JOURNAL OF THE AMERICAN STATISTICAL ASSOCIATION
2018, VOL. 113, NO. 522, 767–779, Theory and Methods
https://doi.org/10.1080/01621459.2017.1285776

Taylor & Francis
Taylor & Francis Group

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

Sebastian Calonico[a], Matias D. Cattaneo[b], and Max H. Farrell[c]

[a]Department of Economics, University of Miami, Coral Gables, FL; [b]Department of Economics, Department of Statistics, University of Michigan, Ann Arbor, MI; [c]Booth School of Business, University of Chicago, Chicago, IL

ABSTRACT

Nonparametric methods play a central role in modern empirical work. While they provide inference procedures that are more robust to parametric misspecification bias, they may be quite sensitive to tuning parameter choices. We study the effects of bias correction on confidence interval coverage in the context of kernel density and local polynomial regression estimation, and prove that bias correction can be preferred to undersmoothing for minimizing coverage error and increasing robustness to tuning parameter choice. This is achieved using a novel, yet simple, Studentization, which leads to a new way of constructing kernel-based bias-corrected confidence intervals. In addition, for practical cases, we derive coverage error optimal bandwidths and discuss easy-to-implement bandwidth selectors. For interior points, we show that the mean-squared error (MSE)-optimal bandwidth for the original point estimator (before bias correction) delivers the fastest coverage error decay rate after bias correction when second-order (equivalent) kernels are employed, but is otherwise suboptimal because it is too "large." Finally, for odd-degree local polynomial regression, we show that, as with point estimation, coverage error adapts to boundary points automatically when appropriate Studentization is used; however, the MSE-optimal bandwidth for the original point estimator is suboptimal. All the results are established using valid Edgeworth expansions and illustrated with simulated data. Our findings have important consequences for empirical work as they indicate that bias-corrected confidence intervals, coupled with appropriate standard errors, have smaller coverage error and are less sensitive to tuning parameter choices in practically relevant cases where additional smoothness is available. Supplementary materials for this article are available online.

ARTICLE HISTORY
Received January 2016
Revised November 2016

KEYWORDS
Coverage error; Edgeworth expansion; Kernel methods; Local polynomial regression

1. Introduction

Nonparametric methods are widely employed in empirical work, as they provide point estimates and inference procedures that are 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 settings. However, these methods require specifying a bandwidth and their performance in applications crucially depends 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 scale in either direction can greatly skew results. This article studies kernel density and local polynomial regression estimation and inference based on the popular Wald-type statistics and demonstrates (via higher-order expansions) that by coupling explicit bias correction 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 is to minimize the asymptotic mean-square error (MSE) of the point estimator, and then use this bandwidth choice even when the 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) wrote: "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 the MSE-optimal bandwidth (in particular) 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 distributional approximations and 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 bias that caused the invalidity of the inference procedure in the first place.

CONTACT Max H. Farrell max.farrell@chicagobooth.edu Booth School of Business, University of Chicago, Chicago, IL 60637.
Color versions of one or more of the figures in the article can be found online at *www.tandfonline.com/r/JASA*.
Supplementary materials for this article are available online. Please go to *www.tandfonline.com/r/JASA*.

It has long been believed that undersmoothing is preferable for two reasons. First, theoretical studies showed inferior asymptotic coverage properties of bias-corrected confidence intervals. The pivotal work was done by Hall (1992b), and has been relied upon since. Second, implementation of bias correction is perceived as more complex because a second (usually different) bandwidth is required, deterring practitioners. However, we show theoretically that bias correction is always as good as undersmoothing, and better in many practically relevant cases, if the new standard errors that we derive are used. Further, our findings have important implications for empirical work because the resulting confidence intervals are more robust to bandwidth choice, including to the bandwidth used for bias estimation. Indeed, the two bandwidths may be set equal, a simple and automatic choice that performs well in practice and is optimal in certain objective senses.

Our proposed robust bias correction method delivers valid confidence intervals (and related inference procedures) even when using the MSE-optimal bandwidth for the original point estimator, the most popular approach in practice. Moreover, we show that at interior points, when using second-order kernels or local linear regressions, the coverage error of such intervals vanishes at the best possible rate. (Throughout, the notion of "optimal" or "best" rate is defined as the fastest achievable coverage error decay for a *fixed* kernel order or polynomial degree; and is also different from optimizing point estimation.) When higher-order kernels are used, or boundary points are considered, we find that the corresponding MSE-optimal bandwidth leads to asymptotically valid intervals, but with suboptimal coverage error decay rates, and must be shrunk (sometimes considerably) for better inference.

Heuristically, employing the MSE-optimal bandwidth for the original point estimator, prior to bias correction, is like undersmoothing the bias-corrected point estimator, though the latter estimator employs a possibly random, n-varying kernel, and requires a different Studentization scheme. It follows that the conventional MSE-optimal bandwidth commonly used in practice need not be optimal, even after robust bias correction, when the goal is inference. Thus, we present new coverage error optimal bandwidths and a fully data-driven direct plug-in implementation thereof, for use in applications. In addition, we study the important related issue of asymptotic length of the new confidence intervals.

Our comparisons of undersmoothing and bias correction are based on Edgeworth expansions for density estimation and local polynomial regression, allowing for different levels of smoothness of the unknown functions. We prove that explicit bias correction, coupled with our proposed standard errors, yields confidence intervals with coverage that is as accurate, or better, than undersmoothing (or, equivalently, yields dual hypothesis tests with lower error in rejection probability). 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 from bias correction. To be more specific, our robust bias correction approach yields higher-order refinements whenever additional smoothness is available, and is asymptotically equivalent to the best undersmoothing procedure when no additional smoothness is available.

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 kernel regression. The key distinction is that their expansions, while imposing the same levels of smoothness as we do, crucially relied on the assumption that the bias correction was first-order negligible, essentially forcing bias correction to remove less bias than undersmoothing. In contrast, we allow the bias estimator to potentially have a first-order impact, an alternative asymptotic experiment designed to more closely mimic the finite-sample behavior of bias correction. Therefore, our results formally show that whenever additional smoothness is available to characterize leading bias terms, as is usually the case in practice where MSE-optimal bandwidth are employed, our robust bias correction approach yields higher-order improvements relative to standard undersmoothing.

Our standard error formulas are based on fixed-n calculations, as opposed to asymptotics, 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 establish higher-order boundary carpentry of local polynomial based confidence intervals that use a fixed-n standard error formula, a result that is of independent (but related) interest.

This article is connected to the well-established literature on nonparametric smoothing, see Wand and Jones (1995), Fan and Gijbels (1996), Horowitz (2009), and Ruppert, Wand, and Carroll (2009) for reviews. For more recent work on bias and related issues in nonparametric inference, see Hall and Horowitz (2013), Calonico, Cattaneo, and Titiunik (2014), Armstrong and Kolesár (2017), Schennach (2015), and references therein. We also contribute to the literature on Edgeworth expansions, which have been used both in parametric and, less frequently, nonparametric contexts; see, for example, Bhattacharya and Rao (1976) and Hall (1992a). Fixed-n versus asymptotic-based Studentization has also captured some recent interest in other contexts, for example, Mykland and Zhang (2017). Finally, see Calonico, Cattaneo, and Farrell (2016) for uniformly valid Edgeworth expansions and optimal inference.

The article proceeds as follows. Section 2 studies density estimation at interior points and states the main results on error in coverage probability and its relationship to bias reduction and underlying smoothness, as well as discussing bandwidth choice and interval length. Section 3 then studies local polynomial estimation at interior and boundary points. Practical guidance is explicitly discussed in Sections 2.4 and 3.3, respectively; all methods are available in R and STATA via the `nprobust` package, see Calonico, Cattaneo, and Farrell (2017). Section 4 summarizes the results of a Monte Carlo study, and Section 5 concludes. Some technical details, all proofs, and

additional simulation evidence are collected in a lengthy online supplement.

2. Density Estimation and Inference

We first present our main ideas and conclusions for inference on the density at an interior point, as this requires relatively little notation. The data are assumed to obey the following.

Assumption 1 (Data-generating process). $\{X_1, \dots, X_n\}$ 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 $f^{(s)}$, $s = 1, 2, \dots, S$, and $f^{(S)}$ is Hölder continuous with exponent ς.

The parameter of interest is $f(x)$ for a fixed scalar point x in the interior of the support. (In the supplemental appendix, we discuss how our results extend naturally to multivariate X_i and derivative estimation.) The classical kernel-based estimator of $f(x)$ is

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

for a kernel function K that integrates to 1 and positive bandwidth $h \to 0$ as $n \to \infty$. The choice of h can be delicate, and our work is motivated in part by the standard empirical practice of employing the MSE-optimal bandwidth choice for $\hat{f}(x)$ when conducting inference.

In this vein, let us suppose for the moment that K is a kernel of order $\mathfrak{k}$, where $\mathfrak{k} \leq S$ so that the MSE-optimal bandwidth can be characterized. The bias is then given by

$$\mathbb{E}[\hat{f}(x)] - f(x) = h^{\mathfrak{k}} f^{(\mathfrak{k})}(x)\mu_{K,\mathfrak{k}} + o(h^{\mathfrak{k}}), \tag{2}$$

where $f^{(\mathfrak{k})}(x) := \partial^{\mathfrak{k}} f(x)/\partial x^{\mathfrak{k}}$ and $\mu_{K,\mathfrak{k}} = \int u^{\mathfrak{k}} K(u)du/\mathfrak{k}!$. Computing the variance gives

$$(nh)\mathbb{V}[\hat{f}(x)] = \frac{1}{h}\left\{\mathbb{E}\left[K\left(\frac{x - X_i}{h}\right)^2\right] - \mathbb{E}\left[K\left(\frac{x - X_i}{h}\right)\right]^2\right\}, \tag{3}$$

which is *nonasymptotic*: n and h are fixed in this calculation. Using other, first-order valid approximations, for example, $(nh)\mathbb{V}[\hat{f}(x)] \approx f(x)\int K(u)^2 du$, will have finite sample consequences that manifest as additional terms in the Edgeworth expansions. In fact, Section 3 shows that using an asymptotic variance for local polynomial regression removes automatic coverage-error boundary adaptivity.

