Bernoulli **28**(4), 2022, 2998–3022
https://doi.org/10.3150/21-BEJ1445

Coverage error optimal confidence intervals for local polynomial regression

SEBASTIAN CALONICO[1,a], MATIAS D. CATTANEO[2,b] and MAX H. FARRELL[3,c]

[1]*Department of Health Policy and Management, Columbia University, New York, New York, U.S.A.*
[a]*sebastian.calonico@columbia.edu*
[2]*Department of Operations Research and Financial Engineering, Princeton University, Princeton, New Jersey, U.S.A.* [b]*cattaneo@princeton.edu*
[3]*Booth School of Business, University of Chicago, Chicago, Illinois, U.S.A.* [c]*max.farrell@chicagobooth.edu*

This paper studies higher-order inference properties of nonparametric local polynomial regression methods under random sampling. We prove Edgeworth expansions for t statistics and coverage error expansions for interval estimators that (i) hold uniformly in the data generating process, (ii) allow for the uniform kernel, and (iii) cover estimation of derivatives of the regression function. The terms of the higher-order expansions, and their associated rates as a function of the sample size and bandwidth sequence, depend on the smoothness of the population regression function, the smoothness exploited by the inference procedure, and on whether the evaluation point is in the interior or on the boundary of the support. We prove that robust bias corrected confidence intervals have the fastest coverage error decay rates in all cases, and we use our results to deliver novel, inference-optimal bandwidth selectors. The main methodological results are implemented in companion R and Stata software packages.

Keywords: Edgeworth expansion; Cramér condition; nonparametric regression; robust bias correction; bandwidth selection; optimal inference; minimax bound

1. Introduction

We study local polynomial inference in the general heteroskedastic nonparametric regression model:

$$Y = \mu(X) + \varepsilon, \qquad \mathbb{E}[\varepsilon|X] = 0, \qquad \mathbb{E}[\varepsilon^2|X] = v(X), \tag{1}$$

where (Y, X) is a pair of random variables with distribution F. The parameter of interest is the level or derivative of the regression function at $X = \mathsf{x}$:

$$\mu^{(\nu)} = \mu^{(\nu)}(\mathsf{x}) := \frac{\mathrm{d}^\nu}{\mathrm{d}x^\nu}\mathbb{E}[Y|X=x]\bigg|_{x=\mathsf{x}}, \qquad \nu \in \mathbb{Z}_+, \tag{2}$$

where the evaluation point x may be in the interior or on the boundary of the support of X. We drop the evaluation point from the notation when possible, and employ the usual convention $\mu = \mu^{(0)}$. Derivatives at boundary points are defined as one-sided derivatives from the interior. Given a random sample $(Y_1, X_1), \ldots, (Y_n, X_n)$ of size n from F, we investigate the quality of statistical inference for $\mu^{(\nu)}$ when using kernel-based local polynomial regression methods [17,18], focusing in particular on higher-order distributional properties of t statistics as well as on coverage error and length of Wald-type confidence interval estimators. We also employ our results to compare and optimize inference procedures for empirical practice and to shed light on the sometimes underappreciated gap between point estimation and inference.

Our main technical contributions are novel Edgeworth expansions for local polynomial based Wald-type t statistics of the form

$$T = \frac{\hat{\theta} - \mu^{(\nu)}}{\hat{\vartheta}}, \tag{3}$$

for different choices of point estimator $\hat{\theta}$ and standard error estimator $\hat{\vartheta}$ detailed in Section 2. We study the accuracy of the Gaussian approximation to the distribution of such T and the error in coverage probability of their dual confidence interval estimators. Our expansions capture the dependence on implementation choices including the polynomial order, the kernel function, and the bandwidth sequence.

Edgeworth expansions are a long-standing tool for more detailed (higher-order) analyses of asymptotic distributional approximations, named after the author of a series of papers on the idea, beginning with [14] and treated more extensively in [15]. See [20] for a textbook review. Informally, an Edgeworth expansion characterizes the leading terms of the difference between the distribution of T and the Gaussian distribution, denoted $\Phi(z)$, for $z \in \mathbb{R}$. That is, an Edgeworth expansion gives the leading terms $E_{T,F}(z)$ and the rate $r_{T,F}$ (which depend on the distribution generating the data, the specific t statistic at issue, along with n, x, and other particulars) such that

$$\lim_{n\to\infty} r_{T,F}^{-1} \sup_{z\in\mathbb{R}} \left| \mathbb{P}_F[T < z] - \Phi(z) - E_{T,F}(z) \right| = 0, \tag{4}$$

where $\mathbb{P}_F$ is the probability law when F is the true data generating process.

We improve on prior work on valid Edgeworth expansions for nonparametric kernel-based regression in three ways: (i) the expansions hold uniformly over a class of data-generating processes (instead of only for one F), (ii) the uniform kernel is allowed (instead of only kernel functions with sufficient variation), and (iii) the expansions hold for any derivative $\nu \geq 0$ (instead of only for the level $\nu = 0$). As discussed below, these improvements offer new theoretical and practical conclusions.

Edgeworth expansions are almost always established pointwise in the underlying distribution, that is, for a single, fixed F, as in (4). Indeed, standard references on the subject [3,20] do not even mention uniformity. However, F is unknown, and a researcher would like some assurances that their inference is equally accurate regardless of the specific underlying data generating process. This motivates expansions that are valid uniformly over a class of plausible distributions for the data, denoted $\mathscr{F}_S$, encoding the researcher's statistical model, accompanying assumptions, and the empirical regularities of the application of interest. Thus, instead of (4), in Section 3, we prove

$$\lim_{n\to\infty} \sup_{F\in\mathscr{F}_S} r_{T,F}^{-1} \sup_{z\in\mathbb{R}} \left| \mathbb{P}_F[T < z] - \Phi(z) - E_{T,F}(z) \right| = 0. \tag{5}$$

We also characterize the worst-case rate $r_T = \inf_{F\in\mathscr{F}_S} r_{T,F}$ of distributional approximation over $\mathscr{F}_S$. The specific class $\mathscr{F}_S$ we consider, defined precisely in Section 2, matches standard empirical settings employing kernel-based nonparametric inference for $\mu^{(\nu)}$, and therefore our theoretical and methodological results speak directly to common practice. Uniformly valid expansions have some precedence in the literature when studying notions of optimality, perhaps originating with [2], but these results are rare and confined to parametric models. Our corresponding uniform results for nonparametric kernel-smoothing do not appear to have a direct antecedent in the literature.

Second, the uniform kernel is ruled out in all prior work on Edgeworth expansions for kernel-based nonparametrics, both for density estimation [19–21] and regression [5,10,11], due to a technical limitation in the proofs that we overcome. Other work on nonparametric regression has assumed away the issue by studying non-random designs [22,25]. In fact, [19, p. 218] conjectured that valid Edgeworth expansions would require techniques for lattice-valued random variables if the uniform kernel was

used. On the contrary, we show that such techniques are not needed. Allowing for the uniform kernel is important for empirical work because it is the optimal kernel shape in terms of minimizing interval length (as discussed in Section 4.2) and because unweighted local least squares regression is a popular choice in some applications.

Finally, inference on derivatives of the regression function, again ignored in prior work, is a common task in empirical work and therefore it is valuable to have valid Edgeworth expansions and implementation guidance specifically for this case, including inference-optimal bandwidth selection. Moreover, considering derivatives yields several interesting theoretical conclusions, highlighting the difference between point estimation and inference: we find not only that the rate of the inference-optimal bandwidth does not depend on the specific derivative order ν being considered, analogous to the well-known result for mean squared error (MSE) optimal bandwidth, but also that the rate for inference itself does not depend on ν, in sharp contrast to the MSE of the point estimator.

The main result, a generic Edgeworth expansion encompassing all three of these contributions, is Theorem 1 in Section 3. We then use this general result to examine the error in coverage probability of confidence interval estimators dual to each t statistic, and the roles of smoothing bias and Studentization in both the distributional approximation and the coverage error of the confidence intervals.

The role of bias is concretized in Section 3.1. Given a level of smoothness of the unknown function μ and polynomial order of the local polynomial procedure $\hat{\theta}$, the nonparametric bias must be removed for valid inference. The method of robust bias correction (RBC) addresses this issue by incorporating explicit bias estimation into the centering $\hat{\theta}$ and then also adjusting the scale $\hat{\vartheta}$ to account for the additional variability introduced by the bias estimation [5,8]. An alternative principled inference method relies on removing the bias by shrinking the bandwidth used when conducting inference, often called undersmoothing. Other ad-hoc inference approaches rely on either upper bounding the bias, inflating the scale of the t statistic, or simply ignoring the bias altogether. Using our higher-order expansions, we show that RBC leads to demonstrable higher-order superior inference for $\mu^{(\nu)}$ relative to the other approaches in the literature.

Our results also show that the choice of Studentization $\hat{\vartheta}$ is crucial for good higher-order properties. This is in contrast to first order approximations, where only consistency of the standard errors is required. An important finding here is that using asymptotic approximations to the variance of $\hat{\theta}$ will increase the leading remainder terms $E_{T,F}(z)$ and hence also coverage error. Using fixed-n Studentization, where $\hat{\vartheta}$ directly estimates the variability of $\hat{\theta}$, completely removes these errors. This result was first proved in [5], but only pointwise in F and excluding the uniform kernel and derivatives of μ. Section 3.2 shows this in full generality. Failure to account for the effect of using asymptotic variance approximations has lead to some confusion in the prior literature: for example, [11] found inflated coverage error at boundary points and [21] found that undersmoothing provides more accurate coverage than bias correction, but both conclusions are due to improper Studentization.

A key practical consequence of the foregoing is that RBC with fixed-n Studentization has leading remainder terms $E_{T,F}(z)$ and rate $r_{T,F}$, of the expansion (5), that vanish at least as fast as, and often strictly faster than, undersmoothing-based approaches, both at interior and boundary evaluation points and for any derivative ν. Intuitively, this holds because RBC exploits all available smoothness to remove bias, but is not punished (in rates) if no additional smoothness is available to remove bias. Section 4 discusses novel implementation of RBC intervals, giving inference-optimal bandwidth and kernel choices that further improve the coverage properties and length of RBC intervals.

More broadly, our results speak to the sometimes neglected gap between point estimation and inference. Implementations focused on optimizing point estimation may not deliver optimal, or even valid, inference. In particular, they need not proceed at the same rate, and perhaps more surprisingly, the inference rate can be *faster*: the rate $r_{T,F}$ at which the distribution of $\hat{\theta}$ collapses to its asymptotic value (namely $\Phi(\cdot)$) can be *faster* than the rate at which $\hat{\theta}$ itself collapses to its asymptotic value ($\mu^{(\nu)}$). Indeed, there are cases where a bandwidth choice yields the fastest possible inference rate $r_{T,F}$ but yields

invalid point estimation. This is the reverse of the better-known fact that using the estimation-optimal bandwidth (minimizing mean squared error) yields invalid inference. Rate optimality is not as well studied for inference as it is for estimation, but Section 5 follows [23] to develop minimax optimal rates in the sense of achieving the fastest (minimal) rate at which the worst-case (maximal) coverage error vanishes and finds that RBC attains this rate.

The paper closes with simulation evidence supporting our theoretical and methodological work reported in Section 6, and a brief conclusion in Section 7. An appendix contains formulas omitted to improve the exposition, while an online supplement [4] gives all proofs, detailed simulation results, and other methodological results. Software implementing our main results is provided in R and Stata [6]. Last but not least, some of the ideas in this paper have been applied to causal inference and treatment effect estimation in the context of regression discontinuity designs in [7].

2. Model assumptions and estimators

We define the class $\mathscr{F}_S$ of distributions for the pair (Y,X) and make precise the local polynomial point estimator $\hat{\theta}$ and scale estimator $\hat{\vartheta}$ of the t statistic (3). The class $\mathscr{F}_S$ is determined through the following assumption. (Recall that derivatives at the boundary of the support of X correspond to one-sided derivatives from the interior of the support.)

Assumption 1. Let $\mathscr{F}_S$ be the set of distributions F for the pair (Y,X) which obey model (1) and for which there exist constants $S \geq \nu$, $s \in (0,1]$, $0 < c < C < \infty$, and a neighborhood of x on the support of X, none of which depend on F, such that for all x, x' in the neighborhood the following hold.

1. The Lebesgue density of (Y,X), $f_{yx}(\cdot)$, the Lebesgue density of X, $f(\cdot)$, and $v(x) := \mathbb{V}[Y|X=x]$, are each continuous and lie inside $[c,C]$, and $\mathbb{E}[|Y|^{8+c}|X=x] \leq C$.
2. $\mu(\cdot)$ is S-times continuously differentiable and $|\mu^{(S)}(x) - \mu^{(S)}(x')| \leq C|x-x'|^s$.

Throughout, $\{(Y_1,X_1),\ldots,(Y_n,X_n)\}$ is a random sample from (Y,X).

These conditions are not materially stronger than usual in kernel-based nonparametric settings. The restrictions on densities and moments are imposed to achieve uniform validity of Edgeworth expansions. The smoothness condition on μ plays a key role: the assumed smoothness, captured by S and s, and its relationship to the smoothness utilized in estimation, will be important for coverage error.

