstrong and Kolesár (2014) for two recent examples in econometrics employing this idea in the context of regression estimation at a boundary point. In this paper, we do not explore smoothness adaptation, but rather take smoothness as given and investigate the implications of employing bias correction techniques and bandwidth selection for confidence intervals using higher-order Edgeworth expansions. The two approaches are complementary as they give different insights on the properties of point estimators and inference procedures for kernel-based nonparametrics.

The discussion given so far, however, assumed that the level of smoothness was large enough to be inconsequential in the analysis. In this section, in contrast, we take the level of smoothness to be binding, so that we can fully utilize the S derivatives *and* the Hölder condition to obtain the best possible rates of decay in coverage error for both undersmoothing and robust bias correction, but at the price of implementability: the leading bias constants can not be characterized, and hence feasible "optimal" bandwidths are not available.

For undersmoothing, the lowest bias is attained by setting $r > S$ (see Eqn. (2)), in which case the bias is only known to satisfy $\mathbb{E}[\hat{f}] - f = O(h^{S+\varsigma})$ (i.e., B_f is identically zero) and bandwidth selection is not feasible. Note that this approach allows for $\sqrt{nh}h^S \not\to 0$, as $\eta_{\mathtt{us}} = O(\sqrt{nh}h^{S+\varsigma})$.

Robust bias correction has several interesting features here. If $r \leq S-2$ (the top two cases in Eqn. (3)), then the bias from approximating $\mathbb{E}[\hat{f}] - f$ by B_f, that is not targeted by bias correction, dominates $\eta_{\mathtt{bc}}$ and prevents robust bias correction from performing as well as the best possible infeasible (i.e., oracle) undersmoothing approach. That is, even bias correction requires a sufficiently large choice of r in order to ensure the fastest possible rate of decay in coverage error: if $r \geq S-1$, robust bias correction can attain error decay rate as the best undersmoothing approach, and allow $\sqrt{nh}h^S \not\to 0$.

Within $r \geq S-1$, two cases emerge. On the one hand, if $r = S-1$ or S, then B_f is nonzero and $f^{(r)}$ must be consistently estimated to attain the best rate. Indeed, more is required. From Eqn. (3), we will need a bounded, positive ρ to equalize the bias terms. This (again) highlights the advantage of robust bias correction, as the classical procedure would enforce $\rho \to 0$, and thus underperform. On the other hand, $\rho \to 0$ will be required if $r > S$ because (from the final case of (3)) we require $\rho^{r-S} = O(h^{\varsigma})$ to attain the same rate as undersmoothing. Note that we can accommodate $b \not\to 0$ (but bounded). Interestingly, B_f is identically zero and $\hat{B}_f$ merely adds noise to the problem, but this noise is fully accounted for by the robust standard errors, and hence does not affect the rates of coverage error (though the constants of course change). The $\hat{f}^{(r)}$ in $\hat{B}_f$ is *inconsistent* ($f^{(r)}$ does not exist), but the nonvanishing bias of $\hat{f}^{(r)}$ is dominated by h^r.

This discussion is summarized by the following result, specialized from Theorem 2.

Corollary 4. *Let the conditions of Theorem 2 hold.*

(a) If $r > S$, then

$$\mathbb{P}[f \in I_{\mathtt{us}}] = 1 - \alpha + \frac{1}{nh}\frac{\phi(z_{\frac{\alpha}{2}})}{f} q_1(z_{\frac{\alpha}{2}}; K) + O\left(nh^{1+2S+2\varsigma} + h^{S+\varsigma}\right) + o\left((nh)^{-1}\right).$$

(b) If $r \geq S - 1$, then

$$\begin{aligned}\mathbb{P}[f \in I_{\mathtt{rbc}}] = 1 - \alpha + \frac{1}{nh}\frac{\phi(z_{\frac{\alpha}{2}})}{f} q_1(z_{\frac{\alpha}{2}}; M) \\ + O\left(nh(h^{S+\varsigma} \vee h^r b^{S-r+\varsigma \mathbb{1}\{r \leq S\}})^2 + (h^{S+\varsigma} \vee h^r b^{S-r+\varsigma \mathbb{1}\{r \leq S\}})\right) + o\left((nh)^{-1}\right).\end{aligned}$$

Remark 6 (Comparison to Hall (1992b))**.** By viewing r as fixed and $B_f = B_f(r)$ as the leading bias we depart from Hall (1992b), who forces both methods to use the same total amount of smoothness. That is, suppose undersmoothing employed a kernel $K_{\mathtt{us}}$ of order $r_{\mathtt{us}}$, then bias correction use a kernel $K_{\mathtt{bc}}$ in $\hat{f}$ whose order obeys $r_{\mathtt{bc}} + s = r_{\mathtt{us}}$ (and generally equal to S, though this is not important). In this case, $B_f(r_{\mathtt{us}})$ is *not* the bias explicitly corrected for, while it is still the bias to be reduced by undersmoothing. This fact, coupled with the requirement that $\rho \to 0$, implicitly constrains explicit bias correction to remove *less* bias than undersmoothing.