Together, the prior two displays are used to characterize the MSE-optimal bandwidth, $h^*_{\mathtt{mse}} \propto n^{-1/(1+2\mathfrak{k})}$. However, using this bandwidth leaves a bias that is too large, relative to the variance, to conduct valid inference for $f(x)$. To address this important practical problem, researchers must either undersmooth the point estimator (i.e., construct $\hat{f}(x)$ with a bandwidth smaller than $h^*_{\mathtt{mse}}$) or bias-correct the point estimator (i.e., subtract an estimate of the leading bias). Thus, the question we seek to answer is this: if the bias is given by (2), is one better off estimating the leading bias (explicit bias correction) or choosing h small enough to render the bias negligible (undersmoothing) when forming nonparametric confidence intervals?

To answer this question, and to motivate our new robust approach, we first detail the bias correction and variance estimators. Explicit bias correction estimates the leading term of Equation (2), denoted by B_f, using a kernel estimator of $f^{(\mathfrak{k})}(x)$, defined as

$$\hat{B}_f = h^{\mathfrak{k}}\hat{f}^{(\mathfrak{k})}(x)\mu_{K,\mathfrak{k}},$$

where

$$\hat{f}^{(\mathfrak{k})}(x) = \frac{1}{nb^{1+\mathfrak{k}}}\sum_{i=1}^{n} L^{(\mathfrak{k})}\left(\frac{x - X_i}{b}\right),$$

for a kernel $L(\cdot)$ of order ℓ and a bandwidth $b \to 0$ as $n \to \infty$. Importantly, $\hat{B}_f$ takes this form for any $\mathfrak{k}$ and S, even if (2) fails; see Sections 2.2 and 2.3 for discussion. Conventional Studentized statistics based on undersmoothing and explicit bias correction are, respectively,

$$T_{\mathtt{us}}(x) = \frac{\sqrt{nh}\big(\hat{f}(x) - f(x)\big)}{\hat{\sigma}_{\mathtt{us}}}$$

and

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

where $\hat{\sigma}^2_{\mathtt{us}} := \hat{\mathbb{V}}[\hat{f}(x)]$ is the natural estimator of the variance of $\hat{f}(x)$, which only replaces the two expectations in (3) with sample averages, thus maintaining the nonasymptotic spirit. These are the two statistics compared in the influential article of Hall (1992b), under the same assumption imposed herein.

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 bias is asymptotically negligible after scaling, whereas $T_{\mathtt{bc}}$ allows for slower decay by virtue of the manual estimation of the leading bias. Second, $T_{\mathtt{bc}}$ requires that the variance of $h^{\mathfrak{k}}\hat{f}^{(\mathfrak{k})}(x)\mu_{K,\mathfrak{k}}$ be first-order asymptotically negligible: $\hat{\sigma}_{\mathtt{us}}$ in the denominator only accounts for the variance of the main estimate, but $\hat{f}^{(\mathfrak{k})}(x)$, being a kernel-based estimator, naturally has a variance controlled by its bandwidth. That is, even though $\hat{\sigma}^2_{\mathtt{us}}$ is based on a fixed-n calculation, the variance of the numerator of $T_{\mathtt{bc}}$ only coincides with the denominator asymptotically. Under this regime, Hall (1992b) showed that the bias reduction achieved in $T_{\mathtt{bc}}$ is too expensive in terms of noise and that undersmoothing dominates explicit bias correction for coverage error.

We argue that there need not be such a "mismatch" between the numerator of the bias-corrected statistic and the Studentization, and thus consider a third option corresponding to the idea of capturing the finite sample variability of $\hat{f}^{(\mathfrak{k})}(x)$ directly. To do so, note that we may write, after setting $\rho = h/b$,

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

We then define the collective variance of the density estimate and the bias correction as $\sigma^2_{\mathtt{rbc}} = (nh)\mathbb{V}[\hat{f}(x) - \hat{B}_f]$, exactly as

in Equation (3), but with $M(\cdot)$ in place of $K(\cdot)$, and its estimator $\hat{\sigma}^2_{\mathtt{rbc}}$ exactly as $\hat{\sigma}^2_{\mathtt{us}}$. Therefore, our proposed robust bias-corrected inference approach is based on

$$T_{\mathtt{rbc}} = \frac{\sqrt{nh}\big(\hat{f}(x) - h^{\hbar}\,\hat{f}^{(\hbar)}(x)\mu_{K,\hbar} - f(x)\big)}{\hat{\sigma}_{\mathtt{rbc}}}.$$

That is, our proposed standard errors are based on a fixed-n calculation that captures the variability of both $\hat{f}(x)$ and $\hat{f}^{(\hbar)}(x)$, and their covariance. As shown in Section 3, the case of local polynomial regression is analogous, but notationally more complicated.

The quantity $\rho = h/b$ is key. If $\rho \to 0$, then the second term of M is dominated by the first, that is, the bias correction is first-order negligible. In this case, $\sigma^2_{\mathtt{us}}$ and $\sigma^2_{\mathtt{rbc}}$ (and their estimators) will be first-order, 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. To more accurately capture finite sample behavior of bias correction we allow ρ to converge to any (nonnegative) finite limit, allowing (but not requiring) the bias correction to be first-order important, unlike prior work. We show that doing so yields more accurate confidence intervals (i.e., higher-order corrections).

2.1. Generic Higher-Order Expansions of Coverage Error

We first present generic Edgeworth expansions for all three procedures (undersmoothing, traditional bias correction, and robust bias correction), which are agnostic regarding the level of available smoothness (controlled by S in Assumption 1). To be specific, we give higher-order expansions of the error in coverage probability of the following $(1-\alpha)$% confidence intervals based on Normal approximations for the statistics $T_{\mathtt{us}}$, $T_{\mathtt{bc}}$, and $T_{\mathtt{rbc}}$:

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

and

$$I_{\mathtt{rbc}} = \left[\hat{f} - \hat{B}_f - z_{1-\frac{\alpha}{2}}\frac{\hat{\sigma}_{\mathtt{rbc}}}{\sqrt{nh}},\ \ \hat{f} - \hat{B}_f - z_{\frac{\alpha}{2}}\frac{\hat{\sigma}_{\mathtt{rbc}}}{\sqrt{nh}}\right], \quad (5)$$

where z_α is the upper α-percentile of the Gaussian distribution. Here and in the sequel, we omit the point of evaluation x for simplicity. Equivalently, our results can characterize the error in rejection probability of the corresponding hypothesis tests. In subsequent sections, we give specific results under different smoothness assumptions and make direct comparisons of the methods.

We require the following standard conditions on the kernels K and L.

Assumption 2 (Kernels). The kernels K and L are bounded, even functions with support $[-1, 1]$, and are of order $\hbar \geq 2$ and $\ell \geq 2$, respectively, where $\hbar$ and ℓ are even integers. That is, $\mu_{K,0} = 1$, $\mu_{K,k} = 0$ for $1 \leq k < \hbar$, and $\mu_{K,\hbar} \neq 0$ and bounded, and similarly for $\mu_{L,k}$ with ℓ in place of $\hbar$. Further, L is $\hbar$-times continuously differentiable. For all integers k and l such that $k + l = \hbar - 1$, $f^{(k)}(x_0)L^{(l)}((x_0 - x)/b) = 0$ for x_0 in the boundary of the support.

The boundary conditions are needed for the derivative estimation inherent in bias correction, even if x is an interior point, and are satisfied if the support of f is the whole real line. Higher-order results also require a standard n-varying Cramér's condition, given in the supplement to conserve space (see Section S.I.3). Altogether, our assumptions are identical to those of Hall (1991, 1992b).

To state the results some notation is required. First, let the (scaled) biases of the density estimator and the bias-corrected estimator be $\eta_{\mathtt{us}} = \sqrt{nh}(\mathbb{E}[\hat{f}] - f)$ and $\eta_{\mathtt{bc}} = \sqrt{nh}(\mathbb{E}[\hat{f} - \hat{B}_f] - f)$. Next, let $\phi(z)$ be the standard Normal density, and for any kernel K define

$$q_1(K) = \vartheta_{K,2}^{-2}\vartheta_{K,4}\left(z_{\frac{\alpha}{2}}^3 - 3z_{\frac{\alpha}{2}}\right)/6 - \vartheta_{K,2}^{-3}\vartheta_{K,3}^2 \times \left[2z^3/3 + \left(z_{\frac{\alpha}{2}}^5 - 10z_{\frac{\alpha}{2}}^3 + 15z_{\frac{\alpha}{2}}\right)/9\right],$$
$$q_2(K) = -\vartheta_{K,2}^{-1}\,z_{\frac{\alpha}{2}},$$

and

$$q_3(K) = \vartheta_{K,2}^{-2}\vartheta_{K,3}\left(2z_{\frac{\alpha}{2}}^3/3\right),$$

where $\vartheta_{K,k} = \int K(u)^k du$. All that is conceptually important is that these functions are known, odd polynomials in z with coefficients that depend only on the kernel, and not on the sample or data-generating process. Our main theoretical result for density estimation is the following.

Theorem 1. Let Assumptions 1, 2, and Cramér's condition hold and $nh/\log(nh) \to \infty$.

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

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

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

$$\mathbb{P}[f \in I_{\mathtt{bc}}] = 1 - \alpha + \left\{\frac{1}{nh}q_1(K) + \eta_{\mathtt{bc}}^2 q_2(K) + \frac{\eta_{\mathtt{bc}}}{\sqrt{nh}}q_3(K)\right\} \frac{\phi(z_{\frac{\alpha}{2}})}{f}\{1 + o(1)\} + \rho^{1+\hbar}(\Omega_1 + \rho^{\hbar}\Omega_2)\phi(z_{\frac{\alpha}{2}})z_{\frac{\alpha}{2}}\{1 + o(1)\},$$

for constants Ω_1 and Ω_2 given precisely in the supplement.

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

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

This result leaves the scaled biases $\eta_{\mathtt{us}}$ and $\eta_{\mathtt{bc}}$ generic, which is useful when considering different levels of smoothness S, the choices of $\hbar$ and ℓ, and in comparing to local polynomial results. In the next subsection, we make these quantities more precise

and compare them, paying particular attention to the role of the underlying smoothness assumed.

