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

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

December 16, 2017

This supplement 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 $T_{\mathtt{us}}$, $T_{\mathtt{bc}}$, and $T_{\mathtt{rbc}}$, 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.

Part S.I contains all material for density estimation at interior points, while Part S.II treats local polynomial regression at both interior and boundary points, as in the main text. Roughly, these have the same generic outline:

- We first present all notation, both for the estimators themselves and the Edgeworth expansions, regardless of when the notation is used, as a collective reference;
- We then discuss optimal bandwidths and other practical matters, expanding on details of the main text;
- Assumptions for validity of the Edgeworth expansions are restated from the main text, and Cramér's condition is discussed;
- Bias properties are discussed in more detail than in the main text, and some things mentioned there are made precise;
- The main Edgeworth expansions are stated, some corollaries are given, and the proofs are given;
- Complete simulation results are presented.

All our methods are implemented in `R` and `STATA` via the `nprobust` package, available from `http://sites.google.com/site/nppackages/nprobust` (see also `http://cran.r-project.org/package=nprobust`). See Calonico et al. (2017) for a complete description.

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

†Department of Economics, University of Miami.

‡Department of Economics and Department of Statistics, University of Michigan.

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

Contents

Part S.I

Kernel Density Estimation and Inference

S.I.1 Notation

Here we collect notation to be used throughout this section, even if it is restated later. Throughout this supplement, let $X_{h,i} = (x - X_i)/h$ and similarly for $X_{b,i}$. The evaluation point is implicit here. In the course of proofs we will frequently write $s = \sqrt{nh}$.

S.I.1.1 Estimators, Variances, and Studentized Statistics

To begin, recall that the original and bias-corrected density estimators are

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

and

$$\hat{f} - \hat{B}_f = \frac{1}{nh}\sum_{i=1}^{n} M\left(X_{h,i}\right), \qquad M(u) := K(u) - \rho^{1+\mathfrak{K}} L^{(\mathfrak{K})}(\rho u)\mu_{K,\mathfrak{K}}, \tag{S.I.1}$$

for symmetric kernel functions $K(\cdot)$ and $L(\cdot)$ that integrate to one on their compact support, h and b are bandwidth sequences that vanish as $n \to \infty$, and where

$$\rho = h/b, \qquad \hat{B}_f = h^{\mathfrak{K}} \hat{f}^{(\mathfrak{K})}(x)\mu_{K,\mathfrak{K}}, \qquad \hat{f}^{(\mathfrak{K})}(x) = \frac{1}{nb^{1+\mathfrak{K}}}\sum_{i=1}^{n} L^{(\mathfrak{K})}\left(X_{b,i}\right),$$

and integrals of the kernel are denoted

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

The three statistics $T_{\mathtt{us}}$, $T_{\mathtt{bc}}$, and $T_{\mathtt{rbc}}$ share a common structure that is exploited to give a unified theorem statement and proof. For $v \in \{1, 2\}$, define

$$\hat{f}_v = \frac{1}{nh}\sum_{i=1}^{n} N_v\left(X_{h,i}\right), \qquad \text{where} \qquad N_1(u) = K(u) \text{ and } N_2(u) = M(u),$$

and M is given in Eqn. (S.I.1). Thus, $\hat{f}_1 = \hat{f}$ and $\hat{f}_2 = \hat{f} - \hat{B}_f$. In exactly the same way, define

$$\sigma_v^2 := nh\mathbb{V}[\hat{f}_v] = \frac{1}{h}\left\{\mathbb{E}\left[N_v\left(X_{h,i}\right)^2\right] - \mathbb{E}\left[N_v\left(X_{h,i}\right)\right]^2\right\}$$

and the estimator

$$\hat{\sigma}_v^2 = \frac{1}{h}\left\{\frac{1}{n}\sum_{i=1}^{n}\left[N_v\left(X_{h,i}\right)^2\right] - \left[\frac{1}{n}\sum_{i=1}^{n}N_v\left(X_{h,i}\right)\right]^2\right\}.$$

The statistic of interest for the generic Edgeworth expansion is, for $1 \leq w \leq v \leq 2$,

$$T_{v,w} := \frac{\sqrt{nh}(\hat{f}_v - f)}{\hat{\sigma}_w}.$$

In this notation,

$$T_{\texttt{us}} = T_{1,1}, \qquad T_{\texttt{bc}} = T_{2,1}, \qquad \text{and} \qquad T_{\texttt{rbc}} = T_{2,2}.$$

S.I.1.2 Edgeworth Expansion Terms

The scaled bias is $\eta_v = \sqrt{nh}(\mathbb{E}[\hat{f}_v] - f)$. The Standard Normal distribution and density functions are $\Phi(z)$ and $\phi(z)$, respectively.

The Edgeworth expansion for the distribution of $T_{v,w}$ will consist of polynomials with coefficients that depend on moments of the kernel(s). To this end, continuing with the generic notation, for nonnegative integers j, k, p, define

$$\gamma_{v,p} = h^{-1}\mathbb{E}\left[N_v\left(X_{h,i}\right)^p\right], \qquad \Delta_{v,j} = \frac{1}{s}\sum_{i=1}^{n}\left\{N_v\left(X_{h,i}\right)^j - \mathbb{E}\left[N_v\left(X_{h,i}\right)^j\right]\right\},$$

and

$$\nu_{v,w}(j,k,p) = \frac{1}{h}\mathbb{E}\left[\left(N_v\left(X_{h,i}\right) - \mathbb{E}\left[N_v\left(X_{h,i}\right)\right]\right)^j\left(N_w\left(X_{h,i}\right)^p - \mathbb{E}\left[N_w\left(X_{h,i}\right)^p\right]\right)^k\right].$$

We abbreviate $\nu_{v,w}(j,0,p) = \nu_v(j)$.

To expand the distribution function, additional polynomials are needed beyond those used in the main text for coverage error. These are

$$p_{v,w}^{(1)}(z) = \phi(z)\sigma_w^{-3}[\nu_{v,w}(1,1,2)z^2/2 - \nu_v(3)(z^2-1)/6],$$
$$p_{v,w}^{(2)}(z) = -\phi(z)\sigma_w^{-3}\mathbb{E}[\hat{f}_w]\nu_{v,w}(1,1,1)z^2, \qquad \text{and} \qquad p_{v,w}^{(3)}(z) = \phi(z)\sigma_w^{-1}.$$

Next, recall from the main text the polynomials used in *coverage error* expansions, here with an explicit argument for a generic quantile z rather than the specific $z_{\alpha/2}$:

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

The corresponding polynomials for expansions of the *distribution function* are

$$q^{(k)}_{v,w}(z) = \frac{1}{2}\frac{\phi(z)}{f} q_k(z; N_w), \qquad k = 1, 2, 3.$$

Finally, the precise forms of Ω_1 and Ω_2 are:

$$\Omega_1 = -2\frac{\mu_{K,\mathfrak{K}}}{\nu_1(2)}\left\{\int f(x-uh)K(u)L^{(\mathfrak{K})}(u\rho)du - b\int f(x-uh)K(u)du\int f(x-ub)L^{(\mathfrak{K})}(u)du\right\}$$

and $\Omega_2 = \mu^2_{K,\mathfrak{K}}\vartheta^{-2}_{K,2}\vartheta_{L^{(\mathfrak{K})},2}$. These only appear for $T_{\mathtt{bc}}$, and so are not indexed by $\{v, w\}$.

All these are discussed in Section S.I.6.

S.I.2 Details of practical implementation

We maintain $\ell = 2$ and recommend $\mathfrak{K} = 2$. For the kernels K and L, we recommend either the second order minimum variance (to minimize interval length) or the MSE-optimal kernels; see Sections S.I.2.3 and S.I.4.2. In the next two subsections we discuss choice of h and ρ.

As argued below in Section S.I.2.3, we shall maintain $\rho = 1$. In the main text we give a direct plug-in (DPI) rule to implement the coverage-error optimal bandwidth. Here we we give complete details for this procedure as well as document a second practical choice, based on a rule-of-thumb (ROT) strategy. Both choices yield the optimal coverage error decay rate of $n^{-(\mathfrak{K}+2)/(1+(\mathfrak{K}+2))}$.

All our methods are implemented in R and STATA via the `nprobust` package, available from http://sites.google.com/site/nppackages/nprobust (see also http://cran.r-project.org/package=nprobust). See Calonico et al. (2017) for a complete description.

Remark S.I.1 (Undercoverage of $I_{\mathtt{us}}(h^*_{\mathtt{mse}})$)**.** It is possible not only to show that $I_{\mathtt{us}}(h^*_{\mathtt{mse}})$ asymptotically undercovers (see Hall and Horowitz (2013) for discussion in the regression context) but also to quantify precisely the coverage. To do so, write $T_{\mathtt{us}} = \sqrt{nh}(\hat{f} - \mathbb{E}[\hat{f}])/\hat{\sigma}_{\mathtt{us}} + \eta_{\mathtt{us}}/\hat{\sigma}_{\mathtt{us}}$, where the first term will be asymptotically standard Normal and the second will be a nonrandom, non-vanishing bias when $h^*_{\mathtt{mse}}$ is used.

To characterize this second term, first we define $h^*_{\mathtt{mse}}$ in in our notation. Recall from Eqn. (S.I.2) and Section S.I.1 that the mean-square error of $\hat{f}$ can be written as $(nh)^{-1}\sigma^2_{\mathtt{us}} + (nh)^{-1}\eta^2_{\mathtt{us}}$. Define $\tilde{\eta}_{\mathtt{us}}$ to be the leading constant of the bias, so that $\eta_{\mathtt{us}} = \sqrt{nh}h^{\mathfrak{K}}[\tilde{\eta}_{\mathtt{us}} + o(1)]$ and the MSE becomes $(nh)^{-1}\sigma^2_{\mathtt{us}} + h^{2\mathfrak{K}}\tilde{\eta}^2_{\mathtt{us}}$. Then optimizing the MSE yields, in this notation,

$$h^*_{\mathtt{mse}} = n^{-\frac{1}{2\mathfrak{K}+1}}\left(\frac{\sigma^2_{\mathtt{us}}}{2\mathfrak{K}\tilde{\eta}^2_{\mathtt{us}}}\right)^{-\frac{1}{2\mathfrak{K}+1}}.$$

Therefore, the second term of $T_{\mathtt{us}}(h^*_{\mathtt{mse}})$ will be

$$\frac{\eta_{\mathtt{us}}}{\hat{\sigma}_{\mathtt{us}}} = \frac{\sqrt{nh^*_{\mathtt{mse}}}(h^*_{\mathtt{mse}})^{\mathfrak{K}}[\tilde{\eta}_{\mathtt{us}} + o(1)]}{\hat{\sigma}_{\mathtt{us}}} = \left(n(h^*_{\mathtt{mse}})^{2\mathfrak{K}+1}\right)^{1/2}\frac{\tilde{\eta}_{\mathtt{us}}}{\hat{\sigma}_{\mathtt{us}}} = \left(\frac{\sigma^2_{\mathtt{us}}}{2\mathfrak{K}\tilde{\eta}^2_{\mathtt{us}}}\right)^{1/2}\frac{\tilde{\eta}_{\mathtt{us}}}{\hat{\sigma}_{\mathtt{us}}} = \left(\frac{1}{2\mathfrak{K}}\right)^{1/2}[1+o_p(1)],$$

using consistency of $\hat{\sigma}_{\mathtt{us}}$ (or if a feasible $h^*_{\mathtt{mse}}$ is used, it is the bias estimate that must be consistent). Hence $T_{\mathtt{us}}(h^*_{\mathtt{mse}}) \to_d \mathcal{N}\left((2\mathscr{k})^{-1/2}, 1\right)$.

The most common empirical case would be $\mathscr{k} = 2$ and $\alpha = 0.05$, and so $T_{\mathtt{us}} \to_d \mathcal{N}(1/2, 1)$ and $\mathbb{P}[f \in I_{\mathtt{us}}(h^*_{\mathtt{mse}})] \approx 0.92$. ■

S.I.2.1 Bandwidth Choice: Rule-of-Thumb (ROT)

Motivated by the fact that estimating $\hat{H}_{\mathtt{dpi}}$ might be difficult in practice, while data-driven MSE-optimal bandwidth selectors are readily-available, the ROT bandwidth choice is to simply rescale any feasible MSE-optimal bandwidth $\hat{h}_{\mathtt{mse}}$ to yield optimal coverage error decay rates (but sub-optimal constants):

$$\hat{h}_{\mathtt{rot}} = \hat{h}_{\mathtt{mse}}\, n^{-(\mathscr{k}-2)/((1+2\mathscr{k})(\mathscr{k}+3))}.$$

When $\mathscr{k} = 2$, $\hat{h}_{\mathtt{rot}} = \hat{h}_{\mathtt{mse}}$, which is optimal (in rates) as discussed previously.

Remark S.I.2 (Integrated Coverage Error)**.** A closer analogue of the Silverman (1986) rule of thumb, which uses the integrated MSE, would be to integrate the coverage error over the point of evaluation x. For point estimation, this approach has some practical benefits. However, in the present setting note that $\int f^{(\mathscr{k})}(x)dx = 0$, removing the third term (of order $h^{\mathscr{k}}$) entirely and thus, for any given point x, yields a lower quality approximation. ■

S.I.2.2 Bandwidth Choice: Direct Plug-In (DPI)

To detail the direct plug-in (DPI) rule from the main text, it is useful to first simplify the problem. Recall from the main text that the optimal choice is $h^*_{\mathtt{rbc}} = H^*_{\mathtt{rbc}}(\rho) n^{-1/(\mathscr{k}+3)}$, where

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

With $\ell = 2$ and $\rho = 1$, and using the definitions of $q_k(M_1)$, $k = 1, 2, 3$, from the main text or Section S.I.1.2, this simplifies to:

$$\begin{aligned} H^*_{\mathtt{rbc}}(K, L, 1) = \arg\min_H \Bigg| & H^{-1} \left\{ \vartheta_{M,4} \frac{z^2 - 3}{6} - \vartheta^2_{M,3} \frac{z^4 - 4z^2 + 15}{9} \right\} \\ & - H^{1+2(\mathscr{k}+2)} \left\{ (f^{(\mathscr{k}+2)})^2 \left(\mu_{K,\mathscr{k}+2} - \mu_{K,\mathscr{k}}\mu_{L,2}\right)^2 \vartheta_{M,2} \right\} \\ & + H^{\mathscr{k}+2} \left\{ f^{(\mathscr{k}+2)} \left(\mu_{K,\mathscr{k}+2} - \mu_{K,\mathscr{k}}\mu_{L,2}\right) \vartheta_{M,3} \frac{2z^2}{3} \right\} \Bigg|, \end{aligned}$$

where $z = z_{\alpha/2}$ the appropriate upper quantile of the Normal distribution. However, $H^*_{\mathtt{rbc}}(\rho)$ still depends on the unknown density through $f^{(\mathscr{k}+2)}$.

Our recommendation is a DPI rule of order one, which uses a pilot bandwidth to estimate $f^{(\mathfrak{K}+2)}$ consistently. A simple and easy to implement choice is the MSE-optimal bandwidth appropriate to estimating $f^{(\mathfrak{K}+2)}$, say $h^*_{\mathfrak{K}+2,\mathtt{mse}}$, which is different from $h^*_{\mathtt{mse}}$ for the level of the function; see e.g., Wand and Jones (1995). Let us denote a feasible MSE-optimal pilot bandwidth by $\hat{h}_{\mathfrak{K}+2,\mathtt{mse}}$. Then we have:

$$\begin{aligned}\hat{H}_{\mathtt{dpi}}(K, L, 1) = \arg\min_H \Bigg| & H^{-1}\left\{\vartheta_{M,4}\frac{z^2-3}{6} - \vartheta_{M,3}^2\frac{z^4-4z^2+15}{9}\right\} \\ & - H^{1+2(\mathfrak{K}+2)}\left\{\hat{f}^{(\mathfrak{K}+2)}(x;\hat{h}_{\mathfrak{K}+2,\mathtt{mse}})^2\left(\mu_{K,\mathfrak{K}+2} - \mu_{K,\mathfrak{K}}\mu_{L,2}\right)^2\vartheta_{M,2}\right\} \\ & + H^{\mathfrak{K}+2}\left\{\hat{f}^{(\mathfrak{K}+2)}(x;\hat{h}_{\mathfrak{K}+2,\mathtt{mse}})\left(\mu_{K,\mathfrak{K}+2} - \mu_{K,\mathfrak{K}}\mu_{L,2}\right)\vartheta_{M,3}\frac{2z^2}{3}\right\}\Bigg|.\end{aligned}$$

This is now easily solved numerically (see note below). Further, if $\mathfrak{K} = 2$, the most common case in practice, and K and L are either the respective second order minimum variance or MSE-optimal kernels (Sections S.I.2.3 and S.I.4.2), then the above may be simplified to:

$$\begin{aligned}\hat{H}_{\mathtt{dpi}}(M, 1) = \arg\min_H \Bigg| & H^{-1}\left\{\vartheta_{M,4}\frac{z^2-3}{6} - \vartheta_{M,3}^2\frac{z^4-4z^2+15}{9}\right\} \\ & - H^{9}\left\{\hat{f}^{(4)}(x;\hat{h}_{\mathfrak{K}+2,\mathtt{mse}})^2\mu_{M,4}^2\vartheta_{M,2}\right\} \\ & + H^{4}\left\{\hat{f}^{(4)}(x;\hat{h}_{\mathfrak{K}+2,\mathtt{mse}})\mu_{M,4}\vartheta_{M,3}\frac{2z^2}{3}\right\}\Bigg|.\end{aligned}$$

Continuing with $\mathfrak{K} = 2$, a second option is a DPI rule of order zero, which uses a reference model to build the rule of thumb, more akin to Silverman (1986). Using the Normal distribution, so that $f(x) = \phi(x)$ and derivatives have known form, we obtain:

$$\begin{aligned}\hat{H}_{\mathtt{dpi}}(M, 1) = \arg\min_H \Bigg| & H^{-1}\left\{\vartheta_{M,4}\frac{z^2-3}{6} - \vartheta_{M,3}^2\frac{z^4-4z^2+15}{9}\right\} \\ & - H^{9}\left\{\left[\left(\tilde{x}^4 - 6\tilde{x}^2 + 3\right)\phi(\tilde{x})\right]^2\mu_{M,4}^2\vartheta_{M,2}\right\} \\ & + H^{4}\left\{\left(\tilde{x}^4 - 6\tilde{x}^2 + 3\right)\phi(\tilde{x})\mu_{M,4}\vartheta_{M,3}\frac{2z^2}{3}\right\}\Bigg|\end{aligned}$$

where $\tilde{x} = (x - \hat{\mu})/\hat{\sigma}_X$ is the point of interest centered and scaled.

Remark S.I.3 (Notes on computation)**.** When numerically solving the above minimization problems, computation will be greatly sped up by squaring the objective function. ■

S.I.2.3 Choice of ρ

First, we expand on the argument that ρ should be bounded and positive. Intuitively, the standard errors $\hat{\sigma}^2_{\mathtt{rbc}}$ control variance up to order $(nh)^{-1}$, while letting $b \to 0$ faster removes more bias. If b

vanishes too fast, the variance is no longer controlled. Setting $\bar{\rho} \in (0, \infty)$ balances these two. Let us simplify the discussion by taking $\ell = 2$, reflecting the widespread use of symmetric kernels. This does not affect the conclusions in any conceptual way, but considerably simplifies the notation. With this choice, Eqn. (S.I.1) yields the tidy expression

$$\eta_{\texttt{bc}} = \sqrt{nh} h^{\mathscr{k}+2} f^{(\mathscr{k}+2)} \left(\mu_{K,\mathscr{k}+2} - \rho^{-2} \mu_{K,\mathscr{k}} \mu_{L,2} \right) \{1 + o(1)\}.$$

Choice of ℓ and b (or ρ) cannot reduce the first term, which represents $\mathbb{E}[\hat{f}] - f - B_f$, and further, if $\bar{\rho} = \infty$, the bias rate is not improved, but the variance is inflated beyond order $(nh)^{-1}$. On the other hand, if $\bar{\rho} = 0$, then not only is a delicate choice of b needed, but $\ell > 2$ is required, else the second term above dominates $\eta_{\texttt{bc}}$, 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 that, on the contrary, when using our proposed Studentization, it is optimal that $\mathbb{E}[\hat{f}] - f - B_f$ is (part of) the dominant bias term. This reasoning is not an artifact of choosing $\mathscr{k}$ even and $\ell = 2$, but in other cases $\rho \to 0$ can be optimal if the convergence is sufficiently slow to equalize the two bias terms.

The following result which makes the above intuition precise.

Corollary S.I.1 (Robust bias correction: $\rho \to 0$). *Let the conditions of Theorem S.I.1(c) hold, with $\bar{\rho} = 0$, and fix $\ell = 2$ and $\mathscr{k} \leq S - 2$. Then*

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

By virtue of our new studentization, the leading variance remains order $(nh)^{-1}$ and the problematic correlation terms are absent, however by forcing $\rho \to 0$, the ρ^{-2} terms of $\eta_{\texttt{bc}}$ are dominant (the bias of $\hat{B}_f$), and in light of our results, unnecessarily inflated. This verifies that $\bar{\rho} = 0$ or ∞ will be suboptimal.

We thus restrict to bounded and positive, ρ. Therefore, ρ impacts only the shape of the "kernel" $M_\rho(u) = K(u) - \rho^{1+\mathscr{k}} L^{(\mathscr{k})}(\rho u) \mu_{K,\mathscr{k}}$, and hence the choice of ρ depends on what properties the user desires for the kernel. It happens that $\rho = 1$ has good theoretical properties and performs very well numerically (see Section S.I.8). As a result, from the practitioner's point of view, choice of ρ (or b) is completely automatic.

To see the optimality of $\rho = 1$, consider two cogent and well-studied possibilities: finding the kernel shape to minimize (i) interval length and (ii) MSE. The following optimal shapes are derived by Gasser et al. (1985) and references therein. Given the above results, we set $\mathscr{k} = 2$. Indeed, the optimality properties here do not extend to higher order kernels.

Minimizing interval length is (asymptotically) equivalent to finding the minimum variance fourth-order kernel, as $\sigma^2_{\texttt{rbc}} \to f\vartheta_{M,2}$. Perhaps surprisingly, choosing K and $L^{(2)}$ to be the second-order minimum variance kernels for estimating f and $f^{(2)}$ respectively, yields an $M_1(u)$ that is exactly the minimum variance kernel. The fourth order minimum variance kernel for estimating f is $K_{\texttt{mv}}(u) = (3/8)(-5u^2 + 3)$, which is identical to $M_1(u)$ when K is the uniform kernel and $L^{(2)} = (15/4)(3u^2 - 1)$, the minimum variance kernels for f and $f^{(2)}$ respectively.

The result is similar for minimizing MSE: choosing K and $L^{(2)}$ to be the MSE-optimal kernels for their respective point estimation problems yields an MSE-optimal $M_1(u)$. The optimal fourth order kernel is $K_{\texttt{mse}}(u) = (15/32)(7u^4 - 10u^2 + 3)$, and the respective second-order MSE optimal kernels are $K(u) = (3/4)(1 - u^2)$ and $L^{(2)}(u) = (105/16)(6u^2 - 5u^4 - 1)$. A practitioner might use the MSE-optimal kernels (along with $h^*_{\texttt{mse}}$) to obtain the best possible point estimate. Our results then give an accompanying measure of uncertainty that both has correct coverage and the attractive feature of using the same effective sample.

In Section S.I.4.2 we numerically compare several kernel shapes, focusing on: (i) interval length, measured by $\vartheta_{M,2}$, (ii) bias, given by $\tilde{\mu}_{M,4}$, and (iii) the associated MSE, given by $(\vartheta^8_{M,2}\tilde{\mu}^2_{M,4})^{1/9}$. These results, and the discussion above, give the foundations for our recommendation of $\rho = 1$, which delivers an easy-to-implement, fully automatic choice for implementing robust bias-correction that performs well numerically, as in Section S.I.8.

Remark S.I.4 (Coverage Error Optimal Kernels)**.** Our results hint at a third notion of optimal kernel shape: minimizing coverage error. This kernel, for a fixed order $\mathfrak{k}$, would minimize the constants in Corollary 1 of the main text. In that result, h is chosen to optimize the rate and the constant $H^*_{\texttt{us}}$ gives the minimum for a fixed kernel K. A step further would be to view $H^*_{\texttt{us}}$ as a function of K, and optimizing. To our knowledge, such a derivation has not been done and may be of interest. ■

S.I.3 Assumptions

The following assumptions are sufficient for our results. The first two are copied directly from the main text (see discussion there) and the third is the appropriate Cramér's condition.

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

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

It will cause no confusion (as the notations never occur in the same place), but in the course of proofs we will frequently write $s = \sqrt{nh}$.

Assumption S.I.3 (Cramér's Condition)**.** *For each $\xi > 0$ and all sufficiently small h*

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

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

Remark S.I.5 (Sufficient Conditions for Cramér's Condition)**.** Assumption S.I.3 is a high level condition, but one that is fairly mild. Hall (1991) provides a primitive condition for Assumption S.I.3 and Lemma 4.1 in that paper verifies that Assumption S.I.3 is implied. Hall (1992a) and Hall (1992b) assume the same primitive condition. This condition is as follows. On their compact support, assumed here to be $[-1,1]$, there exists a partition $-1 = a_0 < a_1 < \cdots < a_m = 1$, such that on each (a_{j-1}, a_j), K and M are differentiable, with bounded, strictly monotone derivatives.

This condition is met for many kernels, with perhaps the only exception of practical importance being the uniform kernel. As Hall (1991) describes, it is possible to prove the Edgeworth expansion for the uniform kernel using different methods than we use in below. The uniform kernel is also ruled out for local polynomial regression, see Section S.II.3. ■

S.I.4 Bias

This section accomplishes three things. First, we first carefully derive the bias of the initial estimator and the bias correction. Second, we explicate the properties of the induced kernel M_ρ in terms of bias reduction and how exactly this kernel is "higher-order". Finally, we examine two other methods of bias reduction: (i) estimating the derivatives without using derivatives of kernels (Singh, 1977), and (ii) the generalized jackknife approach (Schucany and Sommers, 1977). Further methods are discussed and compared by Jones and Signorini (1997). The message from both alternative methods echoes our main message: it is important to account for any bias correction when doing inference, i.e., to avoid the mismatch present in $T_{\mathtt{bc}}$.

S.I.4.1 Precise Bias Calculations

Recall that the biases of the two estimators are as follows:

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

and

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

The following Lemma gives a rigorous proof of these statements.

Lemma S.I.1. *Under Assumptions S.I.1 and S.I.2, Equations* (S.I.2) *and* (S.I.3) *hold.*

Proof. To show Eqn. (S.I.2), begin with the change of variables and the Taylor expansion

$$\begin{aligned} \mathbb{E}[\hat{f}] &= h^{-1} \int K(X_{h,i}) f(X_i) dX_i = \int K(u) f(x - uh) du \\ &= \sum_{k=0}^{S} \left\{ (-h)^k f^{(k)}(x) \int u^k K(u) du / k! \right\} + (-h)^S \int u^S K(u) \left(f^{(S)}(\bar{x}) - f^{(S)}(x) \right) du. \end{aligned}$$

where $\bar{x} \in [x, x - uh]$. By the Hölder condition of Assumption S.I.1, the final term is $O(h^{S+\varsigma})$. If $\Bbbk > S$, then all $\int u^k K(u) du = 0$, and only this remainder is left. In all other cases, $h^{\Bbbk} f^{(\Bbbk)}(x) \mu_{K,\Bbbk}$ is the first nonzero term of the summation, and hence the leading bias term. Further, by virtue of $\Bbbk$ being even and K symmetric, $\int u^{\Bbbk+1} K(u) du = 0$, leaving only $O(h^{S+\varsigma})$ when $\Bbbk = S - 1$, and otherwise, when $\Bbbk \leq S - 2$, leaving $h^{\Bbbk+2} f^{(\Bbbk+2)}(x) \mu_{K,\Bbbk+2} + o(h^{\Bbbk+2})$. This completes the proof of Eqn. (S.I.2).

To establish Eqn. (S.I.3), first write

$$\mathbb{E}[\hat{f} - \hat{B}_f] - f = \mathbb{E}[\hat{f} - f - B_f] + \mathbb{E}[B_f - \hat{B}_f],$$

where B_f follows the convention of being identically zero if $\Bbbk > S$. The first portion is characterized by rearranging Eqn. (S.I.2), so it remains to examine the second term. Let $\tilde{\Bbbk} = \Bbbk \vee S$. By repeated integration by parts, using the boundary conditions of Assumption S.I.2:

$$\begin{aligned} \mathbb{E}[\hat{f}^{(\Bbbk)}] &= \frac{1}{b^{1+\Bbbk}} \int L^{(\Bbbk)}(X_{b,i}) f(X_i) dX_i \\ &= -\frac{1}{b^{1+(\Bbbk-1)}} L^{(\Bbbk-1)}(X_{b,i}) f(X_i) \bigg|_{\mathcal{X}} + \frac{1}{b^{1+(\Bbbk-1)}} \int L^{(\Bbbk-1)}(X_{b,i}) f^{(1)}(X_i) dX_i \\ &= 0 + \frac{1}{b^{1+(\Bbbk-1)}} \int L^{(\Bbbk-1)}(X_{b,i}) f^{(1)}(X_i) dX_i \\ &= -\frac{1}{b^{1+(\Bbbk-2)}} L^{(\Bbbk-2)}(X_{b,i}) f^{(1)}(X_i) + \frac{1}{b^{1+(\Bbbk-2)}} \int L^{(\Bbbk-2)}(X_{b,i}) f^{(2)}(X_i) dX_i \\ &\ \ \vdots \end{aligned}$$

$$= \frac{1}{b^{1+(\hslash-\tilde{\hslash})}} \int L^{(\hslash-\tilde{\hslash})}(X_{b,i}) f^{(\tilde{\hslash})}(X_i) dX_i$$
$$= \frac{1}{b^{\hslash-\tilde{\hslash}}} \int L^{(\hslash-\tilde{\hslash})}(u) f^{(\tilde{\hslash})}(x-ub) du,$$

where the last line follows by a change of variables. We now proceed separately for each case delineated in (S.I.3), from top to bottom. For $\hslash > S$, no reduction is possible, and the final line above is $O(b^{S-\hslash})$, and with $B_f = 0$, we have $\mathbb{E}[B_f - \hat{B}_f] = 0 - h^{\hslash}\mu_{K,\hslash}\mathbb{E}[\hat{f}^{(\hslash)}] = O(h^{\hslash} b^{S-\hslash})$, as shown. For $\hslash \leq S$, by a Taylor expansion, the final line displayed above becomes

$$\sum_{k=\hslash}^{S} \left\{ b^{k-\hslash} f^{(k)}(x) \mu_{L,k-\hslash} \right\} + b^{S-\hslash} \int u^{S-\hslash} L(u) \left(f^{(S)}(\bar{x}) - f^{(S)}(x) \right) du.$$

The second term above is $O(b^{S-\hslash+\varsigma})$ in all cases, and $\mu_{L,0} = 1$, which yields $\mathbb{E}[\hat{f}^{(\hslash)}] = f^{(\hslash)} + O(b^{S-\hslash+\varsigma})$ for $\hslash \in \{S-1, S\}$, using $\mu_{L,1} = 0$ in the former case. Next, if $\hslash + \ell \leq S$, the above becomes $\mathbb{E}[\hat{f}^{(\hslash)}] = f^{(\hslash)} + b^{\ell} f^{(\hslash+\ell)} \mu_{L,\ell} + o(b^{\ell})$, as $\mu_{L,k} = 0$ for $1 < k < \ell$, whereas if $\hslash + \ell > S$, the remainder terms can not be characterized, leaving $\mathbb{E}[\hat{f}^{(\hslash)}] = f^{(\hslash)} + O(b^{S-\hslash+\varsigma})$. Plugging any of these results into $\mathbb{E}[B_f - \hat{B}_f] = h^{\hslash}\mu_{K,\hslash}(f^{(\hslash)} - \mathbb{E}[\hat{f}^{(\hslash)}])$ completes the demonstration of Eqn. (S.I.3). □

S.I.4.2 Properties of the kernel $M_\rho(\cdot)$

As made precise below, M_ρ is a higher-order kernel. The choices of K, L, and ρ determine the shape of M_ρ, which in turn effects the variance and bias constants. In standard kernel analyses, these constants are used to determine optimal kernel shapes for certain problems (see Gasser et al. (1985) and references therein). For several choices of K, L, and ρ, Table S.I.1 shows numerical results for the various constants of the induced kernel M_ρ. The table includes (i) the variance, given by $\vartheta_{M,2}$ and relevant for interval length, (ii) a measure of bias given by $\tilde{\mu}_{M,4}$, and finally (iii) the resulting mean square error constant, $[\vartheta_{M,2}^8 \tilde{\mu}_{M,4}^2]^{1/9}$ ($\tilde{\mu}_{M,4} = (k!)(-1)^k \mu_{M,4}$). These specific constants are due to M_ρ being a fourth order kernel, as discussed next, and would otherwise remain conceptually the same but rely on different moments. A more general, but more cumbersome procedure would be to choose ρ numerically to minimize some notation of distance (e.g., L_2) between the resulting kernel M_ρ and the optimal kernel shape already available in the literature. However, using $\rho = 1$ as a simple rule-of-thumb exhibits very little lost performance, as shown in the Table and discussed in the paper.

It is worthwhile to make precise the sense in which the n-varying "kernel" $M_\rho(\cdot)$ of Eqn. (S.I.1) is a higher-order kernel. Comparing Equations (S.I.2) and (S.I.3) shows exactly what is meant by this statement: the bias rate attained agrees with a standard estimate using a kernel of order $\hslash + 2$ (if $\bar{\rho} > 0$), as $\ell \geq 2$. For example, if $\hslash = \ell = 2$ and $\bar{\rho} > 0$, then $M_{\bar{\rho}}(\cdot)$ behaves as a fourth-order kernel in terms of bias reduction.

However, it is not true in general that $M(\cdot)$ is a higher-order kernel in the sense that its moments

Table S.I.1: Numerical results for bias and variance constants of the induced higher-order kernel M for several choices of K, L, and ρ

		$\rho = 0.5$			$\rho = 1$			$\rho = 1.5$		
Kernel K	Kernel $L^{(2)}$	$\tilde{\mu}_{M,4}$	$\vartheta_{M,2}$	MSE	$\tilde{\mu}_{M,4}$	$\vartheta_{M,2}$	MSE	$\tilde{\mu}_{M,4}$	$\vartheta_{M,2}$	MSE
Epanechnikov	$(105/16)(6u^2 - 5u^4 - 1)$	0.0690	0.6430	0.3729	-0.0476	1.2500	0.6199	-0.3643	5.5992	3.6944
Uniform	$(105/16)(6u^2 - 5u^4 - 1)$	0.1722	0.5152	0.3752	-0.0222	1.4722	0.6052	-0.5500	11.5742	7.7202
Biweight	$(105/16)(6u^2 - 5u^4 - 1)$	0.0357	0.7617	0.3744	-0.0476	1.2500	0.6199	-0.2738	3.9537	2.5448
Triweight	$(105/16)(6u^2 - 5u^4 - 1)$	0.0210	0.8617	0.3715	-0.0438	1.2774	0.6202	-0.2197	3.2395	2.0300
Tricube	$(105/16)(6u^2 - 5u^4 - 1)$	0.0335	0.7542	0.3658	-0.0506	1.2332	0.6207	-0.2786	3.9344	2.5436
Cosine	$(105/16)(6u^2 - 5u^4 - 1)$	0.0629	0.6617	0.3747	-0.0476	1.2503	0.6199	-0.3475	5.2717	3.4651
Epanechnikov	$(15/4)(3u^2 - 1)$	0.0643	0.6410	0.3660	-0.0857	1.1250	0.6432	-0.4929	4.1754	3.0440
Uniform	$(15/4)(3u^2 - 1)$	0.1643	0.5098	0.3678	-0.0857	1.1250	0.6432	-0.7643	7.6191	5.7276
Biweight	$(15/4)(3u^2 - 1)$	0.0323	0.7543	0.3630	-0.0748	1.1352	0.6291	-0.3656	3.0550	2.1579
Triweight	$(15/4)(3u^2 - 1)$	0.0184	0.8517	0.3568	-0.0649	1.1631	0.6229	-0.2911	2.5444	1.7435
Tricube	$(15/4)(3u^2 - 1)$	0.0300	0.7487	0.3547	-0.0780	1.1319	0.6333	-0.3712	3.0729	2.1764
Cosine	$(15/4)(3u^2 - 1)$	0.0584	0.6583	0.3669	-0.0836	1.1254	0.6399	-0.4693	3.9510	2.8668
Biweight	Biweight$^{(2)}$	0.0323	0.7543	0.3630	-0.0748	1.1352	0.6291	-0.3656	3.0550	2.1579
Tricube	Tricube$^{(2)}$	0.0299	0.7516	0.3556	-0.0790	1.1993	0.6687	-0.3746	3.7063	2.5762
Gaussian	Gaussian$^{(2)}$	2.2500	0.3006	0.4113	-3.0000	0.4760	0.6599	-17.2500	1.3606	2.4758

[1] As discussed in Section S.I.4.2, M_ρ behaves as a fourth order kernel in terms of bias reduction, but does not strictly fit within the class of kernels used in derivation of optimal kernel shapes. This explains the super-optimal behavior exhibited by some choices of K, L, and ρ.

[2] The constants $\tilde{\mu}_{M,4}$ and $\vartheta_{M,2}$ measure bias and variance, respectively (the latter also being relevant for interval length). The MSE is measured by $[\vartheta_{M,2}^8 \tilde{\mu}_{M,4}^2]^{1/9}$, owing to M_ρ being a fourth-order kernel.

below $\mathfrak{k}+2$ are zero. That is, for any $k<\mathfrak{k}$, by the change of variables $w=\rho u$,

$$\begin{aligned}\int_{-1}^{1} u^k M(u)du &= \int_{-1}^{1} u^k K(u)du - \rho^{1+\mathfrak{k}}\mu_{K,\mathfrak{k}}\int_{-1}^{1} u^k L^{(\mathfrak{k})}(\rho u)du \\ &= 0 - \rho^{1+\mathfrak{k}}\mu_{K,\mathfrak{k}}\rho^{-1-k}\int_{-\rho}^{\rho} w^k L^{(\mathfrak{k})}(w)du \\ &= 0 - \rho^{\mathfrak{k}-k}\mu_{K,\mathfrak{k}}\int_{-\rho}^{\rho} w^k L^{(\mathfrak{k})}(w)du.\end{aligned}$$

Now, $L(u)=L(-u)$ implies that $L^{(k)}(u)=(-1)^k L^{(k)}(-u)$. Since $\mathfrak{k}$ is even, $L^{(\mathfrak{k})}(w)$ is symmetric, therefore if k is odd $0=\int_{-\rho}^{\rho} w^k L^{(\mathfrak{k})}(w)du$ for any ρ. But this fails for k even, even for $\rho=1$, and hence $\int_{-1}^{1} u^k M(u)du \neq 0$. For example, in the leading case of $\mathfrak{k}=\ell=2$, $\int_{-1}^{1} u^2 M(u)du \neq 0$ in general, and so $M(\cdot)$ is not a fourth-order kernel in the traditional sense.

Instead, the bias reduction is achieved differently. The proof of Lemma S.I.1 makes explicit use of the structure imposed by estimating $f^{(\mathfrak{k})}$ using the *derivative* of the kernel $L(\cdot)$. From a technical standpoint, an integration by parts argument shows how the properties of the kernel $L(\cdot)$ (not the function $L^{(\mathfrak{k})}(\cdot)$) are used to reduce bias. This argument *precedes* the Taylor expansion of f, and thus moments of M are never encountered and there is no requirement that they be zero. This approach is simple, intuitive, and leads to natural restrictions on the kernel L, and for this reason it is commonly employed in the literature and in practice (Hall, 1992b).

S.I.4.3 Other Bias Reduction Methods

We now examine two other methods of bias reduction: (i) estimating the derivatives without using derivatives of kernels (Singh, 1977), and (ii) the generalized jackknife approach (Schucany and Sommers, 1977). Further methods are discussed and compared by Jones and Signorini (1997). Both methods are shown to be tightly connected to our results. Further, a more general message is that it is important to account for any bias correction when doing inference, i.e., to avoid the mismatch present in $T_{\texttt{bc}}$.

The first method, which dates at least to Singh (1977), is to introduce a class of kernel functions directly for derivative estimation, more closely following the standard notion of a higher-order kernel rather than using the derivative of a kernel to estimate the density derivative and proving bias reduction via integration by parts. Jones (1994) expands on this method and gives further references. This class of kernels is used in the derivation of optimal kernel shapes (for derivative estimation) by Gasser et al. (1985). It is worthwhile to show how this class of kernel achieves bias correction and how this approach fits into our Edgeworth expansions.

Consider estimating $f^{(\mathfrak{k})}$ with

$$\tilde{f}^{(\mathfrak{k})}(x) = \frac{1}{nb^{1+\mathfrak{k}}}\sum_{i=1}^{n} J\left(X_{b,i}\right),$$

for some kernel function $J(\cdot)$. Note well that J is generic, it need not itself be a derivative, but this is the only difference here. A direct Taylor expansion (i.e. without first integrating by parts) then gives

$$\mathbb{E}[\tilde{f}^{(\mathfrak{K})}] = b^{-\mathfrak{K}} \sum_{k=0}^{S} b^k \mu_{J,k} f^{(k)} + O(b^{S+\varsigma}).$$

Thus, if J satisfies $\mu_{J,k} = 0$ for $k = 0, 1, \ldots, \mathfrak{K}-1, \mathfrak{K}+1, \mathfrak{K}+2, \ldots, \mathfrak{K}+(\ell-1)$, $\mu_{J,\mathfrak{K}} = 1$, and $\mu_{J,\mathfrak{K}+\ell} \neq 0$, and S is large enough then

$$\mathbb{E}[\tilde{f}^{(\mathfrak{K})}] = f^{(\mathfrak{K})} + b^{\ell} f^{(\mathfrak{K}+\ell)} \mu_{J,\mathfrak{K}+\ell} + o(b^{\ell}),$$

just as achieved by $\hat{f}^{(\mathfrak{K})}$ and exactly matching Eqn. (S.I.2). Note that $\mu_{J,0} = 0$, that is, the kernel J does not integrate to one. In the language of Gasser et al. (1985), J is a kernel of order $(\mathfrak{K}, \mathfrak{K}+\ell)$.

Given this result, bias correction can of course be performed using $\tilde{f}^{(\mathfrak{K})}(x)$ (based on J) rather than $\hat{f}^{(\mathfrak{K})}$ (based on $L^{(\mathfrak{K})}$). Much will be the same: the structure of Eqn. (S.I.1) will hold with J in place of $L^{(\mathfrak{K})}$ and the results in Eqn. (S.I.3) are achieved with modifications to the constants (e.g., in the first line, $\mu_{J,\mathfrak{K}+\ell}$ appears in place of $\mu_{L,\ell}$). In either case, the same bias rates are attained. Our Edgeworth expansions will hold for this class under the obvious modifications to the notation and assumptions, and all the same conclusions are obtained.

When studying optimal kernel shapes, Gasser et al. (1985) actually further restrict the class, by placing a limit on the number of sign changes over the support of the kernel, which ensures that the MSE and variance minimization problems have well-defined solutions. Collectively, these differences in the kernel classes explain why it is possible to demonstrate "super-optimal" MSE and variance performance for certain choices of K, $L^{(\mathfrak{K})}$, and ρ, as in Table S.I.1.

A second alternative is the generalized jackknife method of Schucany and Sommers (1977), and expanded upon by Jones and Foster (1993). To simplify the notation and ease exposition, we describe this approach for second order kernels ($\mathfrak{K} = 2$), but the method, and all the conclusions below, generalize fully. We thank an anonymous reviewer for encouraging us to include these details.

Begin with two estimators $\hat{f}_1$ and $\hat{f}_2$, with (possibly different) bandwidths and second-order kernels h_j and K_j, $j = 1, 2$; thus Eqn. (S.I.2) gives

$$\mathbb{E}[\hat{f}_j] - f(x) = h_j^2 f^{(2)} \mu_{K_j,2} + o(h_j^2), \qquad j = 1, 2.$$

Schucany and Sommers (1977) propose to estimate f with $\hat{f}_{\mathtt{GJ},R} := (\hat{f}_1 - R\hat{f}_2)/(1-R)$, the bias of which is

$$\mathbb{E}[\hat{f}_{\mathtt{GJ},R} - f] = \frac{f^{(2)}}{1-R}\left(h_1^2 \mu_{K_1,2} - R h_2^2 \mu_{K_2,2}\right) + o(h_1^2 + h_2^2).$$

Hence, setting $R = (h_1^2 \mu_{K_1,2})/(h_2^2 \mu_{K_2,2})$ renders the leading bias exactly zero. Moreover, if $S \geq 4$, $\hat{f}_{\mathtt{GJ},R}$ has bias $O(h_1^4 + h_2^4)$; behaving as a single estimator with $\mathfrak{K} = 4$. To put this in context of our results, observe that with this choice of R, if we let $\tilde{\rho} = h_1/h_2$, then

$$\hat{f}_{\texttt{GJ},R} = \frac{1}{nh_1}\sum_{i=1}^{n} \tilde{M}\left(\frac{X_i - x}{h_1}\right), \qquad M(u) = K_1(u) - \tilde{\rho}^{1+2}\left\{\frac{K_2(\tilde{\rho}u) - \tilde{\rho}^{-1}K_1(u)}{\mu_{K_2,2}(1-R)}\right\}\mu_{K_1,2},$$

exactly matching Eqn. (S.I.1). Or equivalently, $\hat{f}_{\texttt{GJ},R} = \hat{f}_1 - h_1^2\tilde{f}^{(2)}\mu_{K_1,2}$, for the derivative estimator

$$\tilde{f}^{(2)} = \frac{1}{nh_2^{1+2}}\sum_{i=1}^{n} \tilde{L}\left(\frac{X_i - x}{h_2}\right), \qquad \tilde{L}(u) = \frac{K_2(u) - \tilde{\rho}^{-1}K_1(\tilde{\rho}^{-1}u)}{\mu_{K_2,2}(1-R)}.$$

Therefore, we can view $\hat{f}_{\texttt{GJ},R}$ as a change in the kernel $M(\cdot)$ or an explicit bias estimation described directly above with a specific choice of $J(\cdot)$ (depending on $\tilde{\rho}$ in either case). Again, Eqn. (S.I.1) holds exactly. Thus, our results cover the generalized jackknife method as well, and the same lessons apply.

Finally, we note that these bias correction methods can be applied to nonparametric regression as well, and local polynomial regression in particular, and that the same conclusions are found. We will not repeat this discussion however.

S.I.5 First Order Properties

Here we briefly state the first-order properties of $T_{\texttt{us}}$, $T_{\texttt{bc}}$, and $T_{\texttt{rbc}}$, using the common notation $T_{v,w}$ defined in Section S.I.1. Recall that $\eta_v = \sqrt{nh}(\mathbb{E}[\hat{f}_v] - f)$ is the scaled bias in either case. With this notation, we have the following result.

Lemma S.I.2. *Let Assumptions S.I.1 and S.I.2 hold. Then if $nh \to \infty$, $\eta_v \to 0$, and if $v = 2$, $\rho \to 0 + \bar{\rho}\mathbb{1}\{v = w\} < \infty$, it holds that $T_{v,w} \to_d \mathcal{N}(0,1)$.*

The conditions on h and b behind the generic assumption that the scaled bias vanishes can be read off of (S.I.2) and (S.I.3): $T_{\texttt{us}}$ requires $\sqrt{nh}h^{\mathfrak{k}} \to 0$ whereas $T_{\texttt{bc}}$ and $T_{\texttt{rbc}}$ require only $\sqrt{nh}h^{\mathfrak{k}}(h^2 \vee b^{\ell}) \to 0$, and thus accommodate $\sqrt{nh}h^{\mathfrak{k}} \not\to 0$ or $b \not\to 0$ (but not both). However, bias correction requires a choice of $\rho = h/b$. One easily finds that $\mathbb{V}[\sqrt{nh}\hat{B}_f] = O(\rho^{1+2\mathfrak{k}})$, whence $\rho \to 0$ is required for $T_{\texttt{bc}}$. But $T_{\texttt{rbc}}$ does not suffer from this requirement because of our proposed, new Studentization. From a first-order point of view, traditional bias correction allows for a larger class of sequences h, but requires a delicate choice of ρ (or b), and Hall (1992b) shows that this constraint prevents $T_{\texttt{bc}}$ from improving inference. Our novel standard errors remove these constraints, allowing for improvements in bias to carry over to improvements in inference. The fact that a wider range of bandwidths is allowed hints at the robustness to tuning parameter choice discussed above and formalized by our Edgeworth expansions.

Remark S.I.6 ($\rho \to \infty$). $T_{\texttt{rbc}} \to_d \mathcal{N}(0,1)$ will hold even for $\bar{\rho} = \infty$, under the even weaker bias rate restriction that $\eta_{\texttt{bc}} = o(\rho^{1/2+\mathfrak{k}})$, provided $nb \to \infty$. In this case $\hat{B}_f$ dominates the first-order approximation, but $\sigma^2_{\texttt{rbc}}$ still accounts for the total variability. However there is no gain for inference: the bias properties can not be improved due to the second bias term ($\mathbb{E}[\hat{f}] - f - B_f$), while variance can only be inflated. Thus, we restrict to bounded $\bar{\rho}$. Section S.I.2.3 has more discussion on the choice of ρ. ■

S.I.6 Main Result: Edgeworth Expansion

Recall the generic notation:

$$T_{v,w} := \frac{\sqrt{nh}(\hat{f}_v - f)}{\hat{\sigma}_w},$$

for $1 \leq w \leq v \leq 2$. The Edgeworth expansion for the distribution of $T_{v,w}$ will consist of polynomials with coefficients that depend on moments of the kernel(s). Additional polynomials are needed beyond those used in the main text for coverage error. These are:

$$p^{(1)}_{v,w}(z) = \phi(z)\sigma_w^{-3}[\nu_{v,w}(1,1,2)z^2/2 - \nu_v(3)(z^2-1)/6],$$
$$p^{(2)}_{v,w}(z) = -\phi(z)\sigma_w^{-3}\mathbb{E}[\hat{f}_w]\nu_{v,w}(1,1,1)z^2, \qquad \text{and} \qquad p^{(3)}_{v,w}(z) = \phi(z)\sigma_w^{-1}.$$

The polynomials $p^{(k)}_{v,w}$ are even, and hence cancel out of coverage probability expansions, but are used in the expansion of the distribution function itself (or equivalently, the coverage of a one-sided confidence interval).

Next, recall from the main text the polynomials used in *coverage error* expansions:

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

The corresponding polynomials for expansions of the *distribution function* are

$$q^{(k)}_{v,w}(z) = \frac{1}{2}\frac{\phi(z)}{f}q_k(z;N_w), \qquad k = 1,2,3.$$

As before, the $q^{(k)}_{v,w}$ are odd and hence do not cancel when computing coverage: the $q_k(z;N_w)$ in the main text are doubled for just this reason.

Note that, despite the notation, $q^{(k)}_{v,w}(z)$ depends only on the "denominator" kernel N_w. The notation comes from the fact that when first computed, the terms which enter into the $q^{(k)}_{v,w}(z)$ depend on both kernels, but the simplifications in Eqn. (S.I.8) reduce the dependence to N_w. This is because for undersmoothing and robust bias correction, $v = w$, and for traditional bias correction $N_2 = M = K + o(1) = N_1 + o(1)$, as $\rho \to 0$ is assumed. Thus, when computing $\vartheta_{M,q}$ the terms with the lowest powers of ρ will be retained. These can be found by expanding

$$\vartheta_{M,q} = \int \left(K(u) - \rho^{1+\Bbbk}\mu_{K,\Bbbk}L^{(\Bbbk)}(u)\right)^q du = \sum_{j=0}^{q}\binom{q}{j}\left(-\mu_{K,\Bbbk}\rho^{1+\Bbbk}\right)^{q-j}\int K(u)^j L^{(\Bbbk)}(\rho u)^{q-j}du,$$

and hence we can write $\vartheta_{M,q} = \vartheta_{K,q} - \rho^{1+\Bbbk} q \mu_{K,\Bbbk} L^{(\Bbbk)}(0) \vartheta_{K,q-1} + O(h + \rho^{2+\Bbbk})$. We can thus write $q_j(z; M) = q_j(z; K) + o(1)$ in this case. If the expansions were carried out beyond terms of order $(nh)^{-1} + (nh)^{-1/2}\eta_v + \eta_v^2 + \mathbb{1}\{v \neq w\}\rho^{1+2\Bbbk}$ this would not be the case.

Finally, for traditional bias correction, there are additional terms in the expansion (see discussion in the main text) representing the covariance of $\hat{f}$ and $\hat{B}_f$ (denoted by Ω_1) and the variance of $\hat{B}_f$ (Ω_2). We now state their precise forms. These arise from the mismatch between the variance of the numerator of $T_{\mathtt{bc}}$ and the standardization used, $\sigma^2_{\mathtt{us}}$, that is $\sigma^2_{\mathtt{rbc}}/\sigma^2_{\mathtt{us}}$ is given by

$$\frac{nh\mathbb{V}[\hat{f} - \hat{B}_f]}{nh\mathbb{V}[\hat{f}]} = \frac{nh\mathbb{V}[\hat{f}] - 2nh\mathbb{C}[\hat{f}, \hat{B}_f] + nh\mathbb{V}[\hat{B}_f]}{nh\mathbb{V}[\hat{f}]} = 1 - 2\frac{nh\mathbb{C}[\hat{f}, \hat{B}_f]}{nh\mathbb{V}[\hat{f}]} + \frac{nh\mathbb{V}[\hat{B}_f]}{nh\mathbb{V}[\hat{f}]}.$$

This makes clear that Ω_1 and Ω_2 are the constant portions of the last two terms. We have

$$-2\frac{nh\mathbb{C}[\hat{f}, \hat{B}_f]}{nh\mathbb{V}[\hat{f}]} = \rho^{1+\Bbbk}\Omega_1,$$

where

$$\Omega_1 = -2\frac{\mu_{K,\Bbbk}}{\nu_1(2)}\left\{\int f(x-uh)K(u)L^{(\Bbbk)}(u\rho)du - b\int f(x-uh)K(u)du \int f(x-ub)L^{(\Bbbk)}(u)du\right\}.$$

Note $\nu_1(2) = \sigma^2_{\mathtt{us}}$. Turning to Ω_2, using the calculations in Section S.I.4.1 (recall $\tilde{\Bbbk} = \Bbbk \vee S$), we find that

$$\frac{nh\mathbb{V}[\hat{B}_f]}{nh\mathbb{V}[\hat{f}]} = \rho^{1+2\Bbbk}\Omega_2 \quad \text{where} \quad \Omega_2 = \frac{\mu^2_{K,\Bbbk}}{\nu_1(2)}\left\{\int f(x-ub)L^{(\Bbbk)}(u)^2 du - b^{1+2\tilde{\Bbbk}}\left(\int L^{(\Bbbk-\tilde{\Bbbk})}(u) f^{(\tilde{\Bbbk})}(x-ub)du\right)^2\right\}.$$

Fully simplifying would yield

$$\Omega_2 = \mu^2_{K,\Bbbk}\vartheta^{-2}_{K,2}\vartheta_{L^{(\Bbbk)},2},$$

which can be used in Theorem S.I.1.

As a last piece of notation, define the scaled bias as $\eta_v = \sqrt{nh}(\mathbb{E}[\hat{f}_v] - f)$.

We can now state our generic Edgeworth expansion, from whence the coverage probability expansion results follow immediately.

Theorem S.I.1. *Suppose Assumptions S.I.1, S.I.2, and S.I.3 hold, $nh/\log(n) \to \infty$, $\eta_v \to 0$, and if $v = 2$, $\rho \to 0 + \bar{\rho}\mathbb{1}\{v = w\}$. Then for*

$$F_{v,w}(z) = \Phi(z) + \frac{1}{\sqrt{nh}}p^{(1)}_{v,w}(z) + \sqrt{\frac{h}{n}}p^{(2)}_{v,w}(z) + \eta_v p^{(3)}_{v,w}(z) + \frac{1}{nh}q^{(1)}_{v,w}(z) + \eta_v^2 q^{(2)}_{v,w}(z) + \frac{\eta_v}{\sqrt{nh}}q^{(3)}_{v,w}(z)$$
$$- \mathbb{1}\{v \neq w\}\rho^{1+\Bbbk}(\Omega_1 + \rho^{\Bbbk}\Omega_2)\frac{\phi(z)}{2}z,$$

we have

$$\sup_{z\in\mathbb{R}} |\mathbb{P}[T_{v,w} < z] - F_{v,w}(z)| = o\left((nh)^{-1} + (nh)^{-1/2}\eta_v + \eta_v^2 + \mathbb{1}\{v \neq w\}\rho^{1+2\mathscr{k}}\right).$$

To use this result to find the expansion of the error in coverage probability of the Normal-based confidence interval, the function $F_{v,w}(z)$ is simply evaluated at the two endpoints of the interval. (Note: if the confidence interval were instead constructed with the bootstrap, a few additional steps are needed, but these do not alter any conclusions or results outside of constant terms.)

S.I.6.1 Undersmoothing vs. Bias-Correction Exhausting all Smoothness

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

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

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

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

This discussion is summarized by the following result:

Corollary S.I.2. *Let the conditions of Theorem S.I.1 hold.*

(a) If $\mathfrak{K} > S$, then

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

(b) If $\mathfrak{K} \geq S - 1$, then

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

S.I.6.2 Multivariate Densities and Derivative Estimation

We now briefly present state analogues of our results, both for distributional convergence and Edgeworth expansions, that cover multivariate data and derivative estimation. The conceptual discussion and implications are similar to those in the main text, once adjusted notationally to the present setting, and are hence omitted.

For a nonnegative integral d-vector q we adopt the notation that: (i) $[q] = q_1 + \cdots + q_d$, (ii) $g^{(q)}(x) = \partial^{[q]} g(x)/(\partial^{q_1} x_1 \cdots \partial^{q_d} x_d)$, (iii) $k! = q_1! \cdots q_d!$, and (iv) $\sum_{[q]=Q}$ for some integer $Q \geq 0$ denotes the sum over all indexes in the set $\{q : [q] = Q\}$.

The parameter of interest is $f^{(q)}(x)$, for $x \in \mathbb{R}^d$ and $[q] \leq S$. The estimator is

$$\hat{f}^{(q)}(x) = \frac{1}{nh^{d+[q]}} \sum_{i=1}^{n} K^{(q)}\left(X_{h,i}\right).$$

Note that here, and below for bias correction, we use a constant, diagonal bandwidth matrix, e.g. $h \times I_d$. This is for simplicity and comparability, and could be relaxed at notational expense.

The bias, for a given kernel of order $\mathfrak{K} \leq S - [q]$ (we restrict attention to the case where S is large enough), is

$$h^{\mathfrak{K}} \sum_{k:[k+q]=\mathfrak{K}} \mu_{K,k} f^{(q+k)}(x) + o(h^{\mathfrak{K}}),$$

exactly mirroring Eqn. (S.I.2), where now $\mu_{K,k}$ represents a d-dimensional integral. Bias estimation is straightforward, relying on estimates $\hat{f}^{(q+k)}(x)$, for all $[k] = \mathfrak{K} - [q]$. The form of $\hat{f}_2^{(q)}(x) = \hat{f}^{(q)}(x) - \hat{B}_{f^{(q)}}(x)$ is now given by

$$\hat{f}_2^{(q)}(x) = \frac{1}{nh^{d+[q]}} \sum_{i=1}^{n} M_{(q)}\left(X_{h,i}\right) \quad \text{where} \quad M_{(q)}(u) = K^{(q)}(u) - (\rho)^{d+[q]+\mathfrak{K}} \sum_{[k]=\mathfrak{K}} \mu_{K,k} L^{(q+k)}(u),$$

exactly analogous to Eqn. (S.I.1).