On the contrary, Corollaries 1 and 3 show that for $r_{\mathtt{bc}} + s = r_{\mathtt{us}} \leq S$, bias correction, with our proposed standard errors and a bounded, positive ρ, attains the same decay rates in coverage error in undersmoothing. Corollary 4 shows the same, exhausting the smoothness completely. Therefore, even when forcing the methods to use the same amount of smoothness, bias correction is not inferior to undersmoothing. This can be seen even more starkly by recalling Eqn. (4), i.e. that the bias-corrected estimator can be reframed as an average of the kernel M. It is straightforward that M is a kernel of order $r_{\mathtt{bc}} + s$, and hence if this sum is set equal to $r_{\mathtt{us}}$ as above, then both methods use kernels of the *same order*, just different shape. It would surprising indeed if intervals based on one kernel (K) attained better *rates* than those based on a different kernel (M). ■

3.7 Inference at the Boundary

We close our discussion of the density with a brief word about inference at the boundary. Local polynomial regression at the boundary, which is of greater practical importance, is discussed at length below. For the density, the results above continue to hold, but we must be careful when imposing Assumption 3.2(b).[10] To illustrate, suppose the support of X is $[0, \infty)$, the parameter of interest is $f(0)$, and we use a second order boundary-corrected kernel K. Then Assumption 3.2(b) requires $f(0)L^{(1)}(0) = f^{(1)}(0)L(0) = 0$, which can only be satisfied with strong restrictions on the data generating process or by using very particular kernels. However, under such assumptions, the conclusions above will still hold (with appropriate modifications to the constants).

This is by no means an exhaustive treatment of refinements at the boundary, and indeed, may not be satisfactory in practice, where other methods may be more appropriate. For a review on

[10] Also, the smoothness of f must be assumed away from the boundary point.

alternative methods see, e.g., Karunamuni and Alberts (2005) and references therein.

4 Local Polynomial Estimation

We now turn to local polynomial regression. This section has two principle aims. First, we show that the conclusions regarding bias correction carry over from the density case to regression. Second, apart from bias issues, we build on the brief discussions in Remark 1 and Section 3.7 to show that with proper fixed-n studentization, local polynomials do not suffer from coverage error problems at the boundary, in contrast to the findings of Chen and Qin (2002). We will be brief so as to keep the discussion focused on what is novel compared to the density case. To avoid repeatition and overwhelming notation, we will implicitly handle both interior point and boundary point simultaneously by employing high-level notation and restricting attention to odd polynomial powers. In the supplemental appendix, we do spell out the main results for either case in some detail.

First, let us carefully define the regression estimator, its bias, and the bias correction. The notational burden is relatively high for local polynomial methods. Given a random sample $\{(Y_i, X_i) : 1 \leq i \leq n\}$, the local polynomial estimator of $m(x) = \mathbb{E}[Y_i|X_i = x]$ is defined as

$$\hat{m}(x) = e_0'\hat{\beta}_p, \qquad \hat{\beta}_p = \underset{b\in\mathbb{R}^{p+1}}{\arg\min} \frac{1}{nh}\sum_{i=1}^{n}(Y_i - r_p(X_i - x)'b)^2 K\left(\frac{x - X_i}{h}\right),$$

where, for an integer $p \geq 1$, e_0 is the $(p+1)$-vector with a one in the first position and zeros in the remaining, and $r_p(u) = (1, u, u^2, \ldots, u^p)'$. See Fan and Gijbels (1996) for a comprehensive review. We restrict attention to p odd, as is standard, due to the theoretical advantages of odd degree fitting. We define $Y = (Y_1, \cdots, Y_n)'$, $R_p = [r_p((X_1 - x)/h), \cdots, r_p((X_n - x)/h)]'$, $W_p = \text{diag}(h^{-1}K((X_i - x)/h) : i = 1, \ldots, n)$, and $\Gamma_p = R_p'W_pR_p/n$. (Here $\text{diag}(a_i : i = 1, \ldots, n)$ denotes the $n\times n$ diagonal matrix constructed using the elements $a_1, a_2, \cdots, a_n$.) Then, the local polynomial estimator is $\hat{m} = e_0'\Gamma_p^{-1}R_p'W_pY/n$. The conditional bias is given by

$$\mathbb{E}[\hat{m}|X_1, \ldots, X_n] - m = h^{p+1}m^{(p+1)}\frac{1}{(p+1)!}e_0'\Gamma_p^{-1}\Lambda_p + o_P(h^{p+1}), \tag{6}$$

where $\Lambda_p = R_p'W_p[((X_1 - x)/h)^{p+1}, \cdots, ((X_n - x)/h)^{p+1}]'/n$. Here, the quantity $e_1'\Gamma_p^{-1}\Lambda_p/(p+1)!$ is random, unlike in the density case, but it is known and bounded in probability.

We will estimate $m^{(p+1)}$ in (6) using a second local polynomial regression, of degree $q > p$, based on a kernel L and bandwidth b. Thus, $r_q(u)$, R_q, W_q, and Γ_q are defined as above, but substituting q, L, and b in place of p, K, and h, respectively. In general, a subscript p will denote quantities involved in $\hat{m}$, while a subscript of q indicates use in $\hat{m}^{(p+1)}$. Denote by e_{p+1} the $(q+1)$-vector with one in the $p+2$ position, and zeros in the rest. Then we estimate the bias as

$$\hat{B}_m = h^{p+1}\hat{m}^{(p+1)}\frac{1}{(p+1)!}e_0'\Gamma_p^{-1}\Lambda_p, \qquad \text{where} \qquad \hat{m}^{(p+1)} = b^{-p-1}e_{p+1}'\Gamma_q^{-1}R_q'W_qY/n.$$

Exactly as in the density case, subtracting $\hat{B}_m$ introduces variance that is controlled by ρ. Robust bias correction will once again capture the variance of $\hat{B}_m$. Recycling notation to emphasis the parallel, the three statistics we are consider are:

$$T_{\mathtt{us}} = \frac{\sqrt{nh}(\hat{m} - m)}{\hat{\sigma}_{\mathtt{us}}}, \qquad T_{\mathtt{bc}} = \frac{\sqrt{nh}(\hat{m} - \hat{B}_m - m)}{\hat{\sigma}_{\mathtt{us}}}, \qquad \text{and} \qquad T_{\mathtt{rbc}} = \frac{\sqrt{nh}(\hat{m} - \hat{B}_m - m)}{\hat{\sigma}_{\mathtt{rbc}}}.$$

Next, we characterize the variances and their estimators, which will complete the description of these three statistics. Notice that all the definitions above, as well as those upcoming, are valid for an evaluation point in the interior and at the boundary of the support of X_i.