At present, the most visually obvious feature of this result is that all the error terms are of the same form, except for the notable presence of $\rho^{1+\mathfrak{k}}(\Omega_1 + \rho^{\mathfrak{k}}\Omega_2)$ in part (b). These are the leading terms of $\sigma^2_{\mathrm{rbc}}/\sigma^2_{\mathrm{us}} - 1$, consisting of the covariance of $\hat{f}$ and $\hat{B}_f$ (denoted by Ω_1) and the variance of $\hat{B}_f$ (denoted by Ω_2), and are entirely due to the "mismatch" in the Studentization of T_{bc}. 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 because the variance is not well-controlled: in any finite sample, $\hat{B}_f$ would inject variability (i.e., $\rho = h/b > 0$ for each n) and thus $\rho \to 0$ may not be a good approximation. Our new Studentization does not simply remove these leading ρ terms; the entire sequence is absent. As explained below, allowing for $\bar{\rho} = \infty$ cannot reduce bias, but will inflate variance; hence restricting to $\bar{\rho} < \infty$ capitalizes fully on the improvements from bias correction.

2.2. Coverage Error and the Role of Smoothness

Theorem 1 makes no explicit assumption about smoothness beyond the requirement that the scaled biases vanish asymptotically. The fact that the error terms in parts (a) and (c) of Theorem 1 take the same form implies that comparing coverage error amounts to comparing bias, for which the smoothness S and the kernel orders $\mathfrak{k}$ and ℓ are crucial. We now make the biases η_{us} and η_{bc} concrete and show how coverage is affected.

For I_{us}, two cases emerge: (a) enough derivatives exist to allow characterization of the MSE-optimal bandwidth ($\mathfrak{k} \leq S$); and (b) no such smoothness is available ($\mathfrak{k} > S$), in which case the leading term of Equation (2) is exactly zero and the bias depends on the unknown Hölder constant. These two cases lead to the following results.

Corollary 1. Let Assumptions 1, 2, and Cramér's condition hold and $nh/\log(nh) \to \infty$.

(a) If $\mathfrak{k} \leq S$ and $\sqrt{nh}h^{\mathfrak{k}} \to 0$,

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

(b) If $\mathfrak{k} > S$ and $\sqrt{nh}h^{S+\varsigma} \to 0$,

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

The first result is most directly comparable to Hall (1992b, sec. 3.4), and many other past articles, which typically take as a starting point that the MSE-optimal bandwidth can be characterized. This shows that T_{us} must be undersmoothed, in the sense the MSE-optimal bandwidth is "too large" for valid inference. In fact, we know that $I_{\mathrm{us}}(h^*_{\mathrm{mse}})$ will asymptotically *under*cover because $T_{\mathrm{us}}(h^*_{\mathrm{mse}}) \to_d \mathcal{N}((2\mathfrak{k})^{-1/2}, 1)$ (see the supplement). Instead, the optimal h for coverage error, which can be characterized and estimated, is equivalent in rates to balancing variance against bias, not squared bias as in MSE. Part (b) shows that a faster rate of coverage error decay can be obtained by taking a sufficiently high-order kernel, relative to the level of smoothness S, at the expense of feasible bandwidth selection.

Turning to robust bias correction, characterization of η_{bc} is more complex as it has two pieces: the second-order bias of the original point estimator, and the bias of the bias estimator itself. The former is the $o(h^{\mathfrak{k}})$ term of Equation (2) and is not the target of explicit bias correction; it depends either on higher derivatives, if they are available, or on the Hölder condition otherwise. To be precise, if $\mathfrak{k} \leq S - 2$, this term is $[h^{\mathfrak{k}+2} + o(1)] f^{(\mathfrak{k}+2)} \mu_{K_{\mathrm{bc}},\mathfrak{k}+2}$, while otherwise is known only to be $O(h^{S+\varsigma})$. Importantly, the bandwidth b and order ℓ do not matter here, and bias reduction beyond $O(\min\{h^{\mathfrak{k}+2}, h^{S+\varsigma}\})$ is not possible; there is thus little or no loss in fixing $\ell = 2$, which we assume from now on to simplify notation.

The bias of the bias estimator also depends on the smoothness available: if enough smoothness is available the corresponding bias term can be characterized, otherwise only its order will be known. To be specific, when smoothness is not binding ($\mathfrak{k} \leq S - 2$), arguably the most practically-relevant case, the leading term of $\mathbb{E}[\hat{B}_f] - B_f$ will be $h^{\mathfrak{k}} b^2 f^{(\mathfrak{k}+2)} \mu_{K,\mathfrak{k}} \mu_{L,2}$. Smoothness can be exhausted in two ways, either by the point estimate itself ($\mathfrak{k} > S$) or by the bias estimation ($S - 1 \leq \mathfrak{k} \leq S$), and these two cases yield $O(h^{\mathfrak{k}} b^{S-\mathfrak{k}})$ and $O(h^{\mathfrak{k}} b^{S+\varsigma-\mathfrak{k}})$, respectively, which are slightly different in how they depend on the total Hölder smoothness assumed. (Complete details are in the supplement.) Note that regardless of the value of $\mathfrak{k}$, we set $\hat{B}_f = h^{\mathfrak{k}} \hat{f}^{(\mathfrak{k})} \mu_{K,\mathfrak{k}}$, even if $\mathfrak{k} > S$ and $B_f \equiv 0$.

With these calculations for η_{bc}, we have the following result.

Corollary 2. Let Assumptions 1, 2, and Cramér's condition hold, $nh/\log(nh) \to \infty$, $\rho \to \bar{\rho} < \infty$, and $\ell = 2$.

(a) If $\mathfrak{k} \leq S - 2$ and $\sqrt{nh}h^{\mathfrak{k}} b^2 \to 0$,

$$\mathbb{P}[f \in I_{\mathrm{rbc}}] = 1 - \alpha + \left\{ \frac{1}{nh} q_1(M_{\bar{\rho}}) + nh^{1+2(\mathfrak{k}+2)} (f^{(\mathfrak{k}+2)})^2 \times \left(\mu_{K,\mathfrak{k}+2} - \bar{\rho}^{-2} \mu_{K,\mathfrak{k}} \mu_{L,2}\right)^2 q_2(M_{\bar{\rho}}) + h^{\mathfrak{k}+2} f^{(\mathfrak{k}+2)} \left(\mu_{K,\mathfrak{k}+2} - \bar{\rho}^{-2} \mu_{K,\mathfrak{k}} \mu_{L,2}\right) q_3(M_{\bar{\rho}}) \right\} \times \frac{\phi(z_{\frac{\alpha}{2}})}{f} \{1 + o(1)\}.$$

(b) If $S - 1 \leq \mathfrak{k} \leq S$ and $\sqrt{nh}\rho^{\mathfrak{k}} b^{S+\varsigma} \to 0$,

$$\mathbb{P}[f \in I_{\mathrm{rbc}}] = 1 - \alpha + \frac{1}{nh} \frac{\phi(z_{\frac{\alpha}{2}})}{f} q_1(M_{\bar{\rho}}) \{1 + o(1)\} + O\left(nh\rho^{2\mathfrak{k}} b^{2(S+\varsigma)} + \rho^{\mathfrak{k}} b^{S+\varsigma}\right).$$

(c) If $\mathfrak{k} > S$ and $\sqrt{nh}\left(h^{S+\varsigma} \vee \rho^{\mathfrak{k}} b^S\right) \to 0$,

$$\mathbb{P}[f \in I_{\mathrm{rbc}}] = 1 - \alpha + \frac{1}{nh} \frac{\phi(z_{\frac{\alpha}{2}})}{f} q_1(M_{\bar{\rho}}) \{1 + o(1)\} + O\left(nh(h^{S+\varsigma} \vee \rho^{\mathfrak{k}} b^S)^2 + (h^{S+\varsigma} \vee \rho^{k} b^S)\right).$$

Part (a) is the most empirically relevant setting, which reflects the idea that researchers first select a kernel order, then conduct inference based on that choice, taking the unknown

smoothness to be nonbinding. 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 $\mathfrak{k}$ replaced by $\mathfrak{k}+2$, and represent the same conceptual objects. By virtue of our new Studentization, the leading variance remains order $(nh)^{-1}$ and the problematic correlation terms are absent. We explicitly discuss the advantages of robust bias correction relative to undersmoothing in the following section.

Part (a) also argues for a bounded, positive ρ. First, because bias reduction beyond $O(h^{\mathfrak{k}+2})$ is not possible, $\rho \to \infty$ will only inflate the variance. On the other hand, $\bar{\rho} = 0$ requires a delicate choice of b and $\ell > 2$, else the second bias term dominates $\eta_{\mathtt{bc}}$, and the full power of the variance correction is not exploited, that is, more bias may be removed without inflating the variance rate. Hall (1992b, p. 682) remarked that if $\mathbb{E}[\hat{f}] - f - B_f$ is (part of) the leading bias term, then "explicit bias correction [...] is even less attractive relative to undersmoothing." We show, on the contrary, that with our proposed Studentization, it is optimal that $\mathbb{E}[\hat{f}] - f - B_f$ is part of the dominant bias term.

Finally, in both Corollaries above the best possible coverage error decay rate (for a given S) is attained by exhausting all available smoothness. This would also yield point estimators attaining the bound of Stone (1982); robust bias correction cannot evade such bounds, of course. In both Corollaries, coverage is improved relative to part (a), but the constants and optimal bandwidths cannot be quantified. For robust bias correction, Corollary 2 shows that to obtain the best rate in part (b) the unknown $f^{(\mathfrak{k})}$ must be consistently estimated and ρ must be bounded and positive, while in part (c), bias estimation merely adds noise, but this noise is fully accounted for by our new Studentization, as long as $\rho \to 0$ ($b \not\to 0$ is allowed).

2.3. Comparing Undersmoothing and Robust Bias Correction

We now employ Corollaries 1 and 2 to directly compare nonparametric inference based on undersmoothing and robust bias correction. To simplify the discussion, we focus on three concrete cases, which illustrate how the comparisons depend on the available smoothness and kernel order; the messages generalize to any S and/or $\mathfrak{k}$. For this discussion, we let $\mathfrak{k}_{\mathtt{us}}$ and $\mathfrak{k}_{\mathtt{bc}}$ be the kernel orders used for point estimation in $I_{\mathtt{us}}$ and $I_{\mathtt{rbc}}$, respectively, and restrict attention to sequences $h \to 0$ where both confidence intervals are first-order valid, even though robust bias correction allows for a broader bandwidth range. Finally, we set $\ell = 2$ and $\bar{\rho} \in (0, \infty)$ based on the above discussion.