With these changes in notation out of the way, we can (re-)define the generic framework for both estimators exactly as above. Dropping the point of evaluation x, for $v \in \{1, 2\}$, define the estimator as

$$\hat{f}_v^{(q)} = \frac{1}{nh^{d+[q]}} \sum_{i=1}^{n} N_v\left(X_{h,i}\right), \qquad \text{where} \qquad N_1(u) = K^{(q)}(u) \text{ and } N_2(u) = M_{(q)}(u);$$

the variance

$$\sigma_v^2 := nh^{d+[q]}\mathbb{V}[\hat{f}_v^{(q)}] = \frac{1}{h^d}\left\{\mathbb{E}\left[N_v\left(X_{h,i}\right)^2\right] - \mathbb{E}\left[N_v\left(X_{h,i}\right)\right]^2\right\}$$

and its estimator as

$$\hat{\sigma}_v^2 = \frac{1}{h^d}\left\{\frac{1}{n}\sum_{i=1}^{n}\left[N_v\left(X_{h,i}\right)^2\right] - \left[\frac{1}{n}\sum_{i=1}^{n} N_v\left(X_{h,i}\right)\right]^2\right\};$$

and the t-statistics, for $1 \le w \le v \le 2$, as,

$$T_{v,w} := \frac{\sqrt{nh^{d+2[q]}}\left(\hat{f}_v^{(q)} - f^{(q)}\right)}{\hat{\sigma}_w}.$$

As before, $T_{\texttt{us}} = T_{1,1}$, $T_{\texttt{bc}} = T_{2,1}$, and $T_{\texttt{rbc}} = T_{2,2}$.

The scaled bias η_v has the same general definition as well: the bias of the numerator of the $T_{v,w}$. In this case, given by

$$\eta_v = \sqrt{nh^{d+2[q]}}\left(\mathbb{E}\left[\hat{f}_v^{(q)}\right] - f^{(q)}(x)\right).$$

The asymptotic order of η_v for different settings can be obtained straightforwardly via the obvious multivariate extensions of Equation (S.I.3) and the corresponding conclusion of Lemma S.I.1.

First-order convergence is now given by the following result. the proof of which is standard.

Lemma S.I.3. *Suppose appropriate multivariate versions of Assumptions S.I.1 and S.I.2 hold,* $nh^{d+2[q]} \to \infty$, $\eta_v \to 0$, *and if* $v = 2$, $\rho \to 0 + \bar{\rho}\mathbb{1}\{v = w\}$. *Then* $T_{v,w} \to_d \mathcal{N}(0,1)$.

For the Edgeworth expansion, redefine

$$\nu_{v,w}(j,k,p) = \frac{1}{h^{d+[q]\mathbb{1}\{j+pk=1\}}}\mathbb{E}\left[\left(N_v(u_i) - \mathbb{E}[N_v(u_i)]\right)^j\left(N_w(u_i)^p - \mathbb{E}[N_v(u_i)^p]\right)^k\right],$$

where $u_i = (x - X_i)/h$. The polynomials $p_{v,w}^{(k)}(z)$ and $q_{v,w}^{(k)}(z)$ are as given above, but using multivariate moments. The analogue of Theorem S.I.1 is given by the following result, which can be proven following the same steps as in Section S.I.7.

Theorem S.I.2. *Suppose appropriate multivariate versions of Assumptions S.I.1, S.I.2, and S.I.3 hold,* $nh^{d+2[q]}/\log(n) \to \infty$, $\eta_v \to 0$, *and if* $v = 2$, $\rho \to 0 + \bar{\rho}\mathbb{1}\{v = w\}$. *Then for*

$$F_{v,w}(z) = \Phi(z) + \frac{1}{\sqrt{nh^d}}p_{v,w}^{(1)}(z) + \sqrt{\frac{h^{d+2[q]}}{n}}p_{v,w}^{(2)}(z) + \eta_v p_{v,w}^{(3)}(z) + \frac{1}{nh^d}q_{v,w}^{(1)}(z) + \eta_v^2 q_{v,w}^{(2)}(z) + \frac{\eta_v}{\sqrt{nh^d}}q_{v,w}^{(3)}(z)$$

$$+\mathbb{1}\{v\neq w\}\rho^{d+\hbar+[q]}(\Omega_1+\rho^{\hbar+[q]}\Omega_2)\frac{\phi(z)}{2}z,$$

we have

$$\sup_{z\in\mathbb{R}}|\mathbb{P}[T_{v,w}<z]-F_{v,w}(z)|=o\left(((nh^d)^{-1/2}+\eta_v)^2+\mathbb{1}\{v\neq w\}\rho^{d+2(\hbar+[q])}\right).$$

The same conclusions reached in the main text continue to hold for multivariate and/or derivative estimation, both in terms of comparing undersmoothing, bias correction, and robust bias correction, as well as for inference-optimal bandwidth choices. In particular, it is straightforward that the MSE optimal bandwidth in general has the rate $n^{-1/(d+2\hbar+2[q])}$, whereas the coverage error optimal choice is of order $n^{-1/(d+\hbar+[q])}$. Note that these two fit the same patter as in the univariate, level case, with $\hbar+[q]$ in place of $\hbar$ and d in place of one. One intuitive reason for the similarity is that the number of derivatives in question does not impact that variance or higher order moment terms of the expansion, *once the scaling is accounted for*. That is, for all averages beyond the first, for example of the kernel squared, $\sqrt{nh^d}$ can be thought of as the effective sample size, since that is the multiplier which stabilizes averages.

S.I.7 Proof of Main Result

Throughout C shall be a generic constant that may take different values in different uses. If more than one constant is needed, C_1, C_2, ..., will be used. It will cause no confusion (as the notations never occur in the same place), but in the course of proofs we will frequently write $s=\sqrt{nh}$, which overlaps with the order of the kernel L.

The first step is to write $T_{v,w}$ as a smooth function of sums of i.i.d. random variables plus a remainder term that is shown to be of higher order. In addition to the notation above, define

$$\gamma_{v,p}=h^{-1}\mathbb{E}\left[N_v\left(X_{h,i}\right)^p\right]\qquad\text{and}\qquad\Delta_{v,j}=\frac{1}{s}\sum_{i=1}^{n}\left\{N_v\left(X_{h,i}\right)^j-\mathbb{E}\left[N_v\left(X_{h,i}\right)^j\right]\right\}.$$

With this notation $\hat{f}_v-\mathbb{E}[\hat{f}_v]=s^{-1}\Delta_{v,1}$, $\sigma_w^2=\mathbb{E}[\Delta_{w,1}^2]=\gamma_{w,2}-h\gamma_{w,1}^2$ and

$$\hat{\sigma}_w^2-\sigma_w^2=s^{-1}\Delta_{w,2}-h2\gamma_{w,1}s^{-1}\Delta_{w,1}-hs^{-2}\Delta_{w,1}^2. \tag{S.I.4}$$

By a change of variables

$$\gamma_{v,p}=h^{-1}\int N_v\left(X_{h,i}\right)^p f(X_i)dX_i=\int N_v(u)^p f(x-uh)du=O(1).$$

Further, by construction $\mathbb{E}[\Delta_{w,j}]=0$ and

$$\mathbb{V}\left[\Delta_{w,j}\right]=h^{-1}\mathbb{E}\left[N_v\left(X_{h,i}\right)^{2j}\right]-h^{-1}\mathbb{E}\left[N_v\left(X_{h,i}\right)^j\right]^2$$

$$\leq h^{-1}\mathbb{E}\left[N_v\left(X_{h,i}\right)^{2j}\right]$$
$$= \gamma_{v,2j} = O(1).$$

Returning to Eqn. (S.I.4) and applying Markov's inequality, we find that $hs^{-2}\Delta^2_{w,1} = n^{-1}\Delta^2_{w,1} = O_p(n^{-1})$ and $\hat{\sigma}^2_w - \sigma^2_w = s^{-1}O_p(1) - hO(1)s^{-1}O_p(1) - hs^{-2}O_p(1) = O_p(s^{-1})$, whence $\left|\hat{\sigma}^2_w - \sigma^2_w\right|^2 = O_p(s^{-2})$. Using these results preceded by a Taylor expansion, we have

$$\left(\frac{\hat{\sigma}^2_w}{\sigma^2_w}\right)^{-1/2} = \left(1 + \frac{\hat{\sigma}^2_w - \sigma^2_w}{\sigma^2_w}\right)^{-1/2} = 1 - \frac{1}{2}\frac{\hat{\sigma}^2_w - \sigma^2_w}{\sigma^2_w} + \frac{3}{8}\frac{(\hat{\sigma}^2_w - \sigma^2_w)^2}{\sigma^4_w} + o_p((\hat{\sigma}^2_w - \sigma^2_w)^2)$$
$$= 1 - \frac{1}{2\sigma^2_w}\left(s^{-1}\Delta_{w,2} - h2\gamma_{w,1}s^{-1}\Delta_{w,1}\right) + O_p(n^{-1} + s^{-2}).$$

Combining this result with the fact that

$$T_{v,w} = \frac{\Delta_{v,1} + \eta_v}{\hat{\sigma}_w} = \frac{\Delta_{v,1}}{\hat{\sigma}_w} + \frac{\eta_v}{\sigma_w}\left(\frac{\hat{\sigma}^2_w}{\sigma^2_w}\right)^{-1/2},$$

we have

$$\mathbb{P}[T_{v,w} < z] = \mathbb{P}\left[\tilde{T}_{v,w} - R_{v,w} < z - \frac{\eta_v}{\sigma_w}\right], \tag{S.I.5}$$

where

$$\tilde{T}_{v,w} = \frac{\Delta_{v,1}}{\hat{\sigma}_w} - \frac{\eta_v}{2\sigma^3_w}\left(s^{-1}\Delta_{w,2} - h2\gamma_{w,1}s^{-1}\Delta_{w,1}\right)$$

and is a smooth function of sums of i.i.d. random variables and the remainder term is

$$R_{v,w} = \frac{\eta_v}{\sigma_w}\left(hs^{-2}\frac{\Delta^2_{w,1}}{2\sigma^2_w} + \frac{3}{8}\frac{(\hat{\sigma}^2_w - \sigma^2_w)^2}{\sigma^4_w} + o_p((\hat{\sigma}^2_w - \sigma^2_w)^2)\right).$$

Next we apply the delta method, see Hall (1992a, Chapter 2.7) or Andrews (2002, Lemma 5(a)). It will be true that

$$\mathbb{P}[T_{v,w} < z] = \mathbb{P}\left[\tilde{T}_{v,w} < z - \frac{\eta_v}{\sigma_w}\right] + o(s^{-2}) \tag{S.I.6}$$

if it can be shown that $s^2\mathbb{P}[|R_{v,w}| > \varepsilon^2 s^{-2}\log(s)^{-1}] = o(1)$.[1] This can be demonstrated by applying Bernstein's inequality to each piece of $R_{v,w}$, as the kernels K and L, and their derivatives, are bounded.

To apply this inequality to the first term of $R_{v,w}$, note that $|N_w((x - X_i)/h)| \leq C_1$ and that

[1] Here, $s^{-2}\log(s)^{-1}$ may be replaced with any sequence that is $o(s^{-2} + \eta_v^2 + s^{-1}\eta_v)$.

$\mathbb{V}[N_w((x - X_i)/h)] \leq C_2 h$, for different constants, and so for $\varepsilon > 0$ we have

$$
\begin{aligned}
&s^2 \mathbb{P}\left[\frac{\eta_v}{\sigma_w} h s^{-2} \frac{\Delta_{w,1}^2}{2\sigma_w^2} > \varepsilon^2 s^{-2} \log(s)^{-1}\right] \\
&\quad = s^2 \mathbb{P}\left[\left|\sum_{i=1}^{n} \left\{N_w\left(X_{h,i}\right) - \mathbb{E}\left[N_w\left(X_{h,i}\right)\right]\right\}\right| > \varepsilon s^{-1} \log(s)^{-1/2} \left(\frac{2\sigma_w^3 n s^2}{\eta_v}\right)^{1/2}\right] \\
&\quad = s^2 \mathbb{P}\left[\left|\sum_{i=1}^{n} \left\{N_w\left(X_{h,i}\right) - \mathbb{E}\left[N_w\left(X_{h,i}\right)\right]\right\}\right| > \varepsilon \left(\frac{2\sigma_w^3 n}{\eta_v \log(s)}\right)^{1/2}\right] \\
&\quad \leq 2s^2 \exp\left\{-\frac{1}{2} \frac{\varepsilon^2 2\sigma_w^3 n \eta_v^{-1} \log(s)^{-1}}{C_2 n h + \frac{1}{3}\varepsilon C_1 \sqrt{2\sigma_w^3 n/[\eta_v \log(s)]}}\right\} \\
&\quad \leq s^2 \exp\left\{-C \frac{\varepsilon^2 \log(s)^{-1}}{\eta h + \varepsilon\sqrt{\eta_v/[n \log(s)]}}\right\} \\
&\quad \leq \exp\left\{C_1 \log(s) \left[1 - C_2 \frac{\varepsilon^2}{\eta h \log(s)^2 + \varepsilon\sqrt{\eta_v \log(s)^3/n}}\right]\right\},
\end{aligned}
$$

which tends to zero because $\eta_v \to 0$ as $n \to \infty$ is assumed. To see why, note first that the second term of the denominator automatically vanishes, as $\eta_v \to 0$ and $\log(s)^3/n \to 0$. Second, suppose $\eta_v^2 \asymp n h^\omega$ (for example, if $\eta_{\mathtt{us}} \asymp s h^{\Bbbk}$, then $\omega = 1 + 2\Bbbk$) and the first term diverges, it must be that h is at least as large (in order) as

$$
\left(\frac{1}{n \log(s)^4}\right)^{1/(2+\omega)},
$$

which makes the requirement that $\eta_v \to 0$ equivalent to

$$
\eta_v^2 \asymp n h^\omega = n^{1-\omega/(2+\omega)} \log(s)^{-4\omega/(2+\omega)} \to 0,
$$

which is impossible. The remaining terms of $R_{v,w}$, characterized using Eqn. (S.I.4), are handled in exactly the same way. This establishes Eqn. (S.I.6).

Next, the proofs of (Hall, 1992a, Chapters 4.4 and 5.5) show that $\tilde{T}_{v,w}$ has an Edgeworth expansion valid through $o(s^{-2} + s^{-1}\eta_v + \eta_v^2)$. Thus, for a smooth function $G(z)$ we can write $\mathbb{P}[\tilde{T}_{v,w} < z] = G(z) + o(s^{-2} + s^{-1}\eta_v + \eta_v^2)$. Therefore

$$
\mathbb{P}\left[\tilde{T}_{v,w} < z - \frac{\eta_v}{\sigma_w}\right] = \mathbb{P}\left[\tilde{T}_{v,w} < z\right] - \frac{\eta_v}{\sigma_w} G^{(1)}(z) + o(s^{-2} + s^{-1}\eta_v + \eta_v^2). \qquad \text{(S.I.7)}
$$

The final result now follows by combining Equations (S.I.5), (S.I.6), and (S.I.7) with the terms of the expansion computed below. □

S.I.7.1 Computing the Terms of the Expansion

Identifying the terms of the expansion is a matter of straightforward, if tedious, calculation. The first four cumulants of $T_{v,w}$ must be calculated, which are functions of the first four moments. In what follows, we give a short summary. Note well that we always discard higher-order terms for brevity, and to save notation we will write $\stackrel{o}{=}$ to stand in for "equal up to $o((nh)^{-1} + (nh)^{-1/2}\eta_v + \eta_v^2 + \mathbb{1}\{v \neq w\}\rho^{1+2k})$".

Referring to the Taylor expansion above, for the purpose of computing moments and cumulants, we can use

$$T_{v,w} \approx \left(\frac{\Delta_{v,1}}{\sigma_w} + \frac{\eta_v}{\sigma_w}\right)\left(1 - \frac{s^{-1}\Delta_{w,2}}{2\sigma_w} + \frac{h\gamma_{w,1}s^{-1}\Delta_{w,1}}{\sigma_w} + \frac{3}{8}\frac{s^{-2}\Delta_{w,2}^2}{\sigma_w^2}\right).$$

Moments of the two sides agree up to the requisite order. Straightforward moment calculations then give

$$\begin{aligned}\mathbb{E}[T_{v,w}] &\stackrel{o}{=} \frac{s^{-1}\mathbb{E}[\Delta_{v,1}\Delta_{w,2}]}{2\sigma_w^3} + \frac{hs^{-1}\gamma_{w,1}\mathbb{E}[\Delta_{v,1}\Delta_{w,1}]}{\sigma_w^3} + \frac{3s^{-2}\mathbb{E}[\Delta_{v,1}\Delta_{w,2}^2]}{8\sigma_w^5} + \frac{\eta_v}{\sigma_w} + \frac{3s^{-2}\eta_v\mathbb{E}[\Delta_{w,2}^2]}{8\sigma_w^5} \\ &\stackrel{o}{=} -s^{-1}\frac{\nu_{v,w}(1,1,2)}{2\sigma_w^3} + \frac{hs^{-1}\gamma_{w,1}\nu_{v,w}(1,1,1)}{\sigma_w^3} + \frac{\eta_v}{\sigma_w},\end{aligned}$$

$$\begin{aligned}\mathbb{E}[T_{v,w}^2] &\stackrel{o}{=} \frac{\mathbb{E}[\Delta_{v,1}^2]}{\sigma_w^2} + s^{-2}\frac{\mathbb{E}[\Delta_{v,1}^2\Delta_{w,2}^2]}{\sigma_w^6} + s^{-1}\frac{\mathbb{E}[\Delta_{v,1}^2\Delta_{w,2}]}{\sigma_w^4} + 2hs^{-1}\frac{\gamma_{w,1}\mathbb{E}[\Delta_{v,1}^2\Delta_{w,1}]}{\sigma_w^2} \\ &\quad - \eta_v s^{-1}\frac{2\mathbb{E}[\Delta_{v,1}\Delta_{w,2}]}{\sigma_w^4} + \eta_v hs^{-1}\frac{4\gamma_{w,1}\mathbb{E}[\Delta_{v,1}\Delta_{w,1}]}{\sigma_w^2} + \frac{\eta_v^2}{\sigma_w^2} \\ &\stackrel{o}{=} \frac{\sigma_v^2}{\sigma_w^2} + s^{-2}\frac{\sigma_v^2\nu_{v,w}(0,2,2)}{\sigma_w^6} + s^{-2}\frac{2\nu_{v,w}(1,1,2)^2}{\sigma_w^6} - s^{-2}\frac{\nu_{v,w}(2,1,2)^2}{\sigma_w^2} - \eta_v s^{-1}\frac{2\nu_{v,w}(1,1,2)}{\sigma_w^2} + \frac{\eta_v^2}{\sigma_w^2},\end{aligned}$$

$$\begin{aligned}\mathbb{E}[T_{v,w}^3] &\stackrel{o}{=} \frac{\mathbb{E}[\Delta_{v,1}^3]}{\sigma_w^3} - 3s^{-1}\frac{\mathbb{E}[\Delta_{v,1}^3\Delta_{w,2}]}{2\sigma_w^5} + 3hs^{-1}\frac{\gamma_{w,1}\mathbb{E}[\Delta_{v,1}^3\Delta_{w,1}]}{\sigma_w^5} + \eta_v\frac{3\mathbb{E}[\Delta_{v,1}^2]}{\sigma_w^3} - \eta_v s^{-1}\frac{9\mathbb{E}[\Delta_{v,1}^2\Delta_{w,2}]}{2\sigma_w^5} \\ &\stackrel{o}{=} s^{-1}\frac{\nu_v(3)}{\sigma_w^3} - s^{-1}\frac{9\nu_{v,w}(1,1,2)\sigma_v^2}{2\sigma_w^5} + hs^{-1}\frac{9\gamma_{w,1}\nu_{v,w}(1,1,1)}{\sigma_w^5} + \eta_v\frac{3\sigma_v^2}{\sigma_w^3},\end{aligned}$$

and,

$$\begin{aligned}\mathbb{E}[T_{v,w}^4] &\stackrel{o}{=} \frac{\mathbb{E}[\Delta_{v,1}^4]}{\sigma_w^4} - s^{-1}\frac{2\mathbb{E}[\Delta_{v,1}^4\Delta_{w,2}]}{\sigma_w^6} + 4hs^{-1}\frac{\gamma_{w,1}\mathbb{E}[\Delta_{v,1}^4\Delta_{w,1}]}{\sigma_w^6} + s^{-2}\frac{3\mathbb{E}[\Delta_{v,1}^4\Delta_{w,1}^2]}{\sigma_w^8} \\ &\quad + \eta_v\frac{4\mathbb{E}[\Delta_{v,1}^3]}{\sigma_w^4} - \eta_v s^{-1}\frac{8\mathbb{E}[\Delta_{v,1}^3\Delta_{w,2}]}{\sigma_w^6} + \eta_v^2\frac{6\mathbb{E}[\Delta_{v,1}^2]}{\sigma_w^4} \\ &\stackrel{o}{=} s^{-2}\frac{\nu_v(4)}{\sigma_w^4} + 3\frac{\sigma_v^4}{\sigma_w^4} - s^{-2}\frac{8\nu_v(3)\nu_{v,w}(1,1,2) + 12\sigma_v^2\nu_{v,w}(2,1,2)}{\sigma_w^6} + s^{-2}\frac{9\sigma_v^4\nu_{v,w}(0,2,2)}{\sigma_w^8}\end{aligned}$$

$$+ s^{-2}\frac{36\sigma_v^2\nu_{v,w}(1,1,2)^2}{\sigma_w^8} + \eta_v s^{-1}\frac{4\nu_v(3)}{\sigma_w^4} - \eta_v s^{-1}\frac{24\sigma_v^2\nu_{v,w}(1,1,2)}{\sigma_w^6} + \eta_v^2\frac{6\sigma_v^2}{\sigma_w^2}.$$

The expansion now follows, formally, from the following steps. First, combining the above moments into cumulants. Second, these cumulants may be simplified using that

$$\frac{\sigma_v^2}{\sigma_w^2} = 1 + \mathbb{1}(w \neq v)\left(\rho^{1+\mathfrak{K}}\Omega_1 + \rho^{1+2\mathfrak{K}}\Omega_2\right)$$

and in all cases present

$$\nu_{v,w}(i,j,p) = f\vartheta_{N_v,i+jp} + o(1). \tag{S.I.8}$$

The second relation is readily proven for $v = w$, as $\nu_{v,v}(i,j,p) = \mathbb{E}[N_v(X_{h,i})^{i+jp}] + O(h)$, where the remainder represents products of expectations. In the case for $v \neq w$, we find $\nu_{2,1}(i,j,p) = f\vartheta_{N_1,i+jp} + O(\rho^{1+\mathfrak{K}} + h)$, and in this case $\rho \to 0$ is assumed. For any term of a cumulant with a rate of $(nh)^{-1}$, $(nh)^{-1/2}\eta_v$, η_v^2, or $\rho^{1+2\mathfrak{K}}$ (i.e., the extent of the expansion), these simplifications may be inserted as the remainder will be negligible. Note that this is exactly why the polynomials $p_{v,w}^{(k)}$ do not simplify, while the $q_{v,w}^{(k)}$ do. Third, with the cumulants in hand, the terms of the expansion are determined as described by e.g., Hall (1992a, Chapter 2).

Finally, for traditional bias correction, there are additional terms in the expansion (see discussion in the main text) representing the covariance of $\hat{f}$ and $\hat{B}_f$ (denoted by Ω_1) and the variance of $\hat{B}_f$ (Ω_2). We now state their precise forms. These arise from the mismatch between the variance of the numerator of $T_{\texttt{bc}}$ and the standardization used, $\sigma^2_{\texttt{us}}$, that is $\sigma^2_{\texttt{rbc}}/\sigma^2_{\texttt{us}}$ is given by

$$\frac{nh\mathbb{V}[\hat{f} - \hat{B}_f]}{nh\mathbb{V}[\hat{f}]} = \frac{nh\mathbb{V}[\hat{f}] - 2nh\mathbb{C}[\hat{f},\hat{B}_f] + nh\mathbb{V}[\hat{B}_f]}{nh\mathbb{V}[\hat{f}]} = 1 - 2\frac{nh\mathbb{C}[\hat{f},\hat{B}_f]}{nh\mathbb{V}[\hat{f}]} + \frac{nh\mathbb{V}[\hat{B}_f]}{nh\mathbb{V}[\hat{f}]}.$$

This makes clear that Ω_1 and Ω_2 are the constant portions of the last two terms. First, for Ω_1,

$$\begin{aligned}
\mathbb{C}[\hat{f},\hat{B}_f] &= \mathbb{E}\left[\left(\frac{1}{nh}\sum_{i=1}^{n} K\left(X_{h,i}\right)\right)\left(h^{\mathfrak{K}}\mu_{K,\mathfrak{K}}\frac{1}{nb^{1+\mathfrak{K}}}\sum_{i=1}^{n} L^{(\mathfrak{K})}\left(X_{b,i}\right)\right)\right] \\
&= h^{\mathfrak{K}}\mu_{K,\mathfrak{K}}\frac{1}{nb^{1+\mathfrak{K}}}\Big\{\mathbb{E}\left[h^{-1}K\left(X_{h,i}\right)L^{(\mathfrak{K})}\left(X_{b,i}\right)\right] \\
&\qquad - b\mathbb{E}\left[h^{-1}K\left(X_{h,i}\right)\right]\mathbb{E}\left[b^{-1}L^{(\mathfrak{K})}\left(X_{b,i}\right)\right]\Big\} \\
&= \frac{\rho^{\mathfrak{K}}\mu_{K,\mathfrak{K}}}{nb}\left\{\int f(x-uh)K(u)L^{(\mathfrak{K})}(u\rho)du - b\int f(x-uh)K(u)du\int f(x-ub)L^{(\mathfrak{K})}(u)du\right\}.
\end{aligned}$$

Therefore

$$-2\frac{nh\mathbb{C}[\hat{f},\hat{B}_f]}{nh\mathbb{V}[\hat{f}]} = \rho^{1+\mathfrak{K}}\Omega_1,$$

where

$$\Omega_1 = -2\frac{\mu_{K,\mathfrak{k}}}{\nu_1(2)}\left\{\int f(x-uh)K(u)L^{(\mathfrak{k})}(u\rho)du - b\int f(x-uh)K(u)du\int f(x-ub)L^{(\mathfrak{k})}(u)du\right\}.$$

Note $\nu_1(2) = \sigma^2_{\mathtt{us}}$. If we did not include Ω_2 in the Edgeworth expansion, i.e. we stopped at order $\rho^{1+\mathfrak{k}}$, then we could capture only the leading terms of Ω_1, as follows, using that kernel integrates to 1 and $\rho \to 0$,

$$\begin{aligned}\Omega_1 &= -2\frac{\mu_{K,\mathfrak{k}}}{\nu_1(2)}\left\{\int f(x-uh)K(u)L^{(\mathfrak{k})}(u\rho)du - b\int f(x-uh)K(u)du\int f(x-ub)L^{(\mathfrak{k})}(u)du\right\}\\ &= -2\frac{\mu_{K,\mathfrak{k}}}{f(x)\vartheta^2_{K,2}+O(h)}\left\{f(x)L^{(\mathfrak{k})}(0)[1+O(h+h\rho)] - bf(x)^2\int L^{(\mathfrak{k})}(u)du[1+O(b+h)]\right\}\\ &\to -2\mu_{K,\mathfrak{k}}\vartheta^{-2}_{K,2}L^{(\mathfrak{k})}(0).\end{aligned}$$

Note that this matches the term Hall (1992b) calls w_2. We do not do this, for completeness. There are no other terms of up to order $\rho^{1+2\mathfrak{k}}$, so capturing the full contribution of $\sigma_2^2/\sigma_1^2 - 1 = \sigma^2_{\mathtt{rbc}}/\sigma^2_{\mathtt{us}} - 1$ is natural and informative.

Turning to Ω_2, using the calculations in Section S.I.4.1 (recall $\tilde{\mathfrak{k}} = \mathfrak{k} \vee S$), we find that

$$\begin{aligned}\mathbb{V}[\hat{B}_f] &= \frac{h^{2\mathfrak{k}}}{n}\mu^2_{K,\mathfrak{k}}\left\{\frac{1}{b^{1+2\mathfrak{k}}}\mathbb{E}\left[b^{-1}L^{(\mathfrak{k})}\left(X_{b,i}\right)^2\right] - \left(\frac{1}{b^{1+\mathfrak{k}}}\mathbb{E}\left[L^{(\mathfrak{k})}\left(X_{b,i}\right)\right]\right)^2\right\}\\ &= \frac{\rho^{2\mathfrak{k}}\mu^2_{K,\mathfrak{k}}}{nb}\left\{\int f(x-ub)L^{(\mathfrak{k})}(u)^2du - b^{1+2\tilde{\mathfrak{k}}}\left(\int L^{(\mathfrak{k}-\tilde{\mathfrak{k}})}(u)f^{(\tilde{\mathfrak{k}})}(x-ub)du\right)^2\right\},\end{aligned}$$

and hence

$$\frac{nh\mathbb{V}[\hat{B}_f]}{nh\mathbb{V}[\hat{f}]} = \rho^{1+2\mathfrak{k}}\Omega_2 \quad \text{where} \quad \Omega_2 = \frac{\mu^2_{K,\mathfrak{k}}}{\nu_1(2)}\left\{\int f(x-ub)L^{(\mathfrak{k})}(u)^2du - b^{1+2\tilde{\mathfrak{k}}}\left(\int L^{(\mathfrak{k}-\tilde{\mathfrak{k}})}(u)f^{(\tilde{\mathfrak{k}})}(x-ub)du\right)^2\right\}.$$

The final piece will be $b^{1+2S}f^{(\mathfrak{k})}(x)^2[1+o(1)]$ if $\mathfrak{k} \leq S$. Substituting this is permitted because $\rho^{1+2\mathfrak{k}}$ is the limit of the expansion, though it is not necessary to do, because this term is always higher order. Fully simplifying would yield

$$\Omega_2 = \mu^2_{K,\mathfrak{k}}\vartheta^{-2}_{K,2}\vartheta_{L^{(\mathfrak{k})},2},$$

which can be used in Theorem S.I.1.

S.I.8 Complete Simulation Results

To illustrate the gains from robust bias correction we conduct a Monte Carlo study to compare undersmoothing, traditional bias correction, and robust bias correction in terms coverage accuracy and interval length using several data-driven procedures to select the bandwidth. We generate

$n = 500$ observations from a density f given by:

Model 1 (Gaussian Density): $x \backsim \mathcal{N}(0,1)$

Model 2 (Skewed Unimodal Density): $x \backsim \frac{1}{5}\mathcal{N}(0,1) + \frac{1}{5}\mathcal{N}\left(\frac{1}{2}, \left(\frac{2}{3}\right)^2\right) + \frac{3}{5}\mathcal{N}\left(\frac{13}{12}, \left(\frac{5}{9}\right)^2\right)$

Model 3 (Bimodal Density): $x \backsim \frac{1}{2}\mathcal{N}\left(-1, \left(\frac{2}{3}\right)^2\right) + \frac{1}{2}\mathcal{N}\left(1, \left(\frac{2}{3}\right)^2\right)$

Model 4 (Asymmetric Bimodal Density): $x \backsim \frac{3}{4}\mathcal{N}(0,1) + \frac{1}{4}\mathcal{N}\left(\frac{3}{2}, \left(\frac{1}{3}\right)^2\right)$

We evaluate the density at $x = \{-2, -1, 0, 1, 2\}$. These models were previously analyzed in Marron and Wand (1992) and they are plotted in Figure S.I.1. In this simulation study we compare the performance of the confidence intervals defined by $T_{\mathtt{us}}$, $T_{\mathtt{bc}}$, and $T_{\mathtt{rbc}}$. For $T_{\mathtt{us}}$, we take K to be the Epanechnikov kernel, while bias correction uses the Epanechnikov and MSE-optimal kernels for K and $L^{(2)}$, respectively. The bandwidth h is chosen in three different ways:

(i) population MSE-optimal choice $h_{\mathtt{mse}}$;

(ii) estimated ROT optimal coverage error rate $\hat{h}_{\mathtt{rot}}$.

(iii) estimated DPI optimal coverage error rate $\hat{h}_{\mathtt{dpi}}$.

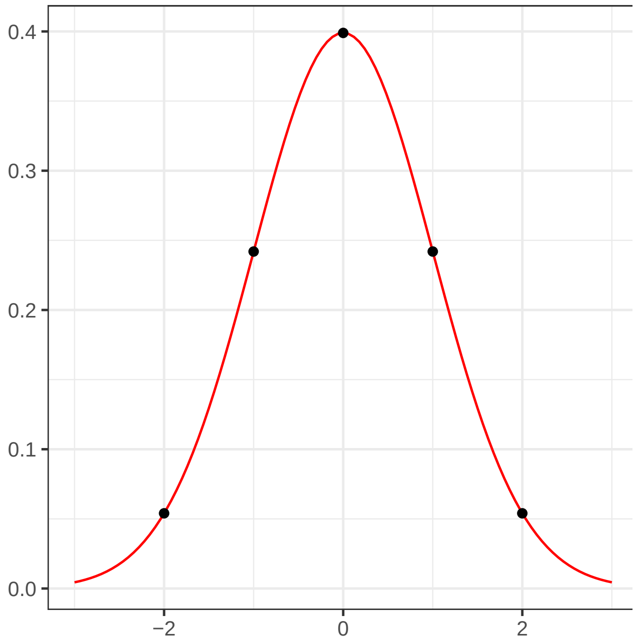
Empirical coverage and length are reported in Tables S.I.2–S.I.5 (Panel A) using our two proposed data-driven bandwidth selectors, as well as the infeasible $h_{\mathtt{mse}}$. The most obvious finding is that robust bias correction has accurate coverage for all bandwidth choices in all models. The intervals are generally longer than for undersmoothing, but neither undersmoothing nor traditional bias correction yield correct coverage outside of a few special cases (e.g., undersmoothing at the infeasible MSE-optimal bandwidth in Model 4). The DPI bandwidth selector generally results in slightly smaller bandwidths (on average). Summary statistics for the two fully data-driven bandwidths are shown in Panel B. The fact that the DPI bandwidth is slightly smaller is born out. It is also, in general, more variable.

To illustrate the robustness to tuning parameter selection, Figures S.I.2–S.I.9 show coverage and length for all four models. The dotted vertical line shows the population MSE-optimal bandwidth for reference. These figures demonstrate the delicate balance required for undersmoothing to provide correct coverage, whereas for a wide range of bandwidths robust bias correction provides correct coverage. Further, interval length is not unduly inflated for bandwidths that provide correct coverage. Recall that robust bias correction can accommodate, and will optimally employ, a larger bandwidth, yielding higher precision. Further emphasizing the point of robustness, we depart from $\rho = 1$ in Figures S.I.10 and S.I.11 to show coverage and length over a grid of h and ρ.

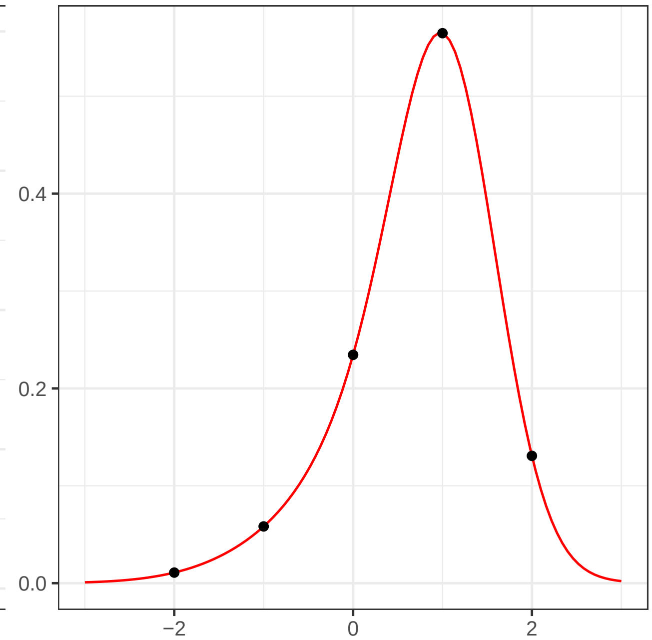
The simulation results for local polynomial regression reported in Section S.II.7 below bear out these same conclusions and study these issues in more detail, in particular interval length.

All our methods are implemented in `R` and `STATA` via the `nprobust` package, available from http://sites.google.com/site/nppackages/nprobust (see also http://cran.r-project.org/package=nprobust). See Calonico et al. (2017) for a complete description.

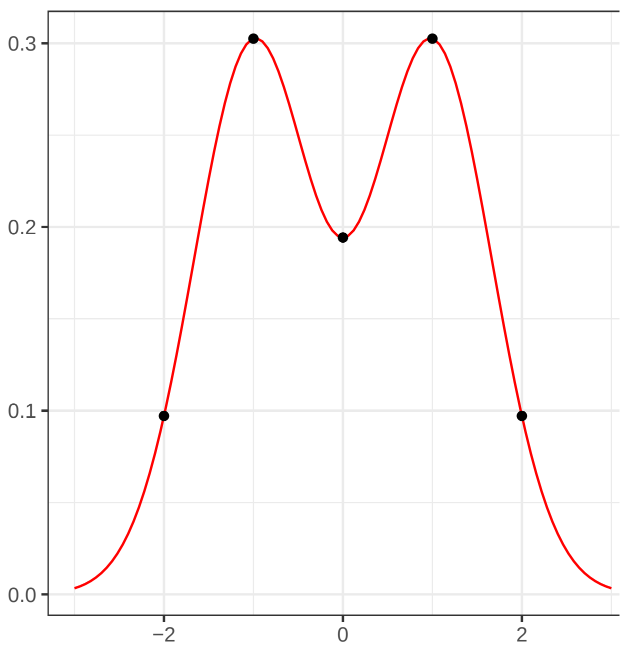
Figure S.I.1: Density Functions



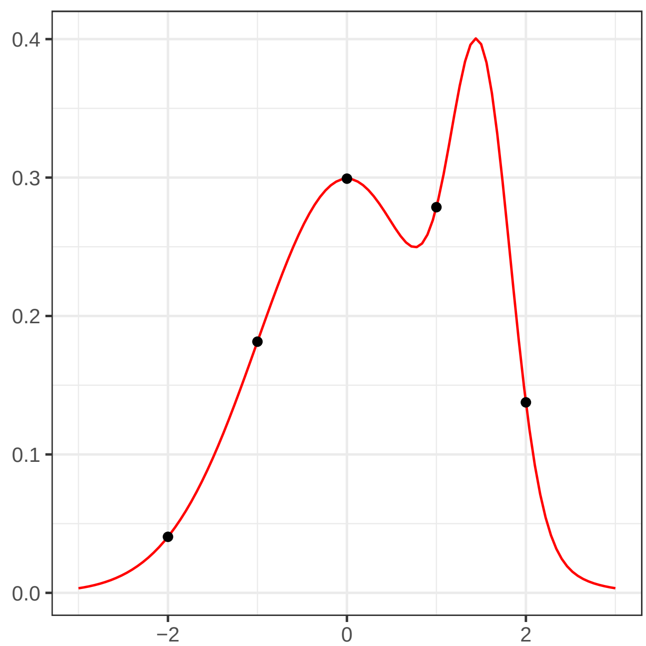
(a) Model 1



(b) Model 2



(c) Model 3



(d) Model 4

Table S.I.2: Simulations Results for Model 1

Panel A: Empirical Coverage and Average Interval Length of 95% Confidence Intervals

	Bandwidth	**Empirical Coverage**			**Interval Length**	
		US	BC	RBC	US	RBC
$x = -2$						
$h_{\mathtt{mse}}$	0.819	82.4	88.0	94.7	0.035	0.042
$\hat{h}_{\mathtt{rot}}$	0.746	82.6	86.1	93.0	0.037	0.044
$\hat{h}_{\mathtt{dpi}}$	0.543	90.1	86.1	92.2	0.043	0.052
$x = -1$						
$h_{\mathtt{mse}}$	-	-	-	-	-	-
$\hat{h}_{\mathtt{rot}}$	1.224	90.1	83.5	93.7	0.044	0.060
$\hat{h}_{\mathtt{dpi}}$	0.665	93.7	86.6	93.8	0.073	0.093
$x = 0$						
$h_{\mathtt{mse}}$	0.842	64.1	78.3	91.3	0.064	0.088
$\hat{h}_{\mathtt{rot}}$	0.775	73.3	79.5	91.5	0.069	0.094
$\hat{h}_{\mathtt{dpi}}$	0.665	80.7	80.7	90.9	0.080	0.107
$x = 1$						
$h_{\mathtt{mse}}$	-	-	-	-	-	-
$\hat{h}_{\mathtt{rot}}$	1.221	90.0	83.5	93.9	0.044	0.060
$\hat{h}_{\mathtt{dpi}}$	0.666	93.9	87.0	94.2	0.073	0.093
$x = 2$						
$h_{\mathtt{mse}}$	0.819	83.0	88.8	94.9	0.035	0.042
$\hat{h}_{\mathtt{rot}}$	0.745	83.2	86.8	93.3	0.037	0.044
$\hat{h}_{\mathtt{dpi}}$	0.541	90.5	87.0	92.4	0.043	0.052

Panel B: Summary Statistics for the Estimated Bandwidths

	Pop. Par.	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.	Std. Dev.
$x = -2$								
$\hat{h}_{\mathtt{rot}}$	0.819	0.546	0.698	0.741	0.746	0.789	1.11	0.07
$\hat{h}_{\mathtt{dpi}}$	-	0.397	0.462	0.493	0.543	0.544	1.95	0.17
$x = -1$								
$\hat{h}_{\mathtt{rot}}$	-	0.898	1.1	1.17	1.22	1.28	9.42	0.27
$\hat{h}_{\mathtt{dpi}}$	-	0.357	0.476	0.588	0.665	0.788	2.01	0.25
$x = 0$								
$\hat{h}_{\mathtt{rot}}$	0.842	0.667	0.756	0.775	0.775	0.795	0.876	0.029
$\hat{h}_{\mathtt{dpi}}$	-	0.425	0.596	0.637	0.665	0.699	1.79	0.11
$x = 1$								
$\hat{h}_{\mathtt{rot}}$	-	0.895	1.1	1.17	1.22	1.28	5.84	0.24
$\hat{h}_{\mathtt{dpi}}$	-	0.356	0.478	0.583	0.666	0.791	2.05	0.25
$x = 2$								
$\hat{h}_{\mathtt{rot}}$	0.819	0.55	0.695	0.741	0.745	0.789	1.11	0.071
$\hat{h}_{\mathtt{dpi}}$	-	0.398	0.462	0.494	0.541	0.545	1.95	0.16

Notes:
(i) US = Undersmoothing, BC = Bias Corrected, RBC = Robust Bias Corrected.
(ii) Columns under "Bandwidth" report the average estimated bandwidths choices, as appropriate, for bandwidth h_n.

Table S.I.3: Simulations Results for Model 2

Panel A: Empirical Coverage and Average Interval Length of 95% Confidence Intervals

	Bandwidth	**Empirical Coverage**			**Interval Length**	
		US	BC	RBC	US	RBC
$x=-2$						
$h_{\mathtt{mse}}$	1.005	90.3	90.4	93.9	0.015	0.018
$\hat{h}_{\mathtt{rot}}$	1.092	94.3	92.4	95.6	0.015	0.017
$\hat{h}_{\mathtt{dpi}}$	1.108	91.8	92.4	96.0	0.015	0.017
$x=-1$						
$h_{\mathtt{mse}}$	0.942	80.9	87.3	93.9	0.034	0.040
$\hat{h}_{\mathtt{rot}}$	0.622	91.5	87.6	93.6	0.041	0.049
$\hat{h}_{\mathtt{dpi}}$	0.685	85.5	85.2	91.9	0.040	0.048
$x=0$						
$h_{\mathtt{mse}}$	0.772	77.2	84.0	93.8	0.063	0.081
$\hat{h}_{\mathtt{rot}}$	2.119	8.6	13.3	19.8	0.025	0.041
$\hat{h}_{\mathtt{dpi}}$	0.357	94.1	88.6	94.4	0.103	0.127
$x=1$						
$h_{\mathtt{mse}}$	0.614	41.9	72.2	88.0	0.088	0.122
$\hat{h}_{\mathtt{rot}}$	0.593	49.7	72.5	87.6	0.091	0.126
$\hat{h}_{\mathtt{dpi}}$	0.457	79.7	81.3	91.5	0.115	0.153
$x=2$						
$h_{\mathtt{mse}}$	0.603	70.6	85.5	92.9	0.061	0.074
$\hat{h}_{\mathtt{rot}}$	0.913	23.3	53.9	63.4	0.049	0.061
$\hat{h}_{\mathtt{dpi}}$	0.324	93.6	88.5	93.9	0.084	0.102

Panel B: Summary Statistics for the Estimated Bandwidths

	Pop. Par.	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.	Std. Dev.
$x=-2$								
$\hat{h}_{\mathtt{rot}}$	1.005	0.775	1	1.09	1.09	1.17	2.44	0.12
$\hat{h}_{\mathtt{dpi}}$	-	0.684	1.01	1.1	1.11	1.2	1.9	0.14
$x=-1$								
$\hat{h}_{\mathtt{rot}}$	0.942	0.472	0.584	0.619	0.622	0.657	0.844	0.055
$\hat{h}_{\mathtt{dpi}}$	-	0.376	0.528	0.656	0.685	0.774	1.84	0.21
$x=0$								
$\hat{h}_{\mathtt{rot}}$	0.772	0.678	1.35	1.69	2.12	2.25	116	2.38
$\hat{h}_{\mathtt{dpi}}$	-	0.268	0.324	0.342	0.357	0.367	1.38	0.074
$x=1$								
$\hat{h}_{\mathtt{rot}}$	0.614	0.513	0.578	0.593	0.593	0.607	0.682	0.022
$\hat{h}_{\mathtt{dpi}}$	-	0.371	0.436	0.453	0.457	0.474	0.776	0.033
$x=2$								
$\hat{h}_{\mathtt{rot}}$	0.603	0.529	0.772	0.864	0.913	0.988	5.83	0.27
$\hat{h}_{\mathtt{dpi}}$	-	0.272	0.309	0.321	0.324	0.336	1.03	0.025

Notes:
(i) US = Undersmoothing, BC = Bias Corrected, RBC = Robust Bias Corrected.
(ii) Columns under "Bandwidth" report the average estimated bandwidths choices, as appropriate, for bandwidth h_n.

Table S.I.4: Simulations Results for Model 3

Panel A: Empirical Coverage and Average Interval Length of 95% Confidence Intervals

	Bandwidth	**Empirical Coverage**			**Interval Length**	
		US	BC	RBC	US	RBC
$x = -2$						
$h_{\mathtt{mse}}$	0.767	82.7	86.6	93.8	0.047	0.057
$\hat{h}_{\mathtt{rot}}$	2.843	1.5	3.6	5.1	0.021	0.029
$\hat{h}_{\mathtt{dpi}}$	0.554	89.8	86.8	92.4	0.056	0.067
$x = -1$						
$h_{\mathtt{mse}}$	0.716	65.6	79.3	89.7	0.070	0.092
$\hat{h}_{\mathtt{rot}}$	1.204	3.1	29.6	45.4	0.046	0.063
$\hat{h}_{\mathtt{dpi}}$	0.663	72.4	79.5	89.7	0.075	0.097
$x = 0$						
$h_{\mathtt{mse}}$	0.695	74.3	83.2	92.6	0.064	0.081
$\hat{h}_{\mathtt{rot}}$	1.096	1.2	44.8	67.2	0.046	0.061
$\hat{h}_{\mathtt{dpi}}$	0.431	92.7	87.6	94.3	0.085	0.105
$x = 1$						
$h_{\mathtt{mse}}$	0.716	66.6	79.3	89.8	0.070	0.092
$\hat{h}_{\mathtt{rot}}$	1.202	2.6	31.0	46.8	0.046	0.063
$\hat{h}_{\mathtt{dpi}}$	0.662	72.4	79.3	89.7	0.075	0.097
$x = 2$						
$h_{\mathtt{mse}}$	0.767	82.1	86.2	93.7	0.047	0.057
$\hat{h}_{\mathtt{rot}}$	2.829	1.4	3.5	5.0	0.021	0.029
$\hat{h}_{\mathtt{dpi}}$	0.554	89.3	86.0	92.2	0.056	0.067

Panel B: Summary Statistics for the Estimated Bandwidths

	Pop. Par.	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.	Std. Dev.
$x = -2$								
$\hat{h}_{\mathtt{rot}}$	0.767	1.16	1.89	2.29	2.84	3.03	46.7	1.98
$\hat{h}_{\mathtt{dpi}}$	-	0.411	0.494	0.527	0.554	0.573	1.82	0.12
$x = -1$								
$\hat{h}_{\mathtt{rot}}$	0.716	0.973	1.14	1.19	1.2	1.25	1.86	0.09
$\hat{h}_{\mathtt{dpi}}$	-	0.572	0.638	0.659	0.663	0.683	0.954	0.037
$x = 0$								
$\hat{h}_{\mathtt{rot}}$	0.695	0.953	1.07	1.09	1.1	1.12	1.31	0.043
$\hat{h}_{\mathtt{dpi}}$	-	0.375	0.416	0.428	0.431	0.443	0.604	0.023
$x = 1$								
$\hat{h}_{\mathtt{rot}}$	0.716	0.968	1.14	1.19	1.2	1.25	1.84	0.09
$\hat{h}_{\mathtt{dpi}}$	-	0.565	0.637	0.658	0.662	0.683	1.21	0.037
$x = 2$								
$\hat{h}_{\mathtt{rot}}$	0.767	1.24	1.89	2.3	2.83	3.02	119	2.50
$\hat{h}_{\mathtt{dpi}}$	-	0.417	0.494	0.526	0.554	0.57	1.83	0.13

Notes:
(i) US = Undersmoothing, BC = Bias Corrected, RBC = Robust Bias Corrected.
(ii) Columns under "Bandwidth" report the average estimated bandwidths choices, as appropriate, for bandwidth h_n.

Table S.I.5: Simulations Results for Model 4

Panel A: Empirical Coverage and Average Interval Length of 95% Confidence Intervals

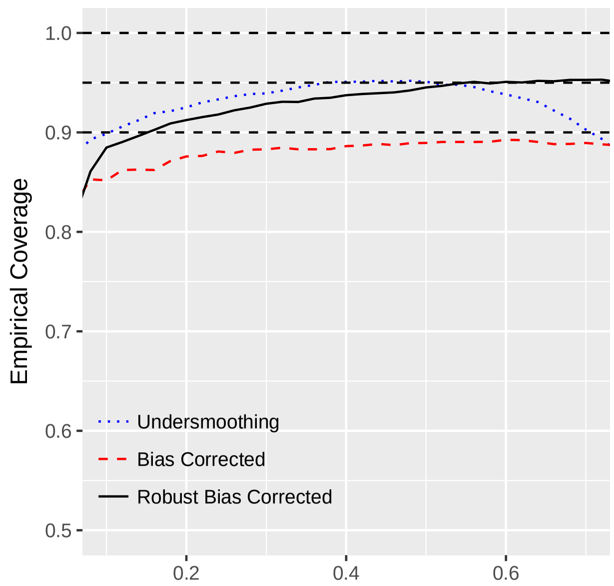
	Bandwidth	**Empirical Coverage**			**Interval Length**	
		US	BC	RBC	US	RBC
$x = -2$						
$h_{\mathtt{mse}}$	0.853	84.3	88.8	94.4	0.030	0.036
$\hat{h}_{\mathtt{rot}}$	0.844	78.9	85.4	91.8	0.030	0.036
$\hat{h}_{\mathtt{dpi}}$	0.579	91.7	87.4	92.5	0.036	0.043
$x = -1$						
$h_{\mathtt{mse}}$	-	-	-	-	-	-
$\hat{h}_{\mathtt{rot}}$	1.751	77.3	79.0	88.3	0.032	0.044
$\hat{h}_{\mathtt{dpi}}$	0.823	93.3	87.2	94.5	0.057	0.072
$x = 0$						
$h_{\mathtt{mse}}$	0.879	74.1	81.1	91.6	0.060	0.080
$\hat{h}_{\mathtt{rot}}$	1.086	44.9	66.8	82.7	0.050	0.068
$\hat{h}_{\mathtt{dpi}}$	0.791	78.4	81.4	92.0	0.067	0.088
$x = 1$						
$h_{\mathtt{mse}}$	0.600	81.0	83.1	92.8	0.079	0.101
$\hat{h}_{\mathtt{rot}}$	0.900	55.5	60.3	80.3	0.058	0.078
$\hat{h}_{\mathtt{dpi}}$	0.804	59.9	64.4	86.0	0.066	0.086
$x = 2$						
$h_{\mathtt{mse}}$	0.526	75.9	85.0	92.5	0.068	0.082
$\hat{h}_{\mathtt{rot}}$	1.872	2.6	1.0	3.7	0.031	0.042
$\hat{h}_{\mathtt{dpi}}$	0.816	36.7	43.2	53.2	0.055	0.067

Panel B: Summary Statistics for the Estimated Bandwidths

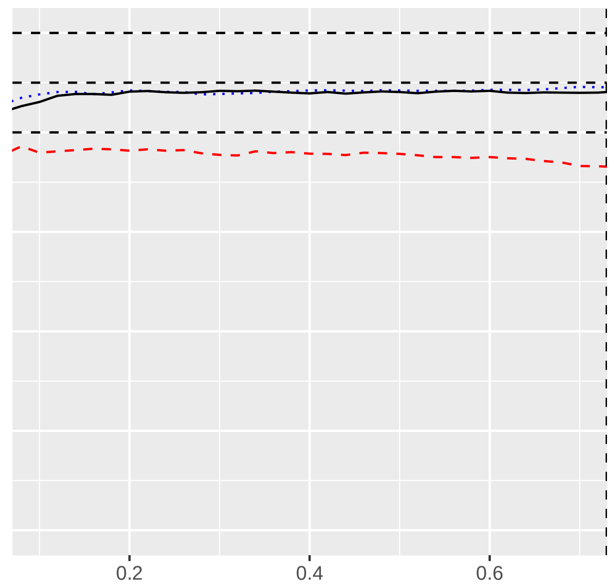
	Pop. Par.	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.	Std. Dev.
$x = -2$								
$\hat{h}_{\mathtt{rot}}$	0.853	0.632	0.781	0.839	0.844	0.896	1.25	0.088
$\hat{h}_{\mathtt{dpi}}$	-	0.447	0.515	0.545	0.579	0.589	1.86	0.13
$x = -1$								
$\hat{h}_{\mathtt{rot}}$	-	1.1	1.4	1.55	1.75	1.8	16.6	0.83
$\hat{h}_{\mathtt{dpi}}$	-	0.395	0.659	0.794	0.823	0.934	2.06	0.24
$x = 0$								
$\hat{h}_{\mathtt{rot}}$	0.879	0.918	1.04	1.08	1.09	1.12	1.53	0.063
$\hat{h}_{\mathtt{dpi}}$	-	0.424	0.635	0.757	0.791	0.893	1.99	0.23
$x = 1$								
$\hat{h}_{\mathtt{rot}}$	0.600	0.787	0.876	0.899	0.9	0.923	1.08	0.036
$\hat{h}_{\mathtt{dpi}}$	-	0.429	0.69	0.768	0.804	0.874	2.03	0.21
$x = 2$								
$\hat{h}_{\mathtt{rot}}$	0.526	1.08	1.43	1.6	1.87	1.89	61	1.57
$\hat{h}_{\mathtt{dpi}}$	-	0.412	0.606	0.795	0.816	0.94	2.01	0.26

Notes:
(i) US = Undersmoothing, BC = Bias Corrected, RBC = Robust Bias Corrected.
(ii) Columns under "Bandwidth" report the average estimated bandwidths choices, as appropriate, for bandwidth h_n.