4.1 Variance Estimation

The studentizations employed in the density case were based on fixed-n expectations, and we aim to show that retaining this spirit is crucial for local polynomials. We will give the corresponding forms for local polynomials and contrast these with their asymptotic counterparts. The fixed-n versus asymptotic distinction is separate from, and more fundamental than, whether we employ feasible versus infeasible quantities.

Let us begin by giving the infeasible variances and their asymptotic counterparts. To work with fixed n, we will condition on the covariates, so that Γ_p^{-1} is fixed. Define $v(\cdot) := \mathbb{V}[Y|X = \cdot]$ and $\Sigma = \mathrm{diag}(v(X_i) : i = 1, \ldots, n)$. Then, straightforward calculation gives

$$\sigma^2_{\mathtt{us}} := (nh)\mathbb{V}[\hat{m}|X_1, \cdots, X_n] = e_0'\Gamma_p^{-1}R_p'W_p\Sigma W_pR_p\Gamma_p^{-1}e_0. \tag{7}$$

It is then easy to find that $\sigma^2_{\mathtt{us}} \to_P v(x)f(x)^{-1}\mathcal{V}(K,p)$, for a known, constant function $\mathcal{V}(K,p)$ of the kernel and polynomial degree. The latter calculation remains the same whether the evaluation point of interest x is on (or near) boundary or in the interior of the support of X_i, though $\mathcal{V}(K,p)$ changes slightly. Nonetheless, Eqn. (7) is always the same.

To first order, one could use $\sigma^2_{\mathtt{us}}$ or the leading asymptotic term; all that remains is to make them feasible. To utilize $\sigma^2_{\mathtt{us}}$, only Σ need be estimated, involving the variance function $v(\cdot)$, whereas the asymptotic form requires a density estimate as well. These two unknown functions, $v(x)$ and $f(x)$, may be difficult to estimate when x is a boundary point, and motivated by this concern, Chen and Qin (2002, p. 93) consider infeasible versions but conclude that "an increased coverage error near the boundary is still the case even when we know the values of $f(x)$ and $v(x)$." Our results show that this is not true in general: using fixed-n Studentization based on Eqn. (7), feasible or infeasible, leads to confidence intervals with the same coverage error rates at the interior and at the boundary of the support of X_i, thereby retaining the celebrated boundary carpentry property.

As in the density case, $\sigma^2_{\mathtt{rbc}}$ will capture the variance of $\hat{m}$ and $\hat{m}^{(p+1)}$ as well as their covariance,

and as in to Eqn. (7), it is based on the fixed-n calculation:

$$\begin{aligned}\sigma^2_{\texttt{rbc}} &:= (nh)V[\hat{m} - \hat{B}_m | X_1, \ldots, X_n] \\ &= e_0'\Gamma_p^{-1}\left(R_p'W_p - \rho^{p+1}\Lambda_p\Gamma_q^{-1}R_q'W_q\right)\Sigma\left(R_p'W_p - \rho^{p+1}\Lambda_p\Gamma_q^{-1}R_q'W_q\right)'\Gamma_p^{-1}e_0.\end{aligned} \tag{8}$$

To make the fixed-n scalings feasible, $\hat{\sigma}^2_{\texttt{us}}$ and $\hat{\sigma}^2_{\texttt{rbc}}$ take the forms (7) and (8) and replace Σ with an appropriate estimator. For $\hat{\sigma}^2_{\texttt{us}}$, we estimate Σ using $\hat{\Sigma}_p = \text{diag}(\hat{v}(X_i) : i = 1, \ldots, n)$, where $\hat{v}(X_i) = (Y_i - r_p(X_i - x)'\hat{\beta}_p)^2$. Robust bias correction takes the same approach, with q in place of p, i.e. $\hat{\Sigma}_q$ is constructed using the plug-in estimated residuals $\hat{v}(X_i) = (Y_i - r_q(X_i - x)'\hat{\beta}_q)^2$. The logic behind these choices is that $r_p(X_i - x)'\beta_p$ is a p-term Taylor expansion of $m(X_i)$ around the point of interest, and $\hat{\beta}_p$ estimates β_p.

One crucial property of this method, in the context of Edgeworth expansions, is that the bias in estimation of $\sigma^2_{\texttt{us}}$ is of the same order as the original $\hat{m}(x)$. Because $q > p$, $\hat{\Sigma}_q$ suffers lower bias than $\hat{\Sigma}_p$, just as $\hat{m} - \hat{B}_m$ is also bias-corrected.[11] Using other methods may result in additional terms, with possibly distinct rates, appearing in the Edgeworth expansions. Some examples that may have this issue are (i) using $\hat{v}(X_i) = (Y_i - \hat{m}(x))^2$; (ii) using local or assuming global homoskedasticity; (iii) using other nonparametric estimators for $v(X_i)$, relying on new tuning parameters.

4.2 Coverage Error

With the statistics $T_{\texttt{us}}$, $T_{\texttt{bc}}$, and $T_{\texttt{rbc}}$ completely (re-)defined, we can now present the expansions of coverage error. We will state only a generic coverage error result, similar to Theorem 2. The implications of the result, the interaction with the smoothness of m, and bandwidth selection, are all analogous to the density case, and we will avoid repeating the discussion. Indeed, much of the notation is intentionally recycled from Theorem 2 to emphasize the similarity.

The following conditions will suffice for our results.