For the first case, assume that f is twice continuously differentiable ($S = 2$) and both methods use second-order kernels ($\mathfrak{k}_{\mathtt{us}} = \mathfrak{k}_{\mathtt{bc}} = \ell = 2$). In this case, both methods target the *same* bias. The coverage errors for $I_{\mathtt{us}}$ and $I_{\mathtt{rbc}}$ then follow directly from Corollaries 1(a) and 2(b) upon plugging in these kernel orders, yielding

$$\left|\mathbb{P}[f \in I_{\mathtt{us}}] - (1-\alpha)\right| \asymp \frac{1}{nh} + nh^5 + h^2$$

and

$$\left|\mathbb{P}[f \in I_{\mathtt{rbc}}] - (1-\alpha)\right| \asymp \frac{1}{nh} + nh^{5+2\varsigma} + h^{2+\varsigma}.$$

Because $h \to 0$ and $\bar{\rho} \in (0, \infty)$, the coverage error of $I_{\mathtt{rbc}}$ vanishes more rapidly by virtue of the bias correction. A higher-order kernel ($\mathfrak{k}_{\mathtt{us}} > 2$) would yield this rate for $I_{\mathtt{us}}$.

Second, suppose that the density is four-times continuously differentiable ($S = 4$) but second-order kernels are maintained. The relevant results are now Corollaries 1(a) and 2(a). Both methods continue to target the *same* leading bias, but now the additional smoothness available allows precise characterization of the improvement shown above, and we have

$$\left|\mathbb{P}[f \in I_{\mathtt{us}}] - (1-\alpha)\right| \asymp \frac{1}{nh} + nh^5 + h^2$$

and

$$\left|\mathbb{P}[f \in I_{\mathtt{rbc}}] - (1-\alpha)\right| \asymp \frac{1}{nh} + nh^9 + h^4.$$

This case is perhaps the most empirically relevant one, where researchers first choose the order of the kernel (here, second order) and then conduct/optimize inference based on that choice. Indeed, for this case optimal bandwidth choices can be derived (Section 2.4).

Finally, maintain $S = 4$ but suppose that undersmoothing is based on a fourth-order kernel while bias correction continues to use two second-order kernels ($\mathfrak{k}_{\mathtt{us}} = 4$, $\mathfrak{k}_{\mathtt{bc}} = \ell = 2$). This is the exact example given by Hall (1992b, p. 676). Now the two methods target *different* biases, but use the *same* amount of smoothness. In this case, the relevant results are again Corollaries 1(a) and 2(a), now with $\mathfrak{k} = 4$ and $\mathfrak{k} = 2$, respectively. The two methods have the same coverage error decay rate:

$$\left|\mathbb{P}[f \in I_{\mathtt{us}}] - (1-\alpha)\right| \asymp \left|\mathbb{P}[f \in I_{\mathtt{rbc}}] - (1-\alpha)\right| \asymp \frac{1}{nh} + nh^9 + h^4.$$

Indeed, more can be said: with the notation of Equation (4), the difference between $T_{\mathtt{us}}$ and $T_{\mathtt{rbc}}$ is the change in "kernel" from K to M, and since $\mathfrak{k}_{\mathtt{bc}} + \ell = \mathfrak{k}_{\mathtt{us}}$, the two kernels are the same order. (M acts as an n-varying, higher-order kernel for bias, but may not strictly fit the definition, as explored in the supplement.) This tight link between undersmoothing and robust bias correction does not carry over straightforwardly to local polynomial regression, as we discuss in more detail in Section 3.

In the context of this final example, it is worth revisiting traditional bias correction. The fact that undersmoothing targets a different, and asymptotically smaller, bias than does explicit bias correction, coupled with the requirement that $\rho \to 0$, implicitly constrains bias correction to remove *less* bias than undersmoothing. This is necessary for traditional bias correction, but on the contrary, robust bias correction attains the *same* coverage error decay rate as undersmoothing under the same assumptions.

In sum, these examples show that under identical assumptions, bias correction is not inferior to undersmoothing and if any additional smoothness is available, can yield improved coverage error. These results are confirmed in our simulations.

2.4. Optimal Bandwidth and Data-Driven Choice

The prior sections established that robust bias correction can equal, or outperform, undersmoothing for inference. We now show how the method can be implemented to deliver these

results in applications. We mimic typical empirical practice where researchers first choose the order of the kernel, then conduct/optimize inference based on that choice. Therefore, we assume the smoothness is unknown but taken to be large and work within Corollary 2(a), that is, viewing $\mathfrak{k} \leq S-2$ and $\ell = 2$ as fixed and ρ bounded and positive. This setup allows characterization of the coverage error optimal bandwidth for robust bias correction.

Corollary 3. Under the conditions of Corollary 2(a) with $\bar{\rho} \in (0, \infty)$, if $h = h^*_{\mathtt{rbc}} = H^*_{\mathtt{rbc}}(\rho) n^{-1/(1+(\mathfrak{k}+2))}$, then $\mathbb{P}[f \in I_{\mathtt{rbc}}] = 1 - \alpha + O(n^{-(\mathfrak{k}+2)/(1+(\mathfrak{k}+2))})$, where

$$H^*_{\mathtt{rbc}}(\bar{\rho}) = \operatorname*{arg\,min}_{H>0} \Big| H^{-1} q_1(M_{\bar{\rho}}) + H^{1+2(\mathfrak{k}+2)} (f^{(\mathfrak{k}+2)})^2 \times \left(\mu_{K,\mathfrak{k}+2} - \bar{\rho}^{-2} \mu_{K,\mathfrak{k}} \mu_{L,2}\right)^2 q_2(M_{\bar{\rho}}) + H^{\mathfrak{k}+2} f^{(\mathfrak{k}+2)} \left(\mu_{K,\mathfrak{k}+2} - \bar{\rho}^{-2} \mu_{K,\mathfrak{k}} \mu_{L,2}\right) q_3(M_{\bar{\rho}}) \Big|.$$

We can use this result to give concrete methodological recommendations. At the end of this section, we discuss the important issue of interval length. Construction of the interval $I_{\mathtt{rbc}}$ from Equation (5) requires bandwidths h and b and kernels K and L. Given these choices, the point estimate, bias correction, and variance estimators are then readily computable from data using the formulas above. For the kernels K and L, we recommend either second-order minimum variance (to minimize interval length) or MSE-optimal kernels (see, e.g., Gasser, Muller, and Mammitzsch 1985, and the supplemental appendix).

The bandwidth selections are more important in applications. For the bandwidth h, Corollary 2(a) shows that the MSE-optimal choice $h^*_{\mathtt{mse}}$ will deliver valid inference, but will be suboptimal in general (Corollary 3). From a practical point of view, the robust bias-corrected interval $I_{\mathtt{rbc}}(h)$ is attractive because it allows for the MSE-optimal bandwidth and kernel, and hence is based on the MSE-optimal point estimate, while using the same effective sample for both point estimation and inference. Interestingly, although $I_{\mathtt{rbc}}(h^*_{\mathtt{mse}})$ is always valid, its coverage error decays as $n^{-\min\{4,\mathfrak{k}+2\}/(1+2\mathfrak{k})}$ and is thus rate optimal only for second-order kernels ($\mathfrak{k} = 2$), while otherwise being suboptimal, with a rate that is lower the larger is the order $\mathfrak{k}$.

Corollary 3 gives the coverage error optimal bandwidth, $h^*_{\mathtt{rbc}}$, which can be implemented using a simple direct plug-in (DPI) rule: $\hat{h}_{\mathtt{dpi}} = \hat{H}_{\mathtt{dpi}}\, n^{-1/(\mathfrak{k}+3)}$, where $\hat{H}_{\mathtt{dpi}}$ is a plug-in estimate of $H^*_{\mathtt{rbc}}$ formed by replacing the unknown $f^{(\mathfrak{k}+2)}$ with a pilot estimate (e.g., a consistent nonparametric estimator based on the appropriate MSE-optimal bandwidth). In the supplement, we give precise implementation details, as well as an alternative rule-of-thumb bandwidth selector based on rescaling already available data-driven MSE-optimal choices.

For the bandwidth b, a simple choice is $b = h$, or, equivalently, $\rho = 1$. We show in the supplement that setting $\rho = 1$ has good theoretical properties, minimizing interval length of $I_{\mathtt{rbc}}$ or the MSE of $\hat{f} - \hat{B}_f$, depending on the conditions imposed. In our numerical work, we found that $\rho = 1$ performed well. As a result, from the practitioner's point of view, the choice of b (or ρ) is completely automatic, leaving only one bandwidth to select.

An extensive simulation study, reported in the supplement, illustrates our findings and explores the numerical performance of these choices. We find that coverage of $I_{\mathtt{rbc}}$ is robust to both h and ρ and that our data-driven bandwidth selectors work well in practice, but we note that estimating bandwidths may have higher-order implications (e.g., Hall and Kang 2001).

Finally, an important issue in applications is whether the good coverage properties of $I_{\mathtt{rbc}}$ come at the expense of increased interval length. When coverage is asymptotically correct, Corollaries 1 and 2 show that $I_{\mathtt{rbc}}$ can accommodate (and will optimally employ) a larger bandwidth (i.e., $h \to 0$ more slowly), and hence $I_{\mathtt{rbc}}$ will have shorter average length in large samples than $I_{\mathtt{us}}$. Our simulation study (see below and the supplement) gives the same conclusion.

2.5. Other Methods of Bias Correction

We study a plug-in bias correction method, but there are alternatives. In particular, as pointed out by a reviewer, a leading alternative is the generalized jackknife method by Schucany and Sommers (1977) (see Cattaneo, Crump, and Jansson (2013) for an application to kernel-based semiparametric inference and for related references). We will briefly summarize this approach and show a tight connection to our results, restricting to second-order kernels and $S \geq 2$ only for simplicity.