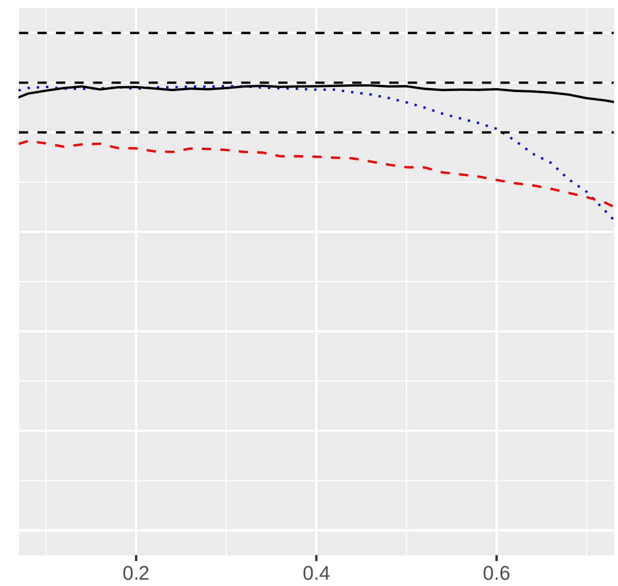
Figure S.I.2: Empirical Coverage of 95% Confidence Intervals - Model 1



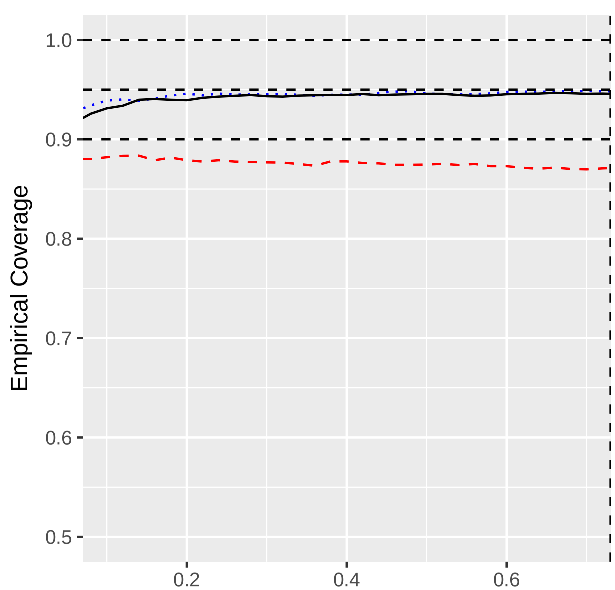
(a) $x = -2$



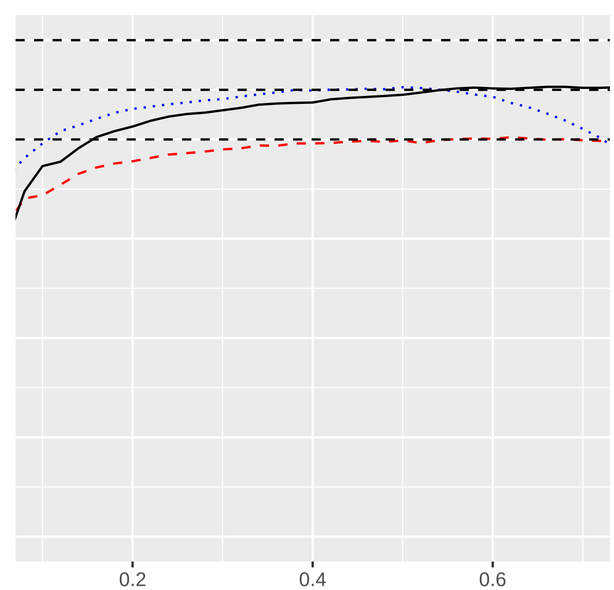
(b) $x = -1$



(c) $x = 0$

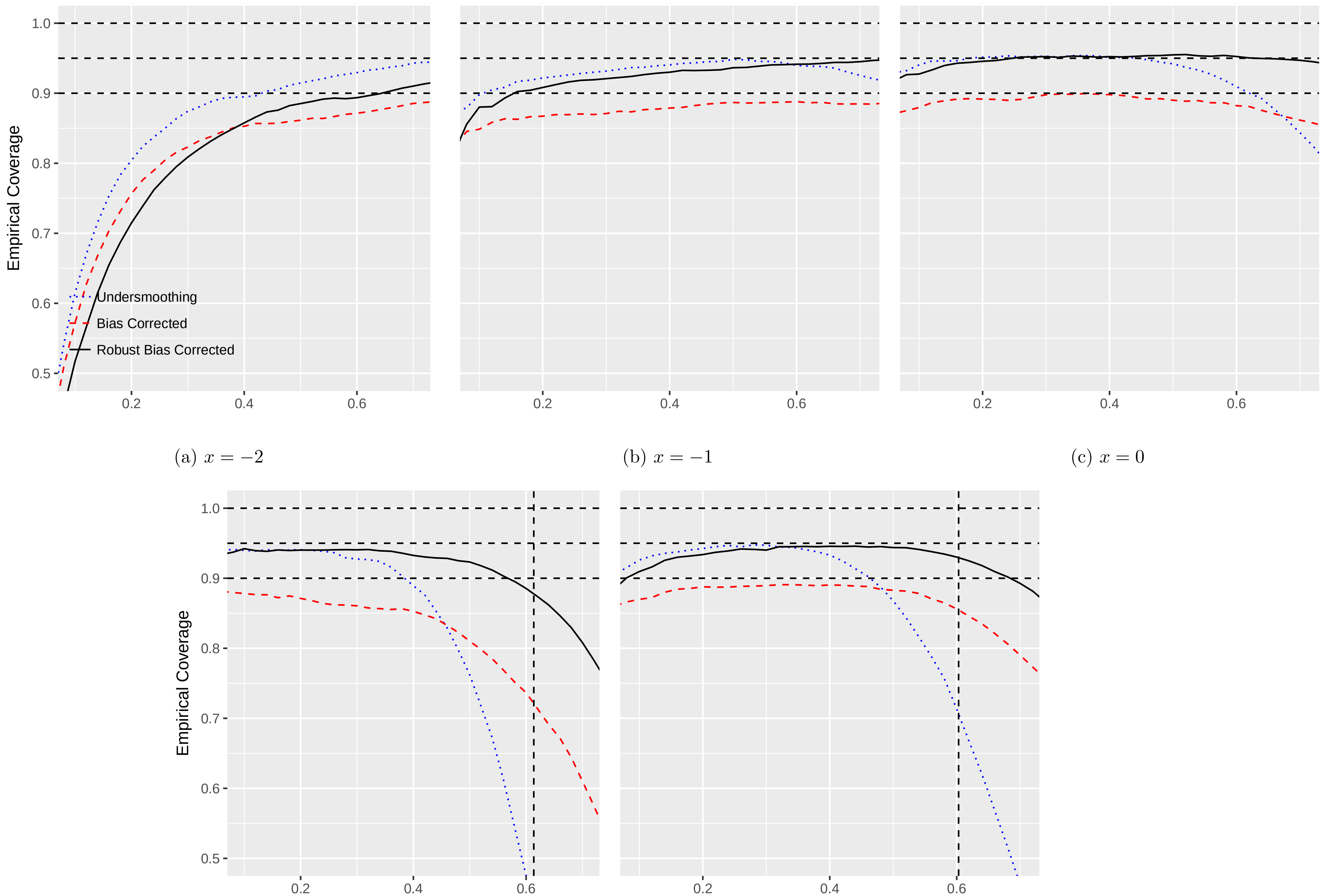


(d) $x = 1$



(e) $x = 2$

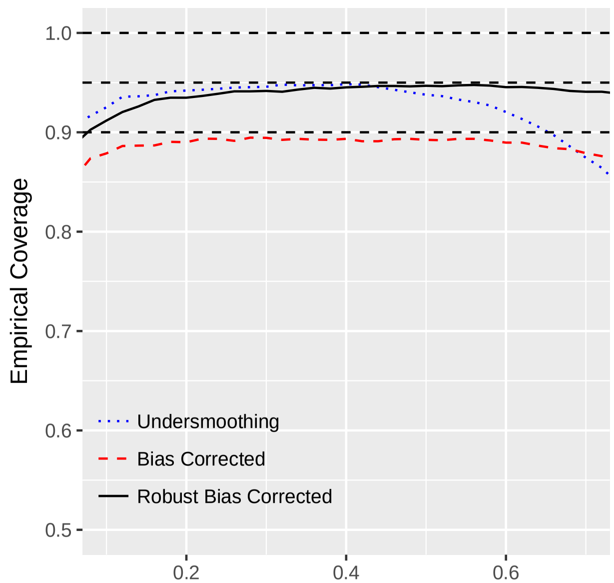
Figure S.I.3: Empirical Coverage of 95% Confidence Intervals - Model 2



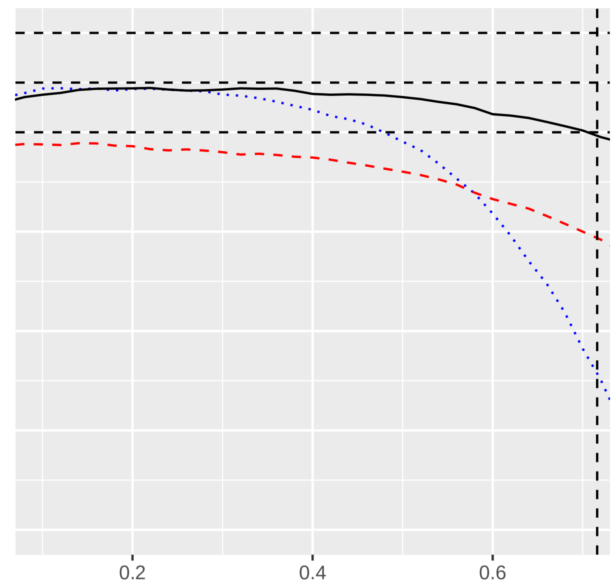
(a) $x = -2$ (b) $x = -1$ (c) $x = 0$

(d) $x = 1$ (e) $x = 2$

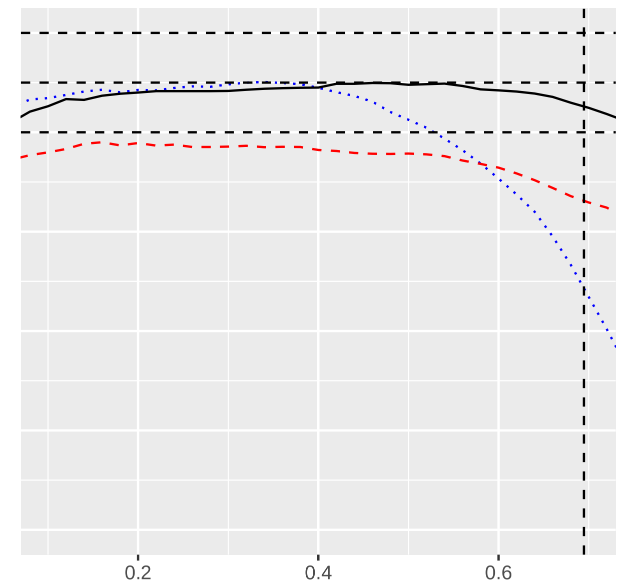
Figure S.I.4: Empirical Coverage of 95% Confidence Intervals - Model 3



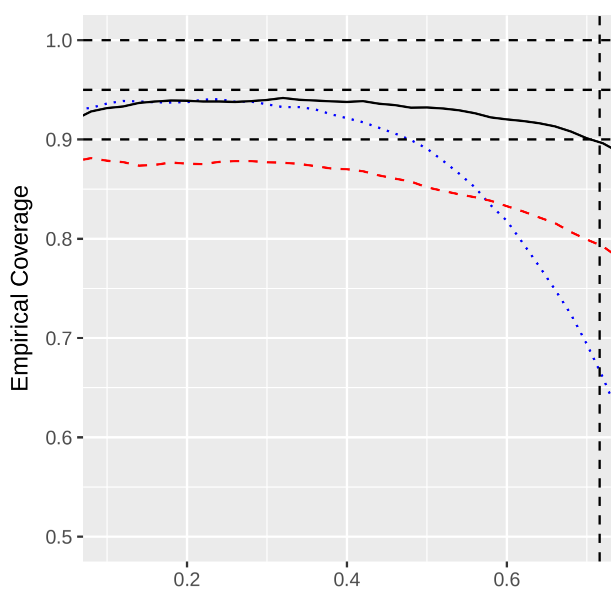
(a) $x = -2$



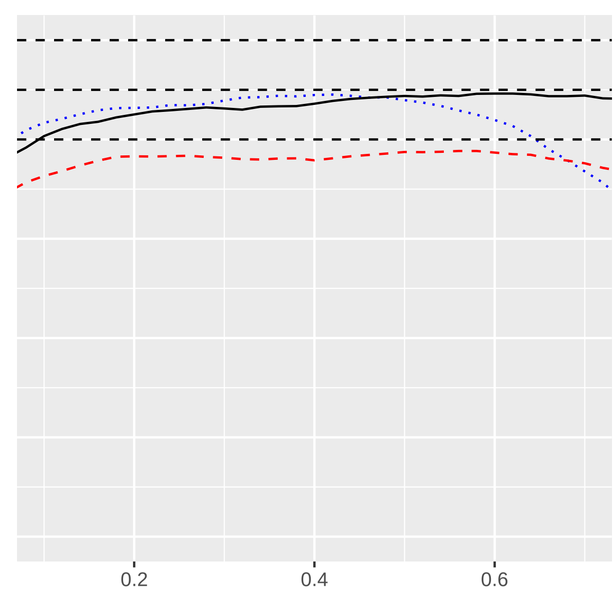
(b) $x = -1$



(c) $x = 0$

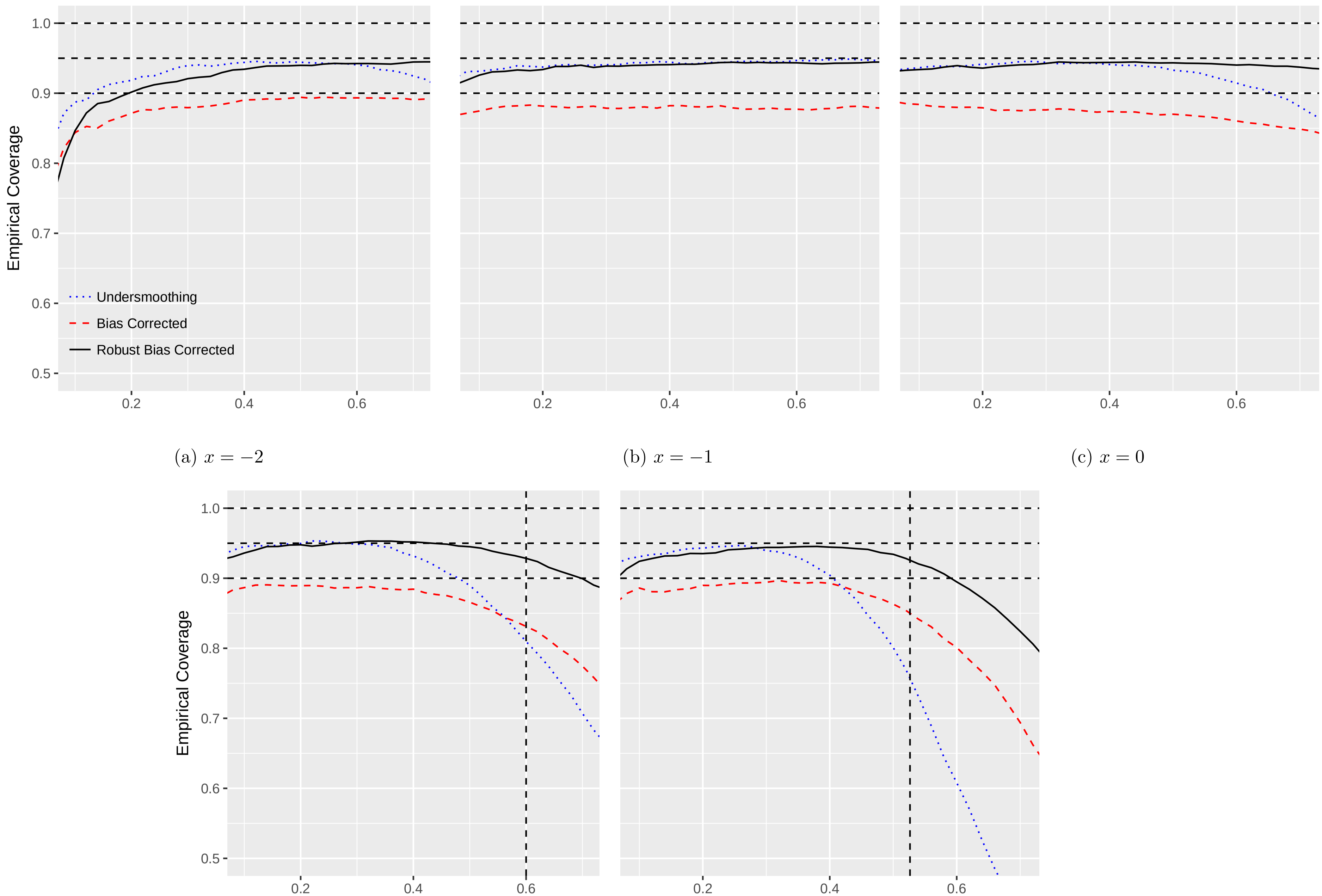


(d) $x = 1$



(e) $x = 2$

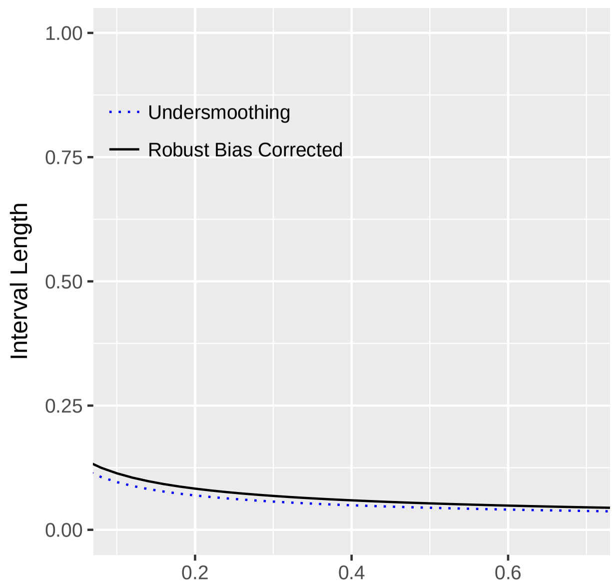
Figure S.I.5: Empirical Coverage of 95% Confidence Intervals - Model 4



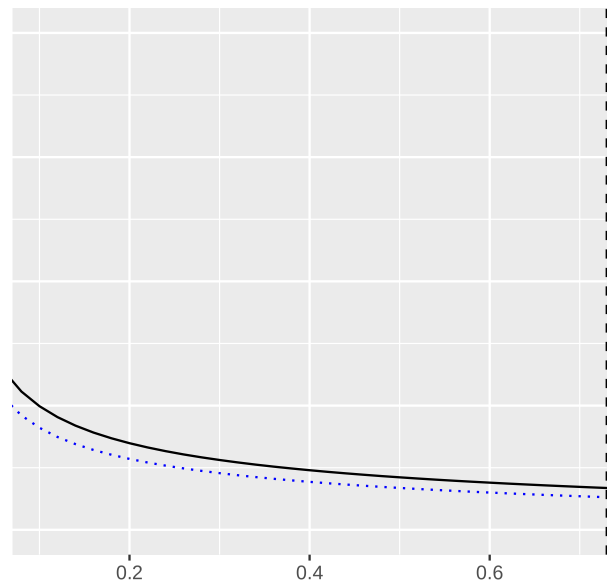
(a) $x = -2$ (b) $x = -1$ (c) $x = 0$

(d) $x = 1$ (e) $x = 2$

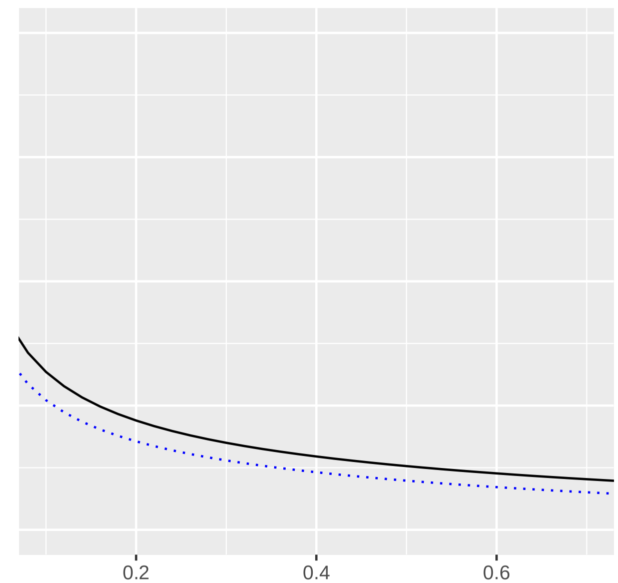
Figure S.I.6: Average Interval Length of 95% Confidence Intervals - Model 1



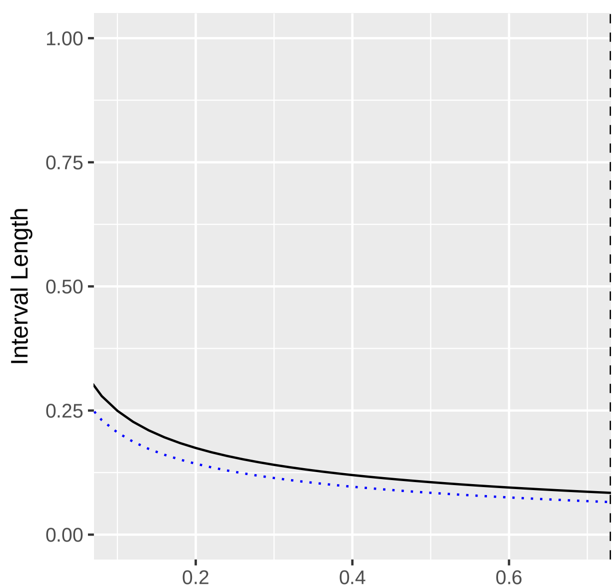
(a) $x = -2$



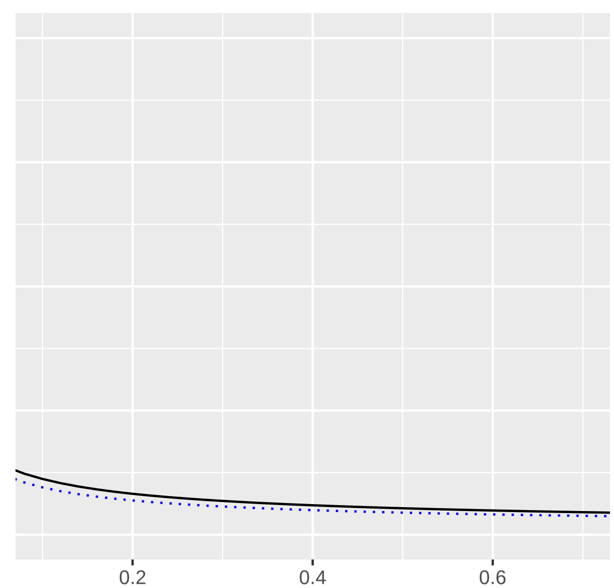
(b) $x = -1$



(c) $x = 0$

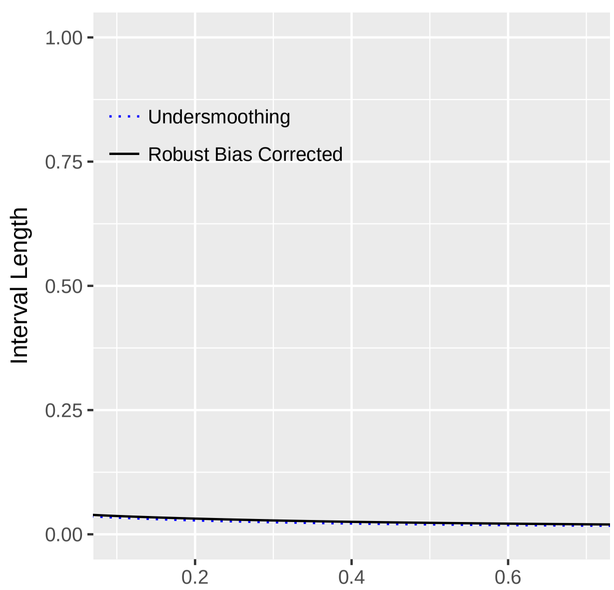


(d) $x = 1$

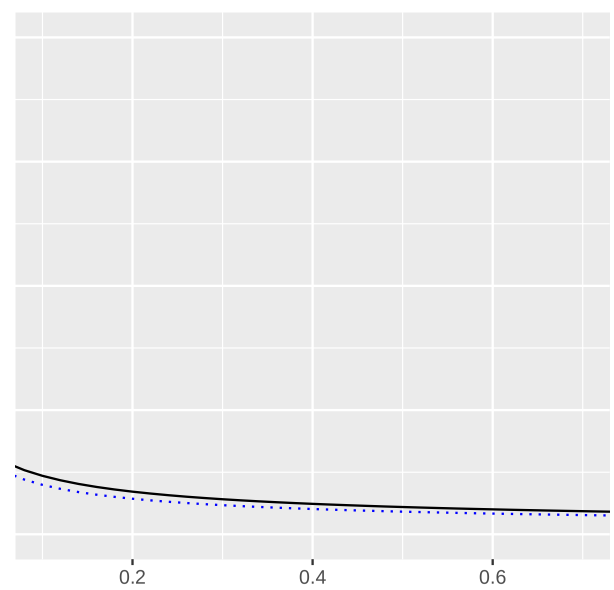


(e) $x = 2$

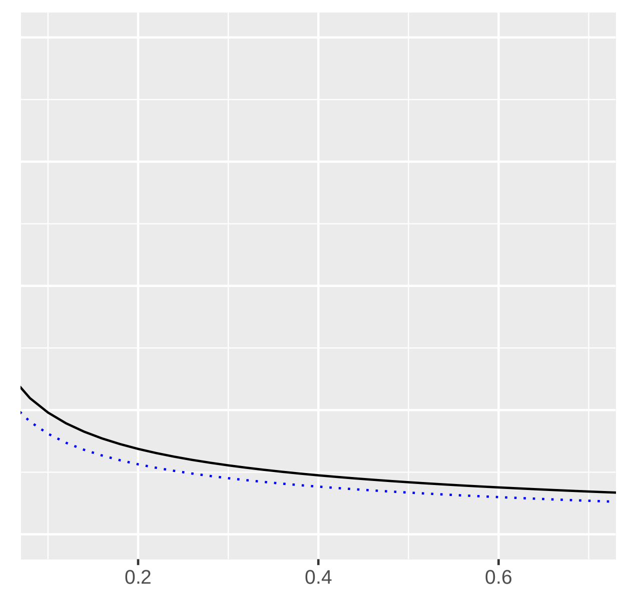
Figure S.I.7: Average Interval Length of 95% Confidence Intervals - Model 2



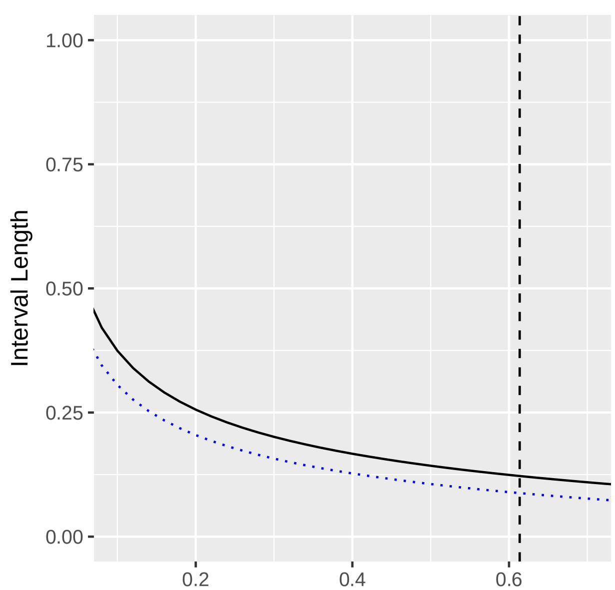
(a) $x = -2$



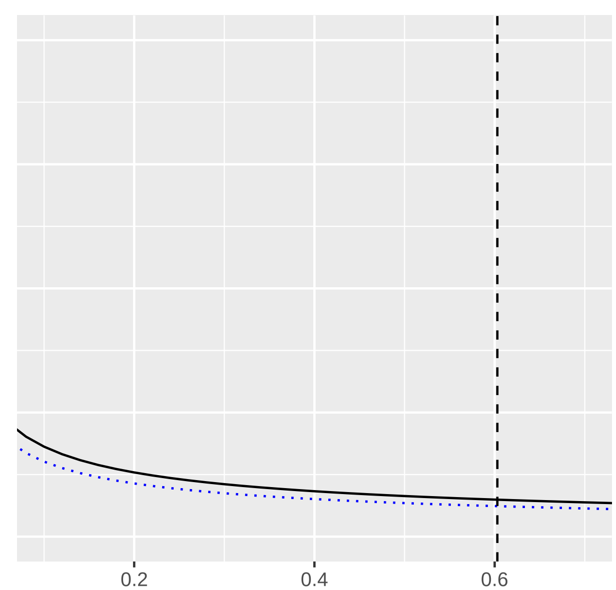
(b) $x = -1$



(c) $x = 0$

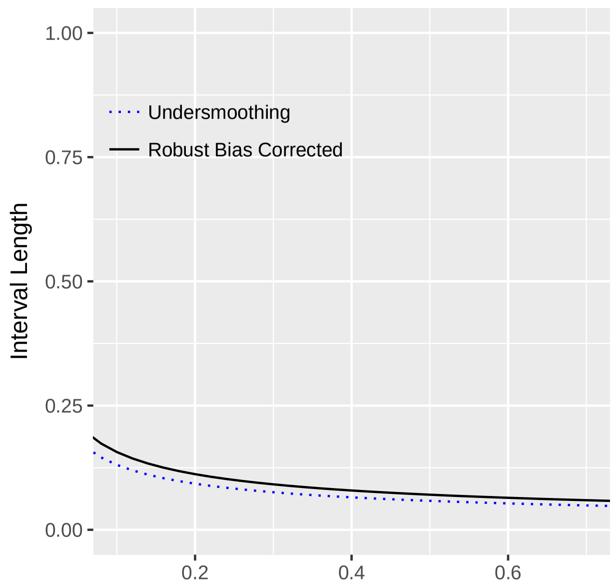


(d) $x = 1$

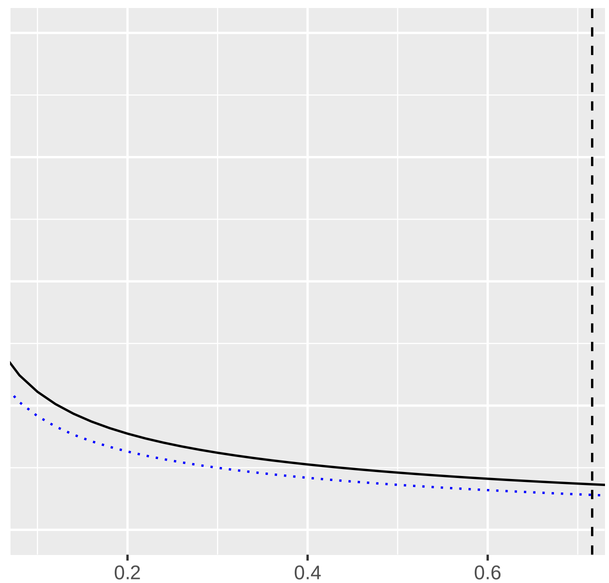


(e) $x = 2$

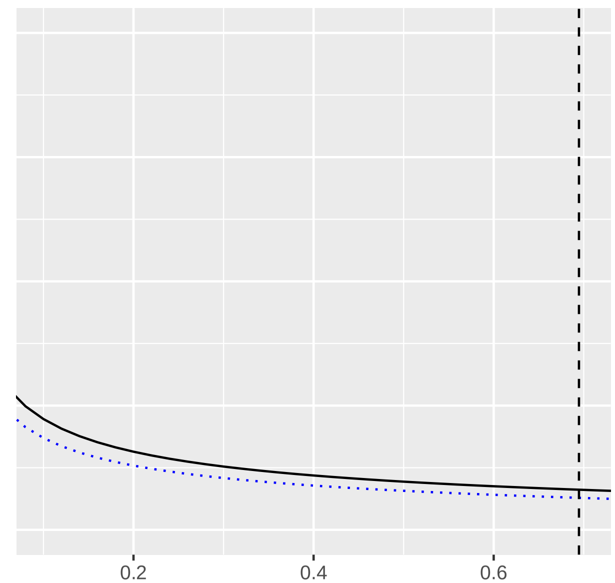
Figure S.I.8: Average Interval Length of 95% Confidence Intervals - Model 3



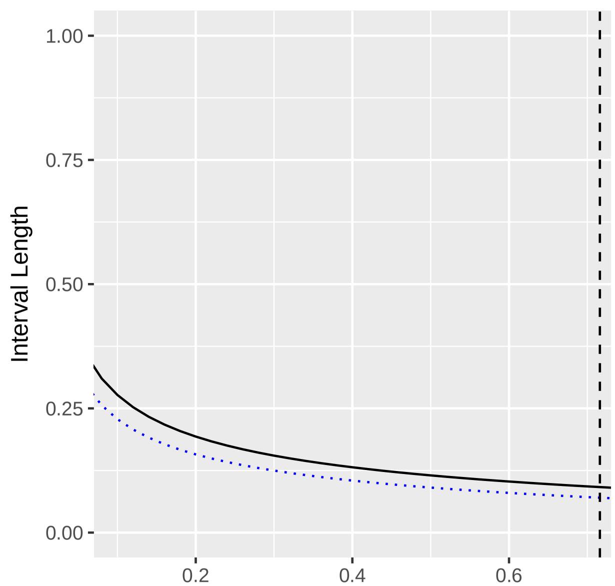
(a) $x = -2$



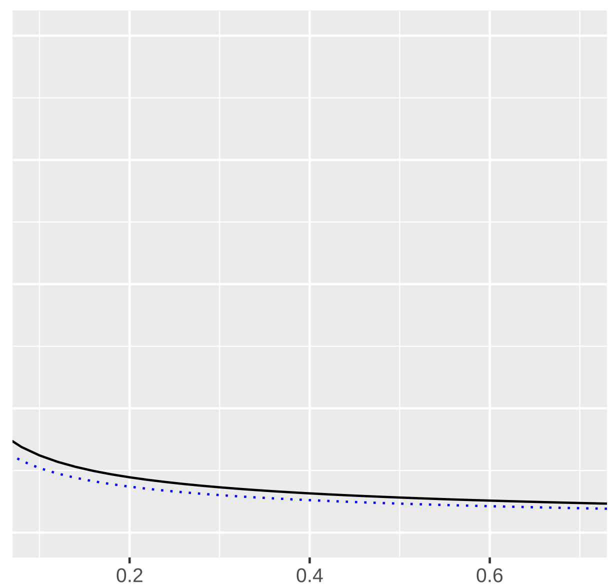
(b) $x = -1$



(c) $x = 0$

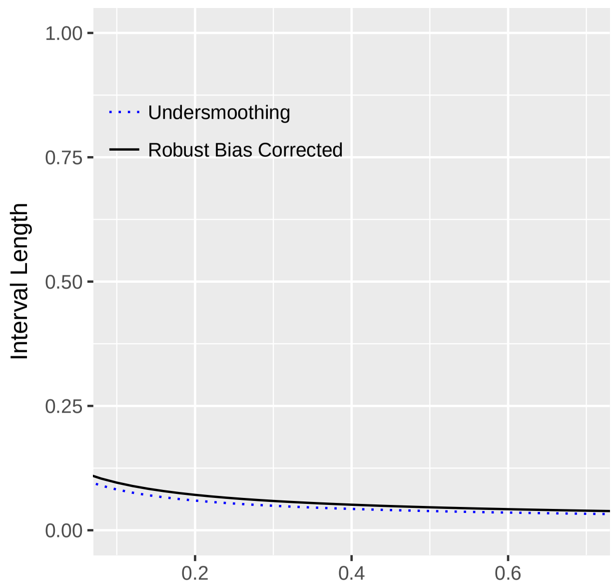


(d) $x = 1$

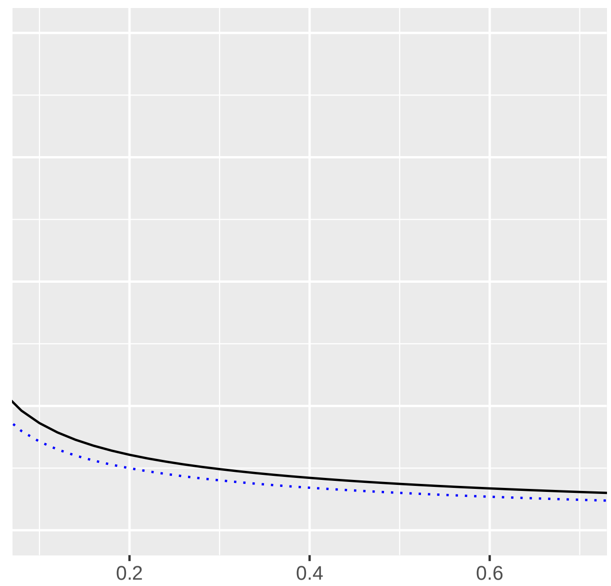


(e) $x = 2$

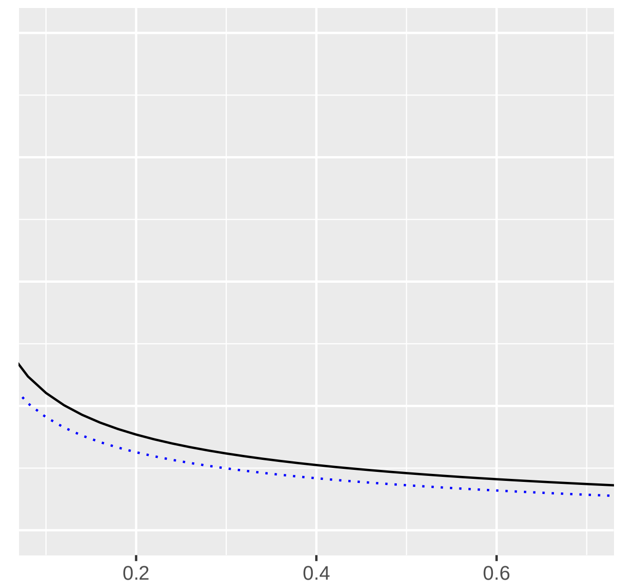
Figure S.I.9: Average Interval Length of 95% Confidence Intervals - Model 4



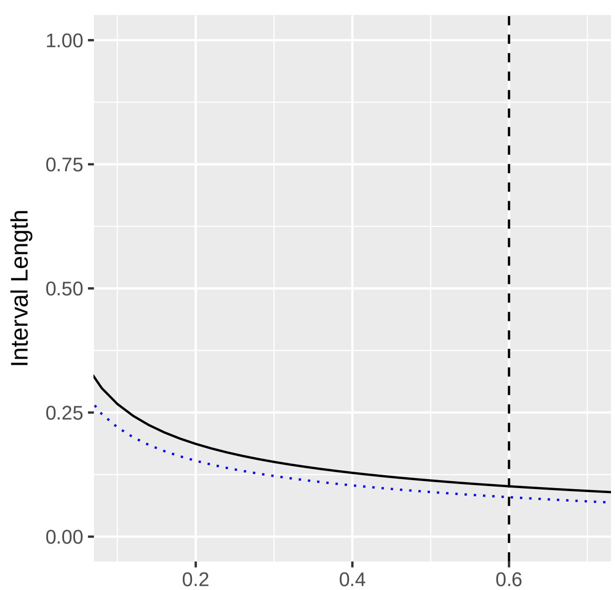
(a) $x = -2$



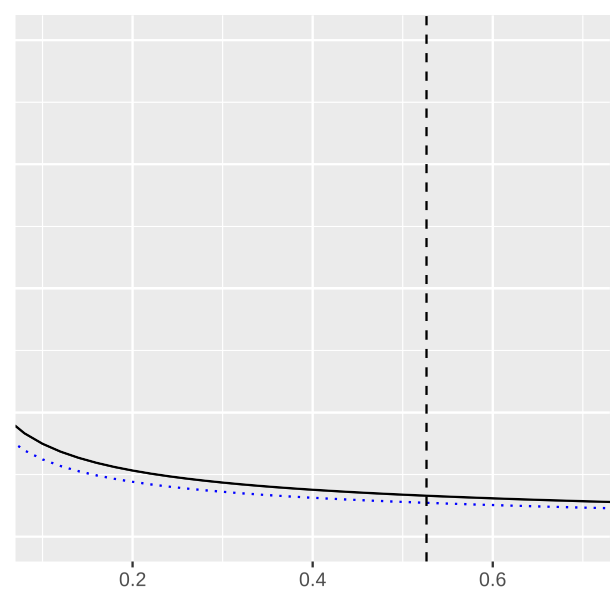
(b) $x = -1$



(c) $x = 0$

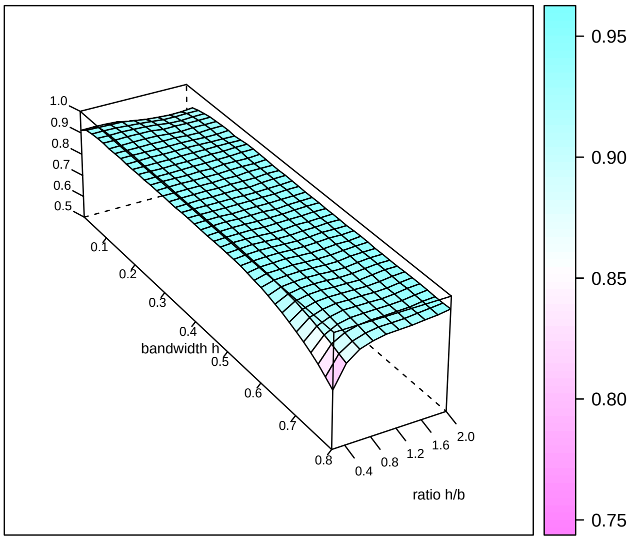


(d) $x = 1$

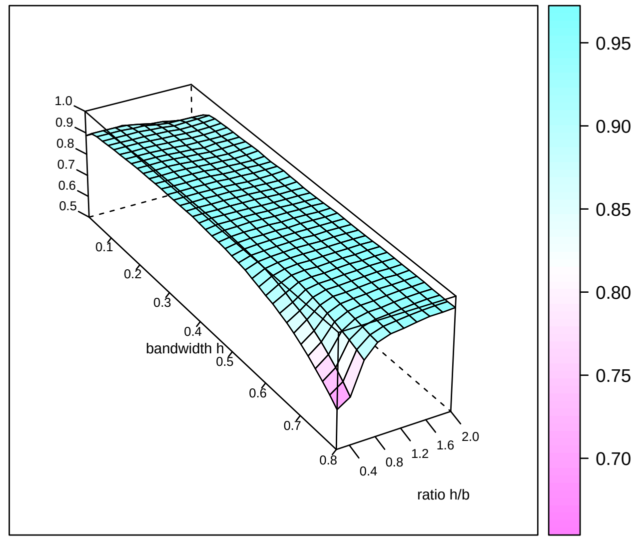


(e) $x = 2$

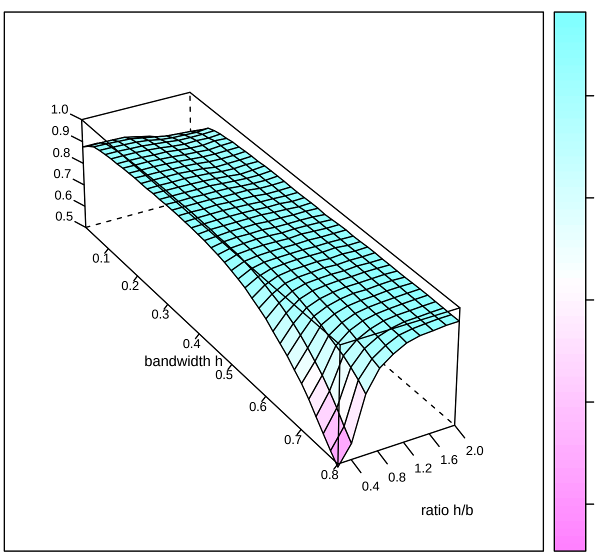
Figure S.I.10: Empirical Coverage of 95% Confidence Intervals ($x = 0$)



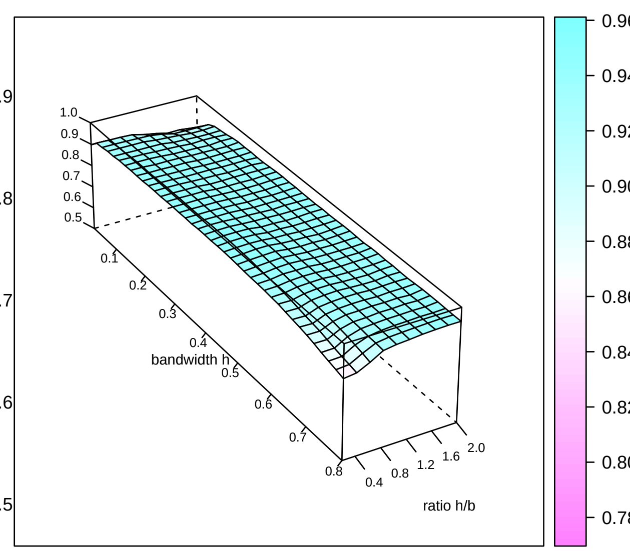
(a) Model 1



(b) Model 2

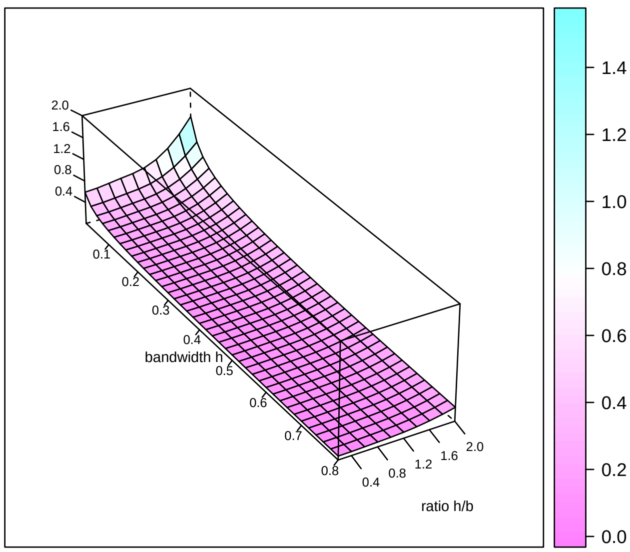


(c) Model 3

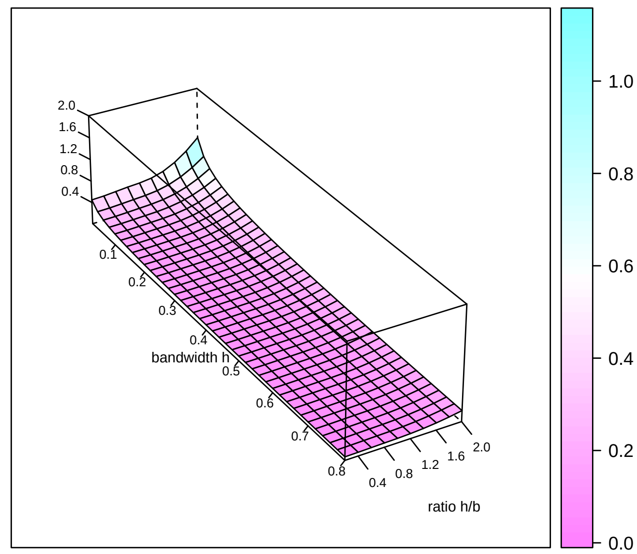


(d) Model 4

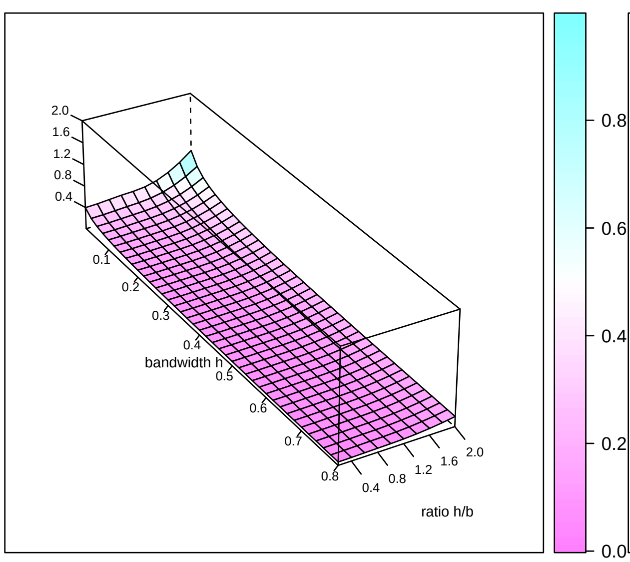
Figure S.I.11: Average Interval Length of 95% Confidence Intervals ($x = 0$)



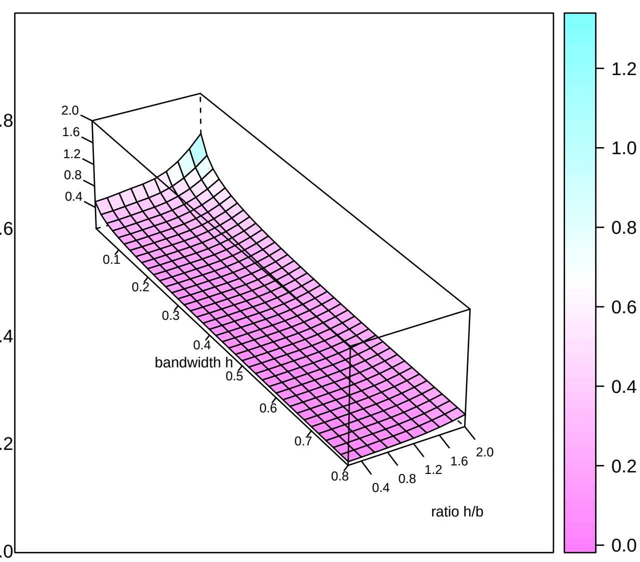
(a) Model 1



(b) Model 2



(c) Model 3



(d) Model 4

Part S.II

Local Polynomial Estimation and Inference

S.II.1 Notation

Local polynomial regression is notationally demanding, and the Edgeworth expansions will be substantially more so. For ease of reference, we collect all notation here regardless of where it is introduced and used. Much of the notation is fully restated later, when needed. As such, this subsection is designed more for reference, and is not easily readable.

Throughout, a subscript p will generally refer to a quantity used to estimate $m(x) = \mathbb{E}[Y_i|X_i = x]$, while a subscript q will refer to the bias correction portion (the vectors e_0 and e_{p+1} below are notable exceptions to this rule). Recall that $p \geq 1$ is odd and $q > p$ may be even or odd.

Throughout this section let $X_{h,i} = (X_i - x)/h$ and similarly for $X_{b,i}$. The evaluation point is implicit here.

To save notation, products of functions will be written together, with only one argument. For example

$$(Kr_p r_p')(X_{h,i}) := K(X_{h,i})r_p(X_{h,i})r_p(X_{h,i})' = K\left(\frac{X_i - x}{h}\right) r_p \left(\frac{X_i - x}{h}\right) r_p \left(\frac{X_i - x}{h}\right)',$$

and similarly for $(Kr_p)(X_{h,i})$, $(Lr_q)(X_{b,i})$, etc.

All expectations are fixed-n calculations. To give concrete examples of this notation ($\Lambda_{p,k}$, R_p, and W_p are redefined below):

$$\Lambda_{p,k} = R_p' W_p[((X_1 - x)/h)^{p+k}, \cdots, ((X_n - x)/h)^{p+k}]'/n = \frac{1}{nh}\sum_{i=1}^{n}(Kr_p)(X_{h,i})X_{h,i}^{p+k}$$

and

$$\tilde{\Lambda}_{p,k} = \mathbb{E}[\Lambda_{p,k}] = h^{-1}\mathbb{E}[(Kr_p)(X_{h,j})X_{h,i}^{p+k}] = h^{-1}\int_{\text{supp}\{X\}} K\left(\frac{X_i - x}{h}\right) r_p\left(\frac{X_i - x}{h}\right)\left(\frac{X_i - x}{h}\right)^{p+k} f(X_i)dX_i.$$

Here the range of integration is explicit, but in general it will not be. This is important for boundary issues, where the notation is generally unchanged, and it is to be understood that moments and moments of the kernel be replaced by the appropriate truncated version. Continuing this example, if $\text{supp}\{X\} = [0, \infty)$ and $x = 0$, then by a change of variables

$$\tilde{\Lambda}_{p,k} = h^{-1}\int_{\text{supp}\{X\}}(Kr_p)(X_{h,j})X_{h,i}^{p+k}f(X_i)dX_i = \int_0^{\infty}(Kr_p)(u)u^{p+k}f(-uh)du,$$

whereas if supp$\{X\} = (-\infty, 0]$ and $x = 0$, then

$$\tilde{\Lambda}_{p,k} = \int_{-\infty}^{0} (Kr_p)(u) u^{p+k} f(-uh) du.$$

For the remainder of this section, the notation is left generic.

For the proofs (Section S.II.6) we will frequently abbreviate $s = \sqrt{nh}$.

S.II.1.1 Estimators, Variances, and Studentized Statistics

To define the estimator $\hat{m}$ of m and the bias correction, begin by defining:

$$\begin{aligned}
&r_p(u) = \left(1, u, u^2, \ldots, u^p\right)', && R_p = \left[r_p(X_{h,1}), \cdots, r_p(X_{h,n})\right]', \\
&W_p = \operatorname{diag}\left(h^{-1}K(X_{h,i}) : i = 1, \ldots, n\right), && H_p = \operatorname{diag}\left(1, h^{-1}, h^{-2}, \ldots, h^{-p}\right), \\
&\Gamma_p = R_p' W_p R_p / n, \quad \text{and} && \Lambda_{p,k} = R_p' W_p \left[X_{h,1}^{p+k}, \cdots, X_{h,n}^{p+k}\right]' / n,
\end{aligned} \tag{S.II.1}$$

where diag$(a_i : i = 1, \ldots, n)$ denote the $n \times n$ diagonal matrix constructed using the elements $a_1, a_2, \cdots, a_n$. Note that in the main text $\Lambda_{p,1}$ is denoted by Λ_p.

Similarly, define

$$\begin{aligned}
&r_q(u) = \left(1, u, u^2, \ldots, u^q\right)', && R_q = \left[r_q(X_{b,1}), \cdots, r_q(X_{b,n})\right]', \\
&W_q = \operatorname{diag}\left(b^{-1}L(X_{b,i}) : i = 1, \ldots, n\right), && H_q = \operatorname{diag}\left(1, b^{-1}, b^{-2}, \ldots, b^{-q}\right), \\
&\Gamma_q = R_q' W_q R_q / n, \quad \text{and} && \Lambda_{q,k} = R_q' W_q \left[X_{b,1}^{q+k}, \cdots, X_{b,n}^{w+k}\right]' / n,
\end{aligned} \tag{S.II.2}$$

These are identical, but substituting q, L, and b in place of p, K, and h, respectively. Note that some dimensions change but other do not: for example, W_p and W_q are both $n \times n$, but Γ_p is $(p+1)$ square whereas Γ_q is $(q+1)$.

Denote by e_0 the $(p+1)$-vector with a one in the first position and zeros in the remaining and $Y = (Y_1, \cdots, Y_n)'$. The local polynomial estimator of $m(x) = \mathbb{E}[Y_i | X_i = x]$ is

$$\hat{m} = e_0' \hat{\beta}_p = e_0' H_p \Gamma_p^{-1} R_p' W_p Y / n,$$

where

$$\hat{\beta}_p = \underset{b \in \mathbb{R}^{p+1}}{\arg\min} \frac{1}{nh} \sum_{i=1}^{n} (Y_i - r_p(X_i - x)' b)^2 K\left(X_{h,i}\right) = H_p \Gamma_p^{-1} R_p' W_p Y / n.$$

If we define $\check{R} = [r_p(X_1 - x), \cdots, r_p(X_n - x)]'$ and $M = [m(X_1), \ldots, m(X_n)]'$, then we can split $\hat{m} - m$ into the variance and bias terms

$$\hat{m} - m = e_0' \Gamma_p^{-1} R_p' W_p (Y - M)/n + e_0' \Gamma_p^{-1} R_p' W_p (M - \check{R}\beta_p)/n.$$

This will be useful in the course of the proofs.

The conditional bias is given by

$$\mathbb{E}[\hat{m}|X_1,\dots,X_n] - m = h^{p+1}m^{(p+1)}\frac{1}{(p+1)!}e_0'\Gamma_p^{-1}\Lambda_{p,1} + o_p(h^{p+1}). \tag{S.II.3}$$

(Recall that in the main paper, $\Lambda_{p,1}$ is denoted Λ_p.) This result is valid for p odd, our main focus, but also for p even at boundary points.

Denote by e_{p+1} the $(q+1)$-vector with one in the $p+2$ position, and zeros in the rest. Then we estimate the bias as

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

The bias corrected estimator can then be written

$$\begin{aligned}\hat{m} - \hat{B}_m &= e_0'H_p\Gamma_p^{-1}R_p'W_pY/n - h^{p+1}e_0'\Gamma_p^{-1}\Lambda_{p,1}e_{p+1}'H_q\Gamma_q^{-1}R_q'W_qY/n \\ &= e_0'\Gamma_p^{-1}\left(R_p'W_p - \rho^{p+1}\Lambda_{p,1}e_{p+1}'\Gamma_q^{-1}R_q'W_q\right)Y/n,\end{aligned}$$

using the fact that $e_{p+1}'H_q = b^{p+1}e_{p+1}'$.

The fixed-n variances are

$$\sigma^2_{\texttt{us}} := (nh)\mathbb{V}[\hat{m}|X_1,\cdots,X_n] = e_0'\Gamma_p^{-1}\left(hR_p'W_p\Sigma W_pR_p/n\right)\Gamma_p^{-1}e_0 \tag{S.II.4}$$

and

$$\begin{aligned}\sigma^2_{\texttt{rbc}} &:= (nh)V[\hat{m} - \hat{B}_m|X_1,\dots,X_n] \\ &= e_0'\Gamma_p^{-1}(h/n)\left(R_p'W_p - \rho^{p+1}\Lambda_{p,1}e_{p+1}'\Gamma_q^{-1}R_q'W_q\right)\Sigma\left(R_p'W_p - \rho^{p+1}\Lambda_{p,1}e_{p+1}'\Gamma_q^{-1}R_q'W_q\right)'\Gamma_p^{-1}e_0,\end{aligned} \tag{S.II.5}$$

where

$$\Sigma = \operatorname{diag}(v(X_i) : i = 1,\dots,n), \qquad \text{with} \qquad v(x) = \mathbb{V}[Y|X=x].$$

These are the closest analogue to the density case, but are still random due to the conditioning on the covariates. Their respective estimators are

$$\hat{\sigma}^2_{\texttt{us}} = e_0'\Gamma_p^{-1}\left(hR_p'W_p\hat{\Sigma}_pW_pR_p\Gamma_p^{-1}/n\right)e_0$$

and

$$\hat{\sigma}^2_{\texttt{rbc}} = e_0'\Gamma_p^{-1}(h/n)\left(R_p'W_p - \rho^{p+1}\Lambda_{p,1}e_{p+1}'\Gamma_q^{-1}R_q'W_q\right)\hat{\Sigma}_q\left(R_p'W_p - \rho^{p+1}\Lambda_{p,1}e_{p+1}'\Gamma_q^{-1}R_q'W_q\right)'\Gamma_p^{-1}e_0.$$

The conditional variance matrixes are estimated as

$$\hat{\Sigma}_p = \operatorname{diag}(\hat{v}(X_i) : i = 1, \ldots, n), \qquad \text{with} \qquad \hat{v}(X_i) = (Y_i - r_p(X_i - x)'\hat{\beta}_p)^2,$$

and

$$\hat{\Sigma}_q = \operatorname{diag}(\hat{v}(X_i) : i = 1, \ldots, n), \qquad \text{with} \qquad \hat{v}(X_i) = (Y_i - r_q(X_i - x)'\hat{\beta}_q)^2.$$

The Studentized statistics of interest are then:

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

The main result of this section is an Edgeworth expansion of the distribution function of these statistics.

S.II.1.2 Edgeworth Expansion Terms

The terms of the Edgeworth expansion require further notation and discussion. The expressions are not nearly as compact as in the density case (cf. Section S.I.6).

Define the expectations of Γ_p, Γ_q, $\Lambda_{p,k}$, and $\Lambda_{q,k}$ as $\tilde{\Gamma}_p$, $\tilde{\Gamma}_q$, $\tilde{\Lambda}_{p,k}$, and $\tilde{\Lambda}_{q,k}$, such as

$$\tilde{\Gamma}_p = \mathbb{E}\left[\Gamma_p\right] = \mathbb{E}\left[h^{-1}(K r_p r_p')(X_{h,i})\right].$$

These will be used to define nonrandom biases and variances that appear in the expansions.

The biases are defined in Eqn. (S.II.7), and are given by

$$\begin{aligned}
\eta_{\mathtt{us}} &= \sqrt{nh} \int e_0' \tilde{\Gamma}_p^{-1} K(u) r_p(u) \left(m(x - uh) - r_p(uh)'\beta_p\right) f(x - uh) du, \\
\eta_{\mathtt{bc}} &= \sqrt{nh} \int e_0' \tilde{\Gamma}_p^{-1} K(u) r_p(u) \left(m(x - uh) - r_{p+1}(uh)'\beta_{p+1}\right) f(x - uh) du \\
&\quad - \sqrt{nh}\rho^{p+1} \int e_0' \tilde{\Gamma}_p^{-1} \tilde{\Lambda}_{p,1} e_{p+1}' \tilde{\Gamma}_q^{-1} L(u) r_q(u) \left(m(x - ub) - r_q(ub)'\beta_q\right) f(x - ub) du.
\end{aligned}$$

Further discussion and leading terms are found in Section S.II.4.

The fixed-n variances are computed conditionally, and we must replace them with their nonrandom analogues (just as $\eta_{\mathtt{us}}$ and $\eta_{\mathtt{bc}}$ must be nonrandom). Recalling Equations (S.II.4) and (S.II.5), define

$$\tilde{\sigma}_{\mathtt{us}}^2 := e_0' \tilde{\Gamma}_p^{-1} \tilde{\Psi}_p \tilde{\Gamma}_p^{-1} e_0,$$

where

$$\tilde{\Psi}_p = \mathbb{E}\left[\check{\Psi}_p\right] \qquad \text{and} \qquad \check{\Psi}_p := h R_p' W_p \Sigma W_p R_p / n,$$

and

$$\tilde{\sigma}^2_{\texttt{rbc}} := e_0'\tilde{\Gamma}_p^{-1}\tilde{\Psi}_q\tilde{\Gamma}_p^{-1}e_0$$

where

$$\tilde{\Psi}_q = \mathbb{E}\left[\check{\Psi}_q\right] \quad \text{and} \quad \check{\Psi}_q := h\left(R_p'W_p - \rho^{p+1}\tilde{\Lambda}_{p,1}\tilde{\Gamma}_q^{-1}R_q'W_q\right)\Sigma\left(R_p'W_p/n - \rho^{p+1}\tilde{\Lambda}_{p,1}\tilde{\Gamma}_q^{-1}R_q'W_q/n\right)'.$$

In the course of the proofs, we will also use $\hat{\Psi}_p = hR_p'W_p\hat{\Sigma}_pW_pR_p/n$ and the analogously-defined $\hat{\Psi}_q$.

We now give the precise forms of the polynomials in the Edgeworth expansion. As with the density, there will be both even and odd polynomials. These are not as compact or simple as the density case. Further, we will not attempt to simplify these functions by making use of limiting versions of moments. For example, we will *not* replace $\tilde{\Lambda}_{p,1}$ by $f(x)\int(Kr_p)(u)u^{p+1}du$, and similarly for other pieces. The only simplification made will be the use of $q_{k,\texttt{us}}(z)$ in the expansion for $T_{\texttt{bc}}$, which otherwise would require further notation than what is below (along the lines of $p_{1,\texttt{us}}(z)$ below).