Assumption 4.1 (Data-generating process)**.** $\{(Y_1, X_1), \ldots, (Y_n, X_n)\}$ *is a random sample, where* X_i *has the absolutely continuous distribution with Lebesgue density* f*,* $\mathbb{E}[Y^4|X = \cdot] < \infty$*, and in a neighborhood of* x*,* f *and* v *are continuous and bounded away from zero,* m *is* S*-times continuously differentiable with bounded derivatives, and* $m^{(S)}$ *is Hölder continuous with exponent* ς*.*

Assumption 4.2 (Kernels)**.** *The kernels* K *and* L *are bounded, even functions with compact support.*

Assumption 4.3 (Cramér's Condition)**.** *For each* $\delta > 0$ *and all sufficiently small* h*, the random variable* $Z = (K(u)r_p(u)', K(u)r_p(u)'\varepsilon, \text{vech}(K(u)r_p(u)r_p(u)')', \text{vech}(K(u)r_p(u)r_p(u)'\varepsilon^2)')'$ *obeys*

$$\sup_{t \in \mathbb{R}^{\dim\{Z\}}, \|t\| > \delta} \left| \int \exp\{it'Z\} f(x - uh) du \right| \leq 1 - C(x, \delta)h,$$

[11] This relies on choosing $\rho \not\to 0$, as the bias in $\hat{\Sigma}_q$ is of order $b^{q+1} = b^{q-p}\rho^{-p-1}$, and thus if $\rho \to 0$ sufficiently fast, then the bias is greater. This is due to the absence here of the "other" portion of the bias, that of the bias estimator itself. The Edgeworth expansions with known studentization (not shown) and from the density case, show that choosing $\rho \not\to 0$ is preferred, and thus, in practice, $\hat{\Sigma}_q$ will indeed be bias-reduced.

where $C(x,\delta) > 0$ *is a fixed constant,* $\|t\|^2 = \sum_{d=1}^{\dim\{Z\}} t_d^2$*, and* $i = \sqrt{-1}$*.*

Assumptions 4.1 and 4.2 are standard in the literature, and Assumption 4.3 imposes the appropriate Cramér's condition for validity of the higher-order expansions. Chen and Qin (2002) discuss primitive conditions for Assumption 4.3 in the local linear case. Disallowing q even for interior points allows weakening the bound on S to $q+1$.

Before stating the result, two differences are worth mentioning, both due to the complexity of local polynomial estimators. First, the polynomials q_1, q_1, and q_3 are notationally cumbersome, and hence we defer their precise forms to the supplemental appendix. However, they retain the important properties from above: they are known, odd, and do not depend on n. Here, in addition to moments of the kernel, they depend on features of the data generating process.

Second, the biases η_{us} and η_{bc} are not as conceptually simple. The closest parallel to the density would be (for example) $\eta_{\mathsf{us}} = \sqrt{nh}(\mathbb{E}[\hat{m}] - m)$, but this can not be used due to the presence of Γ_p^{-1} inside the expectation, and next natural choice, the conditional bias $\sqrt{nh}(\mathbb{E}[\hat{m}|X_1,\dots X_n] - m)$, is still random. Instead, η_{us} and η_{bc} are biases computed after replacing Γ_p^{-1}, Γ_q^{-1}, and Λ_p with their limiting counterparts (again, precise forms are in the supplement). This retains the spirit of the conditional biases while still matching the rates of the density case.

Our main, generic result on coverage error for local polynomials is the following.

Theorem 3 (Coverage error)**.** *Let Assumptions 4.1, 4.2, and 4.3 hold, and* $nh/\log(n) \to \infty$*.*

(a) If $\eta_{\mathsf{us}} \to 0$*, then*

$$\begin{aligned}\mathbb{P}[m \in I_{\mathsf{us}}] = 1 - \alpha + \left\{ \frac{1}{nh} q_{1,\mathsf{us}}(z_{\frac{\alpha}{2}}) + \eta_{\mathsf{us}}^2 q_{2,\mathsf{us}}(z_{\frac{\alpha}{2}}) + \frac{\eta_{\mathsf{us}}}{\sqrt{nh}} q_{3,\mathsf{us}}(z_{\frac{\alpha}{2}}) \right\} \phi(z_{\frac{\alpha}{2}}) \\ + o\left((nh)^{-1} + \eta_{\mathsf{us}}^2 + \eta_{\mathsf{us}}(nh)^{-1/2} \right).\end{aligned}$$

(b) If $\eta_{\mathsf{bc}} \to 0$ *and* $\rho \to 0$*, then*

$$\begin{aligned}\mathbb{P}[m \in I_{\mathsf{bc}}] = 1 - \alpha + \left\{ \frac{1}{nh} q_{1,\mathsf{us}}(z_{\frac{\alpha}{2}}) + \eta_{\mathsf{bc}}^2 q_{2,\mathsf{us}}(z_{\frac{\alpha}{2}}) + \frac{\eta_{\mathsf{bc}}}{\sqrt{nh}} q_{3,\mathsf{us}}(z_{\frac{\alpha}{2}}) \right\} \phi(z_{\frac{\alpha}{2}}) \\ + \rho^{p+2}(\Sigma_1 + \rho^{p+1}\Sigma_2)\phi(z_{\frac{\alpha}{2}}) z_{\frac{\alpha}{2}} \\ + o\left((nh)^{-1} + \eta_{\mathsf{bc}}^2 + \eta_{\mathsf{bc}}(nh)^{-1/2} \right).\end{aligned}$$

(c) If $\eta_{\mathsf{bc}} \to 0$ *and* $\rho \to \bar{\rho} < \infty$*, then*

$$\begin{aligned}\mathbb{P}[m \in I_{\mathtt{rbc}}] = 1 - \alpha + \left\{ \frac{1}{nh} q_{1,\mathtt{rbc}}(z_{\frac{\alpha}{2}}) + \eta_{\mathsf{bc}}^2 q_{2,\mathtt{rbc}}(z_{\frac{\alpha}{2}}) + \frac{\eta_{\mathsf{bc}}}{\sqrt{nh}} q_{3,\mathtt{rbc}}(z_{\frac{\alpha}{2}}) \right\} \phi(z_{\frac{\alpha}{2}}) \\ + o\left((nh)^{-1} + \eta_{\mathsf{bc}}^2 + \eta_{\mathsf{bc}}(nh)^{-1/2} \right).\end{aligned}$$