The generalized jackknife estimator is $\hat{f}_{\mathtt{GJ},R} := (\hat{f}_1 - R\hat{f}_2)/(1-R)$, where $\hat{f}_1$ and $\hat{f}_2$ are two initial kernel density estimators, with possibly different bandwidths (h_1, h_2) and second-order kernels (K_1, K_2). From Equation (2), the bias of $\hat{f}_{\mathtt{GJ},R}$ is $(1-R)^{-1} f^{(2)}(h_1^2 \mu_{K_1,2} - R h_2^2 \mu_{K_2,2}) + o(h_1^2 + h_2^2)$, whence choosing $R = (h_1^2 \mu_{K_1,2})/(h_2^2 \mu_{K_2,2})$ renders the leading bias term exactly zero. Further, if $S \geq 4$, $\hat{f}_{\mathtt{GJ},R}$ has bias $O(h_1^4 + h_2^4)$; behaving as a point estimator with $\mathfrak{k} = 4$. To connect this approach to ours, observe that with this choice of R and $\tilde{\rho} = h_1/h_2$,

$$\hat{f}_{\mathtt{GJ},R} = \frac{1}{n h_1} \sum_{i=1}^n \tilde{M}\left(\frac{X_i - x}{h_1}\right),$$

$$\tilde{M}(u) = K_1(u) - \tilde{\rho}^{1+2} \left\{ \frac{K_2(\tilde{\rho} u) - \tilde{\rho}^{-1} K_1(u)}{\mu_{K_2,2}(1-R)} \right\} \mu_{K_1,2},$$

exactly matching Equation (4); alternatively, write $\hat{f}_{\mathtt{GJ},R} = \hat{f}_1 - h_1^2 \tilde{f}^{(2)} \mu_{K_1,2}$, where

$$\tilde{f}^{(2)} = \frac{1}{n h_2^{1+2}} \sum_{i=1}^n \tilde{L}\left(\frac{X_i - x}{h_2}\right),$$

$$\tilde{L}(u) = \frac{K_2(u) - \tilde{\rho}^{-1} K_1(\tilde{\rho}^{-1} u)}{\mu_{K_2,2}(1-R)},$$

is a derivative estimator. Therefore, we can view $\hat{f}_{\mathtt{GJ},R}$ as a specific kernel M or a specific derivative estimator, and all our results directly apply to $\hat{f}_{\mathtt{GJ},R}$; hence our article offers a new way of conducting inference (new Studentization) for this case as well. Though we omit the details to conserve space, this is equally true for local polynomial regression (Section 3).

More generally, our main ideas and generic results apply to many other bias correction methods. For a second example, Singh (1977) also proposed a plug-in bias estimator, but without using the derivative of a kernel. Our results cover this

approach as well; see the supplement for further details and references. The key, common message in all cases is that to improve inference one must account for the additional variability introduced by any bias correction method (i.e., to avoid the mismatch present in $T_{\mathtt{bc}}$).

3. Local Polynomial Estimation and Inference

This section studies local polynomial regression (Ruppert and Wand 1994; Fan and Gijbels 1996), and has two principal aims. First, we show that the conclusions from the density case, and their implications for practice, carry over to odd-degree local polynomials. Second, we show that with proper fixed-n Studentization, coverage error adapts to boundary points. We focus on what is novel relative to the density, chiefly variance estimation and boundary points. For interior points, the implications for coverage error, bandwidth selection, and interval length are all analogous to the density case, and we will not retread those conclusions.

To be specific, throughout this section we focus on the case where the smoothness is large relative to the local polynomial degree p, which is arguably the most relevant case in practice. The results and discussion in Sections 2.2 and 2.3 carry over, essentially upon changing k to $p+1$ and ℓ to $q-p$ (or $q-p+1$ for interior points with q even). Similarly, but with increased notational burden, the conclusions of Section 2.5 also remain true. The present results also extend to multivariate data and derivative estimation.

To begin, we define the regression estimator, its bias, and the bias correction. Given a random sample $\{(Y_i, X_i) : 1 \le i \le n\}$, the local polynomial estimator of $m(x) = \mathbb{E}[Y_i|X_i = x]$, temporarily making explicit the evaluation point, is

$$\hat{m}(x) = \mathbf{e}_0'\hat{\boldsymbol{\beta}}_p,$$
$$\hat{\boldsymbol{\beta}}_p = \underset{\boldsymbol{b}\in\mathbb{R}^{p+1}}{\arg\min} \sum_{i=1}^{n} (Y_i - \mathbf{r}_p(X_i - x)'\boldsymbol{b})^2 K\left(\frac{X_i - x}{h}\right),$$

where, for an integer $p \ge 1$, $\mathbf{e}_0$ is the $(p+1)$-vector with a one in the first position and zeros in the rest, and $\mathbf{r}_p(u) = (1, u, u^2, \ldots, u^p)'$. We restrict attention to p odd, as is standard, though the qualifier may be omitted. We define $\mathbf{Y} = (Y_1, \ldots, Y_n)'$, $\mathbf{R}_p = [\mathbf{r}_p((X_1 - x)/h), \ldots, \mathbf{r}_p((X_n - x)/h)]'$, $\mathbf{W}_p = \mathrm{diag}(h^{-1}K((X_i - x)/h) : i = 1, \ldots, n)$, and $\boldsymbol{\Gamma}_p = \mathbf{R}_p'\mathbf{W}_p\mathbf{R}_p/n$ (here $\mathrm{diag}(a_i : i = 1, \ldots, n)$ denotes the $n \times n$ diagonal matrix constructed using $a_1, a_2, \ldots, a_n$). Then, reverting back to omitting the argument x, the local polynomial estimator is $\hat{m} = \mathbf{e}_0'\boldsymbol{\Gamma}_p^{-1}\mathbf{R}_p'\mathbf{W}_p\mathbf{Y}/n$.

Under regularity conditions below, the conditional bias satisfies

$$\mathbb{E}[\hat{m}|X_1, \ldots, X_n] - m = h^{p+1}m^{(p+1)}\frac{1}{(p+1)!}\mathbf{e}_0'\boldsymbol{\Gamma}_p^{-1}\boldsymbol{\Lambda}_p + o_P(h^{p+1}), \tag{6}$$

where $\boldsymbol{\Lambda}_p = \mathbf{R}_p'\mathbf{W}_p[((X_1 - x)/h)^{p+1}, \ldots, ((X_n - x)/h)^{p+1}]'/n$. Here, the quantity $\mathbf{e}_0'\boldsymbol{\Gamma}_p^{-1}\boldsymbol{\Lambda}_p/(p+1)!$ is random, unlike in the density case (see (2)), but it is known and bounded in probability. Following Fan and Gijbels (1996, p. 116), we will estimate $m^{(p+1)}$ in (6) using a second local polynomial regression, of degree $q > p$ (even or odd), based on a kernel L and bandwidth b. Thus, $\mathbf{r}_q(u)$, $\mathbf{R}_q$, $\mathbf{W}_q$, and $\boldsymbol{\Gamma}_q$ are defined as above, but substituting q, L, and b in place of p, K, and h, respectively. Denote by $\mathbf{e}_{p+1}$ the $(q+1)$-vector with one in the $p+2$ position, and zeros in the rest. Then we estimate the bias with

$$\hat{B}_m = h^{p+1}\hat{m}^{(p+1)}\frac{1}{(p+1)!}\mathbf{e}_0'\boldsymbol{\Gamma}_p^{-1}\boldsymbol{\Lambda}_p,$$
$$\hat{m}^{(p+1)} = b^{-p-1}(p+1)!\mathbf{e}_{p+1}'\boldsymbol{\Gamma}_q^{-1}\mathbf{R}_q'\mathbf{W}_q\mathbf{Y}/n.$$

Exactly as in the density case, $\hat{B}_m$ introduces variance that is controlled by ρ and will be captured by robust bias correction.

3.1. Variance Estimation

The Studentizations in the density case were based on fixed-n expectations, and we will show that retaining this is crucial for local polynomials. The fixed-n versus asymptotic distinction is separate from, and more fundamental than, whether we employ feasible versus infeasible quantities. The advantage of fixed-n Studentization also goes beyond bias correction.

To begin, we condition on the covariates so that $\boldsymbol{\Gamma}_p^{-1}$ is fixed. Define $\upsilon(\cdot) = \mathbb{V}[Y|X = \cdot]$ and $\boldsymbol{\Sigma} = \mathrm{diag}(\upsilon(X_i) : i = 1, \ldots, n)$. Straightforward calculation gives

$$\sigma_{\mathtt{us}}^2 = (nh)\mathbb{V}[\hat{m}|X_1, \ldots, X_n] = \frac{h}{n}\mathbf{e}_0'\boldsymbol{\Gamma}_p^{-1}\left(\mathbf{R}_p'\mathbf{W}_p\boldsymbol{\Sigma}\mathbf{W}_p\mathbf{R}_p\right)\boldsymbol{\Gamma}_p^{-1}\mathbf{e}_0. \tag{7}$$

One can then show that $\sigma_{\mathtt{us}}^2 \to_P \upsilon(x)f(x)^{-1}\mathcal{V}(K, p)$, with $\mathcal{V}(K, p)$ a known, constant function of the kernel and polynomial degree. Importantly, both the nonasymptotic form and the convergence hold in the interior or on the boundary, though $\mathcal{V}(K, p)$ changes.

To first order, one could use $\sigma_{\mathtt{us}}^2$ or the leading asymptotic term; all that remains is to make each feasible, requiring estimators of the variance function, and for the asymptotic form, also the density. These may be difficult to estimate when x is a boundary point. Concerned by this, Chen and Qin (2002, p. 93) considered feasible and 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 $\upsilon(x)$." Our results show that this is not true in general: using fixed-n Studentization, feasible or infeasible, leads to confidence intervals with the same coverage error decay rates at interior and boundary points, thereby retaining the celebrated boundary carpentry property.