We consider several options for the elements of the t statistic $T = (\hat{\theta} - \mu^{(\nu)})/\hat{\vartheta}$ given in (3). The starting point is the standard local polynomial regression point estimate of $\mu^{(\nu)}$. See [17] for an introduction. We index the classical local polynomial estimate by $p \in \mathbb{Z}_+$, the order of the polynomial used, assumed to be at least ν. Suppressing the dependence on x to simplify notation, we therefore set

$$\hat{\mu}_p^{(\nu)} = \nu!\boldsymbol{e}_\nu'\hat{\boldsymbol{\beta}}_p = \frac{1}{nh^\nu}\nu!\boldsymbol{e}_\nu'\boldsymbol{\Gamma}^{-1}\boldsymbol{\Omega}\boldsymbol{Y}, \qquad \hat{\boldsymbol{\beta}}_p = \arg\min_{\boldsymbol{b}\in\mathbb{R}^{p+1}} \sum_{i=1}^n (Y_i - \boldsymbol{r}_p(hX_{h,i})'\boldsymbol{b})^2 K\left(X_{h,i}\right), \tag{6}$$

where K is a kernel or weighting function, $h = h(n) \longrightarrow 0$ is a bandwidth sequence, $X_{h,i} = (X_i - \mathsf{x})/h$, $\boldsymbol{r}_p(u) = (1,u,u^2,\ldots,u^p)'$,

$$\boldsymbol{\Gamma} = \frac{1}{nh}\sum_{i=1}^n K(X_{h,i})\boldsymbol{r}_p(X_{h,i})\boldsymbol{r}_p(X_{h,i})', \quad \boldsymbol{\Omega} = \frac{1}{h}\left[K(X_{h,1})\boldsymbol{r}_p(X_{h,1}),\ldots,K(X_{h,n})\boldsymbol{r}_p(X_{h,n})\right],$$

$\boldsymbol{e}_\nu$ is the $(p+1)$-vector with a one in the $(\nu+1)^{\text{th}}$ position and zeros in the rest, and $\boldsymbol{Y} = (Y_1,\ldots,Y_n)'$.

The point estimator $\hat{\theta}$ in T is then finalized depending on how the smoothing bias is to be accounted for. The traditional approach takes $\hat{\theta} = \hat{\mu}_p^{(\nu)}$, and then for inference to be valid undersmoothing is required. Explicit bias correction incorporates into $\hat{\theta}$ an estimate of the leading bias term of $\hat{\mu}_p^{(\nu)}$. Both approaches are motivated by the fact that the conditional bias of $\hat{\mu}_p^{(\nu)}$ is of order $h^{p+1-\nu}$ and given by

$$\mathbb{E}\left[\hat{\mu}_p^{(\nu)} \middle| X_1, \ldots, X_n\right] - \mu^{(\nu)} = h^{p+1-\nu}\nu!\boldsymbol{e}_\nu'\boldsymbol{\Gamma}^{-1}\boldsymbol{\Lambda}\frac{\mu^{(p+1)}}{(p+1)!} + o_{\mathbb{P}}(h^{p+1-\nu}), \tag{7}$$

with $\boldsymbol{\Lambda} = \boldsymbol{\Omega}[X_{h,1}^{p+1}, \cdots, X_{h,n}^{p+1}]'/n$, provided $p - \nu$ is odd and $p \leq S - 1$, the standard setting in the literature. Section 3.1 details other cases for p and S. Throughout, asymptotic orders and their in-probability versions always hold uniformly in $\mathscr{F}_S$, as required by our framework: for example, $A_n = o_{\mathbb{P}}(a_n)$ means $\sup_{F \in \mathscr{F}_S} \mathbb{P}_F[|A_n/a_n| > \epsilon] \to 0$ for every $\epsilon > 0$. Limits are taken as $n \to \infty$ unless stated otherwise.

Undersmoothing leaves the center of the interval at $\hat{\theta} = \hat{\mu}_p^{(\nu)}$ unchanged and assumes that the bandwidth h vanishes rapidly enough to render the leading term of (7) negligible relative to the standard error of the point estimator. The term *under*smoothing refers to using less nonparametric smoothing than would be optimal from a mean squared error (MSE) point estimation point of view [17, Section 4]. The MSE-optimal bandwidth choice is the most common by far, and indeed, the default in most software. With $p \leq S - 1$, the MSE-optimal bandwidth for $\hat{\mu}_p^{(\nu)}$ is well-defined whenever $\mu^{(p+1)}(\mathsf{x}) \neq 0$. However, the MSE-optimal bandwidth is too "large" for standard Gaussian inference: the bias remains first-order important when scaled by the standard deviation of the point estimator, and so valid inference requires a bandwidth that vanishes faster.

Explicit bias correction, on the other hand, subtracts an estimate of the leading term of (7), of which only $\mu^{(p+1)}$ is unknown. Thus we have:

$$\hat{\theta}_{\mathtt{rbc}} := \hat{\mu}_p^{(\nu)} - h^{p+1-\nu}\nu!\boldsymbol{e}_\nu'\boldsymbol{\Gamma}^{-1}\boldsymbol{\Lambda}\boldsymbol{e}_{p+1}'\hat{\boldsymbol{\beta}}_{p+1} = \frac{1}{nh^\nu}\nu!\boldsymbol{e}_\nu'\boldsymbol{\Gamma}^{-1}\boldsymbol{\Omega}_{\mathtt{rbc}}\boldsymbol{Y}, \tag{8}$$

where $\boldsymbol{\Omega}_{\mathtt{rbc}} = \boldsymbol{\Omega} - \rho^{p+1}\boldsymbol{\Lambda}\boldsymbol{e}_{p+1}'\bar{\boldsymbol{\Gamma}}^{-1}\bar{\boldsymbol{\Omega}}$ and $\hat{\boldsymbol{\beta}}_{p+1}$, $\bar{\boldsymbol{\Gamma}}$, and $\bar{\boldsymbol{\Omega}}$ are defined akin to $\hat{\boldsymbol{\beta}}_p$, $\boldsymbol{\Gamma}$, and $\boldsymbol{\Omega}$ of (6), but with $p + 1$ in place of p and a bandwidth $b := \rho^{-1}h$ instead of h. The parameter ρ will play a key role in the Edgeworth and coverage error expansions and we will derive optimal choices below.

With the point estimator $\hat{\theta}$ defined, we now define the choice of standard errors $\hat{\vartheta}$. We will focus primarily on "fixed-n" Studentization, also called "preasymptotic" by [18], which means choosing the Studentization to directly estimate $\mathbb{V}[\hat{\theta}|X_1, \ldots, X_n]$, a population quantity but not an asymptotic one. Such choices have superior coverage, as shown below, particularly compared to employing an estimator of an asymptotic representation of $\mathbb{V}[\hat{\theta}|X_1, \ldots, X_n]$. Importantly, when $\hat{\theta} = \hat{\theta}_{\mathtt{rbc}}$, a fixed-$n$ approach makes bias correction robust, because the Studentization accounts for the variability of bias estimation.

These fixed-n variances are easy to compute based on standard least squares logic. Referring to (6), for $\hat{\theta} = \hat{\mu}_p^{(\nu)}$,

$$nh^{1+2\nu}\,\mathbb{V}[\hat{\mu}_p^{(\nu)}|X_1, \ldots, X_n] = \nu!^2\boldsymbol{e}_\nu'\boldsymbol{\Gamma}^{-1}(h\boldsymbol{\Omega}\boldsymbol{\Sigma}\boldsymbol{\Omega}'/n)\boldsymbol{\Gamma}^{-1}\boldsymbol{e}_\nu, \tag{9}$$

where $\boldsymbol{\Sigma}$ is the n-diagonal matrix of conditional variances $v(X_i)$. This formula applies to $\hat{\theta}_{\mathtt{rbc}}$ as well, upon replacing $\boldsymbol{\Omega}$ with $\boldsymbol{\Omega}_{\mathtt{rbc}}$, because the two estimators share the same structure, as shown by comparing the second form in (8) to (6). The fixed-n Studentization is obtained by replacing $\boldsymbol{\Sigma}$ with an

appropriate plug-in estimator, and we then obtain the final $\hat{\vartheta}$ as follows:

$$\begin{aligned} \hat{\vartheta}^2 &= \frac{\hat{\sigma}_p^2}{nh^{1+2\nu}}, & \hat{\sigma}_p^2 &:= \nu!^2 \boldsymbol{e}_\nu' \boldsymbol{\Gamma}^{-1} (h\boldsymbol{\Omega}\hat{\boldsymbol{\Sigma}}_p \boldsymbol{\Omega}'/n) \boldsymbol{\Gamma}^{-1} \boldsymbol{e}_\nu, & \text{and} \\ \hat{\vartheta}^2 &= \hat{\vartheta}^2_{\mathtt{rbc}} := \frac{\hat{\sigma}^2_{\mathtt{rbc}}}{nh^{1+2\nu}}, & \hat{\sigma}^2_{\mathtt{rbc}} &:= \nu!^2 \boldsymbol{e}_\nu' \boldsymbol{\Gamma}^{-1} (h\boldsymbol{\Omega}_{\mathtt{rbc}} \hat{\boldsymbol{\Sigma}}_{\mathtt{rbc}} \boldsymbol{\Omega}_{\mathtt{rbc}}'/n) \boldsymbol{\Gamma}^{-1} \boldsymbol{e}_\nu, \end{aligned} \tag{10}$$

where $\hat{\boldsymbol{\Sigma}}_p$ and $\hat{\boldsymbol{\Sigma}}_{\mathtt{rbc}}$ are the n-diagonal matrices of the squared residuals $\hat{v}(X_i) = (Y_i - \boldsymbol{r}_p(X_i)'\hat{\boldsymbol{\beta}}_p)^2$ and $\hat{v}(X_i) = (Y_i - \boldsymbol{r}_{p+1}(X_i)'\hat{\boldsymbol{\beta}}_{p+1})^2$, respectively. The above variance estimators separate explicitly the "constant" portions, denoted $\hat{\sigma}_p^2$ and $\hat{\sigma}^2_{\mathtt{rbc}}$, which will be used in Section 4.2 for interval length optimization. More precisely, $\hat{\sigma}_p^2$ and $\hat{\sigma}^2_{\mathtt{rbc}}$ will both be bounded and bounded away from zero in probability under our assumptions.

To complete the set of t statistics under consideration, we impose the following standard conditions on the kernel function. This assumption allows for standard choices such as not only the triangular and Epanechnikov kernels, but also the uniform kernel.

Assumption 2. The kernel K is supported on $[-1,1]$, positive, bounded, and even. Further, $K(u)$ is either constant (the uniform kernel) or $(1, K(u)\boldsymbol{r}_{3(k+1)}(u))'$ is linearly independent on $[-1,0]$ and $[0,1]$, where $k = p$ if T is based on $\hat{\mu}_p^{(\nu)}$ and $\hat{\sigma}_p$, and $k = p+1$ if T uses $\hat{\theta}_{\mathtt{rbc}}$ or $\hat{\sigma}_{\mathtt{rbc}}$. The order p is at least ν.

3. Uniformly valid Edgeworth and coverage error expansions

We now give the main technical result of this paper: a uniformly valid, generic Edgeworth expansion as in (5), for the t-statistic T in (3) when using local polynomial regression methods as described in the previous section. To state the result we need some notation. Here we give only what is needed conceptually, leaving cumbersome formulas to the appendix. The terms of the Edgeworth expansion are defined as

$$\begin{aligned} E_{T,F}(z) = \frac{1}{\sqrt{nh}} \omega_{1,T,F}(z) + \Psi_{T,F}\omega_{2,T,F}(z) + \lambda_{T,F}\omega_{3,T,F}(z) \\ + \frac{1}{nh}\omega_{4,T,F}(z) + \Psi_{T,F}^2 \omega_{5,T,F}(z) + \frac{1}{\sqrt{nh}} \Psi_{T,F}\omega_{6,T,F}(z), \end{aligned} \tag{11}$$

where z is the point of evaluation of the distribution, $\Psi_{T,F}$ denotes the generic non-random (fixed-n) bias of the $\sqrt{nh^{1+2\nu}}$-scaled numerator of T, $\lambda_{T,F}$ denotes the mismatch between the variance of the numerator of the t-statistic and the population standardization used, and the six terms $\omega_{k,T,F}(z)$, $k = 1,2,\ldots,6$, are non-random functions bounded uniformly in $\mathscr{F}_S$, and bounded away from zero for at least one $F \in \mathscr{F}_S$. Section 3.1 provides further details on $\Psi_{T,F}$ and Section 3.2 discusses $\lambda_{T,F}$. The quantities $\omega_{k,T,F}(z)$, $k = 1,2,\ldots,6$ are relatively less important, beyond their parity, because they cannot be altered by implementation choices.

We then have the following result (Theorem 1), establishing (5). This result is general, covering interior and boundary points, $p - \nu$ even and odd, any derivative $\nu \geq 0$, and any combination of p and S. Different cases for each of these primarily affect the expansion, and the final rates, through the bias $\Psi_{T,F}$, as explored in the next section. The conditions imposed are strengthened relative to typical pointwise first-order analyses only by $\log(nh)$ factors on the bandwidth(s) and the other uniformity

requirements of Assumption 1. (Recall that asymptotic orders and their in-probability versions are always required to hold uniformly in $\mathscr{F}_S$ throughout.)

Theorem 1. *Let Assumptions 1 and 2 hold, and assume that*

$$\log(nh)^{2+\gamma}/nh = o(1), \quad \Psi_{T,F}\log(nh)^{1+\gamma} = o(1), \quad \lambda_{T,F} = o(1), \quad \rho = O(1),$$

for any γ bounded away from zero uniformly in $\mathscr{F}_S$. Then,

$$\lim_{n\to\infty} \sup_{F\in\mathscr{F}_S} r_{T,F}^{-1} \sup_{z\in\mathbb{R}} \left|\mathbb{P}_F[T < z] - \Phi(z) - E_{T,F}(z)\right| = 0$$

holds with $E_{T,F}(z)$ of (11) and $r_{T,F} = \max\{(nh)^{-1}, \Psi_{T,F}^2, (nh)^{-1/2}\Psi_{T,F}, \lambda_{T,F}\}$.

A crucial piece in the proof of Theorem 1 is establishing that the appropriate Cramér's condition holds under Assumption 2, and in particular the linear independence condition. Such linear independence fails when K is uniform and u runs over the support of $K(u)$, and this failure has prevented the uniform kernel from being covered by past work. Our key insight is that previous approaches ignored the region *outside* the support of $K(\cdot)$ but *inside* the neighborhood of Assumption 1. Loosely speaking, $(1, K(u), uK(u), \ldots)'$ may be linearly *dependent* on $u \in [-1, 1]$ (when K is uniform), but $(1, K(\frac{x-\mathsf{x}}{h}), (\frac{x-\mathsf{x}}{h})K(\frac{x-\mathsf{x}}{h}), \ldots)'$ is linearly *independent* on x in a fixed neighborhood of x. This allows us to verify Cramér's condition. See the supplement for details [4].