First, define the following functions, which depend on n, p, q, h, b, K and L, but this is generally suppressed:

$$\begin{aligned}
\ell^0_{\texttt{us}}(X_i) &= e_0'\tilde{\Gamma}_p^{-1}(Kr_p)(X_{h,i});\\
\ell^0_{\texttt{bc}}(X_i) &= \ell^0_{\texttt{us}}(X_i) - \rho^{p+1}e_0'\tilde{\Gamma}_p^{-1}\tilde{\Lambda}_{p,1}e_{p+1}'\tilde{\Gamma}_q^{-1}(Lr_q)(X_{b,i});\\
\ell^1_{\texttt{us}}(X_i,X_j) &= e_0'\tilde{\Gamma}_p^{-1}\left(\mathbb{E}[(Kr_pr_p')(X_{h,j})] - (Kr_pr_p')(X_{h,j})\right)\tilde{\Gamma}_p^{-1}(Kr_p)(X_{h,i});\\
\ell^1_{\texttt{bc}}(X_i,X_j) &= \ell^1_{\texttt{us}}(X_i,X_j) - \rho^{p+1}e_0'\tilde{\Gamma}_p^{-1}\Big\{\left(\mathbb{E}[(Kr_pr_p')(X_{h,j})] - (Kr_pr_p')(X_{h,j})\right)\tilde{\Gamma}_p^{-1}\tilde{\Lambda}_{p,1}e_{p+1}'\\
&\qquad + \left((Kr_p)(X_{h,j})X_{h,i}^{p+1} - \mathbb{E}[(Kr_p)(X_{h,j})X_{h,i}^{p+1}]\right)e_{p+1}'\\
&\qquad + \tilde{\Lambda}_{p,1}e_{p+1}'\tilde{\Gamma}_q^{-1}\left(\mathbb{E}[(Lr_qr_q')(X_{b,j})] - (Lr_qr_q')(X_{b,j})\right)\Big\}\tilde{\Gamma}_q^{-1}(Lr_q)(X_{b,i}).
\end{aligned}$$

With this notation, we can write

$$\begin{aligned}
\tilde{\sigma}^2_{\texttt{us}} &= \mathbb{E}[h^{-1}\ell^0_{\texttt{us}}(X)^2v(X)],\\
\tilde{\sigma}^2_{\texttt{rbc}} &= \mathbb{E}[h^{-1}\ell^0_{\texttt{bc}}(X)^2v(X)],\\
\eta_{\texttt{us}} &= s\mathbb{E}\left[h^{-1}\ell^0_{\texttt{us}}(X_i)[m(X_i) - r_p(X_i-x)'\beta_p]\right],
\end{aligned}$$

and

$$\begin{aligned}
\eta_{\texttt{bc}} = s\mathbb{E}\Big[&h^{-1}\ell^0_{\texttt{us}}(X_i)[m(X_i) - r_{p+1}(X_i-x)'\beta_{p+1}]\\
&+ h^{-1}\left(\ell^0_{\texttt{bc}}(X_i) - \ell^0_{\texttt{us}}(X_i)\right)[m(X_i) - r_q(X_i-x)'\beta_q]\Big].
\end{aligned}$$

We will define the Edgeworth expansion polynomials first for the undersmoothing case. The standard Normal density is $\phi(z)$. First, the even polynomials are

$$p_{1,\mathtt{us}}(z) = \phi(z)\tilde{\sigma}_{\mathtt{us}}^{-3}\mathbb{E}\left[h^{-1}\ell_{\mathtt{us}}^{0}(X_i)^3\varepsilon_i^3\right]\left\{(2z^2-1)/6\right\}$$

and

$$p_{3,\mathtt{us}}(z) = -\phi(z)\tilde{\sigma}_{\mathtt{us}}^{-1}.$$

The absence of $p^{(2)}(z)$ is noteworthy: there is no version of this term for local polynomial estimation, because ε_i is conditionally mean zero.

Next, the odd polynomials for undersmoothing are defined as follows:

$$\begin{aligned}
q_{1,\mathtt{us}}(z) = {} & \phi(z)\tilde{\sigma}_{\mathtt{us}}^{-6}\mathbb{E}\left[h^{-1}\ell_{\mathtt{us}}^{0}(X_i)^3\varepsilon_i^3\right]^2\left\{z^3/3+7z/4+\tilde{\sigma}_{\mathtt{us}}^{2}z(z^2-3)/4\right\} \\
& + \phi(z)\tilde{\sigma}_{\mathtt{us}}^{-2}\mathbb{E}\left[h^{-1}\ell_{\mathtt{us}}^{0}(X_i)\ell_{\mathtt{us}}^{1}(X_i,X_i)\varepsilon_i^2\right]\left\{-z(z^2-3)/2\right\} \\
& + \phi(z)\tilde{\sigma}_{\mathtt{us}}^{-4}\mathbb{E}\left[h^{-1}\ell_{\mathtt{us}}^{0}(X_i)^4(\varepsilon_i^4-v(X_i)^2)\right]\left\{z(z^2-3)/8\right\} \\
& - \phi(z)\tilde{\sigma}_{\mathtt{us}}^{-2}\mathbb{E}\left[h^{-1}\ell_{\mathtt{us}}^{0}(X_i)^2r_p(X_{h,i})'\tilde{\Gamma}_p^{-1}(Kr_p)(X_{h,i})\varepsilon_i^2\right]\left\{z(z^2-1)/2\right\} \\
& - \phi(z)\tilde{\sigma}_{\mathtt{us}}^{-4}\mathbb{E}\left[h^{-1}\ell_{\mathtt{us}}^{0}(X_i)^3r_p(X_{h,i})'\tilde{\Gamma}_p^{-1}\varepsilon_i^2\right]\mathbb{E}\left[h^{-1}(Kr_p)(X_{h,i})\ell_{\mathtt{us}}^{0}(X_i)\varepsilon_i^2\right]\left\{z(z^2-1)\right\} \\
& + \phi(z)\tilde{\sigma}_{\mathtt{us}}^{-2}\mathbb{E}\left[h^{-2}\ell_{\mathtt{us}}^{0}(X_i)^2(r_p(X_{h,i})'\tilde{\Gamma}_p^{-1}(Kr_p)(X_{h,j}))^2\varepsilon_j^2\right]\left\{z(z^2-1)/4\right\} \\
& + \phi(z)\tilde{\sigma}_{\mathtt{us}}^{-4}\mathbb{E}\left[h^{-3}\ell_{\mathtt{us}}^{0}(X_j)^2r_p(X_{h,j})'\tilde{\Gamma}_p^{-1}(Kr_p)(X_{h,i})\ell_{\mathtt{us}}^{0}(X_i)r_p(X_{h,j})'\tilde{\Gamma}_p^{-1}(Kr_p)(X_{h,k})\ell_{\mathtt{us}}^{0}(X_k)\varepsilon_i^2\varepsilon_k^2\right] \\
& \qquad \times \left\{z(z^2-1)/2\right\} \\
& + \phi(z)\tilde{\sigma}_{\mathtt{us}}^{-4}\mathbb{E}\left[h^{-1}\ell_{\mathtt{us}}^{0}(X_i)^4\varepsilon_i^4\right]\left\{-z(z^2-3)/24\right\} \\
& + \phi(z)\tilde{\sigma}_{\mathtt{us}}^{-4}\mathbb{E}\left[h^{-1}\left(\ell_{\mathtt{us}}^{0}(X_i)^2v(X_i)-\mathbb{E}[\ell_{\mathtt{us}}^{0}(X_i)^2v(X_i)]\right)\ell_{\mathtt{us}}^{0}(X_i)^2\varepsilon_i^2\right]\left\{z(z^2-1)/4\right\} \\
& + \phi(z)\tilde{\sigma}_{\mathtt{us}}^{-4}\mathbb{E}\left[h^{-2}\ell_{\mathtt{us}}^{1}(X_i,X_j)\ell_{\mathtt{us}}^{0}(X_i)\ell_{\mathtt{us}}^{0}(X_j)^2\varepsilon_j^2v(X_i)\right]\left\{z(z^2-3)\right\} \\
& + \phi(z)\tilde{\sigma}_{\mathtt{us}}^{-4}\mathbb{E}\left[h^{-2}\ell_{\mathtt{us}}^{1}(X_i,X_j)\ell_{\mathtt{us}}^{0}(X_i)\left(\ell_{\mathtt{us}}^{0}(X_j)^2v(X_j)-\mathbb{E}[\ell_{\mathtt{us}}^{0}(X_j)^2v(X_j)]\right)\varepsilon_i^2\right]\left\{-z\right\} \\
& + \phi(z)\tilde{\sigma}_{\mathtt{us}}^{-4}\mathbb{E}\left[h^{-1}\left(\ell_{\mathtt{us}}^{0}(X_i)^2v(X_i)-\mathbb{E}[\ell_{\mathtt{us}}^{0}(X_i)^2v(X_i)]\right)^2\right]\left\{-z(z^2+1)/8\right\};
\end{aligned}$$

$$q_{2,\mathtt{us}}(z) = -\phi(z)\tilde{\sigma}_{\mathtt{us}}^{-2}z/2;$$

$$q_{3,\mathtt{us}}(z) = \phi(z)\tilde{\sigma}_{\mathtt{us}}^{-4}\mathbb{E}[h^{-1}\ell_{\mathtt{us}}^{0}(X_i)^3\varepsilon_i^3](z^3/3).$$

For robust bias correction, both the even polynomials, $p_{1,\mathtt{rbc}}(z)$ and $p_{3,\mathtt{rbc}}(z)$, and the odd polynomials, $q_{1,\mathtt{rbc}}(z)$, $q_{2,\mathtt{rbc}}(z)$, and $q_{3,\mathtt{rbc}}(z)$ are defined in the exact same way, but changing the $\tilde{\sigma}_{\mathtt{us}}$ to $\tilde{\sigma}_{\mathtt{rbc}}$, $\ell_{\mathtt{us}}^{k}(\cdot)$ to $\ell_{\mathtt{bc}}^{k}(\cdot)$, K to L, and p to q, and so forth. For $q_{1,\mathtt{us}}(z)$ and $q_{1,\mathtt{rbc}}(z)$, the seventh term can be rewritten by rearranging the terms and factoring the expectation, as follows:

$$\begin{aligned}
\mathbb{E}&\Big[h^{-3}\ell^0_{\texttt{us}}(X_j)^2 r_p(X_{h,j})'\tilde{\Gamma}_p^{-1}(Kr_p)(X_{h,i})\ell^0_{\texttt{us}}(X_i) r_p(X_{h,j})'\tilde{\Gamma}_p^{-1}(Kr_p)(X_{h,k})\ell^0_{\texttt{us}}(X_k)\varepsilon_i^2\varepsilon_k^2\Big] \\
&= \mathbb{E}\Big[h^{-1}\ell^0_{\texttt{us}}(X_i)\varepsilon_i^2(Kr_p')(X_{h,i})\tilde{\Gamma}_p^{-1}\Big]\,\mathbb{E}\Big[h^{-1}\ell^0_{\texttt{us}}(X_j)^2 r_p(X_{h,j})r_p(X_{h,j})'\tilde{\Gamma}_p^{-1}\Big] \\
&\qquad\qquad \times\ \mathbb{E}\left[h^{-1}(Kr_p)(X_{h,k})\ell^0_{\texttt{us}}(X_k)\varepsilon_k^2\right]
\end{aligned} \tag{S.II.6}$$

The polynomials defined here are for *distribution function* expansions, and are different from those used for *coverage error*. The polynomials $q_{1,\texttt{us}}$, $q_{2,\texttt{us}}$, and $q_{3,\texttt{us}}$ and $q_{1,\texttt{rbc}}$, $q_{2,\texttt{rbc}}$, and $q_{3,\texttt{rbc}}$, which do *not* have an argument, used for *coverage error* in the main text and in Corollary S.II.1 below, are defined in terms of those given above, which *do* have an argument. Specifically, the polynomials above should be doubled, divided by the standard Normal density, and evaluated at the Normal quantile $z_{\alpha/2}$, that is,

$$q_{k,\bullet} := \frac{2}{\phi(z)} q_{k,\bullet}(z)\bigg|_{z=z_{\alpha/2}}, \qquad\qquad k=1,2,3, \quad \bullet = \texttt{us},\texttt{rbc}$$

For traditional bias correction, $q_{1,\texttt{us}}(z)$, $q_{2,\texttt{us}}(z)$, and $q_{3,\texttt{us}}(z)$ are used, but such simplification can not be done for $p_{1,\texttt{bc}}(z)$ and $p_{3,\texttt{bc}}(z)$, which must be defined as

$$\begin{aligned}
p_{1,\texttt{bc}}(z) &= \phi(z)\tilde{\sigma}_{\texttt{us}}^{-3}\Big(\mathbb{E}\left[h^{-1}\ell^0_{\texttt{us}}(X_i)^3\varepsilon_i^3\right]\left\{-(z^2-1)/6\right\} + \mathbb{E}\left[h^{-1}\ell^0_{\texttt{us}}(X_i)^2\ell^0_{\texttt{bc}}(X_i)\varepsilon_i^3\right]\left\{-(z^2-3)/4\right\}\Big) \\
&\quad + \phi(z)\tilde{\sigma}_{\texttt{us}}^{2}\tilde{\sigma}_{\texttt{rbc}}^{-5}\mathbb{E}\left[h^{-1}\ell^0_{\texttt{us}}(X_i)^2\ell^0_{\texttt{bc}}(X_i)\varepsilon_i^3\right]\left\{3(z^2-1)/4\right\}
\end{aligned}$$

and

$$p_{3,\texttt{bc}}(z) = -\phi(z)\tilde{\sigma}_{\texttt{us}}^{-1}.$$

Lastly, traditional bias correction also exhibits additional terms in the expansion (see discussion in the main text) representing the covariance of $\hat{m}$ and $\hat{B}_m$ (denoted by $\Omega_{1,\texttt{bc}}$) and the variance of $\hat{B}_m$ ($\Omega_{2,\texttt{bc}}$). We now state their precise forms. These arise from the mismatch between the variance of the numerator of $T_{\texttt{bc}}$ and the standardization used, $\sigma^2_{\texttt{us}}$, but these are random, and so $\Omega_{1\texttt{bc}}$ and $\Omega_{2,\texttt{bc}}$ must be derived from the nonrandom versions, $\tilde{\sigma}^2_{\texttt{rbc}}$ and $\tilde{\sigma}^2_{\texttt{us}}$ (cf. Section S.I.6; for the same reason $\eta_{\texttt{us}}$ and $\eta_{\texttt{bc}}$ must be nonrandom). Recalling the definitions above,

$$\begin{aligned}
\frac{\tilde{\sigma}^2_{\texttt{rbc}}}{\tilde{\sigma}^2_{\texttt{us}}} &= \frac{\mathbb{E}[h^{-1}\ell^0_{\texttt{bc}}(X)^2 v(X)]}{\mathbb{E}[h^{-1}\ell^0_{\texttt{us}}(X)^2 v(X)]} \\
&= \frac{\mathbb{E}[h^{-1}\{\ell^0_{\texttt{us}}(X) + (\ell^0_{\texttt{bc}}(X) - \ell^0_{\texttt{us}}(X))\}^2 v(X)]}{\mathbb{E}[h^{-1}\ell^0_{\texttt{us}}(X)^2 v(X)]} \\
&= 1 - 2\tilde{\sigma}_{\texttt{us}}^{-2}\mathbb{E}[h^{-1}\{\ell^0_{\texttt{us}}(X)(\ell^0_{\texttt{bc}}(X) - \ell^0_{\texttt{us}}(X))\}v(X)] + \tilde{\sigma}_{\texttt{us}}^{-2}\mathbb{E}[h^{-1}\{(\ell^0_{\texttt{bc}}(X) - \ell^0_{\texttt{us}}(X))\}^2 v(X)] \\
&= 1 - 2\rho^{1+(p+1)}\tilde{\sigma}_{\texttt{us}}^{-2}\mathbb{E}[h^{-1}\{\rho^{-p-2}\ell^0_{\texttt{us}}(X)(\ell^0_{\texttt{bc}}(X) - \ell^0_{\texttt{us}}(X))\}v(X)] \\
&\qquad + \rho^{1+2(p+1)}\tilde{\sigma}_{\texttt{us}}^{-2}\mathbb{E}[b^{-1}\{\rho^{-p-2}(\ell^0_{\texttt{bc}}(X) - \ell^0_{\texttt{us}}(X))\}^2 v(X)]
\end{aligned}$$

Therefore

$$\Omega_{1,\texttt{bc}} = -2\tilde{\sigma}_{\texttt{us}}^{-2}\mathbb{E}[h^{-1}\{\rho^{-p-2}\ell^0_{\texttt{us}}(X)(\ell^0_{\texttt{bc}}(X) - \ell^0_{\texttt{us}}(X))\}v(X)]$$

and

$$\Omega_{2,\texttt{bc}} = \tilde{\sigma}_{\texttt{us}}^{-2}\mathbb{E}[b^{-1}\{\rho^{-p-2}(\ell^0_{\texttt{bc}}(X) - \ell^0_{\texttt{us}}(X))\}^2 v(X)].$$

Remark S.II.1 (Simplifications)**.** It is possible for the above-defined polynomials to simplify in special cases. A leading example is in the homoskedastic Gaussian regression model:

$$Y_i = m(X_i) + \varepsilon_i, \qquad \text{where} \qquad \varepsilon_i \sim \mathcal{N}(0, v).$$

This model is a common theoretical baseline to study, though over-simplified from an empirical point of view. In this special case, $\mathbb{E}[\varepsilon_i^3] = 0$ and thus $q_{3,\texttt{us}}(z) \equiv 0$, entirely removing this term from the Edgeworth expansions. This has little bearing on the conceptual conclusions however, and in particular the comparison of undersmoothing and robust bias correction. ■

S.II.2 Details of Practical Implementation

In the main text we give a direct plug-in (DPI) rule to implement the coverage-error optimal bandwidth. Here we we give complete details for this procedure as well as document a second practical choice, based on a rule-of-thumb (ROT) strategy. Both choices yield the optimal coverage error decay rate at interior and boundary points.

All our methods are implemented in R and STATA via the `nprobust` package, available from http://sites.google.com/site/nppackages/nprobust (see also http://cran.r-project.org/package=nprobust). See Calonico et al. (2017) for a complete description.

As in the density case, the MSE-optimal bandwidth undercovers when used in the undersmoothing confidence interval; that is, Remark S.I.1 applies directly. See also Hall and Horowitz (2013).

S.II.2.1 Bandwidth Choice: Rule-of-Thumb (ROT)

As with the density case, a simple rule-of-thumb based on rescaling the MSE-optimal bandwidth is:

$$\hat{h}^{\texttt{int}}_{\texttt{rot}} = \hat{h}^{\texttt{int}}_{\texttt{mse}}\, n^{-(p-1)/((2p+3)(p+4))} \qquad \text{and} \qquad \hat{h}^{\texttt{bnd}}_{\texttt{rot}} = \hat{h}^{\texttt{bnd}}_{\texttt{mse}}\, n^{-p/((2p+3)(p+3))}.$$

where $\hat{h}^{\texttt{int}}_{\texttt{mse}}$ and $\hat{h}^{\texttt{bnd}}_{\texttt{mse}}$ denote readily-available implementations of the MSE-optimal bandwidth for interior and boundary points, respectively. See, e.g., Fan and Gijbels (1996). Again, when $p = 1$ in the interior, no scaling is needed ($\hat{h}^{\texttt{int}}_{\texttt{rot}} = \hat{h}^{\texttt{int}}_{\texttt{mse}}$), but for $p > 1$ any data-driven MSE-optimal bandwidth should always be shrunk to improve inference at the boundary (i.e., reduce coverage errors of the robust bias-corrected confidence intervals).

The ROT selector may be especially attractive for simplicity, if estimating the constants described below in the DPI case is prohibitive.

Remark S.I.2 applies to this case as well, though less transparently and without consequences that are as dramatic.

S.II.2.2 Bandwidth Choice: Direct Plug-In (DPI)

We now detail the required steps to implement the plug-in bandwidth $\hat{h}_{\texttt{dpi}}$ for interior and boundary points. We always set $K = L$, $\rho = 1$, and $q = p + 1$. The steps are:

(1) As a pilot bandwidth, use $\hat{h}_{\texttt{mse}}$: any data-driven version of $h^*_{\texttt{mse}}$.

(2) Using this bandwidth, estimate the regression function $m(X_i)$ as $\hat{m}(X_i; \hat{h}_{\texttt{mse}}) = r_p(X_i - x)'\hat{\beta}_p(\hat{h}_{\texttt{mse}})$, where $\hat{\beta}_p(\hat{h}_{\texttt{mse}})$ is the local polynomial coefficient estimate of order p exactly as defined in the main text, using the bandwidth $\hat{h}_{\texttt{mse}}$.

Form $\hat{\varepsilon}_i = Y_i - \hat{m}(X_i; \hat{h}_{\texttt{mse}})$.

(3) Following Fan and Gijbels (1996, §4.2) we estimate derivatives $m^{(k)}$ using a global least squares polynomial fit of order $k + 2$. That is, estimate $\hat{m}^{(p+3)}(x)$ as

$$\hat{m}^{(p+3)}(x) = [\hat{\gamma}]_{p+4}\,(p+3)! + [\hat{\gamma}]_{p+5}\,(p+4)!\,x + [\hat{\gamma}]_{p+6}\,\frac{(p+5)!}{2}\,x^2,$$

where $[\hat{\gamma}]_k$ is the k-th element of the vector $\hat{\gamma}$ that is estimated as

$$\hat{\gamma} = \operatorname*{arg\,min}_{\gamma \in \mathbb{R}^{p+6}} \sum_{i=1}^{n} \left(Y_i - r_{p+5}(X_i)'\gamma\right)^2.$$

The estimate for $\hat{m}^{(p+2)}(x)$ is similar, with all indexes incremented down once.

For interior points, both are needed, while only $\hat{m}^{(p+2)}(x)$ is required for the boundary.

(4) The estimated polynomials $\hat{q}_{k,\texttt{rbc}}$, $k = 1, 2, 3$ and the bias constants $\hat{\tilde{\eta}}^{\texttt{int}}_{\texttt{bc}}$ and $\hat{\tilde{\eta}}^{\texttt{bnd}}_{\texttt{bc}}$ are defined as follows. The polynomials $q_{1,\texttt{rbc}}$, $q_{2,\texttt{rbc}}$, and $q_{3,\texttt{rbc}}$, which do *not* have an argument, are defined in terms of those given in Section S.II.1.2, which *do* have an argument. Specifically, the polynomials in Section S.II.1.2 should be doubled, divided by the standard Normal density, and evaluated at the Normal quantile $z_{\alpha/2}$, that is, $q_{k,\texttt{rbc}} = \phi(z_{\alpha/2})^{-1} q_{k,\texttt{rbc}}(z_{\alpha/2})$. For $q_{1,\texttt{rbc}}$, the form given in Eqn. (S.II.6) should be used.

Note that with the recommended choice of $K = L$, $\rho = 1$, and $q = p + 1$, the polynomials $\hat{q}_{k,\texttt{rbc}}$, $k = 1, 2, 3$ can be read off the expressions for the undersmoothing versions, $\hat{q}_{k,\texttt{us}}$, $k = 1, 2, 3$, with p replaced by $p + 1$.

The bias terms, for the interior and boundary, are given as follows (dropping remainder terms). With $q = p+1$, and hence even, and $\rho = 1$, the expressions of Section S.II.4 simplify. For the interior: $\eta_{\texttt{bc}}^{\texttt{int}} = \sqrt{nh}h^{p+3}\tilde{\eta}_{\texttt{bc}}^{\texttt{int}}$, with

$$\tilde{\eta}_{\texttt{bc}}^{\texttt{int}} = h^{-1}\frac{m^{(p+2)}}{(p+2)!}\left\{e_0'\tilde{\Gamma}_p^{-1}\left(\tilde{\Lambda}_{p,2} - \tilde{\Lambda}_{p,1}e_{p+1}'\tilde{\Gamma}_q^{-1}\tilde{\Lambda}_{q,1}\right)\right\} + \frac{m^{(p+3)}}{(p+3)!}\left\{e_0'\tilde{\Gamma}_p^{-1}\left(\tilde{\Lambda}_{p,3} - \tilde{\Lambda}_{p,1}e_{p+1}'\tilde{\Gamma}_q^{-1}\tilde{\Lambda}_{q,2}\right)\right\};$$

At the boundary: $\eta_{\texttt{bc}}^{\texttt{bnd}} = \sqrt{nh}h^{p+2}\tilde{\eta}_{\texttt{bc}}^{\texttt{bnd}}$, with

$$\tilde{\eta}_{\texttt{bc}}^{\texttt{bnd}} = \frac{m^{(p+2)}}{(p+2)!}\left\{e_0'\tilde{\Gamma}_p^{-1}\left(\tilde{\Lambda}_{p,2} - \tilde{\Lambda}_{p,1}e_{p+1}'\tilde{\Gamma}_q^{-1}\tilde{\Lambda}_{q,1}\right)\right\}.$$

The estimates of these, $\hat{q}_{k,\texttt{rbc}}$, $k = 1, 2, 3$ and $\hat{\tilde{\eta}}_{\texttt{bc}}^{\texttt{int}}$ and $\hat{\tilde{\eta}}_{\texttt{bc}}^{\texttt{bnd}}$, are defined by replacing:

(i) h with $\hat{h}_{\texttt{mse}}$,

(ii) population expectations with sample averages (see note below),

(iii) residuals ε_i with $\hat{\varepsilon}_i$,

(iv) derivatives $m^{(p+2)}$ and $m^{(p+3)}$ with their estimators from above,

(v) limiting matrices $\tilde{\Gamma}_p$, $\tilde{\Lambda}_{p,2}$, etc, with the corresponding sample versions using the bandwidth $\hat{h}_{\texttt{mse}}$, e.g., $\tilde{\Gamma}_p$ is replaced with $\Gamma_p(\hat{h}_{\texttt{mse}}) = R_p'W_p(\hat{h}_{\texttt{mse}})R_p/n$, where $W_p(\hat{h}_{\texttt{mse}}) = \mathrm{diag}\left(\hat{h}_{\texttt{mse}}^{-1}K\left((X_i - x)/\hat{h}_{\texttt{mse}}\right)\right)$.

(5) Finally $\hat{h}_{\texttt{dpi}}^{\texttt{int}} = \hat{H}_{\texttt{dpi}}^{\texttt{int}}(\hat{h}_{\texttt{mse}})n^{-1/(p+4)}$ and $\hat{h}_{\texttt{dpi}}^{\texttt{bnd}} = \hat{H}_{\texttt{dpi}}^{\texttt{bnd}}(\hat{h}_{\texttt{mse}})n^{-1/(p+3)}$, where

$$\hat{H}_{\texttt{dpi}}^{\texttt{int}}(\hat{h}_{\texttt{mse}}) = \arg\min_H \left|H^{-1}\hat{q}_{1,\texttt{rbc}} + H^{1+2(p+3)}(\hat{\tilde{\eta}}_{\texttt{bc}}^{\texttt{int}})^2\hat{q}_{2,\texttt{rbc}} + H^{p+3}(\hat{\tilde{\eta}}_{\texttt{bc}}^{\texttt{int}})\hat{q}_{3,\texttt{rbc}}\right|,$$

while at (or near) the boundary the optimal bandwidth is $h_{\texttt{rbc}}^* = H_{\texttt{rbc}}^*(\rho)n^{-1/(p+3)}$, where

$$\hat{H}_{\texttt{dpi}}^{\texttt{bnd}}(\hat{h}_{\texttt{mse}}) = \arg\min_H \left|H^{-1}\hat{q}_{1,\texttt{rbc}} + H^{1+2(p+2)}(\hat{\tilde{\eta}}_{\texttt{bc}}^{\texttt{bnd}})^2\hat{q}_{2,\texttt{rbc}} + H^{p+2}(\hat{\tilde{\eta}}_{\texttt{bc}}^{\texttt{bnd}})\hat{q}_{3,\texttt{rbc}}\right|.$$

These numerical minimizations are easily solved; see note below. Code available from the authors' websites performs all the above steps.

Remark S.II.2 (Notes on computation)**.**

- When numerically solving the above minimization problems, computation will be greatly sped up by squaring the objective function.

- For step 4 above, in estimating $q_{1,\texttt{rbc}}$, the form given in Eqn. (S.II.6) should be used. The original form requires evaluating a triple sum, or third order U-statistic, which will be far slower than the right hand side of Eqn. (S.II.6).

- For step 4(ii) above, in estimating $\hat{q}_{1,\texttt{rbc}}$, and specifically when replacing population expectations with sample averages, we use the appropriate U-statistic forms to reduce bias. There are several terms which are expectations over two or three observations, and for these the second or third order U-statistic forms are preferred. For example, when estimating terms such as

$$\mathbb{E}\left[h^{-2}\ell^0_{\texttt{us}}(X_i)^2(r_p(X_{h,i})'\tilde{\Gamma}_p^{-1}(Kr_p)(X_{h,j}))^2\varepsilon_j^2\right]$$

we use

$$\frac{1}{n(n-1)}\sum_{i=1}^n\sum_{j\neq i}\left[\hat{h}_{\texttt{mse}}^{-2}\hat{\ell}^0_{\texttt{rbc}}(X_i)^2(r_p(X_{\hat{h}_{\texttt{mse}},i})'\Gamma_p^{-1}(Kr_p)(X_{\hat{h}_{\texttt{mse}},j}))^2\hat{\varepsilon}_j^2\right],$$

where $\hat{\ell}^0_{\texttt{rbc}}(X_i)$ is made feasible as in step 4(v). ■

S.II.2.3 Alternative Standard Errors

As argued in the main text, using variance forms other than (S.II.4) and (S.II.5) can be detrimental to coverage. Within these forms however, two alternative estimates of Σ are natural. First, motivated by the fact that the least-squares residuals are on average too small, the well-known HCk class of heteroskedasticity consistent estimators can be used; see MacKinnon (2013) for details and a recent review. In our notation, these are defined as follows. First, $\hat{\sigma}^2_{\texttt{us}}$-HC0 is the estimator above. Then, for $k = 1, 2, 3$, the $\hat{\sigma}^2_{\texttt{us}}$-HC$k$ estimator is obtained by dividing $\hat{\varepsilon}_i^2$ by, respectively, $(n-2\,\mathrm{tr}(Q_p)+\mathrm{tr}(Q_p'Q_p))/n$, $(1-Q_{p,ii})$, and $(1-Q_{p,ii})^2$, where $Q_{p,ii}$ is the i-th diagonal element of the projection matrix $Q_p := R_p'\Gamma_p^{-1}R_p'W_p/n$. The corresponding estimators $\hat{\sigma}^2_{\texttt{rbc}}$-HC$k$ are the same way, with q in place of p. As is well-known in the literature, these estimators perform better for small sample sizes, a fact we confirm in our simulation study below.

A second option is to use a nearest-neighbor-based variance estimators with a fixed number of neighbors, following the ideas of Muller and Stadtmuller (1987); Abadie and Imbens (2008). To define these, let J be a fixed number and $j(i)$ be the j-th closest observation to X_i, $j = 1, \ldots, J$, and set $\hat{v}(X_i) = \frac{J}{J+1}(Y_i - \sum_{j=1}^J Y_{j(i)}/J)^2$. This "estimate" is unbiased (but inconsistent) for $v(X_i)$.

Both types of residual estimators could be handled in our results. The constants will change, but the rates will not. This is because, in all cases, the errors in estimating $v(X_i)$ are no greater than in the original $\hat{m}(x)$. Inspection of the proof shows that simple modifications allow for the HCk estimators: only the terms of Eqn. (S.II.12) will change, and indeed, we conjecture that the HCk estimators will result in fewer terms and a reduced coverage error. This is consistent with the improved finite-sample behavior of these estimators and the fact that they are asymptotically equivalent. Accommodating the nearest-neighbor estimates require slightly more work and a modified version of Assumption S.II.3.

One crucial property of our method, in the context of Edgeworth expansions, is that the bias in estimation of Σ is of the same order as the original $\hat{m}(x)$. Using other methods may result in additional terms, with possibly distinct rates, appearing in the Edgeworth expansions. Some examples that may have this issue are (i) using $\hat{v}(X_i) = (Y_i - \hat{m}(x))^2$; (ii) using local or assuming global heteroskedasticity; (iii) using other nonparametric estimators for $v(X_i)$, relying on new tuning parameters.

S.II.3 Assumptions

The following assumptions are sufficient for our results. The first two are copied directly from the main text (see discussion there) and the third is the appropriate Cramér's condition.

Assumption S.II.1 (Data-generating process)**.** *$\{(Y_1, X_1), \ldots, (Y_n, X_n)\}$ is a random sample, where X_i has the absolutely continuous distribution with Lebesgue density f, $\mathbb{E}[Y^{8+\delta}|X] < \infty$ 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 S.II.2 (Kernels)**.** *The kernels K and L are positive, bounded, even functions, and with compact support.*

Assumption S.II.3 (Cramér's Condition)**.** *For each $\delta > 0$ and all sufficiently small h, the random variables $Z_{\texttt{us}}(u)$ and $Z_{\texttt{rbc}}(u)$ defined below obey*

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

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

The random variables of Assumption S.II.3 are defined follows. For two kernels K_1 and K_2, two polynomial orders (i.e. positive integers) p_1 and p_2, a bandwidth b, and a scalar ρ, let

$$Z_m(u; K_1, p_1, p_2, b, \rho) := \Big(K_1(u) r_{p_1}(u)'\varepsilon, \ K_1(u) r_{p_1}(u)'(m(x-ub) - r_{p_2}(ub)'\beta_{p_2}), \ \mathrm{vech}(K_1(u) r_{p_1}(u) r_{p_1}(u)')' \Big)'.$$

and

$$\begin{aligned} Z_\sigma(u; K_1, K_2, p_1, p_2, b, \rho) := \Big(& \mathrm{vech}(K_1(u) K_2(u\rho) r_{p_1}(u) r_{p_2}(u\rho)'\varepsilon^2)', \\ & \mathrm{vech}(K_1(u) K_2(u\rho) r_{p_1}(u) r_{p_2}(u\rho)' v(x - ub))', \\ & \mathrm{vech}(K_1(u) K_2(u\rho) r_{p_1}(u) r_{p_2}(u\rho)'\varepsilon(m(x - ub) - r_{p_2}(ub)'\beta_{p_2}))', \\ & \mathrm{vech}(K_2(u)^2 r_{p_2}(u) r_{p_2}(u)' r_{p_2}(u)')', \\ & \mathrm{vech}(K_1(u) K_2(u\rho) r_{p_1}(u) r_{p_2}(u\rho)' r_{p_2}(u)'\varepsilon)', \end{aligned}$$

$$\text{vech}(K_1(u)K_2(u\rho)r_{p_1}(u)r_{p_2}(u\rho)'r_{p_2}(u\rho)'\varepsilon(m(x-ub)-r_{p_2}(ub)'\beta_{p_2}))' \Big)'.$$

The subscripts are intended to make clear that $Z_m(\cdot)$ collects quantities from the numerator of the Studentized statistic, while $Z_\sigma(\cdot)$ gathers additional variables required for the variance estimation. With this notation, we define

$$Z_{\texttt{us}}(u) = \big(Z_m(u; K, p, p, h, 1)',\ Z_\sigma(u; K, K, p, p, h, 1)'\big)',$$

$$Z_{\texttt{bc}}(u) = \big(Z_m(u; K, p, p{+}1, h, 1)',\ Z_m(u; L, q, q, b, \rho)',\ \text{vech}(K(u)r_p(u)u^{p+1})',\ Z_\sigma(u; K, K, p, p, h, 1)'\big)',$$

and

$$Z_{\texttt{rbc}}(u) = \big(Z_m(u; K, p, p+1, h, 1)',\ Z_m(u; L, q, q, b, \rho)',\ \text{vech}(K(u)r_p(u)u^{p+1})',$$
$$Z_\sigma(u; K, K, p, q, b, \rho)',\ Z_\sigma(u; L, L, q, q, b, 1)',\ Z_\sigma(u; K, L, p, q, b, \rho)'\big)'.$$

Discussion. This notation is quite compact, and while it emphasizes the simplicity of Cramér's condition and the fact that it puts mild restrictions on the kernels, it does obscure the full notational breadth, particularly for $Z_{\texttt{rbc}}$. I is also mostly repetitive: what holds for the kernel K and order p fit must also hold for L and q, and for their squares and cross products. To make this clear, we can expand all the Z_m and Z_σ, to write out the full random variables as

$$Z_{\texttt{us}}(u) = \Big(K(u)r_p(u)'\varepsilon,\ K(u)r_p(u)'(m(x-uh)-r_p(uh)'\beta_p),\ \text{vech}(K(u)r_p(u)r_p(u)')',$$
$$\text{vech}(K(u)^2r_p(u)r_p(u)'\varepsilon^2)',\ \text{vech}(K(u)^2r_p(u)r_p(u)'v(x-uh))',$$
$$\text{vech}(K(u)^2r_p(u)r_p(u)'\varepsilon(m(x-uh)-r_p(uh)'\beta_p))',\ \text{vech}(K(u)^2r_p(u)r_p(u)'r_p(u)')',$$
$$\text{vech}(K(u)^2r_p(u)r_p(u)'r_p(u)'\varepsilon)',\ \text{vech}(K(u)^2r_p(u)r_p(u)'r_p(u)'\varepsilon(m(x-uh)-r_p(uh)'\beta_p))'\Big)',$$

$$Z_{\texttt{bc}}(u) = \Big(K(u)r_p(u)'\varepsilon,\ \text{vech}(K(u)r_p(u)r_p(u)')',$$
$$\text{vech}(K(u)^2r_p(u)r_p(u)'\varepsilon^2)',\ \text{vech}(K(u)^2r_p(u)r_p(u)'v(x-uh))',$$
$$\text{vech}(K(u)^2r_p(u)r_p(u)'\varepsilon(m(x-uh)-r_p(uh)'\beta_p))',\ \text{vech}(K(u)^2r_p(u)r_p(u)'r_p(u)')',$$
$$\text{vech}(K(u)^2r_p(u)r_p(u)'r_p(u)'\varepsilon)',\ \text{vech}(K(u)^2r_p(u)r_p(u)'r_p(u)'\varepsilon(m(x-uh)-r_p(uh)'\beta_p))',$$
$$K(u)r_p(u)'(m(x-uh)-r_{p+1}(uh)'\beta_{p+1}),\ L(u\rho)r_q(u\rho)'\varepsilon,\ \text{vech}(L(u\rho)r_q(u\rho)r_q(u\rho)')',$$
$$\text{vech}(K(u)r_p(u)u^{p+1})',\ L(u\rho)r_q(u\rho)'(m(x-uh)-r_q(uh)'\beta_q)\Big)',$$

and

$$Z_{\texttt{rbc}}(u) = \Big(Z_{\texttt{bc}}(u)',\ \text{vech}(K(u)^2r_p(u)r_p(u)'\varepsilon^2)',\ \text{vech}(K(u)^2r_p(u)r_p(u)'v(x-ub))',$$
$$\text{vech}(K(u)^2r_p(u)r_p(u)'\varepsilon(m(x-ub)-r_q(ub)'\beta_q))',\ \text{vech}(K(u)^2r_p(u)r_p(u)'r_q(u\rho)')',$$

$$
\begin{aligned}
&\mathrm{vech}(K(u)^2 r_p(u) r_p(u)' r_q(u\rho)'\varepsilon)',\ \mathrm{vech}(K(u)^2 r_p(u) r_p(u)' r_q(u\rho)'\varepsilon(m(x-ub) - r_q(ub)'\beta_q))',\\
&\mathrm{vech}(L(u)^2 r_q(u) r_q(u)'\varepsilon^2)',\ \mathrm{vech}(L(u)^2 r_q(u) r_q(u)' v(x-ub))',\\
&\mathrm{vech}(L(u)^2 r_q(u) r_q(u)'\varepsilon(m(x-ub) - r_q(ub)'\beta_q))',\ \mathrm{vech}(L(u)^2 r_q(u) r_q(u)' r_q(u)')',\\
&\mathrm{vech}(L(u)^2 r_q(u) r_q(u)' r_q(u)'\varepsilon)',\ \mathrm{vech}(L(u)^2 r_q(u) r_q(u)' r_q(u)'\varepsilon(m(x-ub) - r_q(ub)'\beta_q))',\\
&\mathrm{vech}(K(u)L(u\rho) r_p(u) r_q(u\rho)'\varepsilon^2)',\ \mathrm{vech}(K(u)L(u\rho) r_p(u) r_q(u\rho)' v(x-ub))',\\
&\mathrm{vech}(K(u)L(u\rho) r_p(u) r_q(u\rho)'\varepsilon(m(x-ub) - r_q(ub)'\beta_q))',\ \mathrm{vech}(L(u)^2 r_q(u) r_q(u)' r_q(u)')',\\
&\mathrm{vech}(K(u)L(u\rho) r_p(u) r_q(u\rho)' r_q(u)'\varepsilon)',\\
&\mathrm{vech}(K(u)L(u\rho) r_p(u) r_q(u\rho)' r_q(u\rho)'\varepsilon(m(x-ub) - r_q(ub)'\beta_q))' \Big)'.
\end{aligned}
$$

Finally, the precise random variables $Z_{\texttt{us}}(u)$, $Z_{\texttt{bc}}(u)$, and $Z_{\texttt{rbc}}(u)$ used can be replaced with slightly different constructions without altering the conclusions of Theorem S.II.1: there are other potential functions $\tilde{T}$ that satisfy Eqn. (S.II.8) in the proof. Such changes necessarily involve asymptotically negligible terms, and do not materially alter the severity of the restrictions imposed.

Remark S.II.3 (Sufficient Conditions for Cramér's Condition)**.** Assumption S.II.3 is a high level condition, but one that is fairly mild. It is essentially a continuity requirement, and is discussed at length by (among others) Bhattacharya and Rao (1976), Bhattacharya and Ghosh (1978), and Hall (1992a). For a recent work in econometrics, the present condition can be compared to that employed by Kline and Santos (2012) for parametric regression (the role of the covariates is here played by $r_p(X_{h,i})$): ours is more complex due to the nonparametric smoothing bias and the fact that the expansion is carried out to higher order.

It is straightforward to provide sufficient conditions for Assumption S.II.3, given that Assumptions S.II.1 and S.II.2 hold. In particular, if we additionally assume that $(1,\ \mathrm{vech}(K(u)r_p(u)r_p(u)')')'$ comprises a linearly independent set of functions on $[-1,1]$, then it holds $Z_{\texttt{us}}(u)$ has components that are nondegenerate and absolutely continuous, and this will imply that Assumption S.II.3 holds for $Z_{\texttt{us}}(u)$, by arguing as in Bhattacharya and Ghosh (1978, Lemma 2.2) and Hall (1992a, p. 65). This is precisely the approach taken by Chen and Qin (2002), when studying undersmoothed local linear regression. If the linear independence continues to hold when the set of functions is augmented with $\mathrm{vech}(L(u)r_q(u)r_q(u)')$, then $Z_{\texttt{bc}}(u)$ and $Z_{\texttt{rbc}}(u)$ satisfy Assumption S.II.3 as well.

At heart, these are requirements on the kernel functions, just as in Assumption S.I.3 in the density case. The uniform kernel is again ruled out. See Section S.I.3. Further, note that if these sets of functions are not linearly independent, there will exist a there exists a smaller set of functions which are linearly independent and can replace the original set while leaving the value of the statistic unchanged (see Bhattacharya and Ghosh (1978, p. 442)). ■

S.II.4 Bias

We will not present a detailed discussion of bias issues, along the lines of Section S.I.4.1, for brevity; we focus only on the case of nonbinding smoothness.

The biases $\eta_{\texttt{us}}$ and $\eta_{\texttt{bc}}$ are not as conceptually simple as in the density case. The closest parallel to the density case would be (for example) $\eta_{\texttt{us}} = \sqrt{nh}(\mathbb{E}[\hat{m}] - m)$, but this can not be used due to the presence of Γ_p^{-1} inside the expectation, and the next natural choice, the conditional bias $\sqrt{nh}(\mathbb{E}[\hat{m}|X_1, \dots X_n] - m)$, is still random. Instead, $\eta_{\texttt{us}}$ and $\eta_{\texttt{bc}}$ are biases computed after replacing Γ_p, Γ_q, and $\Lambda_{p,1}$ with their expectations, denoted $\tilde{\Gamma}_p$, $\tilde{\Gamma}_q$, and $\tilde{\Lambda}_{p,1}$. We thus define

$$\begin{aligned}
\eta_{\texttt{us}} &= \sqrt{nh} \int e_0' \tilde{\Gamma}_p^{-1} K(u) r_p(u) \left(m(x - uh) - r_p(uh)'\beta_p \right) f(x - uh) du, \\
\eta_{\texttt{bc}} &= \sqrt{nh} \int e_0' \tilde{\Gamma}_p^{-1} K(u) r_p(u) \left(m(x - uh) - r_{p+1}(uh)'\beta_{p+1} \right) f(x - uh) du \\
&\quad - \sqrt{nh} \rho^{p+1} \int e_0' \tilde{\Gamma}_p^{-1} \tilde{\Lambda}_{p,1} e_{p+1}' \tilde{\Gamma}_q^{-1} L(u) r_q(u) \left(m(x - ub) - r_q(ub)'\beta_q \right) f(x - ub) du.
\end{aligned} \tag{S.II.7}$$

For the generic results of coverage error or the generic Edgeworth expansions of Theorem S.II.1 below, the above definitions of $\eta_{\texttt{us}}$ and $\eta_{\texttt{bc}}$ are suitable. For the Corollaries detailing specific cases, and to understand the behavior at different points, it is useful to make the leading terms precise, that is, analogues of Equations (S.I.2) and (S.I.3). We must consider interior and boundary point estimation, and even and odd q. We depart slightly from other terms of the expansion in that we do retain only the leading term for some pieces. This is done in order to capture the rate of convergence explicitly and to give practicable results. These results are derived by Fan and Gijbels (1996, Section 3.7) and similar calculations (though our expressions differ slightly as fixed-n expectations are retained as much as possible).

Since p is odd, both at boundary and interior points we have

$$\eta_{\texttt{us}} = \sqrt{nh} h^{p+1} \frac{m^{(p+1)}}{(p+1)!} e_0' \tilde{\Gamma}_p^{-1} \tilde{\Lambda}_{p,1} \left[1 + o(1) \right].$$

Moving to $\eta_{\texttt{bc}}$, consider the first term, which in the present notation is: $\sqrt{nh}\mathbb{E}[h^{-1}\ell_{\texttt{us}}^0(X)(m(X) - r_{p+1}(X - x)'\beta_{p+1})]$. With $p + 1$ even, we find that in the interior the leading terms are

$$\sqrt{nh} h^{p+3} e_0' \tilde{\Gamma}_p^{-1} \left(\frac{m^{(p+2)}}{(p+2)!} \tilde{\Lambda}_{p,2} h^{-1} + \frac{m^{(p+3)}}{(p+3)!} \tilde{\Lambda}_{p,3} \right) \left[1 + o(1) \right],$$

due to the well-known symmetry properties of local polynomials that result in the cancellation of the leading terms of $\tilde{\Gamma}_p^{-1}$ and $\tilde{\Lambda}_{p,2}$. The rate of h^{p+3} accounts for this. At the boundary, no such cancellation occurs and we have only

$$\sqrt{nh} h^{p+2} \frac{m^{(p+2)}}{(p+2)!} e_0' \tilde{\Gamma}_p^{-1} \tilde{\Lambda}_{p,2} \left[1 + o(1) \right].$$

Next, turn to the bias of the bias estimate:

$$\sqrt{nh}\rho^{p+1}e_0'\tilde{\Gamma}_p^{-1}\tilde{\Lambda}_{p,1}e_{p+1}'\tilde{\Gamma}_q^{-1}\int L(u)r_q(u)\left(m(x-ub)-r_q(ub)'\beta_q\right)f(x-ub)du.$$

If q is odd (so that $q-(p+1)$ is also odd), then at the interior or boundary the leading term will be

$$\sqrt{nh}b^{q+1}\rho^{p+1}\frac{m^{(q+1)}}{(q+1)!}e_0'\tilde{\Gamma}_p^{-1}\tilde{\Lambda}_{p,1}e_{p+1}'\tilde{\Gamma}_q^{-1}\tilde{\Lambda}_{q,1}\left[1+o(1)\right]\asymp\sqrt{nh}h^{p+1}b^{q-p}.$$

The same expression applies at the boundary for q even. However, for the interior, if q is even, which it is in the leading case of $q=p+1$, then we again have cancellation of certain leading terms, resulting in the bias of the bias estimate being

$$\sqrt{nh}b^{q+2}\rho^{p+1}e_0'\tilde{\Gamma}_p^{-1}\tilde{\Lambda}_{p,1}e_{p+1}'\tilde{\Gamma}_q^{-1}\left(\frac{m^{(q+1)}}{(q+1)!}\tilde{\Lambda}_{q,1}b^{-1}+\frac{m^{(q+2)}}{(q+2)!}\tilde{\Lambda}_{q,2}\right)\left[1+o(1)\right]\asymp\sqrt{nh}h^{p+1}b^{q+1-p}.$$

Combining all these results, we find the following. For an interior point $\eta_{\texttt{bc}}^{\texttt{int}}=\sqrt{nh}h^{p+3}\left[\tilde{\eta}_{\texttt{bc}}^{\texttt{int}}+o(1)\right]$, where, if q is even

$$\tilde{\eta}_{\texttt{bc}}^{\texttt{int}}=e_0'\tilde{\Gamma}_p^{-1}\left(\frac{m^{(p+2)}}{(p+2)!}\tilde{\Lambda}_{p,2}h^{-1}+\frac{m^{(p+3)}}{(p+3)!}\tilde{\Lambda}_{p,3}\right)\\
-\rho^{-2}b^{q-(p+1)}e_0'\tilde{\Gamma}_p^{-1}\tilde{\Lambda}_{p,1}e_{p+1}'\tilde{\Gamma}_q^{-1}\left(\frac{m^{(q+1)}}{(q+1)!}\tilde{\Lambda}_{q,1}b^{-1}+\frac{m^{(q+2)}}{(q+2)!}\tilde{\Lambda}_{q,2}\right),$$

while if q is odd,

$$\tilde{\eta}_{\texttt{bc}}^{\texttt{int}}=e_0'\tilde{\Gamma}_p^{-1}\left(\frac{m^{(p+2)}}{(p+2)!}\tilde{\Lambda}_{p,2}h^{-1}+\frac{m^{(p+3)}}{(p+3)!}\tilde{\Lambda}_{p,3}\right)-\rho^{-2}b^{q-(p+2)}\frac{m^{(q+1)}}{(q+1)!}e_0'\tilde{\Gamma}_p^{-1}\tilde{\Lambda}_{p,1}e_{p+1}'\tilde{\Gamma}_q^{-1}\tilde{\Lambda}_{q,1}.$$

At the boundary, for any q, $\eta_{\texttt{bc}}^{\texttt{bnd}}=\sqrt{nh}h^{p+2}\left[\tilde{\eta}_{\texttt{bc}}^{\texttt{bnd}}+o(1)\right]$, with

$$\tilde{\eta}_{\texttt{bc}}^{\texttt{bnd}}=\frac{m^{(p+2)}}{(p+2)!}e_0'\tilde{\Gamma}_p^{-1}\tilde{\Lambda}_{p,2}-\rho^{-1}b^{q-(p+1)}\frac{m^{(q+1)}}{(q+1)!}e_0'\tilde{\Gamma}_p^{-1}\tilde{\Lambda}_{p,1}e_{p+1}'\tilde{\Gamma}_q^{-1}\tilde{\Lambda}_{q,1}.$$

S.II.5 Main Result: Edgeworth Expansion

We now state our generic Edgeworth expansion, from whence the coverage probability expansion results follow immediately. We have opted to state separate results for undersmoothing, bias correction, and robust bias correction, rather than the unified statement of Theorem S.I.1, for clarity. The unified structure is still present, and will be used in the proof of the result below, but is too cumbersome to use here. The Standard Normal distribution and density functions are $\Phi(z)$ and $\phi(z)$, respectively.

Theorem S.II.1. *Let Assumptions S.II.1, S.II.2, and S.II.3 hold, and assume* $nh/\log(n) \to \infty$.

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

$$F_{\mathtt{us}}(z) = \Phi(z) + \frac{1}{\sqrt{nh}} p_{1,\mathtt{us}}(z) + \eta_{\mathtt{us}} p_{3,\mathtt{us}}(z) + \frac{1}{nh} q_{1,\mathtt{us}}(z) + \eta_{\mathtt{us}}^2 q_{2,\mathtt{us}}(z) + \frac{\eta_{\mathtt{us}}}{\sqrt{nh}} q_{3,\mathtt{us}}(z),$$

we have

$$\sup_{z\in\mathbb{R}} |\mathbb{P}[T_{\mathtt{us}} < z] - F_{\mathtt{us}}(z)| = o\left((nh)^{-1} + (nh)^{-1/2}\eta_{\mathtt{us}} + \eta_{\mathtt{us}}^2\right).$$

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

$$F_{\mathtt{bc}}(z) = \Phi(z) + \frac{1}{\sqrt{nh}} p_{1,\mathtt{bc}}(z) + \eta_{\mathtt{bc}} p_{3,\mathtt{bc}}(z) + \frac{1}{nh} q_{1,\mathtt{us}}(z) + \eta_{\mathtt{bc}}^2 q_{2,\mathtt{bc}}(z) + \frac{\eta_{\mathtt{bc}}}{\sqrt{nh}} q_{3,\mathtt{bc}}(z)$$
$$- \rho^{p+2}(\Omega_1 + \rho^{p+1}\Omega_2)\frac{\phi(z)}{2} z,$$

we have

$$\sup_{z\in\mathbb{R}} |\mathbb{P}[T_{\mathtt{bc}} < z] - F_{\mathtt{bc}}(z)| = o\left((nh)^{-1} + (nh)^{-1/2}\eta_{\mathtt{bc}} + \eta_{\mathtt{bc}}^2 + \rho^{1+2(p+1)}\right).$$

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

$$F_{\mathtt{rbc}}(z) = \Phi(z) + \frac{1}{\sqrt{nh}} p_{1,\mathtt{rbc}}(z) + \eta_{\mathtt{bc}} p_{3,\mathtt{rbc}}(z) + \frac{1}{nh} q_{1,\mathtt{rbc}}(z) + \eta_{\mathtt{bc}}^2 q_{2,\mathtt{rbc}}(z) + \frac{\eta_{\mathtt{bc}}}{\sqrt{nh}} q_{3,\mathtt{rbc}}(z),$$

we have

$$\sup_{z\in\mathbb{R}} |\mathbb{P}[T_{\mathtt{rbc}} < z] - F_{\mathtt{rbc}}(z)| = o\left((nh)^{-1} + (nh)^{-1/2}\eta_{\mathtt{bc}} + \eta_{\mathtt{bc}}^2\right).$$

S.II.5.1 Coverage Error for Undersmoothing

For undersmoothing estimators, we have the following result, which is valid for both interior and boundary points, with moments appropriately truncated if necessary. This result is the analogue of the robust bias correction corollary in the main text, and follows directly from the generic theorem there or Theorem S.II.1 above. Exponents such as $1 + 2(p + 1)$ are intentionally not simplified to ease comparison to other results, particularly the density case.

The polynomials $q_{1,\mathtt{us}}$, $q_{2,\mathtt{us}}$, and $q_{3,\mathtt{us}}$, which do *not* have an argument, are defined in terms of those given in Section S.II.1.2 and used in Theorem S.II.1, which *do* have an argument. Specifically, the polynomials in Section S.II.1.2 and Theorem S.II.1 should be doubled, divided by the standard Normal density, and evaluated at the Normal quantile $z_{\alpha/2}$, that is,

$$q_{k,\mathtt{us}} := \frac{2}{\phi(z)} q_{k,\mathtt{us}}(z)\bigg|_{z=z_{\alpha/2}}, \qquad k = 1, 2, 3.$$

Corollary S.II.1 (Undersmoothing)**.** *Let the conditions of Theorem S.II.1(a) hold. Then*

$$\mathbb{P}[m \in I_{\mathtt{us}}] = 1 - \alpha + \left\{ \frac{1}{nh} q_{1,\mathtt{us}} + nh^{1+2(p+1)} \left(m^{(p+1)} \right)^2 \left(e_0' \tilde{\Gamma}_p^{-1} \tilde{\Lambda}_{p,1}/(p+1)! \right)^2 q_{2,\mathtt{us}} \right.$$
$$\left. + h^{p+1} \left(m^{(p+1)} \right) \left(e_0' \tilde{\Gamma}_p^{-1} \tilde{\Lambda}_{p,1}/(p+1)! \right) q_{3,\mathtt{us}} \right\} \phi(z_{\frac{\alpha}{2}}) \{1 + o(1)\}.$$

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

$$H^*_{\mathtt{us}} = \arg\min_H \left| H^{-1} q_{1,\mathtt{us}} + H^{1+2(p+1)} \left(m^{(p+1)} \right)^2 \left(e_0' \tilde{\Gamma}_p^{-1} \tilde{\Lambda}_{p,1}/(p+1)! \right)^2 q_{2,\mathtt{us}} \right.$$
$$\left. + H^{p+1} \left(m^{(p+1)} \right) \left(e_0' \tilde{\Gamma}_p^{-1} \tilde{\Lambda}_{p,1}/(p+1)! \right) q_{3,\mathtt{us}} \right|.$$

S.II.6 Proof of Main Result

We will first prove Theorem S.II.1(a), as it is notationally simplest. From a technical and conceptual point of view, proving the remainder of Theorem S.II.1 is identical, simply more involved notationally due to the additional complexity of the bias correction. Outlines of these proofs are found below.

S.II.6.1 Proof of Theorem S.II.1(a)

Let $s = \sqrt{nh}$.

Throughout this proof, we will generally omit the subscripts $\mathtt{us}$ and p when this causes no confusion. This entire proof focuses on the undersmoothing statistic, $T_{\mathtt{us}} = \hat{\sigma}_{\mathtt{us}}^{-1} s(\hat{m} - m)$, and since bias correction is not involved at all, the associated constructions such as Γ_q, W_q, etc, do not appear, and hence there is no need to carry the additional notation to distinguish W_p from W_q, or $\hat{\sigma}_{\mathtt{us}}$ from $\hat{\sigma}_{\mathtt{rbc}}$, for example, and we will simply write Γ for Γ_p, W for W_p, $\hat{\sigma}$ for $\hat{\sigma}_{\mathtt{us}}$, etc.

Our goal is to expand $\mathbb{P}[T_{\mathtt{us}} < z]$, where $T_{\mathtt{us}} = \hat{\sigma}^{-1} s(\hat{m} - m)$. The proof proceeds by identifying a smooth function $\tilde{T} = \tilde{T}(z)$ such that, for the random variable $Z_{\mathtt{us}} := Z_{\mathtt{us}}(u)$ that obeys Cramér's condition (Assumption S.II.3), $\tilde{T}(\mathbb{E}[Z_{\mathtt{us}}]) = 0$ and

$$\mathbb{P}\big[T_{\mathtt{us}} < z\big] = \mathbb{P}\big[\tilde{T}(\bar{Z}_{\mathtt{us}}) < \tilde{z}\big] + o(s^{-2} + s^{-1}\eta + \eta^2), \tag{S.II.8}$$

where $\bar{Z} = \sum_{i=1}^n Z_i/n$ and $\tilde{z}$ is a known, nonrandom quantity that depends on the original quantile z and the remainder $T_{\mathtt{us}} - \tilde{T}$. An Edgeworth expansion for $\tilde{T}$ holds under Assumption S.II.3, and a Taylor expansion of this function around $\tilde{z}$ yields the final result. As in the density case, $\tilde{z}$ will capture the bias terms of $T_{\mathtt{us}}$: in that case $\tilde{z} = z - \eta/\tilde{\sigma}$, but here bias is present in both the numerator and the Studentization.