For robust bias correction, $\sigma_{\mathtt{rbc}}^2 = (nh)V[\hat{m} - \hat{B}_m|X_1, \ldots, X_n]$ captures the variances of $\hat{m}$ and $\hat{m}^{(p+1)}$ as well as their covariance. A fixed-n calculation gives

$$\sigma_{\mathtt{rbc}}^2 = \frac{h}{n}\mathbf{e}_0'\boldsymbol{\Gamma}_p^{-1}\left(\boldsymbol{\Xi}_{p,q}\boldsymbol{\Sigma}\,\boldsymbol{\Xi}_{p,q}'\right)\boldsymbol{\Gamma}_p^{-1}\mathbf{e}_0,$$
$$\boldsymbol{\Xi}_{p,q} = \mathbf{R}_p'\mathbf{W}_p - \rho^{p+2}\boldsymbol{\Lambda}_p\mathbf{e}_{p+1}'\boldsymbol{\Gamma}_q^{-1}\mathbf{R}_q'\mathbf{W}_q. \tag{8}$$

To make the fixed-n scalings feasible, $\hat{\sigma}_{\mathtt{us}}^2$ and $\hat{\sigma}_{\mathtt{rbc}}^2$ take the forms (7) and (8) and replace $\boldsymbol{\Sigma}$ with an appropriate estimator. First, we form $\hat{\upsilon}(X_i) = (Y_i - \mathbf{r}_p(X_i - x)'\hat{\boldsymbol{\beta}}_p)^2$ for $\hat{\sigma}_{\mathtt{us}}^2$ or $\hat{\upsilon}(X_i) = (Y_i - \mathbf{r}_q(X_i - x)'\hat{\boldsymbol{\beta}}_q)^2$ for $\hat{\sigma}_{\mathtt{rbc}}^2$. The latter is bias-reduced because $\mathbf{r}_p(X_i - x)'\boldsymbol{\beta}_p$ is a p-term Taylor expansion of $m(X_i)$ around x, and $\hat{\boldsymbol{\beta}}_p$ estimates $\boldsymbol{\beta}_p$ (similarly with q in place

of p), and we have $q > p$. Next, motivated by the fact that least-square residuals are on average too small, we appeal to the HCk class of estimators (see MacKinnon (2013) for a review), which are defined as follows. First, $\hat{\sigma}^2_{\text{us}}$-HC0 uses $\hat{\mathbf{\Sigma}}_{\text{us}} = \text{diag}(\hat{v}(X_i) : i = 1, \ldots, n)$. Then, $\hat{\sigma}^2_{\text{us}}$-HC$k$, $k = 1, 2, 3$, is obtained by dividing $\hat{v}(X_i)$ by, respectively, $(n - 2\,\text{tr}(\mathbf{Q}_p) + \text{tr}(\mathbf{Q}_p'\mathbf{Q}_p))/n$, $(1 - \mathbf{Q}_{p,ii})$, or $(1 - \mathbf{Q}_{p,ii})^2$, where $\mathbf{Q}_p := \mathbf{R}_p'\mathbf{\Gamma}_p^{-1}\mathbf{R}_p'\mathbf{W}_p/n$ is the projection matrix and $\mathbf{Q}_{p,ii}$ its ith diagonal element. The corresponding estimators $\hat{\sigma}^2_{\text{rbc}}$-HC$k$ are the same, but with q in place of p. For theoretical results, we use HC0 for concreteness and simplicity, though inspection of the proof shows that simple modifications allow for the other HCk estimators and rates do not change. These estimators may perform better for small sample sizes. Another option is to use a nearest-neighbor-based variance estimators with a fixed number of neighbors, following the ideas of Muller and Stadtmuller (1987) and Abadie and Imbens (2008). Note that none of these estimators assume local or global homoscedasticity nor rely on new tuning parameters. Details and simulation results for all these estimators are given in the supplement, see Section S.II.2.3 and Table S.II.9.

3.2. Higher-Order Expansions of Coverage Error

Recycling notation to emphasize the parallel, we study the following three statistics:

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

and their associated confidence intervals I_{us}, I_{bc}, and I_{rbc}, exactly as in Equation (5). Importantly, all present definitions and results are valid for an evaluation point in the interior and at the boundary of the support of X_i. The following standard conditions will suffice, augmented with the appropriate Cramér's condition given in the supplement to conserve space.

Assumption 3 (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, $\mathbb{E}[Y^{8+\delta}|X] < \infty$ for some $\delta > 0$, and in a neighborhood of x, f and v are continuous and bounded away from zero, m is $S > q + 2$ times continuously differentiable with bounded derivatives, and $m^{(S)}$ is Hölder continuous with exponent ς.

Assumption 4 (Kernels). The kernels K and L are positive, bounded, even functions, and have compact support.

We now give our main, generic result for local polynomials, analogous to Theorem 1. For notation, the polynomials q_1, q_2, and q_3 and the biases η_{us} and η_{bc}, are cumbersome and exact forms are deferred to the supplement. All that matters is that the polynomials are known, odd, bounded, and bounded away from zero and that the biases have the usual convergence rates, as detailed below.

Theorem 2. Let Assumptions 3, 4, and Cramér's condition hold and $nh/\log(nh) \to \infty$.

(a) If $\eta_{\text{us}} \log(nh) \to 0$, then

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

(b) If $\eta_{\text{bc}} \log(nh) \to 0$ and $\rho \to 0$, then

$$\mathbb{P}[m \in I_{\text{bc}}] = 1 - \alpha + \left\{ \frac{1}{nh} q_{1,\text{us}} + \eta^2_{\text{bc}} q_{2,\text{us}} + \frac{\eta_{\text{bc}}}{\sqrt{nh}} q_{3,\text{us}} \right\} \phi(z_{\frac{\alpha}{2}}) \{1 + o(1)\} + \rho^{p+2}(\Omega_{1,\text{bc}} + \rho^{p+1}\Omega_{2,\text{bc}})\phi(z_{\frac{\alpha}{2}}) z_{\frac{\alpha}{2}} \{1 + o(1)\}.$$

(c) If $\eta_{\text{bc}} \log(nh) \to 0$ and $\rho \to \bar{\rho} < \infty$, then

$$\mathbb{P}[m \in I_{\text{rbc}}] = 1 - \alpha + \left\{ \frac{1}{nh} q_{1,\text{rbc}} + \eta^2_{\text{bc}} q_{2,\text{rbc}} + \frac{\eta_{\text{bc}}}{\sqrt{nh}} q_{3,\text{rbc}} \right\} \phi(z_{\frac{\alpha}{2}}) \{1 + o(1)\}.$$

This theorem, which covers both interior and boundary points, establishes that the conclusions found in the density case carry over to odd-degree local polynomial regression. (Although we focus on p odd, part (a) is valid in general and (b) and (c) are valid at the boundary for p even.) In particular, this shows that robust bias correction is as good as, or better than, undersmoothing in terms of coverage error. Traditional bias correction is again inferior due to the variance and covariance terms $\rho^{p+2}(\Omega_{1,\text{bc}} + \rho^{p+1}\Omega_{2,\text{bc}})$. Coverage error optimal bandwidths can be derived as well, and similar conclusions are found. Best possible rates are defined for fixed p here, the analog of $\hbar$ above; see Section 2.2 for further discussion on smoothness.

Before discussing bias correction, one aspect of the undersmoothing result is worth mentioning. The fact that Theorem 2 covers both interior and boundary points, without requiring additional assumptions, is in some sense, expected: one of the strengths of local polynomial estimation is its adaptability to boundary points. In particular, from Equation (6) and p odd it follows that $\eta_{\text{us}} \asymp \sqrt{nh}h^{p+1}$ at the interior and the boundary. Therefore, part (a) shows that the decay rate in coverage error does not change at the boundary for the standard confidence interval (but the leading constants will change). This finding contrasts with the result of Chen and Qin (2002) who studied the special case $p = 1$ without bias correction (part (a) of Theorem 2), and is due entirely to our fixed-n Studentization.

Turning to robust bias correction, we will, in contrast, find rate differences between the interior and the boundary, no matter the parity of q. As before, η_{bc} has two terms, representing the higher-order bias of the point estimator and the bias of the bias estimator. The former can be viewed as the bias if $m^{(p+1)}$ were zero, and since $p + 1$ is even, we find that it is of order $\sqrt{nh}h^{p+3}$ in the interior but $\sqrt{nh}h^{p+2}$ at the boundary. The bias of the bias correction depends on both bandwidths h and b, as well as p and q, in exact analogy to the density case. For q odd, it is of order $h^{p+1}b^{q-p}$ at all points, whereas for q even this rate is attained at the boundary, but in the interior the order increases to $h^{p+1}b^{q+1-p}$. Collecting these facts: in the interior, $\eta_{\text{bc}} \asymp \sqrt{nh}h^{p+3}(1 + \rho^{-2}b^{q-p-2})$ for odd q or with b^{q-p-1} for q even; at the boundary, $\eta_{\text{bc}} \asymp \sqrt{nh}h^{p+2}(1 + \rho^{-1}b^{q-p-1})$. Further details are in the supplement.

In light of these rates, the same logic of Section 2.2 leads us to restrict attention to bounded, positive ρ and $q = p + 1$, and thus even. Calonico, Cattaneo, and Titiunik (2014, Remark 7) pointed out that in the special case of $q = p + 1, K = L$, and $\rho = 1$, $\hat{m} - \hat{B}_m$ is identical to a local polynomial estimator of order q; this is the closest analog to M being a higher-order kernel. If the point of interest is in the interior, then $q = p + 2$ yields the same rates.

For notational ease, let $\tilde{\eta}_{\mathtt{bc}}^{\mathtt{int}}$ and $\tilde{\eta}_{\mathtt{bc}}^{\mathtt{bnd}}$ be the leading constants for the interior and boundary, respectively, so that, for example, $\eta_{\mathtt{bc}} = \sqrt{nh}h^{p+3}[\tilde{\eta}_{\mathtt{bc}}^{\mathtt{int}} + o(1)]$ in the interior (exact expressions are in the supplement). We then have the following, precise result; the analog of Corollary 2(a).

Corollary 4. Let the conditions of Theorem 2(c) hold, with $\bar{\rho} \in (0, \infty)$ and $q = p + 1$.

(a) For an interior point,

$$\mathbb{P}[m \in I_{\mathtt{rbc}}] = 1 - \alpha + \left\{ \frac{1}{nh} q_{1,\mathtt{rbc}} + nh^{1+2(p+3)} (\tilde{\eta}_{\mathtt{bc}}^{\mathtt{int}})^2 q_{2,\mathtt{rbc}} + h^{p+3} (\tilde{\eta}_{\mathtt{bc}}^{\mathtt{int}}) q_{3,\mathtt{rbc}} \right\} \phi(z_{\frac{\alpha}{2}}) \{1 + o(1)\}.$$

(b) For a boundary point,

$$\mathbb{P}[m \in I_{\mathtt{rbc}}] = 1 - \alpha + \left\{ \frac{1}{nh} q_{1,\mathtt{rbc}} + nh^{1+2(p+2)} (\tilde{\eta}_{\mathtt{bc}}^{\mathtt{bnd}})^2 q_{2,\mathtt{rbc}} + h^{p+2} (\tilde{\eta}_{\mathtt{bc}}^{\mathtt{bnd}}) q_{3,\mathtt{rbc}} \right\} \phi(z_{\frac{\alpha}{2}}) \{1 + o(1)\}.$$

There are differences in both the rates and constants between parts (a) and (b) of this result, though most of the changes to constants are "hidden" notationally by the definitions of $\tilde{\eta}_{\mathtt{bc}}^{\mathtt{bnd}}$ and the polynomials $q_{k,\mathtt{rbc}}$. Part (a) most closely resembles Corollary 2 due to the symmetry yielding the corresponding rate improvement (recall that $\hbar$ in the density case is replaced with $p + 1$ here), and hence all the corresponding conclusions hold qualitatively for local polynomials.