In practice, the error in coverage probability of two-sided interval estimators may be more directly relevant than the distributional approximation of the Edgeworth expansion. We therefore turn to interval estimators dual to each t statistic, given by

$$I = \left[\hat{\theta} - z_u\,\hat{\vartheta}\,,\ \hat{\theta} - z_l\,\hat{\vartheta}\right], \tag{12}$$

where z_l and z_u denote chosen quantiles. Our starting point is a generic coverage error expansion for confidence intervals I, dual to a given T, which follows immediately from Theorem 1 by evaluating the Edgeworth expansion at the interval quantiles (see [4]).

Corollary 1. *Let the conditions of Theorem 1 hold, assume that $\Phi(z_u) - \Phi(z_l) = 1 - \alpha$, and define $C_{I,F}(z_l, z_u) = E_{T,F}(z_u) - E_{T,F}(z_l) = O(r_I)$ for some sequence r_I. Then,*

$$\lim_{n\to\infty} r_I^{-1} \sup_{F\in\mathscr{F}_S} \left|\mathbb{P}_F\left[\mu^{(\nu)}(\mathsf{x}) \in I\right] - (1-\alpha) - C_{I,F}(z_l, z_u)\right| = 0.$$

This result is as general as Theorem 1. The uniform-in-$\mathscr{F}_S$ rate r_I is the slowest vanishing of the rates of each term in the Edgeworth expansion (11), which without specifying any elements further, can only be known to vanish at least as fast as $r_T = \sup_{F\in\mathscr{F}_S} r_{T,F}$ from Theorem 1. However, even at this level of generality, several conclusions are already evident due to the parity of the functions ω_k making up $E_{T,F}(z)$ and hence $C_{I,F}(z_l, z_u)$. First, regarding the choice of quantiles, we recover the classical finding that symmetric intervals, where $z_l = -z_u$, have superior coverage properties, because ω_1 and ω_2 are even functions of z. Asymmetric choices that still have $\Phi(z_u) - \Phi(z_l) = 1 - \alpha$ can yield correct coverage, but the error will vanish more slowly, whereas other choices will not yield uniformly correct coverage. Bootstrap-based quantiles will, in general, not improve coverage error rates in nonparametric contexts beyond the symmetric case [21], and can in fact be detrimental for coverage error [24]. Second, the remaining w_k functions are odd, and therefore to obtain better coverage properties we should focus

on intervals with small (rapidly vanishing) $\Psi_{T,F}$ and $\lambda_{T,F}$. The upcoming subsections discuss each of these pieces in turn.

Our expansions highlight the conceptual gap between point estimation and inference. The rate at which the distribution of $\hat{\theta}$ collapses to its asymptotic value ($\Phi(\cdot)$) can be *faster* than the rate at which the point estimator $\hat{\theta}$ itself collapses to its asymptotic value ($\mu^{(\nu)}$). Moreover, it is possible that coverage error may vanish even if mean squared error does not, and vice versa. One direction of this phenomenon captures the well-known result that the coverage error of a confidence interval centered at the MSE-optimal point estimator does not vanish. That is, $\hat{\mu}_p^{(\nu)}$ in (6) using the MSE-optimal bandwidth $h_{\texttt{mse}} = H_{\texttt{mse}} n^{-1/(2p+3)}$, for some constant $H_{\texttt{mse}}$, is optimal for point estimation given a fixed p, but

$$\sup_{F \in \mathscr{F}_S} \left| \mathbb{P}_F \left[\mu^{(\nu)} \in \left\{ \hat{\mu}_p^{(\nu)} \pm z_{\alpha/2} \hat{\sigma}_p H_{\texttt{mse}}^{-1/2} n^{-1/2+(1+2\nu)/(4p+6)} \right\} \right] - (1-\alpha) \right| \asymp 1,$$

where $a \asymp b$ denotes that $a \leq C_1 b$ and $b \leq C_1 a$ for some constants C_1 and C_2.

The other direction may be more surprising and novel: we find that the variance of $\hat{\theta}$ can be too large for mean-square consistency, but nonetheless be captured well enough by $\hat{\vartheta}$ for valid inference. For example, consider inference on $\mu^{(1)}(\mathsf{x})$ using I_p with local linear regression ($p = 1$). Choosing $h \asymp n^{-1/3}$ yields $r_{I_p} \asymp n^{-2/3}$, which is the fastest attainable rate for I_p in this case, but also gives $\mathbb{V}[\hat{\mu}_p^{(\nu)} | X_1, \ldots, X_n] \asymp_{\mathbb{P}} (nh^{1+2\nu})^{-1} \asymp 1$, and thus $\hat{\mu}_p^{(1)}$ is not consistent in mean square. Therefore, we found a confidence interval that is *optimal for coverage* of $\mu^{(1)}$, but implicitly relies on a point estimator that is *not even consistent* in mean square.

3.1. Bias details

We now give details for the bias term, $\Psi_{T,F}$, highlighting three main points. First, the rate at which $\Psi_{T,F}$ vanishes does not depend on the derivative ν. Second, we establish that performing bias correction never slows the rate at which $\Psi_{T,F}$ vanishes. The third goal is then practical: we spell out several cases of the rates and constants for the bias of $\hat{\theta}_{\texttt{rbc}}$ so that we may use these for bandwidth and kernel selection later.

To describe $\Psi_{T_p,F}$, the bias term for T_p, let $\boldsymbol{\beta}_p$ be the $p+1$ vector with $(j+1)$ element equal to $\mu^{(j)}(\mathsf{x})/j!$ for $j = 0, 1, \ldots, p$ as long as $j \leq S$, and zero otherwise, and $\boldsymbol{B}_p$ as the n-vector with i^{th} entry $[\mu(X_i) - \boldsymbol{r}_p(X_i - \mathsf{x})'\boldsymbol{\beta}_p]$. Then,

$$\Psi_{T_p,F} = \sqrt{nh}\, \nu! \boldsymbol{e}_\nu' \mathbb{E}[\boldsymbol{\Gamma}]^{-1} \mathbb{E}[\boldsymbol{\Omega} \boldsymbol{B}_p]. \tag{13}$$

Turning to bias correction, define $\boldsymbol{\beta}_{p+1}$ and $\boldsymbol{B}_{p+1}$ as above, but with $p+1$ in place of p in all cases. Then, using the definition of $\boldsymbol{\Omega}_{\texttt{rbc}}$ in (8),

$$\Psi_{T_{\texttt{rbc}},F} = \sqrt{nh}\, \nu! \boldsymbol{e}_\nu' \mathbb{E}[\boldsymbol{\Gamma}]^{-1} \left(\mathbb{E}[\boldsymbol{\Omega} \boldsymbol{B}_{p+1}] - \rho^{p+1} \mathbb{E}[\boldsymbol{\Lambda}] \boldsymbol{e}_{p+1}' \mathbb{E}[\bar{\boldsymbol{\Gamma}}]^{-1} \mathbb{E}[\bar{\boldsymbol{\Omega}} \boldsymbol{B}_{p+1}] \right). \tag{14}$$

These bias terms are non-random but otherwise non-asymptotic: all expectations are fixed-n and we have not done the typical Taylor expansion. The derivative ν only appears in the constant term $\nu! \boldsymbol{e}_\nu$, and therefore the rate at which $\Psi_{T_p,F}$ vanishes does not depend on the derivative being estimated. Intuitively, this can be seen from the second form for $\hat{\mu}_p^{(\nu)}$ in (6), $n^{-1} h^{-\nu} \nu! \boldsymbol{e}_\nu' \boldsymbol{\Gamma}^{-1} \boldsymbol{\Omega} \boldsymbol{Y}$, coupled with rate $\sqrt{nh^{1+2\nu}}$ of the Studentizations of (10): together, these account for the derivative, and distributional properties of $\boldsymbol{\Gamma}^{-1} \boldsymbol{\Omega} \boldsymbol{Y}$ are left independent of ν; the first conclusion of this subsection.

The rate of convergence (to zero) of $\Psi_{T_p,F}$ or $\Psi_{T_{\mathrm{rbc}},F}$ can be deduced by first expanding $\mu(X_i)$ entering $\boldsymbol{B}_p$ and $\boldsymbol{B}_{p+1}$ around x, and then specializing to a given p and S. For any p, we have

$$\mu(X_i) - \boldsymbol{r}_p(X_i-\mathsf{x})'\boldsymbol{\beta}_p = \sum_{k=S\wedge p+1}^{S} \frac{1}{k!}(X_i-\mathsf{x})^k \mu^{(k)}(\mathsf{x}) + \frac{1}{S!}(X_i-\mathsf{x})^S \left(\mu^{(S)}(\bar{x}) - \mu^{(S)}(\mathsf{x})\right),$$

where the summation is taken to be zero if $p \geq S$. To obtain the final rate, this expansion is substituted into $\boldsymbol{B}_p$ and the leading terms are identified by stabilizing the expectation of the terms involving $(X_i - \mathsf{x})^k$ by writing $h^k(X_{h,i})^k$, thus isolating the rate. The rate will depend on the smoothness, location of x, parity of $p-\nu$, and the bandwidth h. For $\boldsymbol{B}_{p+1}$, replace p with $p+1$ everywhere and use b in place of h in the second term. For details see [4].

Our second point is that $\Psi_{T_{\mathrm{rbc}},F} = O(\Psi_{T_p,F})$, which follows from the expansion above and taking ρ bounded and bounded away from zero. First, observe from the Taylor expansion applied to (14) that $\Psi_{T_{\mathrm{rbc}},F}$ depends on higher order derivatives than $\Psi_{T_p,F}$, which follows from applying the Taylor expansion to (13), and therefore stabilizing leads to higher powers of h. Intuitively, the bias of $\hat{\mu}_p^{(\nu)}$ is the product of the rate h^{p+1} and the constant targeted by bias correction. Therefore, the bias of $\hat{\theta}_{\mathrm{rbc}}$ is at most h^{p+1} times the bias of the bias correction plus the higher order term of (7). For a fixed sequence h, neither of these can be greater than h^{p+1}. Second, the rate for $\Psi_{T_{\mathrm{rbc}},F}$ cannot be improved by letting $\rho = h/b$ vanish or diverge: ρ vanishing decreases the second term, but the first term is unchanged, while letting ρ diverge can only inflate the second term. Further, diverging ρ renders the effective sample size nb, which is smaller than nh, which would only inflate the Edgeworth expansion terms without reducing bias (hence the restriction in Theorem 1 to bounded ρ).

Therefore, in optimizing inference later on, we will focus on $\hat{\theta}_{\mathrm{rbc}}$ and take ρ bounded and bounded away from zero. We need the leading bias constants for this case, which follow from carrying on the Taylor expansion completely in (14). The bias is always of the form

$$\Psi_{T_{\mathrm{rbc}},F} = O(\sqrt{nh}h^{\zeta})$$

for an exponent ζ that depends on the location of x, the parity of $p-\nu$, and the smoothness S. A complete list of ζ is shown in Table 1. From there, we see that if p is large enough relative to S (how large depends on the specific case), then $\zeta = S+s$, implying $\Psi_{T_{\mathrm{rbc}},F} = O(\sqrt{nh}h^{S+s})$.

The more empirically relevant case is to treat p as fixed and smaller than S, specifically $p \leq S-3$ for interior x with $p-\nu$ odd and $p \leq S-2$ otherwise (i.e. for boundary points or if x is an interior point with $p-\nu$ even). In these cases, we can use the Taylor expansion above to characterize the leading term, and write

$$\Psi_{T_{\mathrm{rbc}},F} = \sqrt{nh}h^{\zeta}\psi_{T_{\mathrm{rbc}},F}[1+o(1)],$$

where $\zeta = p+3$ for interior x with $p-\nu$ odd and $p+2$ otherwise. The term $\psi_{T_{\mathrm{rbc}},F}$ will be referred to as the constant term for simplicity, though technically it is a non-random sequence with known form, uniformly bounded in $\mathscr{F}_S$, and nonzero for some $F \in \mathscr{F}_S$. Referring to Table 1 for the different cases,

Location of x	Parity of $p-\nu$	Smoothness	ζ	$\psi_{T_{\mathrm{rbc}},F}$
Boundary	Odd or Even	$p+2 \leq S$	$p+2$	Equation (15a)
		$p+2 > S$	$S+s$	N/A
Interior	Even	$p+2 \leq S$	$p+2$	Equation (15b)
		$p+2 > S$	$S+s$	N/A
	Odd	$p+3 \leq S$	$p+3$	Equation (15c)
		$p+2 \geq S$	$S+s$	N/A

Table 1. Bias Terms For Bias-Corrected Centering $\hat{\theta}_{\mathrm{rbc}}$. With ρ bounded and bounded away from zero, $\Psi_{T_{\mathrm{rbc}},F} = O(\sqrt{nh}h^{\zeta})$ and further, if p is small relative to S, $\Psi_{T_{\mathrm{rbc}},F} = \sqrt{nh}h^{\zeta}\psi_{T_{\mathrm{rbc}},F}[1+o(1)]$.