This theorem establishes that the advantages of robust bias correction found in the density case carry over to local polynomial regression, as claimed above. Further, this result covers both

interior and boundary points. In particular, the decay rate in coverage error does not change at the boundary for any of the three estimators, and no unpleasant additional assumptions are needed as we had in Section 3.7. This is, in some sense, expected: one of the strengths of local polynomial estimation is its adaptability to boundary point estimation. This result requires only employing the appropriate fixed-n standard error estimate, i.e. $\hat{\sigma}^2_{\texttt{us}}$ or $\hat{\sigma}^2_{\texttt{rbc}}$. The constants will change near the boundary, replacing integrals over the kernel with appropriate truncated versions (see Fan and Gijbels, 1996). The higher-order bias behavior will also change, depending on choice of q, as considered next.

4.3 Bandwidth Choices and Implications for Empirical Practice

The bandwidth selection discussion for density estimation (Section 3.4) largely carries over to local polynomial regression. In particular, the MSE-optimal bandwidth is valid for robust bias correction and may yield best-possible rates, but is not valid for undersmoothing. Coverage error optimal bandwidths can be derived as well, and similar conclusions are found. And as before, $\rho = 1$ is a simple choice with good properties. However, there are a few important differences, and we will spell these out separately for inference at the boundary and the interior, due to the higher-order bias behavior in the two cases.

First, consider the interior. As before, $\eta_{\texttt{bc}}$ has two pieces. To determine their exact order, we must assume f is differentiable and $S \geq q+2$ in Assumption 4.1. Then, due to well-known symmetries in local polynomial estimation (Fan and Gijbels, 1996, Section 3.7) $\mathbb{E}[\hat{m}|X_1,\ldots,X_n] - m - B_m$ will be of order h^{p+3} and will depend on $f(x)$, $f^{(1)}(x)$, $m^{(p+2}$, and $m^{(p+3)}$. This is due to the restriction that p is odd, and is not related to choice of q. The bias of the bias estimator, on the other hand, will depend on q. In particular, $\mathbb{E}[\hat{m}^{(p+1)}|X_1,\ldots,X_n] - m^{(p+1)} \asymp b^{(q+1)-(p+1)}$ if q is odd, and $b^{(q+2)-(p+1)}$ otherwise. Thus, $\mathbb{E}[\hat{B}_m|X_1,\ldots,X_n] - B_m] \asymp h^{p+1}b^{q-p}$ for odd q and $h^{p+1}b^{q+1-p}$ for even.

At the boundary, there is no symmetry available, and $\eta_{\texttt{bc}} = O(h^{p+2} + h^{p+1}b^{q-p})$ and the only involve the unknown derivatives $m^{(p+2)}$ and $m^{(q+1)}$. Characterization of these constants does not require differentiability of f and only $S \geq q+1$. In this way, the boundary case is simpler than interior point estimation, again, for any q. This will ease implementation, as bandwidth selection procedures will be less complicated. Precise derivations are given in the supplement. Taken together, this result, the automatic boundary carpentry in coverage error, and the coverage error improvements, give strong theoretical evidence for use of robust bias corrected local polynomial regression in the regression discontinuity design, as advocated by Calonico, Cattaneo, and Titiunik (2014).

Remark 7 (Choice of ρ)**.** As in the density case, choosing $\bar{\rho} \in (0, \infty)$ yields the best-possible coverage error decay. As a consequence, given the rates above, there is no gain bias (in terms of rates) from choosing q higher than $p+1$. This provides tight guidance for implementation. Further, regarding the particular choice of ρ, Calonico, Cattaneo, and Titiunik (2014, Remark 7) point out that if $q = p+1$, $K = L$, and $\rho = 1$, then $\hat{m} - \hat{B}_m$ is algebraically identical to a local

polynomial estimator of order q (i.e., $e_0'\Gamma_q^{-1}R_q'W_qY/n$). Thus, any known optimality property of local polynomial estimators automatically justify $\rho = 1$, as this choice will deliver an equivalent kernel representation for the bias-corrected estimator that enjoys those optimality properties by construction. This again advocates for the simple choice of $\rho = 1$ in empirical work, though the discussion in Remark 5 applies here as well. ■

5 Simulation Results

To illustrate the gains from robust bias correction we conducted a small Monte Carlo study to compare undersmoothing, traditional bias correction, and robust bias correction in terms coverage accuracy, interval length, and robustness to bandwidth choice. For simplicity we restrict attention to the density. These results largely reinforce our theory: robust bias correction offers superior coverage accuracy and robustness, with a small price in length.

We generated 500 observations from a true density f and compared $I_{\texttt{us}}$, $I_{\texttt{bc}}$, and $I_{\texttt{rbc}}$ for $f(0)$, implementing a wide range of bandwidths. The true density is first taken to be standard normal and then we repeat the exercise for the mixture $(1/2)\mathcal{N}(0,1) + (1/2)\mathcal{N}(3,1)$ (these are borrowed directly from Hall (1992b)). For undersmoothing, we take K to be the Epanechnikov kernel, while robust bias correction uses the Epanechnikov and MSE-optimal kernels for K and $L^{(2)}$, respectively. Results are based on 10,000 replications.