To begin, define the notation $\check{R} = [r_p(X_1 - x), \cdots, r_p(X_n - x)]'$ and $M = [m(X_1), \ldots, m(X_n)]'$, and use this to split T into variance and bias terms, as follows:

$$T = \hat{\sigma}^{-1} s e_0' \Gamma^{-1} R' W (Y - M)/n + \hat{\sigma}^{-1} s e_0' \Gamma^{-1} R' W (M - \check{R}\beta)/n.$$

We use this decomposition to rewrite $\mathbb{P}[T_{\texttt{us}} < z]$ as

$$\begin{aligned}
\mathbb{P}\left[T_{\texttt{us}} < z\right] &= \mathbb{P}\left[T_{\texttt{us}} - \tilde{\sigma}^{-1}\eta < z - \tilde{\sigma}^{-1}\eta\right] \\
&= \mathbb{P}\left[\left\{\hat{\sigma}^{-1} s e_0' \Gamma^{-1} R' W (Y - M)/n + \hat{\sigma}^{-1} s e_0' \Gamma^{-1} R' W (M - \check{R}\beta)/n - \tilde{\sigma}^{-1}\eta\right\} < z - \tilde{\sigma}^{-1}\eta\right] \\
&= \mathbb{P}\Bigg[\Bigg\{\tilde{\sigma}^{-1} s e_0' \Gamma^{-1} R' W (Y - M)/n \\
&\qquad + \tilde{\sigma}^{-1} s e_0' \tilde{\Gamma}^{-1} R' W (M - \check{R}\beta)/n - \tilde{\sigma}^{-1}\eta \\
&\qquad + \tilde{\sigma}^{-1} s e_0' \left(\Gamma^{-1} - \tilde{\Gamma}^{-1}\right) R' W (M - \check{R}\beta)/n \\
&\qquad + \left(\hat{\sigma}^{-1} - \tilde{\sigma}^{-1}\right) s e_0' \Gamma^{-1} R' W (Y - M)/n \\
&\qquad + \left(\hat{\sigma}^{-1} - \tilde{\sigma}^{-1}\right) s e_0' \Gamma^{-1} R' W (M - \check{R}\beta)/n\Bigg\} < z - \tilde{\sigma}^{-1}\eta\Bigg].
\end{aligned} \tag{S.II.9}$$

The first three lines in the last equality obey the desired properties of $\tilde{T}$ by the orthogonality of ε_i, the definition of $\eta_{\texttt{us}}$ in Eqn. (S.II.7) as $\mathbb{E}\left[s e_0' \tilde{\Gamma}^{-1} R' W (M - \check{R}\beta)/n\right]$, and the fact that $\Gamma^{-1} - \tilde{\Gamma}^{-1} = \tilde{\Gamma}^{-1}\left(\tilde{\Gamma} - \Gamma\right)\Gamma^{-1}$. For the final two (which are $T_{\texttt{us}} - \tilde{\sigma}^{-1} s(\hat{m} - m) = \hat{\sigma}^{-1} - \tilde{\sigma}^{-1} s(\hat{m} - m)$), we must expand the difference $\hat{\sigma}^{-1} - \tilde{\sigma}^{-1}$. Accounting for the resulting terms will constitute the bulk of the remainder of the proof, as well as complete the construction of $\tilde{z}$ and the remainder terms of Eqn. (S.II.8).[2]

To begin, with $\tilde{\sigma}^2 = e_0' \tilde{\Gamma}^{-1} \check{\Psi} \tilde{\Gamma}^{-1} e_0$ defined in Section S.II.1.2,

$$\frac{1}{\hat{\sigma}} = \frac{1}{\tilde{\sigma}} \left(\frac{\hat{\sigma}^2}{\tilde{\sigma}^2}\right)^{-1/2} = \frac{1}{\tilde{\sigma}} \left(1 + \frac{\hat{\sigma}^2 - \tilde{\sigma}^2}{\tilde{\sigma}^2}\right)^{-1/2},$$

and hence a Taylor expansion gives

$$\frac{1}{\hat{\sigma}} = \frac{1}{\tilde{\sigma}} \left[1 - \frac{1}{2}\frac{\hat{\sigma}^2 - \tilde{\sigma}^2}{\tilde{\sigma}^2} + \frac{3}{8}\left(\frac{\hat{\sigma}^2 - \tilde{\sigma}^2}{\tilde{\sigma}^2}\right)^2 - \frac{1}{3!}\frac{15}{8}\left(\frac{\hat{\sigma}^2 - \tilde{\sigma}^2}{\tilde{\sigma}^2}\right)^3 \frac{\tilde{\sigma}^7}{\bar{\sigma}^7}\right],$$

for a point $\bar{\sigma}^2 \in [\tilde{\sigma}^2, \hat{\sigma}^2]$, and so

$$\hat{\sigma}^{-1} - \tilde{\sigma}^{-1} = -\frac{1}{2}\frac{\hat{\sigma}^2 - \tilde{\sigma}^2}{\tilde{\sigma}^3} + \frac{3}{8}\frac{\left(\hat{\sigma}^2 - \tilde{\sigma}^2\right)^2}{\tilde{\sigma}^5} - \frac{5}{16}\frac{\left(\hat{\sigma}^2 - \tilde{\sigma}^2\right)^3}{\bar{\sigma}^7}. \tag{S.II.10}$$

We thus focus on $\hat{\sigma}^2 - \tilde{\sigma}^2$. Recall the definition of $\check{\Psi} = hR'W\Sigma WR/n$. Then define the two terms

[2] Technically, to obtain a $\tilde{T}$ with the desired properties, one need not expand $\hat{\sigma}^{-1} - \tilde{\sigma}^{-1}$ for the variance term: that is, in Eqn. (S.II.9), $\tilde{\sigma}^{-1} s e_0' \Gamma^{-1} R' W (Y - M)/n$ and $\left(\hat{\sigma}^{-1} - \tilde{\sigma}^{-1}\right) s e_0' \Gamma^{-1} R' W (Y - M)/n$ may be collapsed. This requires strengthening Cramér's condition (see Section S.II.3), and since $\hat{\sigma}^{-1} - \tilde{\sigma}^{-1}$ must be accounted for in the final bias term, $\left(\hat{\sigma}^{-1} - \tilde{\sigma}^{-1}\right) s e_0' \Gamma^{-1} R' W (M - \check{R}\beta)/n$, there is little reason not to do both terms.

A_1 and A_2 through the following:

$$\hat{\sigma}^2 - \tilde{\sigma}^2 = e_0'\Gamma^{-1}\left(\hat{\Psi} - \check{\Psi}\right)\Gamma^{-1}e_0 + \left(e_0'\Gamma^{-1}\check{\Psi}\Gamma^{-1}e_0 - e_0'\tilde{\Gamma}^{-1}\tilde{\Psi}\tilde{\Gamma}^{-1}e_0\right) =: A_1 + A_2. \qquad \text{(S.II.11)}$$

For A_1, recall that $\hat{\varepsilon}_i = y_i - r_p(X_i - x)'\hat{\beta}_p$ and so

$$\begin{aligned}
\hat{\Psi} - \check{\Psi} &= \frac{1}{nh}\sum_{i=1}^{n}(K^2 r_p r_p')(X_{h,i})\left\{\hat{\varepsilon}_i^2 - v(X_i)\right\} \\
&= \frac{1}{nh}\sum_{i=1}^{n}(K^2 r_p r_p')(X_{h,i})\left\{\left(y_i - r_p(X_i - x)'\hat{\beta}_p\right)^2 - v(X_i)\right\} \\
&= \frac{1}{nh}\sum_{i=1}^{n}(K^2 r_p r_p')(X_{h,i})\left\{\left(\varepsilon_i + [m(X_i) - r_p(X_i - x)'\beta_p] + r_p(X_i - x)'\left[\beta_p - \hat{\beta}_p\right]\right)^2 - v(X_i)\right\} \\
&=: A_{1,1} + A_{1,2} + A_{1,3} + A_{1,4} + A_{1,5} + A_{1,6} + A_{1,7} + A_{1,8}, \qquad \text{(S.II.12)}
\end{aligned}$$

where

$$A_{1,1} = \frac{1}{nh}\sum_{i=1}^{n}(K^2 r_p r_p')(X_{h,i})\left\{\varepsilon_i^2 - v(X_i)\right\},$$

is due to the approximation of the (average over the) conditional variance by the squared residuals (i.e. $A_{1,1}$ is the sole remainder that would arise if the true residuals were known and used in place of $\hat{\varepsilon}_i^2$), and, using $r_p(X_i - x)'\hat{\beta} = r_p(X_i - x)'H_p\Gamma^{-1}R'WY/n = r_p(X_{h,i})'\Gamma^{-1}R'WY/n$, the terms $A_{1,k}$, $k = 2, 3, \ldots, 8$ are:

$$\begin{aligned}
A_{1,2} &= \frac{1}{nh}\sum_{i=1}^{n}(K^2 r_p r_p')(X_{h,i})\left\{2\varepsilon_i[m(X_i) - r_p(X_i - x)'\beta_p]\right\}, \\
A_{1,3} &= \frac{1}{nh}\sum_{i=1}^{n}(K^2 r_p r_p')(X_{h,i})\left\{-2\varepsilon_i r_p(X_{h,i})'\right\}\Gamma^{-1}R'W(Y - \check{R}\beta)/n, \\
A_{1,4} &= \frac{1}{nh}\sum_{i=1}^{n}(K^2 r_p r_p')(X_{h,i})\left\{-2[m(X_i) - r_p(X_i - x)'\beta_p]r_p(X_{h,i})'\right\}\Gamma^{-1}R'W(Y - M)/n, \\
A_{1,5} &= \frac{1}{nh}\sum_{i=1}^{n}(K^2 r_p r_p' r_p')(X_{h,i})\Gamma^{-1}R'W(Y - M)/n\left[(Y - M)'/n + 2(M - \check{R}\beta)/n\right]WR\Gamma^{-1}r_p(X_{h,i}), \\
A_{1,6} &= \frac{1}{nh}\sum_{i=1}^{n}(K^2 r_p r_p')(X_{h,i})[m(X_i) - r_p(X_i - x)'\beta_p]^2, \\
A_{1,7} &= \frac{1}{nh}\sum_{i=1}^{n}(K^2 r_p r_p' r_p')(X_{h,i})\left\{-2[m(X_i) - r_p(X_i - x)'\beta_p]\right\}\Gamma^{-1}R'W(M - \check{R}\beta)/n,
\end{aligned}$$

and

$$A_{1,8} = \frac{1}{nh} \sum_{i=1}^{n} (K^2 r_p r_p' r_p')(X_{h,i}) \Gamma^{-1} [R' W (M - \check{R}\beta)/n][(M - \check{R}\beta)'/n W R] \Gamma^{-1} r_p(X_{h,i}).$$

With this notation, we can write $A_1 = e_0' \Gamma^{-1} \left(\hat{\Psi} - \check{\Psi} \right) \Gamma^{-1} e_0 = e_0' \Gamma^{-1} \left(\sum_{k=1}^{8} A_{1,k} \right) \Gamma^{-1} e_0$. The terms $A_{1,1}$ to $A_{1,5}$ will be incorporated into $\tilde{T}$: notice that these terms obey $A_{1,k} = A_{1,k}(\bar{Z}_{\texttt{us}})$ and $A_{1,k}(\mathbb{E}[Z_{\texttt{us}}]) = 0$, and hence these properties will be inherited in the final two lines of Eqn. (S.II.9). However, $A_{1,6}$, $A_{1,7}$, and $A_{1,8}$ do not have these properties, and will thus be incorporated into $\tilde{z}$ and the remainder. Details are below.

Turning to A_2 in Eqn. (S.II.11), using the identity $\Gamma^{-1} - \tilde{\Gamma}^{-1} = \tilde{\Gamma}^{-1} \left(\tilde{\Gamma} - \Gamma \right) \Gamma^{-1}$ and that Γ and Ψ are symmetric, we find that

$$\begin{aligned} A_2 &= e_0' \Gamma^{-1} \check{\Psi} \Gamma^{-1} e_0 - e_0' \tilde{\Gamma}^{-1} \tilde{\Psi} \tilde{\Gamma}^{-1} e_0 \\ &= e_0' \Gamma^{-1} \left(\check{\Psi} - \tilde{\Psi} \right) \Gamma^{-1} e_0 + e_0' \left(\Gamma^{-1} - \tilde{\Gamma}^{-1} \right) \tilde{\Psi} \Gamma^{-1} e_0 + e_0' \left(\Gamma^{-1} - \tilde{\Gamma}^{-1} \right) \tilde{\Psi} \tilde{\Gamma}^{-1} e_0 \\ &= e_0' \Gamma^{-1} \left(\check{\Psi} - \tilde{\Psi} \right) \Gamma^{-1} e_0 - e_0' \tilde{\Gamma}^{-1} \left(\Gamma - \tilde{\Gamma} \right) \Gamma^{-1} \tilde{\Psi} \left(\Gamma^{-1} + \tilde{\Gamma}^{-1} \right) e_0. \end{aligned}$$

All of these terms obey the required properties of $\tilde{T}$.

We now collect the terms from expanding $\hat{\sigma}^{-1} - \tilde{\sigma}^{-1}$ and return to Eqn. (S.II.9). Plugging the terms $A_{1,1}$–$A_{1,8}$ and A_2 into the Taylor expansion in Eqn. (S.II.10), by way of Eqn. (S.II.11), and collecting terms appropriately (i.e. those that belong in $\tilde{T}$ as described above), we have the following, which picks up from Eqn. (S.II.9) and is a precursor to Eqn. (S.II.8):

$$\mathbb{P}[T_{\texttt{us}} < z] = \mathbb{P}\left[\tilde{T}(\bar{Z}_{\texttt{us}}) + U < \tilde{z} \right]. \tag{S.II.13}$$

In this statement, we have made the following constructions:

$$\begin{aligned} \tilde{T} &= \tilde{\sigma}^{-1} s e_0' \Gamma^{-1} R' W (Y - M)/n \\ &\quad + \tilde{\sigma}^{-1} s e_0' \tilde{\Gamma}^{-1} R' W (M - \check{R}\beta)/n - \tilde{\sigma}^{-1} \eta \\ &\quad + \tilde{\sigma}^{-1} s e_0' \left(\Gamma^{-1} - \tilde{\Gamma}^{-1} \right) R' W (M - \check{R}\beta)/n \\ &\quad + \left\{ -\frac{1}{2\tilde{\sigma}^3} \left[e_0' \Gamma^{-1} \left(\sum_{k=1}^{5} A_{1,k} \right) \Gamma^{-1} e_0 + A_2 \right] + \frac{3}{8\tilde{\sigma}^5} \left[e_0' \Gamma^{-1} A_{1,1} \Gamma^{-1} e_0 + A_2 \right]^2 \right\} \\ &\qquad \times \left\{ s e_0' \Gamma^{-1} R' W (Y - M)/n + s e_0' \Gamma^{-1} R' W (M - \check{R}\beta)/n \right\}, \end{aligned}$$

$$\begin{aligned} U = &\left\{ -\frac{1}{2\tilde{\sigma}^3} e_0' \Gamma^{-1} \left(A_{1,6} + A_{1,7} + A_{1,8} \right) \Gamma^{-1} e_0 + \frac{3}{8\tilde{\sigma}^5} \left[e_0' \Gamma^{-1} \left(\sum_{k=2}^{8} A_{1,k} \right) \Gamma^{-1} e_0 \right]^2 - \frac{5}{16} \frac{\left(\hat{\sigma}^2 - \tilde{\sigma}^2 \right)^3}{\bar{\sigma}^7} \right\} \\ &\times \left\{ s e_0' \Gamma^{-1} R' W (Y - M)/n + s e_0' \Gamma^{-1} R' W (M - \check{R}\beta)/n \right\} \end{aligned}$$

$$
- \left\{ -\frac{1}{2\tilde{\sigma}^3} e_0' \tilde{\Gamma}^{-1} \left(\tilde{A}_{1,6} + \tilde{A}_{1,7} + \tilde{A}_{1,8} \right) \tilde{\Gamma}^{-1} e_0 \right\} \eta,
$$

and

$$
\tilde{z} = z - \left\{ \tilde{\sigma}^{-1} - \frac{1}{2\tilde{\sigma}^3} e_0' \tilde{\Gamma}^{-1} \left(\tilde{A}_{1,6} + \tilde{A}_{1,7} + \tilde{A}_{1,8} \right) \tilde{\Gamma}^{-1} e_0 \right\} \eta.
$$

In U and $\tilde{z}$, each $\tilde{A}_{1,k}$ is $A_{1,k}$ where all elements have been replaced by their respective fixed-n expected values, that is,

$$
\begin{aligned}
\tilde{A}_{1,6} &= \mathbb{E}[A_{1,6}] = \mathbb{E}\left[h^{-1}(K^2 r_p r_p')(X_{h,i}) \left[m(X_i) - r_p(X_i - x)'\beta_p \right]^2 \right], \\
\tilde{A}_{1,7} &= -2\mathbb{E}\Big[h^{-1}(K^2 r_p r_p' r_p')(X_{h,i}) \left[m(X_i) - r_p(X_i - x)'\beta_p \right] \Big] \\
&\qquad\qquad \times \tilde{\Gamma}^{-1}\mathbb{E}\Big[h^{-1}(K r_p)(X_{h,j}) \left[m(X_j) - r_p(X_j - x)'\beta_p \right] \Big],
\end{aligned}
$$

and

$$
\tilde{A}_{1,8} = \mathbb{E}\left[h^{-1}(K^2 r_p r_p')(X_{h,i}) \mathbb{E}\left[h^{-1} r_p(X_{h,i})' \tilde{\Gamma}^{-1} (K r_p)(X_{h,j}) \left[m(X_j) - r_p(X_j - x)'\beta_p \right] \Big| X_i \right]^2 \right].
$$

The next step in the proof is to show that, for $r_* = \max\{s^{-2}, \eta^2, h^{p+1}\}$ (i.e., the slowest decaying), it holds that

$$
\frac{1}{r_*}\mathbb{P}[|U| > r_n] \to 0, \qquad \text{for some } r_n = o(r_*). \tag{S.II.14}
$$

This result is established by Lemma S.II.4 in Section S.II.6.3 below. This, together with Eqn. (S.II.13), implies Eqn. (S.II.8).

Under Assumption S.II.3, an Edgeworth expansion holds for $\tilde{T}$ up to $o(s^{-2} + s^{-1}\eta + \eta^2)$. Thus, for a smooth function $G(z)$, we have $\mathbb{P}[\tilde{T} < z] = G(z) + o(s^{-2} + s^{-1}\eta + \eta^2)$. Therefore, a Taylor expansion gives

$$
\mathbb{P}[\tilde{T} < \tilde{z}] = G(z) - G^{(1)}(z) \left\{ \tilde{\sigma}^{-1} - \frac{1}{2\tilde{\sigma}^3} e_0' \Gamma^{-1} \left(\tilde{A}_{1,6} + \tilde{A}_{1,7} + \tilde{A}_{1,8} \right) \Gamma^{-1} e_0 \right\} + o(s^{-2} + s^{-1}\eta + \eta^2),
$$

which together with Eqn. (S.II.8) establishes the validity of the Edgeworth expansion. The terms of the expansion are computed in Section S.II.6.4 below. □

S.II.6.2 Proof of Theorem S.II.1(b) & (c)

To prove parts (b) and (c) of Theorem S.II.1 the same steps are required, and so we will not pursue all the details here. Indeed, the same expansions are performed and the same bounds computed on objects which are conceptually similar, only taking into account the bias correction (in the numerator for (b), and also in the denominator for (c)). The bias correction will result in essentially two changes: first, many more terms like $\Gamma - \tilde{\Gamma}$ appear, and second, the bias expressions and rates change. To illustrate, we will list several key points where these changes manifest. This list is not exhaustive, but it will show that the same methods used above still apply.

First, for the numerator of $T_{\texttt{bc}}$ and $T_{\texttt{rbc}}$, recall that the estimator $\hat{m}$ is

$$\hat{m} = \Big\{ e_0' \Gamma_p^{-1} R_p' W_p \Big\} Y/n,$$

while the bias corrected estimator is

$$\hat{m} - \hat{B}_m = \Big\{ e_0' \Gamma_p^{-1} \left(R_p' W_p - \rho^{p+1} \Lambda_{p,1} e_{p+1}' \Gamma_q^{-1} R_q' W_q \right) \Big\} Y/n.$$

Comparing these two expressions, it can be seen that the terms in the proof above that involve $\Gamma_p - \tilde{\Gamma}_p$ will now additionally involve $\Gamma_q - \tilde{\Gamma}_q$ and $\Lambda_{p,1} - \tilde{\Lambda}_{p,1}$, whereas those that with $e_0' \tilde{\Gamma}_p^{-1} R_p' W_p$ will now have $e_0' \tilde{\Gamma}_p^{-1} \left(R_p' W_p - \rho^{p+1} \tilde{\Lambda}_{p,1} e_{p+1}' \tilde{\Gamma}_q^{-1} R_q' W_q \right)$ instead. To give a concrete example, consider the third line of Eqn. (S.II.9),

$$\tilde{\sigma}_{\texttt{us}}^{-1} s e_0' \left(\Gamma_p^{-1} - \tilde{\Gamma}_p^{-1} \right) R_p' W_p (M - \check{R}_p \beta_p)/n,$$

which becomes a piece of the function $\tilde{T}$. For part (b) Theorem S.II.1, treating $T_{\texttt{bc}}$, this will become

$$\begin{aligned} &\tilde{\sigma}_{\texttt{us}}^{-1} s e_0' \left(\Gamma_p^{-1} - \tilde{\Gamma}_p^{-1} \right) R_p' W_p (M - \check{R}_{p+1} \beta_{p+1})/n \\ &\qquad - s e_0' \rho^{p+1} \left(\Gamma_p^{-1} \Lambda_{p,1} e_{p+1}' \Gamma_q^{-1} - \tilde{\Gamma}_p^{-1} \tilde{\Lambda}_{p,1} e_{p+1}' \tilde{\Gamma}_q^{-1} \right) R_q' W_q (M - \check{R}_q \beta_q)/n, \end{aligned}$$

and part (c) will have the same but with $\tilde{\sigma}_{\texttt{rbc}}^{-1}$. Then, since

$$\begin{aligned} \Gamma_p^{-1} \Lambda_{p,1} e_{p+1}' \Gamma_q^{-1} - \tilde{\Gamma}_p^{-1} \tilde{\Lambda}_{p,1} e_{p+1}' \tilde{\Gamma}_q^{-1} &= \left(\Gamma_p^{-1} - \tilde{\Gamma}_p^{-1} \right) \Lambda_{p,1} e_{p+1}' \Gamma_q^{-1} \\ &\quad + \tilde{\Gamma}_p^{-1} \left(\Lambda_{p,1} - \tilde{\Lambda}_{p,1} \right) e_{p+1}' \Gamma_q^{-1} + \tilde{\Gamma}_p^{-1} \tilde{\Lambda}_{p,1} e_{p+1}' \left(\Gamma_q^{-1} - \tilde{\Gamma}_q^{-1} \right), \end{aligned}$$

this term is handled identically, since the appropriate Cramér's condition is assumed.

Consider now the denominator of the Studentized statistics. For part (b), there is no change as $\hat{\sigma}_{\texttt{us}}^2$ is still used, and so the terms involving $A_{1,k}$ and A_2 will be identical. However, for $T_{\texttt{rbc}}$, we must account for changes of the above form, but also that the residuals are estimated with the degree q fit: $\hat{\varepsilon}_i = y_i - r_q(X_i - x)' \hat{\beta}_q$ instead of degree p. With these changes in mind, the analogue of Eqn. (S.II.11) will be

$$\hat{\sigma}_{\texttt{rbc}}^2 - \tilde{\sigma}_{\texttt{rbc}}^2 = e_0' \Gamma_p^{-1} \left(\hat{\Psi}_q - \check{\Psi}_q \right) \Gamma_p^{-1} e_0 + \left(e_0' \Gamma_p^{-1} \check{\Psi}_q \Gamma_p^{-1} e_0 - e_0' \tilde{\Gamma}_p^{-1} \tilde{\Psi}_q \tilde{\Gamma}_p^{-1} e_0 \right). \tag{S.II.15}$$

The second term will proceed as above, though $\check{\Psi}_p - \tilde{\Psi}_p$ will be replaced by

$$\check{\Psi}_q - \tilde{\Psi}_q = \frac{1}{nh} \sum_{i=1}^{n} \left\{ \tilde{\ell}_{\texttt{bc}}^0(X_i) \tilde{\ell}_{\texttt{bc}}^0(X_i)' v(X_i) - \mathbb{E}\left[\tilde{\ell}_{\texttt{bc}}^0(X_i) \tilde{\ell}_{\texttt{bc}}^0(X_i)' v(X_i) \right] \right\},$$

where $\tilde{\ell}^0_{\mathtt{bc}}(X_i) = (Kr_p)(X_{h,i}) - \rho^{p+1}\tilde{\Lambda}_{p,1}\tilde{\Gamma}_q^{-1}(Lr_p)(\rho X_{h,i})$ (cf. Section S.II.1.2, the function $\ell^0_{\mathtt{bc}}$ therein is $\ell^0_{\mathtt{bc}}(X_i) = e_0'\tilde{\Gamma}_p^{-1}\tilde{\ell}^0_{\mathtt{bc}}(X_i)$). To use similar notation,

$$\check{\Psi}_p - \tilde{\Psi}_p = \frac{1}{nh}\sum_{i=1}^n \left\{ \tilde{\ell}^0_{\mathtt{us}}(X_i)\tilde{\ell}^0_{\mathtt{us}}(X_i)'v(X_i) - \mathbb{E}\left[\tilde{\ell}^0_{\mathtt{us}}(X_i)\tilde{\ell}^0_{\mathtt{us}}(X_i)'v(X_i)\right]\right\}.$$

Then, expanding $\tilde{\ell}^0_{\mathtt{bc}}(X_i)$ shows that $\check{\Psi}_q - \tilde{\Psi}_q$ is equal to

$$\left(\check{\Psi}_p - \tilde{\Psi}_p\right) + \rho^{2(p+1)+1}\tilde{\Lambda}_{p,1}\tilde{\Gamma}_q^{-1}\frac{1}{nb}\sum_{i=1}^n \left\{(L^2r_qr_q')(X_{b,i})v(X_i) - \mathbb{E}\left[(L^2r_qr_q')(X_{b,i})v(X_i)\right]\right\}\tilde{\Gamma}_q^{-1}\tilde{\Lambda}_{p,1}$$
$$- \rho^{(p+1)+1}2\frac{1}{nh}\sum_{i=1}^n \left\{(Kr_p)(X_{h,i})(Lr_q')(\rho X_{h,i})v(X_i) - \mathbb{E}\left[(Kr_p)(X_{h,i})(Lr_q')(\rho X_{h,i})v(X_i)\right]\right\}\tilde{\Gamma}_q^{-1}\tilde{\Lambda}_{p,1},$$

and since all these terms still obey the appropriate Cramér's condition, the same steps apply. (The extra factor of ρ in $\rho^{2(p+1)+1}$ and $\rho^{(p+1)+1}$ accounts for the fact that $\hat{\sigma}^2_{\mathtt{rbc}}$ is scaled by (nh) instead of (nb), but the W_q matrixes contribute a b^{-1}.)

The first term of Eqn. (S.II.15) will also follow by the same method as in the prior proof, but more care must be taken as many more terms will be present because $\hat{\Psi}_q - \check{\Psi}_q$ consists of the following three terms, representing the variance of $\hat{m}$, the variance of $\hat{B}_m$, and their covariance, respectively:

$$\begin{aligned}\hat{\Psi}_q - \check{\Psi}_q = {} & hR_p'W_p\left(\hat{\Sigma}_q - \Sigma\right)W_pR_p/n \\ & + h\rho^{2(p+1)}\Lambda_{p,1}\Gamma_q^{-1}\left(R_q'W_q\hat{\Sigma}_qW_qR_q\right)\Gamma_q^{-1}\Lambda_{p,1}'/n - h\rho^{2(p+1)}\tilde{\Lambda}_{p,1}\tilde{\Gamma}_q^{-1}\left(R_q'W_q\Sigma W_qR_q\right)\tilde{\Gamma}_q^{-1}\tilde{\Lambda}_{p,1}'/n \\ & - 2h\rho^{p+1}R_p'W_p\left(\hat{\Sigma}_qW_qR_q\Gamma_p^{-1}\Lambda_{p,1}'\Gamma - \Sigma W_qR_q\tilde{\Gamma}_p^{-1}\tilde{\Lambda}_{p,1}'\right)/n.\end{aligned}$$

The first of these three is as in the prior proof, and yields the same $A_{1,1}$–$A_{1,8}$, only with the bias of a q-degree fit: $m(X_i) - r_q(X_i - x)'\beta_q$. If we define

$$\check{\check{\Psi}}_q := \frac{1}{nb}\sum_{i=1}^n (L^2r_qr_q')(X_{b,i})v(X_i)$$

then the second term of $\hat{\Psi}_q - \check{\Psi}_q$ is equal to

$$\begin{aligned}&\rho^{1+2(p+1)}\Lambda_{p,1}\Gamma_q^{-1}\left\{\frac{1}{nb}\sum_{i=1}^n (L^2r_qr_q')(X_{b,i})\left\{\hat{\varepsilon}_i^2 - v(X_i)\right\}\right\}\Gamma_q^{-1}\Lambda_{p,1} \\ & + \rho^{1+2(p+1)}\left(\Lambda_{p,1} - \tilde{\Lambda}_{p,1}\right)\Gamma_q^{-1}\check{\check{\Psi}}_q\Gamma_q^{-1}\Lambda_{p,1} \\ & + \rho^{1+2(p+1)}\tilde{\Lambda}_{p,1}\left(\Gamma_q^{-1} - \tilde{\Gamma}_q^{-1}\right)\check{\check{\Psi}}_q\Gamma_q^{-1}\Lambda_{p,1} \\ & + \rho^{1+2(p+1)}\tilde{\Lambda}_{p,1}\tilde{\Gamma}_q^{-1}\check{\check{\Psi}}_q\left(\Gamma_q^{-1} - \tilde{\Gamma}_q^{-1}\right)\Lambda_{p,1} \\ & + \rho^{1+2(p+1)}\tilde{\Lambda}_{p,1}\tilde{\Gamma}_q^{-1}\check{\check{\Psi}}_q\tilde{\Gamma}_q^{-1}\left(\Lambda_{p,1} - \tilde{\Lambda}_{p,1}\right).\end{aligned}$$

The first of these terms will also give rise to versions of $A_{1,1}$–$A_{1,8}$, only with the bias of a q-degree fit and changing K to L, p to q, h to b, etc, and will thus be treated exactly as above. The rest of these are incorporated into $\tilde{T}_{\texttt{rbc}}$, similar to how A_2 is treated, because Cramér's condition is satisfied. The third and final piece of $\hat{\Psi}_q - \check{\Psi}_q$ is equal to

$$\begin{aligned}
&-2\rho^{1+(p+1)}\left\{\frac{1}{nh}\sum_{i=1}^{n}(Kr_p)(X_{h,i})(Lr_q')(X_{h,i}\rho)\left\{\hat{\varepsilon}_i^2 - v(X_i)\right\}\right\}\Gamma_q^{-1}\Lambda_{p,1}' \\
&\quad - 2\rho^{1+(p+1)}\check{\check{\Psi}}_q\left(\Gamma_q^{-1} - \tilde{\Gamma}_q^{-1}\right)\Lambda_{p,1}' \\
&\quad - 2\rho^{1+(p+1)}\check{\check{\Psi}}_q\tilde{\Gamma}_q^{-1}\left(\Lambda_{p,1} - \tilde{\Lambda}_{p,1}\right),
\end{aligned}$$

and thus is entirely analogous, with yet another version of $A_{1,1}$–$A_{1,8}$ defined for the remainder in the first line, and the second two easily incorporated into $\tilde{T}_{\texttt{rbc}}$.

From these arguments, it is clear that the analogue of Lemma S.II.4 will hold for these cases as well: the same fundamental pieces are involved, and thus the same arguments will apply, just as above.

S.II.6.3 Lemmas

Our proof of Theorem S.II.1 relies on the following lemmas. The first gives generic results used to derive rate bounds on the probability of deviations of the necessary terms. Some such results are collected in Lemma S.II.2. Lemma S.II.4 shows how to use the previous results to establish negligibility of the remainder terms required for Eqn. (S.II.14).

As above, we will generally omit the details required for Theorem S.II.1 parts (b) and (c), to save space. These are entirely analogous, as can be seen from the steps in Lemma S.II.2. Indeed, the first results are stated in terms of the kernel K and bandwidth h, but continue to hold for L and b under the obvious substitutions and appropriate assumptions.

Throughout proofs C shall be a generic conformable constant that may take different values in different places. If more than one constant is needed, C_1, C_2, ..., will be used.

Lemma S.II.1. *Let the conditions of Theorem S.II.1 hold and let $g(\cdot)$ and $t(\cdot)$ be continuous scalar functions.*

(a) *For some $\delta > 0$,*

$$s^2\mathbb{P}\left[\left|s^{-2}\sum_{i=1}^{n}\left\{(Kt)(X_{h,i})g(X_i) - \mathbb{E}[(Kt)(X_{h,i})g(X_i)]\right\}\right| > \delta s^{-1}\log(s)^{1/2}\right] \to 0.$$

(b) *For some $\delta > 0$,*

$$s^2\mathbb{P}\left[\left|s^{-1}\sum_{i=1}^{n}\left\{(Kt)(X_{h,i})g(X_i)\varepsilon_i\right\}\right| > \delta\log(s)^{1/2}\right] \to 0.$$

The same holds with $\varepsilon_i^2 - v(X_i)$ in place of ε_i, since it is conditionally mean zero and has more than four moments.

(c) *For any $\delta > 0$, an integer k, and any $\gamma > 0$,*

$$\frac{1}{h^{p+1}}\mathbb{P}\left[\left|s^{-2}\sum_{i=1}^{n}(Kt)(X_{h,i})g(X_i)\left[m(X_i) - r_p(X_i - x)'\beta_p\right]^k\right| > \delta h^{(k-1)(p+1)}\log(s)^\gamma\right] \to 0.$$

(d) *For any $\delta > 0$ and any $\gamma > 0$,*

$$s^2\mathbb{P}\left[\left|s^{-2}\sum_{i=1}^{n}(Kt)(X_{h,i})g(X_i)\varepsilon_i\left[m(X_i) - r_p(X_i - x)'\beta_p\right]\right| > \delta h^{p+1}\log(s)^\gamma\right] \to 0.$$

(e) *For any $\delta > 0$, an integer k, and any $\gamma > 0$,*

$$s^2\mathbb{P}\Bigg[\Bigg|s^{-2}\sum_{i=1}^{n}\Big\{(Kt)(X_{h,i})g(X_i)(m(X_i) - r_p(X_i - x)'\beta_p)^k$$
$$- \mathbb{E}\left[(Kt)(X_{h,i})g(X_i)(m(X_i) - r_p(X_i - x)'\beta_p)^k\right]\Big\}\Bigg| > \delta h^{k(p+1)}\log(s)^\gamma\Bigg] \to 0.$$

Proof of Lemma S.II.1(a). Because the kernel function has compact support and t and g are continuous, we have

$$|(Kt)(X_{h,i})g(X_i) - \mathbb{E}[(Kt)(X_{h,i})g(X_i)]| < C_1.$$

Further, by a change of variables and using the assumptions on f, g and t:

$$\mathbb{V}[(Kt)(X_{h,i})g(X_i)] \leq \mathbb{E}\left[(Kt)(X_{h,i})^2 g(X_i)^2\right] = \int f(X_i)(Kt)(X_{h,i})^2 g(X_i)^2 dX_i$$
$$= h\int f(x - uh)g(x - uh)(Kt)(u)^2 du \leq C_2 h.$$

Therefore, by Bernstein's inequality

$$s^2\mathbb{P}\left[\left|\frac{1}{s^2}\sum_{i=1}^{n}\{(Kt)(X_{h,i})g(X_i) - \mathbb{E}[(Kt)(X_{h,i})g(X_i)]\}\right| > \delta s^{-1}\log(s)^{1/2}\right]$$
$$\leq 2s^2\exp\left\{-\frac{(s^4)(\delta s^{-1}\log(s)^{1/2})^2/2}{C_2 s^2 + C_1 s^2\delta s^{-1}\log(s)^{1/2}/3}\right\}$$
$$= 2\exp\{2\log(s)\}\exp\left\{-\frac{\delta^2\log(s)/2}{C_2 + C_1\delta s^{-1}\log(s)^{1/2}/3}\right\}$$
$$= 2\exp\left\{\log(s)\left[2 - \frac{\delta^2/2}{C_2 + C_1\delta s^{-1}\log(s)^{1/2}/3}\right]\right\},$$

which vanishes for any δ large enough, as $s^{-1}\log(s)^{1/2} \to 0$. □

Proof of Lemma S.II.1(b). For a sequence $r_n \to \infty$ to be given later, define

$$H_i = s^{-1}(Kt)(X_{h,i})g(X_i)\left(Y_i\mathbb{1}\{Y_i \le r_n\} - \mathbb{E}[Y_i\mathbb{1}\{Y_i \le r_n\} \mid X_i]\right)$$

and

$$T_i = s^{-1}(Kt)(X_{h,i})g(X_i)\left(Y_i\mathbb{1}\{Y_i > r_n\} - \mathbb{E}[Y_i\mathbb{1}\{Y_i > r_n\} \mid X_i]\right).$$

By the conditions on $g(\cdot)$ and $t(\cdot)$ and the kernel function,

$$|H_i| < C_1 s^{-1} r_n$$

and

$$\begin{aligned}
\mathbb{V}[H_i] = s^{-2}\mathbb{V}[(Kt)(X_{h,i})g(X_i)Y_i\mathbb{1}\{Y_i \le r_n\}] &\le s^{-2}\mathbb{E}\left[(Kt)(X_{h,i})^2 g(X_i)^2 Y_i^2 \mathbb{1}\{Y_i \le r_n\}\right] \\
&\le s^{-2}\mathbb{E}\left[(Kt)(X_{h,i})^2 g(X_i)^2 Y_i^2\right] \\
&= s^{-2}\int (Kt)(X_{h,i})^2 g(X_i)^2 v(X_i) f(X_i) dX_i \\
&= s^{-2} h \int (Kt)(u)^2 (gvf)(x - uh) du \\
&\le C_2/n.
\end{aligned}$$

Therefore, by Bernstein's inequality

$$\begin{aligned}
s^2\mathbb{P}\left[\left|\sum_{i=1}^n H_i\right| > \delta\log(s)^{1/2}\right] &\le 2s^2 \exp\left\{-\frac{\delta^2\log(s)/2}{C_2 + C_1 s^{-1} r_n \delta \log(s)^{1/2}/3}\right\} \\
&\le 2\exp\{2\log(s)\}\exp\left\{-\frac{\delta^2\log(s)/2}{C_2 + C_1 s^{-1} r_n \delta \log(s)^{1/2}/3}\right\} \\
&\le 2\exp\left\{\log(s)\left[2 - \frac{\delta^2/2}{C_2 + C_1 s^{-1} r_n \delta \log(s)^{1/2}/3}\right]\right\},
\end{aligned}$$

which vanishes for δ large enough as long as $s^{-1} r_n \log(s)^{1/2}$ does not diverge.

Next, by Markov's inequality and the moment condition on Y of Assumption S.II.1

$$\begin{aligned}
s^2\mathbb{P}\left[\left|\sum_{i=1}^n T_i\right| > \delta\log(s)^{1/2}\right] &\le s^2\frac{1}{\delta^2\log(s)}\mathbb{E}\left[\left|\sum_{i=1}^n T_i\right|^2\right] \\
&\le s^2\frac{1}{\delta^2\log(s)} n\mathbb{E}\left[T_i^2\right] \\
&\le s^2\frac{1}{\delta^2\log(s)} n\mathbb{V}\left[s^{-1}(Kt)(X_{h,i})g(X_i)Y_i\mathbb{1}\{Y_i > r_n\}\right]
\end{aligned}$$

$$
\begin{aligned}
&\leq s^2 \frac{1}{\delta^2 \log(s)} n s^{-2} \mathbb{E}\left[(Kt)(X_{h,i})^2 g(X_i)^2 Y_i^2 \mathbb{1}\{Y_i > r_n\}\right] \\
&\leq s^2 \frac{1}{\delta^2 \log(s)} n s^{-2} \mathbb{E}\left[(Kt)(X_{h,i})^2 g(X_i)^2 |Y_i|^{2+\xi} r_n^{-\eta}\right] \\
&\leq s^2 \frac{1}{\delta^2 \log(s)} n s^{-2} (C h r_n^{-\xi}) \\
&\leq \frac{C}{\delta^2} \frac{s^2}{\log(s) r_n^{\xi}},
\end{aligned}
$$

which vanishes if $s^2 \log(s)^{-1} r_n^{-\xi} \to 0$.

It thus remains to choose r_n such that $s^{-1} r_n \log(s)^{1/2}$ does not diverge and $s^2 \log(s)^{-1} r_n^{-\xi} \to 0$. This can be accomplished by setting $r_n = s^\gamma$ for any $2/\xi \leq \gamma < 1$, which is possible as $\xi > 2$. □

Proof of Lemma S.II.1(c). By Markov's inequality

$$
\begin{aligned}
&\frac{1}{h^{p+1}} \mathbb{P}\left[\left| s^{-2} \sum_{i=1}^{n} (Kt)(X_{h,i}) g(X_i) \left[m(X_i) - r_p(X_i - x)'\beta_p\right]^k \right| > \delta h^{(k-1)(p+1)} \log(s)^\gamma \right] \\
&\quad \leq \frac{1}{h^{p+1}} \frac{1}{\delta h^{(k-1)(p+1)} \log(s)^\gamma} \mathbb{E}\left[h^{-1} (Kt)(X_{h,i}) g(X_i) \left[m(X_i) - r_p(X_i - x)'\beta_p\right]^k\right] \\
&\quad \leq \frac{1}{\delta h^{k(p+1)} \log(s)^\gamma} h^{k(p+1)} \mathbb{E}\left[h^{-1} (Kt)(X_{h,i}) g(X_i) \left[h^{-p-1}(m(X_i) - r_p(X_i - x)'\beta_p)\right]^k\right] \\
&\quad = O(\log(s)^{-\gamma}) = o(1).
\end{aligned}
$$

This relies on the following calculation, which uses the conditions placed on $m(\cdot)$:

$$
\begin{aligned}
&\mathbb{E}\left[h^{-1} \left((Kt)(X_{h,i}) g(X_i) \varepsilon_i\right) \left[m(X_i) - r_p(X_i - x)'\beta_p\right]^k\right] \\
&\quad = h^{-1} \int (gfv)(X_i)(Kt)(X_{h,i}) \left[m(X_i) - r_p(X_i - x)'\beta_p\right]^k dX_i \\
&\quad = h^{-1} \int (gfv)(X_i)(Kt)(X_{h,i}) \left(\frac{m^{(p+1)}(\bar{x})}{(p+1)!}(X_i - x)^{p+1}\right)^k dX_i \\
&\quad = h^{k(p+1)} h^{-1} \int (gfv)(X_i)(Kt)(X_{h,i}) \left(\frac{m^{(p+1)}(\bar{x})}{(p+1)!} X_{h,i}^{p+1}\right)^k dX_i \\
&\quad = C h^{k(p+1)} h^{-1} \int (gfv)(X_i)(Kt)(X_{h,i}) X_{h,i}^{k(p+1)} dX_i \\
&\quad = C h^{k(p+1)} \int (gfv)(x - uh)(Kt)(u) u^{k(p+1)} du \\
&\quad \asymp h^{k(p+1)}.
\end{aligned}
$$

□

Proof of Lemma S.II.1(d). By Markov's inequality, since ε_i is conditionally mean zero, we have

$$
s^2 \mathbb{P}\left[\left| s^{-2} \sum_{i=1}^{n} (Kt)(X_{h,i}) g(X_i) \varepsilon_i \left[m(X_i) - r_p(X_i - x)'\beta_p\right] \right| > \delta h^{p+1} \log(s)^\gamma \right]
$$

$$\leq s^2 \frac{1}{\delta h^{2(p+1)} \log(s)^{2\gamma}} \frac{1}{s^2} \mathbb{E}\left[h^{-1}\left((Kt)(X_{h,i})g(X_i)\varepsilon_i\right)^2 \left[m(X_i) - r_p(X_i - x)'\beta_p\right]^2\right]$$

$$\leq \frac{s^2 h^{2(p+1)}}{\delta s^2 h^{2(p+1)} \log(s)^{\gamma}} \mathbb{E}\left[h^{-1}\left((Kt)(X_{h,i})g(X_i)\varepsilon_i\right)^2 \left[h^{-p-1}(m(X_i) - r_p(X_i - x)'\beta_p)\right]^2\right]$$

$$\asymp \log(s)^{-2\gamma} \to 0,$$

where we rely on the same argument as above to compute the bias rate. □

Proof of Lemma S.II.1(e). Follows from identical steps to S.II.1(d). □

To illustrate how the above Lemma is used for the objects under study, we present the following collection of results. This is not meant to be an exhaustive list of all such results needed to prove all parts of Theorem S.II.1, but any and all omitted terms follow by identical reasoning.

Lemma S.II.2. *Let the conditions of Theorem S.II.1 hold.*

- **(a)** *For some $\delta > 0$, $r_*^{-1}\mathbb{P}[|\Gamma_p - \tilde{\Gamma}_p| > s^{-1}\log(s)^{1/2}] \to 0$. Consequently, there exists a constant $C_\Gamma < \infty$ such that $\mathbb{P}[\Gamma_p^{-1} > 2C_\Gamma] = o(s^{-2})$ and so the prior rate result holds for $|\Gamma_p^{-1} - \tilde{\Gamma}_p^{-1}|$ as well. Finally, these same results hold for Γ_q as well.*

- **(b)** *For some $\delta > 0$, $r_*^{-1}\mathbb{P}[|\Lambda_{p,1} - \tilde{\Lambda}_{p,1}| > s^{-1}\log(s)^{1/2}] \to 0$.*

- **(c)** *For some $\delta > 0$,*

$$s^2\mathbb{P}\left[\left|s^{-1}\sum_{i=1}^n \{(Kr_p)(X_{h,i})\varepsilon_i\}\right| > \delta\log(s)^{1/2}\right] \to 0.$$

- **(d)** *For any $\delta > 0$ and $\gamma > 0$,*

$$\frac{1}{h^{p+1}}\mathbb{P}\left[\left|s^{-2}\sum_{i=1}^n \left\{(Kr_p)(X_{h,i})\left[m(X_i) - r_p(X_i - x)'\beta_p\right]\right\}\right| > \delta\log(s)^{\gamma}\right] \to 0.$$

- **(e)** *There is some constant C_Ψ such that $\mathbb{P}[\check{\Psi}_p > 2C_\Psi] = o(s^{-2})$.*

Proof of Lemma S.II.2(a). A typical element of $\Gamma_p - \tilde{\Gamma}_p$ is, for some integer $k \leq 2p$,

$$\frac{1}{nh}\sum_{i=1}^n \left\{K(X_{h,i})\mathcal{X}_{h,i}^k - \mathbb{E}\left[K(X_{h,i})\mathcal{X}_{h,i}^k\right]\right\}.$$

Therefore, the result follows by applying Lemma S.II.1(a) to each element. Next, note that under the maintained assumptions

$$\tilde{\Gamma}_p = \mathbb{E}\left[h^{-1}(Kr_pr_p')(X_{h,i})\right] = h^{-1}\int (Kr_pr_p')(X_{h,i})f(X_i)dX_i = \int (Kr_pr_p')(u)f(x-uh)du$$

is bounded away from zero and infinity for n large enough. Therefore, there is a $C_\Gamma < \infty$ such that $|\tilde{\Gamma}_p^{-1}| < C_\Gamma$ and then

$$\begin{aligned}\mathbb{P}\left[\Gamma_p^{-1} > 2C_\Gamma\right] &= \mathbb{P}\left[\left(\Gamma_p^{-1} - \tilde{\Gamma}_p^{-1}\right) + \tilde{\Gamma}_p^{-1} > 2C_\Gamma\right] \\ &\leq \mathbb{P}\left[\Gamma_p^{-1} - \tilde{\Gamma}_p^{-1} > s^{-1}\log(s)^{1/2}\right] + \mathbb{P}\left[\tilde{\Gamma}_p^{-1} > 2C_\Gamma - s^{-1}\log(s)^{1/2}\right] \\ &= o(s^{-2}).\end{aligned}$$

The third result follows from these two and the identity $\Gamma_p^{-1} - \tilde{\Gamma}_p^{-1} = \tilde{\Gamma}_p^{-1}(\tilde{\Gamma}_p - \Gamma_p)\Gamma_p^{-1}$.

Finally, for Γ_q, the identical steps apply with L, q, and b in place of K, p, and h. □

Proof of Lemma S.II.2(b). Follows from identical steps to the previous result. □

Proof of Lemma S.II.2(c). Follows from identical steps, but using Lemma S.II.1(b) in place of Lemma S.II.1(a). □

Proof of Lemma S.II.2(d). Follows from identical steps, but using Lemma S.II.1(c) in place of Lemma S.II.1(a). □

Proof of Lemma S.II.2(e). A typical element of $\check{\Psi}_p$ is

$$\frac{1}{nh}\sum_{i=1}^{n}(K^2 r_p r_p')(X_{h,i})v(X_i),$$

and hence under the maintained assumptions the result follows just as the comparable result on Γ_p. □

We next state, without proof, the following fact about the rates appearing in all these Lemmas, which follows from elementary inequalities.

Lemma S.II.3. *If $r_1 = O(r_1')$ and $r_2 = O(r_2')$, for sequences of positive numbers r_1, r_1', r_2, and r_2' and if a sequence of nonnegative random variables obeys $(r_1)^{-1}\mathbb{P}[U_n > r_2] \to 0$ it also holds that $(r_1')^{-1}\mathbb{P}[U_n > r_2'] \to 0$.*

In particular, since $r_* = \max\{s^{-2}, \eta^2, s^{-1}\eta\}$ is defined as the slowest vanishing of the rates, then $r_1^{-1}\mathbb{P}[|U'| > r_n] = o(1)$ implies $r_*^{-1}\mathbb{P}[|U'| > r_n] = o(1)$, for r_1 equal to any of s^{-2}, η^2, or $s^{-1}\eta$. Similarly, r_n may be chosen as any sequence that obeys $r_n = o(r_*)$. Thus, for different pieces of U defined in Eqn. (S.II.14), we may make different choices for these two sequences, as convenient.

The next Lemma proves Eqn. (S.II.14), a crucial step in the proof of Theorem S.II.1(a). Because this result only involves undersmoothing, we will omit the subscript p as above.

Lemma S.II.4. *Let the conditions of Theorem S.II.1(a) hold. Then Eqn. (S.II.14) holds, namely, for some $r_n = o(r_*)$*

$$\frac{1}{r_*}\mathbb{P}[|U| > r_n] \to 0.$$

Proof. Recall the definition:

$$U = \left\{ -\frac{1}{2\tilde{\sigma}^3} e_0'\Gamma^{-1}\left(A_{1,6} + A_{1,7} + A_{1,8}\right)\Gamma^{-1}e_0 + \frac{3}{8\tilde{\sigma}^5}\left[e_0'\Gamma^{-1}\left(\sum_{k=2}^{8} A_{1,k}\right)\Gamma^{-1}e_0\right]^2 - \frac{5}{16}\frac{\left(\hat{\sigma}^2 - \tilde{\sigma}^2\right)^3}{\bar{\sigma}^7} \right\}$$
$$\times \left\{ s e_0'\Gamma^{-1}R'W(Y-M)/n + s e_0'\Gamma^{-1}R'W(M - \check{R}\beta)/n \right\}$$
$$- \left\{ -\frac{1}{2\tilde{\sigma}^3} e_0'\tilde{\Gamma}^{-1}\left(\tilde{A}_{1,6} + \tilde{A}_{1,7} + \tilde{A}_{1,8}\right)\tilde{\Gamma}^{-1}e_0 \right\} \eta.$$

To fully prove the claim of the lemma, we must fully expand U and bound each piece. First, we present complete details on two terms. The remainder are entirely analogous, as discussed below. Consider the pieces involving $A_{1,6}$, namely:

$$e_0'\Gamma^{-1}A_{1,6}\Gamma^{-1}e_0\left\{ s e_0'\Gamma^{-1}R'W(Y-M)/n + s e_0'\Gamma^{-1}R'W(M-\check{R}\beta)/n \right\} - e_0'\tilde{\Gamma}^{-1}\tilde{A}_{1,6}\tilde{\Gamma}^{-1}e_0\,\eta.$$

The first of these is

$$\begin{aligned}
e_0'\Gamma^{-1}A_{1,6}\Gamma^{-1}e_0 s e_0'\Gamma^{-1}R'W(Y-M)/n &= e_0'\Gamma^{-1}\left(A_{1,6} - \tilde{A}_{1,6}\right)\Gamma^{-1}e_0 s e_0'\Gamma^{-1}R'W(Y-M)/n \\
&\quad + e_0'\left(\Gamma^{-1} - \tilde{\Gamma}^{-1}\right)\tilde{A}_{1,6}\Gamma^{-1}e_0 s e_0'\Gamma^{-1}R'W(Y-M)/n \\
&\quad + e_0'\tilde{\Gamma}^{-1}\tilde{A}_{1,6}\left(\Gamma^{-1} - \tilde{\Gamma}^{-1}\right)e_0 s e_0'\Gamma^{-1}R'W(Y-M)/n \\
&\quad + e_0'\tilde{\Gamma}^{-1}\tilde{A}_{1,6}\tilde{\Gamma}^{-1}e_0 s e_0'\left(\Gamma^{-1} - \tilde{\Gamma}^{-1}\right)R'W(Y-M)/n \\
&\quad + e_0'\tilde{\Gamma}^{-1}\tilde{A}_{1,6}\tilde{\Gamma}^{-1}e_0 s e_0'\tilde{\Gamma}^{-1}R'W(Y-M)/n. \\
&=: U_{1,1} + U_{1,2} + U_{1,3} + U_{1,4} + U_{1,5}
\end{aligned}$$

We now bound each remainder in turn. First, for $r_n = h^{p+1}\log(s)^{-1/2}$, we have

$$\begin{aligned}
s^2\mathbb{P}\left[|U_{1,1}| > r_n\right] &= s^2\mathbb{P}\left[\left|e_0'\Gamma^{-1}\left(A_{1,6} - \tilde{A}_{1,6}\right)\Gamma^{-1}e_0 s e_0'\Gamma^{-1}R'W(Y-M)/n\right| > r_n\right] \\
&\leq s^2\mathbb{P}\left[8C_\Gamma^3\left|A_{1,6} - \tilde{A}_{1,6}\right| > \log(s)^{-1/2}r_n\right] \\
&\qquad + s^2\mathbb{P}\left[\left|s^{-1}\sum_{i=1}^{n}\{(Kr_p)(X_{h,i})\varepsilon_i\}\right| > \log(s)^{1/2}\right] + s^2 3\mathbb{P}\left[\Gamma_p^{-1} > 2C_\Gamma\right] \\
&= s^2\mathbb{P}\left[8C_\Gamma^3\left|A_{1,6} - \tilde{A}_{1,6}\right| > h^{2(p+1)}\log(s)^\gamma\frac{r_n}{h^{2(p+1)}\log(s)^{1/2+\gamma}}\right] + o(1) \\
&= o(1),
\end{aligned}$$

because $h^{-2(p+1)}r_n\log(s)^{-1/2-\gamma} = h^{-(p+1)}\log(s)^{-1-\gamma} \to \infty$.

Next, since $\tilde{A}_{1,6} \asymp h^{2(p+1)}$, for $r_n = h^{p+1}\log(s)^{-1/2}$.

$$s^2\mathbb{P}\left[|U_{1,2}| > r_n\right] = s^2\mathbb{P}\left[\left|e_0'\left(\Gamma^{-1} - \tilde{\Gamma}^{-1}\right)\tilde{A}_{1,6}\Gamma^{-1}e_0 s e_0'\Gamma^{-1}R'W(Y-M)/n\right| > r_n\right]$$

$$
\begin{aligned}
&\leq s^2\mathbb{P}\left[4C_\Gamma^2\left|\tilde{A}_{1,6}\right|\left|s^{-1}\sum_{i=1}^n\{(Kr_p)(X_{h,i})\varepsilon_i\}\right| > s\log(s)^{-1/2}r_n\right] \\
&\qquad + s^2\mathbb{P}\left[\left|\Gamma^{-1}-\tilde{\Gamma}^{-1}\right| > s^{-1}\log(s)^{1/2}\right] + s^2 2\mathbb{P}\left[\Gamma_p^{-1} > 2C_\Gamma\right] \\
&= s^2\mathbb{P}\left[4C_\Gamma^2\left|s^{-1}\sum_{i=1}^n\{(Kr_p)(X_{h,i})\varepsilon_i\}\right| > \log(s)^{1/2}\frac{sr_n}{h^{2(p+1)}\log(s)}\right] + o(1) \\
&= o(1),
\end{aligned}
$$

because $sr_nh^{-2(p+1)}\log(s)^{-1} = sh^{-(p+1)}\log(s)^{-3/2} \to \infty$. Terms $U_{1,3}$ and $U_{1,4}$ are nearly identically treated.

Let $r_n = h^{p+1}\log(s)^{-1/2}$. Then since $\tilde{A}_{1,6} \asymp h^{2(p+1)}$,

$$
\begin{aligned}
s^2\mathbb{P}\left[|U_{1,5}| > r_n\right] &= s^2\mathbb{P}\left[\left|e_0'\tilde{\Gamma}^{-1}\tilde{A}_{1,6}\tilde{\Gamma}^{-1}e_0se_0'\tilde{\Gamma}^{-1}R'W(Y-M)/n\right| > r_n\right] \\
&\leq s^2\mathbb{P}\left[C_\Gamma^3\left|\tilde{A}_{1,6}\right|\left|s^{-1}\sum_{i=1}^n\{(Kr_p)(X_{h,i})\varepsilon_i\}\right| > r_n\right] \\
&\leq s^2\mathbb{P}\left[C_\Gamma^3\left|s^{-1}\sum_{i=1}^n\{(Kt)(X_{h,i})g(X_i)\varepsilon_i\}\right| > \log(s)^{1/2}\frac{\log(s)^{-1/2}r_n}{h^{2(p+1)}}\right] \\
&= o(1),
\end{aligned}
$$

because $h^{-2(p+1)}r_n\log(s)^{-1/2} = h^{-(p+1)}\log(s)^{-1} \to \infty$.

Thus, since $\tilde{\sigma}^{-1}$ is bounded away from zero, we find that

$$
s^2\mathbb{P}\left[\left|\frac{1}{2\tilde{\sigma}^3}e_0'\Gamma^{-1}A_{1,6}\Gamma^{-1}e_0se_0'\Gamma^{-1}R'W(Y-M)/n\right| > r_n\right] \to 0.
$$

Turning our attention to the second term, we have

$$
\begin{aligned}
&e_0'\Gamma^{-1}A_{1,6}\Gamma^{-1}e_0se_0'\Gamma^{-1}R'W(M-\check{R}\beta)/n - e_0'\tilde{\Gamma}^{-1}\tilde{A}_{1,6}\tilde{\Gamma}^{-1}e_0\eta \\
&\quad = e_0'\Gamma^{-1}\left(A_{1,6}-\tilde{A}_{1,6}\right)\Gamma^{-1}e_0se_0'\Gamma^{-1}R'W(M-\check{R}\beta)/n \\
&\qquad + e_0'\Gamma^{-1}\tilde{A}_{1,6}\Gamma^{-1}e_0se_0'\Gamma^{-1}\left(R'W(M-\check{R}\beta)/n - \mathbb{E}\left[R'W(M-\check{R}\beta)/n\right]\right) \\
&\qquad + e_0'\left(\Gamma^{-1}-\tilde{\Gamma}^{-1}\right)\tilde{A}_{1,6}\Gamma^{-1}e_0se_0'\Gamma^{-1}\mathbb{E}\left[R'W(M-\check{R}\beta)/n\right] \\
&\qquad + e_0'\tilde{\Gamma}^{-1}\tilde{A}_{1,6}\left(\Gamma^{-1}-\tilde{\Gamma}^{-1}\right)e_0se_0'\Gamma^{-1}\mathbb{E}\left[R'W(M-\check{R}\beta)/n\right] \\
&\qquad + e_0'\tilde{\Gamma}^{-1}\tilde{A}_{1,6}\tilde{\Gamma}^{-1}e_0se_0'\left(\Gamma^{-1}-\tilde{\Gamma}^{-1}\right)\mathbb{E}\left[R'W(M-\check{R}\beta)/n\right] \\
&\quad =: U_{2,1}+U_{2,2}+U_{2,3}+U_{2,4}+U_{2,5}.
\end{aligned}
$$

For $r_n = h^{p+1}\log(s)^{-1}$, we have

$$
r_*^{-1}\mathbb{P}\left[|U_{2,1}| > r_n\right] = r_*^{-1}\mathbb{P}\left[e_0'\Gamma^{-1}\left(A_{1,6}-\tilde{A}_{1,6}\right)\Gamma^{-1}e_0se_0'\Gamma^{-1}R'W(M-\check{R}\beta)/n > r_n\right]
$$

$$
\begin{aligned}
&\leq r_*^{-1}\mathbb{P}\left[8C_\Gamma^3 s\left|A_{1,6}-\tilde{A}_{1,6}\right| > sh^{2(p+1)}\log(s)^\gamma \frac{r_n}{sh^{2(p+1)}\log(s)^{2\gamma}}\right] \\
&\quad + r_*^{-1}\mathbb{P}\left[\left|\frac{1}{nh}\sum_{i=1}^n \left\{(Kr_p)(X_{h,i})\left[m(X_i)-r_p(X_i-x)'\beta_p\right]\right\}\right| > \log(s)^\gamma\right] \\
&\quad + r_*^{-1}3\mathbb{P}\left[\Gamma_p^{-1} > 2C_\Gamma\right] \\
&\leq s^2\mathbb{P}\left[8C_\Gamma^3 s\left|A_{1,6}-\tilde{A}_{1,6}\right| > sh^{2(p+1)}\log(s)^\gamma \frac{r_n}{sh^{2(p+1)}\log(s)^{2\gamma}}\right] \\
&\quad + h^{-(p+1)}\mathbb{P}\left[\left|\frac{1}{nh}\sum_{i=1}^n \left\{(Kr_p)(X_{h,i})\left[m(X_i)-r_p(X_i-x)'\beta_p\right]\right\}\right| > \log(s)^\gamma\right] \\
&\quad + s^2 3\mathbb{P}\left[\Gamma_p^{-1} > 2C_\Gamma\right] \\
&= o(1),
\end{aligned}
$$

because $sh^{2(p+1)}r_n^{-1}\log(s)^{2\gamma} = sh^{p+1}\log(s)^{1+2\gamma} \to 0$ by the conditions on η placed in the theorem.