$\psi_{T_{\mathrm{rbc}},F}$ can be

$$\begin{cases} \dfrac{\mu^{(p+2)}}{(p+2)!}\nu!\boldsymbol{e}_{\nu}'\mathbb{E}[\boldsymbol{\Gamma}]^{-1}\left\{\mathbb{E}[\boldsymbol{\Lambda}_2]-\rho^{-1}\mathbb{E}[\boldsymbol{\Lambda}_1]\boldsymbol{e}_{p+1}'\mathbb{E}[\bar{\boldsymbol{\Gamma}}]^{-1}\mathbb{E}[\bar{\boldsymbol{\Lambda}}_1]\right\}, & \text{(15a)} \\ \dfrac{\mu^{(p+2)}}{(p+2)!}\nu!\boldsymbol{e}_{\nu}'\mathbb{E}[\boldsymbol{\Gamma}]^{-1}\mathbb{E}[\boldsymbol{\Lambda}_2], \qquad \text{or} & \text{(15b)} \\ \nu!\boldsymbol{e}_{\nu}'\mathbb{E}[\boldsymbol{\Gamma}]^{-1}\left\{\dfrac{\mu^{(p+2)}}{(p+2)!}\left[h^{-1}\mathbb{E}[\boldsymbol{\Lambda}_2]-\rho^{-2}b^{-1}\mathbb{E}[\boldsymbol{\Lambda}_1]\boldsymbol{e}_{p+1}'\mathbb{E}[\bar{\boldsymbol{\Gamma}}]^{-1}\mathbb{E}[\bar{\boldsymbol{\Lambda}}_1]\right]\right. & \\ \qquad\left.+\dfrac{\mu^{(p+3)}}{(p+3)!}\left[\mathbb{E}[\boldsymbol{\Lambda}_3]-\rho^{-2}\mathbb{E}[\boldsymbol{\Lambda}_1]\boldsymbol{e}_{p+1}'\mathbb{E}[\bar{\boldsymbol{\Gamma}}]^{-1}\mathbb{E}[\bar{\boldsymbol{\Lambda}}_2]\right]\right\}, & \text{(15c)} \end{cases}$$

where $\boldsymbol{\Lambda}_k = \boldsymbol{\Omega}[X_{h,1}^{p+k},\cdots,X_{h,n}^{p+k}]'/n$ and $\bar{\boldsymbol{\Lambda}}_k = \bar{\boldsymbol{\Omega}}[X_{b,1}^{p+1+k},\ldots,X_{b,n}^{p+1+k}]'/n$, and hence in particular $\boldsymbol{\Lambda}_1 \equiv \boldsymbol{\Lambda}$ as defined in Section 2.

3.2. Variance details

In contrast to first order distributional analysis, where only consistency is required, the choice of scaling, or Studentization, is crucial for higher order properties. Our detailed expansions show that, in general, there are two types of higher-order terms that arise due to Studentization. One is the unavoidable estimation error incurred when replacing any population quantity with a feasible counterpart. The second error arises from the difference between the population variability of the centering $\hat{\theta}$ and the population standardization chosen as the target. This second type of error is what is captured by $\lambda_{T,F}$, and the most important conclusion is that the fixed-n standard errors in (10) achieve $\lambda_{T,F} \equiv 0$, and are therefore demonstrably superior choices for inference. That is, there should not be a "mismatch" between the population variability of the t statistic numerator and the population standardization.

Using an asymptotic approximation to $\mathbb{V}[\hat{\theta}|X_1,\ldots,X_n]$ may yield nonzero $\lambda_{T,F}$, and thus the distributional approximation (and coverage) will suffer. There are too many options to treat comprehensively, but several points warrant discussion. In general, if the chosen standard errors are consistent, $\lambda_{T,F}$ has the form $\lambda_{T,F} = l_n L$, for a rate $l_n \to 0$ and a sequence L that is bounded and bounded away from zero, a "constant", capturing the difference between the variance of the numerator of the t-statistic and the population standardization chosen.

At boundary points the use of asymptotic approximations can be particularly deleterious for coverage, and this has lead to some confusion in the literature. A headline finding of [10] is that an

empirical likelihood confidence interval estimator has coverage error of the same order at interior and boundary points, which is claimed (in the abstract) to be a "significant improvement over confidence intervals based directly on the asymptotic normal distribution". This claim is based on work by the same authors [11] who study, in our notation, the interval with centering $\hat{\theta} = \hat{\mu}_1^{(0)}$ and scaling $\hat{\vartheta} = (nh)^{-1/2}\hat{v}(\mathsf{x})\hat{f}(\mathsf{x})^{-1}\mathcal{V}$, for $\hat{v}(\mathsf{x})$, $\hat{f}(\mathsf{x})$, and $\mathcal{V}$ given therein, where $v(\mathsf{x})f(\mathsf{x})^{-1}\mathcal{V}$ is the probability limit of $\mathbb{V}[(nh)^{1/2}\hat{\mu}_1^{(0)} \mid X_1,\dots,X_n]$. They find that $\lambda_{T,F} = l_n L$ holds with $l_n = h$ at boundary points, meaning greatly increased coverage error. Concerned that this result is due to estimation error, they confirm that $l_n = h$ holds with the infeasible standardization $\vartheta = (nh)^{-1/2}v(\mathsf{x})f(\mathsf{x})^{-1}\mathcal{V}$. However, this neglects the fact that $\lambda_{T,F}$ captures the "mismatch" error, not estimation error, and their conclusion is entirely due to using an asymptotic standardization as opposed to a fixed-n one, and thus empirical likelihood, in particular, does *not* offer higher-order improvements over normality-based intervals.

Explicit bias correction was claimed by [21] to be inferior to undersmoothing for inference; a finding also based entirely on using an asymptotic standardization. In this case, nonrobust bias correction was studied, which pairs $\hat{\theta}_{\mathtt{rbc}}$ with $\hat{\sigma}_p$. This is valid to first order if $\rho = o(1)$, because then $\mathbb{V}[\hat{\theta}_{\mathtt{rbc}}|X_1,\dots,X_n] = \mathbb{V}[\hat{\mu}_p^{(\nu)} \mid X_1,\dots,X_n] = o_{\mathbb{P}}(n^{-1}h^{-1-2\nu})$. However, higher order expansions find $\lambda_{T,F} = \rho^{p+2}(L_1 + \rho^{p+2}L_2)$, where L_1 captures the (scaled) covariance between $\hat{\mu}^{(\nu)}$ and $\hat{\mu}^{(p+1)}$ and L_2 the variance of $\hat{\mu}^{(p+1)}$. These terms lead [21] to conclude that bias correction is inferior to undersmoothing, which [5] later showed is not true for robust bias correction. Our results extend this conclusion to hold for derivatives, boundary points, all smoothness cases, and uniformly in $\mathscr{F}_S$, while also allowing for the uniform kernel.

4. Optimizing interval estimation in practice

We turn to optimizing inference in practice, using the conclusions from the previous sections. Collectively, the previous sections imply that the best coverage will be from using symmetric RBC intervals, i.e. those with $z_l = -z_u = z_{\alpha/2} = \Phi^{-1}(\alpha/2)$, $\hat{\theta}_{\mathtt{rbc}}$ as in (8), $\hat{\vartheta}_{\mathtt{rbc}}$ as in (10), and ρ bounded and bounded away from zero (implying $h = \rho b$). With an eye toward empirical work, we assume in this section that p is fixed and small compared to S. The other cases detailed in Section 3.1 are of relatively little practical value. In practice researchers first choose p and then conduct inference based on that choice (witness the ubiquity of local linear regression and cubic splines).

Letting

$$I_{\mathtt{rbc}}(h) = \left[\hat{\theta}_{\mathtt{rbc}} + z_{\alpha/2}\,\hat{\vartheta}_{\mathtt{rbc}}\,,\ \hat{\theta}_{\mathtt{rbc}} - z_{\alpha/2}\,\hat{\vartheta}_{\mathtt{rbc}}\right]$$

denote the recommended RBC confidence interval, now with its dependence on the bandwidth h explicit to enhance the exposition, we readily deduce from Corollary 1 that

$$C_{I_{\mathtt{rbc}}(h),F}(z_{\alpha/2},-z_{\alpha/2}) = \frac{1}{nh}2\omega_{4,\mathtt{rbc},F} + 2nh^{1+2\zeta}\psi_{T_{\mathtt{rbc}},F}\omega_{5,\mathtt{rbc},F} + h^{\zeta}2\psi_{T_{\mathtt{rbc}},F}\omega_{6,\mathtt{rbc},F}, \tag{16}$$

where the coverage error rate is $r_{\mathtt{rbc}} = \max\{(nh)^{-1}, nh^{1+2\zeta}, h^{\zeta}\}$, with $\zeta = p+3$ if $p-\nu$ is odd and x is a boundary point, or $\zeta = p+2$ otherwise. Furthermore, its interval length is

$$|I_{\mathtt{rbc}}(h)| = 2z_{\alpha/2}\hat{\vartheta}_{\mathtt{rbc}} = 2z_{\alpha/2}\frac{\hat{\sigma}_{\mathtt{rbc}}}{\sqrt{nh^{1+2\nu}}}. \tag{17}$$

Notice that the rate of contraction of length does depend on ν, while the coverage error rate does not.

In the next two subsection we use the above two displays, (16) and (17), to choose the bandwidth parameters h and $\rho = h/b$, and the kernel shape. Before any choices can be made, the researcher must decide on the usual size versus power trade off. In our context, this translates to the relative value they place on coverage error, the discrepancy from nominal level, versus interval length. Because we give the first characterizations of coverage error in many cases, and the first uniformly valid ones, this issue can now be studied in detail: our theoretical ideas can inform this trade off, providing new insights to consider, as well as guiding implementation given a preference for coverage error and length.

At one extreme is the approach that requires only that the interval is not anti-conservative, and then minimizes (expected) length. In this case, a shorter interval that uniformly over-covers is preferred to an interval that is longer but has correct coverage asymptotically. Our results lead one to consider the other extreme: minimize the coverage error directly, and only after optimize length. That is, seek for the confidence interval I such that, in the notation of Corollary 1, r_I vanishes as fast as possible. In applications, an interval with a faster decaying coverage error may approximate its nominal level more closely in finite samples. Such approach focuses on the accuracy of the Gaussian approximation for coverage error, and thus for inference. However, both of these extremes may be unappealing in practice because neither may be optimal from a coverage-length (or, perhaps, size-power for the dual hypothesis test) perspective. Therefore, we will also consider compromises, trading off between coverage error and interval length. One option is to minimize length among consistent interval estimators: seek the shortest interval such that $r_I = o(1)$. In the context of kernel-based nonparametrics, interval estimators with good control of worst-case coverage are able to use larger bandwidths in general, and are thus shorter in large samples; an analogue to the adage that "similar tests have higher power". In general, we will let the user determine a trade off between the two and thus find a bandwidth choice to implement their preference.

4.1. Optimizing interval estimation: Bandwidth selection

We now focus on choosing the bandwidth h optimally, leaving ρ and K to the next section. With pragmatism in mind, we restrict attention to bandwidth sequences that are polynomial in n, that is, of the form $h = Hn^{-\eta}$ for some constants $H > 0$ and $\eta > 0$. For implementation purposes, we optimize $C_{I_{\mathtt{rbc}}(h),F}(z_{\alpha/2}, -z_{\alpha/2})$ pointwise in F. The optimal bandwidths will be functions of F and their implementations are functions of the data, which are draws from F; neither depend explicitly upon $\mathscr{F}_S$. The resulting coverage error rates still hold uniformly, because the bandwidths are of the form $h = Hn^{-\eta}$, where η does not depend on F and H is well-behaved uniformly in $\mathscr{F}_S$. We will focus on cases where coverage is consistent, leveraging our new higher-order results in this paper.

An obvious candidate for h in applications is the classical MSE-optimal choice, denoted $h_{\mathtt{mse}}$, for the point estimator $\hat{\mu}_p^{(\nu)}(x)$ used as part of the centering of the confidence interval $I_{\mathtt{rbc}}(h)$. This bandwidth choice is popular and readily available in most statistical software. Although designed to optimize point estimation, our theoretical results show that it yields valid robust bias corrected inference, that is, $\sup_{F \in \mathscr{F}_S} |\mathbb{P}_F[\mu^{(\nu)}(\mathsf{x}) \in I_{\mathtt{rbc}}(h_{\mathtt{mse}})] - (1-\alpha)| \to 0$, in contrast to the traditional interval $I_p(h_{\mathtt{mse}})$, which undercovers. This gives a principled endorsement for using $h_{\mathtt{mse}}$ coupled with robust bias correction in applications, if a researcher wishes to optimize point estimation instead of inference when choosing the bandwidth h. To be more precise, our results give formal justification (and demonstrate higher-order coverage improvements) for reporting $\hat{\mu}_p^{(\nu)}(\mathsf{x})$ along with $I_{\mathtt{rbc}}(h_{\mathtt{mse}})$, both implemented using the same bandwidth $h_{\mathtt{mse}}$, that is, pairing an MSE-optimal point estimator with a valid measure of uncertainty that uses the same samples. In fact, an interesting consequence of our results is that for interior points and local linear regression ($p = 1$), $I_{\mathtt{rbc}}(h_{\mathtt{mse}})$ has coverage error that vanishes as fast as possible: for this special case, both the mean squared error and coverage error are optimal in rates

upon setting $h = Hn^{-1/(2p+3)}$ for a constant $H > 0$. In other cases, coverage of confidence intervals implemented using $h_{\mathtt{mse}}$ remains consistent but the coverage rate is suboptimal.

To see this, we now turn to inference-optimal bandwidths. We start with the point of view that minimizing coverage error alone is the goal and therefore we choose h by minimizing the terms of (16). This means setting $h_{\mathtt{rbc}} = Hn^{-\eta_{\mathtt{rbc}}}$ for $\eta_{\mathtt{rbc}} = 1/(p+4)$ for interior x with $p-\nu$ odd and $\eta_{\mathtt{rbc}} = 1/(p+3)$ otherwise: Corollary 1 holds for $I_{\mathtt{rbc}}(h_{\mathtt{rbc}})$ with rates $r_{\mathtt{rbc}} = n^{-(p+3)/(p+4)}$ and $r_{\mathtt{rbc}} = n^{-(p+2)/(p+3)}$, respectively. In terms of rates, $h_{\mathtt{rbc}}$ balances the variance and bias of the point estimator, instead of the squared bias as in MSE optimality.