Figure 1 shows the results when the truth is standard normal. The dashed vertical line shows $h^*_{\texttt{mse}}$. Setting $\rho = 1$, our recommended choice, shows excellent coverage properties (panel (a)), in particular that $h^*_{\texttt{mse}}$ is a valid choice. For any fixed bandwidth h, robust bias correction results in longer intervals, as shown in panel (b), however, recall that $I_{\texttt{rbc}}(h)$ will allow for, and optimally use, a larger bandwidth h, thus offsetting the length inflation. Coverage and length are further explored in the bottom panels. Panel (c) shows the empirical coverage of $I_{\texttt{rbc}}$ as both h and ρ vary, while panel (d) reports length. Again, the excellent performance at $\rho = 1$ is evident. More importantly, one can see that for a wide range of both h and ρ, coverage is accurate but length is not unduly inflated. This perhaps best demonstrates the gain from robust bias correction. Figure 2 repeat these results for the mixture $(1/2)\mathcal{N}(0,1) + (1/2)\mathcal{N}(3,1)$, and yields similar conclusions.

6 Conclusion

This paper has made three distinct, but related points regarding nonparametric inference. First, we showed that bias correction, when coupled with a new standard error formula, performs as well or better than undersmoothing for confidence interval coverage and length. Second, our results justify theoretically the popular empirical practice of using MSE optimal bandwidths, and we gave concrete implementation recommendations surrounding this, and other, choices. Third, our results showed that confidence intervals based on local polynomials do indeed have automatic boundary carpentry, provided proper studentization is used. Indeed, these results are tied together through

the themes of bias correction and higher order expansions, but also through the importance of finite sample variance calculations. Many of these messages resonate in other semi- and nonparametric contexts, and formal study of other areas is underway.

7 References

Andrews, D. W. K. (2002): "Higher-Order Improvements of a Computationally Attractive k-Step Bootstrap for Extremum Estimators," *Econometrica*, 70(1), 119–162.

Armstrong, T. B. (2014): "Adaptive Testing on a Regression Function at a Point," Working paper, Cowles Foundation discussion paper No. 1957.

Armstrong, T. B., and M. Kolesár (2014): "A Simple Adjustment for Bandwidth Snooping," Working paper, Yale University.

Calonico, S., M. D. Cattaneo, and R. Titiunik (2014): "Robust Nonparametric Confidence Intervals for Regression-Discontinuity Designs," *Econometrica, forthcoming*.

Chen, S. X., and Y. S. Qin (2002): "Confidence Intervals Based on Local Linear Smoother," *Scandinavian Journal of Statistics*, 29, 89–99.

Chernozhukov, V., D. Chetverikov, and K. Kato (2013): "Anti-Concentration and Honest Adaptive Confidence Bands," *arXiv:1303.7152*.

Fan, J., and I. Gijbels (1996): *Local polynomial modelling and its applications*. Chapman and Hall, London.

Fan, J., and T.-C. Hu (1992): "Bias Correction and Higher Order Kernel Functions," *Statistics & Probability Letters*, 13(3), 235–243.

Gasser, T., H.-G. Muller, and V. Mammitzsch (1985): "Kernels for Nonparametric Curve Estimation," *Journal of the Royal Statistical Society. Series B*, 47(2), 238–252.

Gine, E., and R. Nickl (2010): "Confidence Bands In Density Estimation," *Annals of Statistics*, 38(2), 1122–1170.

Hall, P. (1991): "Edgeworth Expansions for Nonparametric Density Estimators, with Applications," *Statistics*, 22(2), 215–232.

——— (1992a): *The Bootstrap and Edgeworth Expansion*. Springer-Verlag, New York.

——— (1992b): "Effect of Bias Estimation on Coverage Accuracy of Bootstrap Confidence Intervals for a Probability Density," *Annals of Statistics*, 20(2), 675–694.

——— (1993): "On Edgeworth Expansion and Bootstrap Confidence Bands in Nonparametric Curve Estimation," *Journal of the Royal Statistical Society. Series B*, 55(1), 291–304.

Hall, P., and J. L. Horowitz (1996): "Bootstrap Critical Values for Tests Based on Generalized-Method-of-Moments Estimators," *Econometrica*, 64(4), 891–916.

——— (2013): "A Simple Bootstrap Method for Constructing Nonparametric Confidence Bands for Functions," *Annals of Statistics*, 41(4), 1892–1921.

Hall, P., and K.-H. Kang (2001): "Bootstrapping Nonparametric Density Estimators with Empirically Chosen Bandwidths," *Annals of Statistics*, 29(5), 1443–1468.

Hansen, B. E. (2014): "Robust Inference," *working paper*.

Horowitz, J. L. (2001): "The Bootstrap," in *Handbook of Econometrics*, ed. by J. Heckman, and E. Leamer, vol. 5 of *Handbook of Econometrics*, chap. 52. Elsevier.

——— (2009): *Semiparametric and Nonparametric Methods in Econometrics*. Springer.

Ichimura, H., and P. E. Todd (2007): "Implementing Nonparametric and Semiparametric Estimators," in *Handbook of Econometrics*, ed. by J. Heckman, and E. Leamer, vol. 6B of *Handbook of Econometrics*, chap. 74. Elsevier.

Jansson, M. (2004): "The Error in Rejection Probability of Simple Autocorrelation Robust Tests," *Econometrica*, 72(3), 937–946.

Jones, M. C. (1995): "On Higher Order Kernels," *Journal of Nonparametric Statistics*, 5, 215–221.

Jones, M. C., and P. J. Foster (1993): "Generalized Jackknifing and Higher Order Kernels," *Journal of Nonparametric Statistics*, 3, 81–94.

Jones, M. C., and D. F. Signorini (1997): "A Comparison of Higher-Order Bias Kernel Density Estimators," *Journal of the American Statistical Association*, 92(439), 1063–1073.

Karunamuni, R. J., and T. Alberts (2005): "On boundary correction in kernel density estimation," *Statistical Methodology*, 2, 191–212.

Kiefer, N. M., and T. J. Vogelsang (2005): "A new asymptotic theory for heteroskedasticity-autocorrelation robust tests," *Econometric Theory*, 21(6), 1130–1164.