3.3. Practical Choices and Empirical Consequences

As we did for the density, we now derive bandwidth choices, and data-driven implementations, to optimize coverage error in applications.

Corollary 5. Let the conditions of Corollary 4 hold.

(a) For an interior point, if $h = h^*_{\mathtt{rbc}} = H^*_{\mathtt{rbc}} n^{-1/(p+4)}$, then $\mathbb{P}[m \in I_{\mathtt{rbc}}] = 1 - \alpha + O(n^{-(p+3)/(p+4)})$, where

$$H^*_{\mathtt{rbc}}(\bar{\rho}) = \underset{H>0}{\arg\min} \left| H^{-1} q_{1,\mathtt{rbc}} + H^{1+2(p+3)} (\tilde{\eta}_{\mathtt{bc}}^{\mathtt{int}})^2 q_{2,\mathtt{rbc}} + H^{p+3} (\tilde{\eta}_{\mathtt{bc}}^{\mathtt{int}}) q_{3,\mathtt{rbc}} \right|.$$

(b) For a boundary point, if $h = h^*_{\mathtt{rbc}} = H^*_{\mathtt{rbc}}(\rho) n^{-1/(p+3)}$, then $\mathbb{P}[m \in I_{\mathtt{rbc}}] = 1 - \alpha + O(n^{-(p+2)/(p+3)})$, where

$$H^*_{\mathtt{rbc}}(\bar{\rho}) = \underset{H>0}{\arg\min} \left| H^{-1} q_{1,\mathtt{rbc}} + H^{1+2(p+2)} (\tilde{\eta}_{\mathtt{bc}}^{\mathtt{bnd}})^2 q_{2,\mathtt{rbc}} + H^{p+2} (\tilde{\eta}_{\mathtt{bc}}^{\mathtt{bnd}}) q_{3,\mathtt{rbc}} \right|$$

To implement these results, we first set $\rho = 1$ and the kernels K and L equal to any desired second-order kernel, typical choices being triangular, Epanechnikov, and uniform. The variance estimator $\hat{\sigma}^2_{\mathtt{rbc}}$ is defined in Section 3.1, and is fully implementable, and thus so is $I_{\mathtt{rbc}}$, once the bandwidth h is chosen.

For selecting h at an interior point, the same conclusions from density estimation apply: (i) coverage of $I_{\mathtt{rbc}}$ is quite robust with respect to h and ρ, (ii) feasible choices for h are easy to construct, and (iii) an MSE-optimal bandwidth only delivers the best coverage error for $p = 1$ (i.e., $\hbar = 2$ in the density case). On the other hand, for a boundary point, an interesting consequence of Corollary 5 is that an MSE-optimal bandwidth *never* delivers optimal coverage error decay rates, even for local linear regression: $h^*_{\mathtt{mse}} \propto n^{-1/(2p+3)} \gg h^*_{\mathtt{rbc}} \propto n^{-1/(p+3)}$.

Keeping this in mind, we give a fully data-driven direct plug-in (DPI) bandwidth selector for both interior and boundary points: $\hat{h}^{\mathtt{int}}_{\mathtt{dpi}} = \hat{H}^{\mathtt{int}}_{\mathtt{dpi}}\, n^{-1/(p+4)}$ and $\hat{h}^{\mathtt{bnd}}_{\mathtt{dpi}} = \hat{H}^{\mathtt{bnd}}_{\mathtt{dpi}}\, n^{-1/(p+3)}$, where $\hat{H}^{\mathtt{int}}_{\mathtt{dpi}}$ and $\hat{H}^{\mathtt{bnd}}_{\mathtt{dpi}}$ are estimates of (the appropriate) $H^*_{\mathtt{rbc}}$ of Corollary 5, obtained by estimating unknowns by pilot estimators employing a readily available pilot bandwidth. The complete steps to form $\hat{H}^{\mathtt{int}}_{\mathtt{dpi}}$ and $\hat{H}^{\mathtt{bnd}}_{\mathtt{dpi}}$ are in the supplement, as is a second data-driven bandwidth choice, based on rescaling already-available MSE-optimal bandwidths. All our methods are available in R and STATA via the `nprobust` package, see Calonico, Cattaneo, and Farrell (2017).

4. Simulation Results

We now report a representative sample of results from a simulation study to illustrate our findings. We drew 5000 replicated datasets, each being $n = 500$ iid draws from the model $Y_i = m(X_i) + \varepsilon_i$, with $m(x) = \sin(3\pi x/2)(1 + 18x^2[\mathrm{sgn}(x) + 1])^{-1}$, $X_i \sim \mathcal{U}[-1, 1]$, and $\varepsilon_i \sim \mathcal{N}(0, 1)$. We consider inference at the five points $x \in \{-2/3, -1/3, 0, 1/3, 2/3\}$. The function $m(x)$ and the five evaluation points are plotted in Figure 1; this function was previously used by Berry, Carroll, and Ruppert (2002) and Hall and Horowitz (2013). The supplement gives

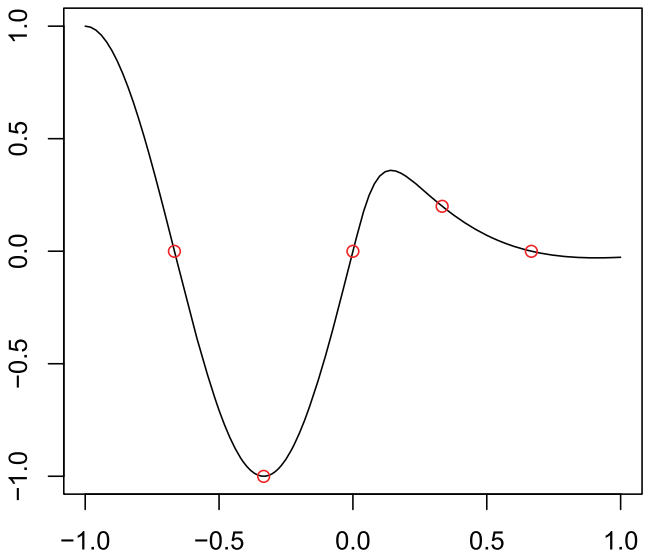


Figure 1. True regression model and evaluation points.

Table 1. Empirical coverage and average interval length of 95% confidence intervals.

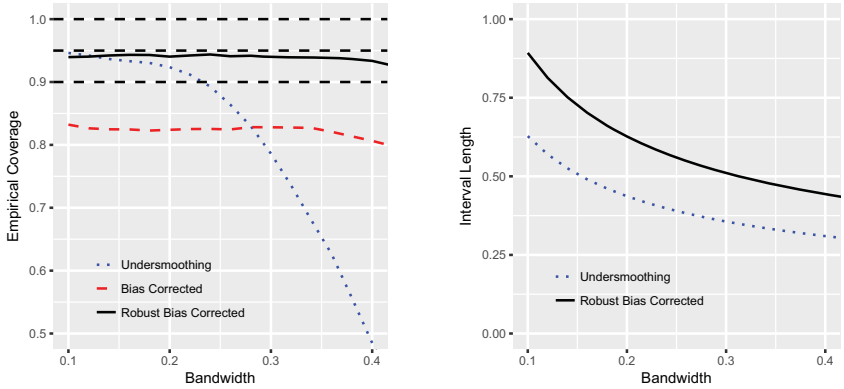
Evaluation point	Average bandwidth	Empirical coverage					Interval length			
		US	Locfit	BC	HH	RBC	US	Locfit	HH	RBC
−2/3	0.203	95.4	95.3	83.4	93.7	95.0	0.437	0.472	0.410	0.627
−1/3	0.307	44.1	66.3	82.1	31.2	94.1	0.357	0.381	0.275	0.507
0	0.320	73.5	83.1	81.0	58.8	93.6	0.348	0.376	0.267	0.498
1/3	0.343	93.3	93.7	82.0	83.1	94.5	0.332	0.361	0.245	0.477
2/3	0.262	94.2	94.5	82.0	89.2	94.3	0.386	0.418	0.329	0.554

NOTES: (i) Column "Average bandwidth" reports simulation average of estimated bandwidths $h = \hat{h}_{\texttt{dpi}} \equiv \hat{h}^{\texttt{int}}_{\texttt{dpi}}$. Simulation distributions for estimated bandwidths are reported in the supplement. (ii) US = Undersmoothing, Locfit = R package `locfit` by Loader (2013), BC = Bias corrected, HH = Hall and Horowitz (2013), RBC = Robust bias corrected.

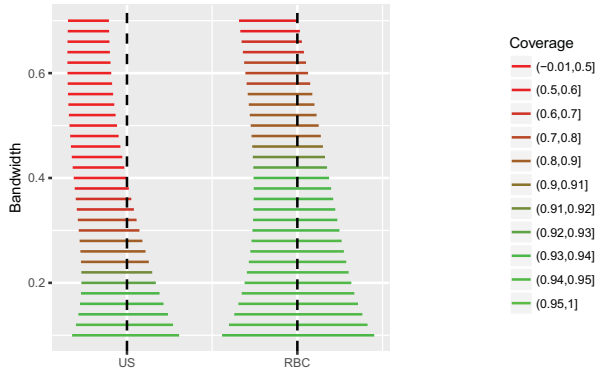
results for other models, bandwidth selectors and their simulation distributions, alternative variance estimators, and more detailed studies of coverage and length.