A natural way of choosing the constant H in practice is to minimize the constant portion of the coverage error of (16). Plugging in $h_{\mathtt{rbc}} = Hn^{-\eta_{\mathtt{rbc}}}$ and factoring out the rate we get

$$H_{\mathtt{rbc}} = \arg\min_{H>0} \left| H^{-1}\left\{2\omega_{4,\mathtt{rbc},F}\right\} + H^{1+2\zeta}\left\{2\psi^2_{T_{\mathtt{rbc}},F}\omega_{5,\mathtt{rbc},F}\right\} + H^{\zeta}\left\{2\psi_{T_{\mathtt{rbc}},F}\omega_{6,\mathtt{rbc},F}\right\}\right|.$$

It is straightforward to give a data-driven version of $H_{\mathtt{rbc}}$, and therefore of $h_{\mathtt{rbc}}$, because all quantities involved can be estimated. We defer the details to [4] to conserve space. In a nutshell, plug-in estimators can be constructed, denoted by $\hat{\omega}_{4,\mathtt{rbc},F}$, $\hat{\omega}_{5,\mathtt{rbc},F}$, and $\hat{\omega}_{6,\mathtt{rbc},F}$, as well as an estimate of the bias constant, $\hat{\psi}_{\mathtt{rbc},F}$. We then numerically solve

$$\hat{H}_{\mathtt{rbc}} = \arg\min_{H>0} \left| H^{-1}\left\{2\hat{\omega}_{4,\mathtt{rbc},F}\right\} + H^{1+2\zeta}\left\{2\hat{\psi}^2_{\mathtt{rbc},F}\hat{\omega}_{5,\mathtt{rbc},F}\right\} + H^{\zeta}\left\{2\hat{\psi}_{\mathtt{rbc},F}\hat{\omega}_{6,\mathtt{rbc},F}\right\}\right|.$$

Because this bandwidth depends on the specific data-generating process F, we view it as a rule-of-thumb implementation.

As discussed above, we can also seek for a shorter interval (more power) by sacrificing coverage error (size control). Interval length (17) is reduced for larger bandwidths, meaning smaller exponents η. Corollary 1, or Equation (16) specifically, shows that the smallest η (i.e., the slowest vanishing bandwidth) such that the coverage of $I_{\mathtt{rbc}}(n^{-\eta})$ to be (uniformly) asymptotically correct is $\eta > (1/(1+2\zeta)$, where recall that $\zeta = p+3$ for interior points with $p-\nu$ odd and $\zeta = p+2$ otherwise. Therefore, taking $h = Hn^{-\eta}$ for any $\eta > (1/(1+2\zeta)$ and $H > 0$ results in the ideal interval given these preferences over coverage error and length.

This same idea can be extended to accomplish a *trade-off* between coverage error and length. Researchers may want to have an interval that is closer to nominal level, and therefore may be concerned that in finite samples an interval with coverage error only known to obey $r_{\mathtt{rbc}} = o(1)$ will not be satisfactory. We can therefore take $h_{\mathtt{to}} = H_{\mathtt{to}}n^{-\eta_{\mathtt{to}}}$ for some $\eta_{\mathtt{to}} \in (1/(1+2\zeta), \eta_{\mathtt{rbc}}]$. Note that if $\eta > \eta_{\mathtt{rbc}}$ (i.e. $h = o(h_{\mathtt{rbc}})$), both the rate of coverage error decay and interval length contraction can be improved. There is no well-defined optimal choice in this range of asymptotically valid options, as the choice must reflect each researcher's preference for length vs. coverage error. This range does not depend on ν, even though the resulting length will, see (17). This may affect how the researcher wishes to trade off the two quantities. The endpoints of the range for $\eta_{\mathtt{to}}$ represent preferences for only optimizing coverage error or only length.

To select the constant for this trade off, $H_{\mathtt{to}}$, note first that for $\eta < \eta_{\mathtt{rbc}}$ the middle term of the coverage error (16) is dominant. This term, $n^{1-\eta_{\mathtt{to}}(1+2\zeta)}\left\{2\psi^2_{T,F}\omega_{5,T,F}\right\}$, shares the rate of the scaled, squared bias. Therefore, it is natural to balance this against the square of interval length, to match the trade off that $h_{\mathtt{to}}$ represents. The feasible choice of this constant, $\hat{H}_{\mathtt{to}}$, will also be a direct plug-in rule that uses the estimators above and a pilot version of $\hat{\sigma}^2_{\mathtt{rbc}}$, as well a researcher's choice of weight $\mathcal{H} \in (0,1)$ capturing their trade off between the two. Put altogether, we can then set

$$\hat{H}_{\mathtt{to}} = \arg\min_{H>0} \left\{ \mathcal{H} \cdot H^{1+2\zeta}\left(2\hat{\psi}^2_{\mathtt{rbc},F}\hat{\omega}_{5,\mathtt{rbc},F}\right) + (1-\mathcal{H}) \cdot 4z^2_{\alpha/2}\frac{\hat{\sigma}^2_{\mathtt{rbc}}}{H^{1+2\nu}} \right\}$$

$$= \left(\frac{(1-\mathcal{H})(1+2\nu)4z_{\alpha/2}^2 \hat{\sigma}_{\mathtt{rbc}}^2}{\mathcal{H}(1+2\zeta)2\hat{\psi}_{\mathtt{rbc},F}^2 \hat{\omega}_{5,\mathtt{rbc},F}} \right).$$

The resulting data-driven bandwidth choice is $\hat{h}_{\mathtt{to}} = \hat{H}_{\mathtt{to}} n^{-\eta_{\mathtt{to}}}$, for a choice $\eta_{\mathtt{to}} \in (1/(1+2\zeta), \eta_{\mathtt{rbc}}]$, and weight $\mathcal{H} \in (0,1)$. The supplement [4] contains details and some additional results.

4.2. Interval length optimality: Choosing ρ and $K(\cdot)$

To complete the implementation of $I_{\mathtt{rbc}}(h)$ we need to select the bias-correction bandwidth b, which we do in the form of $\rho = h/b$, and the kernel function $K(\cdot)$. We choose these to optimize the length (17). With ρ bounded and bounded away from zero, this choice affects only the constant portions of the coverage error expansion of $I_{\mathtt{rbc}}(h_{\mathtt{rbc}})$, in particular changing the shape of the *equivalent kernel* of $\hat{\theta}_{\mathtt{rbc}}$. For more details on equivalent kernels, see [17, Sect. 3.2.2]. To find this equivalent kernel, begin by writing $\hat{\theta}_{\mathtt{rbc}} = \nu! \boldsymbol{e}_\nu' \boldsymbol{\Gamma}^{-1} \boldsymbol{\Omega}_{\mathtt{rbc}} \boldsymbol{Y} / nh^\nu$ as a weighted average of the Y_i. Recall that $X_{h,i} = (X_i - \mathsf{x})/h$ and similarly for $X_{b,i}$. Then,

$$\begin{aligned}
\hat{\theta}_{\mathtt{rbc}} &= \frac{1}{nh^\nu} \nu! \boldsymbol{e}_\nu' \boldsymbol{\Gamma}^{-1} \left(\boldsymbol{\Omega} - \rho^{p+1} \boldsymbol{\Lambda} \boldsymbol{e}_{p+1}' \bar{\boldsymbol{\Gamma}}^{-1} \bar{\boldsymbol{\Omega}} \right) \boldsymbol{Y} \\
&= \frac{1}{nh^{1+\nu}} \sum_{i=1}^n \left\{ \nu! \boldsymbol{e}_\nu' \boldsymbol{\Gamma}^{-1} \left(K(X_{h,i}) \boldsymbol{r}_p(X_{h,i}) - \rho^{p+1} \frac{h}{b} \boldsymbol{\Lambda} \boldsymbol{e}_{p+1}' \bar{\boldsymbol{\Gamma}}^{-1} K(X_{b,i}) \boldsymbol{r}_{p+1}(X_{b,i}) \right) \right\} Y_i.
\end{aligned}$$

The weights here depend on the sample, as $\boldsymbol{\Gamma}$, $\boldsymbol{\Lambda}$, and $\bar{\boldsymbol{\Gamma}}$ are sample quantities. The equivalent kernel replaces these with their limiting versions (not, as elsewhere, their fixed-n expectations), which we shall denote $\mathsf{G} = f(\mathsf{x}) \int K(u) \boldsymbol{r}_p(u) \boldsymbol{r}_p(u)' du$, $\mathsf{L} = f(\mathsf{x}) \int K(u) \boldsymbol{r}_p(u) u^{p+1} du$, and $\bar{\mathsf{G}} = f(\mathsf{x}) \int K(u) \boldsymbol{r}_{p+1}(u) \boldsymbol{r}_{p+1}(u)' du$, respectively. The integrals are over $[-1,1]$ if x is an interior point and appropriately truncated when x is a boundary point. Under our assumptions, convergence to these limits is fast enough that, for the equivalent kernel $\mathcal{K}_{\mathtt{rbc}}(u; K, \rho, \nu)$ defined as

$$\mathcal{K}_{\mathtt{rbc}}(u; K, \rho, \nu) = \nu! \boldsymbol{e}_\nu' \mathsf{G}^{-1} \left(K(u) \boldsymbol{r}_p(u) - \rho^{p+2} \mathsf{L} \boldsymbol{e}_{p+1}' \bar{\mathsf{G}}^{-1} K(u\rho) \boldsymbol{r}_{p+1}(u\rho) \right),$$

and we have the representation

$$\hat{\theta}_{\mathtt{rbc}} = \frac{1}{nh^{1+\nu}} \sum_{i=1}^n \mathcal{K}_{\mathtt{rbc}}\left(X_{h,i}; K, \rho, \nu\right) Y_i \, \{1 + o_{\mathbb{P}}(1)\}.$$

It follows that the (constant portion of the) asymptotic length of $I_{\mathtt{rbc}}(h)$ depends on $K(\cdot)$ and ρ only through the specific functional $\int \left(\mathcal{K}_{\mathtt{rbc}}(u; K, \rho, \nu)\right)^2 du$, which corresponds to the asymptotic variance.

The asymptotic variance of a local polynomial point estimator at a boundary or interior point is minimized by employing the uniform kernel [13]. Therefore, to minimize the constant term of interval length we choose ρ, depending on K, to make $\mathcal{K}_{\mathtt{rbc}}(u; K, \rho, \nu)$ as close as possible to the optimal equivalent kernel, i.e. the $\mathcal{K}_p^*(u)$ induced by the uniform kernel for a given p. If the uniform kernel is used initially, then $\rho^* = 1$ is optimal: that is, $\mathcal{K}_{\mathtt{rbc}}(\cdot; \mathbb{1}\{|u| < 1\}/2, 1, \nu) \equiv \mathcal{K}_{p+1}^*(\cdot)$. This highlights the importance of being able to accommodate the uniform kernel in our higher-order expansions. If a kernel other than uniform is used, we look for the optimal choice of ρ by minimizing the L_2 distance between the induced equivalent kernel and the optimal variance-minimizing equivalent kernel, solving

$$\rho^* = \arg\min_{\rho > 0} \int \left| \mathcal{K}_{\mathtt{rbc}}\left(u; K, \rho, \nu\right) - \mathcal{K}_{p+1}^*(u) \right|^2 du.$$

Table 2. L_2-Optimal Variance-Minimizing ρ

p	Kernel		
	Triangular	Epanechnikov	Uniform
0	0.778	0.846	1.000
1	0.850	0.898	1.000
2	0.887	0.924	1.000
3	0.909	0.940	1.000
4	0.924	0.950	1.000

(a) Boundary point

p	Kernel		
	Triangular	Epanechnikov	Uniform
1	0.798	0.865	1.000
3	0.867	0.915	1.000
5	0.900	0.938	1.000
7	0.919	0.951	1.000

(b) Interior point

Note: Optimal ρ computed by minimizing the L_2 distance between the RBC induced equivalent kernel and the variance-minimizing equivalent kernel (Uniform Kernel).

This is not a sample-dependent problem, only computational. For $p - \nu$ odd, the standard case in practice, Table 2 shows the optimal ρ^*, for boundary and interior points, respectively, the triangular kernel ($K(u) = (1 - |u|)\mathbb{1}(|u| \leq 1)$) and the Epanechnikov kernel ($K(u) = 0.75(1 - u^2)\mathbb{1}(|u| \leq 1)$). These two are popular choices and are MSE-optimal at boundary and interior points, respectively. The shapes of the resulting equivalent kernel, $\mathcal{K}_{\texttt{rbc}}(u; K, \rho^*, \nu)$, are shown in Figure 1 for $\nu = \{0, 1\}$. Note that although ρ^* itself does not vary with ν, the equivalent kernel shape does. Additional choices of p are illustrated in the supplement[4].

5. Minimax coverage error decay rates

In this section we build on [23] and look for a minimax result: characterizing the fastest (minimal) rate at which the worst-case (maximal) coverage error vanishes. The "optimal" interval estimator is one for which this maximal error is minimized. At an intuitive level, this corresponds to the desire for similarity in testing: the confidence interval should have "similar" coverage over the set of plausible distributions. [23] proposed this inference-specific notion of minimax optimality and studied it in the case of one-sided confidence intervals in the i.i.d. parametric location model. This problem is different from the more typical minimaxity considered for point estimation, though the latter is established for robust bias correction by [26] and is discussed more broadly for local polynomials by [13] and [16].