Next, with $r_n = h^{p+1}\log(s)^{-1}$ and using $\tilde{A}_{1,6} \asymp h^{2(p+1)}$, we have

$$
\begin{aligned}
r_*^{-1}\mathbb{P}\left[|U_{2,2}| > r_n\right] &= r_*^{-1}\mathbb{P}\left[\left|e_0'\Gamma^{-1}\tilde{A}_{1,6}\Gamma^{-1}e_0 s e_0'\Gamma^{-1}\left(R'W(M-\check{R}\beta)/n - \mathbb{E}\left[R'W(M-\check{R}\beta)/n\right]\right)\right| > r_n\right] \\
&\leq r_*^{-1}\mathbb{P}\Bigg[8C_\Gamma^3\left|\tilde{A}_{1,6}\right|\Bigg|s^{-1}\sum_{i=1}^n\Big\{(Kr_p)(X_{h,i})\left[m(X_i)-r_p(X_i-x)'\beta_p\right] \\
&\qquad\qquad - \mathbb{E}\left[(Kr_p)(X_{h,i})\left[m(X_i)-r_p(X_i-x)'\beta_p\right]\right]\Big\}\Bigg| > r_n\Bigg] \\
&\quad + r_*^{-1}3\mathbb{P}\left[\Gamma_p^{-1} > 2C_\Gamma\right] \\
&\leq s^2\mathbb{P}\Bigg[8C_\Gamma^3\Bigg|s^{-2}\sum_{i=1}^n\Big\{(Kr_p)(X_{h,i})\left[m(X_i)-r_p(X_i-x)'\beta_p\right] \\
&\qquad - \mathbb{E}\left[(Kr_p)(X_{h,i})\left[m(X_i)-r_p(X_i-x)'\beta_p\right]\right]\Big\}\Bigg| > h^{p+1}\log(s)^\gamma\frac{r_n}{h^{3(p+1)}\log(s)^\gamma}\Bigg] \\
&\quad + s^2 3\mathbb{P}\left[\Gamma_p^{-1} > 2C_\Gamma\right] \\
&= o(1),
\end{aligned}
$$

because $r_n h^{-3(p+1)}\log(s)^{-\gamma} = h^{-2(p+1)}\log(s)^{-1-\gamma} \to \infty$.

Third, as $\tilde{A}_{1,6} \asymp h^{2(p+1)}$ and $\mathbb{E}\left[R'W(M-\check{R}\beta)/n\right] \asymp h^{p+1}$, if we choose $r_n = h^{p+1}\log(s)^{-1}$,

$$
\begin{aligned}
r_*^{-1}\mathbb{P}\left[|U_{2,3}| > r_n\right] &\leq r_*^{-1}\mathbb{P}\left[4C_\Gamma^2 s\left|\Gamma^{-1}-\tilde{\Gamma}^{-1}\right| > s^{-1}\log(s)^{1/2}\frac{sr_n}{h^{3(p+1)}\log(s)^{1/2}}\right] \\
&\quad + r_*^{-1}2\mathbb{P}\left[\Gamma_p^{-1} > 2C_\Gamma\right] \\
&\leq s^2\mathbb{P}\left[4C_\Gamma^2\left|\Gamma^{-1}-\tilde{\Gamma}^{-1}\right| > s^{-1}\log(s)^{1/2}\frac{r_n}{h^{3(p+1)}\log(s)^{1/2}}\right] \\
&\quad + s^2 2\mathbb{P}\left[\Gamma_p^{-1} > 2C_\Gamma\right] \\
&= o(1),
\end{aligned}
$$

because $r_n h^{-3(p+1)} \log(s)^{-1/2} = h^{-2(p+1)} \log(s)^{-1-1/2} \to \infty$. The terms $U_{2,3}$ and $U_{2,5}$ are handled identically.

Thus, since $\tilde{\sigma}^{-1}$ is bounded away from zero, we find that

$$s^2 \mathbb{P}\left[\left|\frac{1}{2\tilde{\sigma}^3} e_0' \Gamma^{-1} A_{1,6} \Gamma^{-1} e_0 s e_0' \Gamma^{-1} R' W (M - \check{R}\beta)/n - e_0' \tilde{\Gamma}^{-1} \tilde{A}_{1,6} \tilde{\Gamma}^{-1} e_0 \eta\right| > r_n\right] \to 0.$$

The same type of arguments, though notationally more challenging, will show that the remainder of U obeys the same bounds. Note that the rest of the terms are even higher order, involving either $A_{1,7}$ and $A_{1,8}$, or the square or cube of the other errors. It is for this reason that only the "leading" three terms need be centered, that is, why only

$$-\left\{-\frac{1}{2\tilde{\sigma}^3} e_0' \tilde{\Gamma}^{-1} \left(\tilde{A}_{1,6} + \tilde{A}_{1,7} + \tilde{A}_{1,8}\right) \tilde{\Gamma}^{-1} e_0\right\} \eta$$

appears in $\tilde{z}$. □

S.II.6.4 Computing the Terms of the Expansion

Identifying the terms of the expansion is a matter of straightforward, if tedious, calculation. The first four cumulants of the Studentized statistics must be calculated (due to James and Mayne (1962)), which are functions of the first four moments. In what follows, we give a short summary. Note well that we always discard higher-order terms for brevity, and to save notation we will write $\overset{o}{=}$ to stand in for "equal up to $o((nh)^{-1} + (nh)^{-1/2}\eta + \eta^2)$", and including $o(\rho^{1+2(p+1)})$ for $T_{\texttt{bc}}$.

The computations will be aided by putting all three estimators into a common structure. In close parallel to the density case, let us define $\hat{m}_1 := \hat{m}$ and $\hat{m}_2 = \hat{m} - \hat{m}_m$, $\sigma_1^2 := \sigma_{\texttt{us}}^2$, and $\sigma_2^2 := \sigma_{\texttt{rbc}}^2$, so that subscripts 1 and 2 generically stand in for undersmoothing and bias correction, respectively. With this in mind, we write

$$T_{\texttt{us}} = T_{1,1}, \qquad T_{\texttt{bc}} = T_{2,1}, \qquad \text{and} \qquad T_{\texttt{rbc}} = T_{2,2},$$

again paralleling the density case, so that the first subscript refers to the numerator and the second to the denominator. In the same vein, with some abuse of notation, we will also use[3] $r_1(u) = r_p(u)$, $r_2(u) = r_q(u)$, $K_1(u) = K(u)$, $K_2(u) = L(u)$, $h_1 = h$, and $h_2 = b$, as well as

$$\begin{aligned}
\ell_1^0(X_i) &\equiv \ell_{\texttt{us}}^0(X_i), \\
\ell_1^1(X_i, X_j) &\equiv \ell_{\texttt{us}}^1(X_i, X_j), \\
\ell_2^0(X_i) &\equiv \ell_{\texttt{bc}}^0(X_i), \\
\ell_2^1(X_i, X_j) &\equiv \ell_{\texttt{bc}}^1(X_i, X_j).
\end{aligned}$$

For the purpose of computing the expansion terms (i.e. moments of the two sides agree up to

[3]Throughout Section S.II, we use only generic polynomial orders p and q, and so this notation will not conflict with the local linear or local quadratic fits, which would also be denoted $r_1(u)$ and $r_2(u)$, respectively.

the requisite order), recalling the Taylor series expansion above, we will use

$$T_{v,w} \approx \left\{1 - \frac{1}{2\tilde{\sigma}_w^2}\left(W_{w,1} + V_{w,1} + V_{w,2}\right) + \frac{3}{8\tilde{\sigma}_w^4}\left(W_{w,1} + V_{w,1} + V_{w,2}\right)^2\right\}$$
$$\tilde{\sigma}_w^{-1}\left\{E_{v,1} + E_{v,2} + E_{v,3} + B_{v,1}\right\},$$

where we define, for $v \in \{1,2\}$,

$$E_{v,1} = s\frac{1}{nh}\sum_{i=1}^{n}\ell_v^0(X_i)\varepsilon_i$$
$$E_{v,2} = s\frac{1}{(nh)^2}\sum_{i=1}^{n}\sum_{j=1}^{n}\ell_v^1(X_i,X_j)\varepsilon_i,$$
$$E_{v,3} =: s\frac{1}{(nh)^3}\sum_{i=1}^{n}\sum_{j=1}^{n}\sum_{k=1}^{n}\ell_v^2(X_i,X_j,X_k)\varepsilon_i,$$

where the final line defines $\ell_{\mathtt{us}}^2(X_i,X_j,X_k)$ in the obvious way following $\ell_{\mathtt{us}}^1$. To concretize the notation, for undersmoothing we are defining

$$E_{1,1} = se_0'\tilde{\Gamma}_p^{-1}R_p'W_p(Y-M)/n,$$
$$E_{1,2} = se_0'\tilde{\Gamma}_p^{-1}(\tilde{\Gamma}_p - \Gamma_p)\tilde{\Gamma}_p^{-1}R_p'W_p(Y-M)/n,$$
$$E_{1,3} = se_0'\tilde{\Gamma}_p^{-1}(\tilde{\Gamma}_p - \Gamma_p)\tilde{\Gamma}_p^{-1}(\tilde{\Gamma}_p - \Gamma_p)\tilde{\Gamma}_p^{-1}R_p'W_p(Y-M)/n.$$

In a similar way,

$$W_{v,1} = \frac{1}{nh}\sum_{i=1}^{n}\left\{\ell_v^0(X_i)^2\left(\varepsilon_i^2 - v(X_i)\right)\right\} - 2\frac{1}{n^2h^2}\sum_{i=1}^{n}\sum_{j=1}^{n}\left\{\ell_v^0(X_i)^2 r_v(X_{h_v,i})'\tilde{\Gamma}_v^{-1}(K_v r_v)(X_{h_v,i})\varepsilon_i\varepsilon_j\right\}$$
$$+ \frac{1}{n^3h^3}\sum_{i=1}^{n}\sum_{j=1}^{n}\sum_{k=1}^{n}\left\{\ell_v^0(X_i)^2 r_v(X_{h_v,i})'\tilde{\Gamma}_v^{-1}(K_v r_v)(X_{h_v,i})\varepsilon_j\varepsilon_k\right\},$$
$$V_{v,1} = \frac{1}{nh}\sum_{i=1}^{n}\left\{\ell_v^0(X_i)^2 v(X_i)^2 - \mathbb{E}[\ell_v^0(X_i)^2 v(X_i)^2]\right\} + 2\frac{1}{n^2h^2}\sum_{i=1}^{n}\sum_{j=1}^{n}\ell_v^2(X_i,X_j)\ell_v^0(X_i)v(X_i),$$
$$V_{v,2} = \frac{1}{n^3h^3}\sum_{i=1}^{n}\sum_{j=1}^{n}\sum_{k=1}^{n}\ell_v^1(X_i,X_j)\ell_v^1(X_i,X_k)v(X_i) + 2\frac{1}{n^3h^3}\sum_{i=1}^{n}\sum_{j=1}^{n}\sum_{k=1}^{n}\ell_v^2(X_i,X_j,X_k)\ell_v^0(X_i)v(X_i),$$

and specifically for undersmoothing and bias correction, let

$$B_{1,1} = s\frac{1}{nh}\sum_{i=1}^{n}\ell_1^0(X_i)[m(X_i) - r_p(X_i - x)'\beta_p]$$

and

$$B_{2,1} = s\frac{1}{nh}\sum_{i=1}^{n}\Big\{h^{-1}\ell^0_{\mathtt{us}}(X_i)[m(X_i) - r_{p+1}(X_i - x)'\beta_{p+1}] \\ - h^{-1}\left(\ell^0_{\mathtt{bc}}(X_i) - \ell^0_{\mathtt{us}}(X_i)\right)[m(X_i) - r_q(X_i - x)'\beta_q]\Big\}.$$

Note that $\eta_{\mathtt{us}} = \mathbb{E}[B_{1,1}]$ and $\eta_{\mathtt{bc}} = \mathbb{E}[B_{2,1}]$.

Straightforward moment calculations yield

$$\mathbb{E}[T_{v,w}] \overset{o}{=} \tilde{\sigma}_w^{-1}\mathbb{E}\left[B_{v,1}\right] - \frac{1}{2\tilde{\sigma}_w^2}\mathbb{E}\left[W_{w,1}E_{v,1}\right],$$

$$\mathbb{E}[T^2_{v,w}] \overset{o}{=} \frac{1}{\tilde{\sigma}_w^2}\mathbb{E}\left[E^2_{v,1} + E^2_{v,2} + 2E_{v,1}E_{v,2} + 2E_{v,1}E_{v,3}\right] \\ - \frac{1}{\tilde{\sigma}_w^4}\mathbb{E}\left[W_{w,1}E^2_{v,1} + V_{w,1}E^2_{v,1} + V_{w,2}E^2_{v,1} + 2V_{w,1}E_{v,1}E_{v,2}\right] \\ + \frac{1}{\tilde{\sigma}_w^6}\mathbb{E}\left[W^2_{w,1}E^2_{v,1} + V^2_{w,1}E^2_{v,1}\right] + \frac{1}{\tilde{\sigma}_w^2}\mathbb{E}\left[B^2_{v,1}\right] - \frac{1}{\tilde{\sigma}_w^4}\mathbb{E}\left[W_{w,1}E_{v,1}B_{v,1}\right],$$

$$\mathbb{E}[T^3_{v,w}] \overset{o}{=} \frac{1}{\tilde{\sigma}_w^3}\mathbb{E}\left[E^3_{v,1}\right] - \frac{3}{2\tilde{\sigma}_w^5}\mathbb{E}\left[W_{w,1}E^3_{v,1}\right] + \frac{3}{\tilde{\sigma}_w^3}\mathbb{E}\left[E^2_{v,1}B_{v,1}\right],$$

and

$$\mathbb{E}[T^4_{v,w}] \overset{o}{=} \frac{1}{\tilde{\sigma}_w^4}\mathbb{E}\left[E^4_{v,1} + 4E^3_{v,1}E_{v,2} + 4E^3_{v,1}E_{v,3} + 6E^2_{v,1}E^2_{v,3}\right] \\ - \frac{2}{\tilde{\sigma}_w^6}\mathbb{E}\left[W_{w,1}E^4_{v,1} + V_{w,1}E^4_{v,1} + 4V_{w,1}E^3_{v,1}E_{v,2} + V_{w,2}E_{v,1}\right] \\ + \frac{3}{\tilde{\sigma}_w^8}\mathbb{E}\left[W^2_{w,1}E^4_{v,1} + V^2_{w,1}E^4_{v,1}\right] \\ + \frac{4}{\tilde{\sigma}_w^4}\mathbb{E}\left[E^3_{v,1}B_{v,1}\right] - \frac{8}{\tilde{\sigma}_w^6}\mathbb{E}\left[W_{w,1}E^3_{v,1}B_{v,1}\right] + \frac{6}{\tilde{\sigma}_w^4}\mathbb{E}\left[E^2_{v,1}B^2_{v,1}\right].$$

Computing each term in turn, we have

$$\mathbb{E}\left[B_{v,1}\right] = \eta_v,$$
$$\mathbb{E}\left[W_{w,1}E_{v,1}\right] \overset{o}{=} s^{-1}\mathbb{E}\left[h^{-1}\ell^0_w(X_i)^2\ell^0_v(X_i)\varepsilon_i^3\right],$$
$$\mathbb{E}\left[E^2_{v,1}\right] \overset{o}{=} \tilde{\sigma}_v^2,$$
$$\mathbb{E}\left[E_{v,1}E_{v,2}\right] \overset{o}{=} s^{-2}\mathbb{E}\left[h^{-1}\ell^1_v(X_i, X_i)\ell^0_v(X_i)\varepsilon_i^2\right],$$
$$\mathbb{E}\left[E^2_{v,2}\right] \overset{o}{=} s^{-1}\mathbb{E}\left[h^{-2}\ell^1_v(X_i, X_j)^2\varepsilon_i^2\right],$$
$$\mathbb{E}\left[E_{v,2}E_{v,3}\right] \overset{o}{=} s^{-2}\mathbb{E}\left[h^{-2}\ell^2_v(X_i, X_j, X_j)\ell^0_v(X_i)\varepsilon_i^2\right],$$
$$\mathbb{E}\left[W_{w,1}E^2_{v,1}\right] \overset{o}{=} s^{-2}\Big\{\mathbb{E}\left[h^{-1}\ell^0_w(X_i)^2\ell^0_v(X_i)^2\left(\varepsilon_i^4 - v(X_i)^2\right)\right]$$

$$
\begin{aligned}
&- 2\tilde{\sigma}_v^2 \mathbb{E}\left[h^{-1}\ell_w^0(X_i)^2 r_w(X_{h_w,i})'\tilde{\Gamma}_w^{-1}(K_w r_w)(X_{h_w,i})\varepsilon_i^2\right] \\
&- 4\mathbb{E}\left[h^{-1}\ell_w^0(X_i)^2\ell_v^0(X_i)^2 r_w(X_{h_w,i})'\tilde{\Gamma}_w^{-1}\varepsilon_i^2\right]\mathbb{E}\left[h^{-1}(K_w r_w)(X_{h_w,i})\ell_v^0(X_i)\varepsilon_i^2\right] \\
&+ \tilde{\sigma}_v^2 \mathbb{E}\left[h^{-2}\ell_w^0(X_i)^2\left(r_w(X_{h_w,i})'\tilde{\Gamma}_w^{-1}(K_w r_w)(X_{h_w,j})\right)^2\varepsilon_j^2\right] \\
&+ \mathbb{E}\left[h^{-1}\ell_{\mathtt{us}}^0(X_j)^2\left(\mathbb{E}\left[h^{-1}r_p(X_{h,j})'\tilde{\Gamma}_p^{-1}(Kr_p)(X_{h,i})\ell_{\mathtt{us}}^0(X_i)\varepsilon_i^2|X_j\right]\right)^2\right]\Bigg\},
\end{aligned}
$$

$$
\begin{aligned}
\mathbb{E}\left[V_{w,1}E_{v,1}^2\right] &\overset{o}{=} s^{-2}\Big\{\mathbb{E}\left[h^{-1}\left(\ell_w^0(X_i)^2 v(X_i) - \mathbb{E}[\ell_w^0(X_i)^2 v(X_i)]\right)\ell_v^0(X_i)^2\varepsilon_i^2\right] \\
&\quad + 2\tilde{\sigma}_v^2\mathbb{E}\left[h^{-1}\ell_w^1(X_i,X_i)\ell_w^0(X_i)v(X_i)\right]\Big\}, \\
\mathbb{E}\left[V_{w,1}E_{v,1}E_{v,2}\right] &\overset{o}{=} s^{-2}\Big\{\mathbb{E}\left[h^{-2}\left(\ell_w^0(X_j)^2 v(X_j) - \mathbb{E}[\ell_w^0(X_j)^2 v(X_j)]\right)\ell_v^1(X_i,X_j)\ell_v^0(X_i)\varepsilon_i^2\right] \\
&\quad + 2\mathbb{E}\left[h^{-3}\ell_w^1(X_i,X_j)\ell_v^1(X_k,X_j)\ell_w^0(X_i)\ell_v^0(X_k)v(X_i)\varepsilon_k^2\right]\Big\}, \\
\mathbb{E}\left[V_{w,2}E_{v,1}^2\right] &\overset{o}{=} s^{-2}\Big\{\tilde{\sigma}_v^2\mathbb{E}\left[h^{-2}\left(\ell_w^1(X_i,X_j)^2 + 2\ell_w^2(X_i,X_j,X_j)\right)v(X_i)\right]\Big\}, \\
\mathbb{E}\left[W_{w,1}^2E_{v,1}^2\right] &\overset{o}{=} s^{-2}\Big\{\tilde{\sigma}_v^2\mathbb{E}\left[h^{-1}\ell_w^0(X_i)^4\left(\varepsilon_i^4 - v(X_i)^2\right)\right] + 2\mathbb{E}\left[h^{-1}\ell_v^0(X_i)\ell_w^0(X_i)^2\varepsilon_i^3\right]^2\Big\}, \\
\mathbb{E}\left[V_{w,1}^2E_{v,1}^2\right] &\overset{o}{=} s^{-2}\tilde{\sigma}_v^2\Big\{\mathbb{E}\left[h^{-1}\left(\ell_w^0(X_i)^2 v(X_i) - \mathbb{E}[\ell_w^0(X_i)^2 v(X_i)]\right)^2\right] \\
&\quad + 4\mathbb{E}\left[h^{-2}\left(\ell_w^0(X_i)^2 v(X_i) - \mathbb{E}[\ell_w^0(X_i)^2 v(X_i)]\right)\ell_w^1(X_j,X_i)\ell_w^0(X_j)v(X_j)\right] \\
&\quad + 4\mathbb{E}\left[h^{-3}\ell_w^1(X_i,X_j)\ell_w^0(X_i)v(X_i)\ell_w^1(X_k,X_j)\ell_w^0(X_k)v(X_k)\right]\Big\}, \\
\mathbb{E}\left[W_{w,1}E_{v,1}B_{v,1}\right] &\overset{o}{=} \mathbb{E}\left[W_{w,1}E_{v,1}\right]\mathbb{E}\left[B_{v,1}\right], \\
\mathbb{E}\left[E_{v,1}^3\right] &\overset{o}{=} s^{-1}\mathbb{E}\left[h^{-1}\ell_v^0(X_i)^3\varepsilon_i^3\right], \\
\mathbb{E}\left[W_{w,1}E_{v,1}^3\right] &\overset{o}{=} \mathbb{E}\left[E_{v,1}^2\right]\mathbb{E}\left[W_{w,1}E_{v,1}\right], \\
\mathbb{E}\left[E_{v,1}^4\right] &\overset{o}{=} 3\tilde{\sigma}_v^4 + s^{-2}\mathbb{E}\left[h^{-1}\ell_v^0(X_i)^4\varepsilon_i^3\right], \\
\mathbb{E}\left[E_{v,1}^3E_{v,2}\right] &\overset{o}{=} s^{-2}6\tilde{\sigma}_v^2\mathbb{E}\left[h^{-1}\ell_v^1(X_i,X_i)\ell_v^0(X_i)\varepsilon_i^2\right], \\
\mathbb{E}\left[E_{v,1}^3E_{v,3}\right] &\overset{o}{=} s^{-2}3\tilde{\sigma}_v^2\mathbb{E}\left[h^{-2}\ell_v^2(X_i,X_j,X_j)\ell_v^0(X_i)\varepsilon_i^2\right], \\
\mathbb{E}\left[E_{v,1}^2E_{v,2}^2\right] &\overset{o}{=} s^{-2}\Big\{\tilde{\sigma}_v^2\mathbb{E}\left[h^{-2}\ell_v^1(X_i,X_j)^2\varepsilon_i^2\right] + 2\mathbb{E}\left[h^{-3}\ell_v^1(X_i,X_j)\ell_v^1(X_k,X_j)\ell_v^0(X_i)\ell_v^0(X_k)\varepsilon_i^2\varepsilon_k^2\right]\Big\}, \\
\mathbb{E}\left[W_{w,1}E_{v,1}^4\right] &\overset{o}{=} s^{-2}\Big\{\mathbb{E}\left[h^{-1}\ell_w^0(X_i)^2\ell_v^0(X_i)\varepsilon_i^3\right]\mathbb{E}\left[h^{-1}\ell_v^0(X_i)^3\varepsilon_i^3\right] + 6\mathbb{E}\left[E_{v,1}^2\right]\mathbb{E}\left[W_{w,1}E_{v,1}^2\right]\Big\}, \\
\mathbb{E}\left[V_{w,1}E_{v,1}^4\right] &\overset{o}{=} s^{-2}\tilde{\sigma}_v^2 6\Big\{\mathbb{E}\left[h^{-1}\left(\ell_w^0(X_i)^2 v(X_i) - \mathbb{E}[\ell_w^0(X_i)^2 v(X_i)]\right)\ell_v^0(X_i)^2\varepsilon_i^2\right] \\
&\quad + 2\mathbb{E}\left[h^{-2}\ell_w^1(X_i,X_j)\ell_w^0(X_i)\ell_v^0(X_j)^2\varepsilon_j^2 v(X_i)\right] + \mathbb{E}\left[h^{-1}\ell_w^1(X_i,X_i)\ell_w^0(X_i)v(X_i)\right]\Big\}, \\
\mathbb{E}\left[V_{w,1}E_{v,1}^3E_{v,2}\right] &\overset{o}{=} 3\mathbb{E}\left[E_{v,1}^2\right]\mathbb{E}\left[V_{w,1}E_{v,1}E_{v,2}\right], \\
\mathbb{E}\left[V_{w,2}E_{v,1}^4\right] &\overset{o}{=} 3\mathbb{E}\left[E_{v,1}^2\right]\mathbb{E}\left[V_{w,2}E_{v,1}^2\right], \\
\mathbb{E}\left[W_{w,1}^2E_{v,1}^4\right] &\overset{o}{=} 3\mathbb{E}\left[E_{v,1}^2\right]\mathbb{E}\left[W_{w,1}^2E_{v,1}^2\right], \\
\mathbb{E}\left[V_{w,1}^2E_{v,1}^4\right] &\overset{o}{=} 3\mathbb{E}\left[E_{v,1}^2\right]\mathbb{E}\left[V_{w,1}^2E_{v,1}^2\right].
\end{aligned}
$$

The expansion now follows, formally, from the following steps. First, combining the above moments into cumulants. Second, these cumulants may be simplified using that

$$\frac{\sigma_v^2}{\sigma_w^2} = 1 + \mathbb{1}(w \neq v) \left(\rho^{1+(p+1)} \Omega_{1,\mathtt{bc}} + \rho^{1+2(p+1)} \Omega_{2,\mathtt{bc}} \right)$$

and that in all cases present products such as $\ell_w^0(X_i)^{k_1} \ell_v^0(X_i)^{k_2}$ and $\ell_w^1(X_i, X_j)^{k_1} \ell_v^1(X_i, X_j)^{k_2}$ may be replaced with $\ell_v^0(X_i)^{k_1+k_2}$ and $\ell_v^1(X_i, X_j)^{k_1+k_2}$, respectively, provided the arguments match. This is immediate for $v = w$, and for $v \neq w$, follows because $\rho \to 0$ is assumed. This is the analogous step to Eqn. (S.I.8) in the density case. For any term of a cumulant with a rate of $(nh)^{-1}$, $(nh)^{-1/2}\eta_v$, η_v^2, or $\rho^{1+2(p+1)}$ (i.e., the extent of the expansion), these simplifications may be inserted as the remainder will be negligible. Third, with the cumulants in hand, the terms of the expansion are determined as described by e.g., (Hall, 1992a, Chapter 2).

S.II.7 Complete Simulation Results

In this section we present the results of a simulation study addressing the finite-sample performance of the methods described in the main paper. As with the density estimator, we report empirical coverage probabilities and average interval length of nominal 95% confidence interval for different estimators of a regression functions $m(x)$ evaluated at values $x = \{-2/3, -1/3, 0, 1/3, 2/3\}$. For each replication, the data is generated as i.i.d. draws, $i = 1, 2, ..., n$, $n = 500$ as follows:

$$Y = m(x) + \varepsilon, \qquad x \sim \mathcal{U}[-1, 1], \qquad \varepsilon \sim \mathcal{N}(0, 1)$$

Model 1: $m(x) = \sin(4x) + 2\exp\{-64x^2\}$

Model 2: $m(x) = 2x + 2\exp\{-64x^2\}$

Model 3: $m(x) = 0.3\exp\{-4(2x+1)^2\} + 0.7\exp\{-16(2x-1)^2\}$

Model 4: $m(x) = x + 5\phi(10x)$

Model 5: $m(x) = \dfrac{\sin(3\pi x/2)}{1 + 18x^2[sgn(x) + 1]}$

Model 6: $m(x) = \dfrac{\sin(\pi x/2)}{1 + 2x^2[sgn(x) + 1]}$

Models 1 to 3 were used by Fan and Gijbels (1996) and Cattaneo and Farrell (2013), while Models 4 to 6 are from Hall and Horowitz (2013), with some originally studied by Berry et al. (2002). The regression functions are plotted in Figure S.II.1 together with the evaluation points used.

We compute confidence intervals for $m(x)$ using five alternative approaches:

US: local-linear estimator using a conventional approach based on undersmoothing ($I_{\mathtt{us}}$).

Locfit: local lineal estimator computed using default options in the `R` package `locfit` (see Loader (2013) for implementation details).

BC: traditional bias corrected estimator using a local-linear estimator with local-quadratic bias-correction, and $\rho = 1$ ($I_{\texttt{bc}}$).

HH: local linear estimator using the bootstrapped confidence bands introduced in Hall and Horowitz (2013) (see Remark S.II.4 below for additional implementation details).

RBC: our proposed local-linear estimator with local-quadratic bias-correction and $\rho = 1$ using robust standard errors ($I_{\texttt{rbc}}$).

In all cases the Epanechnikov kernel is used. The bandwidth h is chosen in three different ways:

(i) population MSE-optimal choice $h_{\texttt{mse}}$;

(ii) estimated ROT optimal coverage error rate $\hat{h}_{\texttt{rot}}$.

(iii) estimated DPI optimal coverage error rate $\hat{h}_{\texttt{dpi}}$.

For the construction of the variance estimators $\hat{\sigma}^2_{\texttt{us}}$ and $\hat{\sigma}^2_{\texttt{rbc}}$ we consider HC3 plug-in residuals when forming the Σ matrix. In Table S.II.9 we report empirical coverage and average interval length of RBC 95% Confidence Intervals (only for Model 5) using $\hat{h}_{\texttt{mse}}$ for different variance estimators. The results reflect the robustness of the findings to this choice.

The results are presented in detail in the tables and figures below to give a complete picture of the performance of robust bias correction. First, Tables S.II.1-S.II.6 show, for each regression model, respectively, the performance of the five methods above, in terms of empirical coverage and interval length, for all evaluation points and bandwidth choices (recall that $I_{\texttt{us}}$ and $I_{\texttt{bc}}$ have the same length). Panel A of each shows the coverage and length, while Panel B gives summary statistics for the two fully data-driven bandwidths. Note that in some cases, the population MSE-optimal bandwidth is not defined or is not computable numerically; usually because the bias is too small or other values are too extreme.

The broad conclusion from these tables is that robust bias correction provides excellent coverage and that the data-driven bandwidths perform well and are numerically stable. In almost all cases robust bias correction provides correct coverage, whereas the other methods often, but not always, fail to do so. In cases where there is little to no bias all the methods give good coverage. This can be seen in results for Models 2 and 4, at $|x| = 2/3$, far enough away from the "hump" in the center of each, where the true regression function is (nearly) linear. But despite the encouraging results away from the center, only robust bias correction yields good coverage closer to the center ($|x| = 1/3$), when there is more bias. Going further, considering $x = 0$, the center of the sharp peak in these models, we see that even robust bias correction fails to provide accurate coverage for $\hat{h}_{\texttt{rot}}$, although $\hat{h}_{\texttt{dpi}}$ performs slightly better. At this point, for these models, the bias is too extreme even for robust bias correction to overcome. The results for the other models yield similar lessons.

It is somewhat more difficult to compare interval length using these tables. The comparison is invited for a fixed bandwidth, in which case, by construction, undersmoothing will have a shorter length. However, this ignores the fact that robust bias correction can accommodate a larger range of bandwidths, and in particular will optimally use a larger bandwidth. For example, robust bias correction has excellent coverage in many cases for $\hat{h}_{\texttt{rot}}$, which is in this case a data-driven MSE-optimal choice (i.e. they coincide). This bandwidth is generally larger than $\hat{h}_{\texttt{dpi}}$, and hence undersmoothing generally covers better with the latter. However, if you compare the length of $I_{\texttt{us}}(\hat{h}_{\texttt{rot}})$ to the length of $I_{\texttt{us}}(\hat{h}_{\texttt{dpi}})$, we see that robust bias correction compares favorably in terms of length.

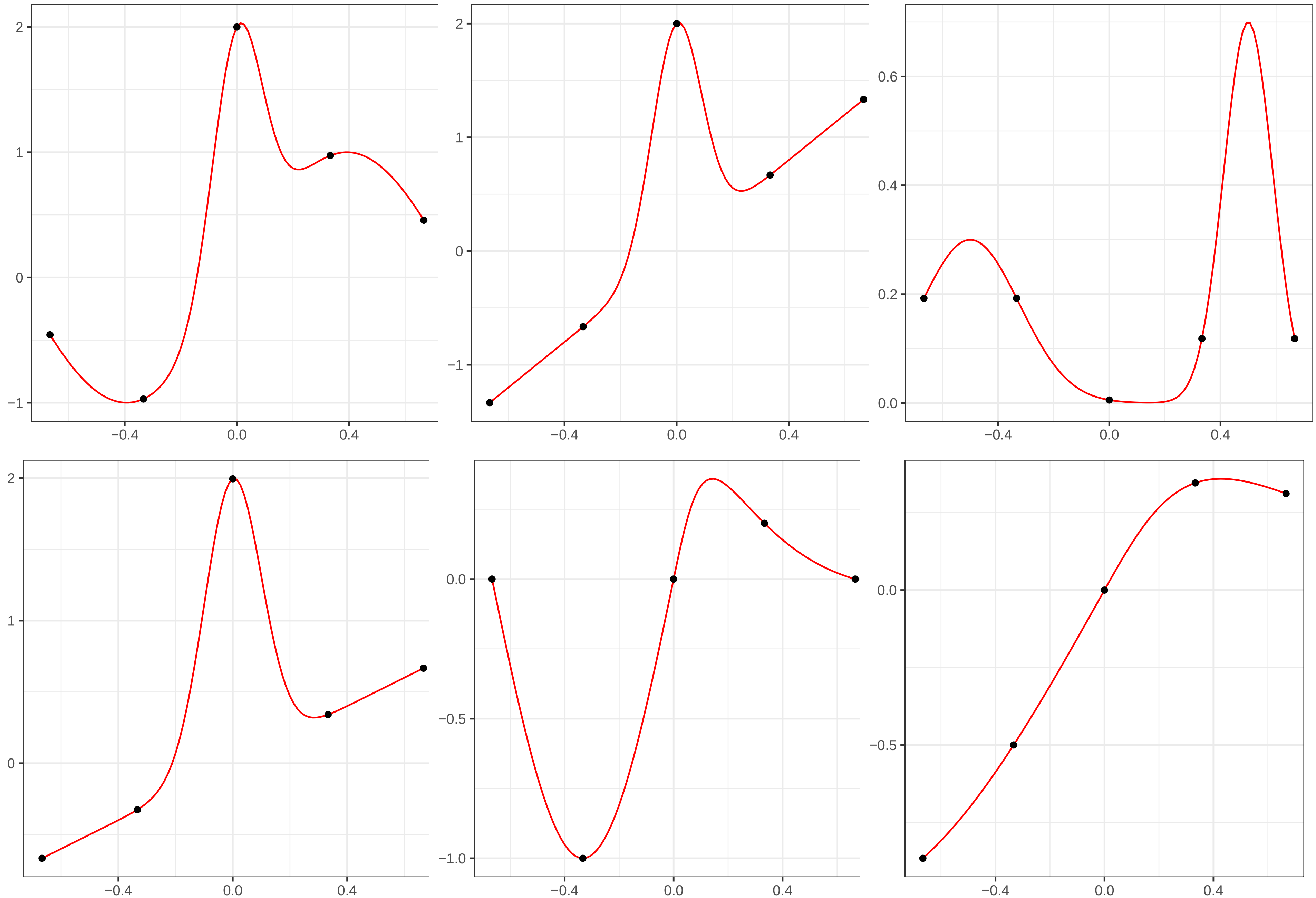
Both to better make this point and to illustrate the robustness of $I_{\texttt{rbc}}$ to tuning parameter selection, Figures S.II.2–S.II.13 show empirical coverage and length for all six models, and all evaluation points, across a range of bandwidths. The dotted vertical line shows the population MSE-optimal bandwidth (whenever available) for reference. The coverage figures highlight the delicate balance required for undersmoothing to provide correct coverage, and the generally poor performance of traditional bias correction, but show that for a wide range of bandwidths robust bias correction provides correct coverage. Further, interval length is not unduly inflated for bandwidths that provide correct coverage. Again, by construction, undersmoothing will yield shorter intervals for a fixed bandwidth, and this is clear from Figures S.II.8–S.II.13, but it is also clear that robust bias correction can use much larger bandwidths while still maintaining correct coverage.

To further illustrate this idea, in Tables S.II.7–S.II.8 we compare average interval length of US and RBC 95% confidence intervals but at different bandwidths. First, in Table S.II.7 we compute average interval length at the largest bandwidth that provides close to correct coverage for each method separately. Note that in all cases these bandwidths are not feasible: these are ex-post findings. Next, in Table S.II.8 we evaluate the performance of US and RBC confidence intervals at certain alternative bandwidths likely to be chosen in practice. First, we evaluate the performance of US confidence intervals at $h = \lambda \hat{h}_{\texttt{mse}}$ for $\lambda = \{0.5; 0.7\}$. We then compare the performance with RBC confidence intervals computed using the optimal, fully data-driven choices $\hat{h}_{\texttt{rot}}$ and $\hat{h}_{\texttt{dpi}}$. Both tables reflect that, once we control for coverage, intervals lengths do not differ systematically between both approaches.

Figures S.II.14-S.II.19 make this same point in a different way. For a range of bandwidths, as in the previous figures, we show the "average position" of $I_{\texttt{us}}$ and $I_{\texttt{rbc}}$, where the center of the bar is placed at the average bias and the length of each bar is the average interval length across the simulations. The bars are then color-coded by coverage (green bars having good coverage, fading to red showing undercoverage). These make visually clear that although undersmoothing provides shorter intervals in general, that this comes at the expense of coverage, while robust bias correction provides good coverage for a range of bandwidths, many of which are "large" enough to yield narrow intervals.

All our methods are implemented in `R` and `STATA` via the `nprobust` package, available from `http://sites.google.com/site/nppackages/nprobust` (see also `http://cran.r-project.org/package=nprobust`). See Calonico et al. (2017) for a complete description.

Remark S.II.4 (Implementation of Hall and Horowitz (2013))**.** The column HH computes the bootstrapped confidence bands introduced in Hall and Horowitz (2013), following as close as possible their implementation choices. First, we estimate $m(x)$ using a local linear estimator using the Epanechnikov kernel for our previously discussed bandwidth choices. Standard errors are calculated using their proposed variance estimator $\hat{\sigma}^2_{HH} = \kappa\hat{\sigma}^2/\hat{f}_X(x)$ where $\kappa = \int K^2$ and $\hat{f}_X(x)$ is a standard kernel density estimator using a data-driven bandwidth choice h_1. Then, we use the same estimator for the error variance $\hat{\sigma}^2 = \sum_{i=1}^n \hat{\varepsilon}_i^2/n$ and $\hat{\varepsilon}_i = \tilde{\varepsilon}_i - \bar{\varepsilon}$, $\tilde{\varepsilon}_i = Y_i - \hat{m}(X_i)$, $\bar{\varepsilon} = n^{-1}\sum_{i=1}^n \tilde{\varepsilon}_i$. Next, we take generate $B = 500$ bootstrap samples $\mathscr{Z}^* = \{(X_i, Y_i^*)\}, 1 \leq i \leq n$, where $Y_i^* = \hat{m}(X_i) + \varepsilon_i^*$, with ε_i^* obtained by sampling with replacement from the $\{\hat{\varepsilon}_i\}, 1 \leq i \leq n$. With these bootstrap samples we can construct the final confidence bands using the adjusted critical values that approximates the estimated coverage error with the selected one. Following their recommendation, the final critical values are taken to be the ξ-level quantile (for $\xi = 0.1$) obtained by repeating this exercise over a grid of evaluation points, which we choose to be the sequence $\{x_1, ..., x_N\} = \{-0.9, -0.8, ..., 0, ..., 0.8, 0.9\}$. ■



(d) Model 4 (e) Model 5 (f) Model 6

Table S.II.1: Simulations Results for Model 1

Panel A: Empirical Coverage and Average Interval Length of 95% Confidence Intervals

	Bandwidth	Empirical Coverage					Interval Length			
		US	Locfit	BC	HH	RBC	US	Locfit	HH	RBC
$x = -2/3$										
$h_{\mathtt{mse}}$	0.478	56.1	76.5	83.6	30.5	94.8	0.302	0.330	0.198	0.422
$\hat{h}_{\mathtt{rot}}$	0.201	93.6	94.3	83.8	94.3	95.2	0.440	0.479	0.468	0.631
$\hat{h}_{\mathtt{dpi}}$	0.177	95.0	95.0	83.6	97.1	94.7	0.467	0.507	0.515	0.669
$x = -1/3$										
$h_{\mathtt{mse}}$	0.331	4.9	31.2	82.1	1.9	93.1	0.357	0.377	0.277	0.488
$\hat{h}_{\mathtt{rot}}$	0.488	3.1	8.8	53.0	2.5	62.7	0.327	0.326	0.199	0.417
$\hat{h}_{\mathtt{dpi}}$	0.319	24.5	47.0	81.0	20.3	91.9	0.366	0.387	0.298	0.504
$x = 0$										
$h_{\mathtt{mse}}$	0.115	52.7	72.9	83.3	61.7	93.6	0.596	0.625	0.665	0.826
$\hat{h}_{\mathtt{rot}}$	0.464	0.0	0.0	0.0	0.0	0.0	0.354	0.328	0.199	0.462
$\hat{h}_{\mathtt{dpi}}$	0.238	1.9	4.1	43.9	2.2	55.9	0.464	0.444	0.398	0.591
$x = 1/3$										
$h_{\mathtt{mse}}$	0.383	92.1	94.2	77.9	82.1	91.4	0.318	0.354	0.239	0.455
$\hat{h}_{\mathtt{rot}}$	0.340	94.1	94.5	79.0	87.4	92.9	0.340	0.378	0.280	0.488
$\hat{h}_{\mathtt{dpi}}$	0.314	95.3	95.5	77.6	90.6	91.9	0.351	0.388	0.298	0.504
$x = 2/3$										
$h_{\mathtt{mse}}$	0.478	58.8	78.2	83.0	32.4	94.5	0.302	0.331	0.198	0.423
$\hat{h}_{\mathtt{rot}}$	0.289	88.6	92.2	82.5	82.4	94.3	0.366	0.403	0.325	0.525
$\hat{h}_{\mathtt{dpi}}$	0.219	92.4	93.7	82.3	92.2	94.4	0.422	0.462	0.431	0.606

Panel B: Summary Statistics for the Estimated Bandwidths

	Pop. Par.	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.	Std. Dev.
$x = -2/3$								
$\hat{h}_{\mathtt{rot}}$	0.478	0.158	0.183	0.191	0.201	0.201	0.671	0.049
$\hat{h}_{\mathtt{dpi}}$	-	0.0513	0.166	0.178	0.177	0.19	0.32	0.024
$x = -1/3$								
$\hat{h}_{\mathtt{rot}}$	0.331	0.223	0.401	0.497	0.488	0.576	0.73	0.109
$\hat{h}_{\mathtt{dpi}}$	-	0.0827	0.284	0.312	0.319	0.343	0.577	0.064
$x = 0$								
$\hat{h}_{\mathtt{rot}}$	0.115	0.32	0.433	0.462	0.464	0.491	0.676	0.046
$\hat{h}_{\mathtt{dpi}}$	-	0.0661	0.212	0.24	0.238	0.265	0.577	0.046
$x = 1/3$								
$\hat{h}_{\mathtt{rot}}$	0.383	0.206	0.281	0.337	0.34	0.39	0.65	0.067
$\hat{h}_{\mathtt{dpi}}$	-	0.0782	0.291	0.313	0.314	0.336	0.576	0.044
$x = 2/3$								
$\hat{h}_{\mathtt{rot}}$	0.478	0.211	0.254	0.279	0.289	0.318	0.505	0.045
$\hat{h}_{\mathtt{dpi}}$	-	0.0667	0.196	0.212	0.219	0.233	0.577	0.044

Notes:
(i) US = Undersmoothing, Locfit = `R` package `locfit` by Loader (2013), BC = Bias Corrected, HH = Hall and Horowitz (2013), RBC = Robust Bias Corrected.
(ii) "Bandwidth" column report the population and average estimated bandwidths choices, as appropriate, for bandwidth h_n.
(iii) The population MSE-optimal choice $h_{\mathtt{mse}}$ coincides with the population ROT optimal coverage error rate $h^{\mathtt{rot}}_{\mathtt{rbc}}$.
(iv) For some evaluation points, $h_{\mathtt{mse}}$ is not well defined so it was left missing.

Table S.II.2: Simulations Results for Model 2

Panel A: Empirical Coverage and Average Interval Length of 95% Confidence Intervals

	Bandwidth	**Empirical Coverage**					**Interval Length**			
		US	Locfit	BC	HH	RBC	US	Locfit	HH	RBC
$x = -2/3$										
$h_{\texttt{mse}}$	-	-	-	-	-	-	-	-	-	-
$\hat{h}_{\texttt{rot}}$	0.325	95.2	95.6	83.7	87.3	95.4	0.351	0.388	0.282	0.504
$\hat{h}_{\texttt{dpi}}$	0.205	95.2	95.5	83.2	94.4	95.3	0.433	0.473	0.421	0.622
$x = -1/3$										
$h_{\texttt{mse}}$	0.706	0.0	0.5	1.5	0.0	4.8	0.254	0.268	0.122	0.356
$\hat{h}_{\texttt{rot}}$	0.461	1.1	18.0	83.5	0.2	94.7	0.304	0.327	0.188	0.418
$\hat{h}_{\texttt{dpi}}$	0.440	13.9	33.6	69.0	8.2	84.2	0.311	0.336	0.203	0.432
$x = 0$										
$h_{\texttt{mse}}$	0.115	52.7	72.8	83.3	49.7	93.6	0.596	0.625	0.576	0.826
$\hat{h}_{\texttt{rot}}$	0.495	0.0	0.0	0.0	0.0	0.0	0.341	0.315	0.174	0.451
$\hat{h}_{\texttt{dpi}}$	0.238	2.3	4.3	43.6	2.1	55.3	0.464	0.444	0.370	0.591
$x = 1/3$										
$h_{\texttt{mse}}$	0.706	0.0	0.4	1.7	0.0	5.1	0.254	0.268	0.122	0.356
$\hat{h}_{\texttt{rot}}$	0.461	1.0	18.4	82.7	0.1	93.7	0.303	0.326	0.188	0.417
$\hat{h}_{\texttt{dpi}}$	0.440	14.0	33.0	68.7	8.4	83.9	0.311	0.336	0.202	0.430
$x = 2/3$										
$h_{\texttt{mse}}$	-	-	-	-	-	-	-	-	-	-
$\hat{h}_{\texttt{rot}}$	0.325	94.8	95.4	82.2	87.0	94.3	0.351	0.388	0.282	0.504
$\hat{h}_{\texttt{dpi}}$	0.205	94.9	94.8	82.1	94.1	93.9	0.434	0.473	0.421	0.623

Panel B: Summary Statistics for the Estimated Bandwidths

	Pop. Par.	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.	Std. Dev.
$x = -2/3$								
$\hat{h}_{\texttt{rot}}$	-	0.205	0.261	0.292	0.325	0.377	0.583	0.083
$\hat{h}_{\texttt{dpi}}$	-	0.0647	0.19	0.205	0.205	0.221	0.356	0.026
$x = -1/3$								
$\hat{h}_{\texttt{rot}}$	0.706	0.32	0.43	0.458	0.461	0.488	0.69	0.043
$\hat{h}_{\texttt{dpi}}$	-	0.151	0.379	0.426	0.44	0.5	0.577	0.081
$x = 0$								
$\hat{h}_{\texttt{rot}}$	0.115	0.395	0.469	0.492	0.495	0.518	0.67	0.037
$\hat{h}_{\texttt{dpi}}$	-	0.0585	0.212	0.241	0.238	0.266	0.385	0.045
$x = 1/3$								
$\hat{h}_{\texttt{rot}}$	0.706	0.309	0.432	0.457	0.461	0.487	0.666	0.043
$\hat{h}_{\texttt{dpi}}$	-	0.178	0.38	0.427	0.44	0.499	0.577	0.081
$x = 2/3$								
$\hat{h}_{\texttt{rot}}$	-	0.207	0.261	0.294	0.325	0.379	0.568	0.083
$\hat{h}_{\texttt{dpi}}$	-	0.0459	0.19	0.205	0.205	0.221	0.373	0.026

Notes:
(i) US = Undersmoothing, Locfit = `R` package `locfit` by Loader (2013), BC = Bias Corrected, HH = Hall and Horowitz (2013), RBC = Robust Bias Corrected.
(ii) "Bandwidth" column report the population and average estimated bandwidths choices, as appropriate, for bandwidth h_n.
(iii) The population MSE-optimal choice $h_{\texttt{mse}}$ coincides with the population ROT optimal coverage error rate $h^{\texttt{rot}}_{\texttt{rbc}}$.
(iv) For some evaluation points, $h_{\texttt{mse}}$ is not well defined so it was left missing.

Table S.II.3: Simulations Results for Model 3

Panel A: Empirical Coverage and Average Interval Length of 95% Confidence Intervals

	Bandwidth	Empirical Coverage					Interval Length			
		US	Locfit	BC	HH	RBC	US	Locfit	HH	RBC
$x = -2/3$										
$h_{\texttt{mse}}$	1.235	86.3	86.3	87.5	29.8	87.7	0.285	0.298	0.078	0.286
$\hat{h}_{\texttt{rot}}$	0.530	91.3	91.7	86.3	64.8	95.5	0.299	0.313	0.166	0.406
$\hat{h}_{\texttt{dpi}}$	0.266	94.9	95.0	83.1	88.1	95.5	0.380	0.412	0.309	0.546
$x = -1/3$										
$h_{\texttt{mse}}$	1.235	83.2	81.2	67.2	32.7	81.8	0.206	0.210	0.070	0.266
$\hat{h}_{\texttt{rot}}$	0.697	80.6	86.4	82.2	43.9	94.4	0.233	0.253	0.116	0.336
$\hat{h}_{\texttt{dpi}}$	0.493	90.2	93.2	82.6	67.7	94.8	0.278	0.303	0.166	0.400
$x = 0$										
$h_{\texttt{mse}}$	0.976	13.8	19.8	40.3	1.3	63.2	0.198	0.215	0.082	0.283
$\hat{h}_{\texttt{rot}}$	0.696	34.2	65.3	84.6	7.9	96.0	0.234	0.254	0.116	0.334
$\hat{h}_{\texttt{dpi}}$	0.354	93.2	94.7	82.8	79.9	95.5	0.327	0.356	0.231	0.470
$x = 1/3$										
$h_{\texttt{mse}}$	0.246	77.8	85.2	79.0	67.4	92.6	0.393	0.424	0.327	0.562
$\hat{h}_{\texttt{rot}}$	0.697	86.0	82.4	49.2	51.2	72.1	0.237	0.253	0.116	0.343
$\hat{h}_{\texttt{dpi}}$	0.491	75.0	68.2	47.9	45.3	71.4	0.282	0.303	0.167	0.406
$x = 2/3$										
$h_{\texttt{mse}}$	0.246	78.3	85.6	79.6	67.1	93.0	0.394	0.425	0.327	0.565
$\hat{h}_{\texttt{rot}}$	0.504	78.2	76.2	46.5	47.5	69.2	0.309	0.321	0.177	0.424
$\hat{h}_{\texttt{dpi}}$	0.267	76.9	84.1	77.8	63.5	91.7	0.381	0.412	0.308	0.547

Panel B: Summary Statistics for the Estimated Bandwidths

	Pop. Par.	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.	Std. Dev.
$x = -2/3$								
$\hat{h}_{\texttt{rot}}$	-	0.25	0.436	0.529	0.53	0.617	0.822	0.119
$\hat{h}_{\texttt{dpi}}$	-	0.0666	0.242	0.262	0.266	0.283	0.576	0.043
$x = -1/3$								
$\hat{h}_{\texttt{rot}}$	-	0.495	0.667	0.703	0.697	0.732	0.833	0.050
$\hat{h}_{\texttt{dpi}}$	-	0.276	0.439	0.492	0.493	0.571	0.577	0.068
$x = 0$								
$\hat{h}_{\texttt{rot}}$	0.976	0.484	0.667	0.704	0.696	0.731	0.826	0.051
$\hat{h}_{\texttt{dpi}}$	-	0.125	0.326	0.347	0.354	0.373	0.577	0.046
$x = 1/3$								
$\hat{h}_{\texttt{rot}}$	0.246	0.469	0.665	0.703	0.697	0.734	0.862	0.052
$\hat{h}_{\texttt{dpi}}$	-	0.201	0.436	0.49	0.491	0.57	0.577	0.069
$x = 2/3$								
$\hat{h}_{\texttt{rot}}$	0.246	0.222	0.392	0.497	0.504	0.609	0.836	0.132
$\hat{h}_{\texttt{dpi}}$	-	0.0659	0.243	0.262	0.267	0.284	0.577	0.045

Notes:
(i) US = Undersmoothing, Locfit = `R` package `locfit` by Loader (2013), BC = Bias Corrected, HH = Hall and Horowitz (2013), RBC = Robust Bias Corrected.
(ii) "Bandwidth" column report the population and average estimated bandwidths choices, as appropriate, for bandwidth h_n.
(iii) The population MSE-optimal choice $h_{\texttt{mse}}$ coincides with the population ROT optimal coverage error rate $h^{\texttt{rot}}_{\texttt{rbc}}$.
(iv) For some evaluation points, $h_{\texttt{mse}}$ is not well defined so it was left missing.

Table S.II.4: Simulations Results for Model 4

Panel A: Empirical Coverage and Average Interval Length of 95% Confidence Intervals

	Bandwidth	Empirical Coverage					Interval Length			
		US	Locfit	BC	HH	RBC	US	Locfit	HH	RBC
$x = -2/3$										
$h_{\mathtt{mse}}$	-	-	-	-	-	-	-	-	-	-
$\hat{h}_{\mathtt{rot}}$	0.309	95.2	95.5	83.5	88.5	95.4	0.358	0.394	0.295	0.515
$\hat{h}_{\mathtt{dpi}}$	0.200	95.2	95.3	83.2	94.7	95.3	0.439	0.478	0.431	0.630
$x = -1/3$										
$h_{\mathtt{mse}}$	0.466	0.5	8.5	76.2	0.0	89.8	0.301	0.323	0.185	0.413
$\hat{h}_{\mathtt{rot}}$	0.441	0.8	14.0	82.6	0.1	94.2	0.309	0.332	0.197	0.426
$\hat{h}_{\mathtt{dpi}}$	0.432	10.5	27.1	67.9	6.3	82.5	0.314	0.337	0.207	0.435
$x = 0$										
$h_{\mathtt{mse}}$	0.128	52.4	73.0	83.4	51.0	93.9	0.564	0.593	0.559	0.785
$\hat{h}_{\mathtt{rot}}$	0.473	0.0	0.0	0.0	0.0	0.1	0.348	0.321	0.183	0.447
$\hat{h}_{\mathtt{dpi}}$	0.233	3.2	7.4	58.5	3.1	72.0	0.457	0.447	0.378	0.592
$x = 1/3$										
$h_{\mathtt{mse}}$	0.466	0.5	9.2	75.0	0.1	89.4	0.301	0.322	0.185	0.412
$\hat{h}_{\mathtt{rot}}$	0.441	0.6	15.1	81.7	0.0	93.2	0.309	0.332	0.197	0.425
$\hat{h}_{\mathtt{dpi}}$	0.433	10.4	26.7	67.0	6.1	82.5	0.313	0.337	0.206	0.433
$x = 2/3$										
$h_{\mathtt{mse}}$	-	-	-	-	-	-	-	-	-	-
$\hat{h}_{\mathtt{rot}}$	0.309	94.6	95.2	82.1	88.2	94.3	0.359	0.394	0.295	0.515
$\hat{h}_{\mathtt{dpi}}$	0.200	94.7	94.7	82.2	94.5	94.1	0.440	0.478	0.431	0.631

Panel B: Summary Statistics for the Estimated Bandwidths

	Pop. Par.	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.	Std. Dev.
$x = -2/3$								
$\hat{h}_{\mathtt{rot}}$	-	0.203	0.254	0.28	0.309	0.341	0.572	0.077
$\hat{h}_{\mathtt{dpi}}$	-	0.0544	0.186	0.2	0.2	0.215	0.354	0.026
$x = -1/3$								
$\hat{h}_{\mathtt{rot}}$	0.466	0.309	0.413	0.438	0.441	0.466	0.643	0.039
$\hat{h}_{\mathtt{dpi}}$	-	0.122	0.373	0.418	0.432	0.487	0.577	0.082
$x = 0$								
$\hat{h}_{\mathtt{rot}}$	0.128	0.382	0.449	0.47	0.473	0.493	0.623	0.033
$\hat{h}_{\mathtt{dpi}}$	-	0.0301	0.21	0.236	0.233	0.259	0.373	0.042
$x = 1/3$								
$\hat{h}_{\mathtt{rot}}$	0.466	0.303	0.414	0.438	0.441	0.465	0.62	0.039
$\hat{h}_{\mathtt{dpi}}$	-	0.13	0.373	0.42	0.433	0.491	0.577	0.082
$x = 2/3$								
$\hat{h}_{\mathtt{rot}}$	-	0.204	0.254	0.281	0.309	0.342	0.566	0.076
$\hat{h}_{\mathtt{dpi}}$	-	0.0448	0.185	0.2	0.2	0.215	0.41	0.026

Notes:
(i) US = Undersmoothing, Locfit = `R` package `locfit` by Loader (2013), BC = Bias Corrected, HH = Hall and Horowitz (2013), RBC = Robust Bias Corrected.
(ii) "Bandwidth" column report the population and average estimated bandwidths choices, as appropriate, for bandwidth h_n.
(iii) The population MSE-optimal choice $h_{\mathtt{mse}}$ coincides with the population ROT optimal coverage error rate $h^{\mathtt{rot}}_{\mathtt{rbc}}$.
(iv) For some evaluation points, $h_{\mathtt{mse}}$ is not well defined so it was left missing.