We compared robust bias correction to undersmoothing, traditional bias correction, the off-the-shelf R package `locfit` (Loader 2013), and the procedure of Hall and Horowitz (2013). In all cases, the point estimator is based on local linear regression with the data-driven bandwidth $\hat{h}^{\texttt{int}}_{\texttt{dpi}}$, which shares the rate of $\hat{h}_{\texttt{mse}}$ in this case, and $\rho = 1$. The `locfit` package has a bandwidth selector, but it was ill-behaved and often gave zero empirical coverage. Hall and Horowitz (2013) did not give an explicit optimal bandwidth, but did advocate a feasible $\hat{h}_{\texttt{mse}}$, following Ruppert, Sheather, and Wand (1995). To implement their method, we used 500 bootstrap replications and we set $1 - \xi = 0.9$ over a sequence $\{x_1, \ldots, x_N\} = \{-0.9, -0.8, \ldots, 0, \ldots, 0.8, 0.9\}$ to obtain the final quantile $\hat{\alpha}_\xi(\alpha_0)$, and used their proposed standard errors $\hat{\sigma}^2_{\text{HH}} = \kappa\hat{\sigma}^2/\hat{f}_X$, where $\hat{\sigma}^2 = \sum_{i=1}^n \hat{\varepsilon}_i^2/n$ for $\hat{\varepsilon}_i = \tilde{\varepsilon}_i - \bar{\varepsilon}$, with $\tilde{\varepsilon}_i = Y_i - \hat{m}(X_i)$ and $\bar{\varepsilon} = \sum_{i=1}^n \tilde{\varepsilon}_i/n$.

Table 1 shows empirical coverage and average length at all five points for all five methods. Robust bias correction yields accurate coverage throughout the support; performance of the other methods varies. For $x = -2/3$, the regression function is nearly linear, leaving almost no bias, and the other methods work quite well. In contrast, at $x = -1/3$ and $x = 0$, all methods except robust bias correction suffer from coverage distortions due to bias. Indeed, Hall and Horowitz (2013, p. 1893) reported that "[t]he 'exceptional' 100ξ% of points that are not covered are typically close to the locations of peaks and troughs, [which] cause difficulties because of bias." Finally, bias is still present, though less of a problem, for $x = 1/3$ and $x = 2/3$, and coverage of the



(a) Empirical Coverage (b) Interval Length



(c) Average Con dence Intervals and Coverage for Undersmoothing (US) and Robust Bias Correction (RBC).

Figure 2. Local polynomial simulation results for $x = 0$.

competing procedures improves somewhat. Motivated by the fact that the data-driven bandwidth selectors may be "too large" for proper undersmoothing, we studied the common practice of ad hoc undersmoothing of the MSE-optimal bandwidth choice $\hat{h}_{\texttt{mse}}$: the results in Table S.II.8 of the supplement show this to be no panacea.

To illustrate our findings further, Figure 2(a) and 2(b) compares coverage and length of different inference methods over a range of bandwidths. Robust bias correction delivers accurate coverage for a wide range of bandwidths, including larger choices, and thus can yield shorter intervals. For undersmoothing, coverage accuracy requires a delicate choice of bandwidth, and for correct coverage, a longer interval. Figure 2(c), in color online, reinforces this point by showing the "average position" of $I_{\texttt{us}}(h)$ and $I_{\texttt{rbc}}(h)$ for a range of bandwidths: each bar is centered at the average bias and is of average length, and then color-coded by coverage (green indicates good coverage, fading to red as coverage deteriorates). These results show that when $I_{\texttt{us}}$ is short, bias is large and coverage is poor. In contrast, $I_{\texttt{rbc}}$ has good coverage at larger bandwidths and thus shorter length.

5. Conclusion

This article 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. Further, such intervals are more robust to bandwidth choice in applications. Second, we showed theoretically when the popular empirical practice of using MSE-optimal bandwidths is justified, and more importantly, when it is not, and we gave concrete implementation recommendations for applications. Third, we proved that confidence intervals based on local polynomials do have automatic boundary carpentry, provided proper Studentization is used. These results are tied together through the themes of higher-order expansions and the importance of finite sample variance calculations and the key, common message that inference procedures must account for additional variability introduced by bias correction.

Supplementary Materials

The supplemental appendix contains technical and notational details omitted from the main text, proofs of all results, further technical details and derivations, and additional simulations results and numerical analyses. The main results are Edgeworth expansions of the distribution functions of the t statistics $T_{\texttt{us}}$, $T_{\texttt{bc}}$, and $T_{\texttt{rbc}}$, for density estimation and local polynomial regression. Stating and proving these results is the central purpose of this supplement. The higher-order expansions of confidence interval coverage probabilities in the main paper follow immediately by evaluating the Edgeworth expansions at the interval endpoints.

Acknowledgments

The authors thank Ivan Canay, Xu Cheng, Joachim Freyberger, Bruce Hansen, Joel Horowitz, Michael Jansson, Francesca Molinari, Ulrich Müller, and Andres Santos for thoughtful comments and suggestions, as well as seminar participants at Cornell, Cowles Foundation, CREST Statistics, London School of Economics, Northwestern, Ohio State University, Princeton, Toulouse School of Economics, University of Bristol, and University College London. The associate editor and three reviewers also provided very insightful comments that improved this manuscript.

Funding

The second author gratefully acknowledges financial support from the National Science Foundation (SES 1357561 and SES 1459931).

References

Abadie, A., and Imbens, G. W. (2008), "Estimation of the Conditional Variance in Paired Experiments," *Annales d'Economie et de Statistique*, 91–92, 175–187. [775]

Armstrong, T. B., and Kolesár, M. (2017), "Optimal Inference in a Class of Regression Models," arxiv preprint arXiv:1511.06028. [768]

Berry, S. M., Carroll, R. J., and Ruppert, D. (2002), "Bayesian Smoothing and Regression Splines for Measurement Error Problems," *Journal of the American Statistical Association*, 97, 160–169. [776]

Bhattacharya, R. N., and Rao, R. R. (1976), *Normal Approximation and Asymptotic Expansions*, New York: Wiley. [768]

Calonico, S., Cattaneo, M. D., and Farrell, M. H. (2016), "Coverage Error Optimal Confidence Intervals for Regression Discontinuity Designs," Working Paper. [768]

——— (2017), "nprobust: Nonparametric Kernel-Based Estimation and Robust Bias-Corrected Inference," Working Paper. [768,776]

Calonico, S., Cattaneo, M. D., and Titiunik, R. (2014), "Robust Nonparametric Confidence Intervals for Regression-Discontinuity Designs," *Econometrica*, 82, 2295–2326. [768,776]

Cattaneo, M. D., Crump, R. K., and Jansson, M. (2013), "Generalized Jackknife Estimators of Weighted Average Derivatives" (with discussion and rejoinder), *Journal of the American Statistical Association*, 108, 1243–1268. [773]

Chen, S. X., and Qin, Y. S. (2002), "Confidence Intervals Based on Local Linear Smoother," *Scandinavian Journal of Statistics*, 29, 89–99. [768,774,775]

Fan, J., and Gijbels, I. (1996), *Local Polynomial Modelling and its Applications*, London: Chapman and Hall. [768,774]

Gasser, T., Muller, H.-G., and Mammitzsch, V. (1985), "Kernels for Nonparametric Curve Estimation," *Journal of the Royal Statistical Society*, Series B, 47, 238–252. [773]

Hall, P. (1991), "Edgeworth Expansions for Nonparametric Density Estimators, with Applications," *Statistics*, 22, 215–232. [770]

——— (1992a), *The Bootstrap and Edgeworth Expansion*, New York: Springer-Verlag. [768]

——— (1992b), "Effect of Bias Estimation on Coverage Accuracy of Bootstrap Confidence Intervals for a Probability Density," *The Annals of Statistics*, 20, 675–694. [768,769,770,771,772]

Hall, P., and Horowitz, J. L. (2013), "A Simple Bootstrap Method for Constructing Nonparametric Confidence Bands for Functions," *The Annals of Statistics*, 41, 1892–1921. [768,777]

Hall, P., and Kang, K.-H. (2001), "Bootstrapping Nonparametric Density Estimators with Empirically Chosen Bandwidths," *The Annals of Statistics*, 29, 1443–1468. [767,773]

Horowitz, J. L. (2009), *Semiparametric and Nonparametric Methods in Econometrics*, New York: Springer. [768]

Loader, C. (2013), locfit: Local Regression, Likelihood and Density Estimation, R Package Version 1.5-9.1, available at *http://cran.rstudio.com/web/packages/locfit/*. [777]

MacKinnon, J. G. (2013), "Thirty Years of Heteroskedasticity-Robust Inference," in *Recent Advances and Future Directions in Causality, Prediction, and Specification Analysis*, eds. X. Chen and N. R. Swanson, New York: Springer, pp. 437–461. [775]

Muller, H.-G., and Stadtmuller, U. (1987), "Estimation of Heteroscedasticity in Regression Analysis," *The Annals of Statistics*, 15, 610–625. [775]

Mykland, P., and Zhang, L. (2017), "Assessment of Uncertainty in High Frequency Data: The Observed Asymptotic Variance," *Econometrica*, 85, 197–231. [768]

Neumann, M. H. (1997), "Pointwise Confidence Intervals in Nonparametric Regression with Heteroscedastic Error Structure," *Statistics*, 29, 1–36. [768]

Ruppert, D., Sheather, S. J., and Wand, M. P. (1995), "An Effective Bandwidth Selector for Local Least Squares Regression," *Journal of the American Statistical Association*, 90, 1257–1270. [777]

Ruppert, D., and Wand, M. P. (1994), "Multivariate Locally Weighted Least Squares Regression," *The Annals of Statistics*, 22, 1346–1370. [774]

Ruppert, D., Wand, M. P., and Carroll, R. (2009), *Semiparametric Regression*, New York: Cambridge University Press. [768]

Schennach, S. M. (2015), "A Bias Bound Approach to Nonparametric Inference," CEMMAP Working Paper CWP71/15, Institute for Fiscal Studies. [768]

Schucany, W., and Sommers, J. P. (1977), "Improvement of Kernel Type Density Estimators," *Journal of the American Statistical Association*, 72, 420–423. [773]

Singh, R. S. (1977), "Improvement on Some Known Nonparametric Uniformly Consistent Estimators of Derivatives of a Density," *The Annals of Statistics*, 5, 394–399. [773]

Stone, C. J. (1982), "Optimal Global Rates of Convergence for Nonparametric Regression," *The Annals of Statistics*, 10, 1040–1053. [772]

Wand, M., and Jones, M. (1995), *Kernel Smoothing*, Boca Raton, FL: Chapman & Hall/CRC. [768]