To state the problem more formally, let $\mathscr{I}_p$ denote a class of confidence interval estimators. We then define the *minimax coverage error* as

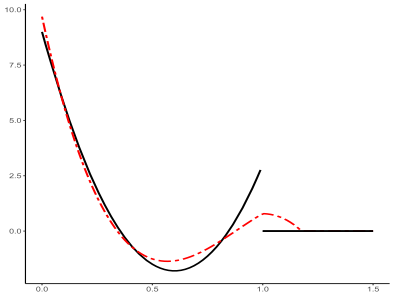
$$\text{MCE}_n := \inf_{I \in \mathscr{I}_p} \sup_{F \in \mathscr{F}_S} \left| \mathbb{P}_F \left[\mu^{(\nu)}(\mathsf{x}) \in I \right] - (1 - \alpha) \right|,$$

where the dependence on the fixed quantities, such as the classes $\mathscr{I}_p$ and $\mathscr{F}_S$ or the level α, are suppressed. Our goal is to characterize the *minimax optimal coverage error decay rate bound*, which is the fastest vanishing sequence $r_\star = r_\star(n)$, $n \in \mathbb{N}$, such that for constants c_1 and c_2,

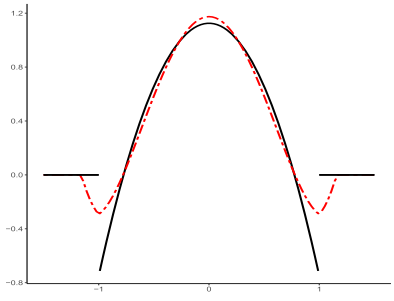
$$0 < c_1 \leq \liminf_{n \to \infty} r_\star^{-1} \text{MCE}_n \leq \limsup_{n \to \infty} r_\star^{-1} \text{MCE}_n \leq c_2 < \infty. \tag{18}$$

We have already characterized the worst-case coverage error in Corollary 1 for the class of distributions defined in Section 2. The key point here is that if we take $\mathscr{I}_p$ to be the class of intervals for which we studied worst-case coverage error in Corollary 1, then we can characterize the minimax rate $r_\star$ as

(a) $\nu = 0$

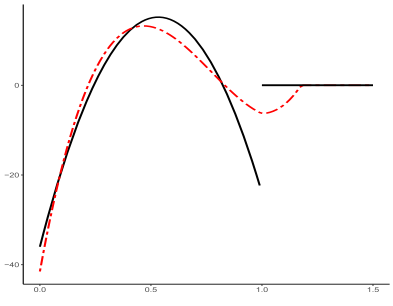


(*i*) Triangular Kernel, Boundary Point, $p = 1$

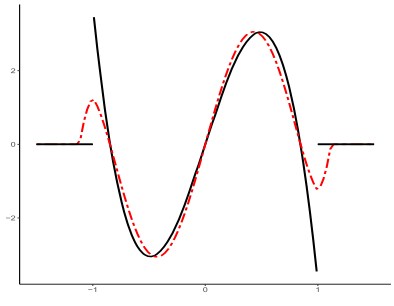


(*ii*) Epanechnikov Kernel, Interior Point, $p = 1$

(b) $\nu = 1$



(*iii*) Triangular Kernel, Boundary Point, $p = 1$



(*iv*) Epanechnikov Kernel, Interior Point, $p = 2$

Notes: $\mathcal{K}^*_{p+1}(u)$, $\mathcal{K}_{\texttt{rbc}}(u;K,\rho^*,\nu)$

Figure 1: $\mathcal{K}^*_{p+1}(u)$ vs. $\mathcal{K}_{\texttt{rbc}}(u;K,\rho^*,\nu)$

well as intervals which attain it. Specifically, we take $\mathscr{I}_p$ to be the Wald-type intervals of the form (12), based on a local polynomial of degree p, with any choice of centering, scaling, bandwidth(s), kernel shape, and quantiles, discussed in Section 2. This includes all those intervals dual to t statistics covered by Theorem 1, but also includes other choices which are not asymptotically level $1-\alpha$. Examples include trivial cases such improper choices of quantiles or inconsistent variance estimators, but also choices such as $I_p(h_{\texttt{mse}})$, i.e., using the MSE-optimal bandwidth sequence for with centering $\hat{\mu}_p^{(\nu)}$ and scaling $\hat{\sigma}_p^2/(nh^{1+2\nu})$. We could also include other procedures, such as bootstrap based quantiles, empirically chosen bandwidths, or empirical likelihood methods, as these will not improve on the worst-case coverage error [10,21,24].

Crucial to proving that such an interval is minimax optimal is that the bias vanishes at the best possible rate, given the smoothness assumed (S) and utilized (p), and this in turn depends on whether x is an interior or boundary point. Collecting all the smoothness cases studied in Section 3.1, we immediately obtain the following result (see [4] for omitted details).

Corollary 2. *Let Assumptions 1 and 2 hold and let $\mathscr{I}_p$ be the class of Wald-type confidence intervals described in the foregoing paragraph.*

(i) Let x *be an interior point in the support of X. If $p-\nu$ is odd, then (18) holds with $r_\star = n^{-(p+3)/(p+4)}$ if $p \leq S-3$ and $r_\star = n^{-(S+s)/(S+s+1)}$ if $p \geq S-2$. If $p-\nu$ is even, then $r_\star = n^{-(p+2)/(p+3)}$ if $p \leq S-2$ and $r_\star = n^{-(S+s)/(S+s+1)}$ if $p \geq S-1$.*

(ii) Let x *be a boundary point of the support of X. Then, (18) holds with $r_\star = n^{-(p+2)/(p+3)}$ if $p \leq S-2$ and $r_\star = n^{-(S+s)/(S+s+1)}$ if $p \geq S-1$.*

For the classes $\mathscr{F}_S$ and $\mathscr{I}_p$ considered herein, this result establishes the minimax rate bounds. The interplay between the two classes is crucial: they should be neither too "large" nor too "small" in order to obtain useful and interesting results. The larger is $\mathscr{F}_S$, the more plausible a given data set is generated by some $F \in \mathscr{F}_S$, but well known results dating back at least to [1] show that if $\mathscr{F}_S$ is too large it is impossible to construct an "effective confidence interval" that controls the worst-case coverage. Our particular $\mathscr{F}_S$ captures common restrictions in the setting of nonparametric regression, and therefore matches empirical practice. The class $\mathscr{I}_p$ is restricted to contain Wald-type interval estimators commonly employed in practice using nonparametric kernel-based regression methods (but can be trivially extended to cover alternatives mentioned above). Recall that our goal is to identify if RBC confidence intervals improve over other options in a uniform sense, and this result is tailored to that goal.

The main message of Corollary 2 is that $I_{\mathtt{rbc}}(h_{\mathtt{rbc}})$ is minimax optimal in all cases. This strengthens the pointwise improvement offered by robust bias correction to optimality within the class $\mathscr{I}_p$ considered here. Intuitively, this is because robust bias correction successfully exploits additional smoothness if it exists, but is not punished (in rates) if there is no such smoothness due to the change in Studentization. This can be compared to I_p, the classical interval that requires undersmoothing. This interval is optimal only in the case when S is known so that p can be chosen large enough; for a fixed p that is small relative to S this interval is dominated in the minimax sense.

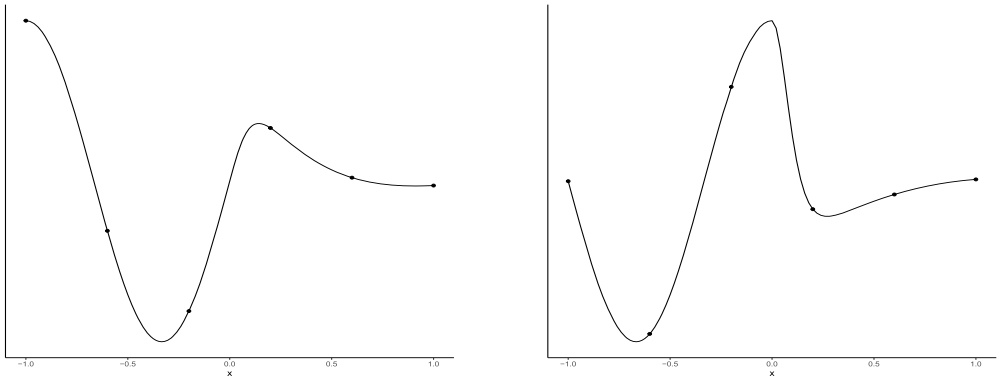
6. Simulation study

This section presents results from a simulation study to examine the finite-sample performance of our methods. Additional results and implementation details can be found in the supplement [4]. We focus on the performance of confidence intervals for $\mu(\mathsf{x})$ and $\mu^{(1)}(\mathsf{x})$ based on robust bias correction and traditional undersmoothing. Data is generated from model (1), with X_i uniformly distributed on $[-1,1]$, ε standard normal, and

$$\mu(x) = \frac{\sin(3\pi x/2)}{1+18x^2(\operatorname{sgn}(x)+1)},$$

where $\operatorname{sgn}(x) = -1$, 0, or -1 according to $x > 0$, $x = 0$ or $x < 0$, respectively. This function, which was also analyzed in [5], is displayed in Figure 2 together with $\mu^{(1)}(x)$. By looking at different evaluation points, we will be able to capture the performance of the methods under different levels of complexity.

We show results for sample sizes $n \in \{100, 250, 500, 750, 1000, 2000\}$, always with 5,000 replications. We study inference at three evaluation points: $\mathsf{x} = -1$ (boundary point), $\mathsf{x} = -0.6$ (low curvature), and $\mathsf{x} = -0.2$ (high curvature). The supplement [4] shows results for $\mathsf{x} \in \{0.2, 0.6, 1\}$. For implementation, we use $p = 1$ (for $\nu = 0$) and $p = 2$ (for $\nu = 1$) with the Epanechnikov kernel ([4] gives results for the uniform kernel). Finally, we evaluate the performance of the confidence intervals using several bandwidth choices. First, following the results from Section 4, we use $\hat{h}_{\mathtt{rbc}}$, a data-driven version of the inference-optimal bandwidth $h_{\mathtt{rbc}}$. We also consider the analogous version for undersmoothed confidence intervals, denoted $\hat{h}_{\mathtt{us}}$ (detailed in [4]), and the standard choice in practice, $\hat{h}_{\mathtt{mse}}$. Robust bias



(a) $\nu = 0$ (b) $\nu = 1$

Figure 2: Conditional mean function and first derivative, $\mu^{(\nu)}(x)$

correction is implemented using $\rho = \rho^*$ according to Table 2. All implementation details are available for R and Stata [6].

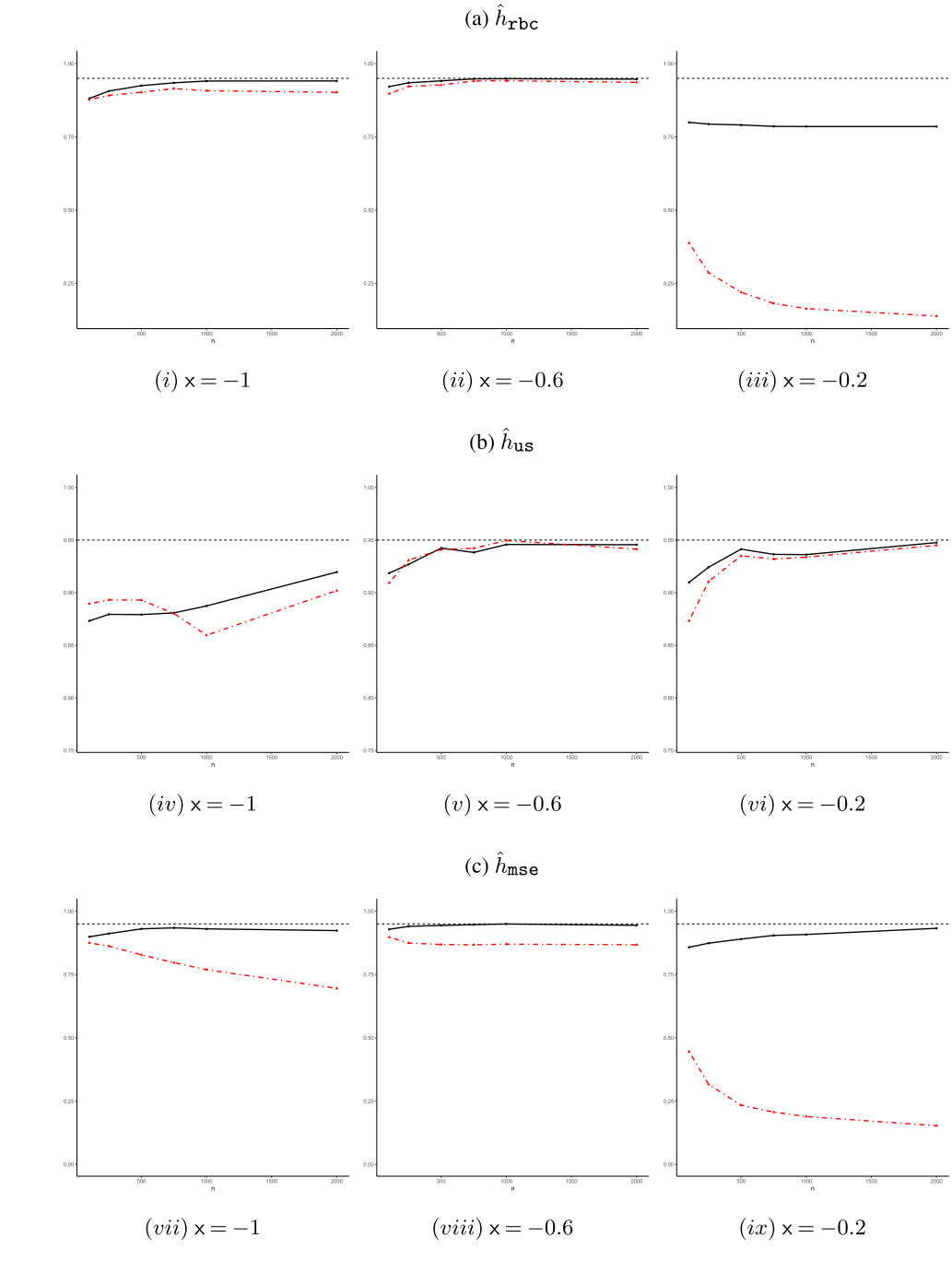
Figures 3 and 4 present empirical coverage probabilities for $\nu = 0$ and $\nu = 1$, respectively, for each evaluation point and choice of bandwidth, as a function of the sample size. Overall, we can see that robust bias correction yields close to accurate coverage, improving over undersmoothing in almost every case. Performance is highly superior at points where the functions present high curvature and also at the boundary. Performance is never worse even when the function is quite linear and optimal bandwidths are (close to) ill-defined.

We also compare confidence interval performance in terms of length in Figure 5. We take coverage into account by looking at RBC and US confidence intervals implemented using their corresponding coverage error optimal bandwidth choices ($\hat{h}_{\texttt{rbc}}$ and $\hat{h}_{\texttt{us}}$, respectively), which is when they perform best in terms of coverage. We also include other valid, but non optimal choices $I_{\texttt{rbc}}(\hat{h}_{\texttt{mse}})$, $I_{\texttt{rbc}}(\hat{h}_{\texttt{us}})$. We find that RBC confidence intervals are, on average, not larger than US, and sometimes even shorter. Lastly, Figure 6 shows the average estimated bandwidths at each point for each sample size, which behave as expected following our theory.