Table S.II.5: Simulations Results for Model 5

Panel A: Empirical Coverage and Average Interval Length of 95% Confidence Intervals

	Bandwidth	**Empirical Coverage**					**Interval Length**			
		US	Locfit	BC	HH	RBC	US	Locfit	HH	RBC
$x = -2/3$										
$h_{\texttt{mse}}$	-	-	-	-	-	-	-	-	-	-
$\hat{h}_{\texttt{rot}}$	0.251	95.1	95.2	83.6	91.0	95.5	0.392	0.424	0.340	0.563
$\hat{h}_{\texttt{dpi}}$	0.203	95.4	95.3	83.4	93.7	95.0	0.437	0.472	0.410	0.627
$x = -1/3$										
$h_{\texttt{mse}}$	0.307	43.5	69.2	82.6	26.4	94.5	0.355	0.380	0.271	0.504
$\hat{h}_{\texttt{rot}}$	0.405	9.9	27.2	81.3	5.4	93.3	0.316	0.334	0.209	0.440
$\hat{h}_{\texttt{dpi}}$	0.307	44.1	66.3	82.1	31.2	94.1	0.357	0.381	0.275	0.507
$x = 0$										
$h_{\texttt{mse}}$	-	-	-	-	-	-	-	-	-	-
$\hat{h}_{\texttt{rot}}$	0.474	24.9	49.6	78.2	5.5	93.4	0.286	0.309	0.177	0.410
$\hat{h}_{\texttt{dpi}}$	0.320	73.5	83.1	81.0	58.8	93.6	0.348	0.376	0.267	0.498
$x = 1/3$										
$h_{\texttt{mse}}$	0.821	3.3	38.3	80.7	0.1	92.7	0.227	0.241	0.102	0.319
$\hat{h}_{\texttt{rot}}$	0.538	72.3	88.1	76.4	44.8	91.1	0.268	0.293	0.158	0.384
$\hat{h}_{\texttt{dpi}}$	0.343	93.3	93.7	82.0	83.1	94.5	0.332	0.361	0.245	0.477
$x = 2/3$										
$h_{\texttt{mse}}$	0.887	91.3	94.1	74.1	46.7	80.0	0.289	0.312	0.107	0.317
$\hat{h}_{\texttt{rot}}$	0.401	93.5	93.8	82.8	78.4	94.6	0.319	0.342	0.218	0.455
$\hat{h}_{\texttt{dpi}}$	0.262	94.2	94.5	82.0	89.2	94.3	0.386	0.418	0.329	0.554

Panel B: Summary Statistics for the Estimated Bandwidths

	Pop. Par.	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.	Std. Dev.
$x = -2/3$								
$\hat{h}_{\texttt{rot}}$	-	0.187	0.225	0.24	0.251	0.262	0.56	0.043
$\hat{h}_{\texttt{dpi}}$	-	0.048	0.186	0.201	0.203	0.217	0.576	0.033
$x = -1/3$								
$\hat{h}_{\texttt{rot}}$	0.307	0.253	0.369	0.41	0.405	0.443	0.631	0.054
$\hat{h}_{\texttt{dpi}}$	-	0.0927	0.287	0.307	0.307	0.327	0.577	0.038
$x = 0$								
$\hat{h}_{\texttt{rot}}$	-	0.362	0.443	0.47	0.474	0.501	0.682	0.044
$\hat{h}_{\texttt{dpi}}$	-	0.0843	0.289	0.312	0.32	0.339	0.577	0.055
$x = 1/3$								
$\hat{h}_{\texttt{rot}}$	0.821	0.311	0.478	0.53	0.538	0.597	0.775	0.078
$\hat{h}_{\texttt{dpi}}$	-	0.0837	0.323	0.342	0.343	0.362	0.576	0.034
$x = 2/3$								
$\hat{h}_{\texttt{rot}}$	0.887	0.251	0.344	0.375	0.401	0.422	0.747	0.089
$\hat{h}_{\texttt{dpi}}$	-	0.0589	0.231	0.251	0.262	0.277	0.576	0.056

Notes:
(i) US = Undersmoothing, Locfit = `R` package `locfit` by Loader (2013), BC = Bias Corrected, HH = Hall and Horowitz (2013), RBC = Robust Bias Corrected.
(ii) "Bandwidth" column report the population and average estimated bandwidths choices, as appropriate, for bandwidth h_n.
(iii) The population MSE-optimal choice $h_{\texttt{mse}}$ coincides with the population ROT optimal coverage error rate $h^{\texttt{rot}}_{\texttt{rbc}}$.
(iv) For some evaluation points, $h_{\texttt{mse}}$ is not well defined so it was left missing.

Table S.II.6: Simulations Results for Model 6

Panel A: Empirical Coverage and Average Interval Length of 95% Confidence Intervals

	Bandwidth	**Empirical Coverage**					**Interval Length**			
		US	Locfit	BC	HH	RBC	US	Locfit	HH	RBC
$x = -2/3$										
$h_{\texttt{mse}}$	0.783	88.8	88.3	91.0	45.2	94.6	0.289	0.299	0.113	0.333
$\hat{h}_{\texttt{rot}}$	0.563	90.7	91.4	85.0	59.5	95.1	0.294	0.303	0.152	0.393
$\hat{h}_{\texttt{dpi}}$	0.359	93.3	94.1	83.2	78.8	95.2	0.334	0.358	0.233	0.479
$x = -1/3$										
$h_{\texttt{mse}}$	0.975	80.3	83.8	77.2	33.4	91.2	0.210	0.218	0.084	0.296
$\hat{h}_{\texttt{rot}}$	0.580	92.0	93.8	83.4	63.3	95.1	0.254	0.276	0.139	0.367
$\hat{h}_{\texttt{dpi}}$	0.475	92.4	93.4	82.5	71.5	94.8	0.283	0.307	0.171	0.408
$x = 0$										
$h_{\texttt{mse}}$	-	-	-	-	-	-	-	-	-	-
$\hat{h}_{\texttt{rot}}$	0.562	87.3	91.2	82.1	59.3	95.4	0.258	0.280	0.143	0.372
$\hat{h}_{\texttt{dpi}}$	0.447	92.3	93.9	82.4	71.6	95.2	0.292	0.317	0.183	0.420
$x = 1/3$										
$h_{\texttt{mse}}$	0.616	52.3	73.3	81.6	19.4	93.8	0.247	0.267	0.129	0.354
$\hat{h}_{\texttt{rot}}$	0.548	66.6	78.9	81.2	36.3	93.1	0.262	0.284	0.146	0.377
$\hat{h}_{\texttt{dpi}}$	0.461	78.8	85.7	81.2	53.6	93.9	0.288	0.312	0.177	0.414
$x = 2/3$										
$h_{\texttt{mse}}$	-	-	-	-	-	-	-	-	-	-
$\hat{h}_{\texttt{rot}}$	0.461	94.3	94.4	83.2	74.6	94.7	0.304	0.318	0.181	0.429
$\hat{h}_{\texttt{dpi}}$	0.347	94.5	94.2	82.4	82.4	94.4	0.340	0.364	0.242	0.487

Panel B: Summary Statistics for the Estimated Bandwidths

	Pop. Par.	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.	Std. Dev.
$x = -2/3$								
$\hat{h}_{\texttt{rot}}$	0.783	0.284	0.515	0.575	0.563	0.621	0.798	0.084
$\hat{h}_{\texttt{dpi}}$	-	0.113	0.302	0.339	0.359	0.398	0.577	0.082
$x = -1/3$								
$\hat{h}_{\texttt{rot}}$	0.975	0.41	0.534	0.574	0.58	0.621	0.801	0.063
$\hat{h}_{\texttt{dpi}}$	-	0.164	0.418	0.47	0.475	0.546	0.577	0.073
$x = 0$								
$\hat{h}_{\texttt{rot}}$	-	0.396	0.52	0.557	0.562	0.6	0.786	0.060
$\hat{h}_{\texttt{dpi}}$	-	0.124	0.387	0.436	0.447	0.506	0.577	0.078
$x = 1/3$								
$\hat{h}_{\texttt{rot}}$	0.616	0.388	0.505	0.542	0.548	0.584	0.764	0.059
$\hat{h}_{\texttt{dpi}}$	-	0.163	0.402	0.449	0.461	0.526	0.577	0.076
$x = 2/3$								
$\hat{h}_{\texttt{rot}}$	-	0.261	0.41	0.452	0.461	0.505	0.786	0.070
$\hat{h}_{\texttt{dpi}}$	-	0.0791	0.291	0.328	0.347	0.384	0.577	0.082

Notes:
(i) US = Undersmoothing, Locfit = `R` package `locfit` by Loader (2013), BC = Bias Corrected, HH = Hall and Horowitz (2013), RBC = Robust Bias Corrected.
(ii) "Bandwidth" column report the population and average estimated bandwidths choices, as appropriate, for bandwidth h_n.
(iii) The population MSE-optimal choice $h_{\texttt{mse}}$ coincides with the population ROT optimal coverage error rate $h^{\texttt{rot}}_{\texttt{rbc}}$.
(iv) For some evaluation points, $h_{\texttt{mse}}$ is not well defined so it was left missing.

Table S.II.7: Empirical Coverage and Average Interval Length of 95% Confidence Intervals

	US			RBC		
	h	EC	IL	h	EC	IL
Model 1						
$x = -2/3$	0.140	94.8	0.523	0.420	94.8	0.442
$x = -1/3$	0.100	94.7	0.625	0.420	94.8	0.434
$x = 0$	0.100	71.3	0.640	0.100	93.7	0.893
$x = 1/3$	0.300	94.6	0.355	0.440	94.3	0.425
$x = 2/3$	0.100	95.0	0.624	0.260	94.9	0.546
Model 2						
$x = -2/3$	0.180	94.9	0.459	0.540	94.9	0.399
$x = -1/3$	0.140	94.8	0.524	0.440	94.9	0.424
$x = 0$	0.100	71.3	0.640	0.100	93.7	0.893
$x = 1/3$	0.140	94.5	0.522	0.440	94.2	0.424
$x = 2/3$	0.260	94.9	0.380	0.280	94.9	0.525
Model 3						
$x = -2/3$	0.140	94.9	0.523	0.420	94.9	0.442
$x = -1/3$	0.200	94.9	0.435	0.400	94.9	0.440
$x = 0$	0.100	94.7	0.628	0.680	94.7	0.337
$x = 1/3$	0.100	93.9	0.623	0.100	94.0	0.887
$x = 2/3$	0.100	94.6	0.624	0.180	94.9	0.658
Model 4						
$x = -2/3$	0.180	94.9	0.459	0.520	94.8	0.406
$x = -1/3$	0.100	94.8	0.625	0.400	94.8	0.444
$x = 0$	0.100	79.3	0.636	0.100	93.9	0.893
$x = 1/3$	0.100	94.4	0.623	0.400	94.2	0.443
$x = 2/3$	0.320	94.9	0.342	0.280	94.9	0.525
Model 5						
$x = -2/3$	0.180	94.9	0.459	0.200	94.8	0.624
$x = -1/3$	0.100	94.7	0.625	0.180	94.6	0.658
$x = 0$	0.100	94.6	0.628	0.240	94.4	0.572
$x = 1/3$	0.140	94.6	0.522	0.260	94.3	0.545
$x = 2/3$	0.200	94.8	0.434	0.280	94.9	0.525
Model 6						
$x = -2/3$	0.140	94.9	0.523	0.600	94.9	0.379
$x = -1/3$	0.140	94.8	0.524	0.420	94.9	0.429
$x = 0$	0.100	94.8	0.628	0.600	94.9	0.359
$x = 1/3$	0.140	94.5	0.522	0.480	94.4	0.401
$x = 2/3$	0.260	94.8	0.380	0.420	94.9	0.442

Notes: Bandwidths are selected ex post as the largest bandwidths yielding good coverage, and as can not be made feasible

Table S.II.8: Empirical Coverage and Average Interval Length of 95% Confidence Intervals

	US ($\lambda = 0.5$)		**US** ($\lambda = 0.7$)		**RBC** ($\hat{h}^{\mathtt{rot}}_{\mathtt{rbc}}$)		**RBC** ($\hat{h}^{\mathtt{dpi}}_{\mathtt{rbc}}$)	
	EC	IL	EC	IL	EC	IL	EC	IL
Model 1								
$x = -2/3$	94.4	0.630	94.7	0.528	94.3	0.630	94.7	0.669
$x = -1/3$	56.5	0.410	21.1	0.362	63.3	0.417	91.9	0.504
$x = 0$	0.0	0.466	0.0	0.414	0.0	0.463	55.9	0.591
$x = 1/3$	93.5	0.479	94.1	0.404	92.4	0.486	91.9	0.504
$x = 2/3$	95.0	0.519	93.3	0.436	94.9	0.522	94.4	0.606
Model 2								
$x = -2/3$	94.9	0.495	95.2	0.416	95.1	0.503	95.3	0.622
$x = -1/3$	92.7	0.408	57.9	0.350	94.4	0.417	84.2	0.432
$x = 0$	0.0	0.455	0.0	0.403	0.0	0.451	55.3	0.591
$x = 1/3$	92.4	0.407	58.0	0.350	93.9	0.417	83.9	0.430
$x = 2/3$	95.3	0.496	95.0	0.417	94.9	0.503	93.9	0.623
Model 3								
$x = -2/3$	94.4	0.384	93.9	0.329	94.9	0.405	95.5	0.546
$x = -1/3$	93.9	0.328	91.4	0.277	94.1	0.336	94.8	0.400
$x = 0$	94.5	0.329	87.5	0.277	95.8	0.334	95.5	0.470
$x = 1/3$	71.2	0.331	77.5	0.281	73.0	0.343	71.4	0.406
$x = 2/3$	81.4	0.399	74.7	0.343	68.9	0.423	91.7	0.547
Model 4								
$x = -2/3$	94.9	0.507	95.1	0.426	95.0	0.513	95.3	0.630
$x = -1/3$	90.2	0.418	51.8	0.358	93.9	0.425	82.5	0.435
$x = 0$	0.0	0.451	0.0	0.403	0.0	0.448	72.0	0.592
$x = 1/3$	90.3	0.417	52.3	0.357	93.5	0.424	82.5	0.433
$x = 2/3$	95.4	0.508	95.0	0.427	94.9	0.514	94.1	0.631
Model 5								
$x = -2/3$	94.6	0.560	95.0	0.470	94.4	0.562	95.0	0.627
$x = -1/3$	85.1	0.437	55.0	0.370	93.1	0.440	94.1	0.507
$x = 0$	90.8	0.402	73.5	0.340	92.0	0.410	93.6	0.498
$x = 1/3$	94.4	0.378	94.1	0.319	92.2	0.385	94.5	0.477
$x = 2/3$	95.2	0.442	94.7	0.373	95.0	0.454	94.3	0.554
Model 6								
$x = -2/3$	94.3	0.368	93.2	0.317	94.9	0.392	95.2	0.479
$x = -1/3$	94.9	0.362	94.4	0.305	94.5	0.366	94.8	0.408
$x = 0$	94.1	0.367	93.0	0.309	94.9	0.372	95.2	0.420
$x = 1/3$	92.6	0.372	86.8	0.313	93.6	0.377	93.9	0.414
$x = 2/3$	94.8	0.407	94.5	0.344	94.7	0.427	94.4	0.487

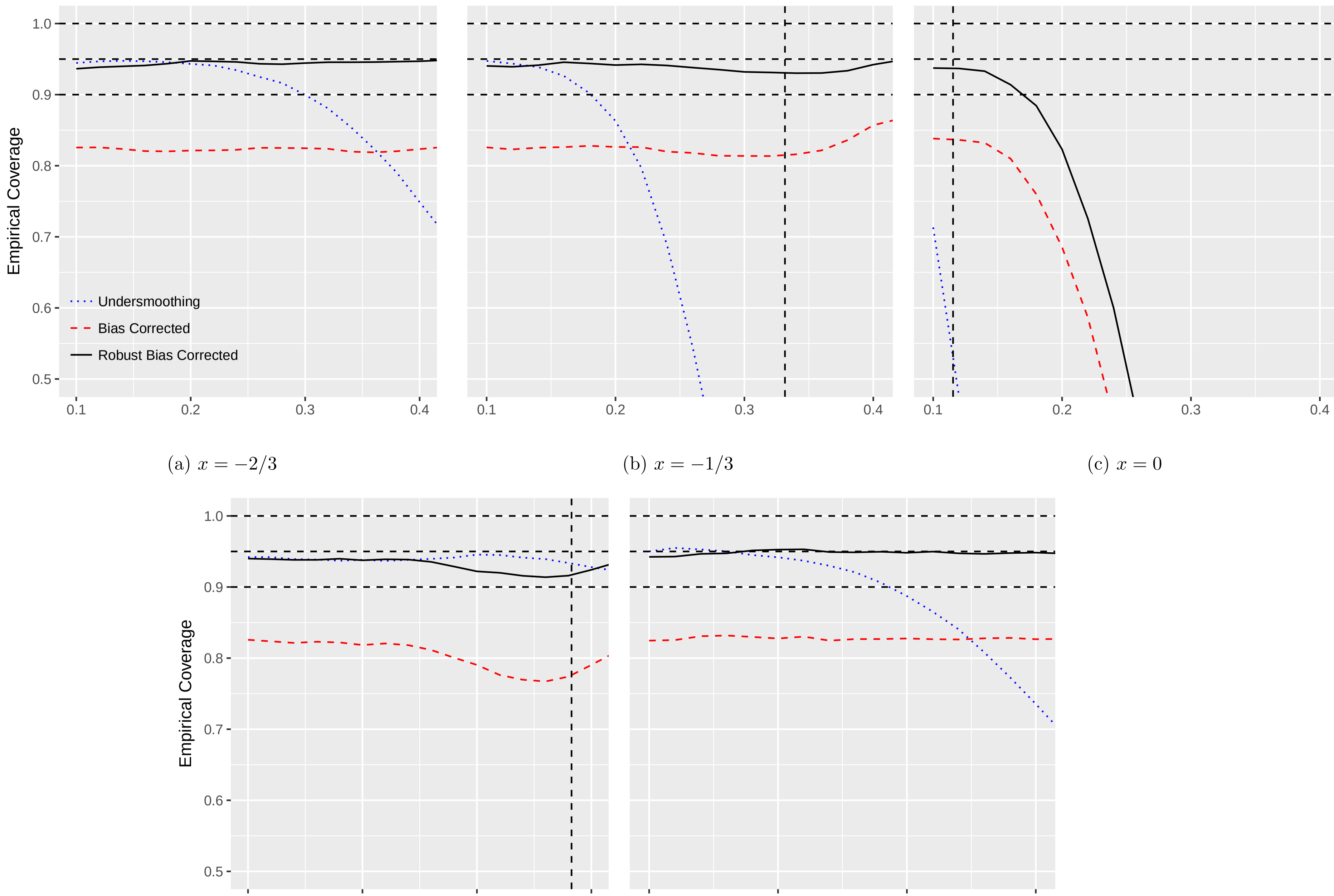
Notes: Undersmoothing is implemented using bandwidths $h = \lambda\hat{h}_{\mathtt{mse}}$ for $\lambda = \{0.5; 0.7\}$, in the columns labeled as such.

Table S.II.9: Empirical Coverage and Average Interval Length of RBC 95% Confidence Intervals for Model 5, for Different Variance Estimators

	h	EC	IL
$x = -2/3$			
HC_0	0.248	94.2	0.555
HC_1	0.249	94.4	0.562
HC_2	0.249	94.4	0.559
HC_3	0.250	94.4	0.562
NN	0.249	93.9	0.560
$x = -1/3$			
HC_0	0.402	92.9	0.437
HC_1	0.403	93.1	0.440
HC_2	0.403	92.9	0.439
HC_3	0.404	93.1	0.440
NN	0.399	92.7	0.441
$x = 0$			
HC_0	0.473	91.9	0.408
HC_1	0.474	92.0	0.410
HC_2	0.474	91.9	0.409
HC_3	0.474	92.0	0.410
NN	0.474	91.7	0.404
$x = 1/3$			
HC_0	0.534	92.0	0.383
HC_1	0.535	92.2	0.385
HC_2	0.535	92.1	0.384
HC_3	0.536	92.2	0.385
NN	0.541	91.9	0.380
$x = 2/3$			
HC_0	0.396	94.8	0.450
HC_1	0.398	95.0	0.454
HC_2	0.397	94.9	0.452
HC_3	0.398	95.0	0.454
NN	0.400	94.7	0.452

Notes:
(i) The h column reports the average estimated bandwidths $\hat{h}_{\texttt{rot}}$.

Figure S.II.2: Empirical Coverage of 95% Confidence Intervals - Model 1



(a) $x = -2/3$

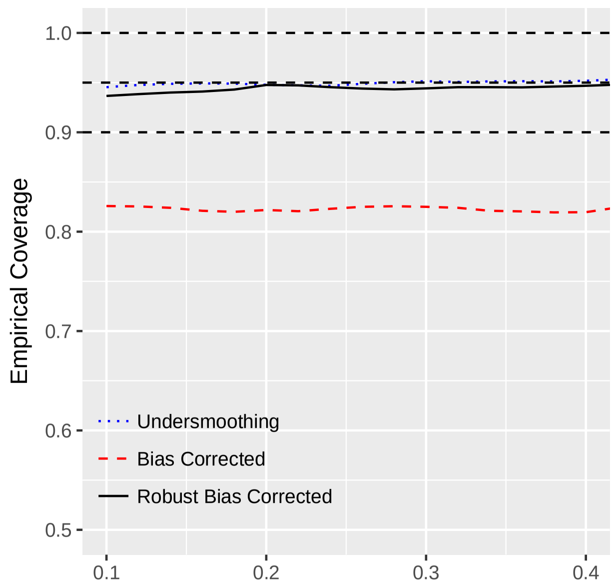
(b) $x = -1/3$

(c) $x = 0$

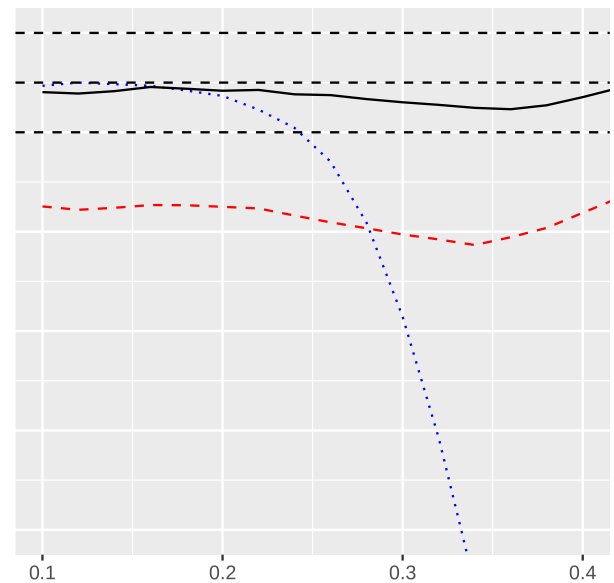
(d) $x = 1/3$

(e) $x = 2/3$

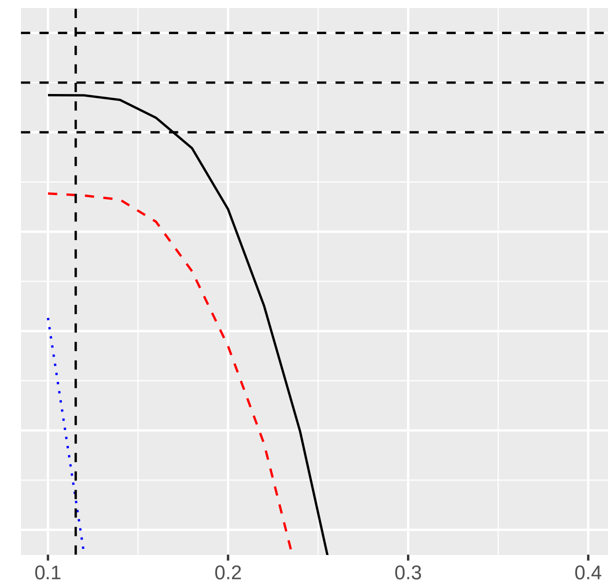
Figure S.II.3: Empirical Coverage of 95% Confidence Intervals - Model 2



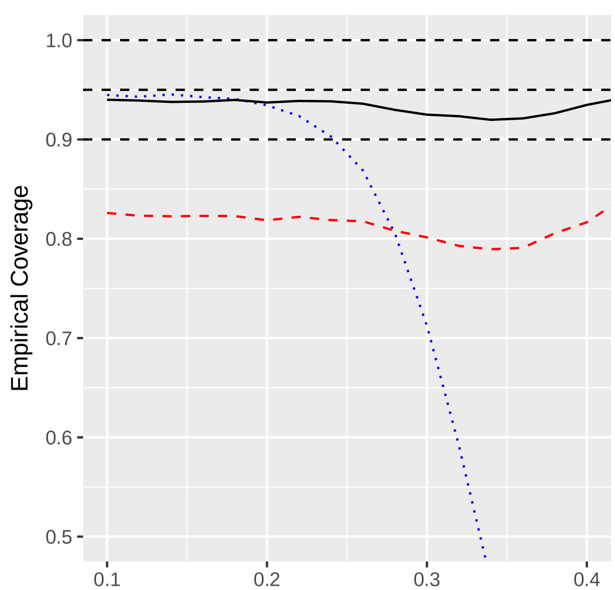
(a) $x = -2/3$



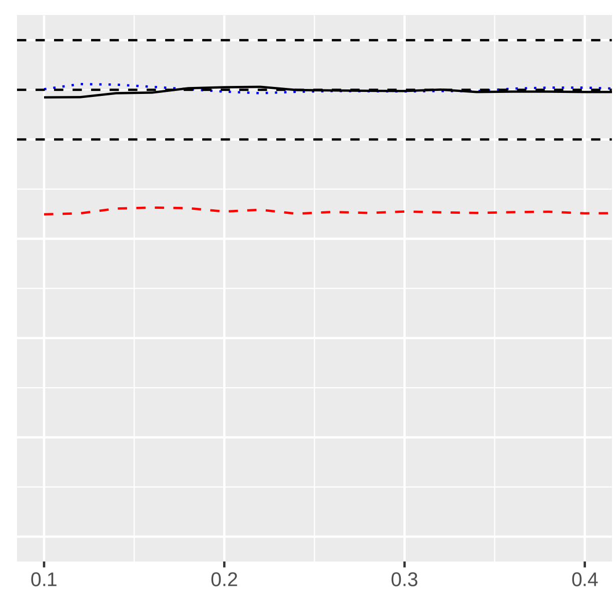
(b) $x = -1/3$



(c) $x = 0$

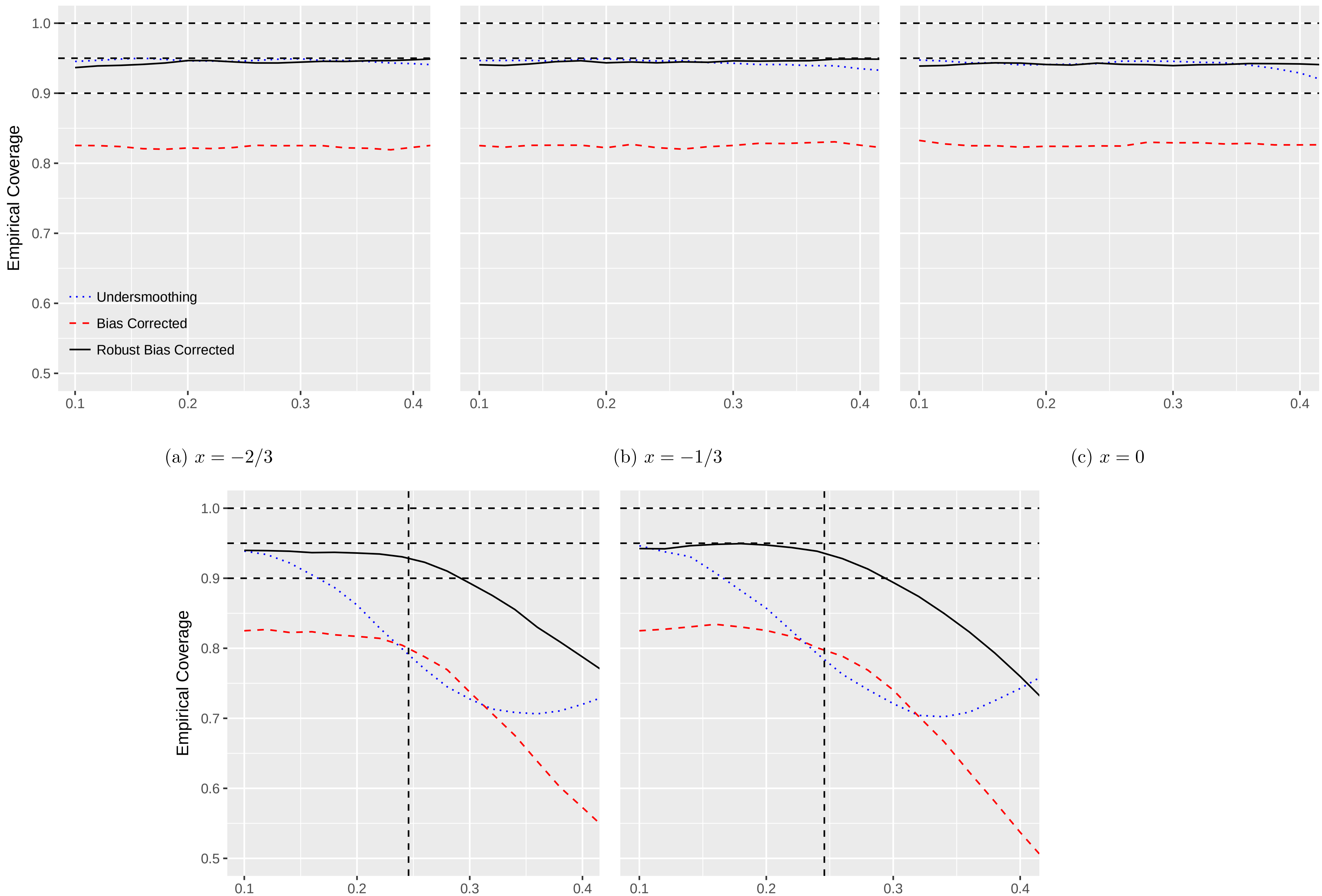


(d) $x = 1/3$



(e) $x = 2/3$

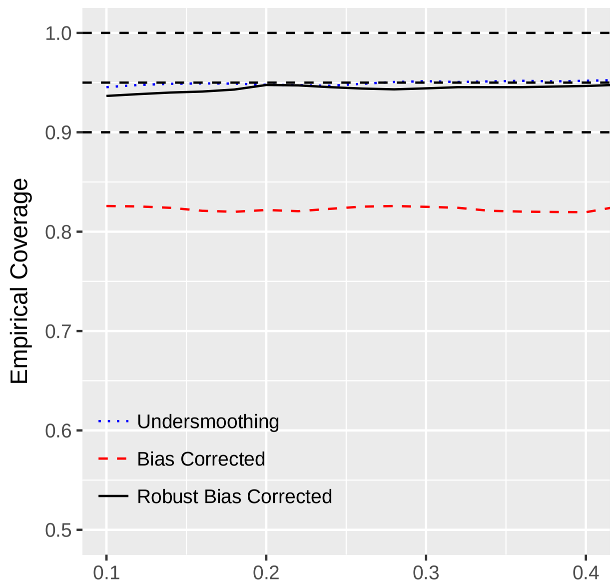
Figure S.II.4: Empirical Coverage of 95% Confidence Intervals - Model 3



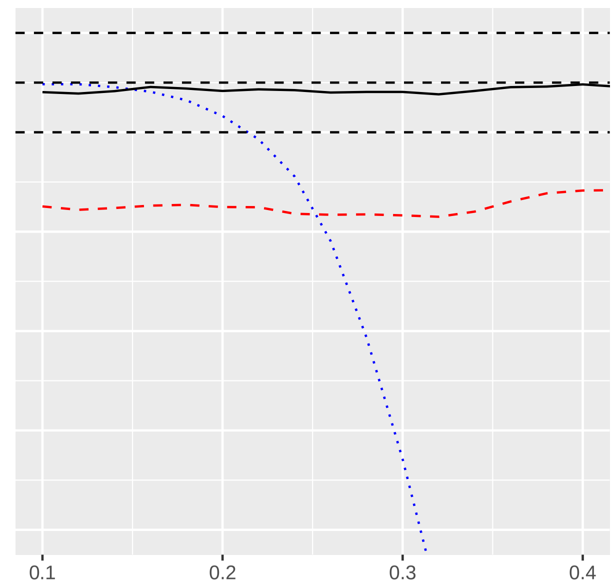
(a) $x = -2/3$ (b) $x = -1/3$ (c) $x = 0$

(d) $x = 1/3$ (e) $x = 2/3$

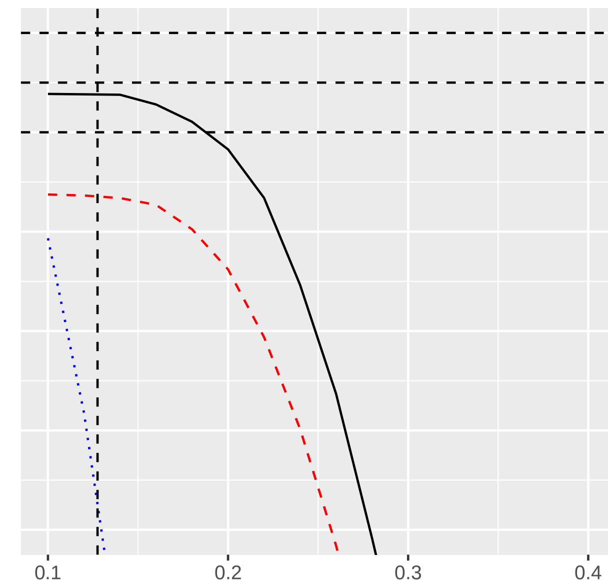
Figure S.II.5: Empirical Coverage of 95% Confidence Intervals - Model 4



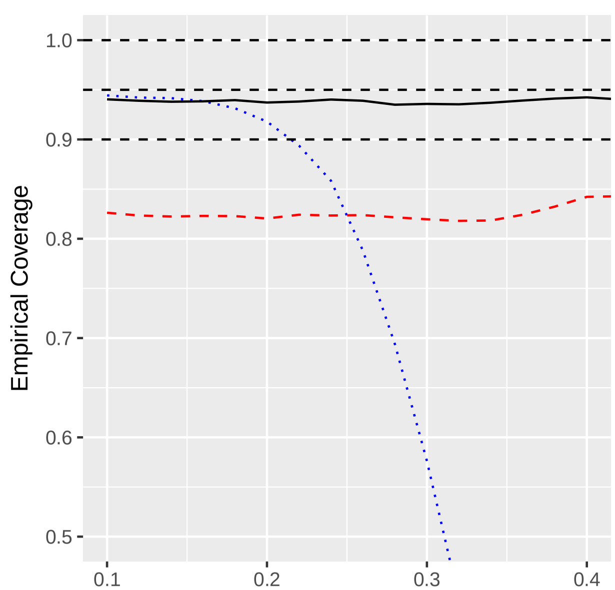
(a) $x = -2/3$



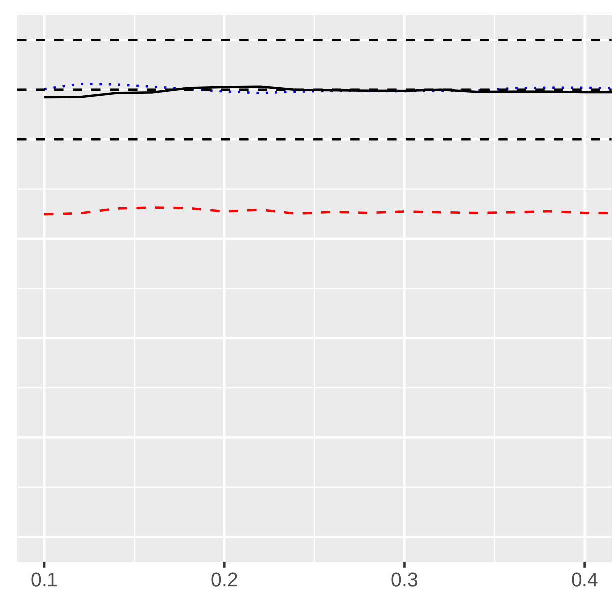
(b) $x = -1/3$



(c) $x = 0$

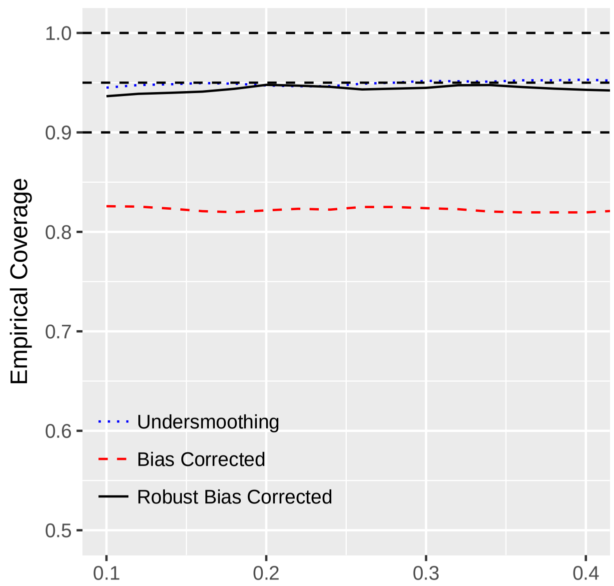


(d) $x = 1/3$

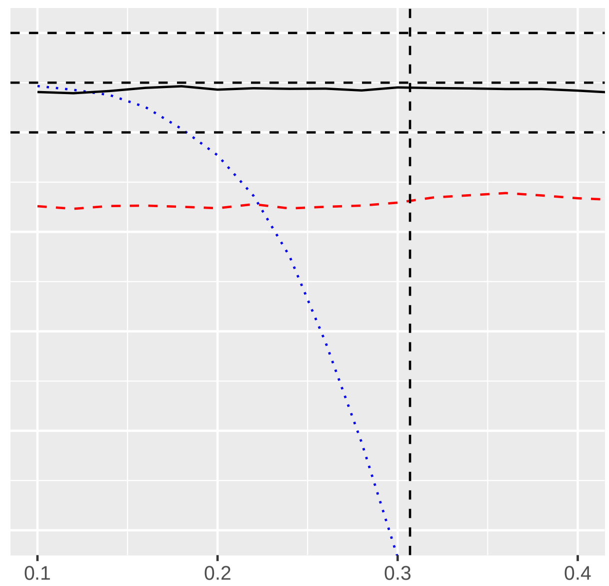


(e) $x = 2/3$

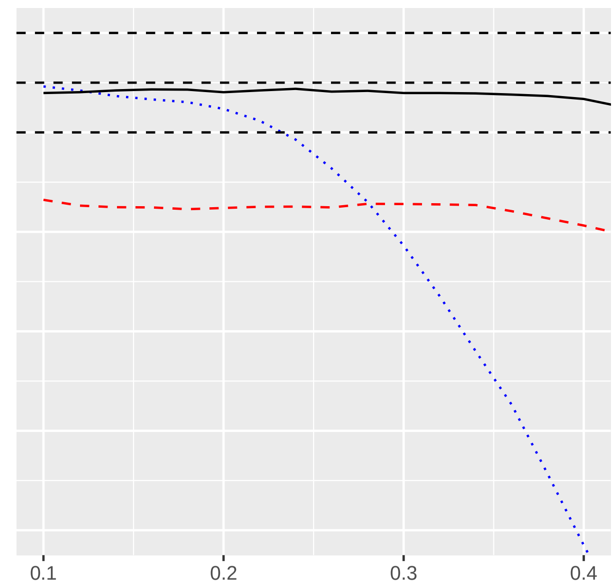
Figure S.II.6: Empirical Coverage of 95% Confidence Intervals - Model 5



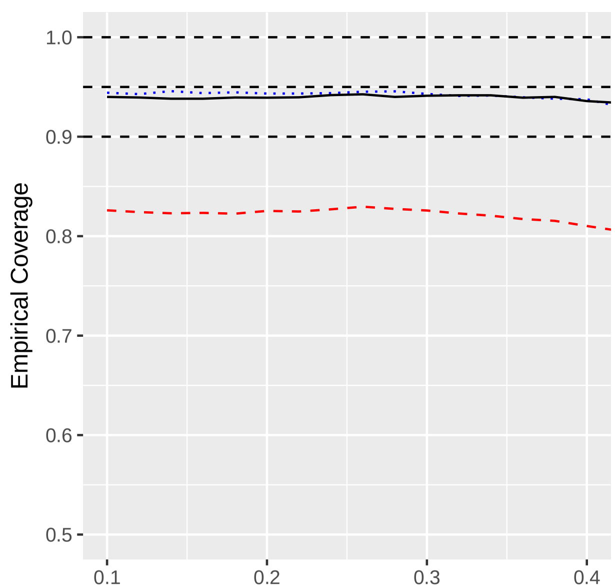
(a) $x = -2/3$



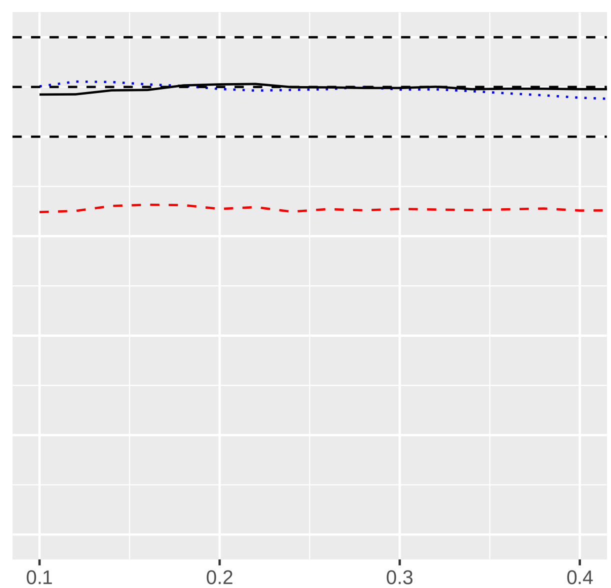
(b) $x = -1/3$



(c) $x = 0$

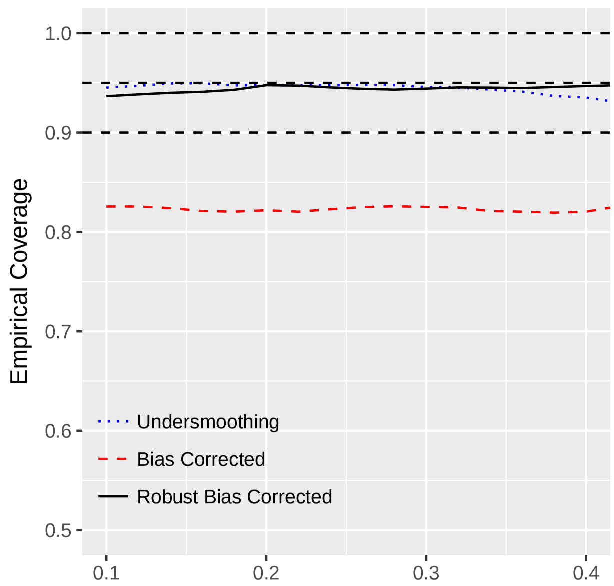


(d) $x = 1/3$

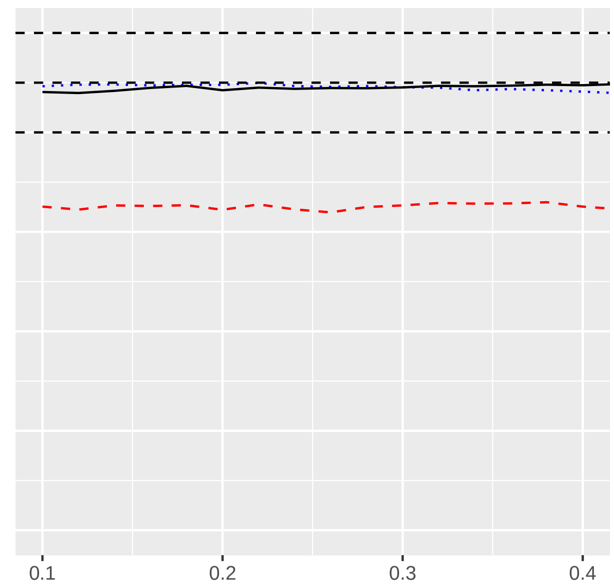


(e) $x = 2/3$

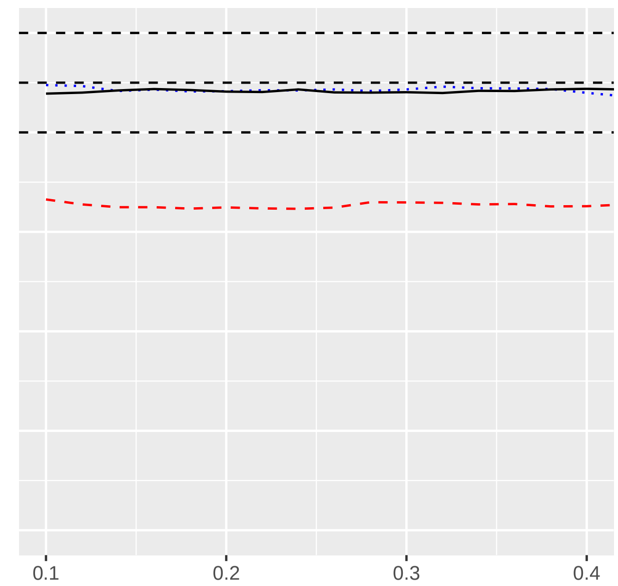
Figure S.II.7: Empirical Coverage of 95% Confidence Intervals - Model 6



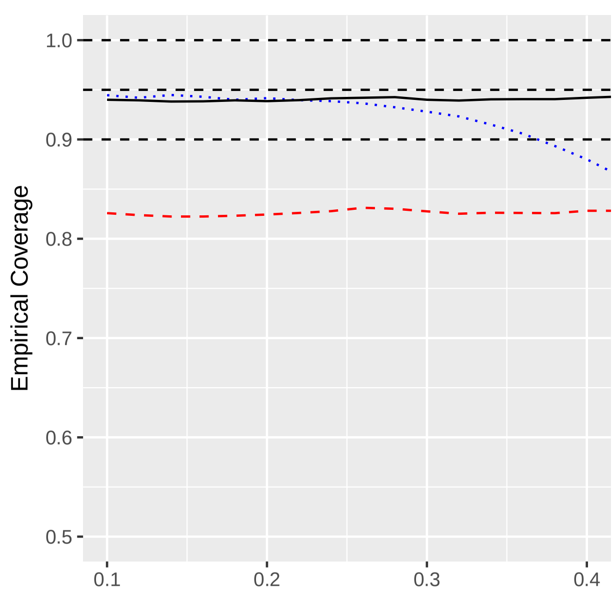
(a) $x = -2/3$



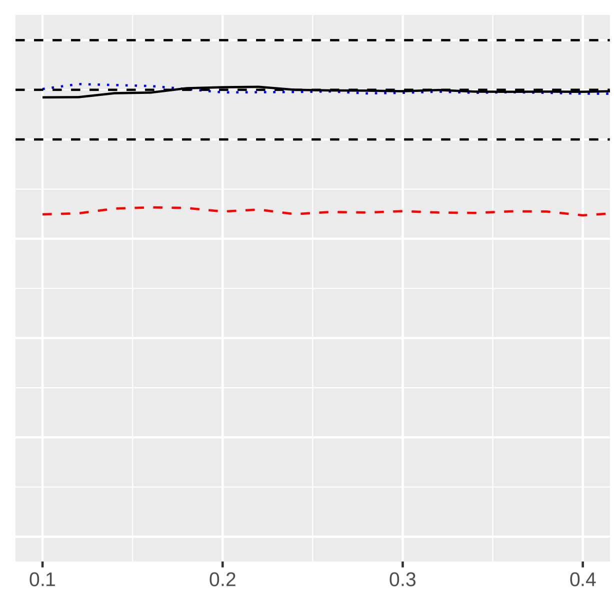
(b) $x = -1/3$



(c) $x = 0$

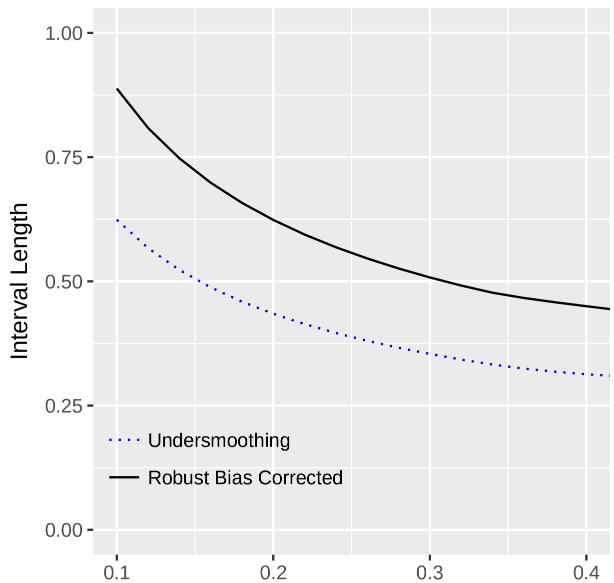


(d) $x = 1/3$

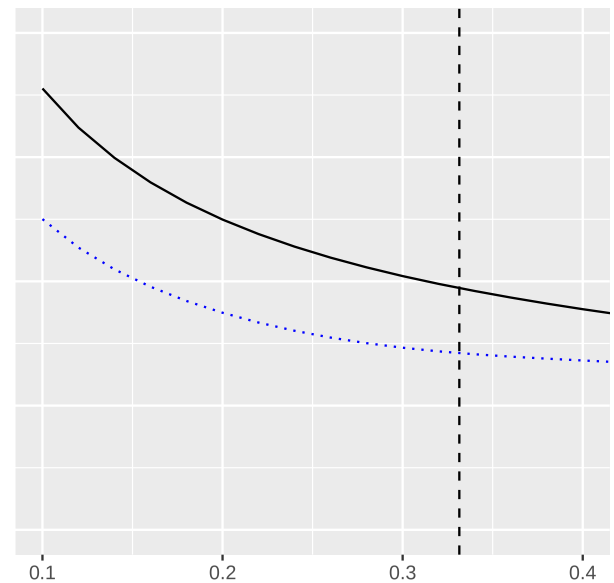


(e) $x = 2/3$

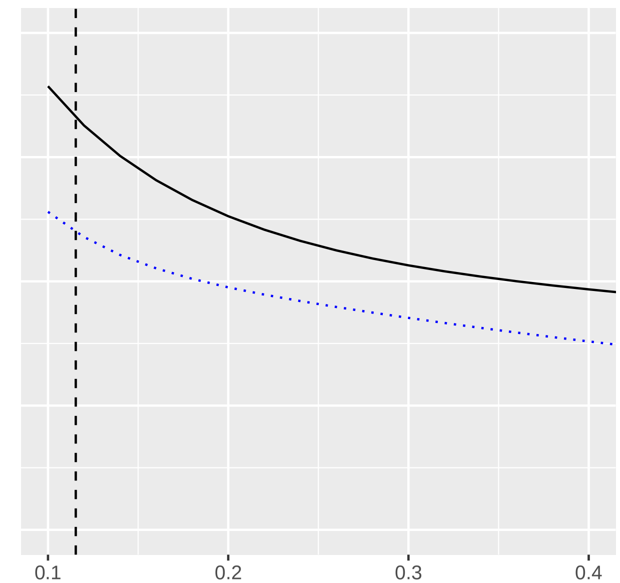
Figure S.II.8: Average Interval Length of 95% Confidence Intervals - Model 1



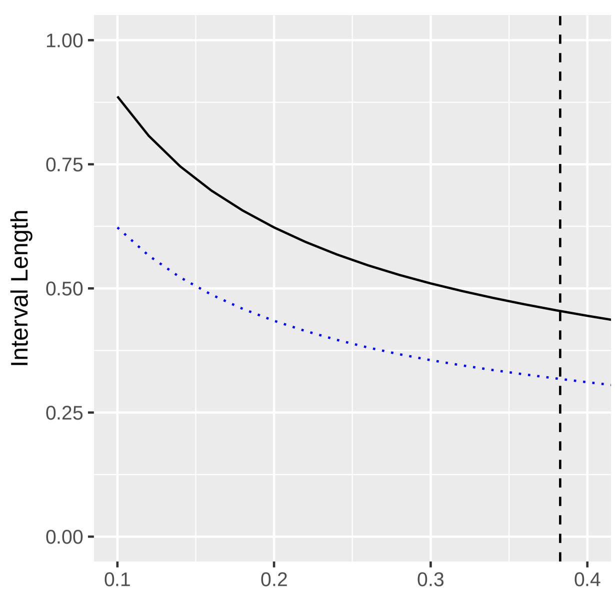
(a) $x = -2/3$



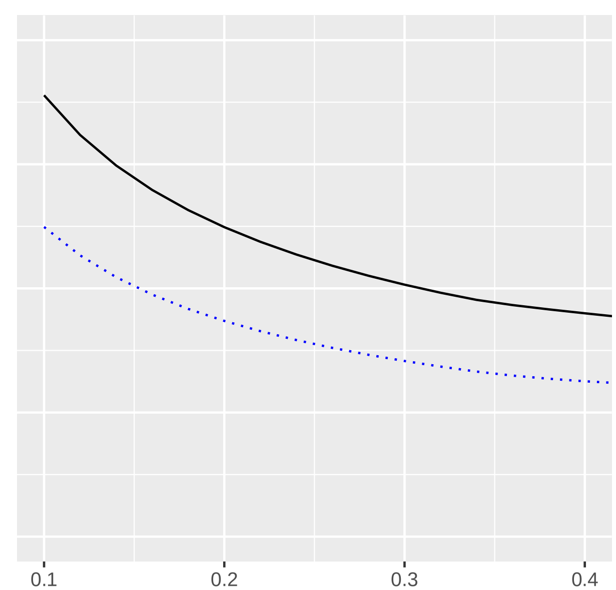
(b) $x = -1/3$



(c) $x = 0$

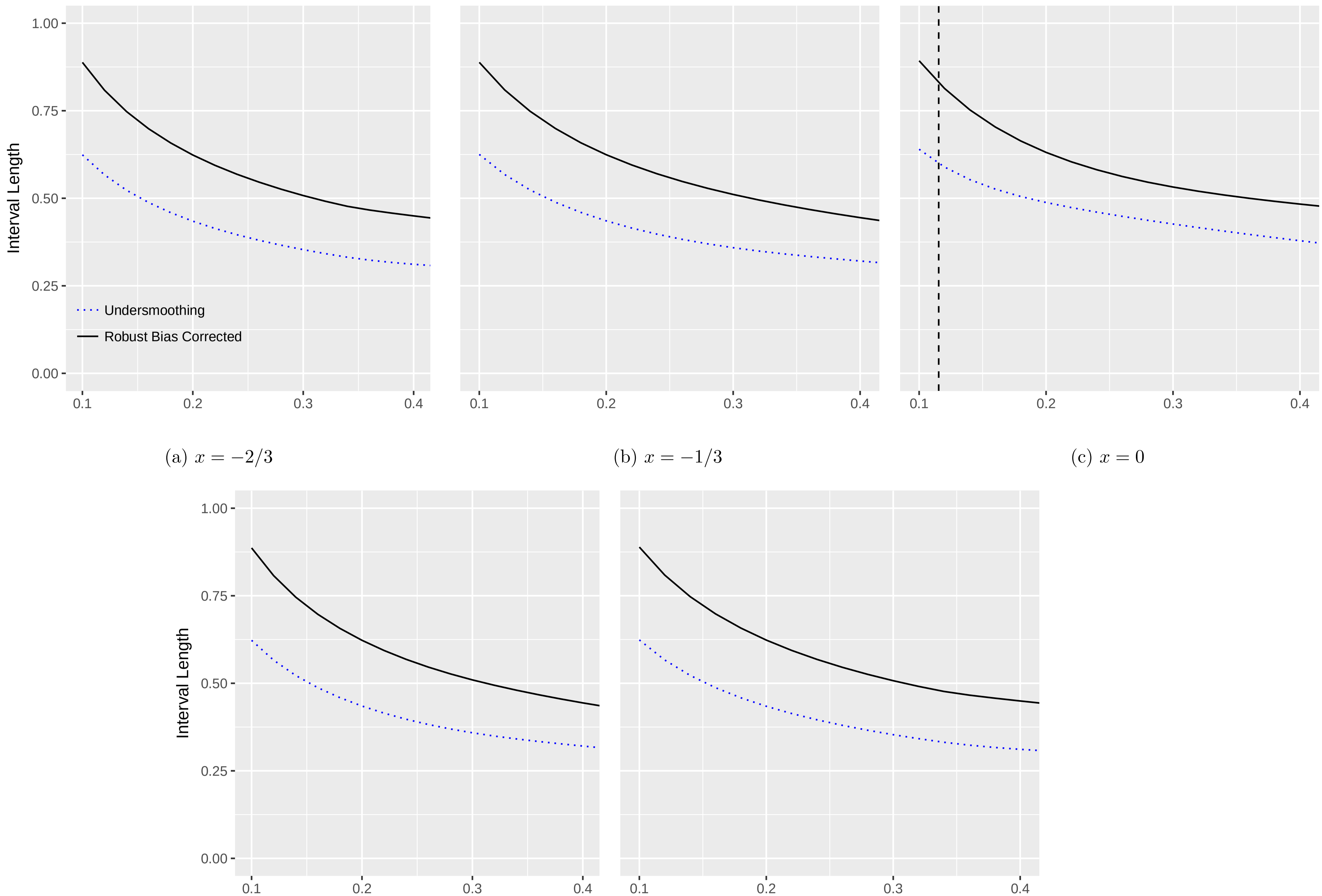


(d) $x = 1/3$



(e) $x = 2/3$

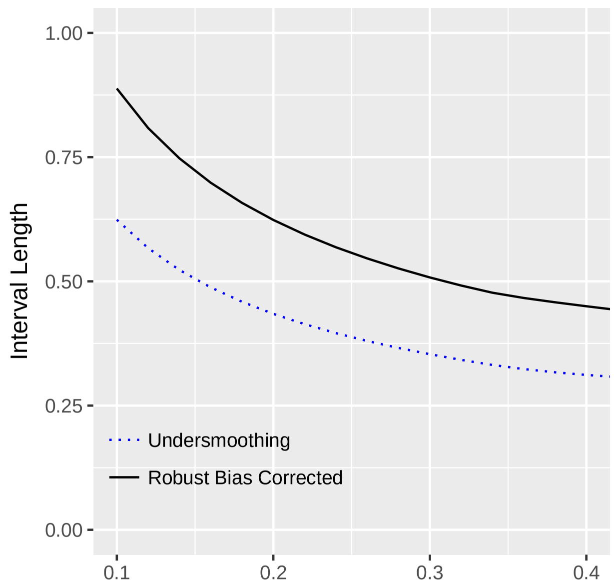
Figure S.II.9: Average Interval Length of 95% Confidence Intervals - Model 2



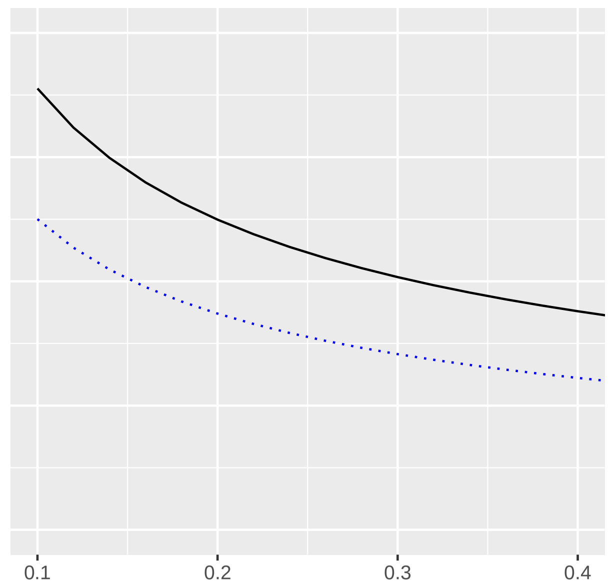
(a) $x = -2/3$ (b) $x = -1/3$ (c) $x = 0$

(d) $x = 1/3$ (e) $x = 2/3$

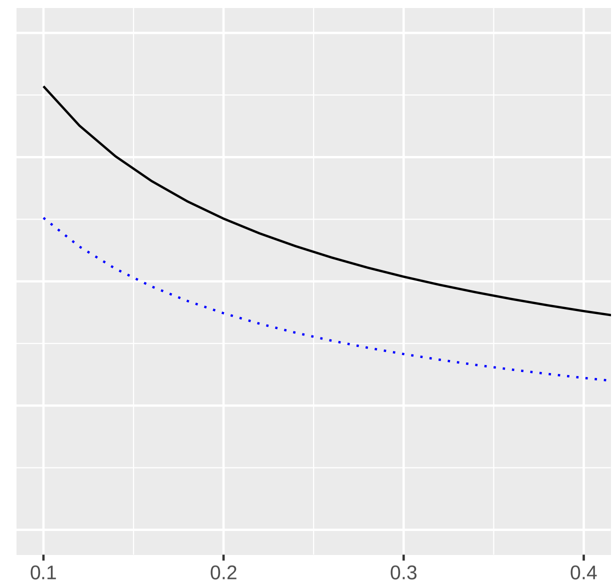
Figure S.II.10: Average Interval Length of 95% Confidence Intervals - Model 3