Kline, P., and A. Santos (2012): "Higher order properties of the wild bootstrap under misspecification," *Journal of Econometrics*, 171, 54–70.

Li, Q., and J. Racine (2007): *Nonparametric Econometrics*. Princeton, Princeton.

Neumann, M. H. (1997): "Pointwise confidence intervals in nonparametric regression with heteroscedastic error structure," *Statistics*, 29, 1–36.

Nishiyama, Y., and P. M. Robinson (2005): "The Bootstrap and the Edgeworth Correction for Semiparametric Averaged Derivatives," *Econometrica*, 73(3), 903–948.

Ruppert, D., M. P. Wand, and R. Carroll (2009): *Semiparametric Regression*. Cambridge University Press, New York.

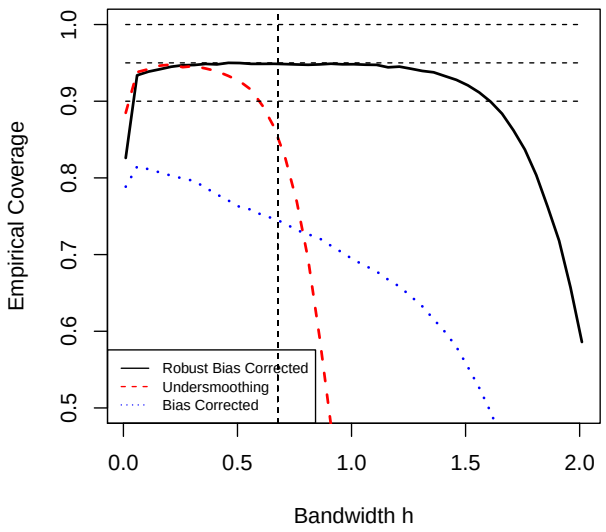
Sun, Y., P. C. B. Phillips, and S. Jin (2008): "Optimal Bandwidth Selection in Heteroskedasticity-Autocorrelation Robust Testing," *Econometrica*, 76(1), 175–194.

Tsybakov, A. B. (2003): *Introduction to Nonparametric Econometrics*. Springer, Paris.

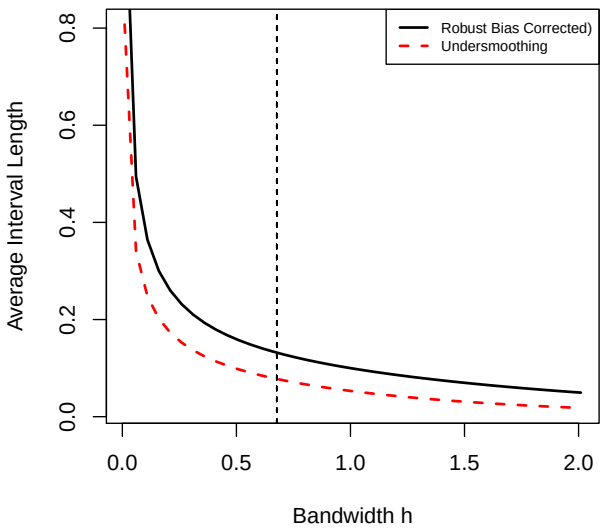
Wand, M., and M. Jones (1995): *Kernel Smoothing*. Chapman & Hall/CRC, Florida.

Xia, Y. (1998): "Bias-corrected confidence bands in nonparametric regression," *Journal of the Royal Statistical Society. Series B*, 60(4), 797–811.

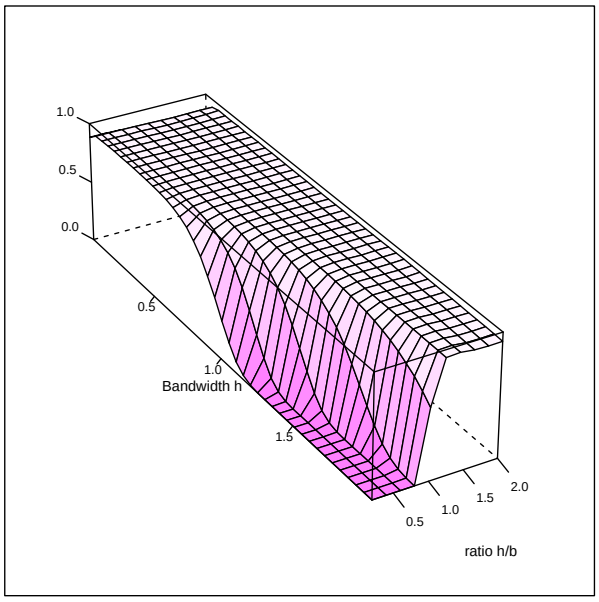
Figure 1: Comparing Undersmoothing, Traditional Bias Correction, and Robust Bias Correction.