7. Conclusion

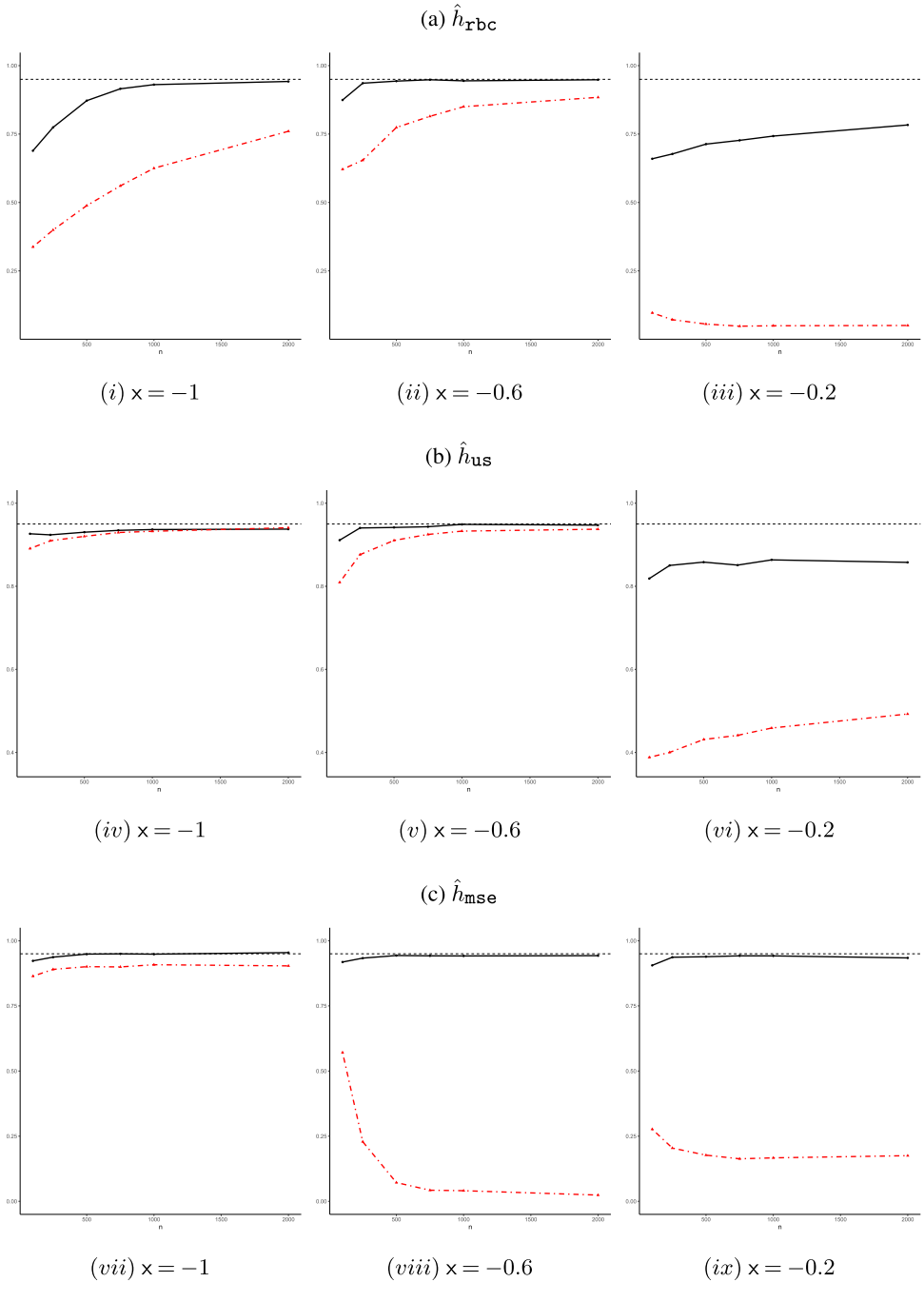
This paper derived higher order expansions for inference in nonparametric local polynomial regression. We provided new Edgeworth expansions and associated error in coverage probability expansions for standard and robust bias corrected methods, showing that the latter have superior coverage properties. Our results hold uniformly in the data generating process, cover derivative estimation, and allow for the uniform kernel. Using our results we developed novel bandwidth selections that target inference directly, achieving lower coverage error and/or shorter length.

Our main results measured coverage error symmetrically, but it is worth mentioning that the absolute loss function may be replaced by the "check" loss function, and thus studying the maximal coverage error $\sup_{F\in\mathscr{F}_S} \mathcal{L}(\mathbb{P}_F[\theta_F \in I] - (1-\alpha))$, with $\mathcal{L}(e) = \mathcal{L}_\tau(e) = e(\tau - \mathbb{1}\{e < 0\})$, and where $\tau \in (0,1)$ encodes the researcher's weight for over- and under-coverage. Setting $\tau = 1/2$ recovers the above, symmetric measure of coverage error. Guarding more against undercoverage (a preference for conservative



Notes: Robust Bias Correction, Undersmoothing; Epanechnikov Kernel

Figure 3: Empirical Coverage for 95% Confidence Intervals, $\nu = 0$



Notes: Robust Bias Correction, Undersmoothing; Epanechnikov Kernel

Figure 4: Empirical Coverage for 95% Confidence Intervals, $\nu = 1$

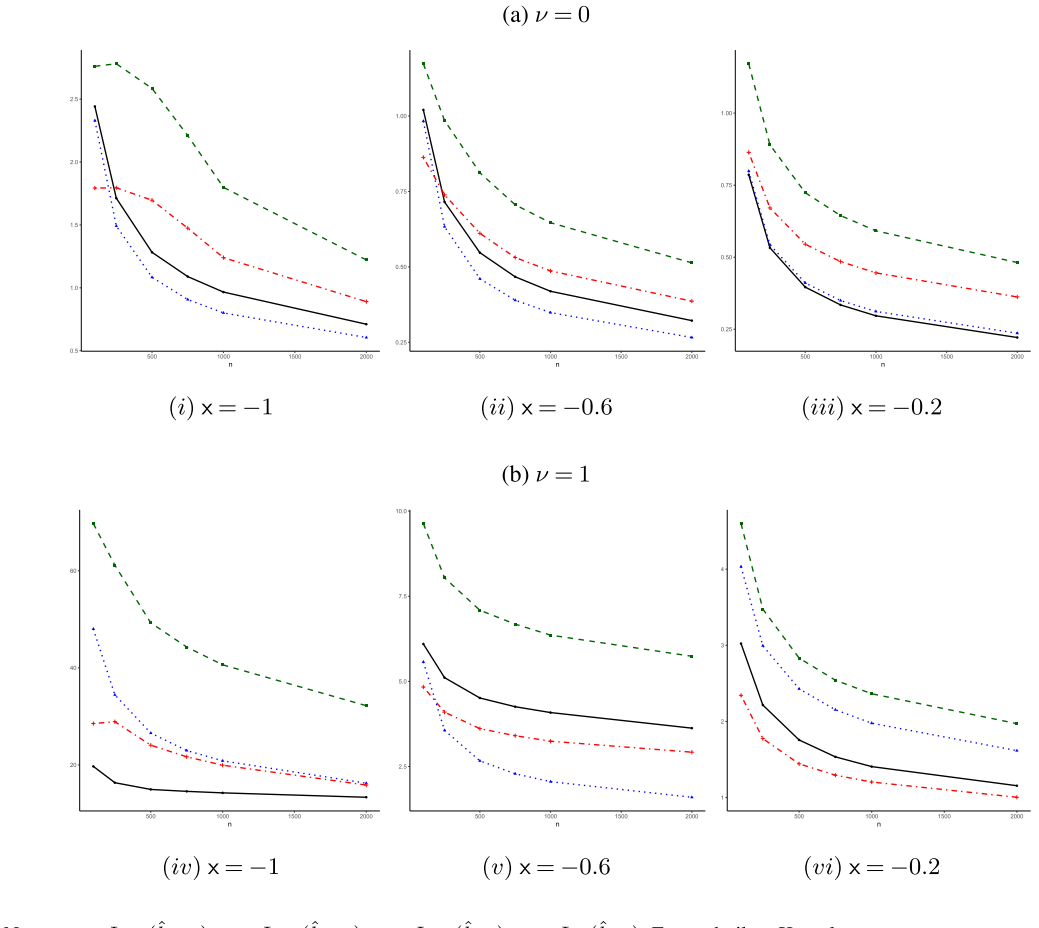


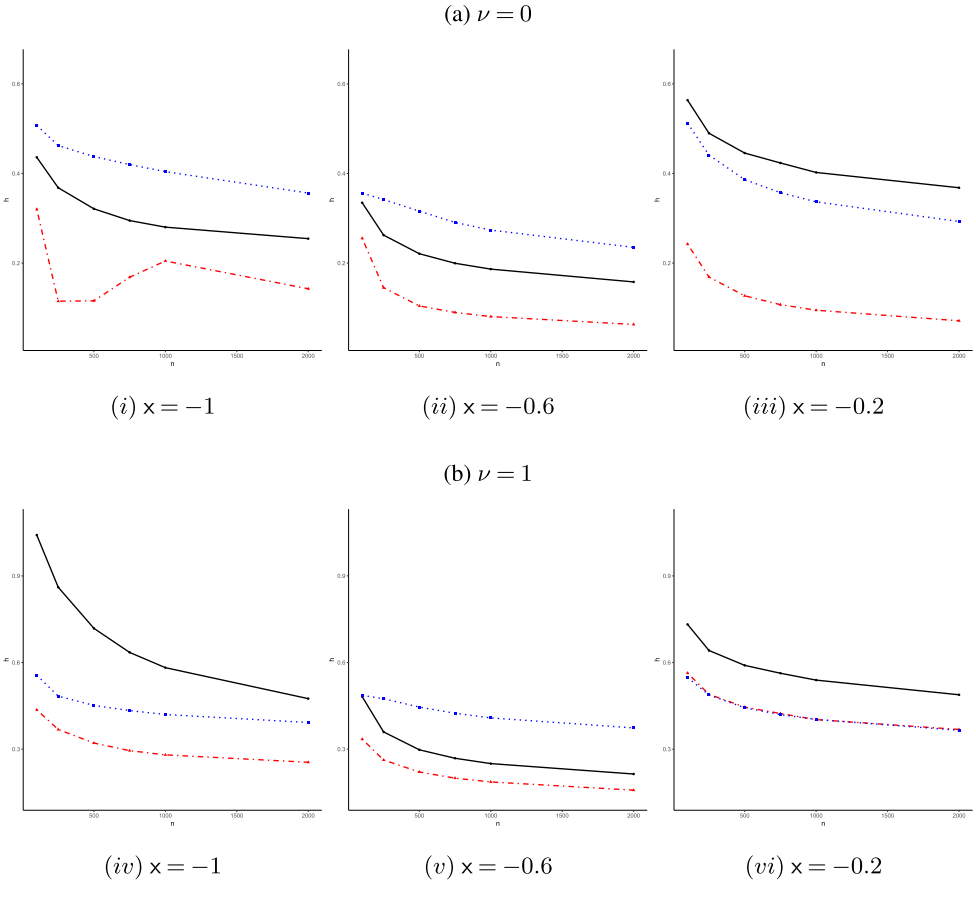
Figure 5: Average Interval Length for 95% Confidence Intervals

intervals) requires choosing a $\tau < 1/2$. For example, setting $\tau = 1/3$ encodes the belief that undercoverage is twice as bad as the same amount of overcoverage. All our results can be established for this loss function.

Finally, this paper studied the properties of confidence intervals at a fixed evaluation point x, but it would be of theoretical and practical interest to extent our results to the case of confidence band construction. Robust bias correction has recently been used to construct valid confidence bands for local polynomial estimation [12] and linear sieve estimation [9]. Because the underlying distributional approximations for confidence band constructions are substantially more complex, obtaining results similar to those presented herein will require substantial extension of our technical work.

Appendix: Terms of the Edgeworth expansion

We give the definition of ω_k, $k = 1,2,\ldots,6$. First, define the following objects, all calculated in a fixed-n sense, bounded uniformly in $\mathscr{F}_S$, and nonzero for some $F \in \mathscr{F}_S$. As shorthand, let a tilde accent denote



Notes: ——— $\hat{h}_{\texttt{rbc}}$, - - - $\hat{h}_{\texttt{us}}$, ····· $\hat{h}_{\texttt{mse}}$; Epanechnikov Kernel

Figure 6: Average Estimated Bandwidth

a fixed-n expectation, so that $\tilde{\mathbf{\Gamma}} = \mathbb{E}[\mathbf{\Gamma}]$, $\tilde{\mathbf{\Lambda}}_1 = \mathbb{E}[\mathbf{\Lambda}_1]$, and so forth. Let

$$\ell^0_{T_p}(X_i) = \nu! \boldsymbol{e}_\nu' \tilde{\mathbf{\Gamma}}^{-1} (K \boldsymbol{r}_p)(X_{h,i});$$

$$\ell^0_{T_{\texttt{rbc}}}(X_i) = \ell^0_{T_p}(X_i) - \rho^{p+1} \nu! \boldsymbol{e}_\nu' \tilde{\mathbf{\Gamma}}^{-1} \tilde{\mathbf{\Lambda}}_1 \boldsymbol{e}_{p+1}' \tilde{\tilde{\mathbf{\Gamma}}}^{-1} (K \boldsymbol{r}_{p+1})(X_{b,i});$$

$$\ell^1_{T_p}(X_i, X_j) = \nu! \boldsymbol{e}_\nu' \tilde{\mathbf{\Gamma}}^{-1} \left(\mathbb{E}[(K \boldsymbol{r}_p \boldsymbol{r}_p')(X_{h,j})] - (K \boldsymbol{r}_p \boldsymbol{r}_p')(X_{h,j}) \right) \tilde{\mathbf{\Gamma}}^{-1} (K \boldsymbol{r}_p)(X_{h,i});$$

$$\begin{aligned} \ell^1_{T_{\texttt{rbc}}}(X_i, X_j) = \ell^1_{T_p}(X_i, X_j) - \rho^{p+1} \nu! \boldsymbol{e}_\nu' \tilde{\mathbf{\Gamma}}^{-1} \Big\{ & \left(\mathbb{E}[(K \boldsymbol{r}_p \boldsymbol{r}_p')(X_{h,j})] - (K \boldsymbol{r}_p \boldsymbol{r}_p')(X_{h,j}) \right) \tilde{\mathbf{\Gamma}}^{-1} \tilde{\mathbf{\Lambda}}_1 \boldsymbol{e}_{p+1}' \\ & + \left((K \boldsymbol{r}_p)(X_{h,j}) X_{h,i}^{p+1} - \mathbb{E}[(K \boldsymbol{r}_p)(X_{h,j}) X_{h,i}^{p+1}] \right) \boldsymbol{e}_{p+1}' \\ & + \tilde{\mathbf{\Lambda}}_1 \boldsymbol{e}_{p+1}' \tilde{\tilde{\mathbf{\Gamma}}}^{-1} \left(\mathbb{E}[(K \boldsymbol{r}_{p+1} \boldsymbol{r}_{p+1}')(X_{b,j})] - (K \boldsymbol{r}_{p+1} \boldsymbol{r}_{p+1}')(X_{b,j}) \right) \Big\} \tilde{\tilde{\mathbf{\Gamma}}}^{-1} (K \boldsymbol{r}_{p+1})(X_{b,i}). \end{aligned}$$