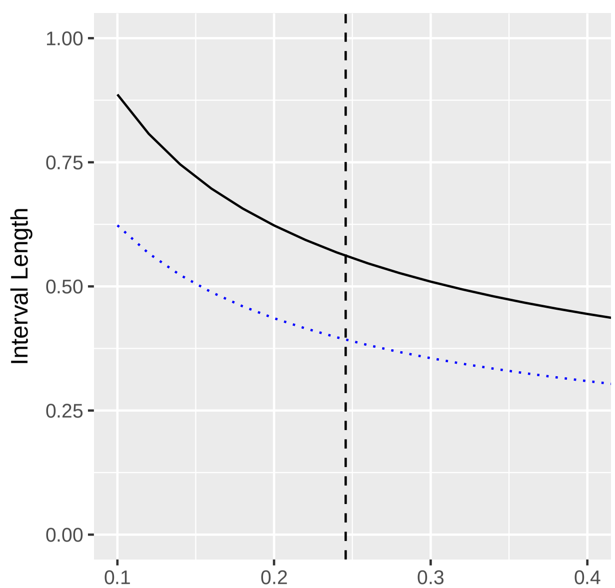
(a) $x = -2/3$



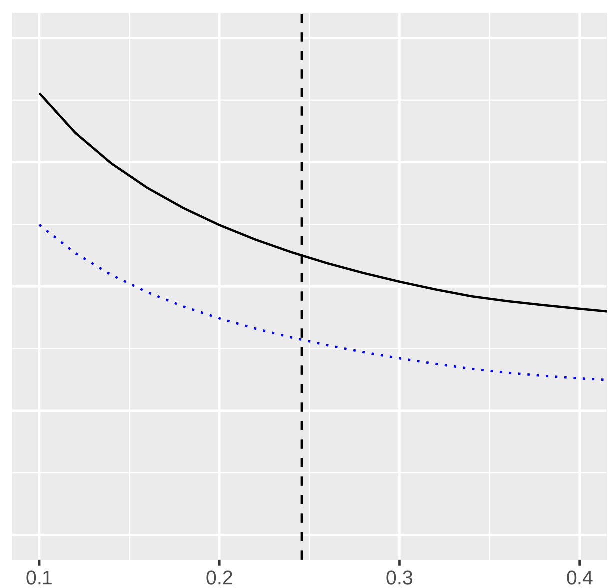
(b) $x = -1/3$



(c) $x = 0$

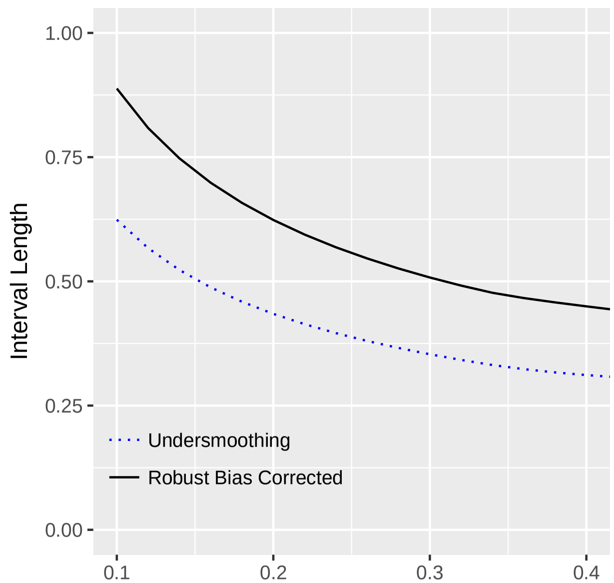


(d) $x = 1/3$

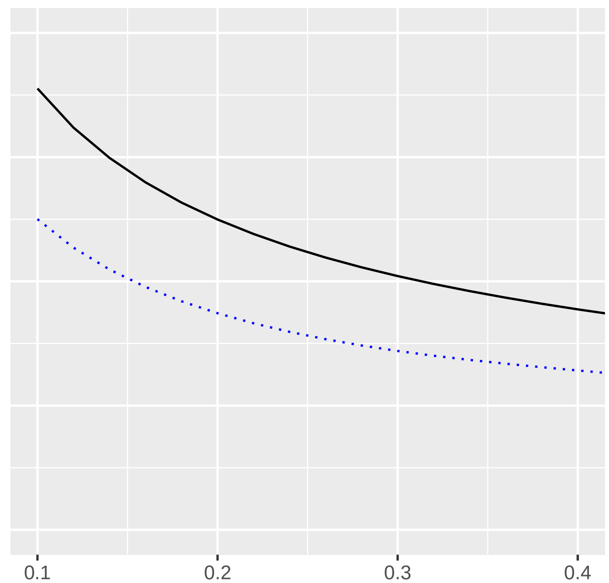


(e) $x = 2/3$

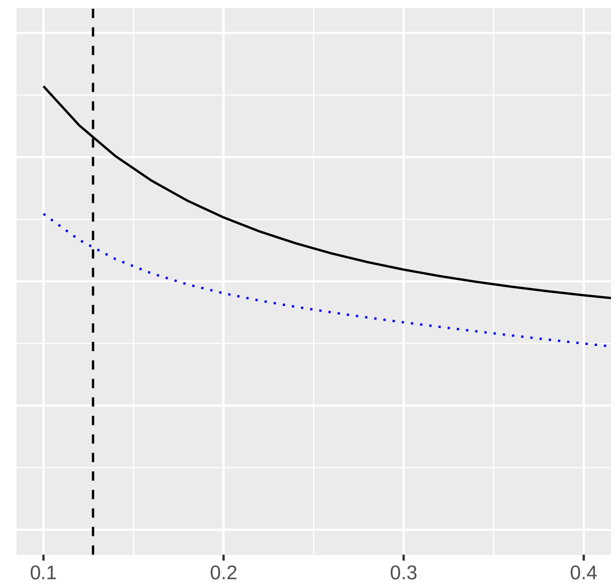
Figure S.II.11: Average Interval Length of 95% Confidence Intervals - Model 4



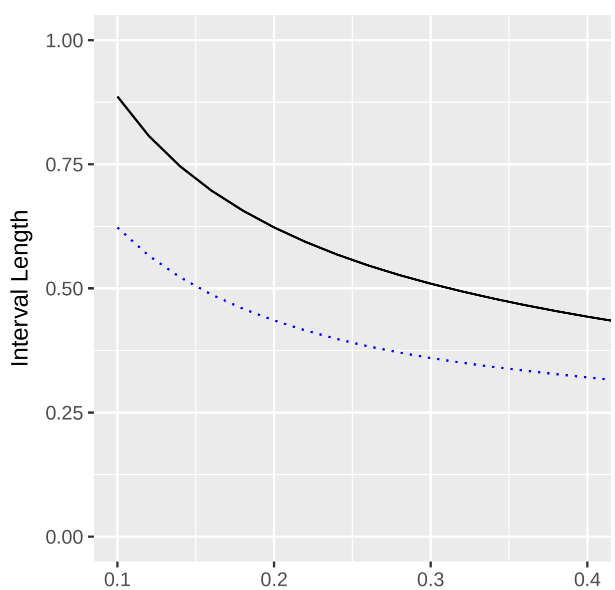
(a) $x = -2/3$



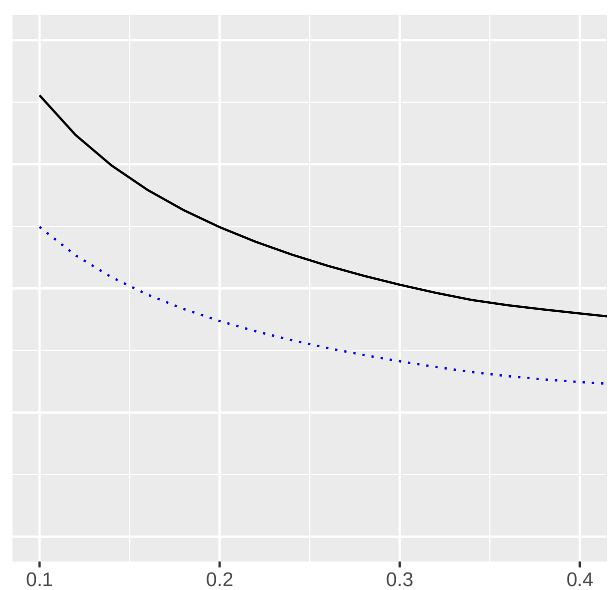
(b) $x = -1/3$



(c) $x = 0$

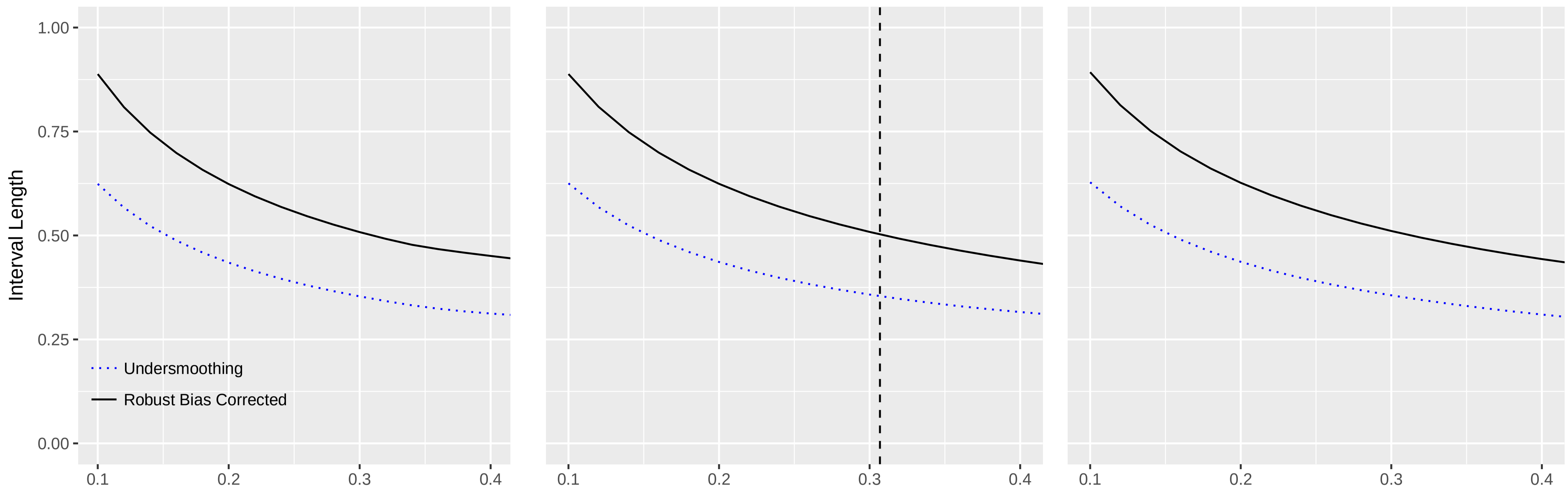


(d) $x = 1/3$

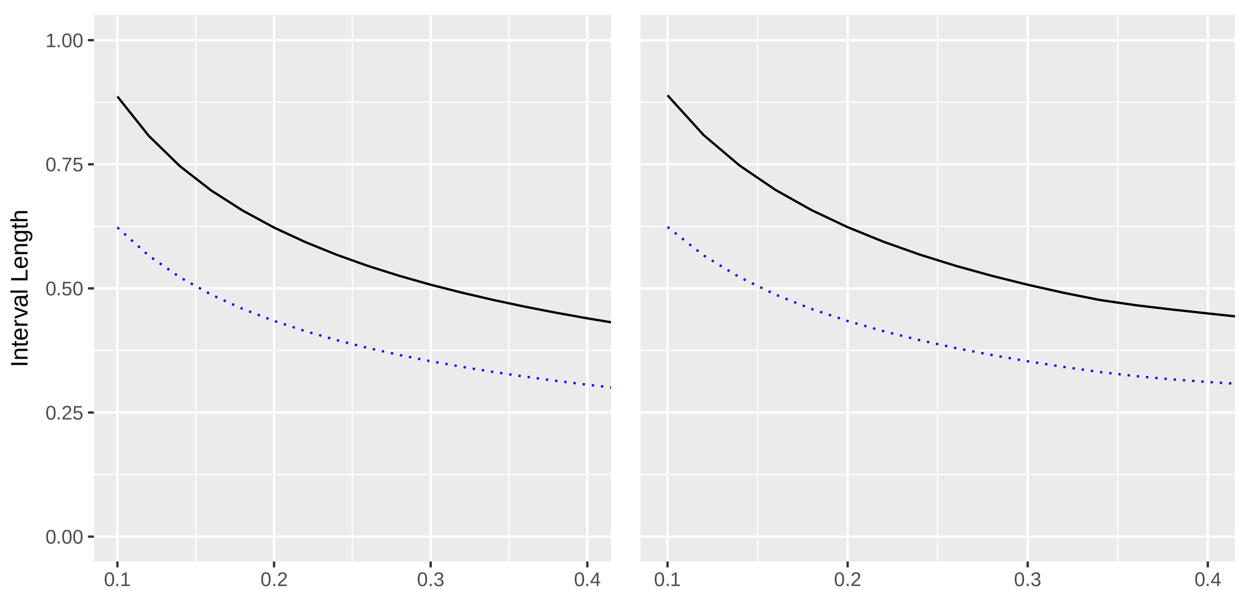


(e) $x = 2/3$

Figure S.II.12: Average Interval Length of 95% Confidence Intervals - Model 5

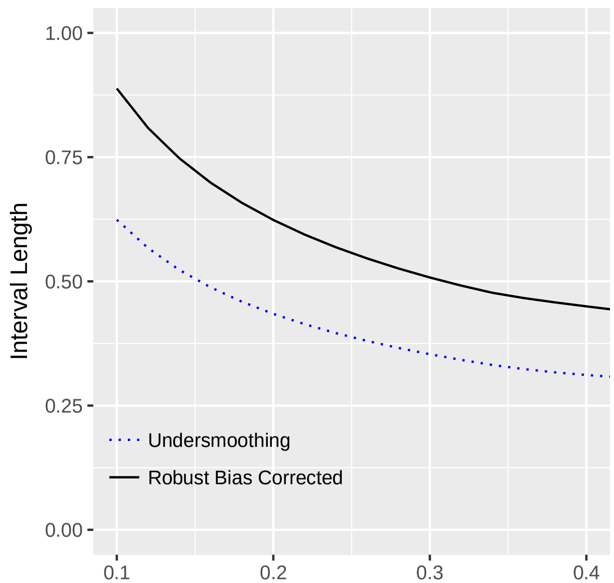


(a) $x = -2/3$ (b) $x = -1/3$ (c) $x = 0$

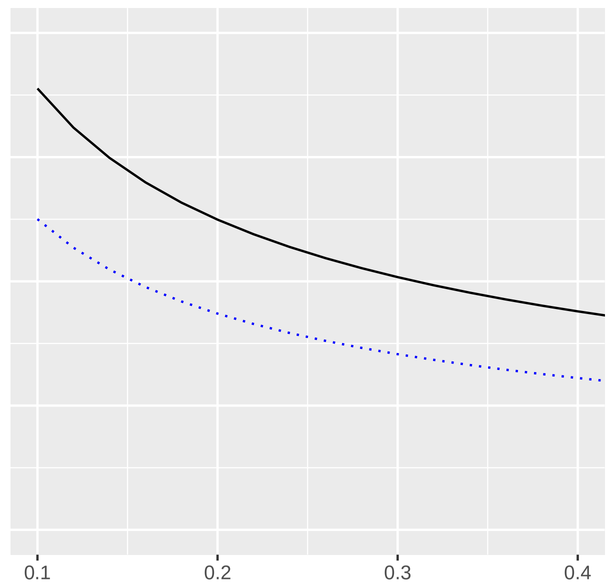


(d) $x = 1/3$ (e) $x = 2/3$

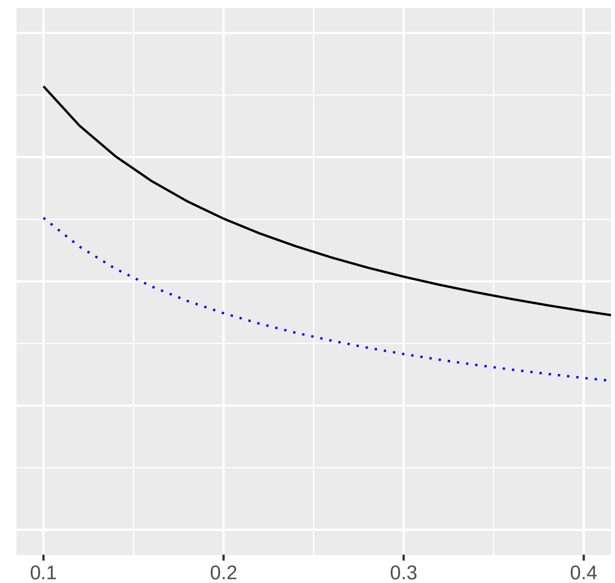
Figure S.II.13: Average Interval Length of 95% Confidence Intervals - Model 6



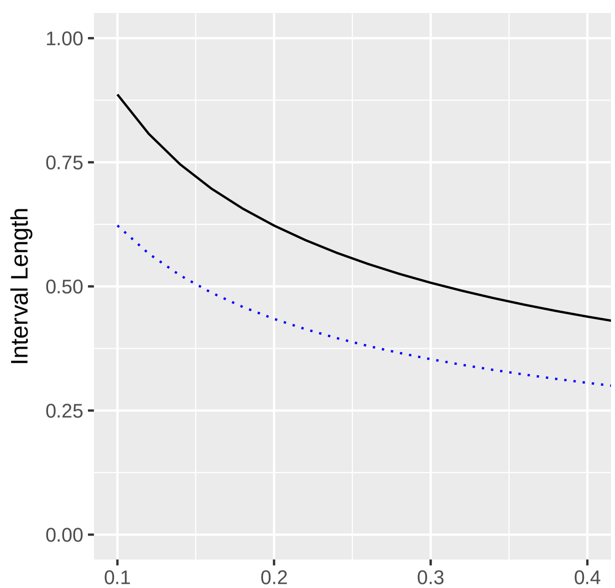
(a) $x = -2/3$



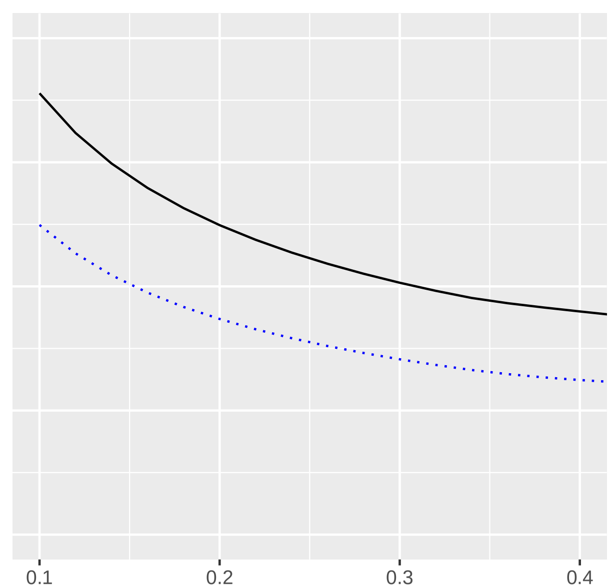
(b) $x = -1/3$



(c) $x = 0$

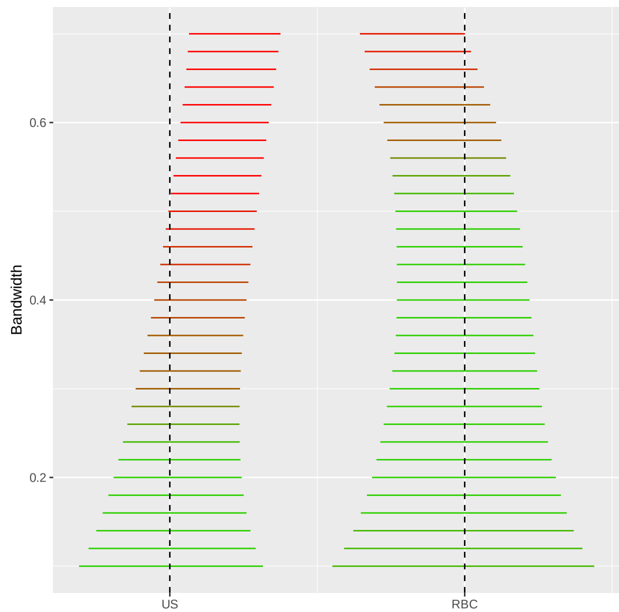


(d) $x = 1/3$

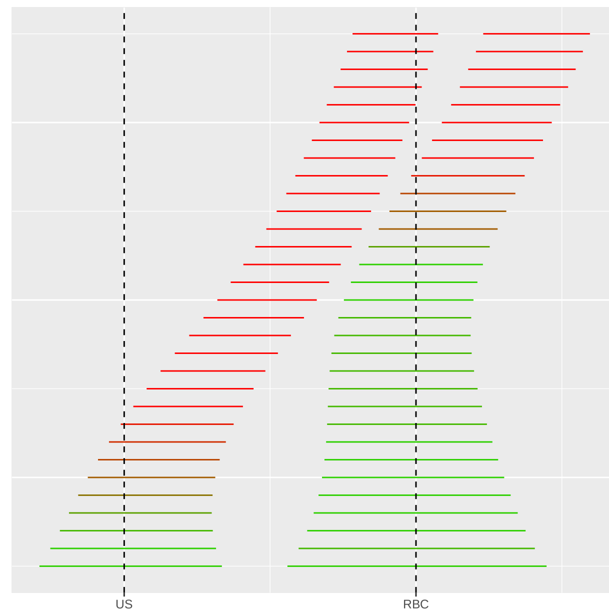


(e) $x = 2/3$

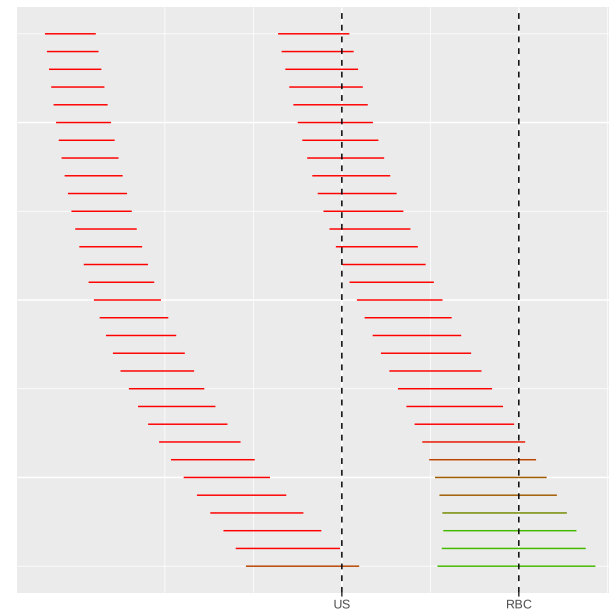
Figure S.II.14: Empirical Coverage and Average Interval Length of 95% Confidence Intervals - Model 1



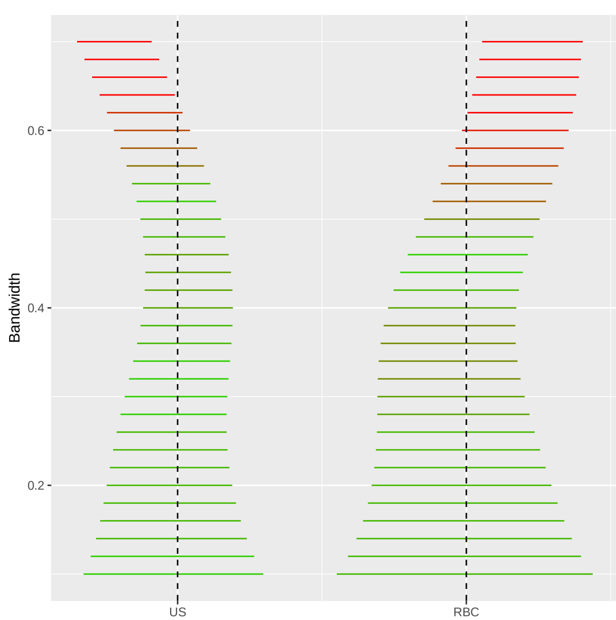
(a) $x = -2/3$



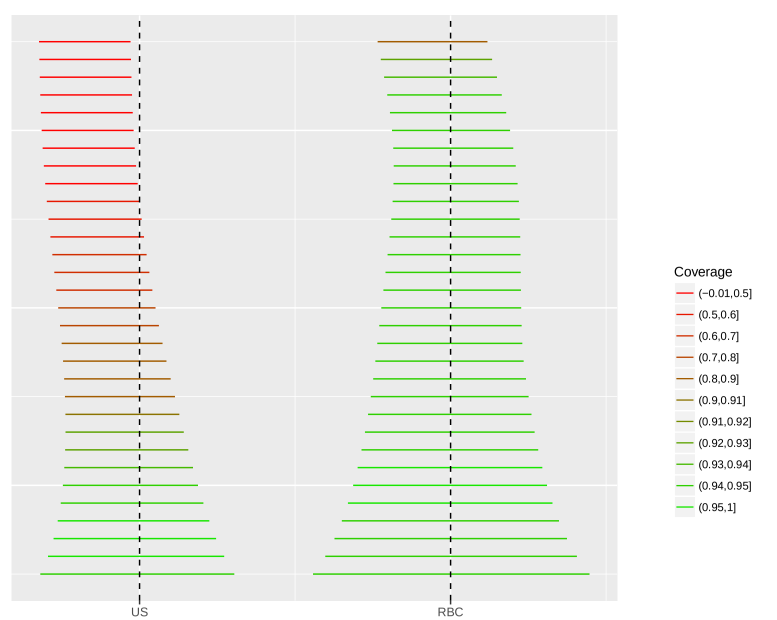
(b) $x = -1/3$



(c) $x = 0$

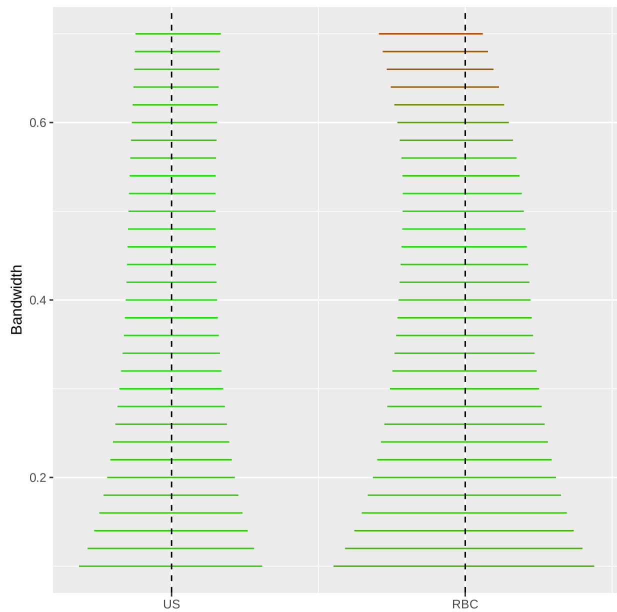


(d) $x = 1/3$

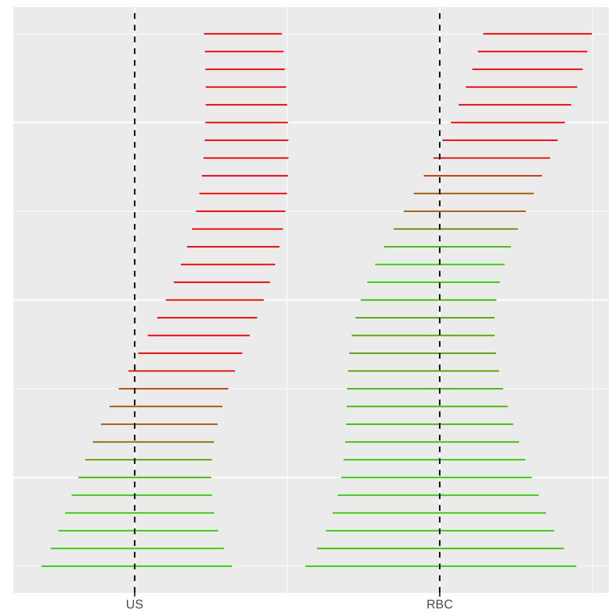


(e) $x = 2/3$

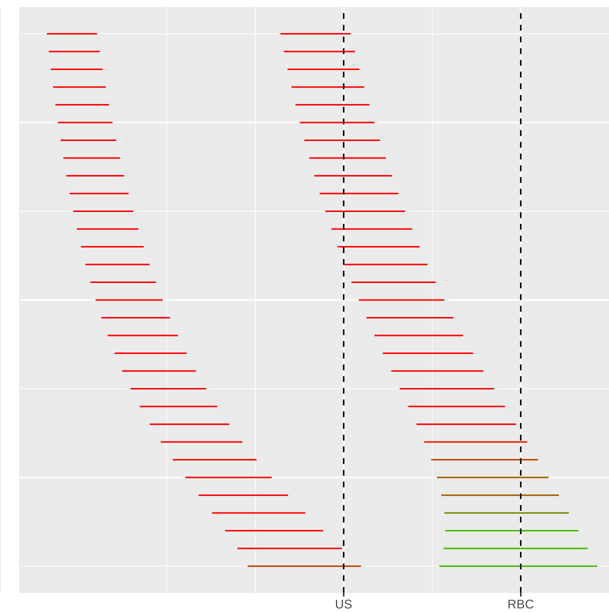
Figure S.II.15: Empirical Coverage and Average Interval Length of 95% Confidence Intervals - Model 2



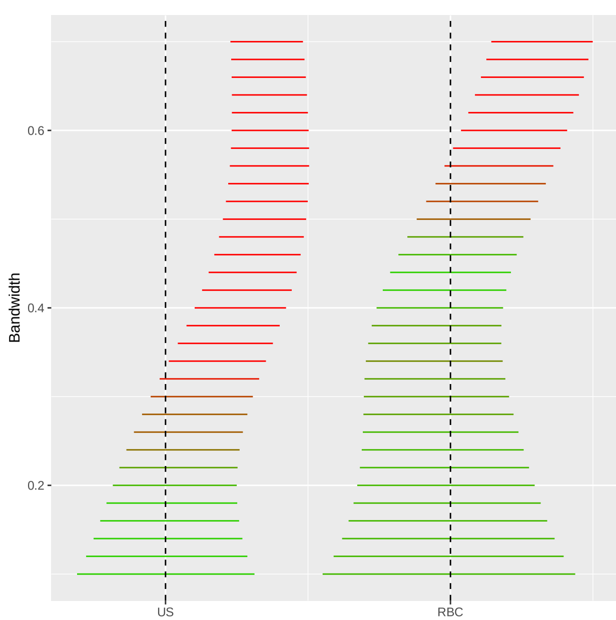
(a) $x = -2/3$



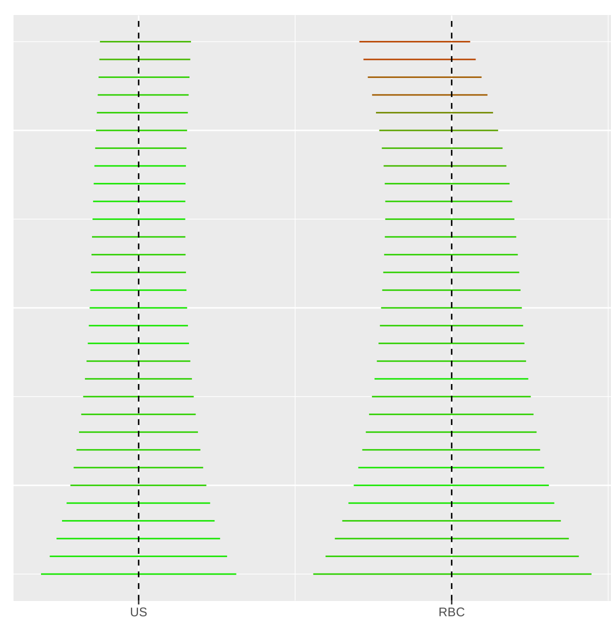
(b) $x = -1/3$



(c) $x = 0$



(d) $x = 1/3$



(e) $x = 2/3$

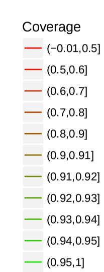
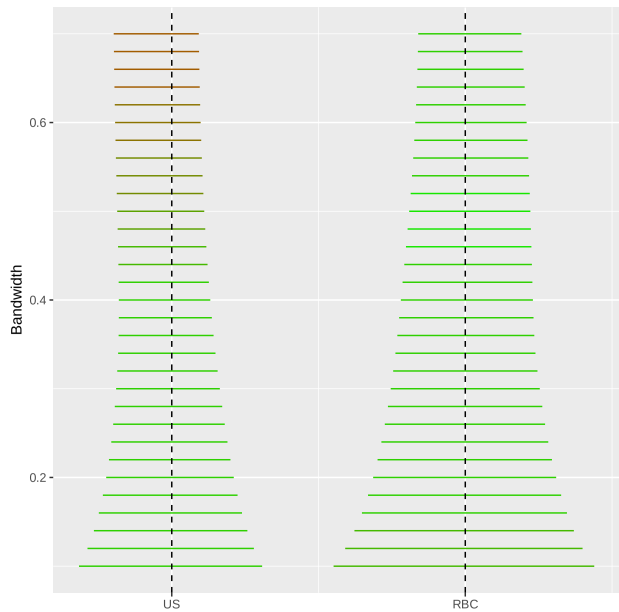
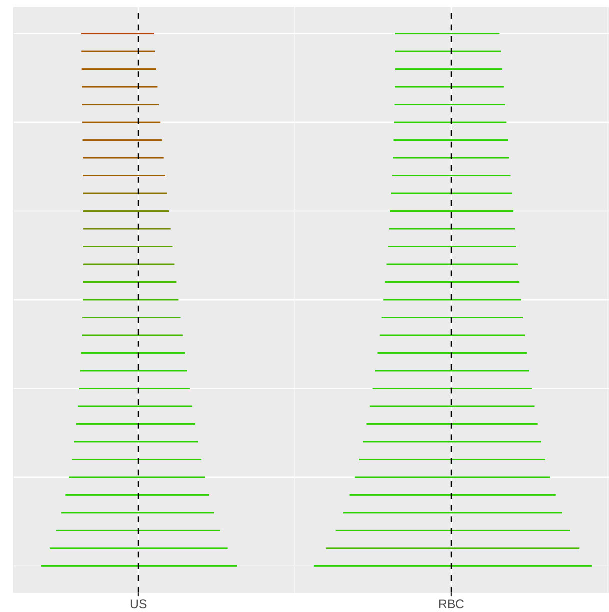


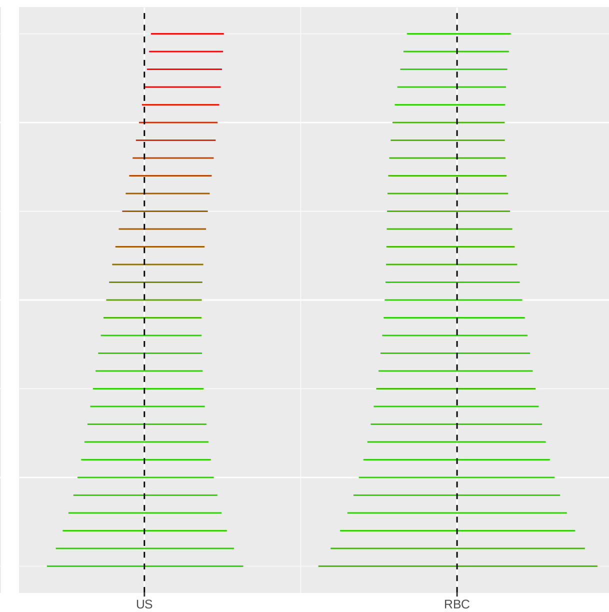
Figure S.II.16: Empirical Coverage and Average Interval Length of 95% Confidence Intervals - Model 3



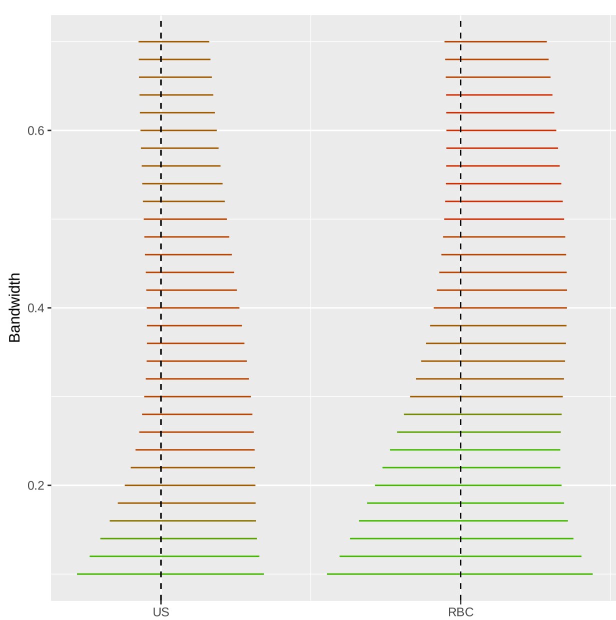
(a) $x = -2/3$



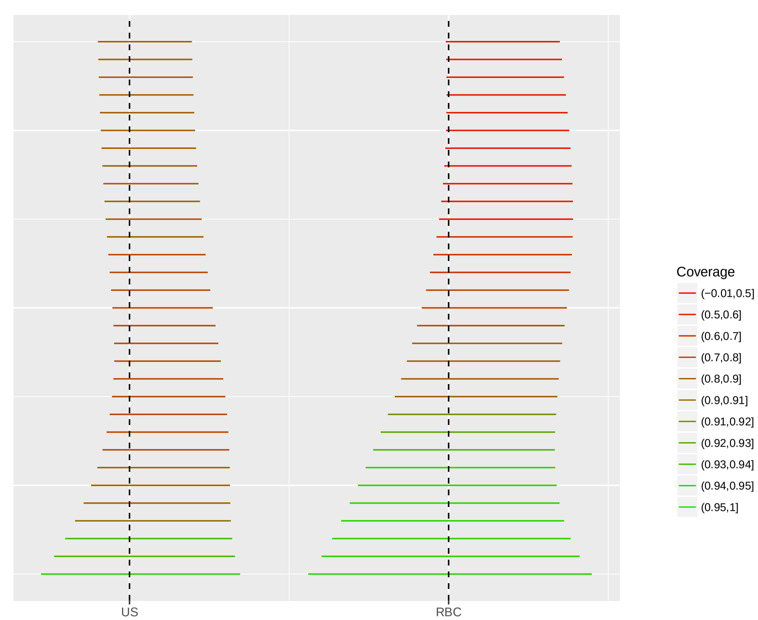
(b) $x = -1/3$



(c) $x = 0$

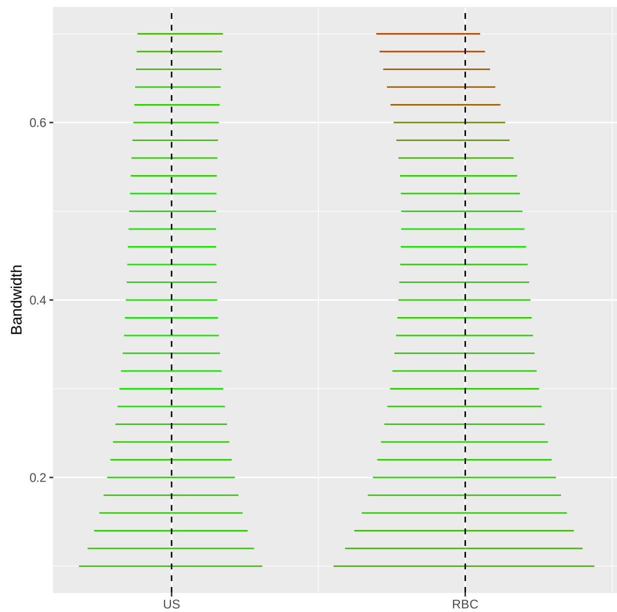


(d) $x = 1/3$

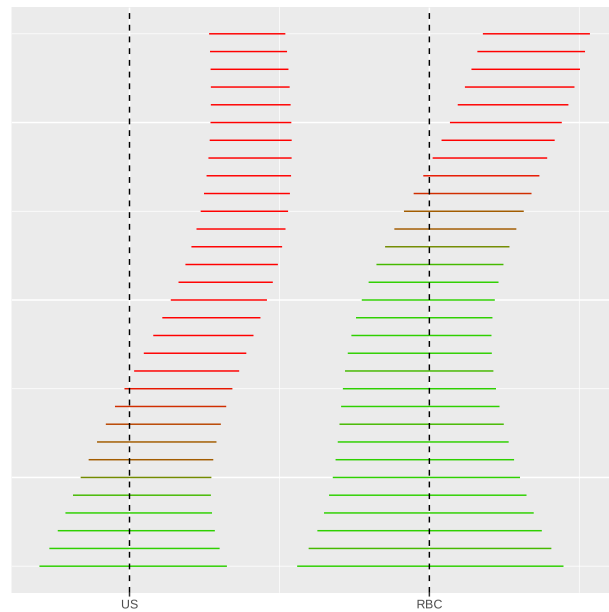


(e) $x = 2/3$

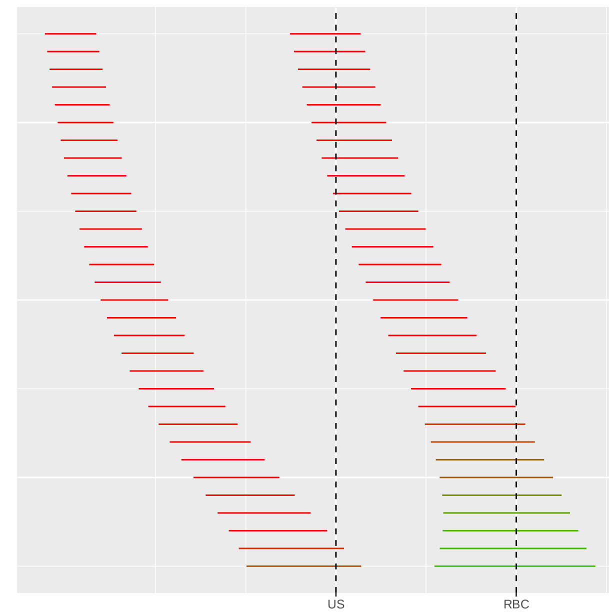
Figure S.II.17: Empirical Coverage and Average Interval Length of 95% Confidence Intervals - Model 4



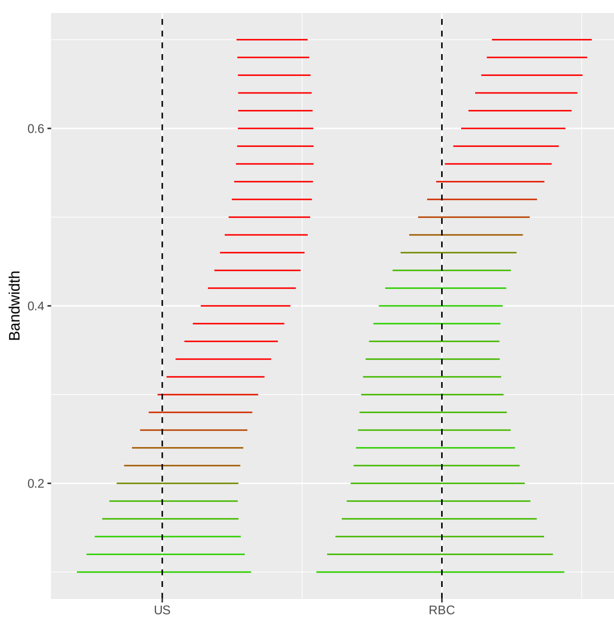
(a) $x = -2/3$



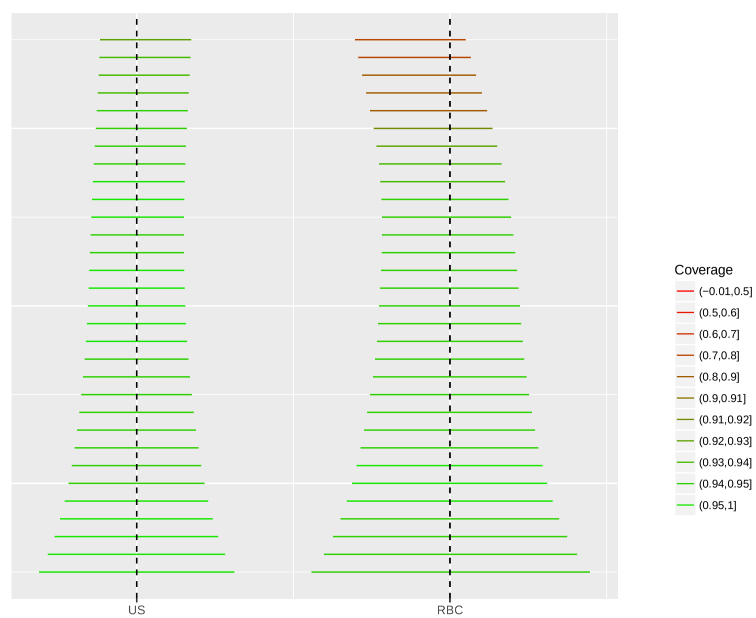
(b) $x = -1/3$



(c) $x = 0$

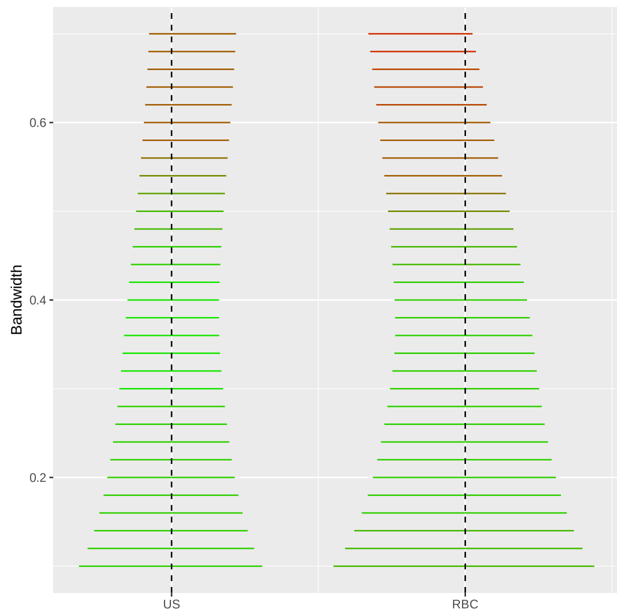


(d) $x = 1/3$

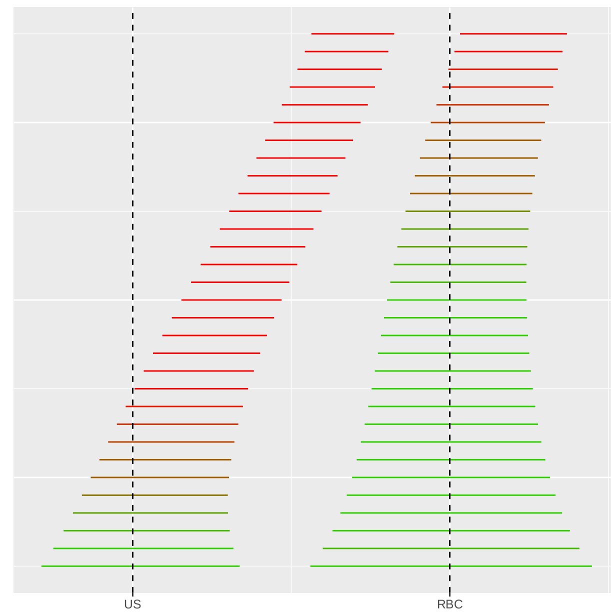


(e) $x = 2/3$

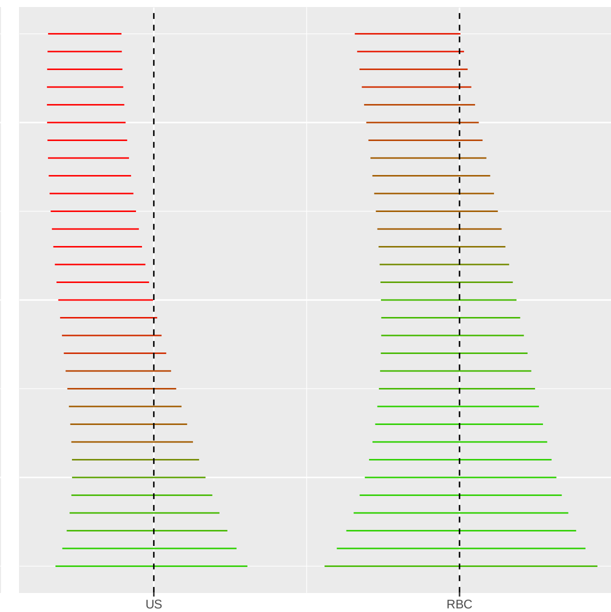
Figure S.II.18: Empirical Coverage and Average Interval Length of 95% Confidence Intervals - Model 5



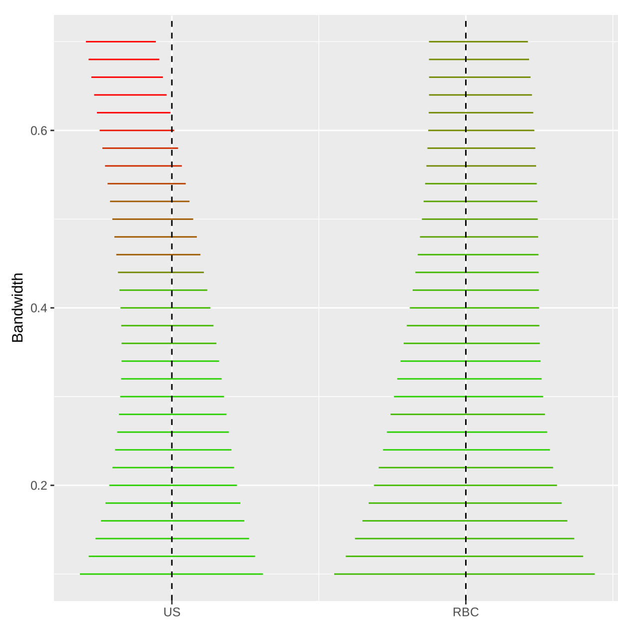
(a) $x = -2/3$



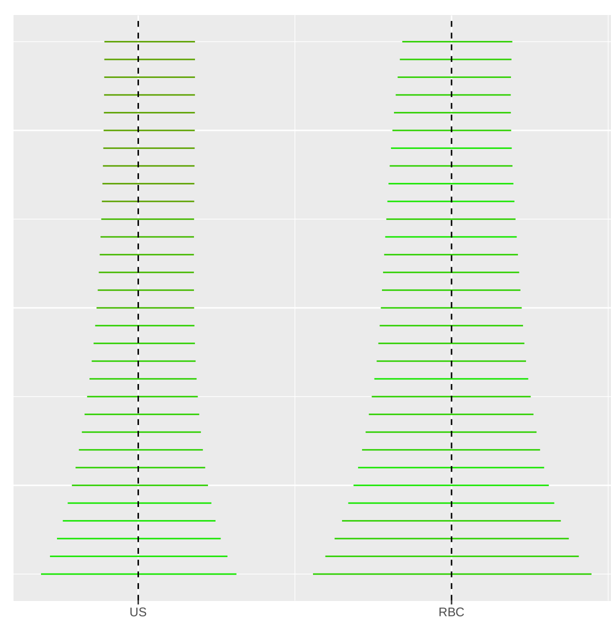
(b) $x = -1/3$



(c) $x = 0$



(d) $x = 1/3$



(e) $x = 2/3$

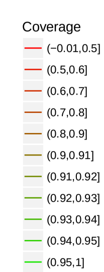
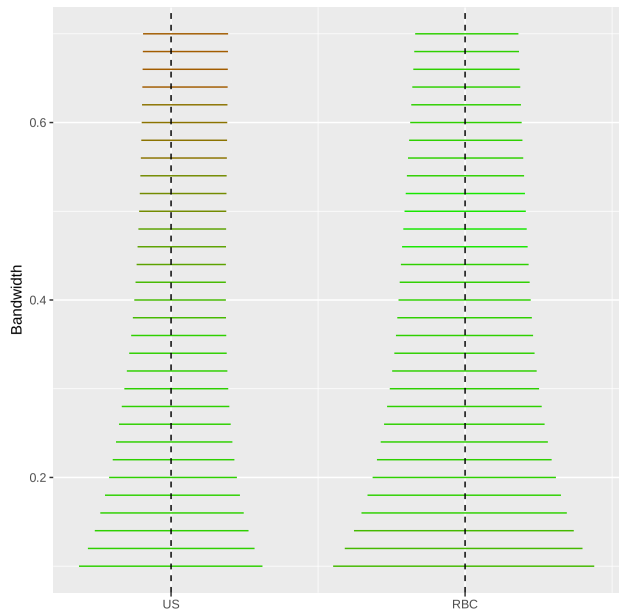
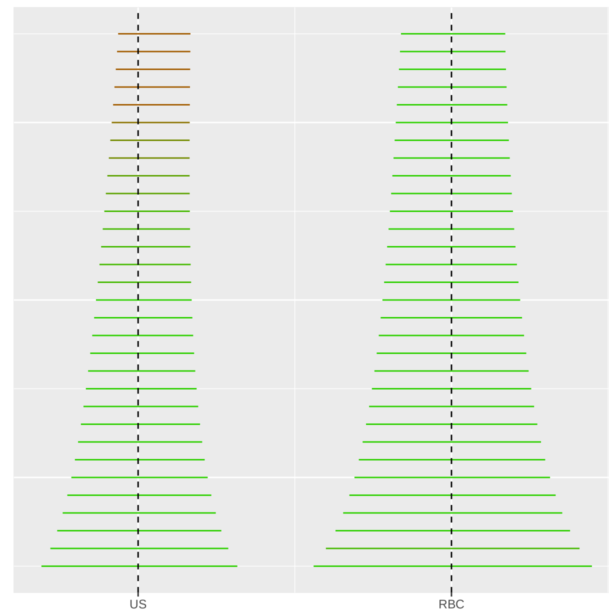


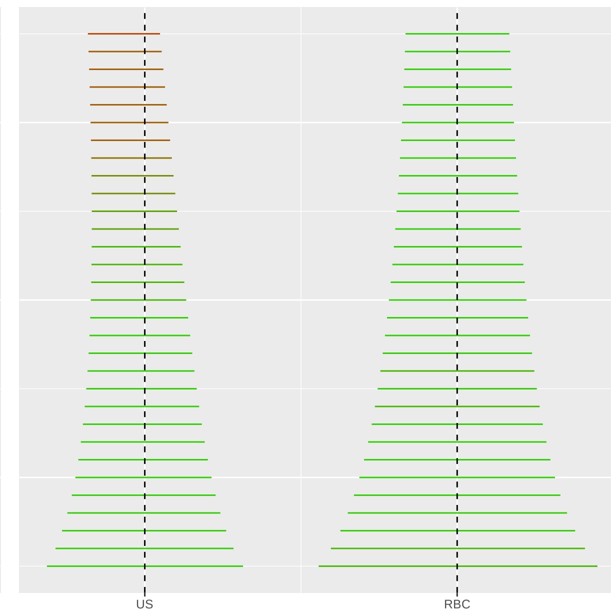
Figure S.II.19: Empirical Coverage and Average Interval Length of 95% Confidence Intervals - Model 6



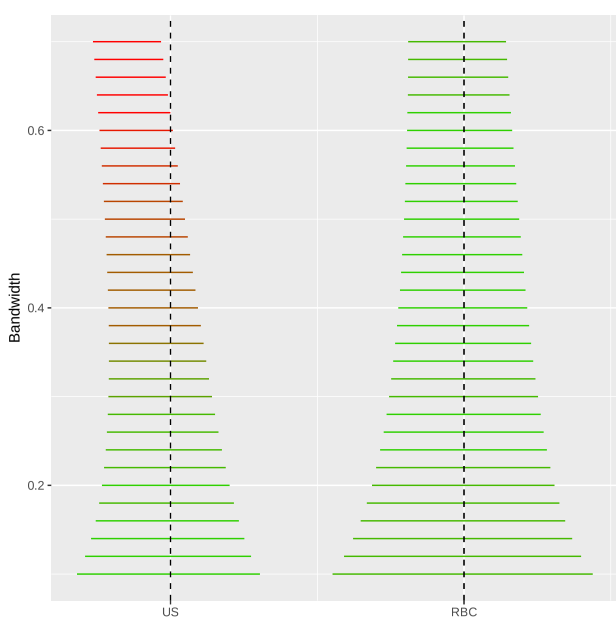
(a) $x = -2/3$



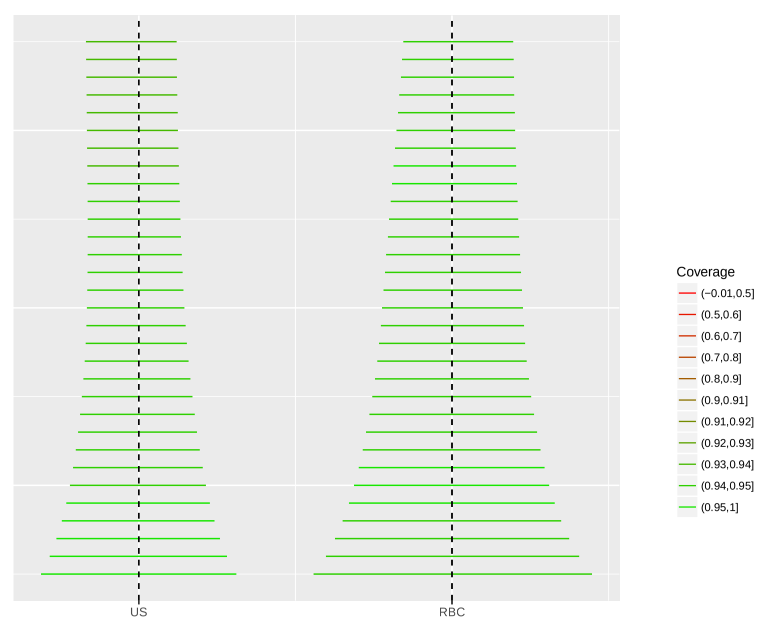
(b) $x = -1/3$



(c) $x = 0$



(d) $x = 1/3$



(e) $x = 2/3$

Part S.III
Supplement References

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

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

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.

Bhattacharya, R. N., and Ghosh, J. K. (1978), "On the Validity of the Formal Edgeworth Expansion," *The Annals of Statistics*, 6, 434–451.

Bhattacharya, R. N., and Rao, R. R. (1976), *Normal Approximation and Asymptotic Expansions*, John Wiley and Sons.

Calonico, S., Cattaneo, M. D., and Farrell, M. H. (2017), "`nprobust`: Nonparametric Kernel-Based Estimation and Robust Bias-Corrected Inference," *working paper*.

Cattaneo, M. D., and Farrell, M. H. (2013), "Optimal Convergence Rates, Bahadur Representation, and Asymptotic Normality of Partitioning Estimators," *Journal of Econometrics*, 174, 127–143.

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

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

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.

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

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

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

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.

James, G. S., and Mayne, A. J. (1962), "Cumulants of Functions of Random Variables," *Sankhyā*, 24, 47–54.

Jones, M. C. (1994), "On Kernel Density Derivative Estimation," *Communications in Statistics - Theory and Methods*, 23, 2133–2139.

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

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

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

Loader, C. (2013), *locfit: Local Regression, Likelihood and Density Estimation.*, R package version 1.5-9.1.

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, Springer, pp. 437–461.

Marron, J. S., and Wand, M. P. (1992), "Exact Mean Integrated Squared Error," *The Annals of Statistics*, 20, 712–736.

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

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

Silverman, B. W. (1986), *Density Estimation for Statistics and Data Analysis*, Chapman & Hall.

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

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