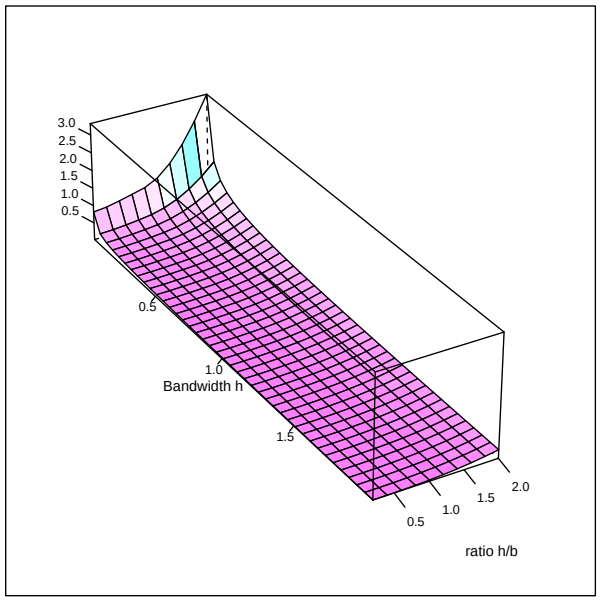
(a) Empirical Coverage



(b) Empirical Length



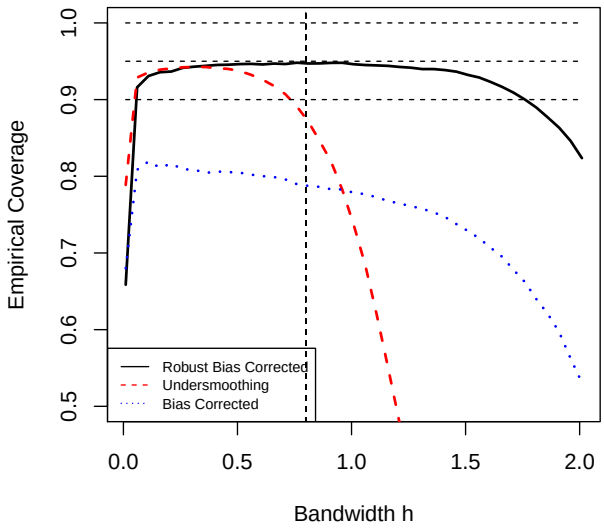
(c) Empirical Coverage



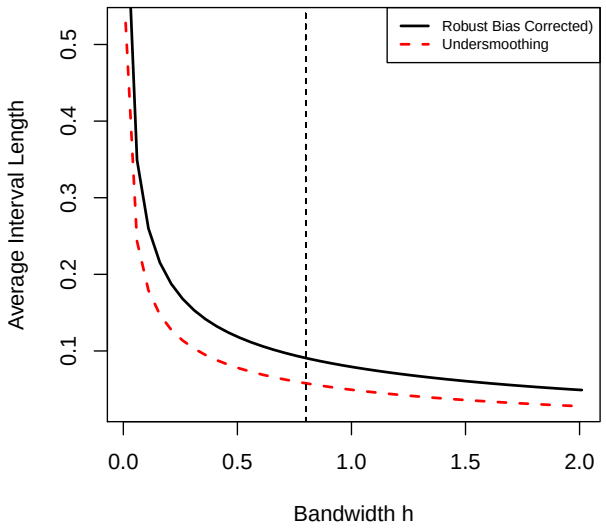
(d) Empirical Length

Notes: (i) The underlying true density is $\mathcal{N}(0,1)$ and the parameter of interest is $f(0)$; (ii) In panels (a) and (b), the vertical dotted line shows the mean-square error optimal bandwidth $h^*_{\texttt{mse}}$; (iii) In panels (c) and (d) we present Robust Bias Correction empirical coverage and interval length as the bandwidth h and ratio $\rho = h/b$ vary.

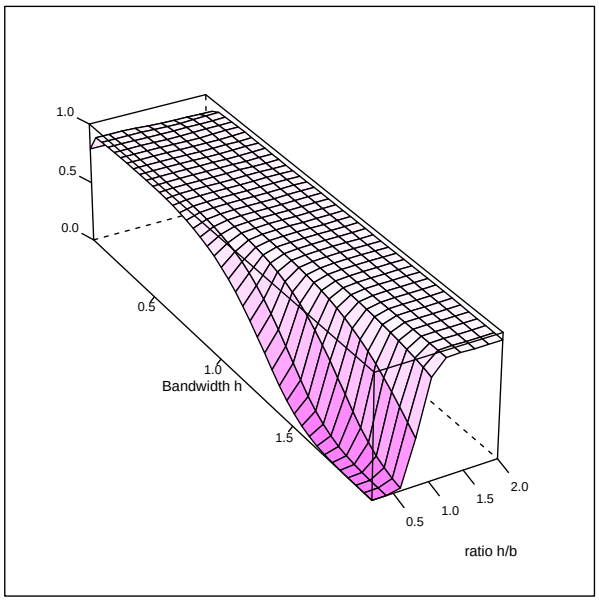
Figure 2: Comparing Undersmoothing, Traditional Bias Correction, and Robust Bias Correction.



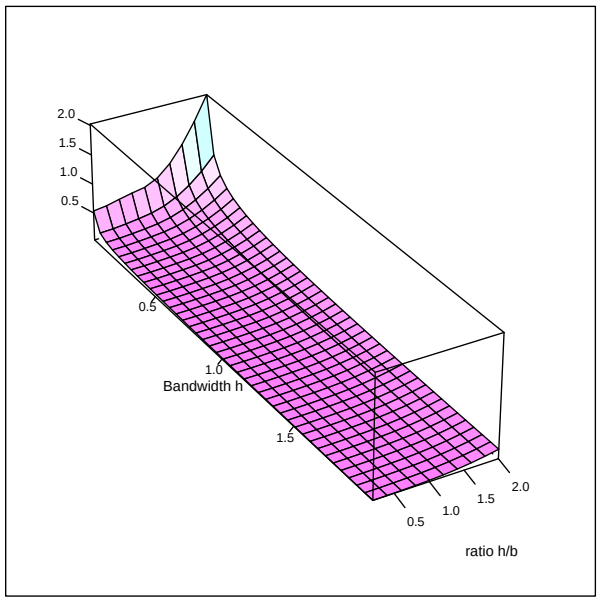
(a) Empirical Coverage



(b) Empirical Length



(c) Empirical Coverage



(d) Empirical Length

Notes: (i) The underlying true density is $(1/2)\mathcal{N}(0,1) + (1/2)\mathcal{N}(3,1)$ and the parameter of interest is $f(0)$; (ii) In panels (a) and (b), the vertical dotted line shows the mean-square error optimal bandwidth $h^*_{\texttt{mse}}$; (iii) In panels (c) and (d) we present Robust Bias Correction empirical coverage and interval length as the bandwidth h and ratio $\rho = h/b$ vary.