Then define $\tilde{\sigma}_T^2 = \mathbb{E}[h^{-1}\ell_T^0(X)^2 v(X)]$ and denote the standard Normal density as $\phi(z)$. Then we define

$$
\begin{aligned}
\omega_{1,T,F}(z) &= \phi(z)\tilde{\sigma}_T^{-3}\mathbb{E}\left[h^{-1}\ell_T^0(X_i)^3\varepsilon_i^3\right]\left\{(2z^2-1)/6\right\},\\
\omega_{2,T,F}(z) &= -\phi(z)\tilde{\sigma}_T^{-1},\\
\omega_{3,T,F}(z) &= -\phi(z)\left\{z/2\right\},\\
\omega_{5,T,F}(z) &= -\phi(z)\tilde{\sigma}_T^{-2}\left\{z/2\right\},\\
\omega_{6,T,F}(z) &= \phi(z)\tilde{\sigma}_T^{-4}\mathbb{E}[h^{-1}\ell_T^0(X_i)^3\varepsilon_i^3]\left\{z^3/3\right\}.
\end{aligned}
$$

For ω_4, it is not quite as simple to state a generic version. Let $\tilde{\boldsymbol{G}}$ stand in for $\tilde{\boldsymbol{\Gamma}}$ or $\tilde{\tilde{\boldsymbol{\Gamma}}}$, $\tilde{p}$ stand in for p or $p+1$, and d_n stand in for h or b, all depending on if $T = T_p$ or $T_{\texttt{rbc}}$. Note however, that h is still used in many places, in particular for stabilizing fixed-n expectations, for $T_{\texttt{rbc}}$. Indexes i, j, and k are always distinct (i.e. $X_{h,i} \neq X_{h,j} \neq X_{h,k}$).

$$
\begin{aligned}
\omega_{4,T,F}(z) = {}& \phi(z)\tilde{\sigma}_T^{-6}\mathbb{E}\left[h^{-1}\ell_T^0(X_i)^3\varepsilon_i^3\right]^2\left\{z^3/3+7z/4+\tilde{\sigma}_T^2 z(z^2-3)/4\right\}\\
&+\phi(z)\tilde{\sigma}_T^{-2}\mathbb{E}\left[h^{-1}\ell_T^0(X_i)\ell_T^1(X_i,X_i)\varepsilon_i^2\right]\left\{-z(z^2-3)/2\right\}\\
&+\phi(z)\tilde{\sigma}_T^{-4}\mathbb{E}\left[h^{-1}\ell_T^0(X_i)^4(\varepsilon_i^4-v(X_i)^2)\right]\left\{z(z^2-3)/8\right\}\\
&-\phi(z)\tilde{\sigma}_T^{-2}\mathbb{E}\left[h^{-1}\ell_T^0(X_i)^2\boldsymbol{r}_{\tilde{p}}(X_{d_n,i})'\tilde{\boldsymbol{G}}^{-1}(K\boldsymbol{r}_{\tilde{p}})(X_{d_n,i})\varepsilon_i^2\right]\left\{z(z^2-1)/2\right\}\\
&-\phi(z)\tilde{\sigma}_T^{-4}\mathbb{E}\left[h^{-1}\ell_T^0(X_i)^3\boldsymbol{r}_{\tilde{p}}(X_{d_n,i})'\tilde{\boldsymbol{G}}^{-1}\varepsilon_i^2\right]\mathbb{E}\left[h^{-1}(K\boldsymbol{r}_{\tilde{p}})(X_{d_n,i})\ell_T^0(X_i)\varepsilon_i^2\right]\left\{z(z^2-1)\right\}\\
&+\phi(z)\tilde{\sigma}_T^{-2}\mathbb{E}\left[h^{-2}\ell_T^0(X_i)^2(\boldsymbol{r}_{\tilde{p}}(X_{d_n,i})'\tilde{\boldsymbol{G}}^{-1}(K\boldsymbol{r}_{\tilde{p}})(X_{d_n,j}))^2\varepsilon_j^2\right]\left\{z(z^2-1)/4\right\}\\
&+\phi(z)\tilde{\sigma}_T^{-4}\mathbb{E}\left[h^{-3}\ell_T^0(X_j)^2\boldsymbol{r}_{\tilde{p}}(X_{d_n,j})'\tilde{\boldsymbol{G}}^{-1}(K\boldsymbol{r}_{\tilde{p}})(X_{d_n,i})\ell_T^0(X_i)\boldsymbol{r}_{\tilde{p}}(X_{d_n,j})'\right.\\
&\left.\times\,\tilde{\boldsymbol{G}}^{-1}(K\boldsymbol{r}_{\tilde{p}})(X_{d_n,k})\ell_T^0(X_k)\varepsilon_i^2\varepsilon_k^2\right]\left\{z(z^2-1)/2\right\}\\
&+\phi(z)\tilde{\sigma}_T^{-4}\mathbb{E}\left[h^{-1}\ell_T^0(X_i)^4\varepsilon_i^4\right]\left\{-z(z^2-3)/24\right\}\\
&+\phi(z)\tilde{\sigma}_T^{-4}\mathbb{E}\left[h^{-1}\left(\ell_T^0(X_i)^2v(X_i)-\mathbb{E}[\ell_T^0(X_i)^2v(X_i)]\right)\ell_T^0(X_i)^2\varepsilon_i^2\right]\left\{z(z^2-1)/4\right\}\\
&+\phi(z)\tilde{\sigma}_T^{-4}\mathbb{E}\left[h^{-2}\ell_T^1(X_i,X_j)\ell_T^0(X_i)\ell_T^0(X_j)^2\varepsilon_j^2v(X_i)\right]\left\{z(z^2-3)\right\}\\
&+\phi(z)\tilde{\sigma}_T^{-4}\mathbb{E}\left[h^{-2}\ell_T^1(X_i,X_j)\ell_T^0(X_i)\left(\ell_T^0(X_j)^2v(X_j)-\mathbb{E}[\ell_T^0(X_j)^2v(X_j)]\right)\varepsilon_i^2\right]\{-z\}\\
&+\phi(z)\tilde{\sigma}_T^{-4}\mathbb{E}\left[h^{-1}\left(\ell_T^0(X_i)^2v(X_i)-\mathbb{E}[\ell_T^0(X_i)^2v(X_i)]\right)^2\right]\left\{-z(z^2+1)/8\right\}.
\end{aligned}
$$

Acknowledgements

We especially thank an Associate Editor, and the reviewers, for insightful comments that improve our manuscript. We also thank Chris Hansen, Michael Jansson, Adam McCloskey, Rocio Titiunik, and participants at various seminars and conferences for comments.

Funding

The second author gratefully acknowledges financial support from the National Science Foundation (SES 1357561, SES 1459931, and SES-1947805) and the National Institutes of Health (R01 GM072611-16). The third author gratefully acknowledges financial support from the Richard N. Rosett and John E. Jeuck Fellowships.

Supplementary Material

Supplement to "Coverage error optimal confidence intervals for local polynomial regression" (DOI: 10.3150/21-BEJ1445SUPP; .pdf). This supplement contains proofs of all results, other technical details, and complete simulation results.

References

[1] Bahadur, R.R. and Savage, L.J. (1956). The nonexistence of certain statistical procedures in nonparametric problems. *Ann. Math. Stat.* **27** 1115–1122. MR0084241 https://doi.org/10.1214/aoms/1177728077

[2] Beran, R. (1982). Estimated sampling distributions: The bootstrap and competitors. *Ann. Statist.* **10** 212–225. MR0642733

[3] Bhattacharya, R.N. and Ranga Rao, R. (1976). *Normal Approximation and Asymptotic Expansions*. *Wiley Series in Probability and Mathematical Statistics*. New York: Wiley. MR0436272

[4] Calonico, S., Cattaneo, M.D. Farrell, M.H. (2022). Supplement to "Coverage error optimal confidence intervals for local polynomial regression." https://doi.org/10.3150/21-BEJ1445SUPP

[5] Calonico, S., Cattaneo, M.D. and Farrell, M.H. (2018). On the effect of bias estimation on coverage accuracy in nonparametric inference. *J. Amer. Statist. Assoc.* **113** 767–779. MR3832225 https://doi.org/10.1080/01621459.2017.1285776

[6] Calonico, S., Cattaneo, M.D. and Farrell, M.H. (2019). `nprobust`: Nonparametric kernel-based estimation and robust bias-corrected inference. *J. Stat. Softw.* **91** 1–33.

[7] Calonico, S., Cattaneo, M.D. and Farrell, M.H. (2020). Optimal bandwidth choice for robust bias-corrected inference in regression discontinuity designs. *Econom. J.* **23** 192–210. MR4108924 https://doi.org/10.1093/ectj/utz022

[8] Calonico, S., Cattaneo, M.D. and Titiunik, R. (2014). Robust nonparametric confidence intervals for regression-discontinuity designs. *Econometrica* **82** 2295–2326. MR3301169 https://doi.org/10.3982/ECTA11757

[9] Cattaneo, M.D., Farrell, M.H. and Feng, Y. (2020). Large sample properties of partitioning-based series estimators. *Ann. Statist.* **48** 1718–1741. MR4124341 https://doi.org/10.1214/19-AOS1865

[10] Chen, S.X. and Qin, Y.S. (2000). Empirical likelihood confidence intervals for local linear smoothers. *Biometrika* **87** 946–953. MR1813987 https://doi.org/10.1093/biomet/87.4.946

[11] Chen, S.X. and Qin, Y.S. (2002). Confidence intervals based on local linear smoother. *Scand. J. Stat.* **29** 89–99. MR1894383 https://doi.org/10.1111/1467-9469.00273

[12] Cheng, G. and Chen, Y.-C. (2019). Nonparametric inference via bootstrapping the debiased estimator. *Electron. J. Stat.* **13** 2194–2256. MR3980957 https://doi.org/10.1214/19-EJS1575

[13] Cheng, M.-Y., Fan, J. and Marron, J.S. (1997). On automatic boundary corrections. *Ann. Statist.* **25** 1691–1708. MR1463570 https://doi.org/10.1214/aos/1031594737
[14] Edgeworth, F.Y. (1883). The law of error. *The London, Edinburgh and Dublin Philosophical Magazine* **5** 300–309.
[15] Edgeworth, F.Y. (1906). The generalised law of error, or law of great numbers. *J. Roy. Stat. Soc.* **69** 497–539.
[16] Fan, J., Gasser, T., Gijbels, I., Brockmann, M. and Engel, J. (1997). Local polynomial regression: Optimal kernels and asymptotic minimax efficiency. *Ann. Inst. Statist. Math.* **49** 79–99. MR1450693 https://doi.org/10.1023/A:1003162622169
[17] Fan, J. and Gijbels, I. (1996). *Local Polynomial Modelling and Its Applications*. *Monographs on Statistics and Applied Probability* **66**. London: CRC Press. MR1383587
[18] Fan, J. and Yao, Q. (2003). *Nonlinear Time Series: Nonparametric and Parametric Methods*. *Springer Series in Statistics*. New York: Springer. MR1964455 https://doi.org/10.1007/b97702
[19] Hall, P. (1991). Edgeworth expansions for nonparametric density estimators, with applications. *Statistics* **22** 215–232. MR1097375 https://doi.org/10.1080/02331889108802305
[20] Hall, P. (1992). *The Bootstrap and Edgeworth Expansion*. *Springer Series in Statistics*. New York: Springer. MR1145237 https://doi.org/10.1007/978-1-4612-4384-7
[21] Hall, P. (1992). Effect of bias estimation on coverage accuracy of bootstrap confidence intervals for a probability density. *Ann. Statist.* **20** 675–694. MR1165587 https://doi.org/10.1214/aos/1176348651
[22] Hall, P. (1992). On bootstrap confidence intervals in nonparametric regression. *Ann. Statist.* **20** 695–711. MR1165588 https://doi.org/10.1214/aos/1176348652
[23] Hall, P. and Jing, B.-Y. (1995). Uniform coverage bounds for confidence intervals and Berry-Esseen theorems for Edgeworth expansion. *Ann. Statist.* **23** 363–375. MR1332571 https://doi.org/10.1214/aos/1176324525
[24] Hall, P. and Kang, K.-H. (2001). Bootstrapping nonparametric density estimators with empirically chosen bandwidths. *Ann. Statist.* **29** 1443–1468. MR1873338 https://doi.org/10.1214/aos/1013203461
[25] Neumann, M.H. (1997). Pointwise confidence intervals in nonparametric regression with heteroscedastic error structure. *Statistics* **29** 1–36. MR1438531 https://doi.org/10.1080/02331889708802572
[26] Tuvaandorj, P. (2020). Regression discontinuity designs, white noise models, and minimax. *J. Econometrics* **218** 587–608. MR4149240 https://doi.org/10.1016/j.jeconom.2020.04.030

Received November 2020 and revised July 2021